



الجمهورية العربية السورية
جامعة دمشق
كلية الاقتصاد
قسم الإحصاء ونظم المعلومات

نموذج إحصائي مقترح لتقدير القيمة الحقيقية لأسعار العقارات
دراسة تطبيقية في مدينة دمشق.

A Proposed Statistical Model to Estimate The Real Value of The Real Estates

A Practical Study in The City of Damascus.

رسالة مقدمة لنيل درجة الماجستير في الأساليب الكمية

إعداد

محمد نضال بسام السقا

إشراف

د. ميساء حمشو

الأستاذ المساعد في قسم الإحصاء ونظم المعلومات، كلية الاقتصاد، جامعة دمشق

2025-2026



الجمهورية العربية السورية
جامعة دمشق
كلية الاقتصاد
قسم الإحصاء ونظم المعلومات

قرار لجنة الحكم

نموذج إحصائي مقترح لتقدير القيمة الحقيقية لأسعار العقارات دراسة تطبيقية في مدينة دمشق.

رسالة مقدمة لنيل درجة الماجستير في الأساليب الكمية

إعداد: محمد نضال بسام السقا

إشراف: د. ميساء حمشو

نوقشت الرسالة بتاريخ 2025/12/28 وأوصت لجنة الحكم الآتية أسماؤهم، بقبول هذه الرسالة دون تعديل والحكم بالإجماع على أهلية الطالب لنيل درجة الماجستير في الأساليب الكمية / قسم الإحصاء ونظم المعلومات.

الاسم والمرتبة العلمية	الصفة التحكيمية	التوقيع
أ.د. شفيق عربش الأستاذ في قسم الإحصاء ونظم المعلومات / كلية الاقتصاد / جامعة دمشق	عضواً	
أ.م.د. ميساء حمشو الأستاذ المساعد في قسم الإحصاء ونظم المعلومات / كلية الاقتصاد / جامعة دمشق	عضواً مشرفاً	
د. سهير بقله المدرس في قسم الإحصاء ونظم المعلومات / كلية الاقتصاد / جامعة دمشق	عضواً	

إهداء

أُهدي هذا العمل إلى والديّ العزيزين، اللذين كانا الداعم الأول في مسيرتي العلمية، وإلى أسرتي الكريمة التي لم

تدّخر جهداً في تقديم الدعم المعنوي والتشجيع المستمر.

كما أُهدي هذا الجهد إلى كل من آمن بأهمية العلم والبحث العلمي سبيلاً للتقدم، وإلى كل من يسعى إلى تطوير

المعرفة رغم التحديات.

شكر وتقدير

أتقدم بخالص الشكر والتقدير والامتنان إلى الدكتورة ميساء حمشو، المشرفة على هذه الرسالة، لما أولته من عناية علمية ومتابعة دقيقة، وما أبدته من حرص أكاديمي أسهم في تنظيم البحث وإخراجه بالصورة السليمة. ولمن كنت لها بمثابة ابن قبل أن أكون طالباً، فكانت سنداً علمياً وإنسانياً، ومثالاً في التفاني والمسؤولية، وكان لدعمها وتوجيهها الأثر الواضح في إنجاز هذه الرسالة.

كما أتوجه بوافر الشكر إلى الأستاذ الدكتور شفيق عربش، لما تفضل به من إرشادات علمية ورؤى منهجية كان لهما الأثر الواضح في تطوير مضمون الرسالة وتعميق معالجتها العلمية.

كما أتقدم بخالص الشكر والتقدير إلى الدكتورة سهير بقله، على تفضلها بتحكيم هذه الرسالة، وما قدّمته من ملاحظات علمية قيّمة أسهمت في تعزيز جودة العمل وصقل مضمونه الأكاديمي.

ولا يفوتني أن أتقدم بالشكر والتقدير إلى أساتذتي في قسم الإحصاء التطبيقي، لما قدّموه من علم وخبرة أسهمت في بناء الأساس المعرفي والمنهجي الذي استندت إليه هذه الدراسة.

إلى والدي، يا من يقف دعاؤكما وراء نجاحي وتوفيقي. رمز العطاء اللامتناهي.

إلى شقيقتي الغاليات، من أشارك معهن لحظات حياتي، ولوجودهن الأثر الكبير في دعمي ومساندتي.

الملخص

تُعد عملية تقدير أسعار العقارات من المسائل المعقدة في الدراسات الاقتصادية والإحصائية، نظراً لتعدد العوامل المؤثرة في القيمة السوقية للعقار، وصعوبة جمعها وتحليلها ضمن نموذج دقيق. وقد هدفت هذه الدراسة إلى بناء نموذج إحصائي لتقدير أسعار العقارات السكنية باستخدام أسلوب انحدار لاسو (LASSO Regression)، مع إجراء مقارنة منهجية بينه وبين نموذج الانحدار الخطي المتعدد (OLS)، لتقييم أداء كل منهما.

واعتمدت الدراسة في مرحلتها الأولى على بيانات مدينة Ames الأمريكية، التي تُعد مرجعاً في أبحاث تحليل أسعار العقارات، لاحتوائها على عدد كبير من المتغيرات المدروسة والموثقة أكاديمياً. وقد أتاح ذلك اختبار قدرة انحدار لاسو على اختيار المتغيرات الأكثر تأثيراً في السعر ضمن بيئة مدروسة، وأظهرت نتائجها دقة هذا الأسلوب في تقليص عدد المتغيرات المدخلة إلى النموذج مع الحفاظ على جودة التنبؤ.

وفي المرحلة الثانية، طُبّق النموذج على بيانات ميدانية من سوق العقارات في مدينة دمشق خلال عام 2024، وهي بيانات جُمعت من مصادر ميدانية غير رسمية، في ظل ندرة البيانات العقارية العامة في سوريا. وبعد إجراء المعالجات اللازمة، بلغ عدد المشاهدات المستخدمة في التحليل 250 مشاهدة، شملت متغيرات كمية ونوعية تتعلق بمواصفات العقار وموقعه وتجهيزه.

وقد أظهرت النتائج التطبيقية تفوق انحدار لاسو في تحقيق توازن أفضل بين دقة النموذج وبساطته مقارنةً بنموذج الانحدار الخطي المتعدد، إذ احتفظ بعدد محدود من المتغيرات المؤثرة (سبعة فقط). كما بيّن التحليل أن المتغيرات الأكثر تأثيراً في تحديد السعر ضمن مدينة دمشق تمثلت في: المساحة، جودة الإكساء، وجود الأثاث، وجود الشرفة، عدد الغرف، وجود المصعد، والارتفاع الطابق.

تُبرز هذه النتائج أهمية استخدام النماذج الإحصائية مثل انحدار لاسو، في حالات تعدد المتغيرات وندرة البيانات، نظراً لقدرتها على تقديم نموذج بسيط، وهو ما يعزز استخدامها كأداة مساندة في التقييم العقاري.

الكلمات المفتاحية: أسعار العقارات، انحدار لاسو، الانحدار الخطي المتعدد، التقييم العقاري، اختيار المتغيرات، خوارزمية الهبوط الإحداثي، معامل العتبة الناعمة.

فهرس المحتويات

XVIII.....	الفصل الأول الإطار التمهيدي للدراسة
XIX.....	1.1 مشكلة البحث
XIX.....	1.2 أهمية البحث
XX	1.3 أهداف البحث
XX	1.4 فرضيات البحث
XXI.....	1.5 منهج البحث
XXI.....	1.6 متغيرات البحث
XXII	1.7 الدراسات السابقة
XXV.....	1.8 هيكلية البحث
2	الفصل الثاني الأسس الرياضية للنماذج المستخدمة
3	2.1 المبحث الأول طريقة تقدير المربعات الصغرى (Least Square)
3	2.1.1 نموذج الانحدار الخطي
4	2.1.2 طريقة تقدير المربعات الصغرى (OLS)
5	2.1.3 خصائص (صفات) مقدر المربعات الصغرى
5	2.1.4 بعض الاشكاليات المحتملة أثناء التقدير
8	2.2 المبحث الثاني أسلوب انحدار لاسو (LASSO)
8	2.2.1 مفاهيم تمهيدية
11	2.2.2 طرائق الانكماش (Shrinkage methods)
11	2.2.3 ما هو انحدار لاسو (مقدر لاسو)
14	2.2.4 صيغة أخرى لانحدار لاسو
15	2.2.5 تقدير انحدار لاسو

2.2.6	تقدير انحدار لاسو لمتغير مستقل واحد فقط	16
2.2.7	تقدير انحدار لاسو في الحالة الخاصة التصميم المتعامد (Orthogonal design)	19
2.2.8	تقدير انحدار لاسو عند وجود أكثر من متغير مستقل	21
2.2.9	تقدير نموذج الانحدار الخطي عن طريق خوارزمية التراجع الإحداثي	22
2.2.10	تقدير انحدار لاسو عن طريق خوارزمية التراجع الإحداثي	24
2.2.11	التفسير الهندسي لمقدر لاسو	25
2.2.12	مزايا ومساوئ مقدر لاسو	28
2.3	المبحث الثالث اختيار معلمة الضبط	29
2.3.1	طريقة التحقق المتقاطع (Cross-Validation)	29
2.3.2	التحقق المتقاطع بأسلوب (K-Folds Cross-Validation)	30
2.3.3	الأخطاء المعيارية للتحقق المتقاطع (K-Folds Cross-Validation)	31
2.3.4	طرائق معايير المعلومات (Information Criteria)	31
35	الفصل الثالث بناء النماذج الإحصائية لتقدير أسعار العقارات	35
3.1	المبحث الأول تطبيق انحدار لاسو لتقدير أسعار العقارات في مدينة Ames	36
	تمهيد	36
3.1.1	الأدوات والبرامج الإحصائية المستخدمة	36
3.1.2	متغيرات الدراسة	37
3.1.3	معالجة البيانات	38
3.1.4	عرض لأهم متغيرات الدراسة	41
3.1.5	الإحصاء الوصفي للمتغير التابع (سعر العقار)	41
3.1.6	مصفوفة الارتباط بين متغير سعر العقار والمتغيرات الكمية	42
3.1.7	معالجة القيم الشاذة	56
3.1.8	بناء النموذج الإحصائي للعلاقة بين متغيرات الدراسة	58

63	خاتمة
64	3.2 المبحث الثاني طرائق التقييم العقاري.....
64	3.2.1 تعريف التقييم العقاري
64	3.2.2 طرائق التقييم العقاري
66	أ. الطريقة التقليدية
66	ب. الطريقة الحديثة
	3.3 المبحث الثالث تطبيق انحدار لاسو لتقدير أسعار العقارات في مدينة دمشق ومقارنته بانحدار OLS
67	
67	تمهيد
67	3.3.1 مجتمع وحدود الدراسة.
68	3.3.2 جمع البيانات من السوق العقاري في مدينة دمشق
68	3.3.3 مصادر البيانات والمنهجية المتبعة
68	3.3.4 المتغيرات المعتمدة في التحليل
69	3.3.5 التحديات والصعوبات
70	3.3.6 معالجة البيانات وتحضيرها للتحليل
71	3.3.7 الأدوات والبرامج الإحصائية المستخدمة
72	3.3.8 التحليل الوصفي لمتغيرات الدراسة.
78	3.3.9 بناء النموذج الإحصائي للعلاقة بين الخصائص العقارية والسعر
81	3.3.10 اختبار افتراضات النموذج
	3.3.11 بناء النموذج الإحصائي للعلاقة بين متغيرات الدراسة الخاصة بمدينة دمشق باستخدام انحدار لاسو
85	لاسو
89	3.3.12 التقييم التطبيقي لأداء النموذج التنبؤي باستخدام عينة من العقارات
90	3.3.13 المقارنة التطبيقية بين أداء انحدار لاسو والانحدار الخطي المتعدد

92	 مناقشة فرضيات الدراسة
93	 4.1 النتائج والتوصيات
93	 أولاً نتائج الدراسة
94	 ثانياً التوصيات
94	 ثالثاً الخلاصة
96	 المراجع ..
101	 الملاحق ..

فهرس الجداول

- الجدول (1): الأدوات والمكتبات الإحصائية المستخدمة في تحليل البيانات الخاصة بمدينة Ames. 37
- الجدول (2): نسب القيم المفقودة لمتغيرات الدراسة الخاصة في مدينة Ames. 39
- الجدول (3): الإحصاء الوصفي لمتغير سعر العقار الخاصة في مدينة Ames. 41
- الجدول (4): نتائج تقدير انحدار لاسو على بيانات التدريب والاختبار الخاصة ببيانات Ames. 61
- الجدول (5): الأدوات والمكتبات الإحصائية المستخدمة في تحليل البيانات الخاصة بمدينة دمشق. 71
- الجدول (6): الإحصاءات الوصفية للمتغيرات الكمية في عينة الدراسة الخاصة لمدينة دمشق. 72
- الجدول (7): توزيع المتغيرات الثنائية وفق الفئات (1/0) من حيث التكرار والنسبة المئوية الخاصة بمدينة دمشق. 76
- الجدول (8): نتائج معامل الارتباط بيرسون بين سعر العقار والمتغيرات المستقلة الخاصة في مدينة دمشق. 78
- الجدول (9): نتائج نموذج الانحدار الخطي المتعدد للعلاقة بين أسعار العقارات والخصائص المتعلقة بها في مدينة دمشق. 79
- الجدول (10): نتائج اختبار شابيرو - ويلك (Shapiro-Wilk) للتحقق من التوزيع الطبيعي للبواقي لنموذج الانحدار الخطي المتعدد الخاص لبيانات مدينة دمشق. 81
- الجدول (11): نتائج اختبار بريوش - بيغان (Breusch-Pagan) للكشف عن تجانس التباين لنموذج الانحدار الخطي المتعدد الخاص لبيانات مدينة دمشق. 83
- الجدول (12): قيم معامل تباين التضخم (VIF) للمتغيرات المستقلة لنموذج الانحدار الخطي المتعدد الخاص لبيانات مدينة دمشق. 84
- الجدول (13): نتائج تقدير انحدار لاسو على بيانات التدريب والاختبار على مدينة دمشق. 88
- الجدول (14): مقارنة بين السعر المعلن والسعر المُقدَّر من انحدار لاسو الخاص ببيانات مدينة دمشق. ... 90

فهرس الأشكال

- الشكل (1) توضيح بياني للتحيز والتباين. 10
- الشكل (2) المعلومات المعيارية المقدرة من خلال انحدار لاسو كدالة في قيم λ 13
- الشكل (3): التمثيل البياني للمعادلة رقم (2.11). 18
- الشكل (4): العلاقة بين مقدر لاسو في حالة وجود متغير مستقل واحد حولت قيمه إلى قيم معيارية، وبين معامل soft thresholding، يمثل الخط العمودي المقطع تقدير المربعات الصغرى، والخط العمودي المتصل تقدير لاسو. 19
- الشكل (5): التفسير الهندسي لانحدار لاسو - حالة حذف أحد المتغيرات. 26
- الشكل (6): التفسير الهندسي لانحدار لاسو - وجود حلين محتملين أحدهما يحذف متغيراً. 27
- الشكل (7): التوزيع التكراري لمتغير سعر العقار في مدينة Ames مع منحنى التوزيع الطبيعي التقريبي. 42
- الشكل (8): خريطة حرارية (Heatmap) لمصفوفة الارتباط بين المتغيرات الكمية وسعر العقار في بيانات Ames. 43
- الشكل (9) شكل الانتشار لجميع المتغيرات الكمية مع متغير سعر العقار. 45
- الشكل (10): التوزيع المشترك وشكل الانتشار لمتغير سعر مبيع العقار مع مساحة الطابق الأول لمدينة Ames. 46
- الشكل (11): التوزيع المشترك وشكل الانتشار لمتغير سعر مبيع العقار وعدد الحمامات الموجودة في الطوابق العلوية في مدينة Ames. 47
- الشكل (12): التوزيع المشترك وشكل الانتشار لمتغير سعر مبيع العقار ومساحة المرآب في مدينة Ames. 48
- الشكل (13): التوزيع المشترك وشكل الانتشار لمتغير سعر مبيع العقار وسعة المرآب من السيارات في مدينة Ames. 49
- الشكل (14): التوزيع المشترك وشكل الانتشار لمتغير سعر مبيع العقار ومساحة العقار الإجمالية في مدينة Ames. 50
- الشكل (15): التوزيع المشترك وشكل الانتشار لمتغير سعر مبيع العقار ومتغير الجودة الشاملة في مدينة Ames. 51
- الشكل (16): التوزيع المشترك وشكل الانتشار لمتغير سعر مبيع العقار مع إجمالي مساحة القبو في مدينة Ames. 52

- الشكل (17): التوزيع المشترك وشكل الانتشار لمتغير سعر مبيع العقار وإجمالي عدد الغرف فوق الطابق الأرضي. 53
- الشكل (18): التوزيع المشترك وشكل الانتشار لمتغير سعر مبيع العقار وتاريخ البناء الأصلي في مدينة Ames. 54
- الشكل (19): التوزيع المشترك وشكل الانتشار لمتغير سعر مبيع العقار وبين تاريخ التجديد في مدينة Ames. 55
- الشكل (20): شكل الانتشار لمتغير سعر مبيع العقار ومتغير المسافة بين العقار والشارع (LotFrontage). 57
- الشكل (21): شكل الانتشار لمتغير سعر مبيع العقار ومتغير مساحة أرض العقار (LotArea). 57
- الشكل (22): شكل الانتشار بين قيمة معلمة الضبط الخاصة بانحدار لاسو مقابل قيمة RMSE الخاص ببيانات مدينة Ames. 59
- الشكل (23): منحنيات معاملات انحدار لاسو كدالة في معلمة الضبط λ الخاصة ببيانات مدينة Ames. 60
- الشكل (24): عدد المتغيرات الداخلة في النموذج كدالة في معلمة الضبط λ الخاصة ببيانات مدينة Ames. 60
- الشكل (25): قيم معاملات إنحدار لاسو لمدينة Ames. 63
- الشكل (26): التوزيع البياني لمتغير المساحة الإجمالية في مدينة دمشق. 73
- الشكل (27): التوزيع البياني لمتغير السعر (بالليرة السورية) في مدينة دمشق. 73
- الشكل (28): التوزيع البياني لمتغير جودة الإكساء في مدينة دمشق. 74
- الشكل (29): التوزيع البياني لمتغير عدد الاتجاهات في مدينة دمشق. 74
- الشكل (30): التوزيع البياني لمتغير الارتفاع الطائفي في مدينة دمشق. 75
- الشكل (31): التوزيع البياني لمتغير عدد الغرف في مدينة دمشق. 75
- الشكل (32): التوزيع البياني لمتغير الشرفة في مدينة دمشق. 77
- الشكل (33): التوزيع البياني لمتغير المصعد في مدينة دمشق. 77
- الشكل (34): التوزيع البياني لمتغير الأثاث في مدينة دمشق. 77
- الشكل (35): التوزيع البياني لمتغير المرآب في مدينة دمشق. 77
- الشكل (36): التوزيع البياني لمتغير التدفئة في مدينة دمشق. 77
- الشكل (37): التوزيع البياني لمتغير الموقع في مدينة دمشق. 77

- الشكل (38): المدرج التكراري للبواقي مع منحنى الكثافة الاحتمالية لنموذج الانحدار الخطي المتعدد لبيانات مدينة دمشق. 82
- الشكل (39): مخطط Q-Q للبواقي لمقارنة التوزيع الفعلي بالتوزيع الطبيعي النظري لنموذج الانحدار الخطي المتعدد لبيانات مدينة دمشق. 82
- الشكل (40): مخطط البواقي مقابل القيم المتوقعة (Residuals vs. Fitted Values) لنموذج الانحدار الخطي المتعدد لبيانات مدينة دمشق. 84
- الشكل (41): شكل الانتشار بين قيمة معلمة الضبط الخاصة بانحدار لاسو مقابل قيمة RMSE الخاصة ببيانات دمشق. 86
- الشكل (42): منحنيات معاملات انحدار لاسو كدالة في معلمة الضبط λ الخاصة ببيانات مدينة دمشق. 87 ..
- الشكل (43): عدد المتغيرات الداخلة في النموذج كدالة في معلمة الضبط λ الخاصة ببيانات مدينة دمشق. 87

قائمة المعادلات

رقم المعادلة	وصف المعادلة	رقم الصفحة
2.1	معادلة الانحدار الخطي المتعدد	3
2.2	معادلة الانحدار الخطي المتعدد في الصيغة المصفوفية	4
2.3	معادلة مجموع مربعات الأخطاء (RSS) في الصيغة المصفوفية	4
2.4	معادلة مقدر معاملات الانحدار الخطي المتعدد بطريقة المربعات الصغرى الاعتيادية (OLS)	5
2.5	معادلة التباين-التغاير لمقدر معاملات الانحدار الخطي المتعدد بطريقة المربعات الصغرى الاعتيادية	5
2.6	معادلة تباين الخطأ المقدر	6
2.7	معادلة متوسط مربعات الأخطاء (MSE)	9
2.8	العلاقة بين متوسط مربعات الخطأ والتباين والتحيز	9
2.9	متوسط مربعات الخطأ المتوقع عند نقطة اختبار جديدة x_0	9
2.10	معادلة تقدير معاملات انحدار لاسو	12
2.11	معادلة حساب الثابت β_0 بعد تقدير معاملات النموذج، عند تحويل القيم إلى قيم معيارية	14
2.12	الصيغة المقيدة لانحدار لاسو	14

15	الصيغة المصفوفية لانحدار لاسو	2.13
15	العلاقة بين مقدر لاسو ومقدر المربعات الصغرى الاعتيادية	2.14
15	دالة الإشارة	2.15
16	مقدّر لاسو باستخدام معامل العتبة الناعمة (Soft-thresholding)	2.16
17	معامل العتبة الناعمة Soft-thresholding) (operator	2.17
24	معادلة الانحدار الخطي عن طريق خوارزمية التراجع الإحداثي	2.18
25	معادلة انحدار لاسو عن طريق خوارزمية التراجع الإحداثي	2.19
31	معادلة الانحراف المعياري لمتوسط MSE	2.20
31	معادلة تقدير الخطأ المعياري للتحقق المتقاطع	2.21
32	معياري أكايكي (AIC) المعدل لانحدار لاسو	2.22
32	معياري شوارتز (BIC) المعدل لانحدار لاسو	2.23
32	صيغة ثانية لمعيار شوارتز (BIC) المعدل لانحدار لاسو	2.24

قائمة المصطلحات

المصطلح باللغة العربية	المصطلح باللغة الإنجليزية
الانحدار التدريجي	Stepwise Regression
الانحدار الخطي	Linear Regression
الانحدار الخطي المتعدد	Multilinear Linear Regression
الانحدار المنتظم/المقيد	Regularized Regression
الانكماش	Shrinkage
البرمجة التربيعية	Quadratic Programming
البواقي	Residuals
التباين	Variance
المثلوية المحدبة	Convex Optimization
التحقق المتقاطع	Cross-Validation
التحيز	Bias
التعدد الخطي	Multicollinearity
التنبؤ خارج العينة	Out-of-Sample Forecasting
التوزيع الطبيعي	Normal Distribution
الجذر التربيعي لمتوسط مربعات الأخطاء	Root Mean Squared Error (RMSE)
الحلول الجبرية المغلقة	Closed-form Solutions
الخريطة الحرارية	Heat Map
الطيات	Folds
القيم الشاذة	Outliers
القيم المتوقعة	Fitted Values
القيم المعيارية	Standardized Values
القيم المفقودة	Missing Values
القيمة الاسمية أو الرتبة	Nominal or Ordinal
المدرج التكراري	Histogram

Least Squares	المربعات الصغرى
subgradient	التناقص المتدرج
Orthonormal Matrix	المصفوفة المتعامدة المعيارية
Coefficients	المعاملات
Magnitude	المقدار
Contours	المنحنيات المحيطية
Bias–Variance Trade–off	الموازنة بين التحيز والتباين
Norm	النظيم
Feature Selection	اختيار المتغيرات
Ridge Regression	انحدار الحرف
LASSO Regression	انحدار لاسو
Heteroscedasticity	انعدام التجانس في تباين حد الخطأ العشوائي
Test Data	بيانات الاختبار
Training Data	بيانات التدريب
High–Dimensional Data	بيانات عالية الأبعاد
Homoscedasticity	تجانس التباين في حد الخطأ العشوائي
ANOVA	تحليل التباين
Data Leakage	تسرب المعلومات/البيانات
Machine Learning	تعلم الآلة
Penalty	الجزاء
Coordinate Descent Algorithm	خوارزمية التراجع الإحداثي
Least Angle Regression (LARS)	خوارزمية انحدار الزوايا الصغرى
Sign Function	دالة الإشارة
Objective Function	دالة الهدف
Vertices	رؤوس
Overfitting	فرط التوفيق
Parsimony Principle	مبدأ الشح

Mean Squared Error (MSE)	متوسط مربعات الأخطاء
Residual Sum of Squares (RSS)	مجموع مربعات البواقي
Correlation Matrix	مصفوفة الارتباط
Pearson Correlation Coefficient	معامل الارتباط بيرسون
R-squared (R^2)	معامل التحديد
Soft-thresholding Operator	معامل العتبة الناعمة
Variance Inflation Factor (VIF)	معامل تباين التضخم
Information Criteria	معايير المعلومات
Tuning Parameter	معلمة الضبط
Resampling Methods	طرائق إعادة المعاينة
Ellipses	منحنيات القطع الناقص

الفصل الأول

الإطار التمهيدي للدراسة

المقدمة.

1.1 مشكلة البحث.

1.2 أهمية البحث.

1.3 أهداف البحث.

1.4 فرضيات البحث.

1.5 منهج البحث.

1.6 متغيرات البحث.

1.7 الدراسات السابقة.

1.8 هيكلية البحث.

المقدمة

مع صدور قانون الضريبة على البيوع العقارية، الذي يحدد مقدار الضريبة المستحقة على البيوع استناداً إلى القيمة الرائجة للعقارات، أصبح من الضروري تطوير نموذج أو أداة لتقدير الأسعار الحقيقية للعقارات بعيداً عن التدخل البشري. يُعد اختيار مجموعة محددة من المتغيرات المؤثرة في أسعار العقارات من أبرز التحديات في بناء نموذج الانحدار المتعدد. حيث يمكن أن يستند اختيار هذه المتغيرات إلى النظريات والأدبيات العلمية ذات الصلة أو إلى خبرة الباحث. وعلى الرغم من أن هذه العملية تبدو بسيطة ظاهرياً، إلا أن قياس بعض المتغيرات قد يكون صعباً نظراً لارتفاع التكلفة المادية، أو لعدم إمكانية قياسها أساساً. بالإضافة إلى ذلك، فإن إدراج عدد كبير من المتغيرات في النموذج المقدر ليس عملياً، ومن غير الصائب إدخالها جميعاً. لذلك، لابد من الاستعانة بطرائق إحصائية تساهم في تحديد المتغيرات الأكثر تأثيراً في أسعار العقارات.

1.1 مشكلة البحث

تُعدّ عملية تقدير أسعار العقارات بالغة التعقيد بسبب تعدّد العوامل المؤثرة فيها؛ إذ يصعب إدراج جميع المتغيرات المحتملة في نموذج واحد، في ظل تواجد عدد كبير من المتغيرات التي من الممكن أن تؤثر في أسعار العقارات والتي من الصعب إدراجها جميعاً في النموذج المقدر حيث ينصح بإدخال عدد محدد من المتغيرات

(Burnham & Anderson, 2002, p. 58)، أيّ عدم إدخال كل المتغيرات الممكنة في النموذج. وبغياض أو عدم وضوح نظرية تحدد بشكل صريح المتغيرات المؤثرة في المتغير المدروس (سعر العقار) مثل موقع العقار والصفات والمميزات التي يتمتع بها والتي تؤدي إلى اختلاف أسعاره يصبح من الضروري الاستعانة بطرائق إحصائية تساعد في اختيار المتغيرات التي ينبغي تضمينها أو استبعادها، وذلك بالاعتماد على معايير إحصائية محددة.

وتتمثل مشكلة البحث في الإجابة عن السؤال التالي: ماهي أبرز العوامل المؤثرة على أسعار العقارات؟

1.2 أهمية البحث

تتبع أهمية هذا البحث من الناحية النظرية من تركيزه على الطرائق الإحصائية المستخدمة في اختيار المتغيرات (Feature Selection) ذات الصلة ضمن نماذج الانحدار، ولا سيما في سياق النماذج متعددة المتغيرات التي تعاني من مشكلات التعددية الخطية. وتبرز الدراسة أهمية انحدار لاسو (LASSO)، باعتباره أحد أبرز أساليب الانحدار المنتظم/المقيد (Regularized Regression)، (Doszyń, 2024, pp. 719–740) لما يتمتع به

من قدرة على تقليص عدد المتغيرات الداخلة في النموذج، واختيار تلك الأكثر تأثيراً في المتغير التابع، الأمر الذي يسهم في تحسين دقة التقدير وتعزيز القدرة التفسيرية للنموذج.

ولا تقتصر أهمية هذا النموذج على الجوانب النظرية، بل تمتد إلى التطبيقات العملية، إذ يمكن اعتماد نتائجه كأساس في تصميم نماذج تنبؤية في مجالات متعددة، لا سيما في التقدير العقاري، ما يفتح المجال أمام تعميم استخدام هذا الأسلوب في ميادين علمية ومهنية مختلفة.

1.3 أهداف البحث

تتجلى أهداف البحث فيما يلي:

- التعريف بانحدار لاسو وتوضيح أهميته وتطبيقاته.
- تحديد أبرز العوامل المؤثرة في أسعار العقارات باستخدام انحدار لاسو.
- بناء نموذج إحصائي لتقدير القيمة الحقيقية لأسعار العقارات في مدينة دمشق.

الجدير بالذكر أنه يمكن لمتخذي القرار في سوريا الاستفادة من النتائج العملية التي سيتم التوصل إليها وتطبيقها على الواقع السوري في تحديد أسعار العقارات، في ظل وجود عدد كبير من العوامل التي قد تؤثر في أسعارها، حيث أن الأساليب الشائعة المتبعة (أسلوب المقارنة مع المبيعات المشابهة، أسلوب رأس المال الداخل، أسلوب التكلفة) في التقييم العقاري تعاني من حالات عدم التأكد والشك الناتجة عن قلة المعرفة وفقر أو نقص المعلومات عن المدخلات التي يمكن أن تُستخدم في عملية التقييم، الأمر الذي دفع إلى إيجاد نموذج إحصائي من شأنه أن يساعد في تقدير أسعار العقارات بعيداً عن التدخل الإنساني الذي قد يؤول إلى نتائج غير صحيحة.

1.4 فرضيات البحث

تقوم هذه الدراسة على الفرضيات الآتية:

1. يمكن من خلال استخدام الطرائق الإحصائية التوصل إلى نموذج يساعد في تقدير أسعار العقارات.
2. تسمح الطرائق الإحصائية بإعطاء نتائج موضوعية جيدة أكثر من الأساليب التقليدية لتقدير أسعار العقارات.

1.5 منهج البحث

اعتمد الباحث في هذه الدراسة الأسلوب الاستنتاجي (الاستنباطي) لتحديد أكثر العوامل تأثيراً في أسعار العقارات، حيث ناقش أولاً الأساليب الإحصائية المستخدمة من الناحية النظرية، بمقدمتها نموذج الانحدار المتعدد وانحدار لاسو، التي تناسب طبيعة المسألة المدروسة، ثم انتقل إلى التطبيق العملي لبناء نموذج إحصائي لتقدير القيمة الحقيقية لأسعار العقارات في مدينة Ames الأمريكية، حيث قسم البيانات إلى مجموعة تدريب يُبنى من خلالها النموذج، ومجموعة اختبار تُستخدم لاختبار النموذج. وبعد التأكد من صحته، سيتم إسقاطه على مدينة دمشق بأخذ عينة من البيوع العقارية في مختلف المناطق الإدارية لتقدير قيمتها الحقيقية، ومقارنتها مع سعر بيع العقار الفعلي.

1.6 متغيرات البحث

تعتمد هذه الدراسة على مجموعتين من البيانات العقارية، بهدف بناء نموذج إحصائي يُستخدم في تقدير أسعار العقارات بناءً على خصائصها الفيزيائية والموقعية:

أولاً - بيانات مدينة Ames (الولايات المتحدة):

استخدمت قاعدة بيانات معيارية غنية¹ تضم 80 متغيراً بحجم (2899) مفردة تتعلق بخصائص العقارات، مثل: مساحة المعيشة، وعدد الغرف، ونوع البناء، وجودة الإكساء، وسنة البناء، ووجود مرآب، وعدد المواقف المتاحة بالمرآب، ومساحة القبو، ونوع التدفئة، وموقع العقار، وغيرها. حيث تُستخدم هذه البيانات لاختبار كفاءة انحدار لاسو في بيئة بيانات ضخمة ومتوازنة.

ثانياً - بيانات مدينة دمشق (سوريا):

جُمِعت البيانات ميدانياً من خلال تسجيل معلومات متعلقة بأسعار بيع العقارات وخصائصها في مدينة دمشق. وتشمل المتغيرات التي وُثِّقت: المساحة، الارتفاع الطابقي، عدد الغرف، نوع الإكساء، اتجاه العقار، موقعه (رئيسي أو فرعي)، وجود شرفة، وجود مرآب، وجود مصعد أو مولدة، حالة التدفئة، وجود أثاث.

ونظراً لعدم توافر قاعدة بيانات معيارية، فإن هذه المتغيرات تمثل ما أمكن جمعه عملياً، وتُستخدم لاختبار قابلية تطبيق النموذج المقترح على بيئة محلية واقعية ذات بيانات محدودة.

¹ <http://jse.amstat.org/v19n3/decock.pdf>

1.7 الدراسات السابقة

1. Kummar, Attri, Singh, 2019, "Comparison of Lasso and stepwise regression technique for wheat yield prediction", India Meteorological Department-New Delhi.

هدفت هذه الدراسة إلى تحديد المتغيرات الأكثر تأثيراً في إنتاج محصول القمح، وذلك من خلال المقارنة بين أسلوب الانحدار التدريجي (Stepwise Regression) وانحدار لاسو باستخدام مجموعة واسعة من متغيرات الطقس، من أبرزها: درجة الحرارة العظمى والصغرى، ومعدل هطول الأمطار، ومستوى الرطوبة النسبية. وقد أظهرت النتائج أن قيمة معامل التحديد (R^2) بلغت 0.84 في انحدار لاسو، مقابل 0.85 في نموذج الانحدار التدريجي. إلا أن قيمة متوسط مربعات الأخطاء (MSE) كانت أقل في انحدار لاسو، ما يشير إلى أداء تنبؤي أفضل. وبناءً على هذه النتائج، خلُصت الدراسة إلى أن انحدار لاسو يُعد أكثر كفاءة في اختيار المتغيرات المؤثرة في إنتاج القمح، نظراً لقدرته على تقليص المتغيرات غير المهمة والحد من التعددية الخطية، دون التأثير السلبي في دقة التقدير.

2. Tibshirani et al, 2020, Best Subset, Forward Stepwise or Lasso? Analysis and Recommendations Based on Extensive Comparisons, Institute of Mathematical Statistics, Statist. Sci, volume 35.

هدفت هذه الدراسة إلى مقارنة ثلاث طرائق لاختيار النماذج الإحصائية، وهي: الاختيار الأمثل لجميع التركيبات الممكنة (Best Subset Regression)، والانحدار التدريجي (Forward Stepwise Regression)، وانحدار لاسو، وذلك باستخدام أسلوب المحاكاة (Simulation). يقوم أسلوب "Best Subset" بتقدير جميع النماذج الممكنة استناداً إلى مجموعة محددة من المتغيرات، ثم اختيار النموذج الأمثل منها بناءً على معايير إحصائية مثل (Akaike Information Criterion (AIC) و Bayesian Information Criterion (BIC)، ومعامل التحديد المعدّل (R^2 Adjusted).

أظهرت نتائج المحاكاة أن خوارزمية الاختيار الأمثل (Best Subset) وانحدار لاسو متعادلان إحصائياً، أي أن أيّاً منهما لا يتفوق على الآخر بشكل مطلق عبر جميع السيناريوهات. كما بينت النتائج أن الاختيار الأمثل والانحدار التدريجي متشابهان من حيث الأداء، ولم تُسجل فروقات جوهرية بينهما تستدعي تفضيل أحدهما على الآخر.

3. Seng, Jia Xin & Khalid, Kamil, 2018, "Modelling House Price Using Ridge Regression and Lasso Regression", International Journal of Engineering & Technology.

هدفت هذه الدراسة إلى بناء نموذج لتقدير أسعار العقارات السكنية باستخدام أسلوب انحدار الحرف (Ridge Regression) وانحدار لاسو، وذلك بالاعتماد على بيانات حقيقية لأسعار العقارات في مدينة Ames الأمريكية خلال الفترة الممتدة بين عامي 2006 و2010. وقد شملت المتغيرات المدروسة: مساحة العقار، عمر المبنى، عدد الغرف، وجود مرآب للسيارات، وموقع العقار. وأظهرت نتائج التحليل تفوق انحدار لاسو على انحدار الحرف من حيث القدرة على التقدير، حيث سجل قيمة أقل للجذر التربيعي لمتوسط مربعات الأخطاء (RMSE)، وقيمة أعلى لمعامل التحديد (R^2); كما ساعد انحدار لاسو على انتقاء المتغيرات الأكثر تأثيراً في السعر واستبعاد المتغيرات الأقل أهمية، ما جعله أكثر كفاءة في بناء نموذج مختصر وفعال لتقدير أسعار العقارات.

4. Liu, Zhihang, 2023, "Real Estate Price Prediction based on Supervised Machine Learning Scenarios", Highlights in Science, Engineering and Technology (CMLAI 2023)

هدفت هذه الدراسة إلى تقييم أداء نموذج الانحدار الخطي البسيط، وانحدار لاسو، وانحدار الحرف في توقع أسعار العقارات باستخدام مجموعة بيانات مفتوحة المصدر حول مبيعات العقارات. بدأ الباحث باستكشاف وصفي للبيانات يوضح علاقات المتغيرات، ثم طبق النماذج الثلاثة، وأظهر التحليل أن انحدار لاسو بلغ أعلى دقة تنبئية (90.15%)، بينما حقق النموذجان الآخران دقة مقاربة (نحو 90%)، إلا أن انحدار لاسو تميز بقدرته على اختيار المتغيرات الأكثر تأثيراً، وبناءً على هذه النتائج استنتج الباحث أن انحدار لاسو هو الخيار الأكثر فاعلية في سياق التنبؤ بأسعار العقارات ضمن البيانات محل الاختبار.

5. سيف علي، 2019، "تقدير معلمات نموذجي انحدار لاسو وانحدار الحرف مع تطبيق عملي"، جامعة كربلاء.

هدفت الدراسة إلى تطبيق انحدار لاسو وانحدار الحرف على بيانات تخص خصوبة المرأة، لدراسة العوامل المؤثرة في عدد الأطفال المولودين (المتغير التابع)، وعدد من المتغيرات المستقلة (عمر الزوج والزوجة، عمر المرأة عند الزواج، التحصيل الدراسي للزوج والزوجة، وزن المرأة، ...). وقد تبين أن البيانات تعاني من مشكلة التعدد الخطي (Multicollinearity)، وتوصلت الدراسة إلى أن انحدار لاسو يعد الأكثر كفاءة من حيث دقة التقدير، إذ حقق أدنى قيمة لمتوسط مربعات الأخطاء (MSE)، يليه انحدار الحرف.

6. منى حمادة، محمد نايفة، 2017، "نموذج رياضي مبني على تحليل الانحدار المتعدد لتقييم العقارات السكنية في دمشق"، جامعة البعث.

هدفت الدراسة إلى تحديد ودراسة المتغيرات الرئيسية التي تؤثر في تحديد أسعار العقارات السكنية في دمشق، حيث جُمعت بيانات 130 عقاراً سكنياً في دمشق من قبل خبراء بالتقييم العقاري لدى البنوك بهدف الرهن العقاري، واختيرت عشرة متغيرات مثل منطقة العقار، مساحة العقار، سعر صرف الدولار، درجة الإكساء الداخلي للعقارات، درجة أمان منطقة العقار، وتاريخ التقييم. وقد بيّنت المعالجة والتحليل أنه لا يمكن الاعتماد على نموذج الانحدار الخطي المتعدد لتقدير سعر العقار إلا بعد تحويله إلى انحدار لا خطي متعدد، وتوصل هذا البحث إلى أن العلاقة بين سعر العقار والمتغيرات المدروسة غير خطية.

تتميز هذه الدراسة عن مجمل الدراسات السابقة بما يلي:

1. معالجة نظرية شاملة باللغة العربية لأسلوب انحدار لاسو

تقدّم هذه الدراسة مرجعاً نظرياً متكاملًا باللغة العربية لأسلوب انحدار لاسو، يتضمّن شرحاً تفصيلياً للأساس الرياضي للنموذج، وخوارزميات تقدير معاملاته، وآلية ضبط مُعامل الانكماش λ ، إلى جانب عرضٍ لمعايير اختيار النموذج الأمثل مثل AIC، وBIC، والتحقق المتقاطع (Cross-Validation). وقد خُصّص الفصل الثاني من هذه الدراسة لهذا الغرض، بما يُغني المكتبة الإحصائية العربية بمادّة علميّة منهجية غير متوفّرة بصورة متكاملة في الأدبيات المنشورة باللغة العربية.

2. مقارنة تجريبية مزدوجة

- المرحلة الأولى (مجموعة بيانات Ames–USA): يُطبّق انحدار لاسو على قاعدة بيانات Ames Housing الشهيرة، وتُستخدم هذه المرحلة لاختبار كفاءة النموذج.
- المرحلة الثانية (بيانات دمشق الميدانية): جمع الباحث بيانات ميدانية حديثة عن سوق العقارات السكنية في مدينة دمشق، ثم يُقارن أداء انحدار لاسو بأداء الانحدار الخطي المتعدد ضمن هذه السوق ذات الخصائص البنيوية المختلفة.

تُسهم هذه الدراسة في إغناء نمذجة أسعار العقارات من جهة، وتسدّ فجوة واضحة في الدراسات العربية المرتبطة بالتحليل الإحصائي الحديث من جهة أخرى.

1.8 هيكليّة البحث

1. الفصل الأول: الإطار العام للدراسة
2. الفصل الثاني: الأسس الرياضية للنماذج المستخدمة. ويتضمن:
 - المبحث الأول: طريقة تقدير المربعات الصغرى.
 - المبحث الثاني: أسلوب انحدار لاسو.
 - المبحث الثالث: اختبار معلمة الضبط الخاصة بانحدار لاسو.
3. الفصل الثالث: بناء النماذج الإحصائية لتقدير أسعار العقارات. ويتضمن:
 - المبحث الأول: تطبيق انحدار لاسو لتقدير أسعار العقارات في مدينة Ames.
 - المبحث الثاني: طرائق التقييم العقاري.
 - المبحث الثالث: تطبيق انحدار لاسو لتقدير أسعار العقارات في مدينة دمشق ومقارنته بانحدار OLS.
4. مناقشة فرضيات البحث.
5. النتائج والتوصيات.
6. المراجع.
7. الملاحق.

الفصل الثاني

الأسس الرياضية للنماذج المستخدمة

المبحث الأول: طريقة تقدير المربعات الصغرى.

المبحث الثاني: أسلوب انحدار لاسو.

المبحث الثالث: اختبار معلمة الضبط الخاصة بانحدار لاسو.

الفصل الثاني

الأسس الرياضية للنماذج المستخدمة

تمهيد

يُعدُّ نموذج الانحدار واحداً من أهم الأدوات في مجال التحليل الإحصائي، حيث يهدف إلى التنبؤ بنتائج مستقبلية أو تفسير العلاقة بين مجموعة من المتغيرات المستقلة التي تؤثر في المتغير التابع محل الدراسة. يسمح هذا التحليل بتحديد أي من المتغيرات المستقلة لها التأثير الأكبر في المتغير التابع.

تُعدُّ طريقة المربعات الصغرى من أكثر طرائق التقدير شيوعاً في تحليل الانحدار، نظراً لسهولة تطبيقها من الناحية الرياضية، وقدرتها على تقديم معاملات قابلة للتفسير بوضوح، فضلاً عن تمتع المقدر الناتج عنها بعدد من الخصائص المرغوبة، مثل عدم التحيز والكفاءة في ظل تحقق فروض النموذج الكلاسيكي (Montgomery et al, 2012, pp. 47-49) إلا أنَّ فعالية هذه الطريقة تبدأ في التراجع عند وجود عدد كبير من المتغيرات المستقلة، حيث تصبح النماذج عُرضة لمشكلة التعدد الخطي، ما يؤدي إلى عدم استقرار التقديرات وتضخم تباينها.

وتتفاقم هذه الإشكالية في الحالة التي يتجاوز فيها عدد المتغيرات عدد المشاهدات (أي عندما يكون $N < P$)، حيث لا يمكن حينها الاعتماد على الطرائق التقليدية لتقدير معاملات النموذج بسبب غياب الحل الوحيد (James et al., 2021, p. 267). في مثل هذه الحالة، يصبح من الضروري اللجوء إلى أساليب تساعد في تقليص عدد المتغيرات الداخلة في النموذج، ومن أبرزها انحدار لاسو، الذي يطبق آلية انتقاء للمتغيرات من خلال إدخال حد جزائي (penalty) في دالة الهدف، ما يسهم في تبسيط النموذج وتعزيز قدرته التنبؤية.

قُسم الفصل الأول إلى ثلاثة مباحث:

1. المبحث الأول: يعرض طريقة المربعات الصغرى، وكيفية حساب المقدر، وبعض خصائصه الرياضية.
2. المبحث الثاني: يتناول أسلوب انحدار لاسو، والخصائص المتعلقة به، وكيفية تقديره.
3. المبحث الثالث: يركز على كيفية اختيار معلمة الضبط الخاصة بانحدار لاسو.

