Факультет компьютерных наук | НИС Python | Москва 2025

Лекция 4
Линейная регрессия

Преподаватель: Меликян Алиса Валерьевна, amelikyan@hse.ru
кандидат наук, доцент Департамента программной инженерии

Регрессионный анализ

Позволяет выявить статистическую взаимосвязь между исследуемыми переменными и математическую формулу функции зависимости переменных. Результаты регрессионного анализа могут быть использованы для прогнозирования изменения значений количественной переменной. Использование результатов регрессионного анализа для прогнозирования предполагает ряд ограничений:

1. Необходимо убедиться не только в статистической, но и в каузальной взаимосвязи двух переменных;

2. Необходимо убедиться, что внешние эффекты не оказывают влияния на изменение одной из переменных.

Простая линейная регрессия

Позволяет выразить линейную взаимосвязь между переменными в виде уравнения прямой:

$$Y = a + b \cdot X$$

где X – независимая переменная (предиктор);
Y – зависимая переменная;
a, b – постоянные величины (параметры модели).

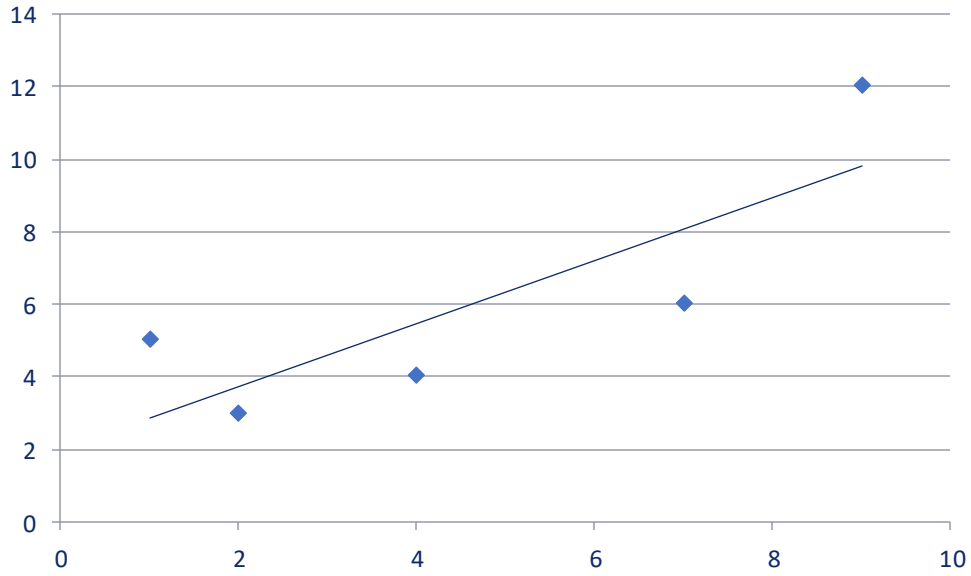
Параметры модели определяются с помощью *метода наименьших квадратов*. Для того, чтобы оценить насколько хорошо прямая выражает связь между переменными, полезно изучить диаграмму рассеяния, чтобы убедиться в наличии линейной взаимосвязи и обнаружить выбросы, которые могут существенно исказить результат.