المبحث الأول

طريقة تقدير المربعات الصغرى (Least Square)

تعتمد معظم الأبحاث العلمية على تحليل وتقييم العلاقات بين مجموعة من المتغيرات بهدف الوصول إلى نماذج تصف هذه العلاقات بدقة. يُعدُّ نموذج الانحدار الخطي واحداً من أهم وأكثر الأدوات الإحصائية شيوعاً واستخداماً، نظراً لسهولة تفسير نتائجه وقدرته على توضيح تأثير المتغيرات المستقلة في المتغير التابع.

في هذا النموذج، تُستخدم المتغيرات المستقلة لاختبار مدى تأثيرها في المتغير التابع، ما يساعد في فهم طبيعة العلاقة بين المتغيرات، وتحديد العوامل الأكثر تأثيراً. هناك عدة طرائق لتقدير معاملات نموذج الانحدار الخطي، ومن أبرز هذه الطرائق طريقة المربعات الصغرى.

وقد اعتمد هذا البحث على هذه الطريقة نظراً لفعاليتها وخصائصها المميزة في تقديم تقديرات دقيقة لمعاملات النموذج.

2.1.1 نموذج الانحدار الخطي

تأخذ معادلة نموذج الانحدار الخطي الذي يضم P متغير مستقل بشكل عام الصيغة الآتية:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_P X_{iP} + \varepsilon_i \\ &= \beta_0 + \sum_{i=1}^n \sum_{j=1}^P X_{ij} \beta_j + \varepsilon_i \end{aligned} \quad (2.1)$$

حيث أن:

Y : المتغير التابع.

$\beta_0, \beta_1, \beta_2, \dots, \beta_P$ معاملات النموذج المجهولة المراد تقديرها.

$X_1, X_2, \dots, X_P$ المتغيرات المستقلة.

ε_i حد الخطأ العشوائي.

n حجم العينة (عدد المشاهدات).

فإذا عُرِفَ كلُّ مما يلي بالشكل الآتي:

$$Y_{n \times 1} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X_{n \times (p+1)} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}, \quad \beta_{(p+1) \times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \text{ and } \varepsilon_{n \times 1} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

فإنه يمكن كتابة ما سبق باستخدام المصفوفات كما يلي:

$$Y = X\beta + \varepsilon \quad (2.2)$$

Y : متجه عمودي من الدرجة $(n \times 1)$ يمثل قيم المتغير التابع.

X : مصفوفة من الدرجة $(n \times (P + 1))$ تمثل قيم المتغيرات المستقلة.

β : متجه عمودي من الدرجة $(p + 1) \times 1$ يمثل قيم المعلمات.

ε : متجه عمودي من الدرجة $n \times 1$ يمثل قيم حد الخطأ العشوائي.

2.1.2 طريقة تقدير المربعات الصغرى (OLS)

يتلخص جوهر هذه الطريقة بجعل الفروق بين القيم الفعلية Y_i للمتغير التابع وبين القيم المقدرة الناتجة عن مستقيم الانحدار $\hat{Y}_i$ أصغر ما يمكن كما في العلاقة الآتية:

$$RSS(\beta_0, \beta_1, \dots, \beta_p) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p x_{ij}\beta_j \right)^2$$

او بالشكل المصفوفي كما يلي:

$$\begin{aligned} RSS(\beta) &= \varepsilon^T \varepsilon \\ &= (Y - X\beta)^T (Y - X\beta) \end{aligned} \quad (2.3)$$

وعند اشتقاق المعادلة (2.3) بالنسبة لـ β نحصل على:

$$\begin{aligned} \nabla RSS(\beta) &= (Y - X\beta)^T (Y - X\beta) \\ &= -2X^T (Y - X\beta) \end{aligned}$$

وعند حل المعادلة نحصل على:

$$-2X^T Y + 2X^T X \hat{\beta} = 0$$

$$X^T Y = X^T X \hat{\beta}$$

وبضرب الطرفين بـ $(X^T X)^{-1}$ نحصل على:

$$\hat{\beta}^{OLS} = (X^T X)^{-1} X^T Y \quad 2.4$$

2.1.3 خصائص (صفات) مقدر المربعات الصغرى

يتميز مقدر المربعات الصغرى $\hat{\beta}^{ols}$ بما يلي:

1- الخطية، أي أن المقدّر $\hat{\beta}$ عبارة عن دالة خطية بالنسبة للمتغير التابع Y .

2- عدم التحيز: حيث يحقق العلاقة $E(\hat{\beta}) = \beta$

3- الكفاءة: يتميز المقدّر بأقل تباين ممكن بين جميع المقدّرات غير المتحيزة.

يُعطى تباين - تغاير المقدّر $\hat{\beta}$ بالعلاقة

$$var - cov(\hat{\beta}) = \hat{\sigma}(X^T X)^{-1} \quad 2.5$$

إن الغرض الأساسي من هذه الفقرة هو توضيح الشروط التي تجعل المصفوفة $(X^T X)$ في المعادلة (2.4) قابل للعكس. وتتحقق قابلية العكس لهذه المصفوفة إذا وفقط إذا:

1- عدم وجود ارتباط خطي تام بين المتغيرات المستقلة: أي أن المتغيرات المستقلة ليست مزيجاً خطياً من بعضها البعض. بمعنى آخر، لا يمكن التعبير عن أي متغير مستقل كمعادلة خطية تعتمد على المتغيرات المستقلة الأخرى.

2- عدد المشاهدات N أكبر من عدد المتغيرات المستقلة $(N > P)$.

وفي حال عدم تحقق هذه الشروط، تصبح المصفوفة $(X^T X)$ غير قابلة للعكس، مما يعني أنه لا يمكن حساب معكوسها، الأمر الذي يؤدي إلى عدم وجود حل (قيمة) وحيد أصغري للمقدّر $\hat{\beta}$ (unique solution) والذي يستدعي إيجاد طريقة أخرى تساعد في بناء النموذج المطلوب.

تبرز طريقة إنحدار لاسو كإحدى الأساليب الشائعة والفعالة عندما يكون عدد المفردات أقل من عدد المتغيرات المستقلة في النموذج، أو عند الرغبة في تقليص عدد المتغيرات واختيار الأكثر تأثيراً منها (Feature Selection).

2.1.4 بعض الإشكاليات المحتملة أثناء التقدير.

من الإشكاليات التي قد تعيق عملية تقدير معالم النموذج وفق طريقة المربعات الصغرى: انعدام التجانس في تباين حد الخطأ العشوائي (Heteroscedasticity)، والارتباط الذاتي بين البواقي (Autocorrelation)، والتعددية الخطية (Multicollinearity) بين المتغيرات المستقلة الناتجة عن كثرتها، بالإضافة إلى إشكالية فرط التوفيق

(Overfitting). ونظراً لارتباط كل من التعددية الخطية وفرط توفيق النموذج بشكل مباشر بكثرة المتغيرات المستقلة، سيتم في الفقرتين التاليتين التركيز على شرح إشكالية وجود عدد كبير من المتغيرات، ثم إشكالية فرط توفيق النموذج الناتجة عنها، دون التطرق لبقية الإشكاليات.

2.1.4.1 وجود عدد كبير من المتغيرات

تعتمد طريقة تقدير المربعات الصغرى بشكل أساسي على حجم العينة (N) وعدد المتغيرات المستقلة الداخلة في النموذج (P)، فعندما يكون حجم العينة أكبر من عدد المتغيرات المستقلة ($N > P$) يمكن تقدير $\hat{\beta}$ بشكل دقيق وفعال من خلال المعادلة (2.4).

أما عندما يكون حجم العينة أصغر من عدد المتغيرات المستقلة أي ($N < P$) عندها نكون بحالة البيانات عالية الأبعاد (High-Dimensional Data) (James, et al, 2021, pp. 59–60) بالتالي لا يوجد حل وحيد أمثل للمقدر $\hat{\beta}$ وإنما مجموعة غير منتهية من الحلول التي جميعها تحقق نفس القيمة الدنيا لمجموع مربعات البواقي. قد يؤدي تعدد الحلول للاعتقاد بوجود علاقة مثالية بين المتغيرات المستقلة والمتغير التابع، حتى في الحالات التي لا تكون فيها تلك العلاقة موجودة فعلياً. وتُعد هذه الظاهرة من أهم التحديات التي تواجه النماذج التقليدية في حالة البيانات عالية الأبعاد، ما يبرز الحاجة إلى اعتماد أساليب تقدير بديلة قادرة على التعامل مع هذا النوع من البيانات.

2.1.4.2 فرط توفيق النموذج بعدد من المتغيرات أكثر من المطلوب (Overfitting)

تحدث مشكلة فرط التوفيق (Overfitting) عندما يتضمن النموذج عدداً من المتغيرات المستقلة يفوق العدد اللازم لتمثيل العلاقة الحقيقية مع المتغير التابع (Everitt & Skrondal, 2010, p. 308). ففي هذه الحالة، قد يبدو النموذج قادراً على تمثيل العلاقة لبيانات العينة بشكل كبير، لكنه يفشل في التعميم عند تطبيقه على بيانات جديدة، مما يؤدي إلى ضعف القدرة التنبؤية للنموذج. لذلك، يفضل عادةً اتباع مبدأ الشح (Parsimony) في اختيار عدد المتغيرات الداخلة في النموذج بمعنى أقل عدد ممكن من المتغيرات، لأن أثر فرط-التوفيق (Overfitting) يؤدي إلى ازدياد تباين المقدر $\hat{\beta}^{OLS}$ (وفق المعادلة رقم 2.4) وذلك لأن العلاقة بين عدد المتغيرات (P) وتباين المقدر المعطى بالعلاقة:

$$\hat{\sigma}^2 = RSS(\hat{\beta}) / (n - p - 1) \quad 2.6$$

هي علاقة طردية أي بمعنى إذا اقترب عدد المتغيرات (P) من حجم العينة (n) فإن قيمة $\hat{\sigma}^2$ تكون كبيرة جداً، بالتالي إعطاء نتائج خاطئة عند الاستدلال الإحصائي. بالإضافة لذلك، تعتمد معظم الطرائق الحديثة على فكرة تقسيم البيانات إلى بيانات تدريب (training data) وبيانات اختبار (test data)، فعند وجود عدد كبير جداً من المتغيرات الداخلة في النموذج أكثر من العدد الحقيقي الممثل للعلاقة المدروسة فإن النموذج يُعطي نتائج جيدة

على بيانات التدريب ويعطي نتائج سيئة على بيانات الاختبار، وهو ما يعرف أيضاً بمشكلة فرط توفيق النموذج (Overfitting) (Hawkins, 2024, pp. 1–12).

المبحث الثاني أسلوب انحدار لاسو (LASSO)

تمهيد

شهدت النماذج الإحصائية تطوراً ملحوظاً من حيث الأساليب المستخدمة لتحليل البيانات، خصوصاً مع توسع حجم البيانات، وتعقيد العلاقات بين المتغيرات. ومن بين هذه الأساليب، يبرز انحدار لاسو، الذي طُوِّر لأول مرة في تسعينيات القرن الماضي على يد تيبشيري (Tibshirani, 1996) كأسلوب يجمع بين التقدير واختيار المتغيرات تلقائياً في آن واحد. وعلى الرغم من ظهوره المبكر، فقد اكتسب انتشاراً واسعاً في السنوات الأخيرة بفضل فعاليته في التعامل مع البيانات عالية الأبعاد، ومعالجة التعدد الخطي، وتحسين القدرة التنبؤية للنماذج.

نعرض في هذا المبحث الأسس النظرية والتطبيقية لانحدار لاسو، بدءاً من المفاهيم الأساسية المرتبطة بتحليل النماذج مثل بيانات التدريب وبيانات الاختبار، والموازنة بين التحيز والتباين (Bias–Variance Trade–off)، وطرائق الانكماش (Shrinkage)، وصولاً إلى تعريف دالة الهدف الخاصة بانحدار لاسو وخوارزميات التقدير المستخدمة. كما يتناول المبحث حالات خاصة مثل تصميم المصفوفات المتعامدة، ويعرض التفسير الرياضي والهندسي لمُقدِّر لاسو، مع التوضيح البياني لآلية عمله وتأثير معلمة الضبط λ على بنية النموذج. ويُختتم المبحث ببيان أبرز مزايا ومحدوديات هذا المُقدِّر، ما يُمهّد لفهم أعمق لتطبيقه العملي في المباحث اللاحقة.

2.2.1 مفاهيم تمهيدية

2.2.1.1 بيانات التدريب وبيانات الاختبار

يُعتبر تقسيم البيانات إلى مجموعتين: بيانات تدريب وبيانات اختبار، من المفاهيم الأساسية في تحليل البيانات وتقدير النماذج الإحصائية. ومع تطور تعلم الآلة (Machine learning) أصبح استخدام مصطلحي بيانات التدريب وبيانات الاختبار أكثر شيوعاً وانتشاراً، حيث يُصار إلى تقسيم مجموعة البيانات الكاملة إلى مجموعتين: تُستخدم المجموعة الأولى، المعروفة ببيانات التدريب، في بناء النموذج، بينما تختص المجموعة الثانية المسماة ببيانات الاختبار لتقييم أداء النموذج، ومدى قدرته على التنبؤ بدقة خارج نطاق العينة المستخدمة في بناءه.

ومن الجدير بالذكر أن هذا النهج مستخدم أيضاً في الاقتصاد القياسي تحت ما يُعرف بـ"التنبؤ خارج العينة" (Out-of-Sample Forecasting)، (C. Porter & Gujarati, 2008)، ويقوم هذا الأسلوب على استبعاد جزء من البيانات أثناء بناء النموذج، ثم إعادة استخدامه لاحقاً لاختبار مدى دقة النموذج وملاءمته في التنبؤ بقيم لم تُستخدم ضمن بيانات التقدير الأصلية.

2.2.1.2 المفاضلة (الموازنة) بين التباين والتحيز

يُعدّ متوسط مربعات الأخطاء من المقاييس الأساسية لتقييم أداء المقدّرات الإحصائية، ويُعرّف على النحو الآتي:

$$MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2 \quad 2.7$$

ويمثّل هذا التعبير القيمة المتوقعة لمربع الفرق بين المقدّر $\hat{\theta}$ والقيمة الحقيقية للمعلمة θ ويمكن تفكيك هذا المؤشر إلى مكونين أساسيين:

$$MSE(\hat{\theta}) = var(\hat{\theta}) + (bias(\hat{\theta}))^2 \quad 2.8$$

أي أن متوسط مربعات الأخطاء يتكوّن من تباين المقدّر بالإضافة إلى مربع التحيز.

ليكن $x_t \in \mathbb{R}^p$ متجه يمثل قيم جميع المتغيّرات المستقلّة للمشاهدة t ، وأن y_t هو المتغير التابع.

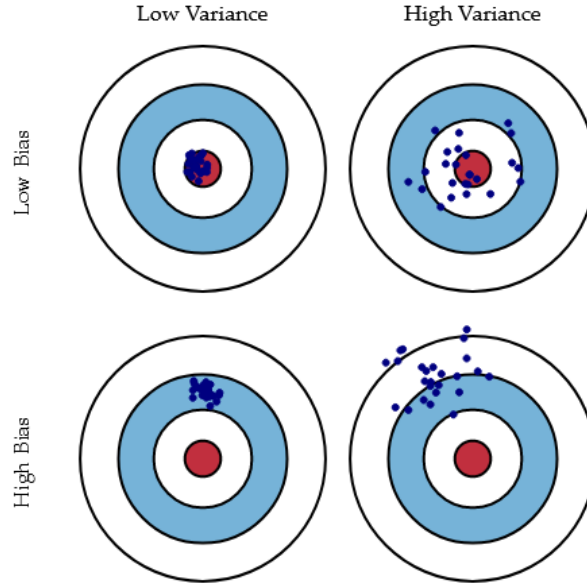
عندها يمكن التعبير عن متوسط مربعات الخطأ المتوقع عند نقطة اختبار جديدة x_0 كما يلي:

$$E(y_0 - \hat{f}(x_0))^2 = Var(\hat{f}(x_0)) + bias(\hat{f}(x_0))^2 \quad 2.9$$

حيث يقصد بالتباين، المقدار الذي تتغير به التنبؤات ($\hat{f}$) عند استخدام عينات تدريب مختلفة لا تُستخدم أصلاً في عملية التقدير. ويُفترض أن تكون نتائج النموذج مستقرة نسبياً عند تغيير مجموعة التدريب؛ إذ يُعدّ تغيّر النتائج بشكل كبير مؤشراً على ارتفاع التباين.

أما التحيز، فهو الفرق بين القيم الحقيقية والمتوقعة الناتجة عن النموذج، ويعكس مدى قدرته على تمثيل العلاقة الفعلية بين المتغيرات. فعند تقدير معلمات نموذج الانحدار الخطي، يُفترض وجود علاقة خطية بين المتغيرات المدروسة، في حين أن هذه العلاقة قد تكون غير خطية في الواقع، وهو ما يُضيف قدراً من التحيز إلى المقدّرات الناتجة عن هذا النموذج في كثير من الحالات.

وفيما يلي الشكل البياني رقم (1) الذي يوضح العلاقة بين التباين والتحيز:



الشكل (1) توضيح بياني للتحيز والتباين.

المصدر : <http://scott.fortmann-roe.com/docs/BiasVariance.html>

يوضح الشكل أعلاه العلاقة بين التحيز (Bias) والتباين (Variance) في أداء المقدّر الإحصائي، من خلال تشبيه النتائج بدقة التصويب على هدف. في الجزء العلوي الأيسر (Low Bias – Low Variance)، تتركز القيم المقدّرة قرب المركز، ما يعكس أن المقدّر يتمتع بدقة عالية، أي أن كل من التحيز والتباين منخفضان.

أما في الجزء العلوي الأيمن (Low Bias – High Variance)، فإن التقديرات لا تزال تدور حول القيمة الحقيقية (المركز) لكن بانتشار واسع، ما يشير إلى انخفاض التحيز مع ارتفاع التباين. في المقابل، يبيّن الجزء السفلي الأيسر (High Bias – Low Variance) أن التقديرات مستقرة لكنها تنحرف بشكل منتظم عن القيمة الحقيقية، أي أن النموذج يعاني من تحيز واضح. أما في الجزء السفلي الأيمن (High Bias – High Variance)، فتمثل هذه الحالة الأداء الأسوأ.

بناءً على ما سبق، فإن الحالة المثلى هي عندما يكون كل من التحيز والتباين منخفضين، لأن ذلك يؤدي إلى تقليل متوسط مربعات الأخطاء (MSE)، الذي يُعد مؤشراً لجودة أداء المقدّر. ولذلك، تسعى النماذج الإحصائية إلى تحقيق هذه الحالة قدر الإمكان من خلال التوازن بين التعقيد والتبسيط، والتقدير الدقيق دون فرط في التوفيق في تمثيل البيانات.

2.2.2 طرائق الانكماش (Shrinkage methods)

تكمّن فكرة طرائق الانكماش في تقليص قيمة (Magnitude) ² معاملات النموذج المقدّرة نحو قيمة مركزية، وغالباً ما يكون الهدف من ذلك الحدّ من التأثير النسبي للمتغيرات التي لا تُسهم بشكل فعّال في تفسير العلاقة المدروسة. وبهذه الطريقة، تُحفظ المعاملات الأكثر أهمية، ويُقلّص قيمة المعاملات الأقل تأثيراً. وبعبارة أخرى، كلما اقتربت قيمة معامل ما من الصفر، كان تأثيره التفسيري أضعف، في حين يُعدّ الابتعاد عن الصفر مؤشراً على أهميته في تفسير المتغير التابع.

تعتمد هذه الطرائق على إدخال مقدار معيّن من التحيز (Bias) بشكل متعمّد في عملية التقدير بهدف تقليل التباين (Variance)، مما يُسهم في تحسين دقة النموذج. وكنتيجة لذلك، يُقلّل متوسط مربع الخطأ (MSE) الخاص بالنتبؤات.

ومن الملاحظ أن تقليص قيم المعاملات يمكن أن يساهم بشكل فعّال في تحسين القدرة التنبؤية للنموذج، لا سيما في الحالات التي تعاني من مشكلات مثل التعددية الخطية أو البيانات عالية الأبعاد (James, et al, 2021, pp. 61-63). وتتدرج تحت مفهوم الانكماش مجموعة من الطرائق الإحصائية، إلا أن هذا البحث سيركّز تحديداً على طريقة انحدار لاسو، نظراً لتمييزها في الجمع بين التقدير وانتقاء المتغيرات تلقائياً ضمن إطار رياضي واحد.

2.2.3 ما هو انحدار لاسو (مقدر لاسو)

اقترح (Tibshirani) عام 1996 دالة (مقدر) Lasso وهي اختصار لـ (Least Absolute Shrinkage and Selection Operator) لتقدير معاملات نموذج الانحدار الخطي، والتي تعمل على تقدير واختيار المعلمات الداخلة في النموذج في الوقت ذاته. يتم الحصول على مُقدّر لاسو، $\hat{\beta}^{lasso} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p)$ وفقاً للصيغة الآتية:

² يشير مصطلح magnitude في سياق طرائق الانكماش إلى قيمة المعلمة (coefficient) أو حجمها بشكل مطلق، أي ما يُعرف بـ "النظم الرياضية" (norms)، وتحديدًا $\ell_1 norm - \ell_2 norm$ ، ويُقصد به مدى كبر أو صغر المعلمة بالمقارنة مع غيرها في النموذج. وغالباً ما يُستخدم هذا المفهوم لتحديد أهمية المتغيرات، حيث يُعدّ اقتراب قيمة المعلمة من الصفر مؤشراً على ضعف أهميتها التفسيرية.

$$\begin{aligned}\hat{\beta}^{lasso} &= \operatorname{argmin} \left\{ \frac{1}{2} \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2 + \lambda \sum_{j=1}^p |\beta_j| \right\} \\ &= RSS + \lambda \sum_{j=1}^p |\beta_j|\end{aligned}\tag{2.10}$$

نلاحظ أن المعادلة تتألف من القسم الأول وهو مقدر المربعات الصغرى بالإضافة لدالة الجزاء والتي تتألف من:

λ : تدعى بمعلمة الضبط (Tuning Parameter) وهي مقدار الجزاء المفروض على عملية التقدير.

$\lambda \sum_{j=1}^p |\beta_j|$: تدعى دالة الجزاء (Penalty Function) أو دالة الضبط (Regularization Function).

يقوم مبدأ انحدار لاسو على تقليل مجموع مربعات البواقي، مع إدراج قيد رياضي يتمثل في وضع حد أقصى لمجموع القيم المطلقة لمعاملات النموذج، ويُعرف هذا القيد باسم L1-norm. ويُراد بهذا القيد منع المعاملات من اتخاذ قيم كبيرة، ما يدفع النموذج إلى تقليص قيم بعض المعاملات والاقتراب بها من الصفر. وبهدف تحقيق ذلك، يعتمد انحدار لاسو على تقليص قيمة معاملات الانحدار باتجاه الصفر. حيث عندما تكون مساهمة بعض المتغيرات في تقليل مربعات البواقي ضئيلة، فإن هذا القيد يؤدي إلى تصغير المعاملات المرتبطة بها تماماً، وبالتالي استبعادها من النموذج أي حذفها. وبهذه الآلية، يُسهم انحدار لاسو في بناء نموذج أكثر بساطة وكفاءة، من خلال الاحتفاظ فقط بالمتغيرات الأكثر مساهمة في تقليل مربعات البواقي.

وتُحدد درجة التقليل وفقاً لقوة معامل الجزاء المفروض λ ، حيث يؤثر هذا المعامل على درجة الانكماش المطبقة على المعاملات. وبذلك، تحقق طريقة لاسو هدفين:

1- تحسين القدرة التنبؤية للنموذج: وذلك من خلال تقليل متوسط مربعات الأخطاء. إذ تُعرف مقدرات

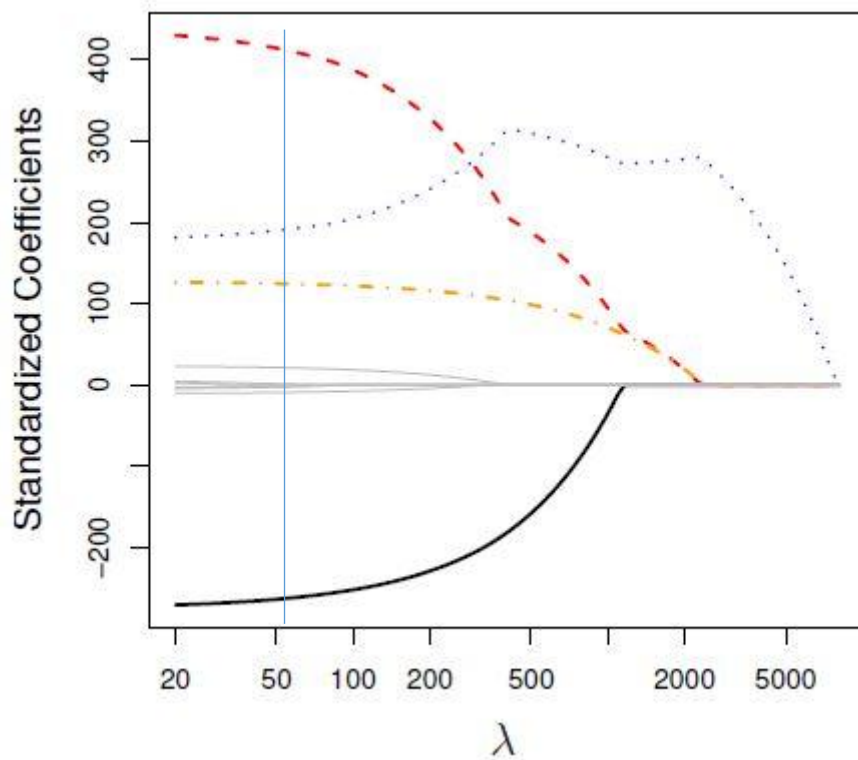
المربعات الصغرى بأنها ذات تحيز منخفض وتباين مرتفع، ما قد يؤدي إلى ضعف التنبؤ خارج العينة. في المقابل، يعمل لاسو على التضحية بجزء بسيط من التحيز مقابل تقليل ملحوظ في التباين، ما ينعكس إيجاباً على دقة التنبؤ الإجمالية.

2- تعزيز القدرة التفسيرية للنموذج: عن طريق استبعاد بعض المتغيرات غير المؤثرة، ما يؤدي إلى بناء

نموذج أبسط وأكثر وضوحاً. وهذا يُعد من مزايا انحدار لاسو مقارنةً بالأساليب التقليدية التي تُدرج جميع المتغيرات بصرف النظر عن أهميتها التفسيرية.

يرتبط مقدار الجزاء المفروض في تقدير انحدار لاسو بقيمة المعامل λ ، حيث يُمثل هذا المعامل درجة القيود المفروضة على معاملات النموذج. فإذا كانت $\lambda = 0$ ، فإن عملية التقدير تناظر تماماً طريقة المربعات الصغرى،

أي لا يُفرض أي قيد على المعاملات، وبالتالي لا يُستبعد أي منها. ومع ازدياد قيمة λ ، تزداد درجة الانكماش المفروضة على معاملات النموذج المقدرة، ما يؤدي إلى تقليص قيمها تدريجياً نحو الصفر. وعندما تصبح قيمة λ كبيرة جداً، فإن الجزء المفروض يكون شديداً إلى درجة أن جميع معاملات النموذج تساوي الصفر أي تُستبعد جميع المتغيرات من النموذج، ويُصبح النموذج خالياً تماماً من المتغيرات التفسيرية (المستقلة). وتوضح العلاقة بين مقدار λ وعدد المتغيرات المختارة في الشكل رقم (2)، حيث يمكن ملاحظة تأثير ازدياد λ على سلوك النموذج وعدد المتغيرات الفعالة داخله.



الشكل (2) المعلمات المعيارية المقدرة من خلال انحدار لاسو كدالة في قيم λ

المصدر : Tibshirani, et all (2021), An Introduction to Statistical Learning

ويوضح الشكل رقم (2) العلاقة بين قيمة معلمة الجزء λ وقيم معاملات المتغيرات المقدرة في انحدار لاسو. حيث يمثل كل منحنى في الشكل مسار أحد معاملات المتغيرات المستقلة، في حين يُمثل المحور الأفقي القيم المختلفة لـ λ والمحور العمودي القيم المقدرة المعيارية للمعاملات. ويلاحظ أنه عندما تكون λ صغيرة، فإن النموذج يشمل جميع المتغيرات تقريباً، وهو ما يتطابق مع نموذج المربعات الصغرى. أما مع ازدياد قيمة λ ، فإن معاملات بعض المتغيرات تبدأ بالانكماش تدريجياً نحو الصفر، إلى أن تصل بعض المسارات إلى الصفر تماماً، ما يعني استبعاد المتغيرات المقابلة لها من النموذج. وعندما تصبح قيمة λ كبيرة جداً، فإن جميع المعاملات تتوّل إلى الصفر، ما يؤدي إلى نموذج لا يحتوي على أي متغيرات مستقلة. ويُبرز هذا الشكل الوظيفة الأساسية لإنحدار لاسو في انتقاء

المتغيرات تلقائياً بناءً على أهمية كل منها في تقليل مربعات البواقي، كما يُظهر بوضوح تأثير قوة الجزء المفروض λ على بنية النموذج.

يُفضل، عند وجود عدد كبير من المتغيرات الداخلة في النموذج، والتي قد تختلف فيما بينها من حيث وحدات القياس، تحويل قيم المتغيرات (سواء المستقلة أو التابعة) إلى قيم معيارية (standardized values)، بحيث يصبح المتوسط الحسابي لها صفراً، ويساوي الانحراف المعياري (أو التباين) القيمة 1.

وبالتالي بعد حذف الثابت من النموذج يمكن إيجاده بعد عملية التقدير من خلال العلاقة:

$$\widehat{\beta}_0 = \bar{y} - \sum_{j=1}^p \bar{x}_j \widehat{\beta}_j \quad 2.11$$

2.2.4 صيغة أخرى لانحدار لاسو

يمكن كتابة المعادلة رقم (2.10) بالشكل الآتي:

$$\text{minimize } \left\{ \frac{1}{2} \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2 \right\} \text{ subject to } \sum_{j=1}^p |\beta_j| \leq s \quad 2.12$$

وتُعد هذه الصيغة (2.12) مكافئة رياضياً للصيغة (2.10)، حيث توجد علاقة مباشرة بين معلمة الجزء λ في (2.10) والحد الأعلى s في (2.12)، إذ أن لكل قيمة من λ توجد قيمة مقابلة لـ s تعطي نفس التقدير لمعاملات لاسو.

نسعى في هذه الصيغة، إلى إيجاد مجموعة القيم المثلى لمعاملات النموذج، بحيث يكون مجموع مربعات البواقي في أدنى قيمة ممكنة، مع الالتزام بالقيد الذي يفرض حداً أقصى على مجموع القيم المطلقة للمعاملات $\sum_{j=1}^p |\beta_j|$. هذا القيد يُجبر بعض المعاملات على الاقتراب من الصفر أو التلاشي تماماً، ما يؤدي إلى استبعاد بعض المتغيرات من النموذج تلقائياً.

وعليه، فإن انحدار لاسو يُحقق وظيفتين في آن واحد كما ذكرنا سابقاً، فهو يعمل على تقدير معاملات النموذج، ويقوم في الوقت ذاته باختيار المتغيرات ذات التأثير الأكثر أهمية، مستبعداً المتغيرات غير الفعالة من خلال آلية الانكماش الرياضي المفروضة.

2.2.5 تقدير انحدار لاسو

لتكن X مصفوفة تمثل قيم المتغيرات المستقلة، و Y متجه يمثل قيم المتغير التابع. بالتالي يمكن كتابة المعادلة رقم (2.10) بالشكل المصفوفي وبدلالة النظم (norm)³ كما يلي:

$$\hat{\beta}^{lasso} = \operatorname{argmin} \frac{1}{N} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 \quad 2.13$$

بالعودة إلى نتيجة تقدير المربعات الصغرى من المعادلة رقم (2.4) يمكن الوصول إلى:

$$\hat{\beta}^{lasso} = (X^T X)^{-1} X^T Y + \lambda \operatorname{sign}(\hat{\beta})$$

$$\hat{\beta}^{lasso} = \hat{\beta}^{OLS} + \lambda \operatorname{sign}(\hat{\beta}) \quad 2.14$$

أي أن مقدر $\hat{\beta}^{lasso}$ هو عبارة عن مقدر المربعات الصغرى $\hat{\beta}^{OLS}$ بالإضافة إلى $\lambda \operatorname{sign}(\hat{\beta})$. حيث أن دالة الإشارة (Sign Function) $\operatorname{sign}(x)$ هي عبارة عن مشتق دالة $f = |x|$ المعرفة كما يلي:

$$\partial|x| = \operatorname{sign}(x) = \begin{cases} -1 & \text{if } x < 0 \\ [-1,1] & \text{if } x = 0 \\ 1 & \text{if } x > 0 \end{cases} \quad 2.15$$

يتبين من هذا التعريف أن الدالة $f = |x|$ ليست قابلة للاشتقاق عند $x = 0$ ، مما يعني أن دالة الهدف في انحدار لاسو تكون غير قابلة للتفاضل عند هذه النقطة. وفي سياق مسائل المثلية المحدبة (convex optimization) يستعاض عن مفهوم المشتق العادي بمفهوم التناقص المتدرج (subgradient) والذي يُعرّف عند هذه النقطة على أنه جميع القيم الواقعة ضمن المجال المغلق $[-1,1]$.

ونتيجة لعدم قابلية دالة الهدف للتفاضل في هذه النقطة، لا يمكن الاعتماد على الحلول الجبرية المغلقة (Closed-form Solutions) في إيجاد قيم المعاملات، كما هو الحال في المربعات الصغرى، الأمر الذي يستدعي اللجوء إلى طرائق عددية تكرارية/طرائق البحث التكرارية (Iterative Numerical Methods). وتقوم هذه الطرائق على اختيار نقطة ابتدائية، ثم تحسينها بشكل تدريجي عبر سلسلة من الخطوات المحسوبة، حتى الوصول إلى القيمة التي تحقق الحد الأدنى لدالة الهدف. وفي هذا السياق، يُقدّم أولاً حل مبسّط لانحدار لاسو في حالة وجود متغير مستقل واحد، قبل التعميم إلى الحالة العامة التي تشمل أكثر من متغير مستقل.

³ إن رمز $\|\beta\|_2$ عبارة عن نظم المتجه β والذي يساوي $\|\beta\|_2 = \sqrt{\sum_{j=1}^p \beta_j^2}$ وأن $\|\beta\|_1 = \sum |\beta_j|$ $\ell_1 \text{ norm}$

وعليه، يُجرى في البداية إيجاد مقدر لاسو $\hat{\beta}^{lasso}$ بالنسبة لمتغير مستقل واحد، ومن ثم للحالة العامة بوجود أكثر من متغير مستقل.

2.2.6 تقدير انحدار لاسو لمتغير مستقل واحد فقط

لتكن X مصفوفة قيم المتغير المستقل، و Y متجه لقيم المتغير التابع، وقد حُولت جميع المتغيرات إلى القيم المعيارية بمتوسط يساوي 0 وتباين يساوي 1، بالتالي فإن

$$\hat{\beta}^{lasso} = \operatorname{argmin} \frac{1}{N} \|Y - X_1 \beta_1\|_2^2 + \lambda \|\beta_1\|_1$$

وباشتقاق المعادلة بالنسبة لـ β_1 نحصل على:

$$-\frac{2}{N} X_1^T (Y - X_1 \hat{\beta}_1) + \lambda \operatorname{sign}(\hat{\beta}_1) = 0$$

$$\frac{1}{N} X_1^T (Y - X_1 \hat{\beta}_1) = \frac{\lambda}{2} \operatorname{sign}(\hat{\beta}_1)$$

$$\frac{1}{N} X_1^T Y - \frac{1}{N} X_1^T X \hat{\beta}_1 = \frac{\lambda}{2} \operatorname{sign}(\hat{\beta}_1)$$

بما أن $\frac{1}{N} X_1^T X = 1$ (نتيجة تحويل قيم المتغيرات إلى قيم معيارية) بالتالي:

$$\hat{\beta}_1 = \frac{1}{N} X_1^T Y - \frac{\lambda}{2} \operatorname{sign}(\hat{\beta}_1)$$

وبالتعويض بالمعادلة رقم (2.13) نحصل على:

$$\hat{\beta}_1 = \begin{cases} \frac{1}{N} X_1^T Y + \frac{\lambda}{2} & \text{if } \frac{1}{N} X_1^T Y < -\frac{\lambda}{2} \\ 0 & \text{if } \frac{1}{N} X_1^T Y \leq \frac{\lambda}{2} \\ \frac{1}{N} X_1^T Y - \frac{\lambda}{2} & \text{if } \frac{1}{N} X_1^T Y > \frac{\lambda}{2} \end{cases}$$

وإذا اعتُبر أن $\frac{1}{N} X_1^T Y = \hat{\beta}^{OLS}$ فإنه يمكن كتابة مقدر لاسو $\hat{\beta}^{lasso}$ كما يلي:

$$\hat{\beta}_1^{lasso} = S_{\frac{\lambda}{2}} \left(\frac{X_1^T Y}{N} \right) = S_{\frac{\lambda}{2}} (\hat{\beta}^{OLS}) \quad 2.16$$

حيث أن S عبارة عن معامل العتبة الناعمة (soft-thresholding operator) والذي يأخذ الصيغة الآتية:

$$S_{\frac{\lambda}{2}}(\beta) = \text{sign}(\beta)(|\beta| - \frac{\lambda}{2})$$

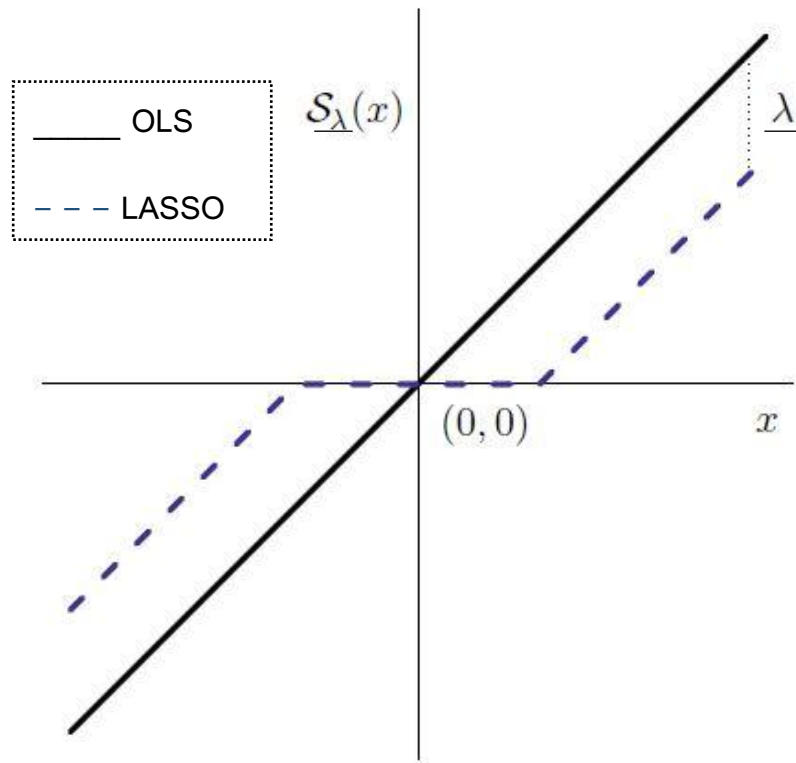
$$= \begin{cases} \beta + \lambda/2 & \text{if } \beta < -\frac{\lambda}{2} \\ 0 & \text{if } |\beta| \leq \frac{\lambda}{2} \\ \beta - \lambda/2 & \text{if } \beta > \frac{\lambda}{2} \end{cases} \quad 2.17$$

وعليه يمكن القول، يعتمد انحدار لاسو في جوهره على تطبيق ما يُعرف بمعامل العتبة الناعمة (Vandenberghe, 2011, p. 188)، وهو أسلوب يُستخدم لتقليص قيم معاملات النموذج عن طريق فرض حدّ معين يُعرف بعتبة الانكماش.

تقوم هذه الآلية على أساس تقييم القيمة المطلقة للمعلمة $|\beta|$ فإذا كانت هذه القيمة أقل من أو تساوي نصف قيمة معامل الجزاء $\frac{\lambda}{2}$ ، تُعدّ المعلمة غير مؤثرة بدرجة كافية وتُزال من النموذج، وذلك من خلال تعيين قيمتها إلى الصفر. أما إذا تجاوز مقدار المعلمة هذا الحد، فنُقَلِّص باتجاه الصفر بمقدار ثابت قدره $\frac{\lambda}{2}$ ، دون أن تُحذف تماماً.

وبذلك، يعمل معامل العتبة الناعمة على التوازن بين الحفاظ على المتغيرات المهمة وتبسيط النموذج، من خلال استبعاد المتغيرات ذات التأثير الضعيف. وتُعدّ هذه الخاصية من الخصائص المميزة لانحدار لاسو، حيث يدمج بين تقدير المعلمات واختيار المتغيرات تلقائياً ضمن عملية واحدة.

ويمكن تمثيل المعادلة رقم (2.17) بالشكل البياني الآتي:

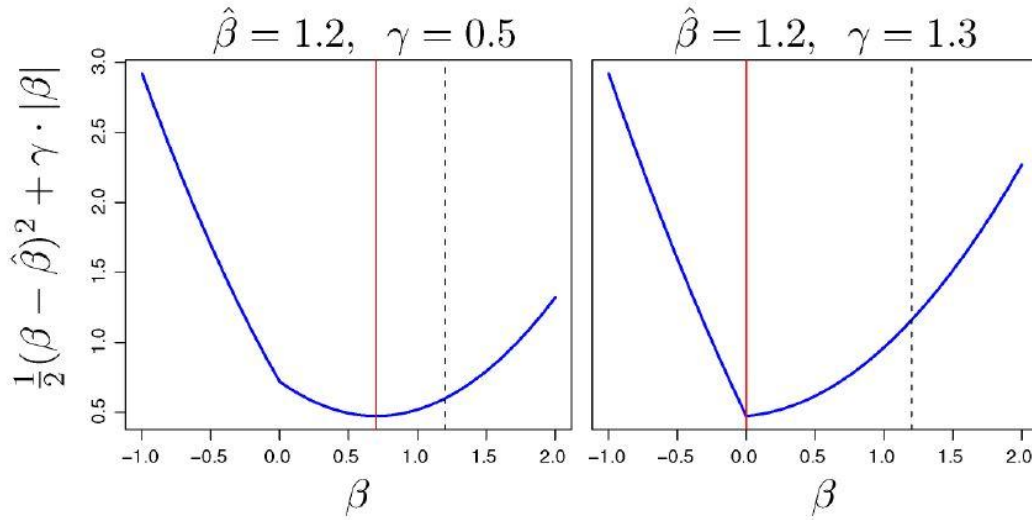


الشكل (3): التمثيل البياني للمعادلة رقم (2.11).

المصدر: من إعداد الباحث بالاعتماد على لغة البرمجة Python.

يوضح الشكل (3) آلية تأثير انحدار لاسو حيث يُمثل الخط الأسود المستقيم ناتج مقدر المربعات الصغرى الاعتيادية، دون تطبيق أي تقليص على قيم المعامل. أما الخط الأزرق المنقطع فيُعبّر عن مُقدّر لاسو.

ويلاحظ أن انحدار لاسو يقوم بتقليص قيم مقدر المربعات الصغرى بمقدار ثابت قدره $\frac{\lambda}{2}$ ، ويعمل على جعل بعض المعاملات تساوي الصفر (أي حذف المتغيرات المرتبطة بها من النموذج) إذا كانت قيم مقدر المربعات الصغرى أصغر من $\frac{\lambda}{2}$ من حيث القيمة المطلقة. وبالتالي، يمكن القول إنه في حال وجود متغير مستقل واحد، وبعد تحويل المتغيرات إلى القيم المعيارية، فإن مقدر لاسو $\hat{\beta}^{LASSO}$ يُمكن الحصول عليه من خلال تطبيق معامل العتبة الناعمة على مقدر المربعات الصغرى، ما يجمع بين تقدير المعاملات واختيار المتغيرات في آنٍ واحد. ويوضح الشكل البياني الآتي كل ما سبق:



الشكل (4): العلاقة بين مقدر لاسو في حالة وجود متغير مستقل واحد حولت قيمه إلى قيم معيارية، وبين معامل soft

thresholding، يمثل الخط العمودي المقطع تقدير المربعات الصغرى، والخط العمودي المتصل تقدير لاسو.

المصدر: FRIEDMAN, JEROME. Et al. 2007, "PATHWISE COORDINATE OPTIMIZATION"

فإذا كُتبت المعادلة (2.16) بشكل مختصر أكثر كما يلي:

$$\begin{aligned}\hat{\beta}^{\text{lasso}}(\gamma) &= S(\hat{\beta}, \gamma) \equiv \text{sign}(\hat{\beta})(|\hat{\beta}| - \gamma) \\ &= \begin{cases} \hat{\beta} - \gamma, & \text{if } \hat{\beta} > 0 \text{ and } \gamma < |\hat{\beta}|, \\ \hat{\beta} + \gamma, & \text{if } \hat{\beta} < 0 \text{ and } \gamma < |\hat{\beta}|, \\ 0, & \text{if } \gamma \geq |\hat{\beta}|. \end{cases}\end{aligned}$$

فإن التعبير يُختزل إلى علاقة تعتمد على الفرق بين مقدر المربعات الصغرى (OLS) ومقدار ثابت γ ، فإذا كانت قيمة المقدّر $\hat{\beta}$ أكبر من الصفر وكانت γ أصغر من القيمة المطلقة لـ $\hat{\beta}$ ، فإن مقدر لاسو يُساوي $\hat{\beta}^{\text{LASSO}} = \hat{\beta} - \gamma$ ، وهو ما يتوافق مع الجزء الأيسر من الشكل البياني (4). أما في الحالات التي تكون فيها γ أكبر من أو تساوي القيمة المطلقة لـ $\hat{\beta}$ ، فإن مقدر لاسو يُعطي صفراً، كما هو مبين في الجزء الأيمن من الشكل ذاته. توضح هذه المعادلة كيف يقوم انحدار لاسو بتقليص بعض المعاملات إلى الصفر بناءً على موقعها النسبي من العتبة، ما يتيح للنموذج حذف المتغيرات غير المؤثرة.