Простая линейная регрессия: пример расчёта

X	Y
1	5
2	3
4	4
7	6
9	12

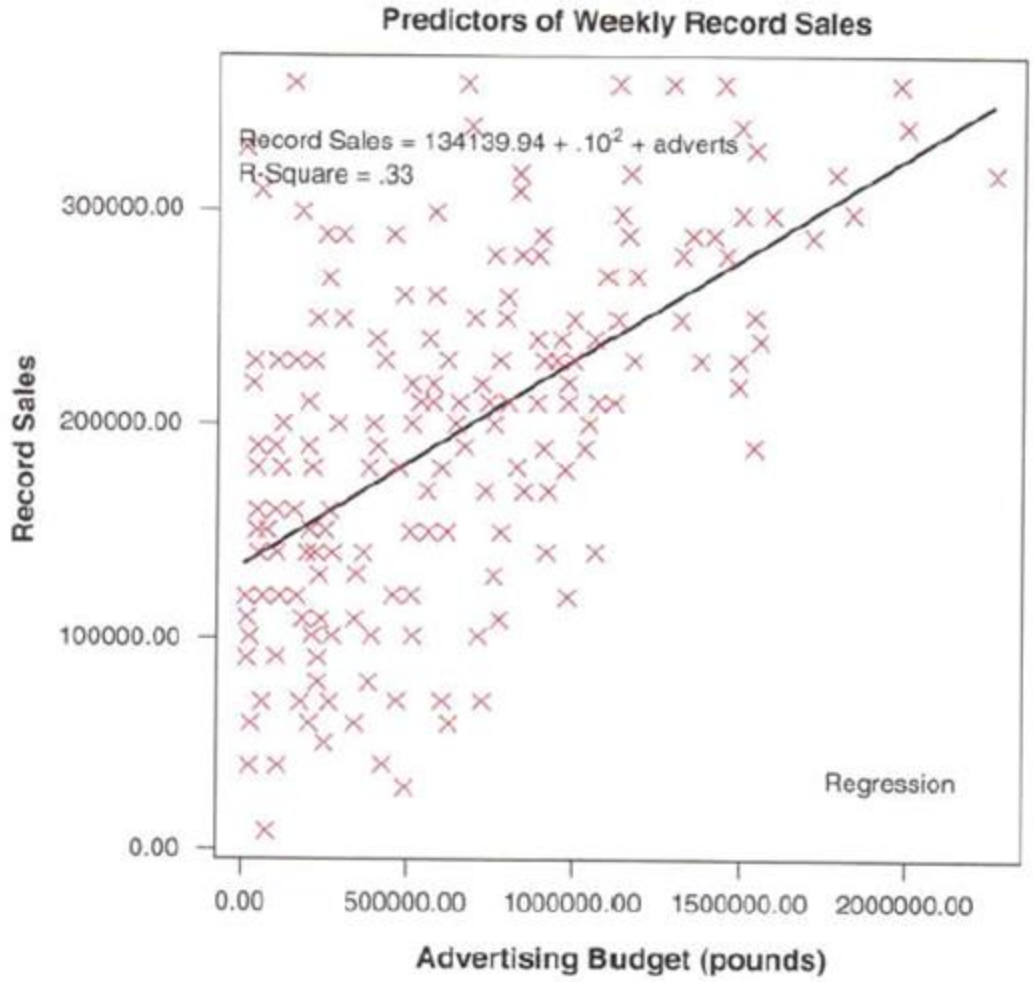
$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$a = \bar{y} - b\bar{x}$$

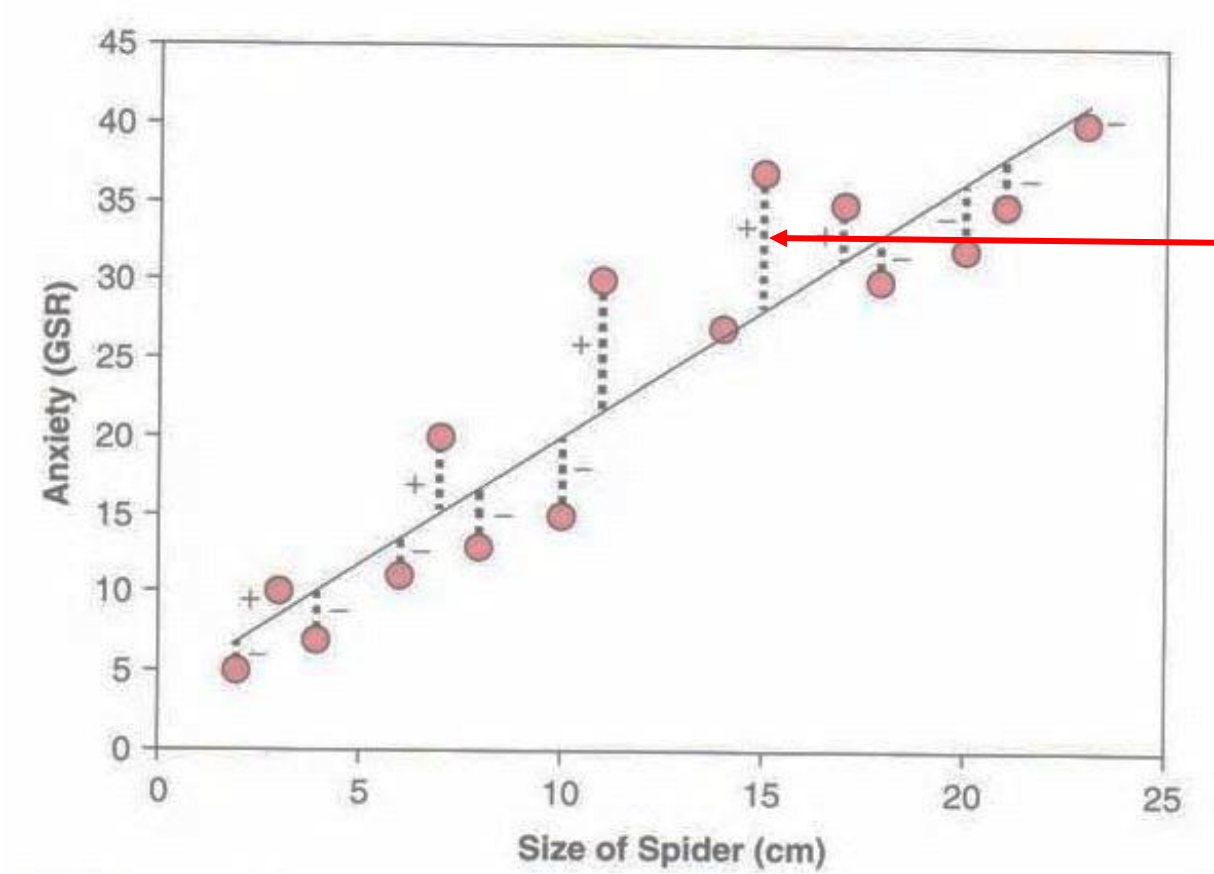


$$Y = 0{,}8*X + 2$$

Диаграмма рассеяния с регрессионной прямой



Остатки



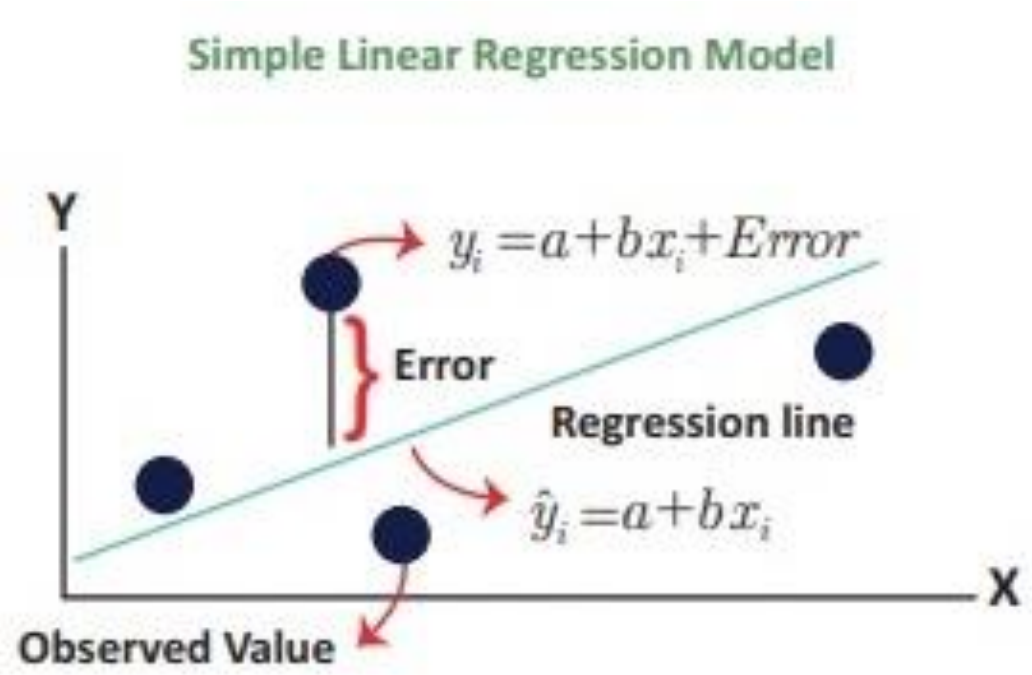
Остатки – отклонения между реальными и предсказанными моделью значениями зависимой переменной. Могут принимать нулевые, отрицательные и положительные значения.