2.2.7 تقدير انحدار لاسو في الحالة الخاصة التصميم المتعامد (Orthogonal design).

لتبيان آلية عمل انحدار لاسو بشكل عام، من المفيد افتراض أن المتغيرات الداخلة في النموذج متغيرات متعامدة بمعنى أن التغاير (Covariance) بين كل زوج من المتغيرات يساوي الصفر، أي كل متغير مستقل لا يرتبط

خطياً بأي متغير مستقل آخر ويُفترض كذلك أن المتوسط الحسابي لكل متغير يساوي صفراً. في هذه الحالة، تتحقق الشروط الآتية:

$$1. \quad x_j^T = 0 \text{ وذلك لجميع المتغيرات } (j = 1, 2, 3, \dots, p)$$

$$2. \quad x_j^T x_k = 0 \text{ وذلك لجميع } j \neq k$$

حيث أن تقدير المربعات الصغرى من المعادلة رقم (2.4) يساوي إلى

$$\beta_j^{ols} = X_j^T Y, \dots, j = 1, 2, 3 \dots p$$

أي بمعنى أن نتيجة مقدر المربعات الصغرى β_j^{ols} لا تعتمد على نتيجة أي متغير آخر سوى X_j ، ولا يتغير أبداً إذا جعلنا $\beta_k = 0$ وذلك من أجل $j \neq k$.

لتكن X مصفوفة قيم المتغيرات المستقلة، و Y متجه لقيم المتغير التابع، وقد حُولت جميع المتغيرات إلى القيم المعيارية بمتوسط يساوي 0 وتباين يساوي 1. وعند افتراض أن المتغيرات في المصفوفة X متعامدة وقد خضعت لعملية التحويل إلى القيم المعيارية⁵، فإن X تصبح مصفوفة متعامدة معيارية (orthonormal matrix). عندها يمكن كتابة الحل الناتج عن انحدار لاسو كما يلي:

$$-\frac{2}{N} X^T (Y - X\hat{\beta}) + \lambda \text{sign}(\hat{\beta}) = 0$$

$$\frac{1}{N} X^T (Y - X\hat{\beta}) = \frac{\lambda}{2} \text{sign}(\hat{\beta})$$

بالخطوات السابقة نفسها نحصل على:

$$\hat{\beta} = \frac{1}{N} X^T Y - \frac{\lambda}{2} \text{sign}(\hat{\beta})$$

⁴ يرمز 1 إلى متجه طوله N (عدد المشاهدات) وجميع عناصره تساوي واحداً، وعليه، فإن $1^T x_j$ تمثل حاصل جمع قيم المتغير x_j ، ويؤدي هذا الشرط إلى أن يكون المتوسط الحسابي لذلك المتغير مساوياً للصفر.

⁵ يُستخدم افتراض تعامد ومعايرة المتغيرات هنا لأغراض التبسيط الرياضي واشتقاق الحل التحليلي لانحدار لاسو. في التطبيق العملي، تم التعامل مع المتغيرات الرتبوية بعد ترميزها عددياً، في حين حُولت المتغيرات الاسمية إلى متغيرات ثنائية (0-1) قبل إدخالها في النموذج. وقد جرى معايرة المتغيرات الكمية فقط، بينما أُبقيت المتغيرات الثنائية دون معايرة.

$$\hat{\beta}_j = \begin{cases} \frac{1}{N}X^TY + \frac{\lambda}{2} & \text{if } \frac{1}{N}X^TY < -\frac{\lambda}{2} \\ 0 & \text{if } \frac{1}{N}X^TY \leq \frac{\lambda}{2} \\ \frac{1}{N}X^TY - \frac{\lambda}{2} & \text{if } \frac{1}{N}X^TY > \frac{\lambda}{2} \end{cases}$$

بالتالي يمكن ملاحظة أن الحل الناتج عن مقدر لاسو هو الحل نفسه الناتج في الفقرة السابقة أي أن:

$$\hat{\beta}_j^{lasso} = S_{\frac{\lambda}{2}}\left(\frac{X_1^TY}{N}\right)_j = S_{\frac{\lambda}{2}}(\hat{\beta}^{OLS})_j$$

أي بمعنى أنه تم تقدير المعلمة $\hat{\beta}_j^{lasso}$ من خلال تطبيق معامل العتبة الناعمة على مقدر المربعات الصغرى $\hat{\beta}_j^{ols}$ ، وبعبارة أخرى، يمكن الحصول على تقدير المعلمة لمتغير ما لانحدار لاسو عن طريق تطبيق معامل العتبة الناعمة على مقدر المربعات الصغرى لذلك المتغير. وبالتالي، نرى مدى أهمية معامل العتبة الناعمة في حال كون المتغيرات الداخلة في النموذج متعامدة، حيث يمكن تقدير انحدار لاسو بالإستعانة بمقدرات المربعات الصغرى عن طريق معامل العتبة الناعمة.

2.2.8 تقدير انحدار لاسو عند وجود أكثر من متغير مستقل.

توجد العديد من الخوارزميات التي يمكن استخدامها لحل مسألة انحدار لاسو، من أبرزها خوارزمية الزوايا الصغرى (Least Angle Regression LARS)، وأساليب البرمجة التربيعية (Quadratic Programming)، وغيرها من الأساليب الأخرى. يُعتمد في هذا البحث على خوارزمية التراجع الإحداثي (الهبوط الإحداثي) Coordinate descent التي تتدرج تحت مسمى الطرائق التكرارية، وذلك لسهولة استخدامها وقدرتها على التعامل مع عدد كبير جداً من المتغيرات (Chih-Jen Hsieh, et al, 2008, p. 2).

تعتمد خوارزمية التراجع الإحداثي على تقليل دالة الهدف بطريقة تكرارية، حيث يُحدَّث معامل واحد (أو: معلمة واحدة)، بينما تُثبَّت باقي المعاملات (coefficients). وبدلاً من تقليل دالة الهدف لجميع المعاملات في الوقت نفسه، تُقلَّل دالة الهدف لمعامل واحد فقط في كل تكرار، ثم تُحدَّث القيم تباعاً حتى يتحقَّق التقارب إلى الحل الأمثل (Tseng, 2001, p. 480)، أي الوصول إلى الحل الذي يعطي أصغر قيمة ممكنة لدالة الهدف.

يمكن تلخيص خطوات الحل بشكل عام كما يلي:

- 1- افترض حلاً أولياً، تبدأ الخوارزمية بقيم ابتدائية لجميع معاملات نموذج الانحدار ولتكن 0.
- 2- صغّر دالة الهدف وذلك باختيار احداثية (معلمة - متغير) واحدة فقط في كل تكرار، وثبت قيم باقي المعاملات.

3- حدّث قيمة الاحداثيّة المختارة التي تساهم في تقليل دالة الهدف.

4- كرر الخطوة رقم 2 و 3 لكل معامل (إحداثيّة) بالتناوب، إلى أن يتحقّق شرط التقارب⁶.

5- تُمثّل قيم المعاملات الناتجة عند توقّف الخوارزمية معاملات انحدار لاسو.

كما تبين في الفقرة السابقة، في حالة وجود متغير مستقل واحد في النموذج، فإنه من السهل إيجاد مقدّر لاسو ($\hat{\beta}^{LASSO}$) من خلال معامل العتبة الناعمة لمقدّر المربعات الصغرى، وبالتالي يمكن تطبيق خوارزمية التراجع الإحداثي لإيجاد مُقدّر لاسو في حال وجود عدة متغيرات مستقلة بالاستفادة من المعلومة السابقة.

لتكن X مصفوفة قيم المتغيرات المستقلة، و Y متجه لقيم المتغير التابع وقد حولت جميع المتغيرات إلى القيم المعيارية بمتوسط يساوي 0 وتباين يساوي 1، فإنه يمكن كتابة لاسو كما يلي⁷:

$$\beta^{lasso} = \operatorname{argmin} \frac{1}{2} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1$$

$$\beta^{lasso} = \operatorname{RSS}(\beta) + \lambda \|\beta\|_1$$

تطبق أولاً خوارزمية التراجع الإحداثي على الحد الأول لتبسيط خطوات الحل، أي على دالة مجموع مربعات البواقي (RSS) فقط، ثم ننقل إلى تطبيقها على الحد الثاني، وهو القيد المفروض على معادلة التقدير.

2.2.9 تقدير نموذج الانحدار الخطي عن طريق خوارزمية التراجع الإحداثي.

فيما يلي خطوات الخوارزمية بشكل عام:

المدخلات: مجموعة البيانات (Y, X)
المخرجات: متجه للمعاملات المقدرة عن طريق انحدار لاسو $\hat{\beta}^{lasso}$
افتراض حل اولي لقيم معاملات نموذج الانحدار الخطي ولتكن $\beta = 0$.

⁶ في التطبيق العملي، يُستخدم حدّ أقصى كبير لعدد التكرارات كإجراء احترازي لأسباب حسابية في حال عدم تحقّق التقارب، وهو نادراً ما يكون سبب التوقف الفعلي للخوارزمية، وغالباً ما يكون محدداً افتراضياً ضمن المكتبات البرمجية المستخدمة.

⁷ يمكن إضافة أو حذف الكسر $\frac{1}{2}$ أو $\frac{1}{2N}$ إلى المعادلة لتبسيط الحل.

repeat for <i>each</i> $j \in \{1, \dots, p\}$ do Compute the partial residual r_j , where $r_j = \mathbf{Y} - \sum_{l \neq j} \mathbf{X}^l \beta_l$ Compute OLS coefficient for single predictor $(\hat{\beta}_{OLS})_j = \frac{1}{n} r_j^T \mathbf{X}_j$ Update β_j (LASSO solution: single variable case) $\beta_j = \mathbb{S}_{\frac{\lambda}{2}}((\hat{\beta}_{OLS})_j)$ end until <i>convergence</i> ; $\hat{\beta}^{lasso} = \beta$ Return $\hat{\beta}^{lasso}$

وقد أوجدنا سابقاً أن:

$$RSS(\beta) = \frac{1}{2} \sum_{i=1}^N (y_i - \sum_{j=1}^p \beta_j x_{ij})^2$$

تعمل خوارزمية التراجع الإحداثي في كل دورة من الحلقة التكرارية على تقليل مجموع مربعات البواقي (RSS) بالنسبة إلى أحد المعاملات β_j ، من خلال اشتقاق دالة الخطأ اشتقاقاً جزئياً بالنسبة لـ β_j كما يلي:

$$\begin{aligned} \frac{\partial}{\partial \beta_j} RSS(\beta) &= \frac{-2}{2} \sum_{i=1}^N x_{ij} (y_i - \sum_{k \neq j} \beta_k x_{ik} - \beta_j x_{ij}) \\ &= - \sum_{i=1}^N x_{ij} (y_i - \sum_{k \neq j} \beta_k x_{ik}) + \beta_j \sum_{i=1}^N (x_{ij})^2 \end{aligned}$$

يُفترض لأغراض التبسيط، أن $z_j = (x_{ij})^2$ وأن $\rho_j = \sum_{i=1}^N x_{ij}(y_i - \sum_{k \neq j}^N \beta_k x_{ik})$ وهي تعبر عن حاصل ضرب قيم المتغير x_j في البواقي الجزئية للنموذج، أي تلك البواقي التي تم حسابها بعد استبعاد تأثير المتغير x_j . وبذلك يشير $(y_i - \sum_{k \neq j}^N \beta_k x_{ik})$ إلى البواقي الجزئية الناتجة عن النموذج بعد حذف المتغير x_j . بالتالي تصبح العلاقة على الشكل الآتي:

$$^8 \triangleq -\rho_j + \beta_j z_j \quad 2.18$$

2.2.10 تقدير انحدار لاسو عن طريق خوارزمية التراجع الإحداثي.

في الفقرة السابقة، جرى اشتقاق الحد الأول باستخدام خوارزمية التراجع الإحداثي، كما هو موضح في المعادلة رقم (2.18). ويمكن كذلك اشتقاق الحد الثاني بالأسلوب نفسه المستخدم سابقاً، والمبين في المعادلة رقم (2.17)، كما يلي:

$$\lambda \partial_{\beta_j} |\beta_j| = \begin{cases} -\lambda & \text{if } \beta_j < 0 \\ [-\lambda, \lambda] & \text{if } \beta_j = 0 \\ \lambda & \text{if } \beta_j > 0 \end{cases}$$

وعليه بضم النتائج السابقة فإن:

$$\partial_{\beta_j}(\text{lasso}) = \beta_j z_j - \rho_j + \begin{cases} -\lambda & \text{if } \beta_j < 0 \\ [-\lambda, \lambda] & \text{if } \beta_j = 0 \\ \lambda & \text{if } \beta_j > 0 \end{cases}$$

$$= \begin{cases} \beta_j z_j - \rho_j - \lambda & \text{if } \beta_j < 0 \\ [-\rho_j - \lambda, -\rho_j + \lambda] & \text{if } \beta_j = 0 \\ \beta_j z_j - \rho_j + \lambda & \text{if } \beta_j > 0 \end{cases}$$

$$\hat{\beta}_j = \begin{cases} \frac{(\rho_j + \lambda)}{z_j} & \text{if } \rho_j < -\lambda \\ 0 & \text{if } \rho_j \in [-\lambda, \lambda] \\ \frac{(\rho_j - \lambda)}{z_j} & \text{if } \rho_j > \lambda \end{cases}$$

⁸ الرمز $\triangleq$ "يُعرف على أنه" ويُستخدم عندما نُقدم تعريفاً رياضياً جديداً، وليس مجرد مساواة ناتجة عن تحويل أو تبسيط.

يمكن كتابة النتيجة المُستخرجة من خلال معامل العتبة الناعمة، كما في الفقرات السابقة، على النحو الوارد في المعادلة رقم (2.16):

$$\hat{\beta}_j^{lasso} = S_{\frac{1}{z_j}}(\rho_j, \lambda)_j \quad 2.19$$

2.2.11 التفسير الهندسي لمقدر لاسو

يتمثل التفسير الهندسي لمقدر لاسو في رؤية الحل كنقطة تقاطع بين منحنى دالة الهدف ومجموعة القيود المفروضة على معاملات النموذج.

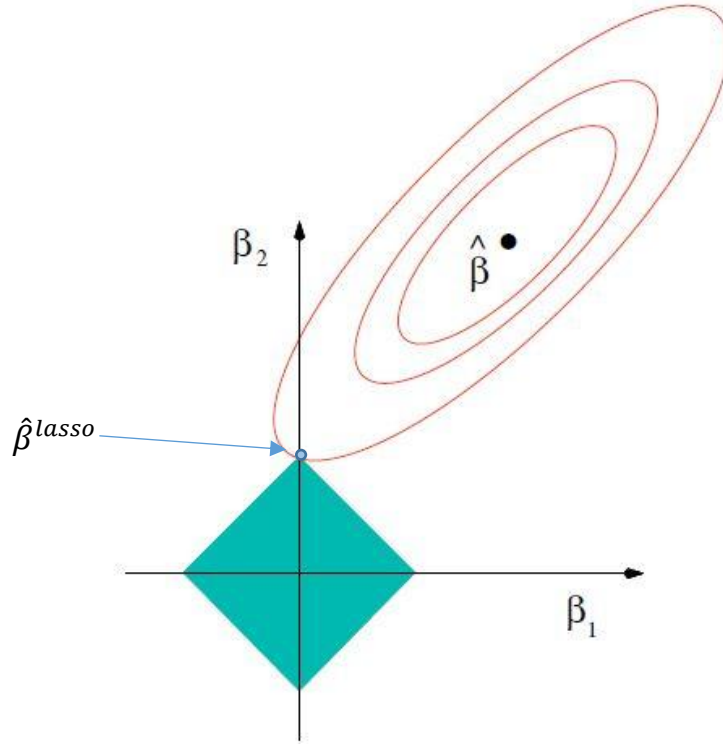
وجدنا سابقاً إنه يمكن كتابة مقدر لاسو كما يلي:

$$\hat{\beta}^{lasso} = \hat{\beta}^{OLS} + \lambda \sum_{j=1}^p |\beta_j|$$

وللسهولة، بإفترض وجود متغيرين مستقلين في النموذج (أي $p = 2$)، ففي هذه الحالة يتمثل الحد الأول في دالة الهدف بمربع الفرق بين القيم الحقيقية والمتوقعة $(y - X\beta)^T(y - X\beta)$ وهو ما يُنتج منحنيات محيطية (Contours) على شكل قطع ناقص (Ellipses) تحيط بنقطة مقدر المربعات الصغرى (OLS). ويرجع هذا الشكل إلى الطبيعة التربيعية لدالة الهدف، حيث إن المنحنيات المحيطية لدوال تربيعية تأخذ عادةً شكل قطع ناقص أو قطوع زائدة، وهو ما يُمثل شكلاً يشبه الوعاء في فضاء ثلاثي الأبعاد.

أما الحد الثاني، والذي يتمثل القيد المفروض على معاملات النموذج، فهو مجموع القيم المطلقة للمعاملات $\lambda \sum_{j=1}^p |\beta_j|$ ، يأخذ شكلاً هندسياً يُعرف باسم المعين (Diamond أو Rhombus) في فضاء البارامترات عندما يكون $p = 2$.

يتمثل الحل الأمثل لمقدر لاسو بنقطة التماس بين القطع الناقص (الذي يمثل دالة الهدف) والمعين (الذي يمثل القيد)، وعندما تقع هذه النقطة على أحد رؤوس المعين (Vertices)، فإن هذا يؤدي إلى أن تأخذ بعض معاملات النموذج القيمة صفر، مما يعني حذف المتغير المرتبط بها من النموذج تلقائياً. يُظهر الشكل الآتي هذا المفهوم الهندسي بوضوح.

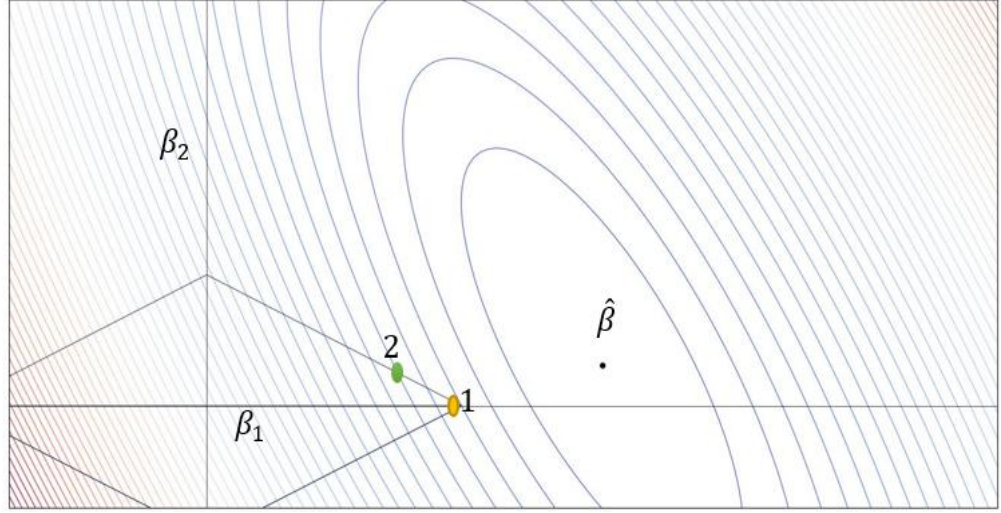


الشكل (5): التفسير الهندسي لانحدار لاسو - حالة حذف أحد المتغيرات.

المصدر : Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical

Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.

نلاحظ وجود عدة حلول ممكنة لمقدر المربعات الصغرى، والممثلة بمنحنيات باللون الأحمر. تمثل النقطة السوداء $\hat{\beta}$ القيمة المقدرة لـ المعلمة β_1 و β_2 التي تحقق أقل قيمة لمجموع مربعات البواقي، وكلما ابتعدنا عن الحل الأمثل (النقطة السوداء) كلما ازداد مجموع مربعات البواقي. أما المنطقة المظللة على شكل مُعَيَّن، فتُمثل مجموعة القيود المفروضة على المعلمات وفقاً لانحدار لاسو، والتي تُحدّد بأن يكون مجموع القيم المطلقة للمعاملات أقل من حد مُعَيَّن. وعليه، فإن الحل المُمثّل بالنقطة الزرقاء، يُمثّل نقطة التماس بين القيود المفروضة على دالة الهدف ومجموع المربعات الصغرى. يتضح من الشكل أن تقاطع القيود مع منحنيات المربعات الصغرى قد يحدث في رأس المُعَيَّن، ما يؤدي إلى تصفير أحد المعاملات (كما في الحالة الموضّحة حيث $\beta_1 = 0$). وهذه الخاصية تميز انحدار لاسو بقدرته على اختيار المتغيرات المهمة واستبعاد غير المؤثرة منها. مع ذلك، تجدر الإشارة إلى أنه ليس من الضروري دائماً أن يقع الحل عند رأس المُعَيَّن، إذ يمكن أن تكون نقطة التقاطع في أي موضع ضمن الحدود، وهو ما يعني أن بعض التقديرات قد لا تُحذف بالكامل، والشكل الآتي يوضح ما سبق:



الشكل (6): التفسير الهندسي لانحدار لاسو - وجود حلين محتملين أحدهما يحذف متغيراً.

المصدر : Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical

.Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer

حيث يؤدي الحل الناتج عند النقطة (1) إلى حذف المتغير β_1 من النموذج، في حين الحل الناتج عند النقطة (2) لن يؤدي إلى حذف أي متغير من النموذج لكن ستكون دالة الهدف أكبر مقارنة بالحل الناتج عند النقطة (1).

بالإضافة إلى التفسير الهندسي، يمكن النظر بشكل عام إلى انحدار لاسو على النحو الآتي، إذا فُرض أن المجموعة النشطة (active set) $\mathcal{A}$ تضم المتغيرات الداخلة في النموذج والتي لم تحذف عند تطبيق انحدار لاسو، أي أن $\mathcal{A} = \{j : \hat{\beta}_j^* \neq 0\}$ ، بالتالي فإن $\hat{\beta}^{LASSO}$ يقلل دالة الهدف إذا حقق الشروط التالية (Tibshirani R. J., 2012, p. 4):

$$j \in \mathcal{A} \text{ if } -2X^T(Y - X\hat{\beta}) + \lambda \text{sign}(\hat{\beta}) = 0$$

$$j \notin \mathcal{A} \text{ if } |-2X^T(Y - X\hat{\beta})| \leq \lambda$$

أي أن الشرط الكافي لعدم حذف متغير ما من النموذج عند تطبيق طريقة لاسو، هو أن يكون حاصل ضرب $|-2X^T(Y - X\hat{\beta})|$ أكبر من قيمة معلمة الضبط λ ، وإلا سيتم استبعاده من النموذج، أي لا يدخل في المجموعة النشطة. وبعبارة أخرى، إذا فُرضت قيمة ما لمعلمة الضبط λ ، فيمكن تطبيق إنحدار لاسو استناداً للشرط السابق، بحيث يُدرج متغير ما ضمن المجموعة النشطة إذا كان حاصل الضرب $|-2X^T(Y - X\hat{\beta})|$ أكبر من قيمة معلمة الضبط λ . وبالتالي، يتّضح مدى أهمية تحديد قيمة معلمة الضبط λ ، والتي سيتم التطرق إلى طرائق تحديدها في المبحث التالي.

2.2.12 مزايا ومساوئ مقدر لاسو

مزايا مقدر لاسو

- 1- تبسيط النماذج والتفسير: يساهم مقدر لاسو في حذف المتغيرات غير الهامة من النموذج عن طريق تقليص بعض معاملات الانحدار إلى الصفر، مما ينتج عنه نموذج يحتوي على عدد أقل من المتغيرات وبالتالي يُسهل فهم النتائج وتفسيرها.
- 2- استقرار إشارة المعاملات: لا يمكن أن تختلف إشارة المعلمات المقدرة لأكثر من مجموعة من الحلول الناتجة عند استخدام انحدار لاسو عندما يكون عدد المتغيرات الداخلة في النموذج المراد تقديرها أكبر من حجم العينة ($P > N$)، فعلى سبيل المثال، إذا كان β_j^1 و β_j^2 يمثلان مجموعتي حلول تقديرية ناتجتين عن خوارزميات مختلفة، فإن حاصل ضربيهما يكون دائماً غير سالب ($\beta_j^1 \cdot \beta_j^2 \geq 0$) (Tibshirani R. J., 2012, p. 14) وذلك عند كل قيمة ثابتة لـ λ . بالتالي لا توجد تناقضات في إشارة المتغير المقدر بالرغم من اختلاف الخوارزمية المستخدمة في عملية التقدير.
- 3- القدرة على التعامل مع البيانات ذات الأبعاد العالية: يتميز مقدر لاسو بفعالته في معالجة الحالات التي يكون فيها عدد المتغيرات كبيراً مقارنة بعدد المشاهدات، مما يجعله أداة قوية لتحليل البيانات في تلك الحالة.

مساوئ مقدر لاسو

- الاختيار الاعباطي للمتغيرات المترابطة: في حال وجود ارتباط خطي قوي بين اثنين أو أكثر من المتغيرات المستقلة، فإن مقدر لاسو لا يحتفظ بها جميعاً في النموذج، بل يميل إلى اختيار أحدها وحذف البقية، بطريقة تكون اعتباطية لا تعتمد على تمايز جوهري بينها. ويعود ذلك إلى طبيعة القيد المفروض (الجزء - العقوبة) لإنحدار لاسو، التي تشجع على حذف أكبر عدد ممكن من المعاملات. ونتيجة لذلك، قد يُستبعد أحد المتغيرات المهمة فقط لأنه يرتبط بمتغير آخر تم الاحتفاظ به، ما قد يؤثر سلباً على التفسير.

المبحث الثالث

اختيار معلمة الضبط

تمهيد

تلعب معلمة الضبط (λ) دوراً أساسياً في طريقة انحدار لاسو، إذ تتيح التحكم بدرجة تقلص معاملات نموذج الانحدار نحو الصفر، ومن ثم إمكانية حذفها كلياً من النموذج. فكلما كانت قيمة λ كبيرة، ازداد مقدار الجزء (التقليص) المفروض على المعادلة، أي الحصول على عدد أقل من المتغيرات الداخلة في النموذج. في المقابل، كلما صغرت قيمة λ ، تراجع أثر تقليص (أو حذف) بعض المعاملات من نموذج الانحدار. وعند مساواة λ للصفر، يُقدّر النموذج باستخدام جميع المتغيرات المستقلة المدروسة، أي دون حذف أي متغير منها، وتُقدر معاملات نموذج الانحدار كاملة.

وبناءً على ما سبق، يتّضح مدى أهمية اختيار القيمة المناسبة لـ λ ، نظراً للدور الحاسم الذي تؤديه في طريقة انحدار لاسو. لذلك، سَيتطرق في الفقرات التالية إلى أسلوب التحقق المتقاطع (Cross-Validation)، بالإضافة إلى طرائق معايير المعلومات (Information Criteria)، مثل معياري AIC و BIC، لاختيار معلمة الضبط λ المناسبة.

2.3.1 طريقة التحقق المتقاطع (Cross-Validation)

تُعدّ طريقة التحقق المتقاطع إحدى طرائق إعادة المعاينة (Resampling Methods)، حيث تعتمد على تقسيم البيانات المتاحة إلى عينات جزئية تُستخدم بشكل متكرر لبناء النموذج أو تقييم أدائه. وتكمن الفكرة الأساسية في سحب عينات فرعية متعددة من مجموعة البيانات الأصلية، ثم استخدام كل عينة لتدريب النموذج واختباره على جزء مختلف من البيانات.

من أبرز مزايا هذه الطريقة أنها تُستخدم عندما يكون حجم البيانات محدوداً، إذ تتيح الاستفادة القصوى من البيانات المتوفرة عن طريق إعادة استخدامها في عمليات التقدير والتقييم.

ومع ذلك، تُعدّ من الناحية الحسابية طريقة مكلفة، خصوصاً إذا كان النموذج المستخدم يتطلب وقتاً أو موارد كبيرة في كل مرة يُعاد فيها بناؤه، نظراً لأن التحقق المتقاطع يتطلب بناء النموذج مرات متعددة.

يمكن استخدام طرائق إعادة المعاينة بشكل عام في عدة سياقات، مثل قياس الدقة من خلال حساب الأخطاء المعيارية للمقدرات، أو في اختيار النموذج عبر المفاضلة بين طرائق تقدير مختلفة، أو أيضاً في تقييم النموذج

من خلال التحقق من جودة أدائه. وتُعدّ الميزة الرئيسية لهذه الطرائق أنها لا تقتصر على أي شكل توزيعي مسبق للبيانات، ما يجعلها مناسبة في الحالات التي لا تتوفر فيها معلومات كافية عن التوزيع الإحصائي الأساسي.

تكمن فكرة التحقق المتقاطع في تقسيم البيانات المدروسة إلى مجموعتين: مجموعة تدريب يُبنى من خلالها النموذج الإحصائي، ومجموعة اختبار تُستخدم لتقييم مدى ملاءمة وكفاءة هذا النموذج، وذلك باستخدام معيار معين مثل متوسط مربعات الأخطاء. وتُكرّر هذه العملية عدة مرات، بحيث يُدرّب النموذج في كل مرة على مجموعة تدريب مختلفة، ويُقيّم أدائه على مجموعة الاختبار المقابلة.

تُعدّ تقنية التحقق المتقاطع من الأدوات الفعّالة لتحسين أداء النماذج الإحصائية عموماً، والمساهمة في اتخاذ قرارات مناسبة تعزّز من كفاءة النموذج. ويُطبّق التحقق المتقاطع غالباً من خلال تقسيم البيانات إلى K جزءاً، تُعرف باسم "الطيات" (Folds)، حيث يُخصّص جزء واحد كمجموعة اختبار، وتُستخدَم الأجزاء المتبقية وعددها $K-1$ كمجموعات تدريب. ثم يُحسب متوسط مربعات الأخطاء لكل مجموعة اختبار، وتُكرّر هذه العملية عبر جميع الطيات، ليُحسب في النهاية متوسط عام لمربعات الأخطاء (Tibshirani R. J., 2021, p. 1).

وتوجد عدة أساليب لتقسيم البيانات، ويُعدّ أسلوب "K-Folds" من أكثرها استخداماً وشيوعاً، وسيُتناول هذا الأسلوب بالتفصيل في الفقرة الآتية.

2.3.2 التحقق المتقاطع بأسلوب (K-Folds Cross-Validation)

فيما يلي خطوات تطبيق التحقق المتقاطع:

1. تقسيم البيانات إلى K جزء متساوية الحجم (4 أو 5 أو 10 أجزاء) (Kohavi, 1995).
2. لكل جزء $K = 1 \dots K$:

 - a. اعتباره بمثابة مجموعة الاختبار، والأجزاء الأخرى بمثابة مجموعات (بيانات) التدريب.
 - b. تطبيق طريقة انحدار لاسو على بيانات التدريب وذلك لجميع قيم λ المختلفة، وحساب MSE لمجموعة الاختبار.

3. حساب متوسط قيم متوسط مربعات الأخطاء التي تم الحصول عليها في الخطوة السابقة، وذلك لكل قيمة من قيم λ .
4. الحصول على منحنى خطأ التحقق المتقاطع (cross-validation error curve)، وهو عبارة عن دالة تربط قيم λ المختلفة بقيم الخطأ المحسوبة سابقاً.
5. اختيار قيمة λ التي يكون عندها المنحنى أقل ما يمكن.

بالتالي بعد الحصول على قيمة λ في الخطوة الأخيرة، التي تجعل منحني خطأ التحقق المتقاطع أقل ما يمكن، يُطبَّق انحدار لاسو على البيانات المدروسة باستخدام تلك القيمة.

2.3.3 الأخطاء المعيارية للتحقق المتقاطع (K-Folds Cross-Validation)

يتيح حساب الخطأ المعياري للتحقق المتقاطع ميزة إضافية لاختيار معلمة الضبط λ . إذ يُحسب أولاً متوسط مربعات الأخطاء (MSE) لكل جزء (FOLD) من أجزاء التحقق المتقاطع، وذلك لكل قيمة $\lambda \in (\lambda_1, \dots, \lambda_m)$ الموضحة في الخطوات السابقة، ويُرمز لقيمة متوسط مربعات الأخطاء في الجزء k بالرمز $MSE_k(\lambda)$.

ثم يُحسب متوسط قيم $MSE_k(\lambda)$ عبر الأجزاء للحصول على خطأ التحقق المتقاطع، ويُرمز له بـ $CV(\lambda)$.

وبعد ذلك يُحسب الانحراف المعياري لقيم متوسط مربعات الأخطاء السابق وذلك لكل قيمة λ :

$$SD(\lambda) = \sqrt{\text{var}(MSE_1(\lambda), \dots, MSE_k(\lambda))} \quad 2.20$$

وأخيراً يمكن تقدير الخطأ المعياري للتحقق المتقاطع كما يلي

$$SE(\lambda) = SD(\lambda)/\sqrt{K} \quad 2.21$$

يمكن استخدام قاعدة الخطأ المعياري الواحد (The One Standard Error Rule) في التحقق المتقاطع كوسيلة بديلة لاختيار معلمة الضبط λ ، موازية للطريقة السابقة. وتقوم هذه القاعدة على اختيار النموذج الذي تقع فيه قيمة λ ضمن مجال لا يتجاوز خطأ معيارياً واحداً ($SE \mp 1$) من القيمة التي تحقق أدنى متوسط لمربعات الأخطاء (MSE).

2.3.4 طرائق معايير المعلومات (Information Criteria)

تعتبر معايير المعلومات من أدوات اختيار النموذج، التي يمكن على أساسها الاختيار ما بين النماذج أو المقارنة بينهما واختيار أفضل نموذج يحقق أغراض التنبؤ، حيث تهدف أغلب المعايير المستخدمة إلى تقليل حجم مجموع مربعات البواقي (RSS)، وعلى وجه التحديد سيذكر معيار AIC ومعيار BIC لاختيار معلمة الضبط λ في انحدار لاسو.

تتميز معايير المعلومات بسهولة حسابها واستخدامها موارد حسابية أقل مقارنةً من التحقق المتقاطع، فبعد تقدير النماذج يمكن بسهولة حساب معايير المعلومات والمقارنة بينهم واختيار النموذج الذي يتميز بأقل قيمة ممكنة للمعايير المستخدمة. ومع ذلك، تكمن الصعوبة في حساب درجات الحرية المستخدمة في تقدير المعايير. إن خصائص معيار AIC (Akaike, 1973) مدروسة ومستخدمة في الدراسات الأكاديمية بشكل جيد وهو مناسب

لأغراض التنبؤ والمقارنة بين النماذج، ولكن اقترح الباحث Zou تعديلاً بسيطاً في طريقة حساب المعيار لاستخدامه في انحدار لاسو (Hui Zou, 2007) وفق الصيغة الآتية:

$$AIC(\lambda) = \frac{\|y - \hat{\mu}\|^2}{n\sigma^2} + \frac{2}{n} \widehat{df}(\hat{\mu}) \quad 2.22$$

حيث أن:

y : متجه لقيم المتغير التابع.

μ : عبارة عن النتائج المقدرة من خلال تطبيق طريقة لاسو $(X\hat{\beta}(\lambda))$.

df : عبارة عن درجات حرية النموذج، وهي عبارة عن عدد المتغيرات التي لم تحذف من النموذج بتطبيق طريقة لاسو.

σ^2 : يُمثل التباين المقدّر، ويُحسب باستخدام العلاقة $\|y - X\beta\|^2/n$

حيث أُضيف الحد $\frac{2}{n} \widehat{df}(\hat{\mu})$ إلى المعادلة الأساسية باعتباره عامل تصحيح يعتمد على عدد المعاملات المدرجة في النموذج فقط.

وبالمثل، تم اقتراح تعديل معيار BIC (Schwarz) ليناسب انحدار لاسو كما يلي:

$$BIC(\lambda) = \frac{\|y - \hat{\mu}\|^2}{n\sigma^2} + \frac{\log(n)}{n} \widehat{df}(\hat{\mu}) \quad 2.23$$

مع استخدام نفس الرموز والتعاريف المذكورة سابقاً.

حيث أُضيف الحد $\frac{\log(n)}{n} \widehat{df}(\hat{\mu})$ إلى المعادلة الأساسية باعتباره عامل تصحيح يعتمد على عدد المعاملات المدرجة في النموذج فقط.

وأيضاً اقترح الباحث Wang وآخرون صيغة معدلة للمعيار السابق (Wang, Li & Leng, 2009) كما يلي:

$$Modified\ BIC(\lambda) = \log\left(\frac{\|y - \hat{\mu}\|^2}{n}\right) + \frac{\log(n)}{n} \widehat{df}(\hat{\mu}) \times C_n \quad 2.24$$

مع استخدام نفس الرموز والتعاريف المذكورة سابقاً.

C_n : عبارة عن قيمة ثابتة موجبة أكبر من الصفر.

وقد بيّن Zou، في الدراسة نفسها المتعلقة بتعديل معايير اختيار النموذج في سياق انحدار لاسو، أن معيار BIC المعدل، متسق عندما يكون حجم العينة أكبر من عدد المتغيرات الداخلة في النموذج وأستخدم قيمة الثابت

الفصل الثالث: القسم العملي

بناء النماذج الإحصائية لتقدير أسعار العقارات.

المبحث الأول: تطبيق انحدار لاسو لتقدير أسعار العقارات في مدينة Ames.

المبحث الثاني: طرائق التقييم العقاري.

المبحث الثالث: تطبيق انحدار لاسو لتقدير أسعار العقارات في مدينة دمشق ومقارنته بانحدار OLS.

الفصل الثالث

بناء النماذج الإحصائية لتقدير أسعار العقارات

يُركّز هذا الفصل على الجانب العملي من الدراسة، حيث طُبقت نماذج إحصائية لتقدير أسعار العقارات استناداً إلى بيانات فعلية من بيئتين مختلفتين: بيئة دولية وبيئة محلية. ويهدف الفصل إلى بناء نماذج قادرة على تحديد العوامل الأكثر تأثيراً في السعر، وتقييم قدرتها على التنبؤ الدقيق، مع تسليط الضوء على مزايا وقيود كل طريقة في السياقات المختلفة.

يتكوّن هذا الفصل من ثلاثة مباحث رئيسية:

- **المبحث الأول** يعرض خطوات بناء انحدار لاسو باستخدام قاعدة بيانات مفتوحة المصدر من مدينة Ames الأمريكية. وأُختيرت هذه البيانات لأنها تُستخدم بشكل واسع لاختبار نماذج تسعير العقارات، وتحتوي على معلومات مفصلة عن خصائص العقارات.
 - **المبحث الثاني** يتناول الطرائق المعتمدة في تقييم العقارات، بما في ذلك طرائق المقارنة والتكلفة، بهدف توضيح الفروق بينها وبين النماذج الإحصائية من حيث المنهج والنتائج.
 - **المبحث الثالث** يقدّم تطبيقاً عملياً على سوق العقارات في مدينة دمشق، استناداً إلى بيانات ميدانية جمعها الباحث. ويتضمن هذا المبحث بناء انحدار لاسو وتقديره، إلى جانب انحدار خطي متعدد لمقارنة النتائج بين النموذجين في بيئة محلية تتسم بخصوصية معينة في طبيعة السوق والبيانات.
- يعتمد هذا الفصل على أدوات ومكتبات Python لبناء النماذج وتحليل النتائج، ويهدف إلى تقديم قراءة واضحة للعوامل المؤثرة في أسعار العقارات، ومدى إمكانية الاعتماد على النماذج الإحصائية الحديثة في هذا المجال.

المبحث الأول

تطبيق انحدار لاسو لتقدير أسعار العقارات في مدينة Ames

تمهيد

يشكل هذا المبحث التطبيق العملي للدراسة، ويهدف إلى بناء نموذج إحصائي يُستخدم في تقدير القيمة الحقيقية لأسعار العقارات، استناداً إلى بيانات عقارية فعلية. وفي هذا السياق، أُعتمدت قاعدة بيانات مدينة Ames الأمريكية لما تتميز به من شمولية وتنوع في الخصائص العقارية، فضلاً عن استخدامها الواسع كمرجعية معيارية في البحوث المتعلقة بنمذجة أسعار المساكن.

يتضمن هذا المبحث عرضاً مفصلاً لمراحل تحليل البيانات بدءاً من قراءة البيانات الخام، مروراً بخطوات تنظيفها ومعالجتها، ثم تقسيمها إلى مجموعات تدريب واختبار. كما يُقدم ملخصات إحصائية لمتغيرات الدراسة، وتحليل التوزيع والارتباط بينها، مع التركيز على العلاقة بين متغير السعر والمتغيرات المستقلة ذات العلاقة.

بعد ذلك، يُقدّر انحدار لاسو باستخدام خوارزمية التراجع الإحداشي مع توظيف أدوات ومكتبات بايثون الإحصائية. وتُجرى عملية ضبط المعامل λ عبر أسلوب التحقق المتقاطع لاختيار النموذج الأمثل من حيث القدرة التنبؤية وتقليل عدد المتغيرات غير الجوهرية.

يختتم المبحث الأول بتحليل نتائج النموذج المقدّر، وبيان مدى كفاءته في تفسير سلوك أسعار العقارات استناداً إلى مؤشرات مثل RMSE وعدد المتغيرات المُحتفظ بها، مع إبراز مساهمة انحدار لاسو في معالجة التعددية الخطية وانتقاء الخصائص الأكثر تأثيراً.

3.1.1 الأدوات والبرامج الإحصائية المستخدمة

تم الاعتماد بشكل أساسي على لغة البرمجة Python الإصدار 3.0، لما توفره من بيئة متكاملة لتحليل البيانات وتقدير النماذج الإحصائية، ولاسيما انحدار لاسو، الذي يعتمد على خوارزميات تكرارية والتحقق المتقاطع في اختيار معلمة الضبط، إلى جانب إمكانية تقسيم البيانات وتقييم الأداء التنبؤي للنماذج، وهو ما يصعب تحقيقه بالكفاءة نفسها باستخدام بعض البرمجيات الإحصائية التقليدية. وفيما يلي مجموعة من الأدوات الإحصائية والمكتبات الجاهزة:

الجدول (1): الأدوات والمكتبات الإحصائية المستخدمة في تحليل البيانات الخاصة بمدينة Ames.

اسم المكتبة	الوصف - الغرض من المكتبة	المصدر
Pandas	مكتبة تُستخدم بشكل أساسي للتعامل مع البيانات وتحليلها. توفر هذه المكتبة أدوات قوية لقراءة وكتابة البيانات من وإلى مصادر متعددة وتقديم وظائف متقدمة لتنظيف وتحليل البيانات.	https://pandas.pydata.org/
Numpy	مكتبة تُستخدم للعمليات الرقمية والحسابية. توفر هذه المكتبة هياكل بيانات متعددة الأبعاد مثل المصفوفات (arrays) وغيرها.	https://numpy.org/
Sklearn	مكتبة شاملة من خوارزميات تعلم الآلة، واعتمد عليها لتطبيق انحدار لاسو.	https://scikit-learn.org/stable/
Seaborn	مكتبة لإعداد الرسومات البيانية بأشكال مختلفة.	https://seaborn.pydata.org/
Matplotlib.Pyplot	مكتبة لإعداد الرسومات البيانية بأشكال مختلفة.	https://matplotlib.org/

المصدر: إعداد الباحث بالاعتماد على الوثائق الرسمية لكل مكتبة برمجية.

3.1.2 متغيرات الدراسة

تتألف البيانات من 80 متغيراً بالمجمل بحجم (2899) مشاهدة وهي عبارة عن جميع المبيعات التي حدثت داخل مدينة Ames، من عام 2006 إلى عام 2010. حيث تدل البيانات على جودة وخصائص العديد من السمات المادية للعقارات، وتمثل معظم المتغيرات الخصائص التي يرغب مشتري العقار في معرفتها عن عقار محتمل (سنة بناء العقار، مساحة العقار، عدد الحمامات..).

يوجد بشكل عام 34 متغيراً كمياً (quantitative)، ركزت معظم المتغيرات على مساحة العقار الإجمالية، ومساحة الكراج لكل عقار، ومساحة الطابق السفلي والعلوي، مساحة القبو، بالإضافة إلى عدد المطابخ وغرف النوم والحمامات، وغيرها الكثير.

تضمّنت قاعدة البيانات عدداً كبيراً من المتغيرات الفئوية بلغ 46 متغيراً، ركّز معظمها على الخصائص البيئية المحيطة بالعقار، ونوع العقار، ووجود بركة سباحة، والشكل الخارجي للبناء، بالإضافة إلى تقييم جودة الإكساء وعدد من السمات الأخرى ذات الصلة بالتصميم.

3.1.3 معالجة البيانات

إن معالجة البيانات خطوة حاسمة لضمان جودة وموثوقية النتائج المستخلصة. تشمل معالجة البيانات تنظيف البيانات والتي تهدف إلى إزالة القيم المفقودة، أو غير الصحيحة، أو غير المنطقية من البيانات، وتحويلها بطريقة مناسبة للاستخدام في تحليل البيانات. وفيما يلي، الخطوات الأساسية لعملية تنظيف البيانات:

3.1.3.1 تقسيم البيانات

تُعدّ عملية تقسيم مجموعة البيانات إلى مجموعتي تدريب واختبار من الخطوات الأساسية، التي جرى التطرّق إليها سابقاً، لما لها من أهمية في تقييم أداء النماذج وقدرتها على التعميم بدقة (Daniel T. Larose, 2014). وفي هذا السياق، قُسمت مجموعة بيانات Ames إلى مجموعتين: مجموعة تدريب بحجم 80% (2318 مشاهدة)، ومجموعة اختبار بحجم 20% (581 مشاهدة).

والجدير بالذكر أن جميع العمليات المرتبطة بمعالجة البيانات نُفذت لكل مجموعة بشكل مستقل، وذلك بهدف الحدّ من احتمالية تسرب المعلومات/البيانات (Data Leakage) بين المجموعتين (M. A. Bouke, 2024).

3.1.3.2 التعامل مع القيم المفقودة

يُعدّ وجود القيم المفقودة من الظواهر الشائعة في مجموعات البيانات، وقد يؤثر سلباً على نتائج التحليل الإحصائي (van Buuren, 2018, p. 10). ولمعالجة هذه المشكلة، جرى تحديد المتغيرات التي تحتوي على قيم مفقودة من خلال حساب نسبة القيم المفقودة في كل متغير. واستُبعدت من الدراسة المتغيرات التي تجاوزت نسبة القيم المفقودة فيها 60%، وهي: تقييم جودة بركة السباحة (POOLQC)، ميزات أخرى (MiscFeature)، نوع الزقاق (ALLEY)، وجودة سياج للعقار (FENCE). وقد أُجري اختبار تضمين هذه المتغيرات واستبعادها، ولم تُظهر النتائج اختلافاً ملحوظاً في الحاليتين.

أما بالنسبة للمتغيرات التي تقل فيها نسبة القيم المفقودة عن 60%، فقد استُبدلت القيم المفقودة في المتغيرات الكمية بقيمة الوسيط، بينما استُبدلت القيم المفقودة في المتغيرات الفئوية بالقيمة الأكثر تكراراً. وتجدر الإشارة إلى أن النتائج لم تختلف بشكل كبير عند استبعاد المتغيرات التي بلغت نسبة القيم المفقودة فيها أقل من 17%.

وفيما يلي جدول يوضح النسب التي جرى حسابها لكل متغير يحتوي على قيم مفقودة:

الجدول (2): نسب القيم المفقودة لمتغيرات الدراسة الخاصة في مدينة Ames.

المتغير	نوع المتغير	نسبة القيم المفقودة	المتغير	نوع المتغير	نسبة القيم المفقودة
PoolQC جودة بركة السباحة	فئوي	99.6%	BsmtCond تقييم حالة القبو العامة.	فئوي	3%
MiscFeature ميزات أخرى.	فئوي	96%	BsmtFinType1 تقييم درجة تشطيب المساحة المنتهية من القبو.	فئوي	3%
Alley نوع الزقاق	فئوي	93%	MasVnrArea مساحة قشرة البناء.	كمي	1%
Fence جودة سياج العقار	فئوي	80%	BsmtFullBath عدد حمامات القبو	كمي	0.07%
MasVnrType نوع قشرة البناء	فئوي	61%	BsmtHalfBath عدد حمامات القبو النصفية.	كمي	0.07%
FireplaceQu جودة الموقد.	فئوي	49%	GarageCars عدد المواقف المتاحة في المرآب	كمي	0.03%
LotFrontage مساحة الفناء	كمي	17%	GarageArea مساحة المرآب	كمي	0.03%

0.03%	كمي	TotalBsmtSF مساحة القبو	5%	كمي	GarageYrBltd تاريخ بناء المرآب
0.03%	كمي	BsmtUnfSF مساحة القبو غير المكتملة	5%	فئوي	GarageCond حالة المرآب
0.03%	كمي	BsmtFinSF2 مساحة إضافية مُجهَّزة في القبو	5%	فئوي	GarageQual جودة المرآب
0.03%	كمي	BsmtFinSF1 المساحة الأساسية المُجهَّزة في القبو	5%	فئوي	GarageFinish ديكور المرآب
0.03%	فئوي	Electrical النظام الكهربائي	5%	فئوي	GarageType نوع المرآب
			3%	فئوي	BsmtExposure مستوى الحديقة
			3%	فئوي	BsmtFinType2 تقييم القبو
			3%	فئوي	BsmtQual تقييم ارتفاع القبو.

المصدر: إعداد الباحث اعتماداً على قاعدة بيانات مدينة AMES باستخدام لغة البرمجة PYTHON

3.1.3.3 ترميز المتغيرات الفئوية

رُمِّزت جميع المتغيرات الفئوية، والبالغ عددها 46 متغيراً، والتي كان مقياسها إما اسمياً (Nominal) أو ترتيبياً (Ordinal)، إذ تم استخدام الترميز البوليني للمتغيرات الاسمية، سواءً كانت ثنائية أم متعددة الفئات، في حين استخدم الترميز الترتيبي للمتغيرات ذات المستويات المرتبة.

3.1.3.4 عمليات التحويل على المتغيرات

تُظهر نتائج الإحصاء الوصفي للمتغير التابع، التي سيتم عرضها في الفقرة اللاحقة، أن توزيع أسعار مبيع العقار ملتوي نحو اليمين، الأمر الذي يبرر استخدام التحويل اللوغاريتمي.

بالإضافة لتحويل قيم المتغيرات المستقلة إلى القيم المعيارية وذلك وفقاً لأفضل الممارسات المنصوح باتباعها (James, et al, 2021, pp. 70–79).

3.1.4 عرض لأهم متغيرات الدراسة.

تعرض هذه الفقرة الإحصاء الوصفي لبعض المتغيرات، والتي تشمل مقاييس النزعة المركزية والتشتت، بالإضافة لعرض خصائص التوزيع لكل متغير ذي الصلة. وأيضاً عرض بعض الاختبارات الإحصائية مثل معامل ارتباط بيرسون، لقياس درجة الارتباط بين المتغيرات المستقلة والمتغير التابع (سعر العقار)، حيث يمكن من خلاله ليس فقط فهم العلاقات بين المتغيرات، ولكن أيضاً التعرف على إمكانية حدوث التعددية الخطية.

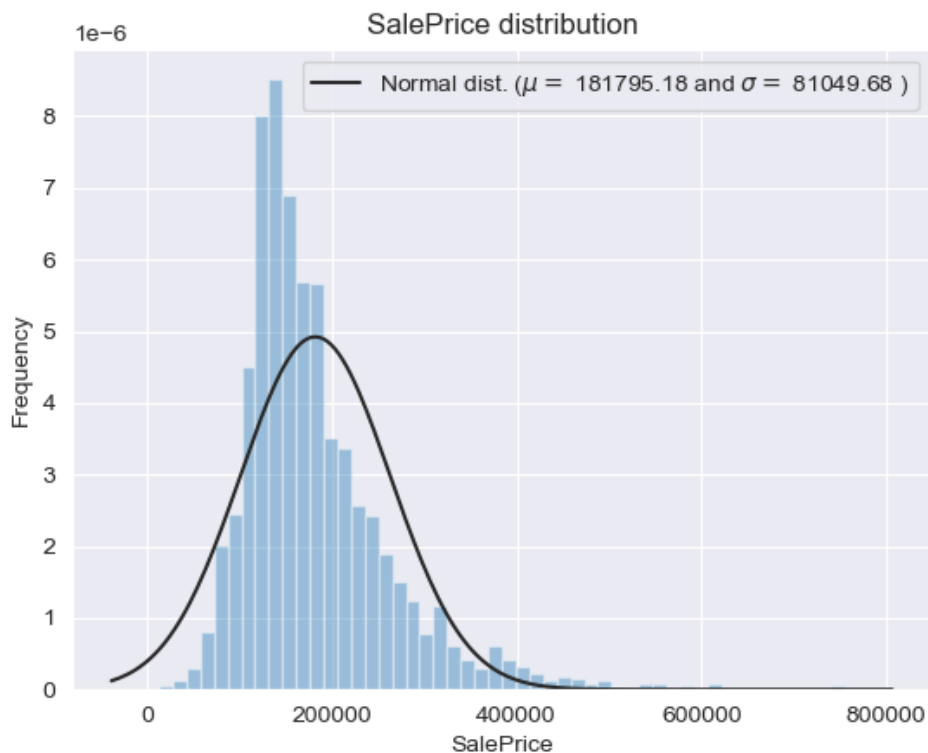
3.1.5 الإحصاء الوصفي للمتغير التابع (سعر العقار)

الجدول (3): الإحصاء الوصفي لمتغير سعر العقار الخاصة في مدينة Ames.

المتوسط	الانحراف المعياري	MIN	MAX
181795	81067	12789	755000

المصدر: إعداد الباحث اعتماداً على قاعدة بيانات مدينة AMES باستخدام لغة البرمجة PYTHON

وفيما يلي شكل التوزيع التكراري لمتغير سعر العقار



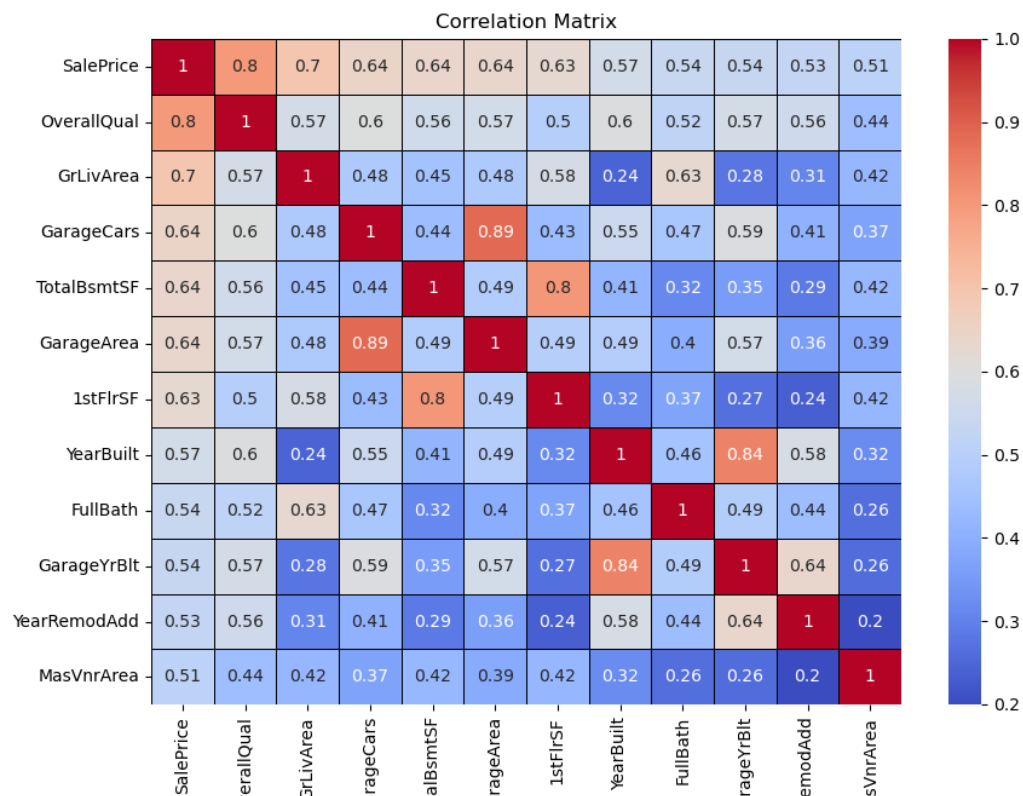
الشكل (7): التوزيع التكراري لمتغير سعر العقار في مدينة Ames مع منحنى التوزيع الطبيعي التقريبي.

المصدر: إعداد الباحث اعتماداً على قاعدة بيانات مدينة Ames باستخدام لغة البرمجة Python

ويلاحظ أن قيمة متوسط سعر العقار \$181795 بانحراف معياري \$81067، وأن شكل التوزيع لقيم المتغير ملتوٍ نحو اليمين، أي بمعنى هناك عدد قليل من الأشخاص الذين يستطيعون شراء عقارات بأسعار مرتفعة.

3.1.6 مصفوفة الارتباط بين متغير سعر العقار والمتغيرات الكمية.

اعتمد الباحث على معامل ارتباط بيرسون لاختبار شدة العلاقة بين المتغيرات المدروسة. ويوضح الشكل البياني التالي (خريطة حرارية / Heat Map) لقيم معاملات الارتباط، مع التركيز على العلاقات التي تجاوزت قيمة ارتباط قدرها 0.50.



الشكل (8): خريطة حرارية (Heatmap) لمصفوفة الارتباط بين المتغيرات الكمية وسعر العقار في بيانات Ames.

المصدر: إعداد الباحث اعتماداً على قاعدة بيانات مدينة Ames باستخدام لغة البرمجة Python

تُبين مصفوفة الارتباط بين المتغيرات المدروسة وجود علاقة ارتباطية قوية بين 11 متغيراً وبين سعر مبيع العقار، حيث تجاوز معامل الارتباط في كل منها القيمة 0.50. وقد كانت أقوى هذه العلاقات مع متغير الجودة الشاملة (Overall Quality)، الذي يُستخدم لتقييم إكساء العقار وجودة مواد الإنشائية، إذ بلغ معامل الارتباط نحو 0.80. تلي ذلك العلاقة مع مساحة المعيشة الإجمالية (GrLivArea)، التي بلغ معامل ارتباطها نحو 0.70، ثم العلاقة مع استيعاب المرآب لعدد المواقف المتاحة (GarageCars) نحو 0.64، ومساحة المرآب (GarageArea) نحو 0.64، إضافةً إلى إجمالي مساحة القبو (TotalBsmtSF) التي سجلت المعامل نفسه. أما المتغيرات الأخرى التي أظهرت ارتباطاً أعلى من 0.50 مع سعر المبيع، فهي كما يلي:

1stFlrSF: مساحة الطابق الأول بالقدم المربعة.

FullBath: عدد الحمامات في الطوابق العلوية.

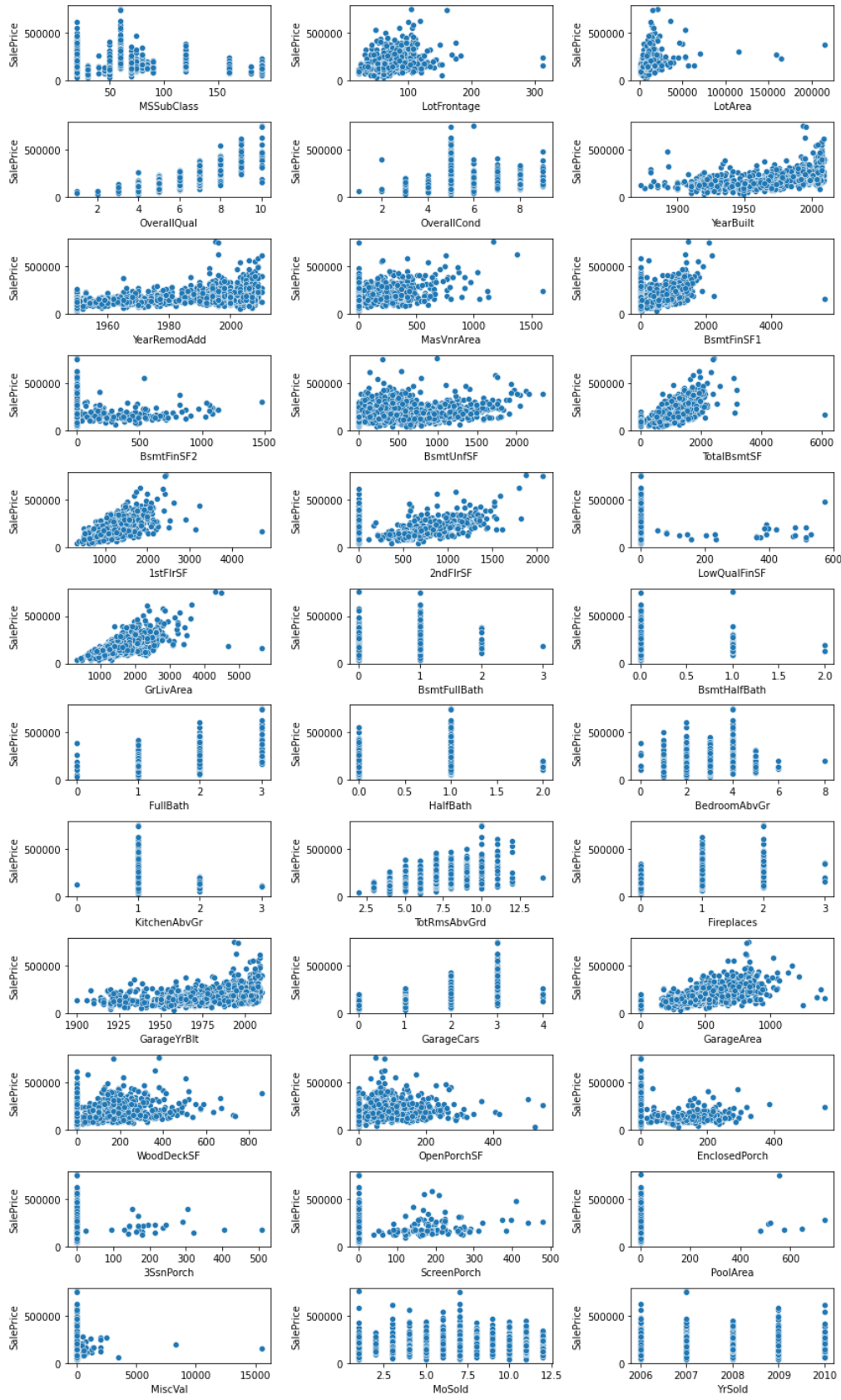
TotRmsAbvGrd: إجمالي عدد الغرف فوق الطابق الأرضي (لا يتضمن الحمامات).

YearBuilt: تاريخ البناء الأصلي.

YearRemodAdd: تاريخ التجديد (وهو يساوي تاريخ البناء الأصلي في حال عدم إجراء أي تجديدات أو إضافات لاحقة).

ومن جهة أخرى، تشير نتائج المصفوفة إلى وجود مؤشرات على التعددية الخطية بين بعض المتغيرات المستقلة، ومن أبرز الأمثلة على ذلك العلاقة بين عدد المواقف المتاحة في المرآب (GarageCars) ومساحة المرآب (GarageArea)، إذ بلغ معامل الارتباط بينهما نحو 0.89، في حين يرتبط كلا المتغيرين بسعر المبيع بشكل متقارب وعالٍ، ما يعزز احتمال وجود تكرار في التمثيل الإحصائي داخل النموذج.

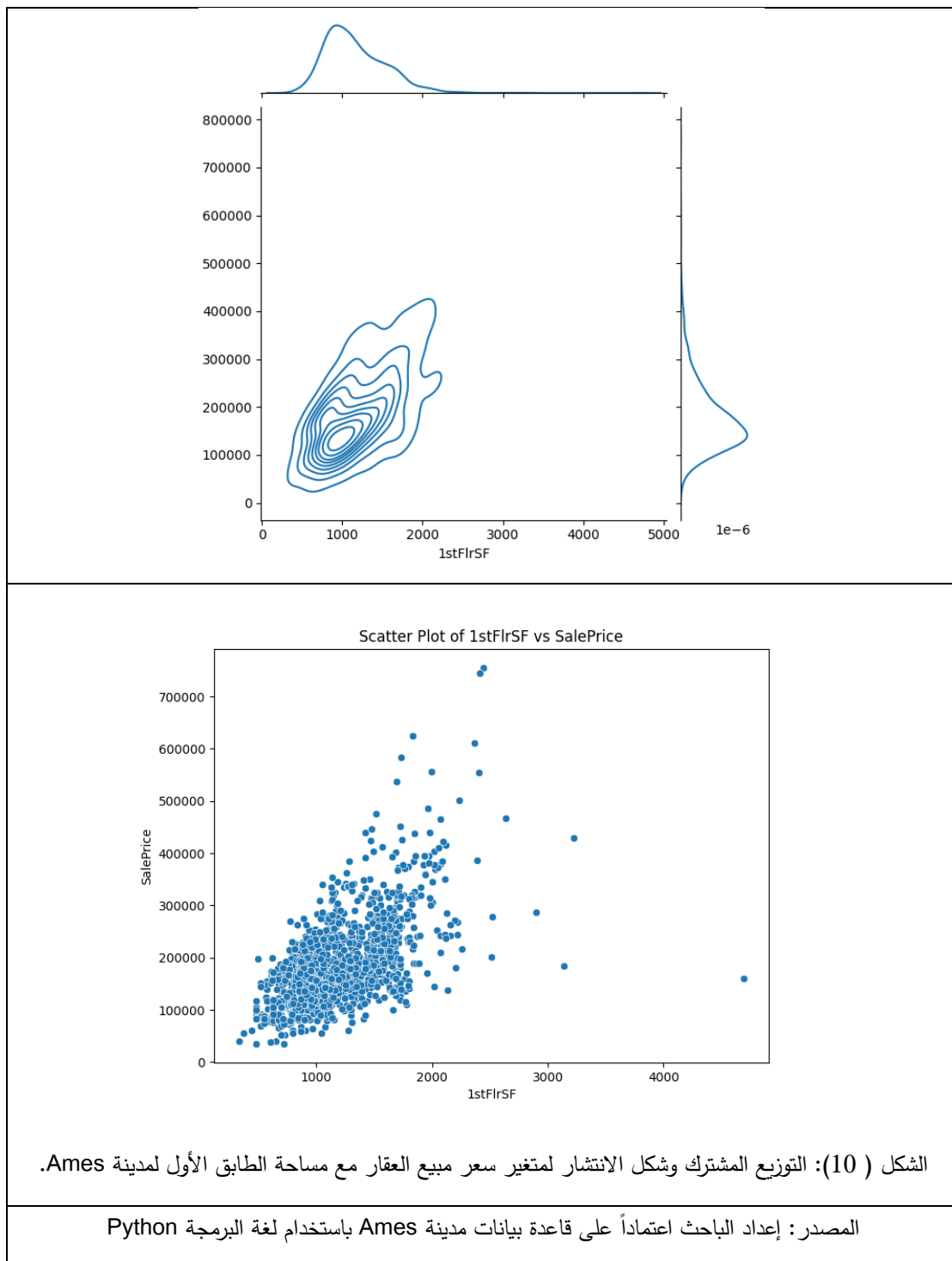
وفيما يلي شكل الانتشار لجميع المتغيرات الكمية مع متغير سعر العقار:



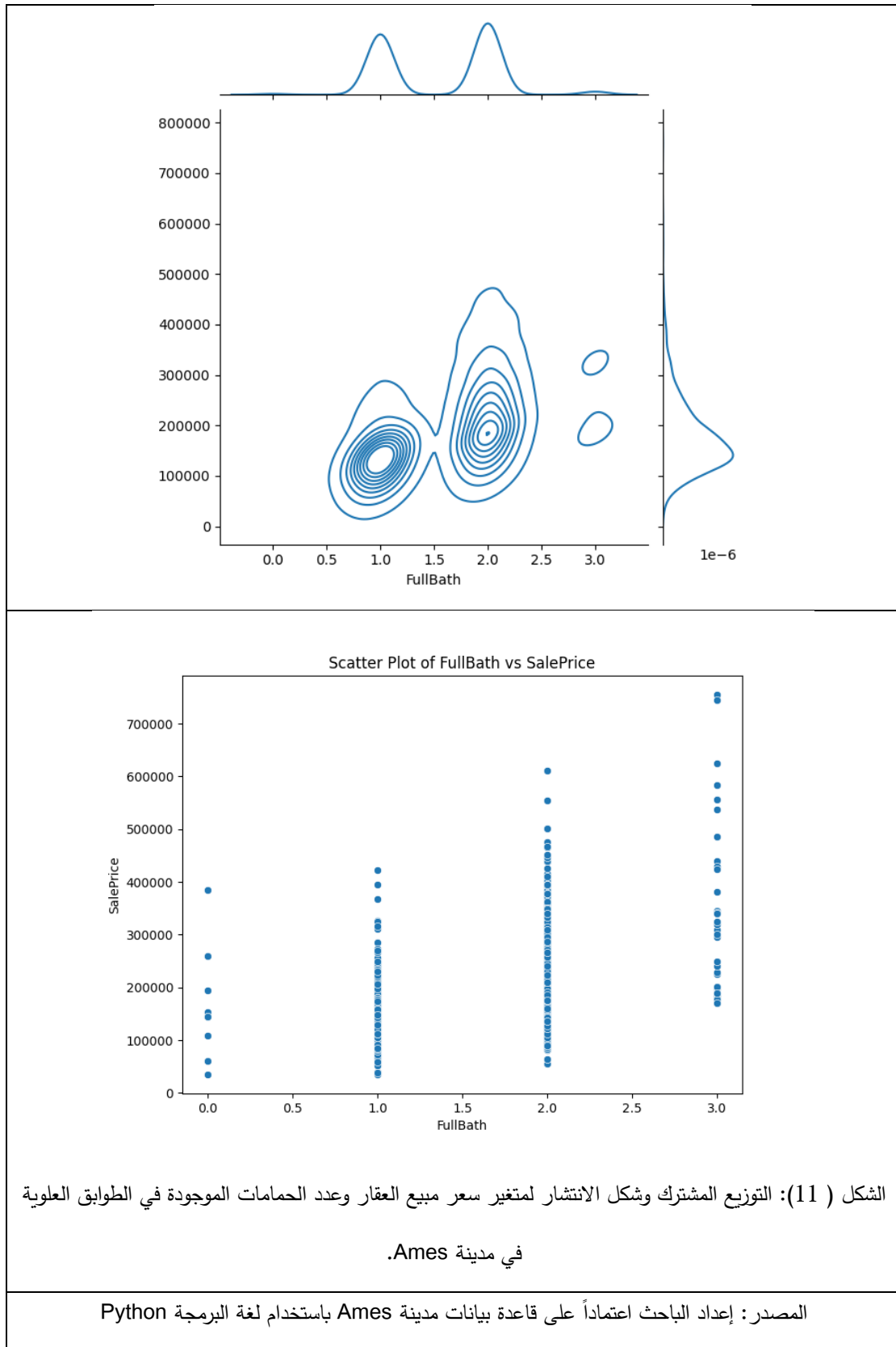
الشكل (9) شكل الانتشار لجميع المتغيرات الكمية مع متغير سعر العقار .

المصدر: إعداد الباحث اعتماداً على قاعدة بيانات مدينة Ames باستخدام لغة البرمجة Python

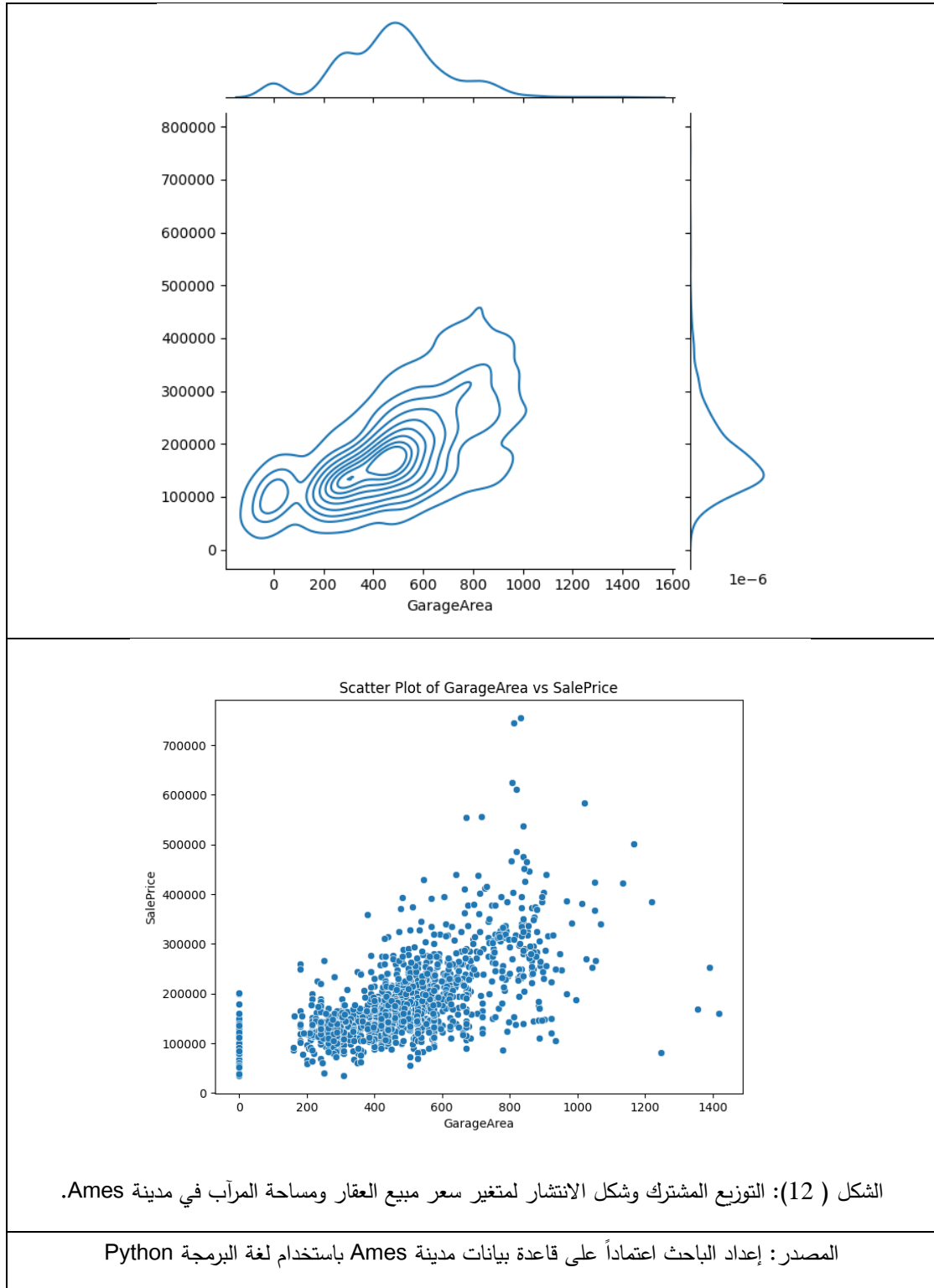
وفيما يلي شكل التوزيع المشترك وشكل الإنتشار لكل متغير من المتغيرات التي كانت ارتباطها أكبر من 50% مع متغير سعر العقار:



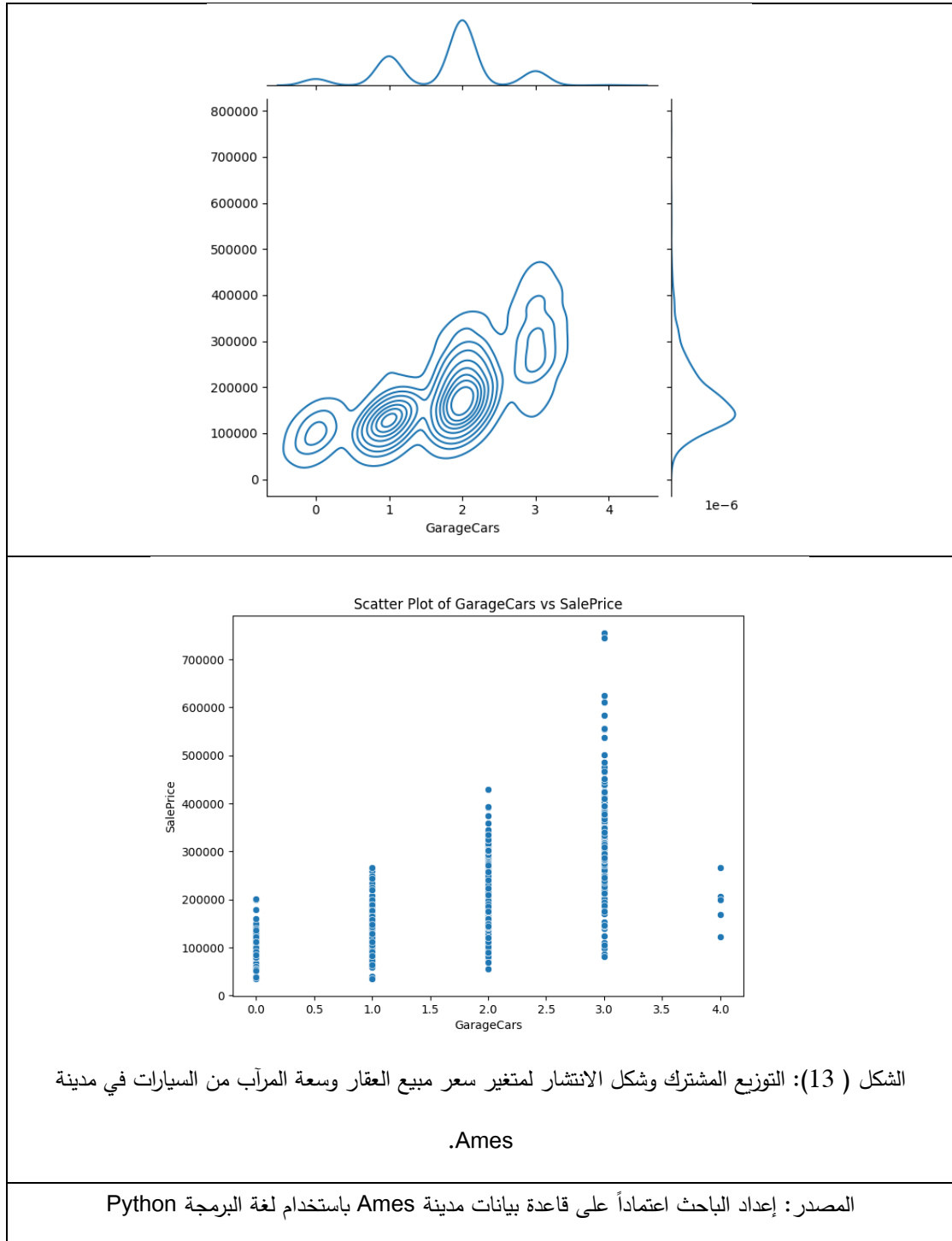
يُظهر الشكل وجود علاقة طردية واضحة بين مساحة الطابق الأول وسعر العقار، حيث تميل الأسعار إلى الارتفاع مع زيادة المساحة. كما يمكن ملاحظة وجود بعض القيم الشاذة التي قد تؤثر على دقة النموذج، ما يستدعي التعامل معها لاحقاً في مرحلة المعالجة أو التقدير.



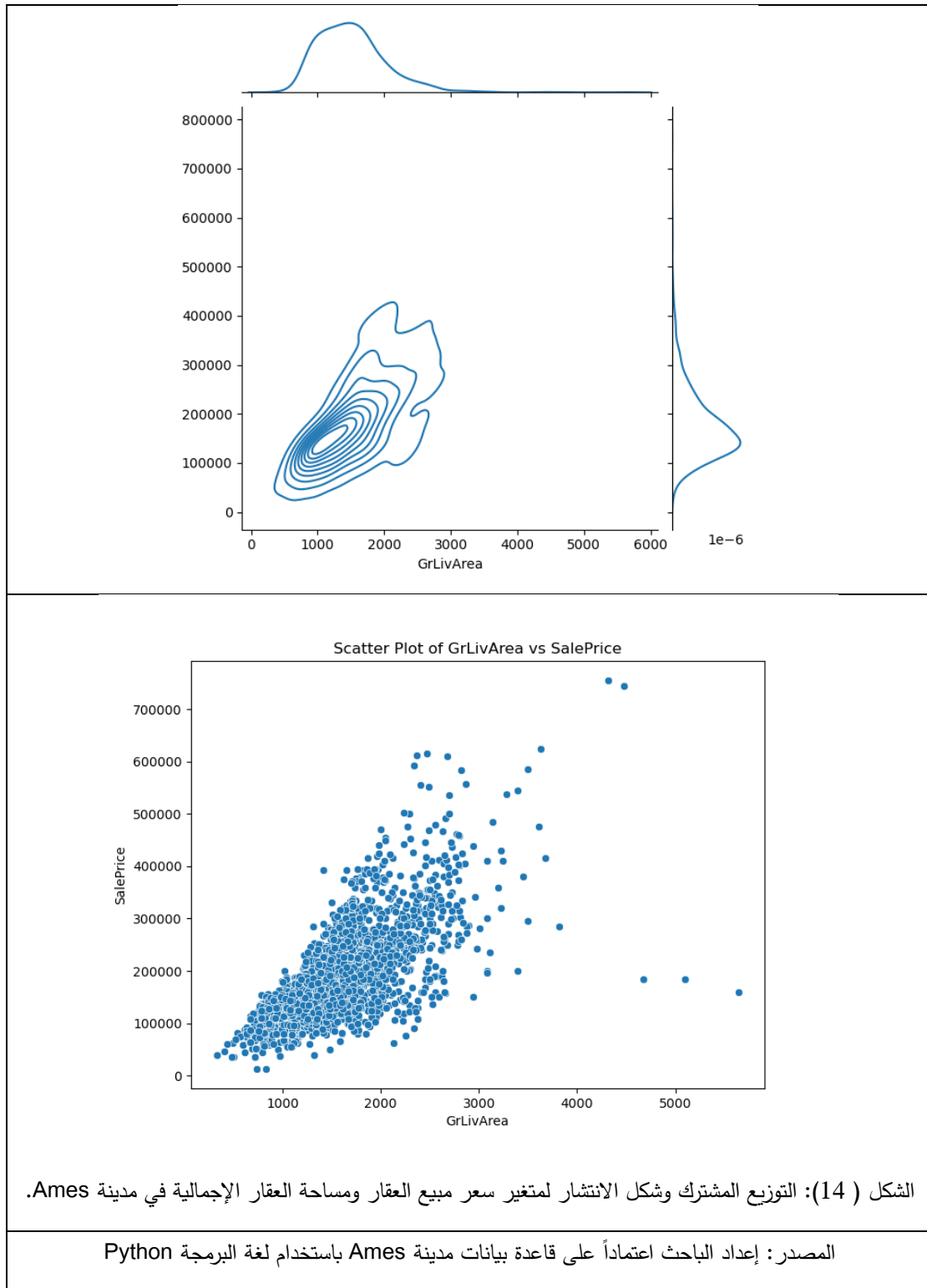
يُظهر الشكل وجود علاقة طردية بين عدد الحمامات الكاملة في الطوابق العلوية وسعر العقار، إذ تميل الأسعار إلى الارتفاع مع زيادة عدد الحمامات. إلا أن هذه العلاقة تبدو محدودة نسبياً.



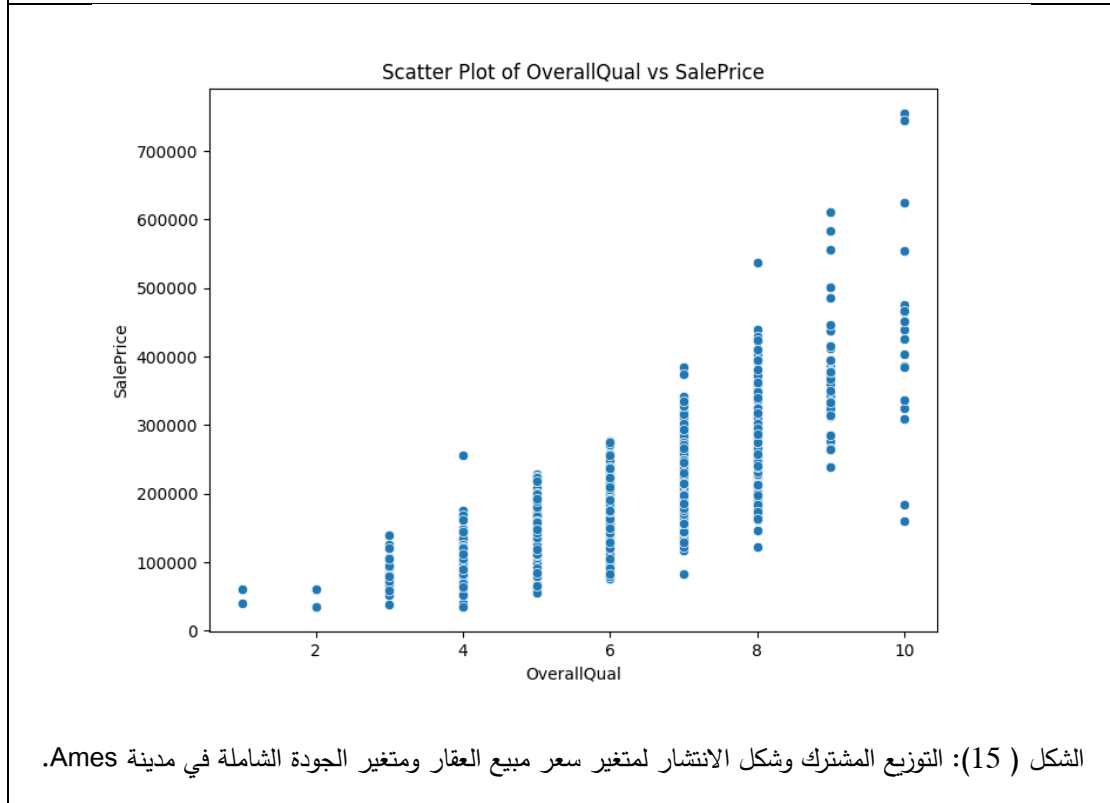
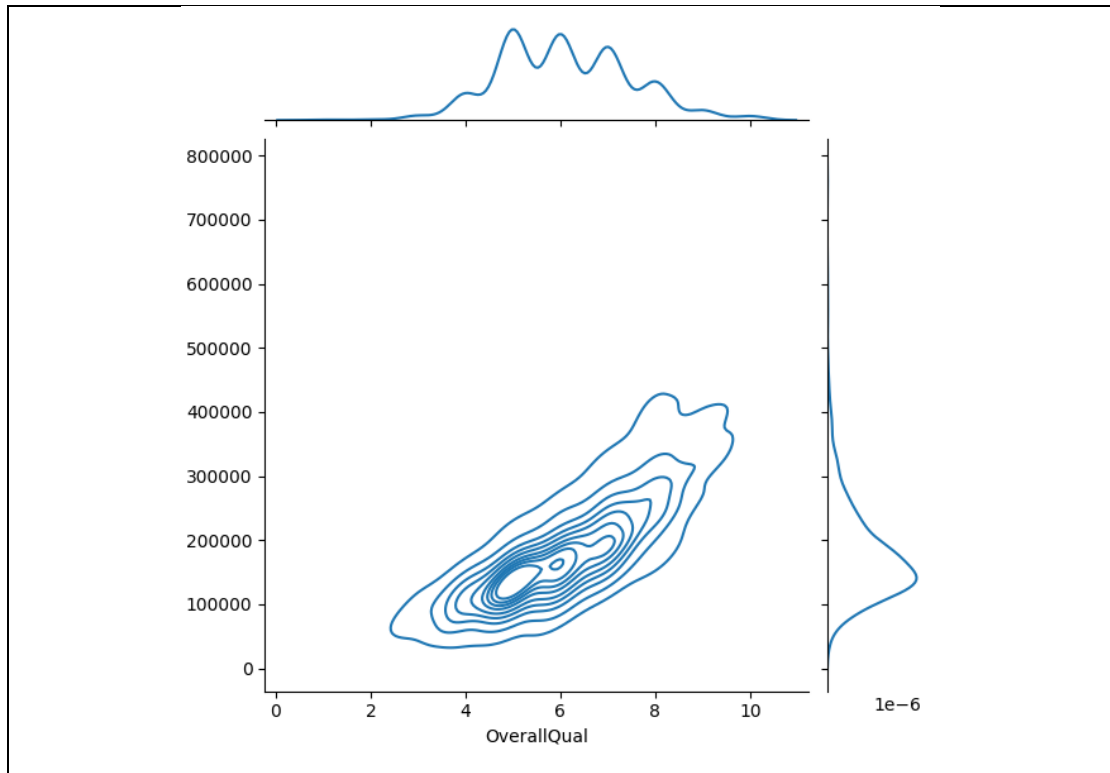
يُظهر الشكل وجود علاقة طردية واضحة بين مساحة المرآب وسعر العقار، حيث تميل الأسعار إلى الارتفاع مع زيادة المساحة. كما يمكن ملاحظة وجود عدد من القيم الشاذة، وخاصة في المساحات الكبيرة، ما يستدعي التحقق من تأثيرها لاحقاً على نتائج التقدير.



يُظهر الشكل وجود علاقة طردية بين سعة المرآب من السيارات وسعر العقار، حيث ترتفع الأسعار بشكل ملحوظ مع ازدياد عدد المواقف المتاحة في المرآب حتى حدٍ معين. إلا أن الأسعار تبدأ بالانخفاض مجدداً عند تجاوز ثلاث سيارات، ما قد يشير إلى تأثير متغيرات أخرى يفضلها المشترون في تلك الفئة.