Метод наименьших квадратов

Цель – подобрать такие параметры модели a и b, чтобы минимизировать квадраты отклонений реальных и предсказанных значений зависимой переменной.

$$E(a,b) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$i.e.,\ E(a,b) = \sum_{i=1}^{n} (y_i - a - bx_i)^2 .$$



Три вида сумм квадратов

Сумма квадратов остатков (Residual Sum of Squares) – RSS

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

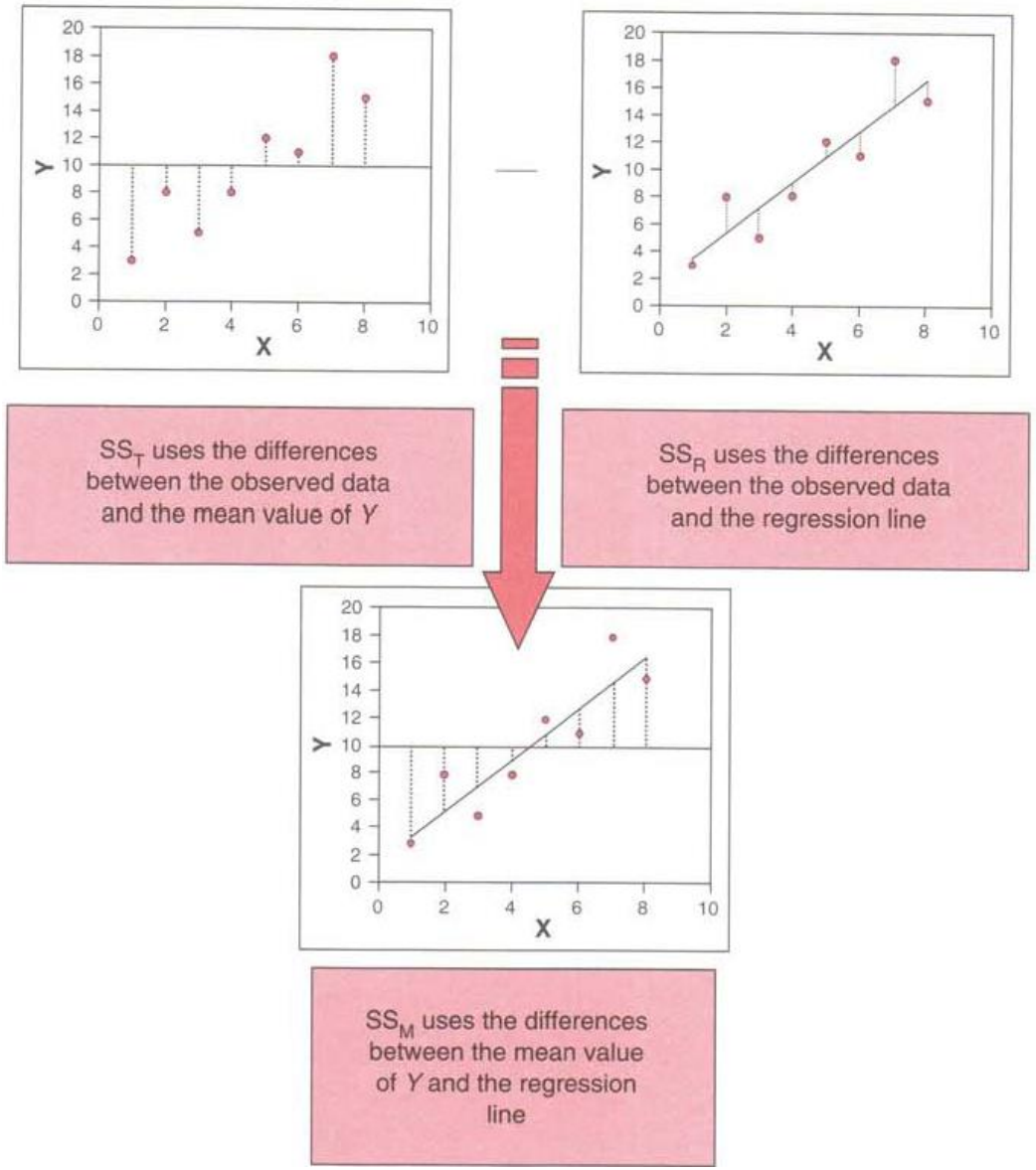
Общая сумма квадратов (Total Sum of Squares) – TSS