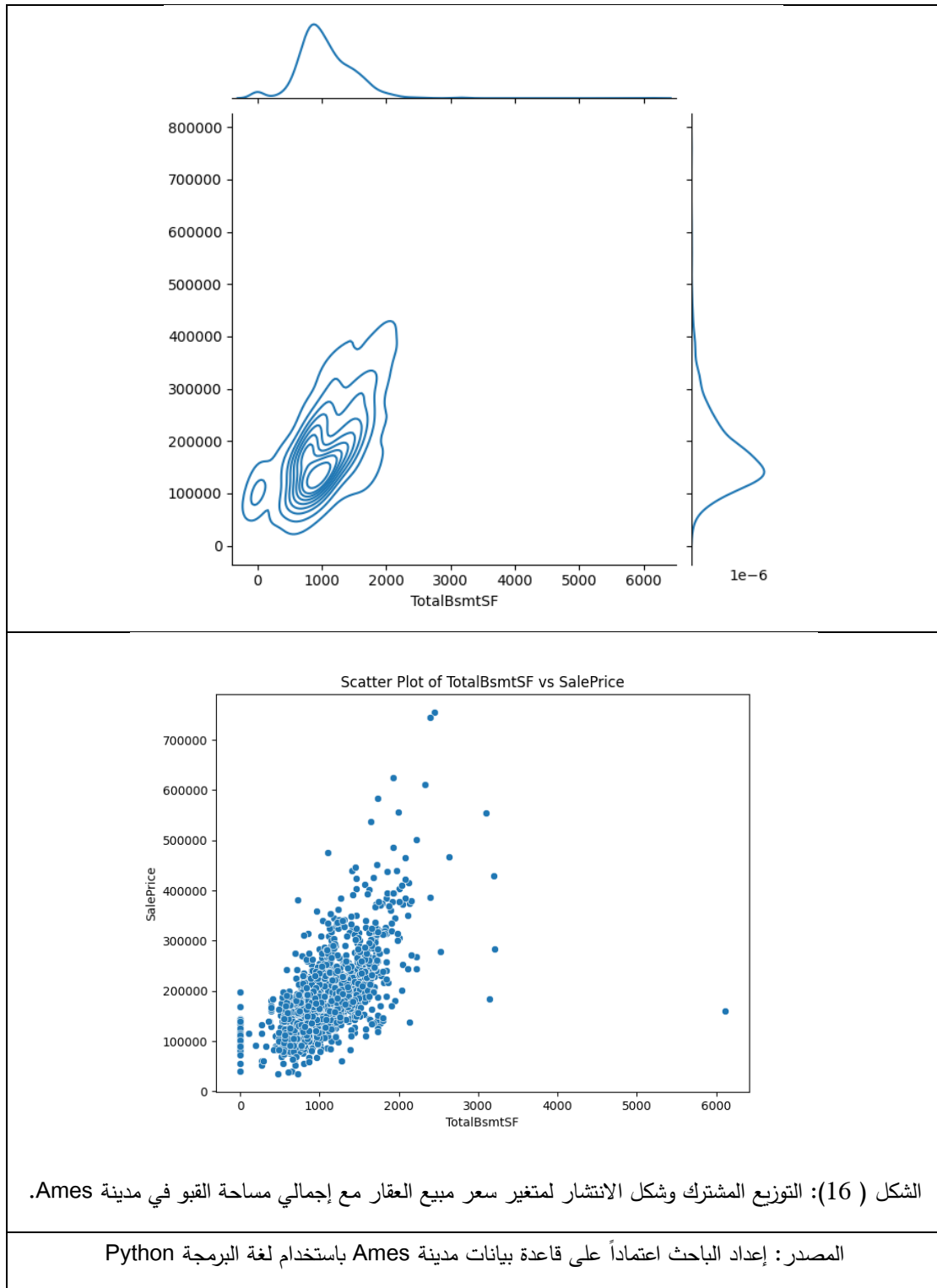


يُظهر الشكل وجود علاقة طردية سعر مبيع العقار ومساحة العقار الإجمالية، حيث ترتفع الأسعار كلما ازدادت المساحة. إلا أنه يمكن ملاحظة بعض القيم الشاذة التي تمثل عقارات ذات مساحات كبيرة جداً، لكن بأسعار غير متناسبة، وهو ما قد يشير إلى تأثير عوامل أخرى غير مفسرة في هذا النموذج، ما يستوجب أخذ هذه القيم بعين الاعتبار أثناء التقدير لاحقاً.

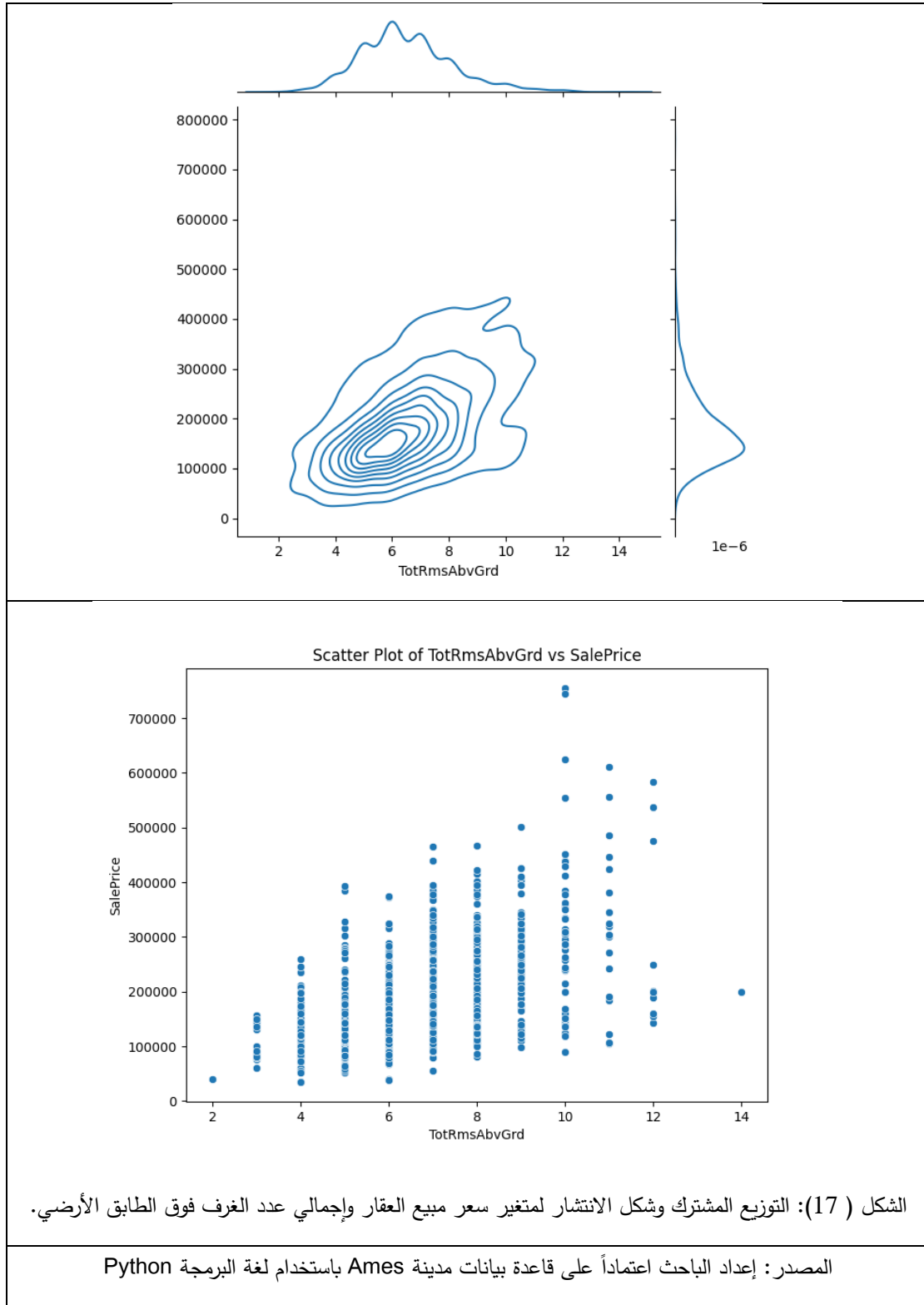


المصدر: إعداد الباحث اعتماداً على قاعدة بيانات مدينة Ames باستخدام لغة البرمجة Python

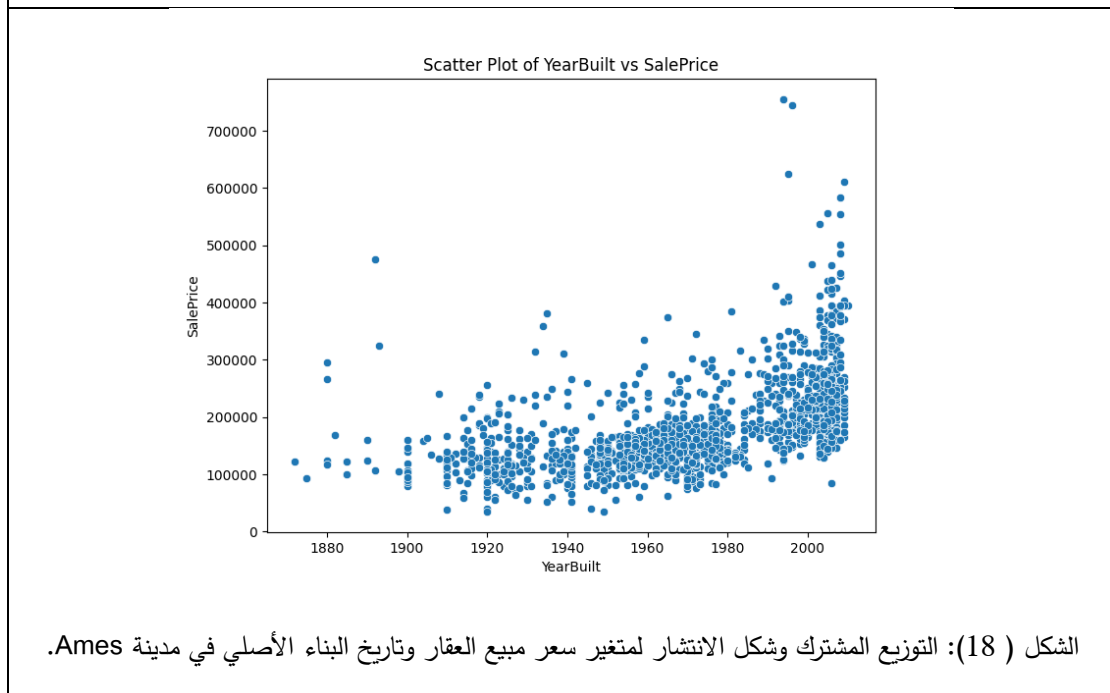
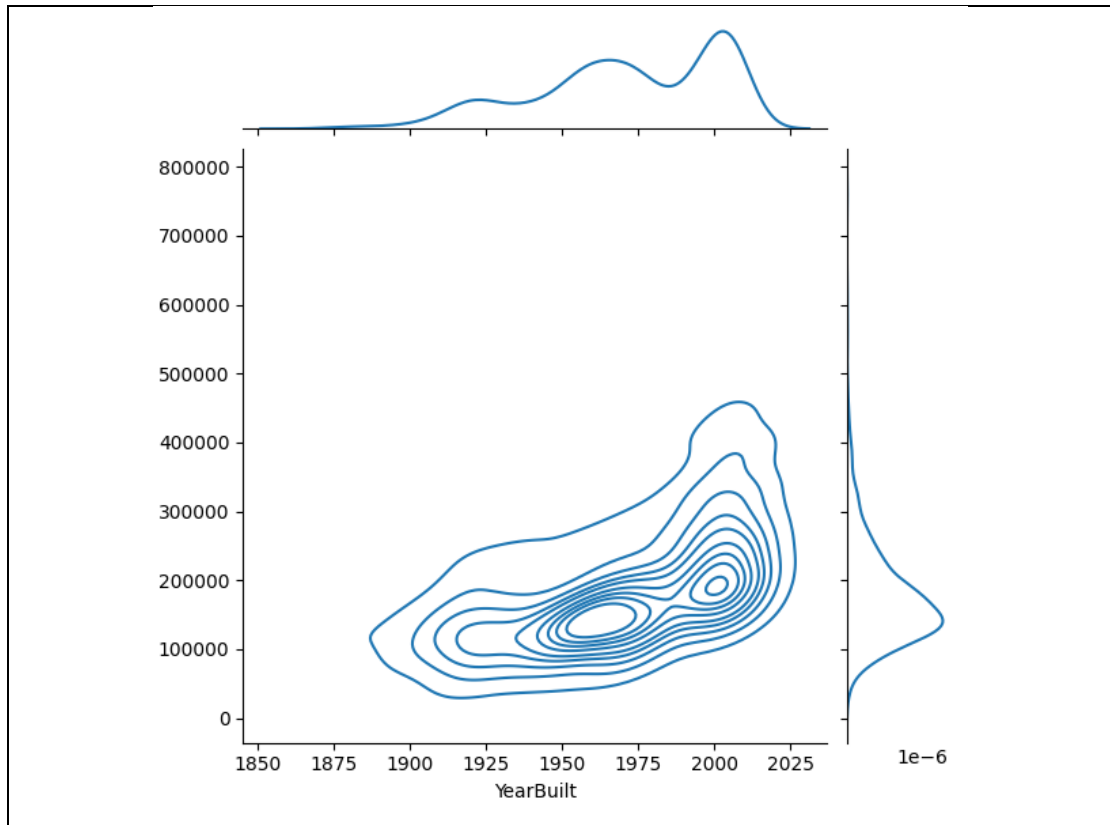
يُظهر الشكل وجود علاقة طردية واضحة بين سعر مبيع العقار ومتغير الجودة الشاملة، إذ نلاحظ ارتفاعاً تدريجياً في الأسعار كلما زادت قيمة تقييم الجودة. وهذا ما ينسجم منطقياً مع الواقع، حيث تميل العقارات ذات الجودة الأعلى من حيث الإكساء والمواد المستخدمة إلى تحقيق أسعار أعلى في السوق العقاري.



يُلاحظ وجود علاقة طردية بين إجمالي سعر مبيع العقار ومساحة القبو، حيث تزداد أسعار العقارات مع تزايد مساحة القبو بشكل عام. وتشير هذه العلاقة إلى أن المساحات الإضافية في الطابق السفلي تلعب دوراً مهماً في رفع القيمة السوقية للعقار، إلا أن وجود بعض القيم الشاذة يدعو إلى مزيد من التحقق لضمان دقة النتائج وتحسين أداء النماذج الإحصائية.

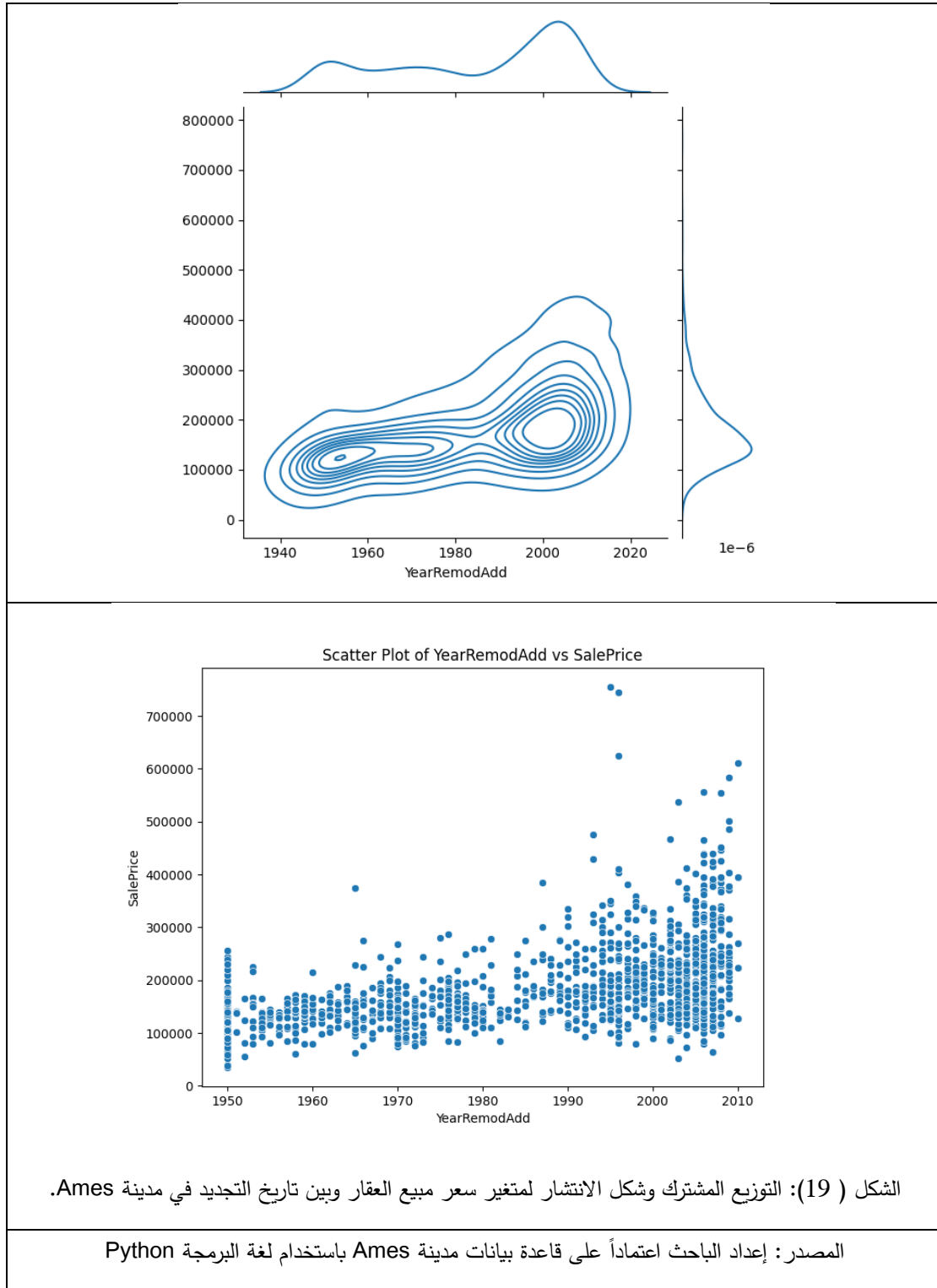


يُلاحظ وجود علاقة طردية بين إجمالي عدد الغرف الموجودة فوق الطابق الأرضي وسعر العقار، فكلما زاد عدد الغرف ارتفع سعر العقار بشكل عام. إلا أن هذه العلاقة لا تبدو مستمرة بعدد معين من الغرف، حيث تبدأ الأسعار بالثبات أو الانخفاض نسبياً، ما يشير إلى احتمال وجود خصائص أخرى تؤثر في السعر بعد حد معين من الغرف.



المصدر: إعداد الباحث اعتماداً على قاعدة بيانات مدينة Ames باستخدام لغة البرمجة Python

يُظهر الرسم البياني وجود علاقة طردية بين سعر العقار وسنة البناء، إلا أن العلاقة لا تبدو قوية للغاية. ومن اللافت للانتباه أن هناك بعض الحالات التي تظهر فيها أسعار عالية مع بناءات أحدث، ما يشير إلى وجود عوامل إضافية تؤثر، مثل المميزات المعمارية، أو التجديدات.



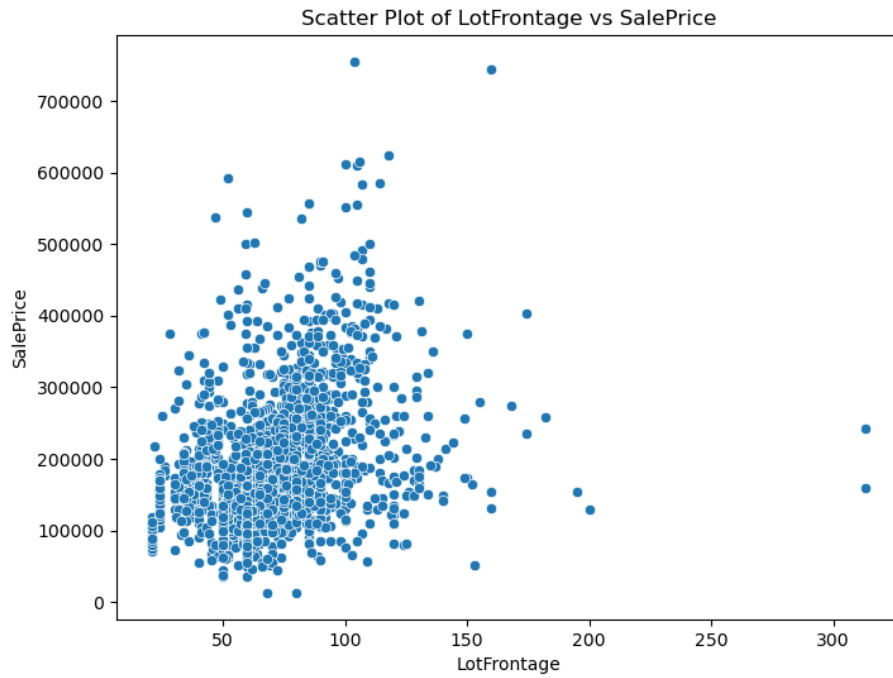
يُظهر الشكل تبايناً واسعاً في أسعار العقارات بغض النظر عن سنة التجديد، ما يدل على أن سنة التجديد وحدها، قد لا تكون مؤشراً كافياً لتفسير الفروق في الأسعار. ورغم وجود بعض العقارات المجددة حديثاً بأسعار مرتفعة، إلا أن الاتجاه العام لا يُظهر علاقة طردية قوية أو منتظمة. ويشير ذلك إلى احتمال وجود عوامل أخرى أكثر تأثيراً في تحديد السعر، مثل جودة التجديد أو خصائص العقار الأخرى.

3.1.7 معالجة القيم الشاذة

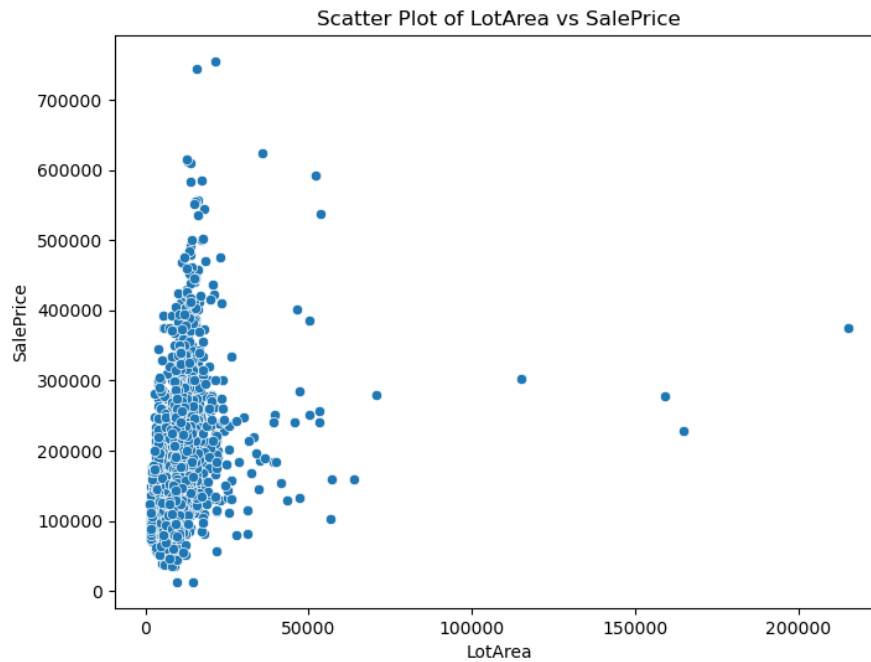
تُعدّ القيم الشاذة من العوامل التي قد تؤثر سلباً على دقة النماذج الإحصائية، ونتائج التحليل في حال عدم التعامل معها بحذر (Aggarwal, 2017, p. 63). لذلك، كان من الضروري تحديد هذه القيم بشكل دقيق، وتقييم مدى تأثيرها، بدلاً من حذفها بشكل تلقائي دون دراسة.

من خلال مراجعة الكتيب التوثيقي للبيانات وتحليل الرسوم البيانية، تبين وجود بعض القيم التي تبتعد كثيراً عن التجمع الرئيسي للنقاط، ما قد يؤثر على التقديرات والانحدار (De Cock, 2011, p. 4)، فعلى سبيل المثال، يظهر في الشكل رقم (14) أن متغير مساحة العقار الإجمالية (GrLivArea) يحتوي على خمس حالات تتجاوز المساحة فيها 4000 قدم مربع، وهي حالات نادرة نسبياً وتُظهر أسعار بيع غير متناسقة مع الاتجاه العام، الأمر الذي يُرجّح تصنيفها كقيم شاذة أثّرت بشكل مباشر في جودة النموذج.

كما يُبين الشكل (20) شكل العلاقة بين متغير المسافة بين العقار والشارع (LotFrontage) وسعر المبيع، ويُظهر الشكل (21) العلاقة بين متغير مساحة أرض العقار (LotArea) وسعر المبيع، وذلك من خلال مخططات الانتشار. ويُلاحظ في كلا الشكلين وجود عدد من القيم المتطرفة المتمركزة على أطراف المخطط، والتي تُشير إلى حالات عقارية ذات خصائص استثنائية من حيث المساحة، أو المسافة بين العقار والشارع. وعلى الرغم من غياب نمط ارتباط واضح بين كلٍّ من هذين المتغيرين وسعر المبيع، إلا أنّ تلك القيم أسهمت في انحراف شكل التوزيع، ما قد يؤثر على دقة بعض التحليلات الإحصائية.



الشكل (20): شكل الانتشار لمتغير سعر مبيع العقار ومتغير المسافة بين العقار والشارع (LotFrontage).



الشكل (21): شكل الانتشار لمتغير سعر مبيع العقار ومتغير مساحة أرض العقار (LotArea).

المصدر: إعداد الباحث اعتماداً على قاعدة بيانات مدينة Ames باستخدام لغة البرمجة Python

بناءً على ما سبق، تقرّر حذف:

- ❖ 5 حالات من متغير مساحة العقار الإجمالية (GrLivArea)،
- ❖ قيمتين من متغير المسافة بين العقار والشارع (LotFrontage)،
- ❖ 4 حالات من متغير مساحة أرض العقار (LotArea).

وبالإضافة إلى ذلك، تم التعامل مع عدد محدود من القيم الشاذة في بعض المتغيرات الأخرى، وذلك وفق معايير مماثلة لما سبق، كما هو موضح في الكود البرمجي المرفق في الملحق.

وقد أدى هذا الحذف إلى تحسّن ملحوظ في نتائج التحليل وتقليل تأثير التشوّه الناتج عن القيم الشاذة، دون التأثير على الاتجاه العام أو فقدان معلومات أساسية.

3.1.8 بناء النموذج الإحصائي للعلاقة بين متغيرات الدراسة.

سيتم نمذجة العلاقة بين سعر العقار والمتغيرات الخاصة بمدينة Ames باستخدام انحدار لاسو. وفيما يلي الخطوات المعتمدة في هذا التحليل:

3.1.8.1 اختيار معلمة الضبط الخاصة بانحدار لاسو

تلعب معلمة الضبط λ الخاصة بانحدار لاسو دوراً محورياً في عملية التقدير، إذ تتحكم في حجم الجزء المفروض على معاملات النموذج، وبالتالي تؤثر على عدد المتغيرات الداخلة في النموذج النهائي. وكما وُضح سابقاً، فإن اختيار القيمة المثلى لـ λ يُعدّ خطوة أساسية للحصول على نموذج يتميز بالتوازن بين البساطة والدقة.

ولتحقيق ذلك، فقد طُبّق أسلوب التحقق المتقاطع باستخدام طريقة K-Folds بعدد $K=4$ ، حيث قُسمت مجموعة بيانات التدريب إلى أربع مجموعات متساوية، تُستخدم ثلاث منها في كل تكرار لأغراض التدريب، فيما تُخصّص المجموعة الرابعة للتحقق. واعتمد التحليل على متوسط مربعات الأخطاء كمؤشر أساسي لاختيار القيمة التي تُحقّق أفضل أداء. حيث تم الاعتماد على أسلوبين منهجيين لاختيار مجموعة من القيم التجريبية للمعلمة λ وفق ما يلي:

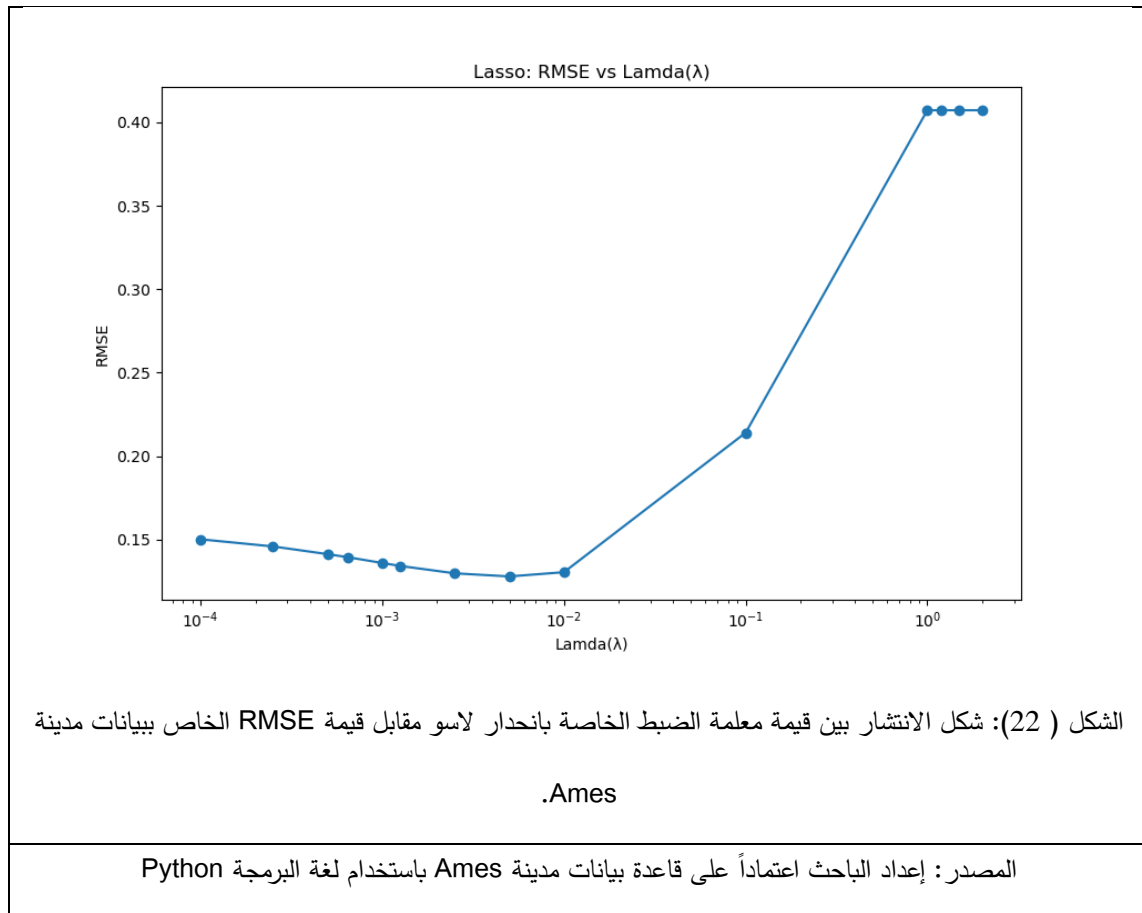
الأسلوب الأول: إنشاء متتالية حسابية لقيم λ تبدأ من 0.0001 وتزداد تدريجياً حتى تصل إلى 1، وذلك لتغطية نطاق واسع من القيم الصغيرة إلى الكبيرة. يُعد هذا الأسلوب شائعاً في الدراسات التطبيقية، وقد اعتمد ضمن مكنبات

علمية مفتوحة المصدر مثل Scikit-learn (Pedregosa, et al, 2011)

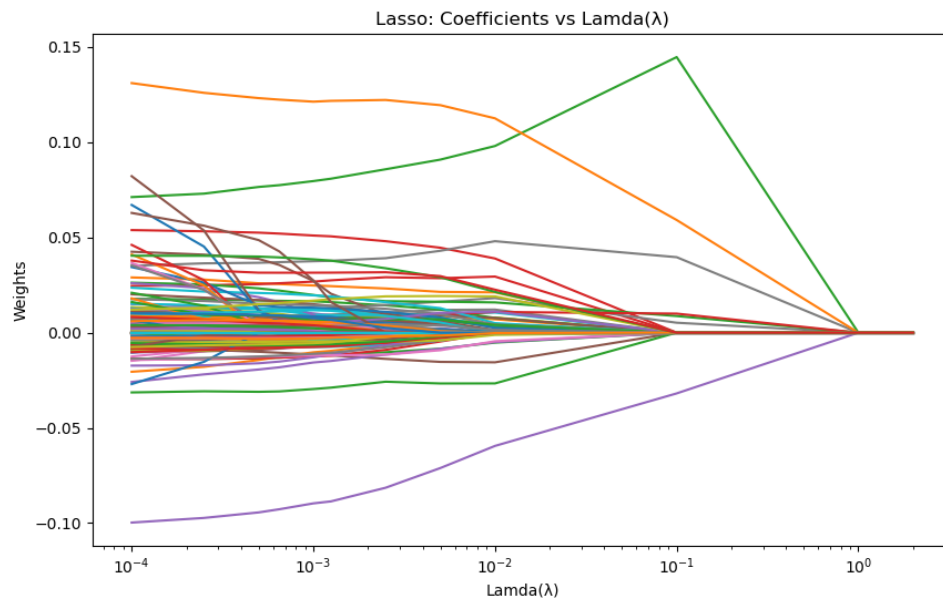
الأسلوب الثاني: اعتمد الأسلوب على مجموعة مختارة من القيم بلغ عددها 14 قيمة، اقترحت استناداً إلى نتائج دراسات سابقة وتجارب تطبيقية منشورة (Friedman, 2010). وقد شملت هذه القيم درجات متفاوتة من الجزء لضمان اختبار النموذج تحت ظروف متعددة (Ian Goodfellow, 2016)، وأدرجت القيم في الملحق (2).

تجدر الإشارة إلى أن كلا الأسلوبين أظهرتا نتائج مقاربة جداً، وقد اعتمدت نتائج الأسلوب الثاني (القائمة المختارة يدوياً) في العرض النهائي.

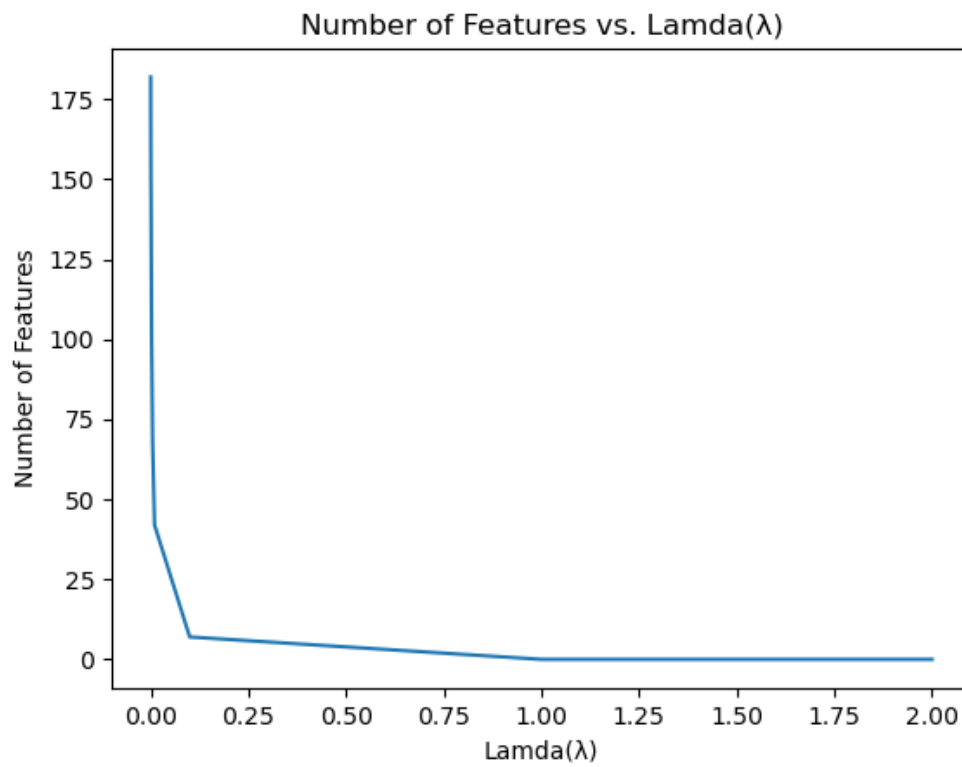
كما أُجري تقدير لـ 14 نموذجاً مختلفاً باستخدام القيم المختارة لمعلمة الضبط λ ، وتبين أن القيمة الأدنى لـ λ بلغت نحو 0.0025، وهي القيمة التي تحقق عندها أدنى مستوى لجذر متوسط مربعات الأخطاء، والذي سجل ما يقارب 0.1232. ويوضح الرسم البياني الآتي هذا السلوك والانخفاض التدريجي في الخطأ عبر القيم المختلفة لـ λ .



بالإضافة لما سبق، يوضح الشكلان (23) و (24) الدور الهام لمعلمة الضبط λ لانحدار لاسو، فكما أُشير سابقاً، كلما ازدادت قيمة معلمة الضبط λ ازداد كمية الجزء المفروض على المعلمات، بالتالي انخفض عدد المتغيرات الداخلة في النموذج. وعندما تكون قيمة معلمة الضبط λ مساوية إلى الصفر، فإن جميع المتغيرات تُدرج في عملية التقدير دون أي تقليص أو حذف.



الشكل (23): منحنيات معاملات انحدار لاسو كدالة في معلمة الضبط λ الخاصة ببيانات مدينة Ames.



الشكل (24): عدد المتغيرات الداخلة في النموذج كدالة في معلمة الضبط λ الخاصة ببيانات مدينة Ames.

المصدر: إعداد الباحث اعتماداً على قاعدة بيانات مدينة Ames باستخدام لغة البرمجة Python

نلاحظ في الشكل (23)، كيف تؤدي الزيادة التدريجية في قيمة λ إلى انكماش معاملات النموذج نحو الصفر، إذ تبدأ بعض المعاملات بالتناقص حتى تختفي تماماً مع تجاوز λ لقيم معينة. يعبر هذا السلوك عن الدور الأساسي لانحدار لاسو في تقليص عدد المتغيرات عبر الجزاء المفروض على حجم المعاملات.

أما الشكل (24)، فيُظهر العلاقة بين قيمة λ وعدد المتغيرات المتبقية في النموذج بعد عملية التقدير. نلاحظ بوضوح أن ازدياد قيمة λ يؤدي إلى انخفاض حاد في عدد المتغيرات المختارة، إلى أن يصل النموذج في النهاية إلى حالة استبعاد جميع المتغيرات عند القيم المرتفعة جداً لـ λ .

تؤكد هذه الأشكال على أهمية اختيار قيمة λ بعناية لتحقيق التوازن المطلوب بين تبسيط النموذج (parsimony) ودقة التنبؤ.

3.1.8.2 تقدير معلمات النموذج على بيانات التدريب والاختبار

بعد تحديد القيمة المثلى لمعلمة الضبط λ الخاصة بانحدار لاسو، جرى الانتقال إلى مرحلة تقدير معاملات النموذج باستخدام بيانات التدريب، ومن ثم تقييم أدائه على بيانات الاختبار. وفيما يلي النتائج التي تم الحصول عليها وفق الآتي:

الجدول (4): نتائج تقدير انحدار لاسو على بيانات التدريب والاختبار الخاصة ببيانات Ames.		
المؤشر	مجموعة البيانات	القيمة
RMSE	بيانات التدريب	0.1035
	بيانات الاختبار	0.1153
عدد المتغيرات المتبقية الداخلة في النموذج	114	
المصدر: إعداد الباحث اعتماداً على قاعدة بيانات مدينة Ames باستخدام لغة البرمجة Python		

- تشير قيمة RMSE المنخفضة لكل من بيانات التدريب والاختبار إلى جودة النموذج المقدر بشكل عام على التنبؤ لأسعار العقارات بناءً على المتغيرات الداخلة في النموذج. كما يظهر النموذج عدد المتغيرات الداخلة في النموذج والتي بلغت 114 متغير من أصل 240 متغيراً والذي يدل على قدرة النموذج على تقدير أسعار العقارات، مع عدد قليل نسبياً من المتغيرات وتبسيط وفهم العلاقة بشكل أبسط. بالإضافة إلى ذلك، فإن انحدار لاسو لديه القدرة على تحديد المتغيرات الأهم في التنبؤ في أسعار العقارات، حيث حذف بعض المتغيرات المرتبطة ببعضها بشكل عالي، وأبقى على أحدهما، وأيضاً القدرة على معالجة مشكلة التعددية الخطية بين المتغيرات.

3.1.8.3 الشكل النهائي للنموذج المقدر.

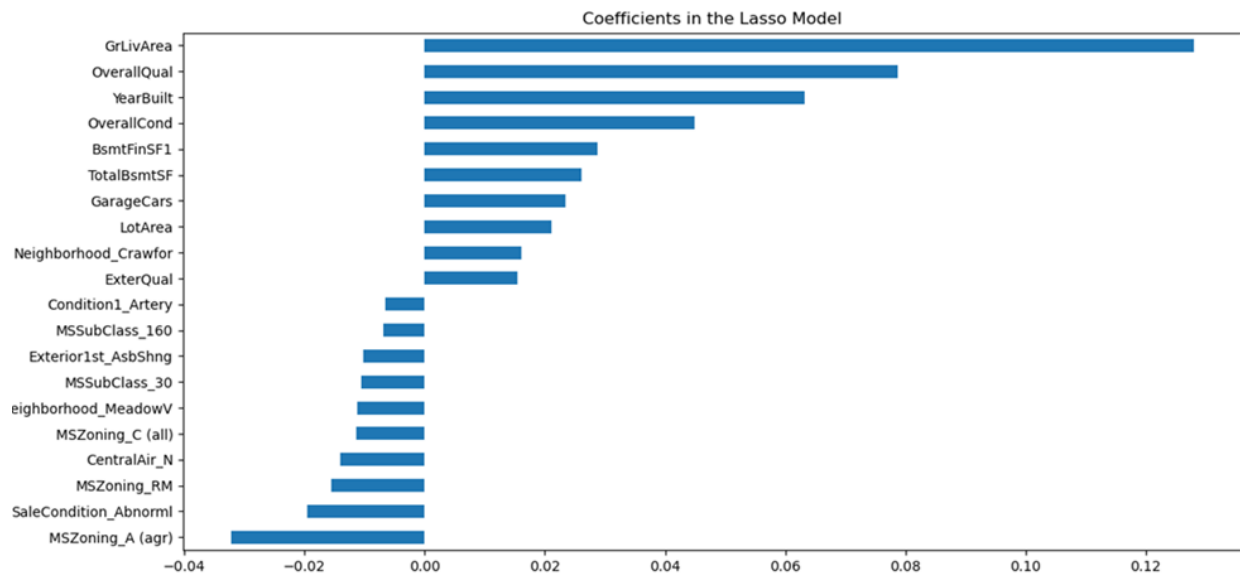
يعرض الجدول الآتي قيم المعاملات المعيارية التي جرى تقديرها باستخدام انحدار لاسو، وذلك بعد اختيار القيمة المثلى لمعلمة الضبط λ ، كما أُشير إليه سابقاً. وتُعكس هذه المعاملات تأثير كل متغير مستقل في المتغير التابع (سعر العقار) ضمن النموذج المعياري المُعتمد.

ونظراً لكثرة المتغيرات الداخلة في التقدير، فقد تم إدراج الصيغة النهائية الكاملة للنموذج في الملحق (1).

الجدول (6): القيم التقديرية للمعاملات المعيارية في انحدار لاسو لتقدير أسعار العقارات في مدينة Ames.							
Var Name	Value	Var Name	Value	Var Name	Value	Var Name	Value
GrLivArea	0.1280	BsmtExposure	0.0104	MasVnrType_Stone	0.0056	Condition1_PosA	0.0018
OverallQual	0.0786	1stFlrSF	0.0103	SaleCondition_AdjLand	0.0055	BsmtFinSF2	0.0017
YearBuilt	0.0633	Exterior1st_BrkFace	0.0103	MSZoning_FV	0.0052	Neighborhood_Greens	0.0016
OverallCond	0.0450	Neighborhood_GrnHill	0.0103	PavedDrive_Y	0.0051	HalfBath	0.0012
BsmtFinSF1	0.0288	Foundation_PConc	0.0103	FullBath	0.0047	ExterCond	0.0009
TotalBsmtSF	0.0261	Exterior1st_PreCast	0.0097	OpenPorchSF	0.0047	MSSubClass_75	0.0006
GarageCars	0.0234	Fireplaces	0.0089	GarageCond	0.0043	SaleType_CWD	0.0003
LotArea	0.0212	BsmtFullBath	0.0087	SaleType_Con	0.0039	Neighborhood_ClearCr	0.0002
Neighborhood_Crawfor	0.0160	Functional_Typ	0.0082	WoodDeckSF	0.0036	Neighborhood_Blueste	0.0002
ExterQual	0.0155	LotFrontage	0.0079	MasVnrArea	0.0034	LotConfig_CulDSac	0.0002
BsmtQual	0.0134	EnclosedPorch	0.0078	LandContour_HLS	0.0030	SaleCondition_Alloca	0.0000
HeatingQC	0.0126	BsmtFinType1	0.0077	MSSubClass_70	0.0030	Neighborhood_NPkVill	0.0000
KitchenQual	0.0123	GarageArea	0.0077	Condition1_PosN	0.0026	Exterior2nd_PreCast	0.0000
Neighborhood_NridgHt	0.0121	Neighborhood_StoneBr	0.0072	Electrical_SBrkr	0.0024	CentralAir_Y	0.0000
ScreenPorch	0.0117	BldgType_1Fam	0.0068	FireplaceQu	0.0023	Neighborhood_Names	-0.0001
Neighborhood_BrkSide	0.0110	Neighborhood_Somerst	0.0067	Exterior2nd_CBlock	0.0021	Exterior1st_BrkComm	-0.0002
SaleType_New	0.0105	Condition1_Norm	0.0067	Neighborhood_NoRidge	0.0018	MSSubClass_180	-0.0002
المصدر: إعداد الباحث اعتماداً على قاعدة بيانات مدينة Ames باستخدام لغة البرمجة Python							

بعد عرض جميع قيم المعاملات الداخلة في النموذج المقدر، يمكن عرض لأول 10 متغيرات والتي كانت متغير مساحة المعيشة فوق الأرض (المساحة الداخلية للسكن الفعلي) GrLivArea، والتقييم الإجمالي للعقار (للاكساء) OverallQual، وعام البناء YearBuilt، والتقييم الإجمالي لحالة للعقار OverallCond، و المساحة الأساسية المُجهّزة في القبو (BsmtFinSF1)، ومساحة القبو TotalBsmtSF، وعدد المواقف المتاحة في المرآب GarageCars، ومساحة أرض العقار LotArea، وموقع العقار ضمن حي Crawford في مدينة Ames، و تقييم جودة المواد المستخدمة في الواجهة الخارجية ExterQual.

وفيما يلي الشكل (25) الذي يوضح القيم التقديرية المعيارية للمعاملات في انحدار لاسو:



الشكل (25): قيم معاملات إنحدار لاسو لمدينة Ames.

المصدر: إعداد الباحث اعتماداً على قاعدة بيانات مدينة Ames باستخدام لغة البرمجة Python

والجدير بالذكر أن نتائج هذه الدراسة قد اشتركت مع ما توصلت إليه دراسة (Seng & Khalid, 2018) في بعض المتغيرات الأساسية المؤثرة في أسعار العقارات، مثل المساحة الداخلية، وجودة الإكساء وسنة البناء. غير أن دراستهم لم تقدم توضيحاً كافياً حول كيفية معالجة البيانات كما لم تعتمد التحويل اللوغاريتمي لمتغير السعر، على غرار ما تم تطبيقه في هذا البحث، ولم تُفصل في آلية ترميز بقية المتغيرات النوعية أو استبعاد بعضها. وبالتالي، فقد تميز هذا البحث بطرح أكثر دقة وشفافية في خطوات التحضير والمعالجة، ما يجعل نتائجه أكثر وضوحاً وقابلية للتطبيق والمقارنة.

خاتمة

خلص هذا المبحث إلى أن انحدار لاسو أتاح بناء نموذج إحصائي فعال لتفسير العلاقة بين أسعار العقارات في مدينة Ames ومجموعة من الخصائص المعمارية التي تميز كل عقار. وقد احتفظ النموذج بـ 114 متغيراً من أصل 240 متغيراً بعد عملية التقدير، مساهماً في تقليص التعددية الخطية وتعزيز القدرة التفسيرية. كما سجل النموذج قيمة منخفضة لجذر متوسط مربعات الأخطاء ($RMSE \approx 0.115$)، ما يعكس دقته في التنبؤ بالسعر اعتماداً على المتغيرات الأكثر تأثيراً.

بالإضافة لما سبق، يُظهر التحليل مدى قدرة انحدار لاسو على انتقاء أكثر المتغيرات تأثيراً، وهو مناسب خصوصاً في الحالات التي تتعدد فيها الخصائص ويصعب تحديد الأكثر تأثيراً منها في السعر.

المبحث الثاني طرائق التقييم العقاري

3.2.1 تعريف التقييم العقاري

يُمثل التقييم العقاري عملية أساسية في قطاع العقارات والاستثمار، تهدف إلى تحديد القيمة العادلة للعقار، استناداً إلى أسس ومعايير معتمدة. ويُعدّ هذا التقييم ركناً محورياً في النشاط العقاري، لما له من تأثير مباشر على قرارات البيع والشراء والاستثمار، إضافةً إلى انعكاساته على الحياة الاقتصادية للمدينة والدولة بشكل عام.

يُعرّف التقييم العقاري بأنه عملية تقدير القيمة السوقية لعقار معيّن في زمن محدد، لأغراض متعددة كالتصرف أو الرهن أو التأجير، ويعتمد على عوامل مثل الموقع والحالة القانونية والوظيفية للعقار، وذلك ضمن منظومة تشمل التنظيم الرسمي لقطاع العقارات (حافظ، 2019، ص 135). وباختصار هو عبارة عن التنبؤ بأسعار العقارات مبنياً على مميزات وصفات العقار. ويعد التنبؤ بخطأ مقبول مسألة معقدة جداً، لأن جميع العقارات تختلف عن بعضها البعض، وخاصة موقع العقار موضوع التقييم (Bulut, 2011).

هناك حاجة لعملية التقييم مهما كان سببها؛ التمويل والتأمين، أو الضرائب، أو إقراض الرهن العقاري، أو إعداد التقارير المالية، أو مبيعات الأصول، حيث تعتبر التقييمات الموثوقة والمعتمدة بالغة الأهمية والتي بدورها تساعد على تعزيز الاستقرار المالي والنمو الاقتصادي المستدام⁹.

3.2.2 طرائق التقييم العقاري

يوجد عدة طرائق لتقييم العقارات لتحديد القيمة الحقيقية للعقارات المختلفة، ونذكر منها ما يلي:

1- طريقة البيوع المقارنة

تعتمد هذه الطريقة على تحديد قيمة العقار المراد تقييمه بواسطة مقارنته مع عقارات مشابهة جرى بيعها حديثاً في المنطقة العقارية نفسها أو قريبة منها، وفي نفس الفترة أو قريبة منها، مع الأخذ بعين الاعتبار الفوارق بين العقار المراد تقييمه والعقارات المباعة والمقارنة بها. وهذه الطريقة هي أكثر وأبسط الطرائق المعروفة لتحديد قيمة العقارات، إذ أنها تعطي مؤشرات واقعية لقيمة العقار المراد تقييمه وذلك بشرط توافر

⁹ IVSC - معايير التقييم الدولية.

المعلومات اللازمة عن العقار والعقارات المقارنة بها. ومن أحد سلبياتها صعوبة معرفة ما إذا كانت العقارات المقارنة قد بيعت في ظروف طبيعية أم في ظروف خاصة. وأهم عناصر المقارنة هي:

- الموقع العام والخاص للعقار.
- حالة العقار (عمر العقار - تمديدات التدفئة وغيرها الكثير).
- مواصفات العقار.
- الإضافات الموجودة في العقار.
- الوضع التنظيمي للعقار.
- وضع السوق والظروف الاقتصادية القائمة بتاريخ البيع.
- نوع ملكية العقار.
- قرب المراكز التجارية والسياحية والطبية من العقار.

2- طريقة حساب التكلفة

هي التكلفة الحالية لإنشاء مبنى آخر، أو بديل مشابه للمبنى القائم، وخصم الاهتلاكات المستحقة من تكلفة الإنشاء، وإضافة قيمة الأرض، بالإضافة إلى ربح المكاوول بعد ذلك يمكن القيام بتعديلات على قيمة الملكية التامة المبينة للعقار المعني، لعكس قيمة العقار الجاري تقييمه. تقوم هذه الطريقة بمقارنة قيمة العقار الموجود وتسويه الفروقات بينه وبين العقار الجديد المشابه له من حيث:

- العمر.
- حالة البناء وصلاحيته للاستخدام، بالاعتماد على مواصفات السوق للعقار الجديد، وعلى أن يجري تقييم الأرض طبقاً لقيمتها السوقية.

تعتبر الحالة المثالية لاستعمال هذه الطريقة هي للمباني قيد الإنشاء، أو المباني الحديثة، والمباني التي لا سوق لها.

إجراءات تنفيذ هذه الطريقة:

1. تقييم سعر الأرض وكأنها خالية من البناء لتحديد، أفضل استعمال لها وفق نظام ضابطة البناء.
2. حساب كلفة البناء الحالية لتكاليف الإنشاء.
3. حساب الاهتلاك.
4. قيمة البناء الحالية = كلفة البناء - الاهتلاك.

5. قيمة العقار = تقييم سعر الأرض + قيمة البناء الحالية.

3- طريقة تقييم قيمة العقار عن طريق الدخل

هي عملية تقدير قيمة العقار من خلال العوائد المادية، أو الدخل الذي سيوفره العقار المراد تقييمه، أي تحليل مقدرة العقار على جلب فوائد مادية وتحويل هذه الفوائد إلى مؤشر لقيمة تقديرية، ووفق هذه الطريقة تكون قيمة العقار مرتبطة بالدخل الإيجاري للعقار.

مجالات استخدام هذه الطريقة:

تُستخدم هذه الطريقة في تقييم العقارات الاستثمارية المدرة للدخل، مثل الشقق السكنية المؤجرة والمكاتب التجارية، كما تُطبّق أيضاً في تقدير قيمة الفنادق والمراكز التجارية، التي تعتمد قيمتها على التدفقات النقدية المستمرة. وهناك أسلوبين لطريقة الدخل: الطريقة التقليدية والطريقة الحديثة (طريقة التدفق النقدي).

أ. الطريقة التقليدية

القيمة = الدخل (الإيجار) / معدل العائد (الرسلة).

حيث أن معدل العائد = $1/\text{عدد السنوات المطلوبة لكي يغطي الإيجار قيمة العقار}$ ، وكلما كان معدل العائد أقل كلما لزم زمن أطول لاسترجاع المال والعكس صحيح.

ب. الطريقة الحديثة

وتعتمد هذه الطريقة على حساب القيمة الحالية لصافي التدفقات النقدية التي ينتجها العقار خلال عمره الاقتصادي. وتتطلب معلومات عن الإيرادات كالإيجارات الحالية، ونمو الإيجارات المتوقع، ومصروفات العقار، بالإضافة إلى معدل سعر خصم مقبول وذلك للوصول إلى صافي القيمة الحالية وهو سعر اليوم.

صدرت في سوريا خلال السنوات الأخيرة مجموعة من القوانين النازمة لمهنة التقييم العقاري، وما يتفرّع عنها من اختصاصات، ومن أبرزها: القانون رقم (15) لعام 2008 القاضي بإحداث هيئة التطوير والاستثمار العقاري، والقانون رقم (39) لعام 2009 الذي نصّ على إحداث الهيئة العامة للإشراف على التمويل العقاري، إضافةً إلى القانون رقم (8) لعام 2012 الخاص بترخيص وتنظيم ومزاولة مهنة التقييم العقاري، وأخيراً القانون رقم (15) لعام 2021 المتعلّق بالضرائب على البيوع العقارية.

وقد أنيطت بالهيئة العامة للإشراف على التمويل العقاري مسؤولية تأهيل المختمين العقاريين، الذين باتوا معنيين بتقديم التقييمات والاستشارات العقارية أمام جميع الجهات العامة والخاصة. ويُشترط على من يرغب في مزاولة هذه المهنة اجتياز امتحان معيّن، بعد الخضوع لفترة تدريب لدى خبير عقاري معتمد من قبل الهيئة.

المبحث الثالث

تطبيق انحدار لاسو لتقدير أسعار العقارات في مدينة دمشق ومقارنته بانحدار OLS

تمهيد

يناقش هذا المبحث نتائج التحليل الإحصائي، الذي أُجري على عيّنة من البيانات العقارية، المأخوذة من مدينة دمشق، وذلك بهدف دراسة العوامل المؤثرة في تحديد القيمة السوقية للعقارات. وتأتي هذه الخطوة ضمن السياق التطبيقي للدراسة، حيث خلّلت البيانات بعد معالجتها وتنظيفها وتحويلها إلى صيغ قابلة للمعالجة الكمية.

وُظف عدد من الأدوات الإحصائية لتحليل البيانات، شملت الإحصاءات الوصفية، وتحليل الارتباط الخطي عن طريق معامل ارتباط بيرسون، والانحدار الخطي المتعدد مع تحويل لوغاريتمي للمتغير التابع (السعر)، بالإضافة إلى استخدام انحدار لاسو لتقليص عدد المتغيرات وتحسين كفاءة النموذج .

يعرض هذا المبحث النتائج ضمن تسلسل منهجي يبدأ بالإحصاءات الوصفية للعيّنة، مروراً بقياس العلاقات بين المتغيرات، وانتهاءً ببناء نموذج للعلاقة بين أسعار العقارات وخصائصها، مدعوماً بالجدول والرسوم البيانية التوضيحية.

3.3.1 مجتمع وحدود الدراسة.

يتكوّن مجتمع الدراسة من سوق العقارات السكنية في مدينة دمشق، وقد شملت العينة شققاً سكنية معروضة للبيع في عدد من الأحياء ذات الطابع الاجتماعي والاقتصادي المتنوع. وقد جُمعت البيانات من خلال الرصد المباشر لإعلانات البيع المنشورة عبر الإنترنت ومنصات التواصل الاجتماعي، بالإضافة إلى التواصل مع مكاتب وساطة عقارية ضمن مناطق مختلفة في المدينة. وقد تم اختيار العقارات من مناطق متعدّدة داخل مدينة دمشق بما يحقق تنوعاً مكانياً في خصائص العينة.

تقتصر هذه الدراسة على تحليل بيانات عقارات سكنية ضمن الحدود المكانية لمدينة دمشق. أما من حيث الحدود الزمانية، فقد جُمعت البيانات خلال الفترة الممتدة بين 1 كانون الثاني 2024 و 1 تشرين الثاني 2024.

بلغ عدد السجلات العقارية المُجمّعة خلال مرحلة التحضير للتحليل 254 سجلاً. وتبيّن وجود أربع سجلات واردة من منطقة المالكى تحتوي على أسعار بيع شاذة للغاية (تجاوزت 15 مليار ليرة سورية)، وهي أسعار تفوق المعدلات السائدة ضمن العينة. وبناءً على ذلك، استُبعدت هذه السجلات باعتبارها قيماً شاذة (Outliers)، قد تؤثر على دقة التحليل. وبذلك أصبح الحجم النهائي للعيّنة المستخدمة في الدراسة 250 سجلاً عقارياً.

3.3.2 جمع البيانات من السوق العقاري في مدينة دمشق

تتطلب عملية جمع البيانات المتعلقة بالسوق العقاري في مدينة دمشق جهداً مكثفاً، نظراً لغياب سجلٍ عقاري رقميٍّ موحدٍ، أو قاعدة بيانات مركزية مفتوحة المصدر. وقد استدعى هذا الواقع الاعتماد على منهجية بحثية ميدانية متعدّدة المصادر، بغية الحصول على عيّنة تعبر عن خصائص السوق العقارية في المدينة. فيما يلي توضيح لأبرز الخطوات والمكونات التي شملتها عملية جمع البيانات، والمنهجية التي جرى اعتمادها، إضافة إلى أبرز التحديات التي ظهرت خلال سير العمل.

3.3.3 مصادر البيانات والمنهجية المتبعة

جمعت البيانات خلال الفترة الممتدة من [2024/01/01] إلى [2024/11/01]، بالاعتماد على عدد من المصادر العامة وشبه الرسمية، منها:

- **المواقع الإلكترونية المحلية المتخصصة بالعقارات**، حيث تُنشر مئات الإعلانات اليومية لبيع وشراء الشقق والعقارات ضمن مدينة دمشق. وعلى الرغم من أن هذه المواقع غير خاضعة لرقابة رسمية، إلا أنها توفر كمّاً كبيراً من البيانات الثانوية، وإن كانت غير موحّدة من حيث الشكل والمحتوى.
- **مجموعات وصفحات فيسبوك** المخصصة للإعلانات العقارية، والتي تُستخدم بكثافة من قبل الوسطاء العقاريين والمواطنين العاديين لعرض ممتلكاتهم.
- **التواصل المباشر مع مكاتب الوساطة العقارية** في مختلف الأحياء، حيث وفّرت هذه المكاتب معلومات إضافية عن الأسعار، وأساليب التسعير، والمواصفات الأكثر طلباً.

3.3.4 المتغيرات المعتمدة في التحليل

اختيرت مجموعة من المتغيرات التي ظهرت بتكرار كافٍ ضمن الإعلانات العقارية، ويُفترض أن لها أثراً مباشراً أو غير مباشر في تحديد السعر المعلن للعقار. وقد جرى تصنيف هذه المتغيرات إلى متغير تابع ومتغيرات مستقلة، كما يلي:

المتغير التابع

- **السعر المعلن (Price)** يُمثّل المتغير الرئيس في التحليل، ويعبّر عن السعر المطلوب بالليرة السورية، كما ورد في الإعلان.

المتغيرات الكمية المستقلة

- **المساحة (Area)** وهي المساحة الإجمالية للعقار بالمتر المربع، وتُعد من أبرز محددات السعر.

- عدد الغرف (Number of Rooms) مؤشر على حجم العقار الداخلي وقدرته الاستيعابية.
- الارتفاع الطابق (Floor) يؤثر على سهولة الوصول، خاصة في الأبنية التي تفتقر إلى المصاعد.
- جودة الإكساء (Quality) مقياس رتبي من 1 إلى 5 يعبر عن الحالة العامة للعقار من حيث الإنشاء والتشطيب.
- الاتجاه (Direction) يعكس عدد الواجهات (من 1 إلى 3).
- وجود مصعد (Elevator) خاصية مؤثرة خصوصاً في الطوابق العليا.
- الفرش (Furnishing) ما إذا كان العقار مفروشاً بالكامل أو جزئياً.
- وجود تدفئة مركزية (Heating) سواء أكانت مركزية أم فردية.
- وجود شرفة (Balcony) عامل جاذب للمشتريين ضمن الأحياء المكتظة.
- الموقع داخل الحي (Location) يشير إلى ما إذا كان العقار يقع على شارع رئيسي أو داخلي.
- وجود مرآب (Garage) يشير إلى ما إذا كان العقار يحوي على مرآب.

الجدير بالذكر أنّ بعض المتغيرات التي ظهرت بتكرار محدود جداً — مثل وجود طاقة شمسية، قبو مشترك، سطح مشترك، وخط كهرباء ممتاز — قد استُبعدت من التحليل الإحصائي، نظراً لعدم كفاية تمثيلها داخل العينة. وقد اعتُمد معيار الحد الأدنى للتكرار، كشرط أساسي لضمان توازن العينة، وتجنّب التأثير المحتمل للمتغيرات النادرة في نتائج النموذج.

3.3.5 التحديات والصعوبات

واجهت عملية جمع البيانات عدة صعوبات على المستويين العملي والمنهجي، أبرزها:

- غياب التوحيد في شكل ومحتوى الإعلانات، ما جعل عملية استخراج المتغيرات عملية مجهدة، وتتطلب جهداً يدوياً كبيراً في التنسيق والتصنيف.
- عدم دقة الأسعار المعروضة، نظراً لأن السعر المعلن قد يكون مبالغاً فيه أو خاضعاً للتفاوض، ما يخلق فجوة بين السعر المعلن والسعر الحقيقي للصفقة.
- نقص المعلومات في بعض الإعلانات، مثل عدم ذكر الطابق، أو الإكساء، أو حالة البناء، ما اضطر إلى استبعاد بعض السجلات التي كان من الصعب الحصول عليها.
- الافتقار إلى بيانات رسمية أو قابلة للتحقق، بسبب عدم وجود سجل عقاري رقمي، ما يجعل من الصعب التأكد من صحة بعض الإعلانات، أو مقارنة الأسعار مع بيانات حكومية موثوقة.

- **التباين الحاد بين الأحياء**، حيث تختلف القيمة العقارية داخل الحي نفسه، بحسب الشارع أو الإطلالة أو مستوى البناء، ما يضيف طبقة إضافية من التعقيد عند تصميم نموذج التنبؤ.
- **غياب إطار معاينة رسمي للسوق العقاري**، شكّل هذا الغياب تحدياً إضافياً، إذ فرض الاعتماد على عينة غير احتمالية جُمعت من مصادر متعددة، مع السعي إلى تحقيق تنوّع مكاني واقتصادي في خصائص العقارات المدروسة.

وعلى الرغم من هذه التحديات، تُعد هذه العينة من البيانات محاولة نادرة لرصد ملامح السوق العقاري الدمشقي، وتحويلها إلى صورة كمية قابلة للمعالجة الإحصائية. وتُشكل البيانات الناتجة الأساس الذي سيُبنى عليه النموذج الإحصائي لاحقاً.

3.3.6 معالجة البيانات وتحضيرها للتحليل

بعد استكمال عملية جمع البيانات الأولية من مصادر متعدّدة، برزت الحاجة إلى تنفيذ معالجة منهجية ودقيقة، تضمن جاهزية البيانات للتحليل الإحصائي. حيث شملت هذه المعالجة عدداً من الخطوات الأساسية، بدءاً من التعامل مع القيم المفقودة، مروراً بترميز المتغيرات النوعية، ووصولاً إلى توحيد القيم الكمية بما يتوافق مع متطلبات النماذج الإحصائية المستخدمة.

3.3.6.1 معالجة القيم المفقودة (Missing Values)

لوحظ وجود غياب جزئي في عدد من المتغيرات، مثل "إكساء المطبخ" أو "وجود المولدة"، وذلك إمّا لقصور في المعلومات الواردة في الإعلانات العقارية نفسها، أو لعدم أهمية هذه التفاصيل في تقدير المعلن. وفي هذا الإطار، جرى التعامل مع القيم المفقودة بأسلوب وحيد، يقضي باستبعاد السجلات بشكل كامل من قاعدة البيانات التي تحتوي على عدد كبير من القيم المفقودة، بحيث يتعدّر تقديرها أو يؤدّي تقديرها إلى تشويه البيانات.

3.3.6.2 تحويل البيانات النوعية (Categorical Encoding)

حوّلت المتغيرات النوعية باستخدام الترميز المناسب وفقاً لطبيعة كل متغير:

- استخدم الترميز الثنائي (Binary Encoding) بالنسبة للمتغيرات التي تملك حالتين متقابلتين (مثل وجود مصعد أو عدمه، وجود كراج أو عدمه)، حيث أُسندت القيمة (1) لوجود الخاصية والقيمة (0) لغيابها.
- اعتمد الترميز الرتبي (Ordinal Encoding) عندما تتخذ المتغيرات قيماً مرتّبة (مثل مستوى جودة الإكساء: سيئ جداً، سيئ، متوسط، جيد، ممتاز)، حيث جرى إسناد أرقام تمثّل رتب كل فئة (1، 2، 3، 4، 5) بُغية الحفاظ على البنية الرتبية بين القيم.

3.3.6.3 توحيد القيم الكمية (Standardization & Scaling)

وُحدت المتغيرات الكمية (مثل المساحة وعدد الغرف) باستخدام مقياس Z-Score وتُعد هذه الخطوة ضرورية عند استخدام انحدار لاسو حيث يكون لحجم المتغير تأثير مباشر في معامل التقدير.

3.3.6.4 معالجة متغير السعر (Logarithmic Transformation)

خضع متغير السعر لتحويل باستخدام اللوغاريتم الطبيعي، وذلك بهدف تعزيز قابلية تفسير النموذج الإحصائي في المراحل اللاحقة، مع الحفاظ على السمات الأساسية للمتغير دون إحداث تغييرات جوهرية في طبيعته.

3.3.6.5 تقسيم البيانات (Train-Test Split)

قُسمت مجموعة البيانات الخاصة بمدينة دمشق إلى مجموعتين: مجموعة التدريب بحجم 80% (200) مشاهدة ومجموعة الاختبار بحجم 20% (50) مشاهدة. وتجدر الإشارة إلى أنه أُجريت جميع العمليات الخاصة بمعالجة البيانات لكل مجموعة بشكل منفصل، وذلك للتقليل من فرص حدوث تسرب البيانات بين المجموعتين، كما وُضح سابقاً. الأمر الذي سيساهم في ضمان تقييم موضوعي لأداء النماذج الإحصائية لاحقاً.

3.3.7 الأدوات والبرامج الإحصائية المستخدمة

تم الاعتماد بشكل أساسي على لغة البرمجة Python الإصدار 3.0 وفيما يلي مجموعة من الأدوات الإحصائية والمكتبات الجاهزة وهي نفسها الموجودة في الجدول رقم (1) التي استخدمت في تحليل أسعار العقارات الخاصة بمدينة Ames، بالإضافة إلى المكتبات الآتية:

الجدول (5): الأدوات والمكتبات الإحصائية المستخدمة في تحليل البيانات الخاصة بمدينة دمشق.		
اسم المكتبة	الوصف - الغرض من المكتبة	المصدر
Statsmodels	مكتبة تُستخدم في تحليل الانحدار الخطي، وتقدير النماذج الإحصائية، بالإضافة إلى إجراء التحليلات التشخيصية مثل (Anderson-Darling)، واختبار التجانس (Breusch-Pagan)،	https://www.statsmodels.org

	وحساب معامل تباين التضخم (VIF).	
https://docs.scipy.org/doc/scipy	تُستخدم لإجراء اختبارات إحصائية متعددة، وجرى استخدامها في هذه الدراسة لتطبيق اختبار Shapiro-Wilk للتحقق من التوزيع الطبيعي للبواقي.	scipy.stats
المصدر: إعداد الباحث بالاعتماد على الوثائق الرسمية لكل مكتبة برمجية.		

3.3.8 التحليل الوصفي لمتغيرات الدراسة.

3.3.8.1 الإحصاءات الوصفية للمتغيرات الكمية

فيما يلي جدول يلخص أهم المؤشرات الإحصائية الوصفية لأبرز المتغيرات الكمية المعتمدة في التحليل، يلي ذلك تفصيل نصي لكل متغير على حدى:

الجدول (6): الإحصاءات الوصفية للمتغيرات الكمية في عينة الدراسة الخاصة لمدينة دمشق.					
المتغير	المتوسط	الوسيط	الانحراف المعياري	الحد الأدنى	الحد الأعلى
السعر	1,108,020,000	1,100,000,000	346,898,400	270,000,000	2,100,000,000
المساحة	90	90	25	50	180
عدد الغرف	–	2	1*	1	5
الارتفاع الطابقي	–	4	4*	0	8
جودة الإكساء	–	3	2*	1	5
الاتجاه	–	2	1*	1	3
* = عبارة عن المدى الربيعي.					
المصدر: إعداد الباحث اعتماداً على قاعدة بيانات مدينة دمشق باستخدام لغة البرمجة Python					

يوضح الجدول رقم (6) أن متوسط السعر المعلن للعقارات في العينة يبلغ نحو 1.1 مليار ليرة سورية، وهو ما يتقاطع تقريباً مع الوسيط البالغ 1.1 مليار أيضاً، ما يشير إلى توزيع متقارب نسبياً وقد بلغت قيمة الانحراف المعياري نحو 347 مليون ل.س.

تراوحت المساحة الكلية للعقارات بين 50 و 180 متراً مربعاً، بمتوسط 90 متراً مربعاً، بينما بلغ متوسط عدد الغرف نحو 2 غرفة، ما يدل على تنوع نسبي في خصائص العقارات من حيث الحجم والتقسيم الداخلي.

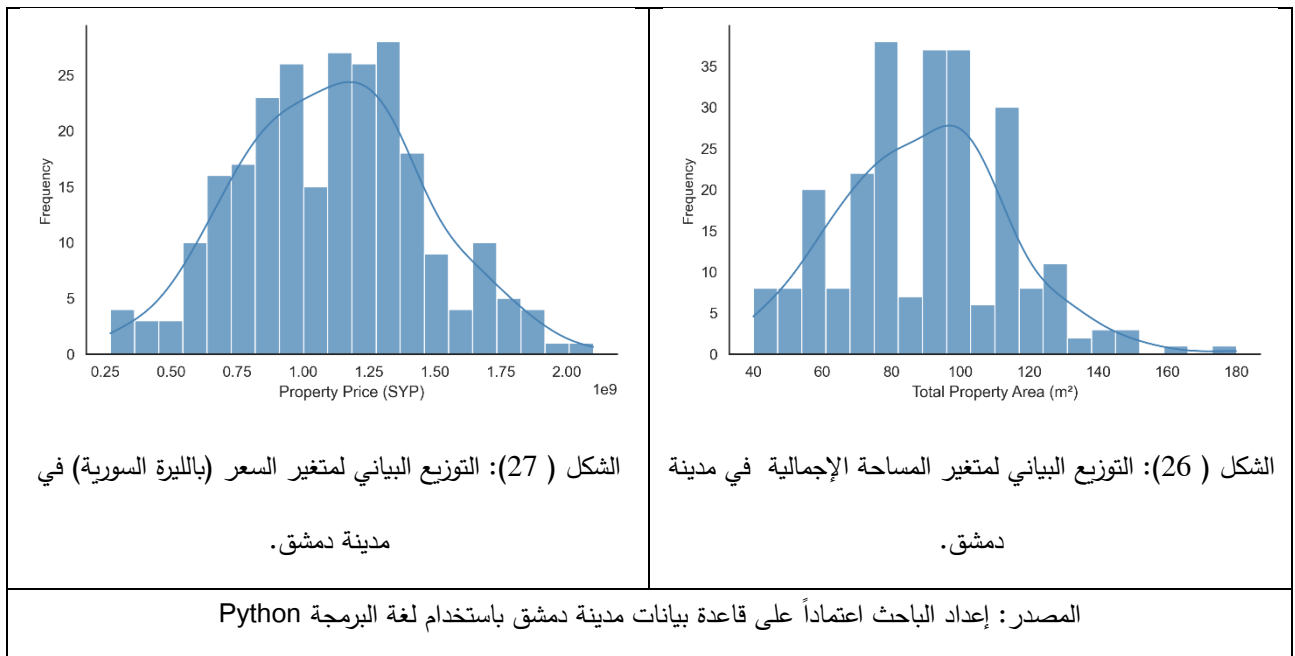
أما متغير عدد الغرف، فقد بلغ الوسيط غرفتين، وبلغ المدى الربيعي نحو غرفة واحدة، ما يدل على تقارب في التقسيم الداخلي للعقارات.

أما متغير الارتفاع الطائقي، فقد بلغ الوسيط 4 طوابق، وبلغ المدى الربيعي نحو 4 طوابق.

وبلغ وسيط جودة الإكساء 3 من 5، وبلغ المدى الربيعي نحو 2، ما يشير إلى تباين في مستوى درجة الإكساء بين العقارات.

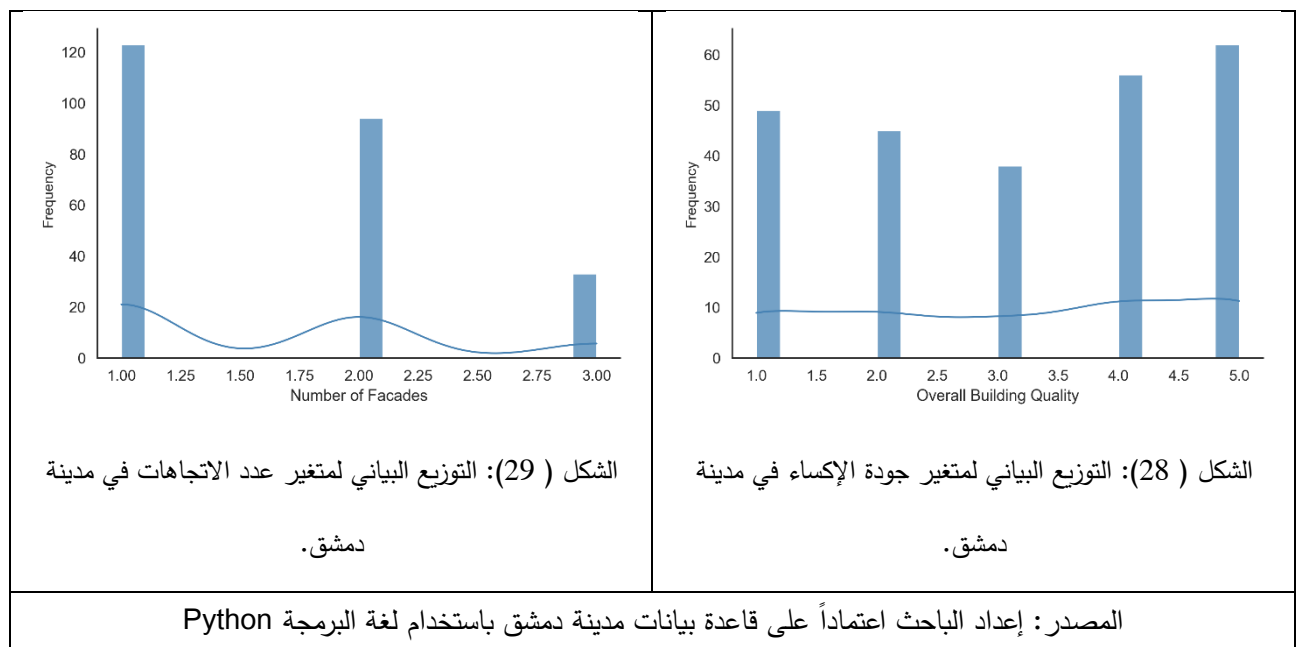
في حين بلغ وسيط عدد الاتجاهات (الواجهات) 2، وبلغ المدى الربيعي نحو 1، وهو ما يعكس تفاوتاً محدوداً في عدد الإطلاقات بين العقارات.

وفيما يلي مجموعة من الرسوم البيانية التوضيحية للمتغيرات المدروسة الخاصة بمدينة دمشق.



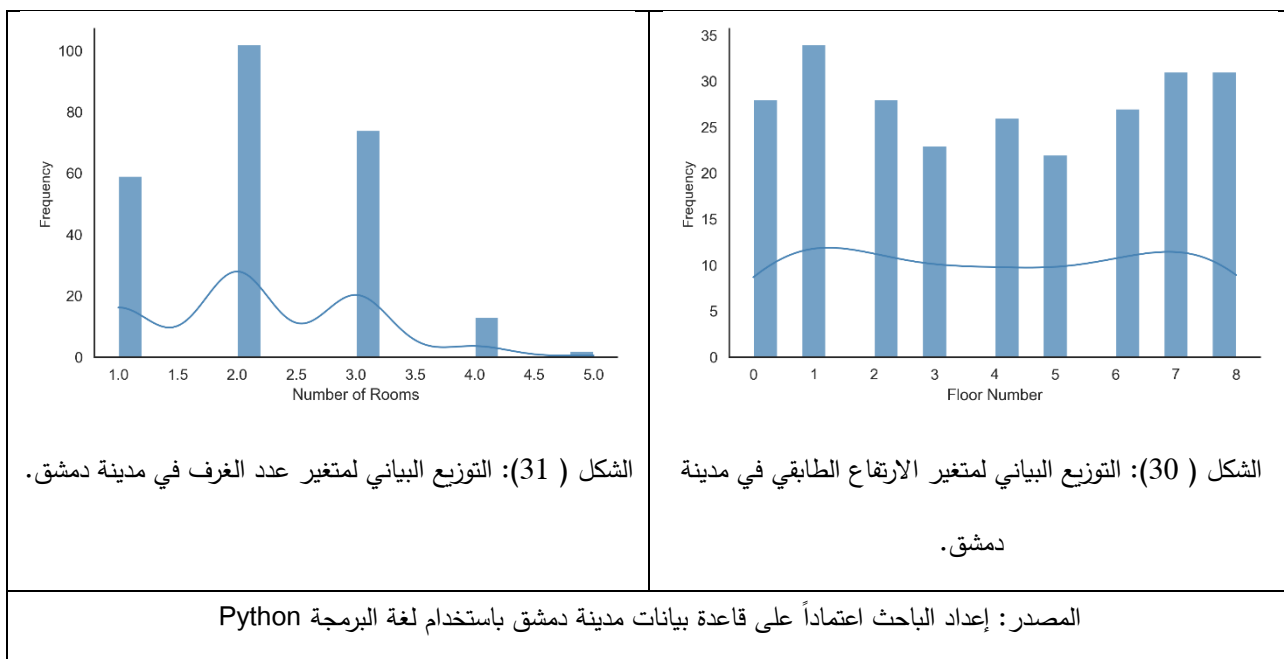
يُظهر الشكل رقم (26) أن معظم الشقق تتراوح مساحتها بين 80 إلى 110 متر مربع، وهو ما يُعد منطقياً بالنظر إلى طبيعة الطلب في دمشق، خصوصاً في الفئات المتوسطة. كما يلاحظ وجود ميل بسيط نحو اليمين، مما يدل على وجود عدد محدود من العقارات ذات المساحات الكبيرة (تتجاوز 150 م²)، والتي تُعد أقل شيوعاً ضمن العينة.

أما الشكل رقم (27) فيوضح أسعار معظم العقارات تتركز حول 1.2 مليار ليرة سورية، مع توزّع قريب من الشكل الطبيعي، وإن كان يميل قليلاً نحو اليمين. وهذا يشير إلى وجود بعض العقارات ذات الأسعار المرتفعة نسبياً (تتجاوز 1.8 مليار)، والتي تُعد من الفئة الأعلى ضمن سوق دمشق السكني، ولكنها لا تشكّل النسبة الأكبر من السوق. كما يشير التوزيع إلى تجانس نسبي في الأسعار مع بعض القيم الأعلى التي قد تعود لمناطق مرغوبة أو مواصفات مميزة.



يُظهر الشكل رقم (28) أن القيم 4 و5 سجلت التكرار الأعلى بين فئات جودة الإكساء، حيث تجاوز عددها 50، في حين أن القيم من 1 إلى 3 سجلت تكرارات أقل تراوحت بين 38 و48 تقريباً.

أما الشكل رقم (29) فيظهر أن معظم القيم تركزت عند الواجهة الواحدة والواجهتين، حيث سُجّل أعلى تكرار عند القيمة 1 (بأكثر من 110 حالة تقريباً)، تليها القيمة 2، بينما كانت القيمة 3 الأقل تكراراً.



يُظهر الشكل رقم (30) أن الطوابق الأكثر تكراراً هي الطابق الأول (1)، والسابع (7)، والثامن (8)، حيث تجاوز عدد المشاهدات في كل منها 30 حالة، وهو ما يعكس ميلاً جزئياً نحو السكن في الطوابق المرتفعة نسبياً والطابق الأول. بالمقابل، كانت أقل التكرارات في الطابق الرابع (4) والخامس (5)، حيث سجل كل منهما عدداً أدنى من التكرارات مقارنة ببقية الطوابق.

بشكل عام، يتوزع السكن على مختلف الطوابق دون انحياز حاد نحو طابق معين، باستثناء القيم الملحوظة في الطابق الأول والطابقين العلويين (7 و8).

يبين الشكل رقم (31) أن الشقق الأكثر شيوعاً هي التي تحتوي على غرفتين، حيث تجاوز تكرارها حاجز 100 حالة، تليها الشقق ذات ثلاث غرف، ثم الشقق ذات غرفة واحدة. وفي المقابل، يظهر انخفاض واضح في عدد الشقق التي تحتوي على أربع غرف أو أكثر، حيث تراجعت التكرارات بشكل حاد.

يعكس هذا التوزيع ميلاً عاماً نحو الوحدات السكنية ذات الحجم المتوسط (غرفتين إلى ثلاث غرف)، وهو أمر متوقع في سياقات عمرانية تستهدف العائلات الصغيرة أو الأفراد، مع محدودية في انتشار الشقق الواسعة ذات الأربع غرف أو أكثر.

3.3.8.2 الإحصاءات الوصفية للمتغيرات الثنائية

تعرض هذه الفقرة التوزيع التكراري للمتغيرات الثنائية التي رُمزت بصيغة (0 = لا، 1 = نعم)، والتي تشمل: وجود الأثاث، ونظام التدفئة، والشرفة، والمصعد، وموقع العقار. ويوضح الجدول أدناه عدد التكرارات والنسب المئوية لكل

فئة ضمن هذه المتغيرات، ما يسهم في تكوين صورة أولية عن خصائص العينة من حيث توفر الخدمات والمواصفات العقارية.

الجدول (7): توزيع المتغيرات الثنائية وفق الفئات (1/0) من حيث التكرار والنسبة المئوية الخاصة بمدينة دمشق.				
المتغير	الفئة 0 (لا يوجد)		الفئة 1 (يوجد)	
	التكرار	النسبة	التكرار	النسبة
وجود الأثاث	22	8.8%	228	91.2%
وجود تدفئة مركزية	70	28%	180	72%
وجود شرفة	148	59.2%	102	40.8%
وجود مصعد	194	77.6%	56	22.4%
وجود مرآب	60	24%	190	76%
الموقع	88	35.2%	162	64.8%
المصدر: إعداد الباحث اعتماداً على قاعدة بيانات مدينة دمشق باستخدام لغة البرمجة Python				

يُظهر الجدول (7) توزيع المتغيرات الثنائية للعقارات في مدينة دمشق وفق الفئات (1/0)، حيث تبين أن نسبة كبيرة من العقارات تحتوي على أثاث (91.2%)، ما يعكس توجهاً واسعاً نحو عرض الوحدات السكنية وهي مفروشة. كما أن 72% من العقارات تتوفر فيها وسيلة تدفئة مركزية.

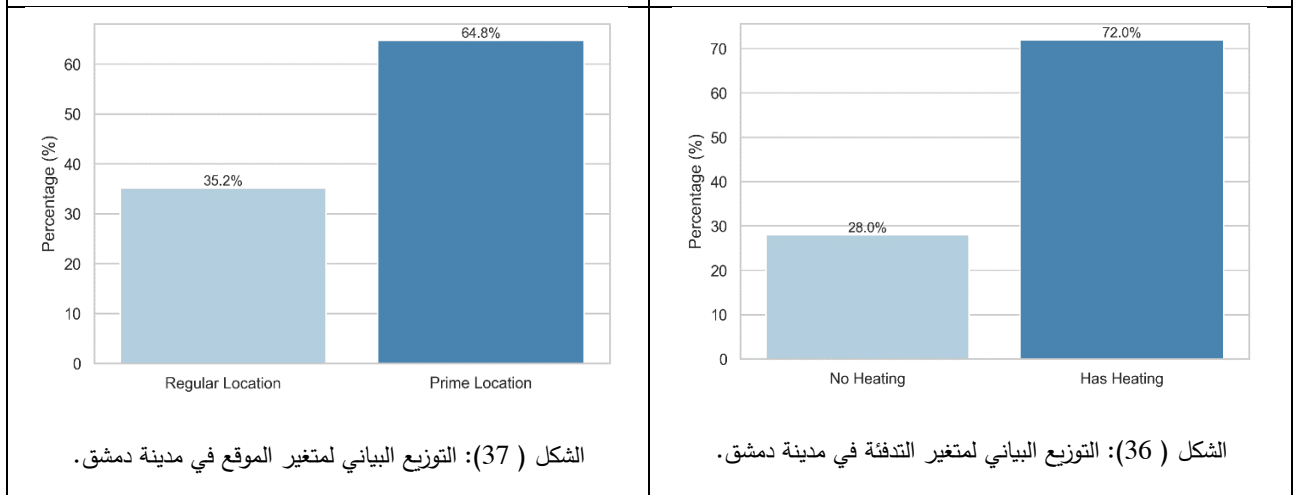
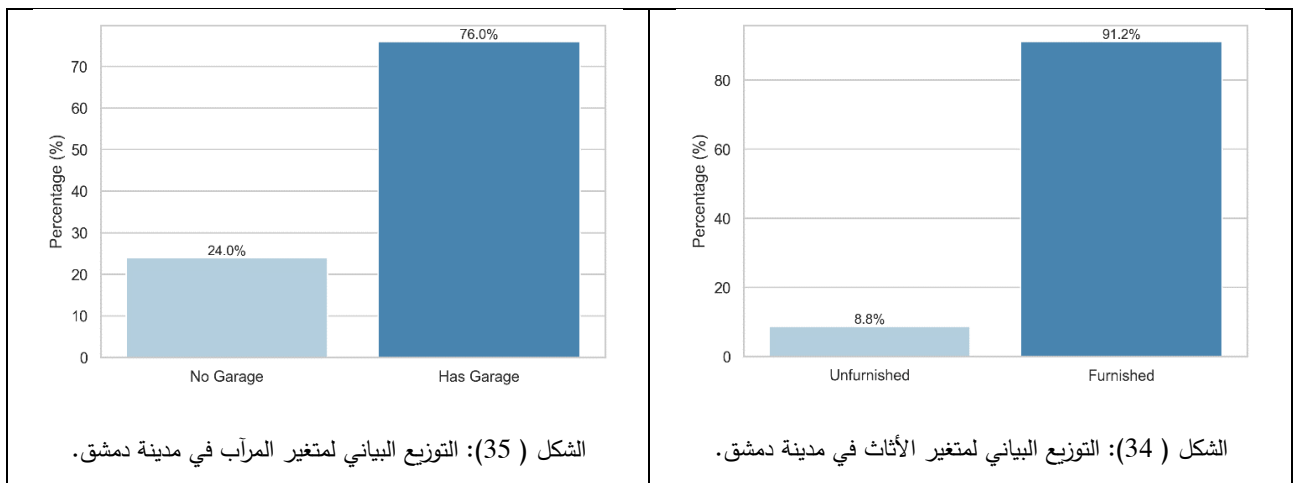
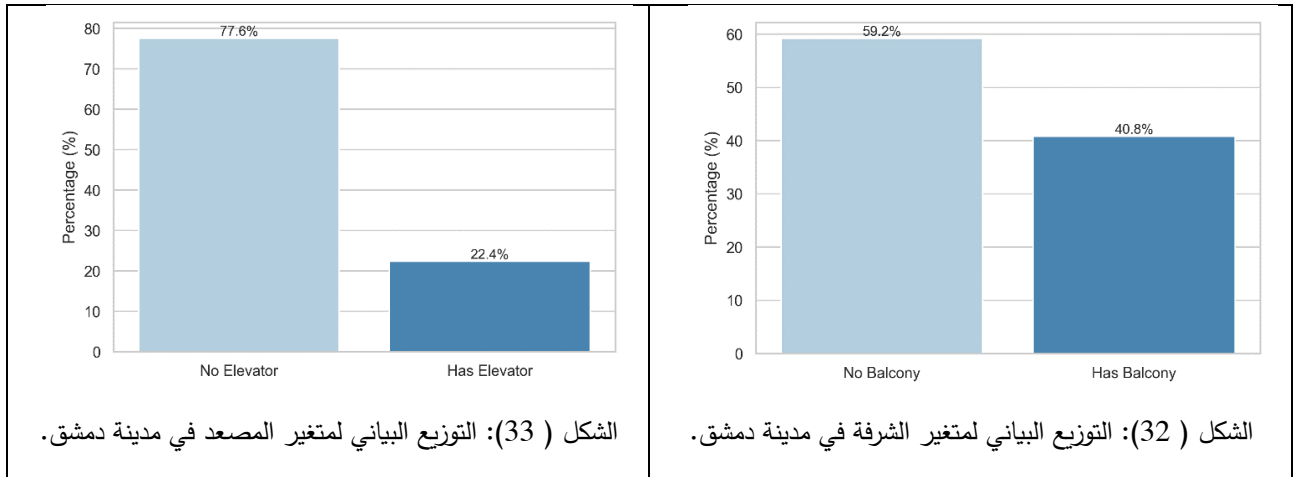
أما بالنسبة لوجود شرفة، فتتوفر في 40.8% من العقارات فقط، وهي نسبة تقل عن النصف، في حين أن وجود المصعد سُجِّل بنسبة 22.4%، وهو ما يعكس قلة الأبنية المجهزة بهذا العنصر، وربما يرتبط ذلك بارتفاع عدد الطوابق في بعض الأبنية دون وجود تجهيزات حديثة.