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Объяснённая сумма квадратов (Explained Sum of Squares) – ESS

$$ESS = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

TSS = RSS + ESS


Y
X
SST uses the differences between the observed data and the mean value of Y
SSR uses the differences between the observed data and the regression line
SSM uses the differences between the mean value of Y and the regression line

Множественная линейная регрессия

В регрессионном анализе участвуют несколько независимых переменных. Уравнение регрессии принимает вид:

$$Y = a + b_1 \cdot X_1 + b_2 \cdot X_2 + \ldots + b_n \cdot X_n$$

Независимые переменные могут коррелировать между собой (мультиколлинеарность), что необходимо учитывать при определении коэффициентов уравнения регрессии.

Добавление категориальных переменных в модель

Если переменная дихотомическая, то можно её добавить в регрессионную модель.

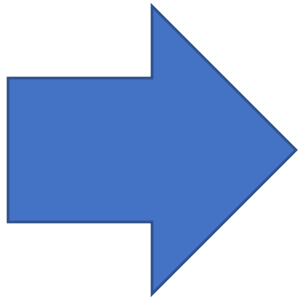
Если переменная имеет больше двух значений необходимо перекодировать её в набор фиктивных переменных, принимающих значения 0 и 1. Создаётся n-1 дихотомическая переменная, где n – число значений, принимаемых переменной.

Не создаётся отдельная переменная для «базовой» или «референтной» категории, т.е. категории, с которой будет происходить сравнение остальных категорий. Эта может быть любая категория. Если нет предпочтений, то можно выбрать в качестве референтной самую многочисленную группу.

Создание дамми-переменных

Исходная переменная «уровень образования» принимает следующие значения: нет (1), среднее (2), высшее (3), послевузовское (4). Можно выбрать высшее образование как референтную группу.

Уровень образования
Высшее
Нет
Среднее
Послевузовское
Высшее
Среднее



Ed1	Ed2	Ed3
0	0	0
1	0	0
0	1	0
0	0	1
0	0	0
0	1	0

Коэффициент детерминации

R-squared

Лежит в диапазоне от 0 до 1. Интерпретируется как доля дисперсии зависимой переменной, «объяснённой» независимыми переменными. Иначе говоря – доля объяснённого разброса в общем разбросе значений зависимой переменной.