وبالنسبة لمتغير "الموقع"، فهو يعكس تموضع العقار داخل الحي السكني، وصُنف إلى فئتين وفق الآتي:

- الفئة (1): العقارات الواقعة على شارع رئيسي.
 - الفئة (0): العقارات الواقعة في شارع فرعي داخلي.
- وتشير البيانات إلى أن 64.8% من العقارات تقع على شارع رئيسي، ما يعكس ميلاً في السوق نحو العقارات ذات الواجهة على الشوارع الأوسع أو الأكثر حركة.

وبالنسبة لوجود المرآب، فقد سُجِّل في 76% من الحالات، وهي نسبة مرتفعة نسبياً، وقد ترتبط بطبيعة الأبنية الحديثة أو الفيلات، ما قد يكون له أثر على القيمة السوقية للعقار، ضمن النماذج التنبؤية المستخدمة لاحقاً.

وفيما يلي، مجموعة من الرسوم البيانية التوضيحية التي تدعم ما ورد في الجدول (9)، حيث توفر تمثيلاً للتوزيع النسبي للفئات الثنائية (1/0) لكل متغير. وتُظهر هذه الأشكال بوضوح هيمنة الفئة (1) في معظم المتغيرات، لاسيما في متغيرات وجود الأثاث، والموقع. في المقابل، تُسجل الفئة (0) النسبة الأعلى في متغيري وجود المصعد ووجود الشرفة، ما يعكس محدودية توفر هذه الميزات في جزء كبير من العينة المدروسة.



المصدر: إعداد الباحث اعتماداً على قاعدة بيانات مدينة دمشق باستخدام لغة البرمجة Python

3.3.9 بناء النموذج الإحصائي للعلاقة بين الخصائص العقارية والسعر

يهدف هذا الجزء إلى بناء نموذج إحصائي يُعبّر عن العلاقة بين سعر العقار من جهة، ومجموعة من الخصائص الكمية والرتبية التي تُمثّل خصائص العقار من جهة أخرى. يُعتمد أولاً نموذج الانحدار الخطي المتعدد (OLS) لتقدير هذه العلاقة وتحليل تأثير المتغيرات المختلفة، ثم يُستخدم انحدار لاسو بهدف تقليص عدد المتغيرات واختيار الأكثر تأثيراً منها. وفي المرحلة الأخيرة، تُجرى مقارنة بين النموذجين.

3.3.9.1 الارتباط الخطي بين المتغيرات الكمية والسعر

حُسب معامل الارتباط بيرسون لقياس قوة واتجاه العلاقة بين المتغير التابع (السعر) وكل من المتغيرات الكمية المستقلة: المساحة، عدد الغرف، الارتفاع الطائقي، جودة الإكساء، وعدد الاتجاهات (الواجهات). وفيما يلي الجدول الذي يوضح ذلك:

الجدول (8): نتائج معامل الارتباط بيرسون بين سعر العقار والمتغيرات المستقلة الخاصة في مدينة دمشق.					
المتغير	المساحة	عدد الغرف	الارتفاع الطائقي	جودة الإكساء	عدد الواجهات.
معامل الارتباط	0.66	0.59	-0.07	0.41	0.04
P-value	< 0.05	< 0.05	0.23	< 0.05	0.49

المصدر: إعداد الباحث اعتماداً على قاعدة بيانات مدينة دمشق باستخدام لغة البرمجة Python

وقد أظهرت النتائج وجود ارتباطات متفاوتة القوة، كانت أقوىها بين السعر والمساحة، حيث بلغ معامل الارتباط حوالي (0.66)، وهو ارتباط موجب متوسط إلى قوي، يعكس العلاقة المنطقية المباشرة بين مساحة العقار وارتفاع سعره. كما ظهر ارتباط موجب متوسط بين السعر وعدد الغرف (0.59)، ما يشير إلى أن زيادة عدد الغرف غالباً ما يقترن بزيادة القيمة السوقية للعقار.

أما جودة الإكساء فقد سجّلت ارتباطاً موجباً أيضاً (0.41)، ما يدل على أن العقارات ذات الجودة الأعلى تميل لأن تكون أغلى، وهو أمر متوقع. بينما كان معامل الارتباط بين السعر والارتفاع الطائقي (Floor) سالباً وضعيفاً (-0.07)، ما يشير إلى ميل طفيف لانخفاض السعر مع ارتفاع الطابق، وهو ما قد يرتبط بعدم توفر المصعد أو بتفضيلات المشتري.

وأخيراً، أظهر متغير عدد الاتجاهات (واجهات العقار) ارتباطاً ضعيفاً جداً مع السعر (0.04)، ما يدل على أن هذا المتغير لا يُعد من العوامل المؤثرة بقوة في تحديد السعر ضمن العينة المدروسة.

3.3.9.2 تحليل الانحدار الخطي المتعدد

بعد الانتهاء من تحليل الارتباط الخطي بين السعر وعدد من المتغيرات المستقلة الكمية والرتبية، قُسمت البيانات إلى عيّنتين: عينة تدريبية بنسبة 80%، وأخرى للاختبار بنسبة 20%، وذلك بهدف ضمان دقة التقييم واختبار قدرة النموذج على التعميم. بعد ذلك، جرى بناء نموذج الانحدار الخطي المتعدد لتفسير التغيرات في السعر بالاستناد إلى مجموعة من الخصائص العقارية. وفيما يلي نعرض النتائج التي أسفر عنها هذا النموذج:

الجدول (9): نتائج نموذج الانحدار الخطي المتعدد للعلاقة بين أسعار العقارات والخصائص المتعلقة بها في مدينة دمشق.						
المتغير	β	$SE \beta$	Standized β	$P - value$	Lower 95% CI	Upper 95% CI
const	19.5856	0.082		0.00	19.424	19.747
المساحة	0.0084	0.001	0.2	0.00	0.007	0.01
عدد الغرف	0.0475	0.021	0.04	0.022	0.007	0.088
الارتفاع الطائقي	-0.0339	0.005	-0.09	0.00	-0.044	-0.024
جودة الإكساء	0.0998	0.009	0.15	0.00	0.083	0.117
الاتجاهات	-0.0054	0.018	0.01	0.768	-0.041	0.03
وجود الأثاث	0.0335	0.045	0.01	0.457	-0.055	0.122
وجود تدفئة	0.0145	0.029	0.01	0.612	-0.042	0.071
وجود شرفة	0.0449	0.026	0.02	0.09	-0.007	0.097
وجود مصعد	0.2216	0.031	0.09	0.00	0.161	0.282
الاتجاه	0.0263	0.027	0.01	0.327	-0.026	0.079
وجود مرآب	0.0352	0.03	0.02	0.248	-0.025	0.095
$R^2 = 0.70, Adj - R^2 = 0.68, F(ANOVA) P - value = 0.00$						
RMSE (train) = 0.1958, RMSE (test) = 0.2006						
المصدر: إعداد الباحث اعتماداً على قاعدة بيانات مدينة دمشق باستخدام لغة البرمجة Python						

بلغ معامل التحديد $R^2 = 0.70$ في نموذج الانحدار الخطي المتعدد، ما يدل على أن المتغيرات المستقلة المدروسة تفسّر ما نسبته 70% من التباينات الحاصلة في أسعار العقارات، وهي نسبة مرتفعة نسبياً تشير إلى جودة النموذج وقدرته التفسيرية العالية. كما بلغت قيمة معامل التحديد المعدل ($Adjusted R^2$) نحو 0.68، وهي قريبة جداً من القيمة السابقة، بمعنى أن النموذج جيد التوفيق.

أما الخطأ المعياري للتقدير (Standard Error of the Estimate) فقد بلغ 0.200، مما يشير إلى درجة محدودة من التشتت حول خط الانحدار، ويدل على ملاءمة النموذج وتمثيله الجيد للبيانات.

تشير نتائج اختبار تحليل التباين (ANOVA) F إلى قيمة معنوية جداً عند مستوى دلالة أقل من 0.01، ما يعني أن النموذج ككل ذو دلالة إحصائية، وأن هناك تأثيراً مشتركاً للمتغيرات المستقلة في تفسير تغيرات السعر. بلغت قيمة جذر متوسط مربعات الأخطاء نحو 0.1958 على بيانات التدريب، و0.2006 على بيانات الاختبار، وهي قيم منخفضة نسبياً، مما يعكس قدرة النموذج على التعميم وتقديم تقديرات مقبولة لأسعار العقارات خارج عينة التدريب.

وقد تبين أن المتغير الأكثر تأثيراً في تحديد السعر هو المساحة (Area)، حيث تُظهر النتائج أن كل زيادة بمقدار 10 أمتار مربعة في المساحة تؤدي إلى زيادة متوقعة في السعر بنسبة تقارب 8.4%، وهو تأثير كبير ومنطقي في سياق السوق العقاري.

كما أظهرت النتائج أن جودة الإكساء (Quality) لها تأثير واضح، حيث أن الانتقال من مستوى جودة إلى المستوى الأعلى يؤدي إلى زيادة متوقعة في السعر بنسبة تبلغ حوالي 10%، مما يؤكد أهمية درجة الإكساء في تقييم العقارات.

في المقابل، تبين أن الارتفاع الطابق (Floor) يرتبط بالسعر بعلاقة عكسية، إذ تؤدي كل زيادة بمقدار طابق واحد إلى انخفاض متوقع في السعر بنسبة تقارب 3.3%، ما يعكس تفضيل السكان للطبقات السفلية أو المتوسطة في المباني السكنية.

أما بالنسبة إلى عدد الغرف (Nrooms)، فقد وُجد أن كل زيادة في عدد الغرف تؤدي إلى ارتفاع في السعر بنسبة تقارب 4.8%، وهو ما يعكس ارتباط الحجم الوظيفي للعقار بقيمته السوقية.

وأخيراً، أظهر وجود المصعد (Elevator) تأثيراً معنوياً واضحاً، إذ يؤدي توفر المصعد في البناء إلى ارتفاع في السعر بنسبة تقارب 22%، ما يؤكد الأهمية التي يوليها المشترون لعامل الراحة وسهولة التنقل داخل الأبنية السكنية.

في المقابل، لم تُظهر باقي المتغيرات مثل اتجاه العقار، والأثاث، والتدفئة، والشرفة، والموقع، ووجود مرآب تأثيراً دالاً إحصائياً في هذا النموذج، ما يشير إلى أن دورها في تفسير تباين الأسعار كان محدوداً ضمن عينة الدراسة، أو أنه قد يتأثر بعوامل سياقية أخرى لم تُدرج في هذا التحليل.

بناءً على ما سبق، يمكن الاستنتاج أن الخصائص الجوهرية للعقار مثل المساحة، الجودة، عدد الغرف، الارتفاع الطابق، ووجود المصعد تُعد من العوامل الأساسية المحددة للسعر في سوق العقارات، وأن التركيز على هذه العناصر يمكن أن يسهم في بناء نماذج أكثر دقة لتقدير القيمة السوقية للعقارات.

تجدر الإشارة إلى أنه جرى أيضاً تطبيق أسلوب Stepwise لبناء نموذج الانحدار، وقد أظهرت النتائج تطابقاً بين النموذج الناتج عن هذا الأسلوب والنموذج المُقدَّر باستخدام أسلوب Enter، حيث احتُفظت المتغيرات نفسها التي أثبتت دلالة إحصائية. وبناءً على ذلك، جرى اعتماد نموذج Enter في عرض النتائج وتحليلها.

3.3.10 اختبار افتراضات النموذج

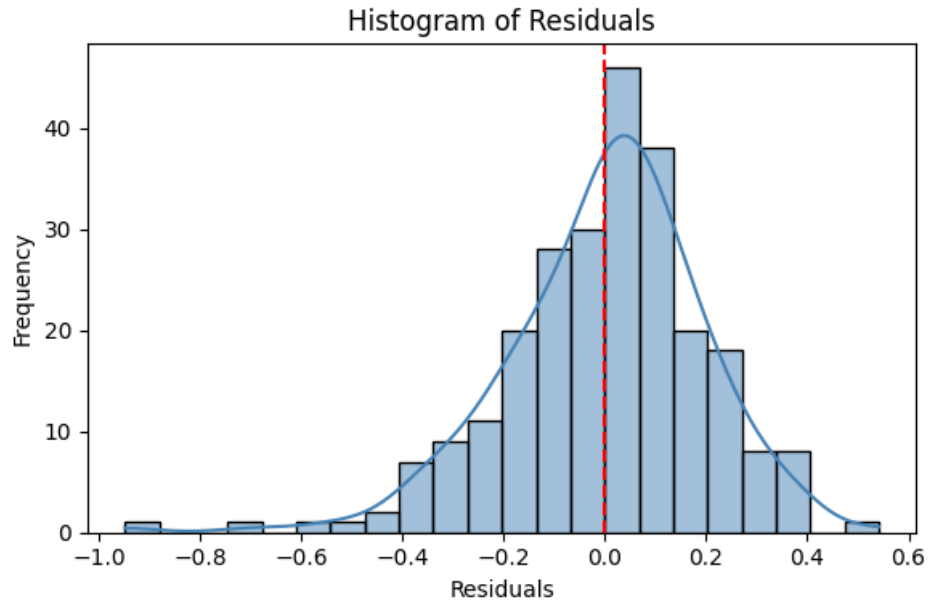
لضمان موثوقية نتائج نموذج الانحدار الخطي المتعدد، أُجريت مجموعة من الاختبارات للتحقق من الفرضيات الأساسية المرتبطة بطريقة المربعات الصغرى العادية (OLS).

3.3.10.1 التحقق من توزيع البواقي (Residuals Normality)

جرى استخدام اختبار شابيرو - ويلك (Shapiro-Wilk) لاختبار فرضية التوزيع الطبيعي للبواقي، ويعرض الجدول الآتي ملخص نتائج هذا الاختبار:

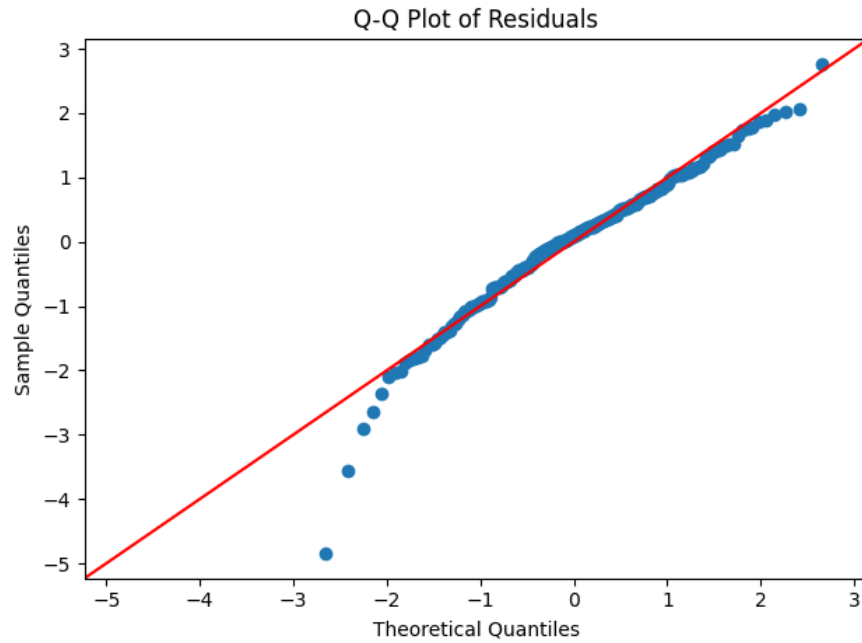
الجدول (10): نتائج اختبار شابيرو - ويلك (Shapiro-Wilk) للتحقق من التوزيع الطبيعي للبواقي لنموذج الانحدار الخطي المتعدد الخاص لبيانات مدينة دمشق.		
الاختبار	الإحصاءة (statistic)	القيمة الاحتمالية (P-value)
Shapiro-Wilk	0.968	0.002
المصدر: إعداد الباحث اعتماداً على قاعدة بيانات مدينة دمشق باستخدام لغة البرمجة Python		

تشير القيمة الاحتمالية (P-value) للاختبار التي كانت أقل من 5% إلى رفض فرضية التوزيع الطبيعي للبواقي. ورغم ذلك، فإن هذه النتيجة تُعد مقبولة في ضوء حجم العينة المستخدم (Wooldridge, 2016, p. 168). في حين أظهرت مخطط Q-Q عن اصطفاك معظم النقاط على الخط القطري 45°، خاصة في وسط التوزيع، مع انحراف بسيط فقط في الأطراف، وهو أمر شائع ومقبول في البيانات الميدانية (Field, 2013)، بالإضافة إلى شكل المدرج التكراري الذي أظهر تركزاً واضحاً للبواقي حول الصفر، مع شكل أقرب إلى التوزيع الطبيعي، والأشكال البيانية الآتية توضح ما سبق.



الشكل (38): المدرج التكراري للبقاوي مع منحنى الكثافة الاحتمالية لنموذج الانحدار الخطي المتعدد لبيانات مدينة دمشق.

المصدر: إعداد الباحث اعتماداً على قاعدة بيانات مدينة دمشق باستخدام لغة البرمجة Python



الشكل (39): مخطط Q-Q للبقاوي لمقارنة التوزيع الفعلي بالتوزيع الطبيعي النظري لنموذج الانحدار الخطي المتعدد لبيانات مدينة دمشق.

المصدر: إعداد الباحث اعتماداً على قاعدة بيانات مدينة دمشق باستخدام لغة البرمجة Python

3.3.10.2 اختبار تجانس التباين (Homoscedasticity)

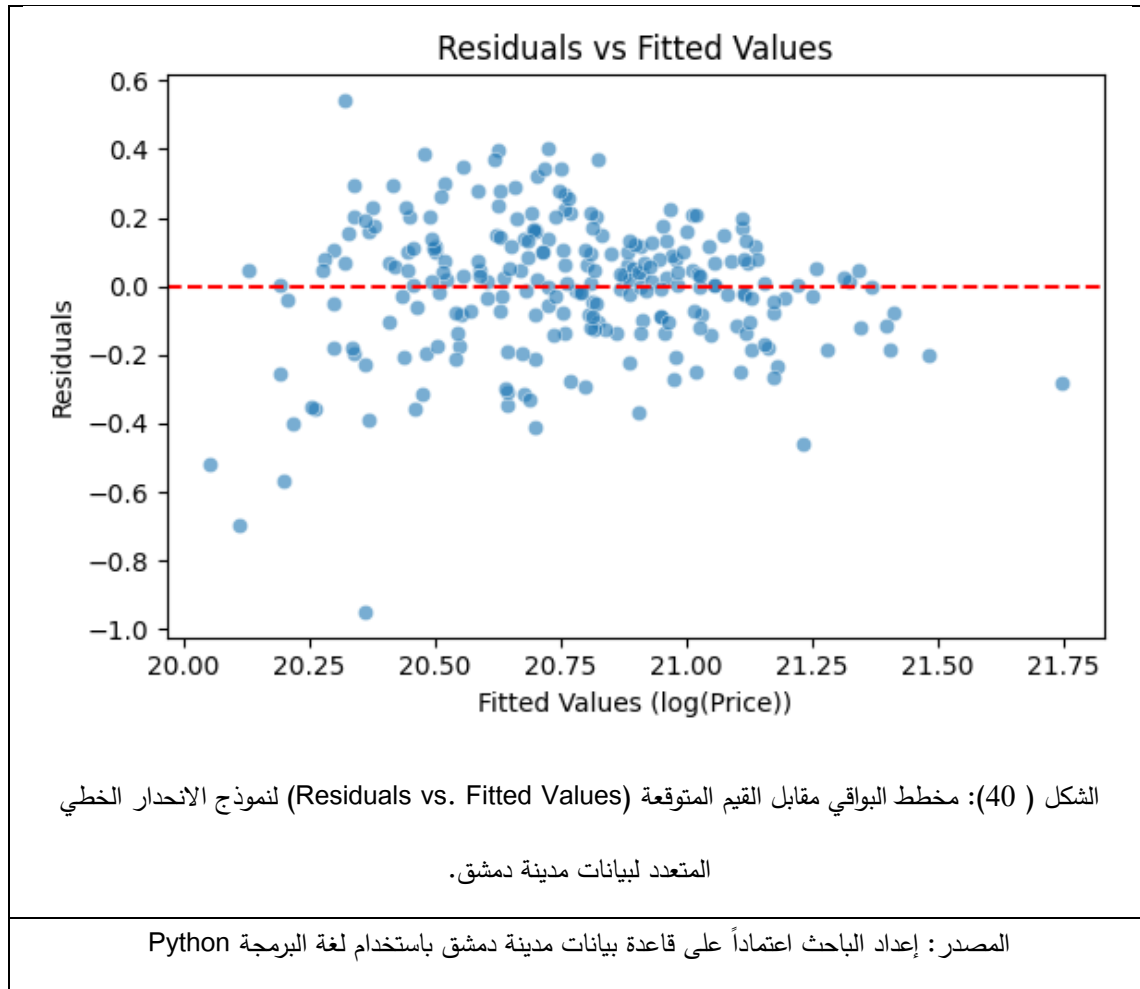
أُجري اختبار بروش – بيغان (Breusch–Pagan test) للتحقق من وجود تجانس في تباين البواقي الناتجة عن النموذج المقدّر، وفيما يلي ملخص النتائج التي أسفر عنها هذا الاختبار:

الجدول (11): نتائج اختبار بريوش – بيغان (Breusch–Pagan) للكشف عن تجانس التباين لنموذج الانحدار الخطي المتعدد الخاص لبيانات مدينة دمشق.		
الاختبار	الإحصاءة (statistic)	القيمة الاحتمالية (P-value)
Lagrange Multiplier (LM)	22.79	0.021
F–statistic	2.17	0.0267
المصدر: إعداد الباحث اعتماداً على قاعدة بيانات مدينة دمشق باستخدام لغة البرمجة Python		

تشير النتائج إلى أن كل من إحصائية LM و F كانت دالة إحصائياً عند مستوى 0.05، مما يدل على رفض فرضية تجانس التباين، وانعدام التجانس في تباين حد الخطأ العشوائي (البواقي) (Heteroscedasticity). وبالتالي، صُحح هذا القصور عبر استخدام أخطاء معيارية متنسقة (Heteroscedasticity–Consistent Standard Errors – HC1) بهدف الحصول على تقديرات أكثر دقة للثقة والمعنوية الإحصائية، كما أوصى بذلك (White, 1980).

3.3.10.3 تحقق الخطية (Linearity)

من خلال فحص مخطط البواقي مقابل القيم المتوقعة (Residuals vs. Fitted Values)، لم يظهر أي نمط غير خطي أو شكل منحنى واضح، ما يدل على تحقق افتراض العلاقة الخطية بين المتغيرات المستقلة والتابع.



3.3.10.4 التعدد الخطي (Multicollinearity)

تم حساب معامل تباين التضخم (Variance Inflation Factor – VIF) لجميع المتغيرات المستقلة، وذلك بهدف فحص وجود مشكلة التعدد الخطي. وفيما يلي ملخص نتائج هذا التحليل:

الجدول (12): قيم معامل تباين التضخم (VIF) للمتغيرات المستقلة لنموذج الانحدار الخطي المتعدد الخاص لبيانات مدينة دمشق.

المتغير	قيمة VIF
الاتجاه	1.01904
جودة الإكساء	1.03690
الارتفاع الطابقي	1.07687
المساحة	2.05674
وجود أثاث	1.01051

1.02342	وجود تدفئة
1.04445	وجود شرفة
2.03951	عدد الغرف
1.03304	وجود مصعد
1.01643	الموقع
1.04842	وجود مرآب
المصدر: إعداد الباحث اعتماداً على قاعدة بيانات مدينة دمشق باستخدام لغة البرمجة Python	

أظهرت النتائج أن جميع القيم كانت أقل من 5، ما يدل على عدم وجود تعدد خطي بين المتغيرات داخل النموذج.

3.3.10.5 اختبار الارتباط الذاتي للبواقي (Durbin-Watson)

بلغت قيمة اختبار Durbin-Watson نحو 2.03، وهي قريبة جداً من القيمة المرجعية 2، مما يشير إلى غياب الارتباط الذاتي بين البواقي (No Autocorrelation)، ويؤكد استقلال الأخطاء.

3.3.11 بناء النموذج الإحصائي للعلاقة بين متغيرات الدراسة الخاصة بمدينة دمشق باستخدام انحدار لاسو

نمذج الباحث العلاقة بين مبيع أسعار العقارات وبين المتغيرات المستقلة الخاصة بمدينة دمشق عن طريق انحدار لاسو، باستخدام الأسلوب المنهجي نفسه المعتمد سابقاً في تحليل بيانات مدينة Ames، حيث قُسمت البيانات إلى 80% للتدريب و20% للاختبار.

وفيما يلي الخطوات التي أُعتمد عليها:

• اختيار معلمة الضبط الخاصة بانحدار لاسو

اعتمد الباحث تطبيق نفس الأسلوبين السابقين لاختيار مجموعة من القيم التجريبية للمعلمة λ :

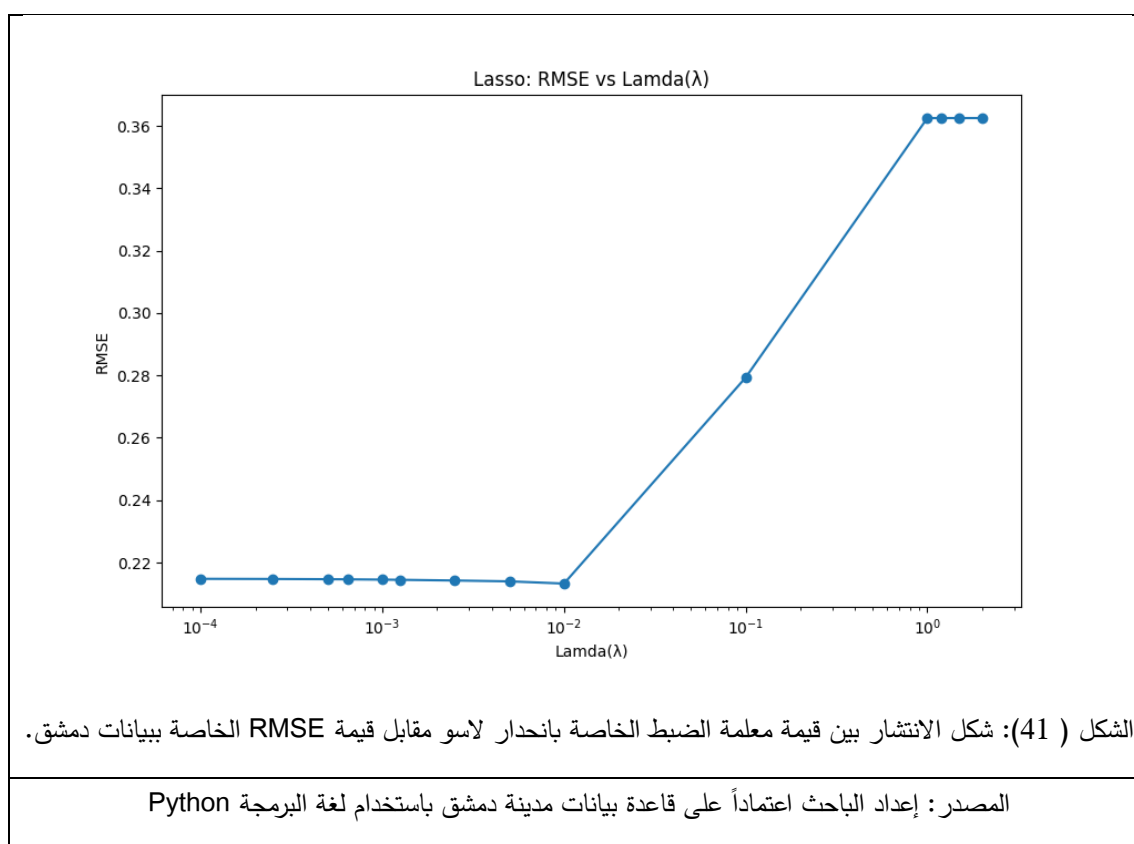
الأسلوب الأول: إنشاء متتالية حسابية لقيم λ تبدأ من 0.0001 وتزداد تدريجياً حتى تصل إلى 1، وذلك لتغطية نطاق واسع من القيم الصغيرة إلى الكبيرة

الأسلوب الثاني: جرى تحديد مجموعة مختارة من القيم (بلغ عددها 14 قيمة) استناداً إلى نتائج دراسات سابقة وتجارب تطبيقية منشورة.

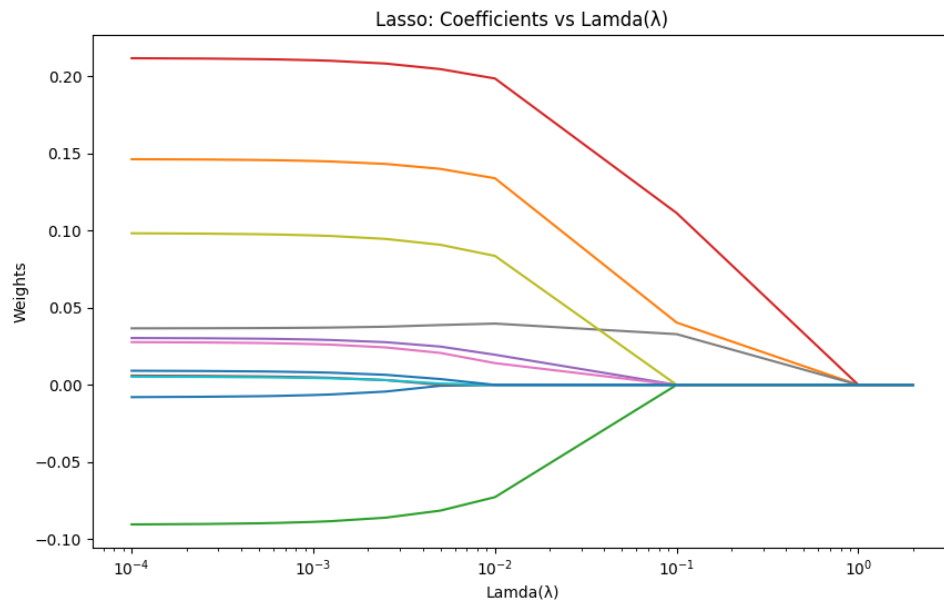
طُبِّق أسلوب التحقّق المتقاطع باستخدام طريقة K-Folds بعدد $K=4$ ، حيث قُسمت مجموعة بيانات التدريب إلى أربع مجموعات متساوية، تُستخدَم ثلاث منها في كل تكرار لأغراض التدريب، بينما تُخصّص المجموعة الرابعة للتحقّق. واعتمد التحليل على متوسط مربعات الأخطاء، كمؤشر لاختيار القيمة التي تُحقّق أفضل أداء تنبؤي.

وقد أسفر هذا التحليل عن النتائج الآتية:

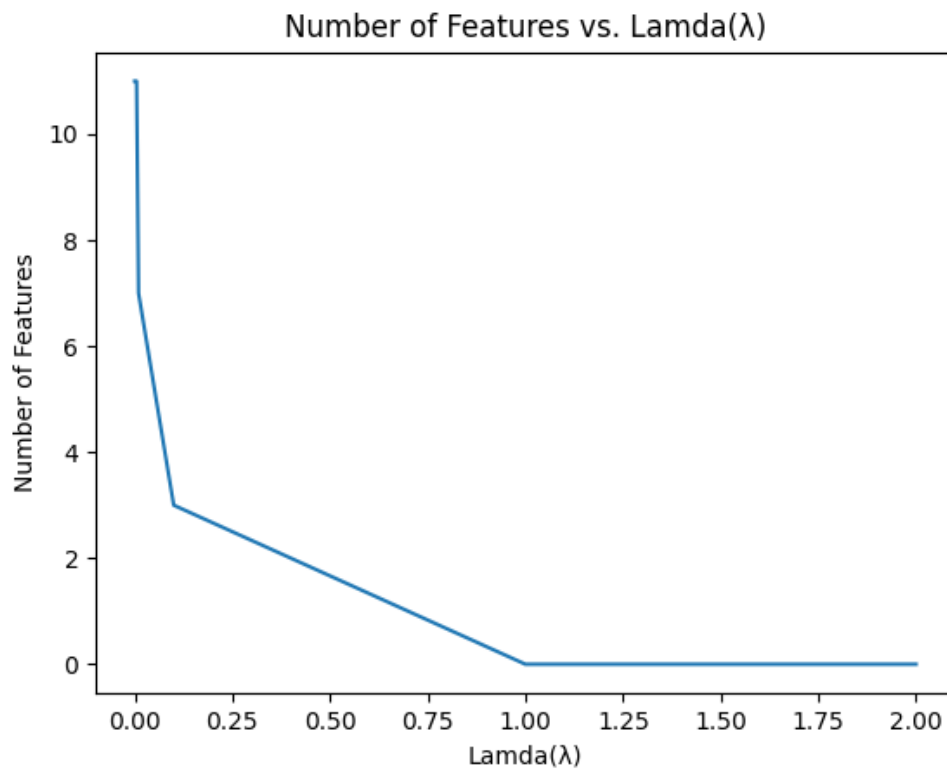
أُجري تقدير لـ 14 نموذجاً بناءً على القيم المختارة لمعلمة الضبط λ ، وتبيّن أن القيمة الأدنى لها بلغت نحو 0.010، وهي القيمة التي حقق عندها منحى جذر متوسط مربعات الأخطاء أدنى مستوياته، مسجلاً ما يقارب 0.2132. ويوضّح الرسم البياني الآتي سلوك النموذج عبر القيم المختلفة لـ λ .



بالإضافة إلى ما سبق، يوضّح الشكل البياني رقم (41) الدور المحوري الذي تؤديه معلمة الضبط λ في انحدار لاسو. وكما أُشير سابقاً، كلما ازدادت قيمة λ ، ازداد مقدار الجزاء المفروض على معاملات النموذج، مما يؤدي إلى تقليص عدد المتغيرات الداخلة في التقدير. أما عندما تكون قيمة λ مساوية للصفر، فإن جميع المتغيرات تُدرج في النموذج دون أي تقليص.



الشكل (42): منحنيات معاملات انحدار لاسو كدالة في معلمة الضبط λ الخاصة ببيانات مدينة دمشق.



الشكل (43): عدد المتغيرات الداخلة في النموذج كدالة في معلمة الضبط λ الخاصة ببيانات مدينة دمشق.

المصدر : إعداد الباحث اعتماداً على قاعدة بيانات مدينة دمشق باستخدام لغة البرمجة Python

يوضح الشكل (42) كيفية تغير معاملات انحدار لاسو مع اختلاف قيمة معلمة الضبط λ باستخدام بيانات مدينة دمشق، حيث يُلاحظ أن زيادة λ تؤدي تدريجياً إلى انكماش المعاملات نحو الصفر، حتى تختفي بعض المتغيرات تماماً عند تجاوز قيم معينة. أما الشكل (43)، فيعرض العلاقة بين λ وعدد المتغيرات التي يحتفظ بها النموذج بعد التقدير، حيث يتبين أن ازدياد قيمة λ يؤدي إلى انخفاض حاد في عدد المتغيرات المختارة، وصولاً إلى حالة شبه معدومة عند القيم الكبيرة جداً لـ λ .

• تقدير معاملات النموذج على بيانات التدريب والاختبار

بعد تحديد قيمة معلمة الضبط λ الخاصة بانحدار لاسو، والتي بلغت 0.010، جرى تقدير معاملات النموذج باستخدام مجموعة التدريب، ثم اختُبر أدائه على مجموعة الاختبار. وفيما يلي تُعرض النتائج التي أُسفر عنها هذا التقدير:

الجدول (13): نتائج تقدير انحدار لاسو على بيانات التدريب والاختبار على مدينة دمشق.		
المؤشر	مجموعة البيانات	القيمة
RMSE	بيانات التدريب	0.1983
	بيانات الاختبار	0.1976
عدد المتغيرات المتبقية الداخلة في النموذج		7
المصدر: إعداد الباحث اعتماداً على قاعدة بيانات مدينة دمشق باستخدام لغة البرمجة Python		

تشير قيمة جذر متوسط مربعات الأخطاء المنخفضة لكل من بيانات التدريب (0.1983) وبيانات الاختبار (0.1976) إلى أن انحدار لاسو يتمتع بقدرة جيدة على التنبؤ بأسعار العقارات السكنية في مدينة دمشق، وذلك بالاعتماد على مجموعة محدودة من المتغيرات التفسيرية. وتجدر الإشارة إلى أن هذه القيم محسوبة بعد تطبيق التحويل اللوغاريتمي الطبيعي على متغير السعر، وعند إعادة تفسيرها على المقياس الأصلي باستخدام العلاقة $1 - \exp(\text{RMSE})$ ، يتبين أن متوسط الانحراف النسبي للقيم المقدرة يبلغ نحو 22%. وتشير هذه النتيجة إلى أن النموذج يحقق أداءً جيداً، بهامش خطأ نسبي مقبول بالنظر إلى طبيعة السوق العقاري المدروس، والقيود المفروضة على البيانات المتوافرة.

وقد احتفظ النموذج بسبعة متغيرات فقط من أصل أحد عشر متغيراً، ما يعكس كفاءته في تبسيط العلاقة الإحصائية، والتركيز على المتغيرات الأكثر فاعلية. ومن أبرز مزايا انحدار لاسو أنه لا يكتفي بتقدير المعاملات، بل يقوم أيضاً

بتحديد المتغيرات الأكثر أهمية عبر تفسير المعاملات غير المؤثرة، ما يقلل من تعقيد النموذج ويحد من مخاطر التعدد الخطي.

وفقاً للنتائج النهائية للنموذج، تم الاحتفاظ بالمتغيرات الآتية: المساحة، جودة الإكساء، وجود الأثاث، وجود الشرفة، عدد الغرف، وجود المصعد، والارتفاع الطابقي. وفي المقابل، تم استبعاد متغيرات نظام التدفئة، والموقع، ووجود مرآب، وعدد الاتجاهات (الواجهات)، وذلك لانخفاض أثرها التنبؤي ضمن العينة المستخدمة.

ويُشار إلى أن قيمة جذر متوسط مربعات الأخطاء والتي بلغت 0.1983 لعينة التدريب و0.1976 لعينة الاختبار، تتدرج ضمن نطاق القيم المحسوبة بعد تطبيق التحويل اللوغاريتمي الطبيعي على متغير السعر. وعند إعادة تفسير هذه القيم في سياقها الأصلي.

• الشكل النهائي للنموذج المقدر.

بناءً على نتائج انحدار لاسو وبعد تحديد القيمة المثلى لمعامل الضبط (λ) التي بلغت 0.010، تحدد الشكل الرياضي للنموذج، الذي يربط بين السعر (بعد تحويله إلى اللوغاريتم الطبيعي) والمتغيرات المستقلة المختارة، وذلك وفق العلاقة الآتية:

$$\log(\text{Price}) = \text{Area} * 0.198 + \text{Quality} * 0.134 + \text{Elevator} 0.0836 - \text{Floor} * 0.072 + \text{Nrooms} * 0.039 + \text{Furniture} * 0.0195 + \text{Belcony} * 0.014$$

حيث:

- جرت عملية توحيد لجميع المتغيرات باستخدام Z-Score قبل التقدير.
- المتغير التابع هو اللوغاريتم الطبيعي للسعر ($\ln(\text{Price})$).
- المتغيرات المستقلة المشار إليها أعلاه هي الوحيدة التي احتفظ بها انحدار لاسو.

3.3.12 التقييم التطبيقي لأداء النموذج التنبؤي باستخدام عينة من العقارات

بهدف تقييم دقة النموذج بشكل تطبيقي، تم اختبار قدرة انحدار لاسو على تقديم تقديرات قريبة من الأسعار المعلنة، من خلال مقارنة القيم المقدرة من النموذج مع القيم الحقيقية، وذلك لعشر مشاهدات تم اختيارها لا على التعيين من قاعدة البيانات. ويعرض الجدول الآتي نتائج هذه المقارنة، والتي تُبرز الفروقات بين السعر المُعلن والسعر المُقدّر لكل عقار:

الجدول (14): مقارنة بين السعر المُعلن والسعر المُقدَّر من انحدار لاسو الخاص ببيانات مدينة دمشق.			
العقار	السعر المُعلن (ل.س.)	السعر المُقدَّر (ل.س.)	الفرق (ل.س.)
عقار 1	835,000,000	831,991,900	(3,008,100)
عقار 2	945,000,000	942,869,354	(2,130,646)
عقار 3	1,150,000,000	1,149,089,820	(910,180)
عقار 4	1,300,000,000	1,299,976,482	(23,518)
عقار 5	580,000,000	579,990,665	(9,335)
عقار 6	715,000,000	718,505,278	3,505,278
عقار 7	1,650,000,000	1,654,333,783	4,333,783
عقار 8	1,300,000,000	1,304,773,134	4,773,134
عقار 9	1,050,000,000	1,055,759,023	5,759,023
عقار 10	590,000,000	595,855,514	5,855,514
المصدر: إعداد الباحث اعتماداً على قاعدة بيانات مدينة دمشق باستخدام لغة البرمجة Python			

تشير النتائج إلى أن النموذج قدّم تقديرات قريبة من الأسعار المُعلنة في معظم الحالات، حيث تراوحت الفروقات في 8 من أصل 10 حالات بين ± 6 ملايين ليرة سورية، وهي فروقات تُعد مقبولة نسبياً. فعلى سبيل المثال، بلغ الفرق في عقار رقم 4 نحو 23 ألف ليرة فقط، وهو فارق يُعد رمزياً. كما حقق النموذج دقة عالية في حالات أخرى مثل عقار 5، حيث كان الفارق أقل من 10 آلاف ليرة.

وفي المقابل، لم تُسجل سوى حالتين تجاوز فيهما الفرق حدود 3.5 مليون ليرة سورية، حيث بلغ في الحالة السادسة نحو 3.5 مليون ليرة سورية، وفي الحالة العاشرة قرابة 5.8 مليون ليرة. ورغم ذلك، تبقى هذه الفروقات ضمن هامش مقبول نسبياً في ضوء التباين الطبيعي لأسعار العقارات.

3.3.13 المقارنة التطبيقية بين أداء انحدار لاسو والانحدار الخطي المتعدد

تمت مقارنة الانحدار الخطي المتعدد بانحدار لاسو، من خلال تطبيق كلا النموذجين على بيانات سوق العقارات في مدينة دمشق، بعد توحيد أسلوب التحليل الإحصائي القائم على تقسيم البيانات إلى عينة تدريب بنسبة 80% وعينة اختبار بنسبة 20%، وتطبيق المعايير (Standardization) لضمان نتائج قابلة للمقارنة.

حقق نموذج OLS معامل تحديد (R^2) بلغ حوالي 0.65 على بيانات الاختبار، بينما بلغ RMSE على بيانات الاختبار 0.2006، ما يشير إلى أداء مقبول في تفسير التباين وتوقع الأسعار. ومع ذلك، احتفظ النموذج بجميع

المتغيرات التفسيرية تقريباً، بما في ذلك المتغيرات غير المؤثرة إحصائياً ($p > 0.05$). كما أظهرت اختبارات التحقق من الفرضيات بعض التحديات، بما في ذلك احتمالية وجود مشكلات في تجانس التباين والتوزيع الطبيعي للبواقي.

في المقابل، بلغ معامل التحديد R^2 على بيانات الاختبار في نموذج انحدار لاسو حوالي 0.660، و $RMSE \approx 0.1976$ ، وهي قيمة قريبة من OLS لكن بنموذج أكثر تبسيطاً، إذ احتفظ فقط بـ 7 متغيرات ذات تأثير فعلي، ما يعزز من سهولة تفسير النموذج ويقلل من مخاطر الإفراط في التوفيق (overfitting). تُظهر هذه النتيجة أهمية انحدار لاسو في تقليص المتغيرات تلقائياً، وتحسين القدرة التنبؤية في الحالات التي تتعدد فيها المتغيرات وتتقاطع علاقاتها.

بناءً على هذه النتائج، يمكن القول إن انحدار لاسو يوفر بديلاً فعالاً وأكثر مرونة من OLS، في السياقات التي تهدف إلى التنبؤ الدقيق وتقليل التعقيد، بينما يظل OLS مفيداً لأغراض التفسير الإحصائي الشامل، بشرط التحقق من سلامة فروضه الأساسية.

مناقشة فرضيات الدراسة

انطلقت هذه الدراسة من فرضيتين رئيسيتين عُرضتا في الفصل الأول، وقد سعى التحليل التطبيقي، سواء على بيانات مدينة Ames أم مدينة دمشق، إلى التحقق من مدى صحتها. وفيما يلي استعراض لمدى تحقق كل فرضية بناءً على النتائج المستخلصة:

1. يمكن من خلال استخدام الطرائق الإحصائية التوصل إلى نموذج يساعد في تقدير أسعار العقارات.
2. تسمح الطرائق الإحصائية بإعطاء نتائج جيدة وموضوعية أكثر من الأساليب التقليدية لتقدير أسعار العقارات.

الفرضية الأولى: يمكن من خلال استخدام الطرائق الإحصائية التوصل إلى نموذج يساعد في تقدير أسعار العقارات.

أظهرت نتائج الدراسة تحقق هذه الفرضية بشكل واضح، حيث تبين من خلال استخدام الانحدار وخاصة انحدار لاسو إمكانية بناء نماذج إحصائية قادرة على تقدير العلاقة بين السعر وبين عدد من الخصائص المميزة للعقار. إذ بلغت قدرة التفسير في نموذج OLS بمدينة دمشق حوالي 70%، في حين بلغت قدرة النموذج التنبؤية باستخدام لاسو نحو 66%، وهو ما يشير إلى أن النماذج الإحصائية قدّمت تمثيلاً جيداً للعلاقة بين المتغيرات.

الفرضية الثانية: تسمح الطرائق الإحصائية بإعطاء نتائج جيدة وموضوعية أكثر من الأساليب التقليدية لتقدير أسعار العقارات.

ناقشت الدراسة عدداً من الأساليب التقليدية المعتمدة في التقييم العقاري، مثل أسلوب المقارنة بالمبيعات وأسلوب التكلفة، وبيّنت ما تعانيه من محدودية ناجمة عن نقص البيانات، وتباين التقديرات البشرية. وعلى الرغم من أن الدراسة لم تُجرِ مقارنة كمية مباشرة بين هذه الأساليب والنماذج الإحصائية، إلا أن النتائج المستخلصة من تحليل نموذجي الانحدار (OLS و LASSO)، أظهرت قدرة كبيرة على تفسير التباين في الأسعار، تستند إلى البيانات. وبناءً عليه، يمكن تأكيد تحقق فرضيتي الدراسة، الأمر الذي يدعم المنهجية المعتمدة، ويعزز من موثوقية النماذج المستخدمة في تقدير الأسعار العقارية.

النتائج والتوصيات

أولاً نتائج الدراسة

أسفرت الدراسة عن عدد من النتائج المهمة، وذلك بعد تطبيق انحدار لاسو على بيانات مدينتين مختلفتين، من حيث الخصائص العقارية والتخطيط العمراني، هما: مدينة Ames الأمريكية، ومدينة دمشق السورية. وقد مكن هذا النهج المقارن من اختبار مدى كفاءة النموذج في بيئتين عقاريّتين مختلفتين، وتحديد المتغيرات الأكثر تأثيراً في أسعار العقارات في كل حالة.

أ. نتائج مدينة Ames

1. حقق انحدار لاسو نتائج تمثلت في الاحتفاظ بـ 114 متغيراً من أصل 240 متغيراً، مع تسجيل أدنى قيمة لجذر متوسط مربعات الخطأ $0.113 \approx (RMSE)$ ، ما يعكس كفاءته في التقدير.
2. أبرز المتغيرات التي أثّرت في السعر العقاري في مدينة Ames شملت: متغير مساحة المعيشة فوق الأرض (المساحة الداخلية للمسكن الفعلي) GrLivArea، والتقييم الإجمالي للعقار (للإكساء) OverallQual، وعام البناء YearBuilt، والتقييم الإجمالي لحالة للعقار OverallCond، و المساحة الأساسية المُجهّزة في القبو (BsmtFinSF1)، ومساحة القبو TotalBsmtSF، وعدد المواقف المتاحة في المرآب GarageCars، ومساحة أرض العقار LotArea، وموقع العقار ضمن حي Crawford في مدينة Ames، و تقييم جودة المواد المستخدمة في الواجهة الخارجية ExterQual.
3. أظهر النموذج قدرة فعالة على معالجة التعددية الخطية، من خلال حذف المتغيرات ذات الارتباط العالي فيما بينها، ما ساهم في تبسيط النموذج للعلاقة بين السعر العقاري وخصائص العقارات.

ب. نتائج مدينة دمشق

1. رغم محدودية البيانات، تمكن النموذج من اختيار 7 متغيرات فقط من أصل 11 متغيراً، مع تحقيق قيمة $RMSE \approx 0.1976$ ، وذلك من خلال انحدار لاسو.
2. المتغيرات الأهم التي أثّرت في السعر في دمشق شملت: المساحة (Area)، جودة الإكساء (Quality)، عدد الغرف (Number of Rooms)، وجود مصعد أو مولدة (Elevator/Generator)، الارتفاع الطابقي (Floor)، توافر الأثاث (Furniture)، توافر الشرفة (Balcony).
3. أثبت النموذج فعاليته في تقديم تفسير كمي للعوامل المؤثرة في السعر العقاري في دمشق، رغم غياب قاعدة بيانات معيارية، ما يُظهر إمكانية تطبيقه على نطاق أوسع مستقبلاً، عند توافر بيانات منظمة وأكثر شمولاً.

4. وبذلك، تحققت أهداف الدراسة في اختبار جدوى انحدار لاسو في تفسير العلاقة بين السعر وخصائص العقارات، وتقديم أداة تحليلية فعالة يمكن استخدامها في بيئات ذات خصائص متعددة أو بيانات محدودة، مثل السوق العقاري الدمشقي.

ثانياً التوصيات

استناداً إلى نتائج الدراسة وتحقيقاً لأهدافها، يمكن تقديم التوصيات الآتية:

1. تشجيع تطوير قواعد بيانات عقارية معيارية في المدن السورية، تتضمن خصائص تفصيلية وموحدة للعقارات (المساحة، الجودة، الطابق، التجهيزات... إلخ)، بما يُمكن الباحثين وصانعي القرار من إجراء تحليلات كمية دقيقة، وتقدير القيمة السوقية للعقارات بطريقة علمية.
2. استخدام نتائج الدراسة كأساس لتطوير نظام دعم قرار لصالح جهات حكومية أو خاصة (مثل مديريات الأملاك، مكاتب العقارات، الجهات الضريبية)، بهدف توفير مؤشرات مرجعية مبنية على خصائص العقار، بدلاً من الاعتماد فقط على التقديرات الشخصية أو الخبرة السوقية.
3. توسيع نطاق الدراسة مستقبلاً ليشمل مدناً سورية أخرى مثل حلب أو اللاذقية، أو إجراء مقارنة بين أحياء مختلفة ضمن المدينة نفسها، ما يُعزّز فهم تأثير العوامل المحلية الدقيقة (مثل البنية التحتية والخدمات) في أسعار العقارات.
4. الاستفادة من نماذج الانحدار الإحصائي، ولاسيما انحدار لاسو، في تحليل الأثر السعري لمشاريع الإسكان أو التطوير العقاري، من خلال إدراج متغيرات تمثل هذه المشاريع ضمن النماذج، بما يتيح قياس مدى تأثيرها الفعلي على أسعار العقارات ودعم تقييم السياسات العقارية على أسس كمية.
5. يُوصى باعتماد انحدار لاسو كأداة تحليلية في تقييم أسعار العقارات وكذلك أسعار السلع المعمّرة، سواء في البيئات التي تتعدد فيها المتغيرات أم التي تفتقر إلى بيانات منظمة، نظراً لقدرة هذا النموذج على تبسيط العلاقات المعقدة واختيار المتغيرات الأكثر تأثيراً. كما يُنصح مستقبلاً بتطبيقه على بيانات عقارية في مدينة دمشق أو غيرها من المدن السورية، عند توافر قواعد بيانات معيارية أكثر تفصيلاً، بما يساهم في تعزيز دقة التقدير ودعم صناعة القرار العقاري بمخرجات كمية موثوقة.

ثالثاً الخلاصة

سعت هذه الدراسة إلى تقديم نموذج إحصائي مقترح يُستخدم في تقدير القيمة الحقيقية لأسعار العقارات في مدينة دمشق، في ظل غياب قواعد بيانات معيارية وموثوقة يمكن الرجوع إليها في التقييم العقاري. ولغرض التحقق من كفاءة النموذج الإحصائي المقترح، طُبّق أولاً انحدار لاسو على قاعدة بيانات مدينة Ames الأمريكية، التي تُعد

مرجعية في الأدبيات العالمية، نظراً لغناها وتنوع خصائصها العقارية. وقد أظهر النموذج فعالية كبيرة في انتقاء المتغيرات المؤثرة والتقليل من التعددية الخطية، مع المحافظة على دقة تنبؤية مرتفعة، وهو ما يؤكد ملاءمته في البيانات التي تتعدد فيها الخصائص وتتشعب العوامل.

وبناءً على ثبوت فعالية هذا الأسلوب، جرى الانتقال إلى التطبيق الميداني في مدينة دمشق، على الرغم من غياب قاعدة بيانات رسمية. ولتحقيق ذلك، قام الباحث بجمع بيانات ميدانية من السوق المحلي، وجرى توثيق مجموعة من الخصائص المتعلقة بالموقع، والمساحة، والجودة، والارتفاع الطائقي، والتجهيزات، وغيرها من السمات العقارية.

استُخدم في هذا التحليل نموذجان إحصائيان: الانحدار الخطي المتعدد (OLS) وانحدار لاسو (LASSO)، وذلك بالاستناد إلى البيانات الميدانية المُجمّعة. وقد أظهرت النتائج أنّ كلا النموذجين قادر على تفسير التفاوت في أسعار العقارات بدرجة مقبولة، مع تسجيل أفضلية لانحدار لاسو في هذه الحالة. ويبرز هذا التفوق خصوصاً في البيانات التي تتسم بقلتها وتعدد المتغيرات، حيث يسهم انحدار لاسو في تبسيط النموذج .

إن ما توصّلت إليه الدراسة يُثبت إمكانية بناء نموذج تقديري لسعر العقار في مدينة دمشق، باستخدام أدوات التحليل الإحصائي، حتى في ظل غياب قاعدة بيانات مركزية. كما يُظهر أنّ انحدار لاسو سيكون مناسباً بدرجة أكبر حين تتوافر قاعدة بيانات أكثر شمولاً وتنوعاً، إذ يتيح تحديد العوامل الأكثر تأثيراً على السعر العقاري من بين عدد كبير من الخصائص.

بناءً على ما سبق، يمكن القول إن هذه الدراسة تمثل خطوة أولى نحو توظيف أدوات التحليل الكمي الحديث في السوق العقاري المحلي، وتُمهد الطريق أمام تطوير نماذج أكثر دقة وشمولاً مستقبلاً، شرط توافر بيانات غنية وموثوقة، تغطي مختلف أحياء مدينة دمشق وخصائص عقاراتها. كما تبرز إمكانية توسيع تطبيق انحدار لاسو ليشمل مجالات أخرى.

المراجع

المراجع باللغة العربية:

1. حافظ، ر. م. (2019). هيكلية القطاع العقاري في مصر. مجلة بحوث العمران - العدد 32، أبريل، 6-5. Retrieved from https://jur.journals.ekb.eg/article_90822_800ced83fe3209ca7f87bdd2bda61920.pdf
2. سيف علي، 2019، "تقدير معلمات نموذجي انحدار لاسو وانحدار الحرف مع تطبيق عملي"، جامعة كربلاء
3. منى حمادة، محمد نايفة، 2017، "نموذج رياضي مبني على تحليل الانحدار المتعدد لتقييم العقارات السكنية في دمشق"، جامعة البعث.

المراجع باللغة الإنجليزية:

1. Aggarwal, C. C. (2017). Outlier Analysis. Springer.
2. Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), Second International Symposium on Information Theory, Akademiai Kiado, Budapest, 267-281.
3. Bouke, M. A., & A. A. (2024). An empirical assessment of ML models for 5G network intrusion detection: A data leakage-free approach. e-Prime – Advances in Electrical Engineering, Electronics and Energy, 2.
4. Bulut, B. A. (2011). A Residential Real-Estate Valuation Model with Reduced Attributes. International Journal of Mathematical Models and Methods in Applied Sciences, 5(3), 586-593.
5. Burnham, K. P., & Anderson, D. (2002). Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. Springer.

6. Chand, S. (2012). On tuning parameter selection of LASSO-type methods – A Monte Carlo study. In Proceedings of the 9th International Bhurban Conference on Applied Sciences & Technology (IBCAST 2012), 120–129.
7. Chih–Jen Hsieh, K.–W. C.–J. (2008). A dual coordinate descent method for large–scale linear SVM. 2. Retrieved from <https://icml.cc/Conferences/2008/papers/166.pdf>
8. Daniel T. Larose, C. D. (2014). *Discovering Knowledge in Data: An Introduction to Data Mining*. Wiley.
9. De Cock, D. (2011). Ames, Iowa: Alternative to the Boston housing data as an end of semester regression project. *Journal of Statistics Education*, Volume 19, Issue 3.
10. Doszyń, M. (2024). Might expert knowledge improve econometric real estate mass appraisal? *Journal of Real Estate Finance and Economics*, 719–740.
11. Everitt, B., & Skrondal, A. (2010). *The Cambridge Dictionary of Statistics*. Cambridge University Press.
12. Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics*, 4th Edition. Sage Publications.
13. Gujarati, D. N., & Porter, D. C. (2008). *Basic Econometrics* (5th edition). McGraw–Hill Education.
14. Hansheng Wang & Bo Li & Chenlei Leng. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Volume 71, Issue 3, 671–683.
15. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.

- 16.Hawkins, D. M. (2024). The problem of overfitting. *Journal of Chemical Information and Modeling*, 1–12.
- 17.Hui Zou, T. H. (2007). On the "degrees of freedom" of the LASSO. *The Annals of Statistics*, Volume 35, Issue 5.
- 18.Ian Goodfellow, Y. B. (2016). *Deep Learning*. MIT Press.
- 19.James, G., Witten, D., Hastie, T., & Robert , T. (2021). *An Introduction to Statistical Learning (2nd edition) With Applications in R*. Springer.
- 20.Jerome Friedman, T. H. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 1–22.
- 21.Kohavi, R. (1995). A Study of Cross–Validation and Bootstrap for Accuracy Estimation and Model Selection. *IJCAI*.
- 22.Kumar, A., Attri, S. D., & Singh, D. (2019). Comparison of lasso and stepwise regression technique for wheat yield prediction. *India Meteorological Department, New Delhi*
- 23.Liu, Z. (2023). Real estate price prediction based on supervised machine learning scenarios. *Highlights in Science, Engineering and Technology, CMLAI 2023*
- 24.Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis (5th edition)*. John Wiley & Sons.
- 25.Pedregosa, F. V. (2011). Scikit–learn: Machine learning in Python. *Journal of Machine Learning Research*, 2825–2830.
- 26.Seng, J. X., & Khalid, K. (2018). Modelling house price using ridge regression and lasso regression. *International Journal of Engineering & Technology*

27. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 267–288.
28. Tibshirani, R. J. (2012). The Lasso Problem and Uniqueness. *arXiv*, 25.
29. Tibshirani, R., Hastie, T., Witten, D., & James, G. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman and Hall/CRC.
30. Tibshirani, R., et al. (2020). Best subset, forward stepwise or lasso? Analysis and recommendations based on extensive comparisons. *Statistical Science*, 35(4), Institute of Mathematical Statistics
31. Tibshirani, R. J. (2021). Cross-validation: what does it estimate and how well does it do it. *arXiv preprint arXiv:2104.00673*, 1–3.
32. Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 480. Retrieved from https://www.mit.edu/~dimitrib/PTseng/papers/archive/bcr_jota.pdf
33. van Buuren, S. (2018). *Flexible Imputation of Missing Data*. CRC Press.
34. Vandenberghe, S. P. (2011). *Proximal Algorithms*. Stanford University. Retrieved from https://web.stanford.edu/~boyd/papers/pdf/prox_algs.pdf
35. White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica*, 817–838.
36. Wooldridge, J. M. (2016). *Introductory Econometrics: A Modern Approach*, 5th Edition. Cengage Learning. Retrieved from <https://jaimeadv.com/eco/4c1-economet/jeffrey-m-wooldridge--econometrics--book.pdf>

1. <https://machinelearningmastery.com/data-leakage-machine-learning/>
2. <https://www.kaggle.com/wiki/Leakage>
3. <https://stats.stackexchange.com/questions/45643/why-l1-norm-for-sparse-models/45644#45644>
4. <https://stats.stackexchange.com/questions/17781/derivation-of-closed-form-lasso-solution/645265#645265>
5. <https://www.stat.cmu.edu/%7Eryantibs/convexopt/lectures/prox-grad.pdf>
6. <https://math.stackexchange.com/questions/471339/derivation-of-soft-thresholding-operator-proximal-operator-of-l-1-norm>
7. <https://www.chioka.in/differences-between-l1-and-l2-as-loss-function-and-regularization/>
8. <https://medium.com/data-science/regularization-in-machine-learning-connecting-the-dots-c6e030bfadd>
9. <https://stats.stackexchange.com/questions/74542/why-does-the-lasso-provide-variable-selection>
10. <https://stats.stackexchange.com/questions/17251/what-is-the-lasso-in-regression-analysis/17254#17254>
11. <http://scott.fortmann-roe.com/docs/BiasVariance.html>
12. <https://pandas.pydata.org/>
13. <https://numpy.org/>
14. <https://scikit-learn.org/stable/>
15. <https://seaborn.pydata.org/>
16. <https://matplotlib.org/>
17. <https://www.statsmodels.org>
18. <https://docs.scipy.org/doc/scipy>

الملاحق

الملحق (1): المعاملات المعيارية الناتجة عن انحدار لاسو الخاص ببيانات مدينة Ames.

$$\begin{aligned}
\hat{y} = & 0.1280 \times GrLivArea + 0.0786 \times OverallQual + 0.0633 \times YearBuilt \\
& + 0.0450 \times OverallCond - 0.0322 \times MSZoning_A (agr) \\
& + 0.0288 \times BsmtFinSF1 + 0.0261 \times TotalBsmtSF \\
& + 0.0234 \times GarageCars + 0.0212 \times LotArea \\
& - 0.0196 \times SaleCondition_Abnorml \\
& + 0.0160 \times Neighborhood_Crawfor - 0.0155 \times MSZoning_RM \\
& + 0.0155 \times ExterQual - 0.0140 \times CentralAir_N \\
& + 0.0134 \times BsmtQual + 0.0126 \times HeatingQC \\
& + 0.0123 \times KitchenQual + 0.0121 \times Neighborhood_NridgHt \\
& + 0.0117 \times ScreenPorch - 0.0113 \times MSZoning_C (all \\
& - 0.0112 \times Neighborhood_MeadowV \\
& + 0.0110 \times Neighborhood_BrkSide - 0.0105 \times MSSubClass_30 \\
& + 0.0105 \times SaleType_New + 0.0104 \times BsmtExposure \\
& + 0.0103 \times 1stFlrSF + 0.0103 \times Exterior1st_BrkFace \\
& + 0.0103 \times Neighborhood_GrnHill + 0.0103 \times Foundation_PConc \\
& - 0.0102 \times Exterior1st_AsbShng + 0.0097 \times Exterior1st_PreCast \\
& + 0.0089 \times Fireplaces + 0.0087 \times BsmtFullBath \\
& + 0.0082 \times Functional_Typ + 0.0079 \times LotFrontage \\
& + 0.0078 \times EnclosedPorch + 0.0077 \times BsmtFinType1 \\
& + 0.0077 \times GarageArea + 0.0072 \times Neighborhood_StoneBr \\
& - 0.0068 \times MSSubClass_160 + 0.0068 \times BldgType_1Fam \\
& + 0.0067 \times Neighborhood_Somerst + 0.0067 \times Condition1_Norm \\
& - 0.0065 \times Condition1_Artery - 0.0062 \times Functional_Sev \\
& - 0.0061 \times Neighborhood_SawyerW \\
& + 0.0056 \times MasVnrType_Stone + 0.0055 \times SaleCondition_AdjLand \\
& - 0.0052 \times Neighborhood_NWAmes - 0.0052 \times Functional_Mod \\
& + 0.0052 \times MSZoning_FV + 0.0051 \times PavedDrive_Y \\
& + 0.0047 \times FullBath + 0.0047 \times OpenPorchSF \\
& - 0.0046 \times SaleCondition_Family - 0.0044 \times Functional_Maj2 \\
& + 0.0043 \times GarageCond - 0.0039 \times YrSold \\
& + 0.0039 \times SaleType_Con - 0.0037 \times Neighborhood_Edwards \\
& - 0.0037 \times BedroomAbvGr - 0.0037 \times BldgType_Twnhs \\
& + 0.0036 \times WoodDeckSF + 0.0034 \times MasVnrArea \\
& - 0.0033 \times Exterior1st_HdBoard - 0.0033 \times Condition1_RRAe \\
& - 0.0033 \times MSSubClass_90 - 0.0032 \times Neighborhood_Blmngtn \\
& + 0.0030 \times LandContour_HLS - 0.0030 \times Alley_None \\
& + 0.0030 \times MSSubClass_70 - 0.0028 \times PoolQC_None \\
& - 0.0026 \times Exterior2nd_AsbShng + 0.0026 \times Condition1_PosN \\
& + 0.0024 \times Electrical_SBrkr + 0.0023 \times FireplaceQu \\
& - 0.0022 \times MSZoning_I (all) - 0.0021 \times HouseStyle_2.5Fin \\
& + 0.0021 \times Exterior2nd_CBlock - 0.0019 \times Neighborhood_OldTow \\
& + 0.0018 \times Neighborhood_NoRidge + 0.0018 \times Condition1_PosA
\end{aligned}$$

$$\begin{aligned}
& + 0.0017 \times \text{BsmtFinSF2} - 0.0017 \times \text{SaleType_Oth} \\
& + 0.0016 \times \text{Neighborhood_Greens} - 0.0015 \times \text{Foundation_BrkTil} \\
& - 0.0015 \times \text{Condition1_RRNn} - 0.0015 \times \text{BldgType_Duplex} \\
& - 0.0014 \times \text{Exterior2nd_Plywood} - 0.0013 \times \text{LotConfig_FR3} \\
& - 0.0013 \times \text{Electrical_FuseF} - 0.0012 \times \text{Fence_GdWo} \\
& + 0.0012 \times \text{HalfBath} - 0.0009 \times \text{LotShape_IR3} \\
& + 0.0009 \times \text{ExterCond} - 0.0007 \times \text{Exterior1st_Stone} \\
& - 0.0007 \times \text{Exterior2nd_Wd Shng} + 0.0006 \times \text{MSSubClass_75} \\
& - 0.0005 \times \text{MasVnrType_BrkCmn} - 0.0005 \times \text{HouseStyle_1.5Unf} \\
& - 0.0003 \times \text{LandContour_Low} + 0.0003 \times \text{SaleType_CWD} \\
& - 0.0003 \times \text{Functional_Sal} - 0.0003 \times \text{GarageType_CarPort} \\
& + 0.0002 \times \text{Neighborhood_ClearCr} \\
& + 0.0002 \times \text{Neighborhood_Blueste} - 0.0002 \times \text{MSSubClass_180} \\
& - 0.0002 \times \text{Exterior1st_BrkComm} + 0.0002 \times \text{LotConfig_CulDSac} \\
& - 0.0001 \times \text{Neighborhood_NAMES} \\
& + 0.0000 \times \text{SaleCondition_Alloca} \\
& + 0.0000 \times \text{Neighborhood_NPkVill} \\
& + 0.0000 \times \text{Exterior2nd_PreCast} + 0.0000 \times \text{CentralAir_Y}
\end{aligned}$$

الملحق (2): قائمة قيم λ المستخدمة في النموذج وفق الأسلوب الثاني

$$\lambda = [0.0001, 0.00025, 0.0005, 0.00065, 0.001, 0.00125, 0.0025, 0.005, 0.01, 0.1, 1, 1.2, 1.5, 2]$$

الملحق (3): كود Python المستخدم لتقدير انحدار لاسو

– مدينة Ames

```

- import pandas as pd
- import numpy as np
- import matplotlib.pyplot as plt
- import seaborn as sns
- import matplotlib
- from scipy.stats import boxcox
- from sklearn.metrics import r2_score
- from sklearn.metrics import mean_squared_error, make_scorer

```

```

- from sklearn.linear_model import Ridge, Lasso
- from scipy.special import inv_boxcox
- from sklearn.model_selection import GridSearchCV, KFold,
  train_test_split
- from sklearn.preprocessing import StandardScaler
- from sklearn.utils import resample
- import warnings
- warnings.filterwarnings("ignore")
-
- # =====
- # Load data
- # =====
- train = pd.read_csv("train.csv")
- test = pd.read_csv("test.csv")
- X = train.copy()
-
- X['train'] = 1
- test['train'] = 0
-
- # Combine train & test
- df = pd.concat([test, X], axis=0, ignore_index=True)
-
- # Quick missing overview
- missing = pd.DataFrame({
-     'types': df.dtypes,
-     'percentange_of_missing': df.isna().sum() / len(df) * 100
- })
- missing = missing[missing['percentange_of_missing'] != 0]
- print(missing.sort_values(by='percentange_of_missing',
-     ascending=False))
-
- # =====
- # Missing values handling
- # =====
- df['PoolQC'] = df['PoolQC'].fillna('None')
- df['MiscFeature'] = df['MiscFeature'].fillna('None')
- df['Alley'] = df['Alley'].fillna('None')
- df['Fence'] = df['Fence'].fillna('None')
- df['FireplaceQu'] = df['FireplaceQu'].fillna('None')
- train_mask = (df['train'] == 1)
- lf_med =
- df.loc[train_mask].groupby('Neighborhood')['LotFrontage'].median()
- df['LotFrontage'] =
- df['LotFrontage'].fillna(df['Neighborhood'].map(lf_med))
- df['LotFrontage'] = df['LotFrontage'].fillna(df.loc[train_mask,
-     'LotFrontage'].median())
-
- # Garage
- garage_cols = [c for c in df.columns if c.startswith('Garage')]

```

```

- for c in df[garage_cols].select_dtypes(exclude='object').columns:
-     df[c] = df[c].fillna(0)
- for c in df[garage_cols].select_dtypes(include='object').columns:
-     df[c] = df[c].fillna('None')
-
- # Basement
- bsmt_cols = [c for c in df.columns if c.startswith('Bsmt')]
- for c in df[bsmt_cols].select_dtypes(exclude='object').columns:
-     df[c] = df[c].fillna(0)
- for c in df[bsmt_cols].select_dtypes(include='object').columns:
-     df[c] = df[c].fillna('None')
-
- # Masonry
- mas_cols = [c for c in df.columns if c.startswith('Mas')]
- for c in df[mas_cols].select_dtypes(exclude='object').columns:
-     df[c] = df[c].fillna(0)
- for c in df[mas_cols].select_dtypes(include='object').columns:
-     df[c] = df[c].fillna('None')
- ms_mode_by_nb = (
-     df.loc[train_mask]
-         .groupby('Neighborhood')['MSZoning']
-         .agg(lambda s: s.mode().iloc[0] if not s.mode().empty else
-             np.nan)
- )
- df['MSZoning'] =
- df['MSZoning'].fillna(df['Neighborhood'].map(ms_mode_by_nb))
-
- df['MSZoning'] = df['MSZoning'].fillna(df.loc[train_mask,
- 'MSZoning'].mode().iloc[0])
- mode_row_train = df.loc[train_mask].mode().iloc[0]
- df = df.fillna(mode_row_train)
-
- # =====
- # Ordinal mappings
- # =====
- df['MSSubClass'] = df['MSSubClass'].astype(object)
- df['ExterQual'] =
- df['ExterQual'].map({'Po':0, 'Fa':1, 'TA':2, 'Gd':3, 'Ex':4})
- df['ExterCond'] =
- df['ExterCond'].map({'Po':0, 'Fa':1, 'TA':2, 'Gd':3, 'Ex':4})
- df['BsmtQual'] =
- df['BsmtQual'].map({'None':0, 'Po':1, 'Fa':2, 'TA':3, 'Gd':4, 'Ex':5})
- df['BsmtCond'] =
- df['BsmtCond'].map({'None':0, 'Po':1, 'Fa':2, 'TA':3, 'Gd':4, 'Ex':5})
- df['BsmtExposure'] =
- df['BsmtExposure'].map({'None':0, 'No':1, 'Mn':2, 'Av':3, 'Gd':4})
- df['BsmtFinType1'] =
- df['BsmtFinType1'].map({'None':0, 'Unf':1, 'LwQ':2, 'Rec':3, 'BLQ':4,
- ALQ':5, 'GLQ':6})

```

```

- df['BsmtFinType2'] =
  df['BsmtFinType2'].map({'None':0, 'Unf':1, 'LwQ':2, 'Rec':3, 'BLQ':4, '
  ALQ':5, 'GLQ':6})
- df['HeatingQC'] =
  df['HeatingQC'].map({'Po':0, 'Fa':1, 'TA':2, 'Gd':3, 'Ex':4})
- df['KitchenQual'] =
  df['KitchenQual'].map({'Po':0, 'Fa':1, 'TA':2, 'Gd':3, 'Ex':4})
- df['GarageFinish'] =
  df['GarageFinish'].map({'None':0, 'Unf':1, 'RFn':2, 'Fin':3})
- df['GarageQual'] =
  df['GarageQual'].map({'None':0, 'Po':1, 'Fa':2, 'TA':3, 'Gd':4, 'Ex':5}
  )
- df['GarageCond'] =
  df['GarageCond'].map({'None':0, 'Po':1, 'Fa':2, 'TA':3, 'Gd':4, 'Ex':5}
  )
- df['FireplaceQu'] =
  df['FireplaceQu'].map({'None':0, 'Po':1, 'Fa':2, 'TA':3, 'Gd':4, 'Ex':5
  })
-
- # =====
- # Outlier rules
- # =====
- train_mask = (df['train'] == 1)
- df = df.drop(df[train_mask & (df['LotFrontage'] > 200)].index)
- df = df.drop(df[train_mask & df['1stFlrSF'] > 4000].index)
- df = df.drop(df[train_mask & df['LotArea'] > 100000].index)
- df = df.drop(df[train_mask & df['BsmtFinSF1'] > 4000].index)
- df = df.drop(df[train_mask & df['TotalBsmtSF'] > 5000].index)
- df = df.drop(df[train_mask & df['GrLivArea'] > 4000].index)
- # =====
- # Price Variable
- # =====
- y = df.loc[df['train'] == 1, 'SalePrice']
- y_test_real = df.loc[df['train'] == 0, 'SalePrice']
- # Drop columns
- df.drop([
-     'SalePrice', 'Id', 'LowQualFinSF', 'KitchenAbvGr',
-     'YearRemodAdd',
-     '3SsnPorch', 'PoolArea', 'MiscVal', 'Street', 'Utilities',
-     'Condition2', 'RoofMatl', 'Heating'
- ], axis=1, inplace=True, errors='ignore')
-
- # =====
- # One-hot encoding
- # =====
- categ_cols = df.select_dtypes(include=['object']).columns.tolist()
- df_enc = pd.get_dummies(df, columns=categ_cols, drop_first=False)
-
- # Split back to train/test

```

```

- X      = df_enc[df_enc['train'] == 1].drop(columns=['train'])
- testX = df_enc[df_enc['train'] == 0].drop(columns=['train'])
-
- # =====
- # Train / Test split and scaling
- # =====
- X_train, X_test, y_train, y_test = train_test_split(
-     X, y, test_size=0.3, random_state=12345
- )
-
- scaler = StandardScaler()
- X_train_sc = pd.DataFrame(scaler.fit_transform(X_train),
-     columns=X_train.columns, index=X_train.index)
- X_test_sc = pd.DataFrame(scaler.transform(X_test),
-     columns=X_test.columns, index=X_test.index)
-
- # =====
- # CV & training
- # =====
- folds = KFold(n_splits=4, shuffle=True, random_state=12345)
-
- def ModelTrainAndEvaluation(model, params, fit_intercept=False):
-     model_cv = GridSearchCV(
-         estimator=model,
-         param_grid=params,
-         scoring='neg_mean_squared_error',
-         cv=folds,
-         return_train_score=True,
-         verbose=1
-     )
-     model_cv.fit(X_train_sc, y_train)
-     alpha = model_cv.best_params_["alpha"]
-     print("Optimum lambda for %s is %f" % (model, alpha))
-     final_model = model_cv.best_estimator_
-
-     # Fit best model
-     final_model.fit(X_train_sc, y_train)
-     y_train_pred = final_model.predict(X_train_sc)
-     y_test_pred = final_model.predict(X_test_sc)
-
-     # Evaluation
-     print('R2 score (train) : ', round(r2_score(y_train,
- y_train_pred), 4))
-     print('R2 score (test) : ', round(r2_score(y_test,
- y_test_pred), 4))
-     print('RMSE (train) : ',
- round(np.sqrt(mean_squared_error(y_train, y_train_pred)), 4))

```

```

-     print('RMSE (test)          : ',
- round(np.sqrt(mean_squared_error(y_test, y_test_pred)), 4))
-
-     # Coefficients vs alpha
-     alphas = params['alpha']
-     coefs = []
-     for a in alphas:
-         model.set_params(alpha=a, max_iter=10000)
-         model.fit(X_train_sc, y_train)
-         coefs.append(model.coef_)
-     plt.figure(figsize=(10, 6))
-     ax = plt.gca()
-     ax.plot(alphas, coefs)
-     ax.set_xscale('log')
-     plt.xlabel('Lamda( $\lambda$ )')
-     plt.ylabel('Weights')
-     plt.title('Lasso: Coefficients vs Lambda( $\lambda$ )')
-     plt.show()
-
-     # Number of features vs alpha
-     num_features = []
-     for a in alphas:
-         model.set_params(alpha=a, max_iter=10000)
-         model.fit(X_train_sc, y_train)
-         num_selected = np.sum(model.coef_ != 0)
-         num_features.append(num_selected)
-     plt.figure()
-     plt.plot(alphas, num_features)
-     plt.xlabel('Lambda( $\lambda$ )')
-     plt.ylabel('Number of Features')
-     plt.title('Number of Features vs. Lambda( $\lambda$ )')
-     plt.show()
-
-     cv_results = pd.DataFrame(model_cv.cv_results_)
-     cv_results['param_alpha'] =
- cv_results['param_alpha'].astype('float32')
-     return final_model, model_cv.cv_results_['mean_test_score']
-
- params = {'alpha': [0.0001, 0.00025, 0.0005, 0.00065, 0.001,
- 0.00125, 0.0025,
-
-                 0.005, 0.01, 0.1, 1, 1.2, 1.5, 2]}
- params = {'alpha': 0.0001 * np.arange(1, 100)}
- lasso, cv_results = ModelTrainAndEvaluation(Lasso(max_iter=10000),
- params, fit_intercept=False)
-
- # Plot CV RMSE vs lambda
- alphas = params['alpha']

```

```

- cv_rmse = np.sqrt(-cv_results)
- plt.figure(figsize=(10, 6))
- plt.plot(alphas, cv_rmse, marker='o')
- plt.title('Lasso: RMSE vs Lambda( $\lambda$ ) ')
- plt.xlabel('Lamda( $\lambda$ ) ')
- plt.ylabel('RMSE')
- plt.xscale('log')
- plt.grid(False)
- plt.show()
- print("Lowest RMSE:", cv_rmse.min())
-
- # =====
- # Coefficients table
- # =====
- model_coefficients = pd.DataFrame(index=X_train_sc.columns)
- model_coefficients['Lasso'] = lasso.coef_
- pd.set_option('display.max_rows', None)
- print(len(model_coefficients))
-
- print("Total no of features : ",
-       model_coefficients[model_coefficients['Lasso'] !=
- 0][['Lasso']].shape[0])
-
- model_coef = (model_coefficients[model_coefficients['Lasso'] !=
- 0][['Lasso']]
-               .sort_values(by='Lasso', ascending=False,
- key=np.abs))
- print(model_coef)
-
- TOP_10 = model_coef[:10]
- print(TOP_10)
-
- # =====
- # Evaluation on the test set
- # =====
- test_aligned = testX.reindex(columns=X.columns, fill_value=0)
- X_Test_Real_sc = pd.DataFrame(scaler.transform(test_aligned),
-                               columns=X.columns,
-                               index=test_aligned.index)
-
- y_test_pred_real = lasso.predict(X_Test_Real_sc)
-
- rmse_real = np.sqrt(mean_squared_error(y_test_real,
- y_test_pred_real))
- print("Root Mean Squared Error (RMSE) on real test alignment:",
-       rmse_real)
-
- # Plot important coefficients

```

```

- coefs = pd.Series(lasso.coef_, index=X_train_sc.columns)
- print("Lasso picked " + str((coefs != 0).sum()) + " features and
  eliminated the other " +
-       str((coefs == 0).sum()) + " features")
- imp_coefs = pd.concat([coefs.sort_values().head(10),
  coefs.sort_values().tail(10)])
- imp_coefs.plot(kind="barh")
- plt.title("Coefficients in the Lasso Model")
- plt.show()

```

- مدينة دمشق

```

# -----
# 0) Imports (collected once)
# -----
import os
import warnings
import numpy as np
import pandas as pd
import matplotlib
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split, GridSearchCV, KFold
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LinearRegression, Lasso
from sklearn.metrics import r2_score, mean_squared_error

warnings.filterwarnings("ignore")

# -----
# Load data
# -----
CSV_PATH = "DamascusData.csv"
df = pd.read_csv(CSV_PATH)

# =====
# 2) Descriptive Analysis
# =====

quant_vars_updated = ['Price', 'Area', 'Nrooms', 'Floor', 'Quality',
'direction']

sns.set(style="whitegrid")
plt.figure(figsize=(15, 18))

for i, var in enumerate(quant_vars_updated):
    plt.subplot(3, 2, i + 1)
    sns.histplot(df[var], kde=True, bins=20, color='steelblue')

    if var == 'Quality':
        title = "Overall Building Quality"
    elif var == 'direction':

```

```

        title = "Number of Facades"
    elif var == 'Nrooms':
        title = "Number of Rooms"
    elif var == 'Floor':
        title = "Floor Number"
    elif var == 'Area':
        title = "Total Property Area (m²)"
    elif var == 'Price':
        title = "Property Price (SYP)"
    else:
        title = var

    plt.title(f"Distribution of {title}")
    plt.xlabel(title)
    plt.ylabel("Frequency")

plt.tight_layout()
plt.show()

# =====
# 3) Scatter Plots
# =====

quant_vars = ['Area', 'Nrooms', 'Floor', 'Quality', 'direction']
output_dir = "scatter_price_vs_variables"
os.makedirs(output_dir, exist_ok=True)

for var in quant_vars:
    plt.figure(figsize=(6, 4))
    plt.scatter(df[var], df['Price'], s=15, alpha=0.6, color='blue')
    plt.xlabel(var, fontsize=10)
    plt.ylabel("Price (SYP)", fontsize=10)
    plt.title(f"Scatter Plot of {var} vs Price", fontsize=11)
    plt.xticks(fontsize=9)
    plt.yticks(fontsize=9)
    plt.tight_layout()
    plt.savefig(f"{output_dir}/scatter_{var}_vs_Price.png", dpi=300)
    plt.close()

# =====
# 4) Linear Regression (train/test)
#     Price Variable uses loglp(Price)
# =====

X_lr = df.drop(columns=['Price'])
y_lr = np.loglp(df['Price'])

X_train_lr, X_test_lr, y_train_lr, y_test_lr = train_test_split(
    X_lr, y_lr, test_size=0.2, random_state=12345
)

scaler_lr = StandardScaler()
X_train_lr_sc = scaler_lr.fit_transform(X_train_lr)
X_test_lr_sc = scaler_lr.transform(X_test_lr)

lr_model = LinearRegression()
lr_model.fit(X_train_lr_sc, y_train_lr)

y_train_lr_pred = lr_model.predict(X_train_lr_sc)
y_test_lr_pred = lr_model.predict(X_test_lr_sc)

```

```

r2_train = r2_score(y_train_lr, y_train_lr_pred)
r2_test = r2_score(y_test_lr, y_test_lr_pred)
rmse_train = np.sqrt(mean_squared_error(y_train_lr, y_train_lr_pred))
rmse_test = np.sqrt(mean_squared_error(y_test_lr, y_test_lr_pred))

print("\n[Linear Regression - log(1+Price)]")
print("R2 (train):", r2_train)
print("R2 (test) :", r2_test)
print("RMSE (train):", rmse_train)
print("RMSE (test) :", rmse_test)

# =====
# 5) LASSO Regression
# =====

# Prepare features/target
X = df.drop(columns=['Price'])
y = np.log1p(df['Price'])

# Split
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=12345
)

# Standardize
scaler = StandardScaler()
scaler.fit(X_train)
X_train_sc = pd.DataFrame(scaler.transform(X_train),
    columns=X_train.columns, index=X_train.index)
X_test_sc = pd.DataFrame(scaler.transform(X_test),
    columns=X_test.columns, index=X_test.index)

# Cross-validation setup and training
folds = KFold(n_splits=4)
def ModelTrainAndEvaluation(model, params, fit_intercept=False):
    model_cv = GridSearchCV(
        estimator=model,
        param_grid=params,
        scoring='neg_mean_squared_error',
        cv=folds,
        return_train_score=True,
        verbose=1
    )
    model_cv.fit(X_train_sc, y_train)
    alpha = model_cv.best_params_["alpha"]
    print("\n[LASSO] Optimum alpha for %s is %f" % (model, alpha))
    final_model = model_cv.best_estimator_

# Fit & evaluate
final_model.fit(X_train_sc, y_train)
y_train_pred = final_model.predict(X_train_sc)
y_test_pred = final_model.predict(X_test_sc)

print('R2 score (train) : ', round(r2_score(y_train, y_train_pred), 4))
print('R2 score (test) : ', round(r2_score(y_test, y_test_pred), 4))
print('RMSE (train) : ', round(np.sqrt(mean_squared_error(y_train,
y_train_pred)), 4))
print('RMSE (test) : ', round(np.sqrt(mean_squared_error(y_test,
y_test_pred)), 4))

```

```

# Coefficients vs alpha
alphas_local = params['alpha']
coefs = []
for a in alphas_local:
    model.set_params(alpha=a)
    model.fit(X_train_sc, y_train)
    coefs.append(model.coef_)
plt.figure(figsize=(10, 6))
ax = plt.gca()
ax.plot(alphas_local, coefs)
ax.set_xscale('log')
plt.xlabel('Lambda ( $\lambda$ )')
plt.ylabel('Weights')
plt.title('LASSO: Coefficients vs Lambda ( $\lambda$ )')
plt.show()

# Number of features vs alpha
num_features = []
for a in alphas_local:
    model.set_params(alpha=a)
    model.fit(X_train_sc, y_train)
    num_selected_features = np.sum(model.coef_ != 0)
    num_features.append(num_selected_features)
plt.figure()
plt.plot(alphas_local, num_features)
plt.xlabel('Lambda ( $\lambda$ )')
plt.ylabel('Number of Features')
plt.title('Number of Features vs. Lambda ( $\lambda$ )')
plt.show()

# CV results (RMSE curve)
cv_results = pd.DataFrame(model_cv.cv_results_)
cv_results['param_alpha'] = cv_results['param_alpha'].astype('float32')
return final_model, model_cv.cv_results_['mean_test_score']

params = {
    'alpha': [0.0001, 0.00025, 0.0005, 0.00065, 0.001, 0.00125, 0.0025,
              0.005, 0.01, 0.1, 1, 1.2, 1.5, 2]
}

lasso, cv_results = ModelTrainAndEvaluaion(Lasso(), params,
fit_intercept=False)

# CV RMSE vs lambda
alphas = [0.0001, 0.00025, 0.0005, 0.00065, 0.001, 0.00125, 0.0025, 0.005,
0.01, 0.1, 1, 1.2, 1.5, 2]
cv_rmse = np.sqrt(-cv_results)
plt.figure(figsize=(10, 6))
plt.plot(alphas, cv_rmse, marker='o')
plt.title('LASSO: RMSE vs Lambda ( $\lambda$ )')
plt.xlabel('Lambda ( $\lambda$ )')
plt.ylabel('RMSE')
plt.xscale('log')
plt.grid(False)
plt.show()
print("Lowest RMSE:", cv_rmse.min())

# Coefficients table
model_coefficients = pd.DataFrame(index=X_train_sc.columns)

```

```

model_coefficients['LASSO'] = lasso.coef_
pd.set_option('display.max_rows', None)
print("Total features (including zeros):", len(model_coefficients))
kept = model_coefficients[model_coefficients['LASSO'] !=
0][['LASSO']].sort_values(
    by='LASSO', ascending=False, key=np.abs
)
print("Total non-zero features:", kept.shape[0])
print(kept.head(10))

# -----
# 10 random-case comparison on original price scale
# -----
y_test_pred_log = lasso.predict(X_test_sc)
y_test_real      = np.expml(y_test)
y_pred_real      = np.expml(y_test_pred_log)

np.random.seed(42)
sample_indices = np.random.choice(len(y_test), size=10, replace=False)

comparison_lasso_df = pd.DataFrame({
    "Listed Price": y_test_real.values[sample_indices].round(0),
    "Predicted Price (LASSO)": y_pred_real[sample_indices].round(0),
    "Difference": (y_pred_real[sample_indices] -
y_test_real.values[sample_indices]).round(0)
}, index=[f"Property {i+1}" for i in range(10)])

print("\n[LASSO] Random 10 comparisons (original price scale):")
print(comparison_lasso_df)

```

Abstract

Estimating real estate prices is a complex task in economic and statistical studies due to the multitude of factors influencing market value and the difficulty of collecting and analyzing them within an accurate model. This study aims to construct a statistical model for estimating residential property prices using the LASSO (Least Absolute Shrinkage and Selection Operator) regression method and to conduct a systematic comparison with the traditional Multiple Linear Regression (OLS) model to evaluate the performance of each.

The first phase of the study utilized data from Ames, Iowa—widely recognized as a benchmark in housing price analysis—due to its richness in thoroughly documented and academically validated variables. This allowed for testing the LASSO model's ability to identify the most influential predictors within a well-defined environment. Results demonstrated the model's effectiveness in reducing the number of included variables while maintaining strong predictive accuracy.

In the second phase, the model was applied to field data collected from the Damascus real estate market in 2024. Given the scarcity of official housing data in Syria, the study relied on non-official field sources. After preprocessing, the final dataset consisted of 250 observations, including both quantitative and qualitative variables related to property features, location, and amenities.

The empirical results showed that the LASSO model outperformed OLS by producing a more parsimonious model (with only seven predictors) without significant loss in performance. The most influential variables in determining housing prices in Damascus were: area, finishing quality, furniture availability, balcony presence, number of rooms, elevator access, and floor level. These findings underscore the importance of using statistical methods like LASSO in contexts where data are high-dimensional or scarce, as they provide simpler models with high interpretability, making them valuable tools in real estate valuation.

Key words:

Real estate prices, LASSO regression, Multiple linear regression, Property valuation, Variable selection, Coordinate descent, Soft-Threshold.

Syrian Arab Republic

Damascus University

Faculty of Economic

Department of Statistics and

Information Systems



**A Proposed Statistical Model to Estimate The Real Value
of The Real Estates**

A Practical Study in The City of Damascus.

**A Dissertation Submitted in Partial Fulfillment of The
Requirements For The Degree of Master in
Quantitative Methods**

By

Muhammad Nidal Bassam AlSaqqqa

Supervisor

Prof. Maisaa Hamsho

Professor in the Department of Statistics and Information Systems

Faculty of Economics Damascus University

2025–2026