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2}$$

Скорректированный коэффициент детерминации

Adj R-squared

Для сравнения моделей с разным числом факторов (независимых переменных) так, чтобы число факторов не влияло на статистику R-квадрат используется скорректированный коэффициент детерминации.

$$R^2_{adj} = 1 - (1 - R^2)\frac{(n-1)}{(n-k)} \leqslant R^2$$

где *n* — количество наблюдений, а k — количество параметров.

F-статистика

F-статистика и соответствующий ей уровень значимости позволяют оценить статистическую значимость регрессионной модели. Проверяется нулевая гипотеза о том, что все коэффициенты регрессии равны нулю. В этом случае модель не имеет предсказательной силы. По сути, F-тест сравнивает регрессионную модель с моделью среднего (содержит только константу, предсказывает все значения зависимой переменной на основе среднего значения) и позволяет понять произойдёт ли значимое улучшение точности предсказания при добавлении в модель предикторов.

$$F = \frac{R^2}{1 - R^2} \frac{(n - m - 1)}{m}$$

Средняя абсолютная ошибка (Mean Absolute Error)

Рассчитывается как среднее абсолютных разностей между фактическим значением зависимой переменной и значением, предсказанным моделью.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|,$$

где N — число наблюдений,

y_i — ффактическое значение зависимой переменной для i-го наблюдения,

$\hat{y}_i$ — предсказанное моделью значение зависимой переменной .

Среднеквадратичная ошибка (Mean Squared Error)

Применяется в случаях, когда требуется подчеркнуть большие ошибки и выбрать модель, которая дает меньше именно больших ошибок. Большие значения ошибок становятся заметнее за счет квадратичной зависимости.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

где n — количество наблюдений,

y_i — ффактическое значение зависимой переменной для i-го наблюдения,

$\hat{y}_i$ — предсказанное моделью значение зависимой переменной.

Корень из среднеквадратичной ошибки (Root Mean Squared Error)

Вычисляется как квадратный корень из MSE.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}$$

RMSE также штрафует за большие ошибки, но в отличие от MSE, масштаб ошибки аналогичен исходным данным, что облегчает интерпретацию. Это делает RMSE хорошим выбором для многих практических задач, где важна интерпретируемость результата.

Информационный критерий Акаике

Применяется для выбора из нескольких альтернативных статистических моделей. Критерий вознаграждает за качество приближения и штрафует за использование излишнего количества параметров модели. Считается, что наилучшей будет модель с наименьшим значением критерия AIC. Абсолютное значение AIC не имеет смысла — он указывает только на относительный порядок сравниваемых моделей.

$$RSS = \sum_{i=1}^{n} \hat{\varepsilon}_i^2.$$

$$AIC = 2k + n[\ln(RSS)]$$

RSS – Residual Sum of Squares, n – число наблюдений, k – число параметров модели

Оценка регрессионной модели

Если модель хорошо отражает реальность, то остатки должны появляться случайно (не систематически), подчиняться нормальному распределению и не коррелировать с зависимой переменной.

Мультиколлинеарность – это состояние высокой степени корреляции между независимыми переменными, входящими в модель. Рекомендуется не включать в регрессионную модель связанные между собой независимые переменные. Одним из способов решения проблемы мультиколлинеарности может быть проведение предварительного факторного анализа предикторов.

Мультиколлинеарность

Ситуация высокой корреляции между независимыми переменными. Оценить можно посчитав корреляции между переменными (не должны превышать 0,75) или коэффициент вздутия дисперсии VIF (variance-inflation factor). Показывает силу корреляции переменной с другими независимыми переменными. Для каждой независимой переменной строится регрессия, в которой она является зависимой переменной, а остальные предикторы модели независимыми переменными. Таким образом оценивается процент вариации предиктора, объясняемой другими предикторами. VIF не должен превышать 10. Tolerance = 1 / VIF.

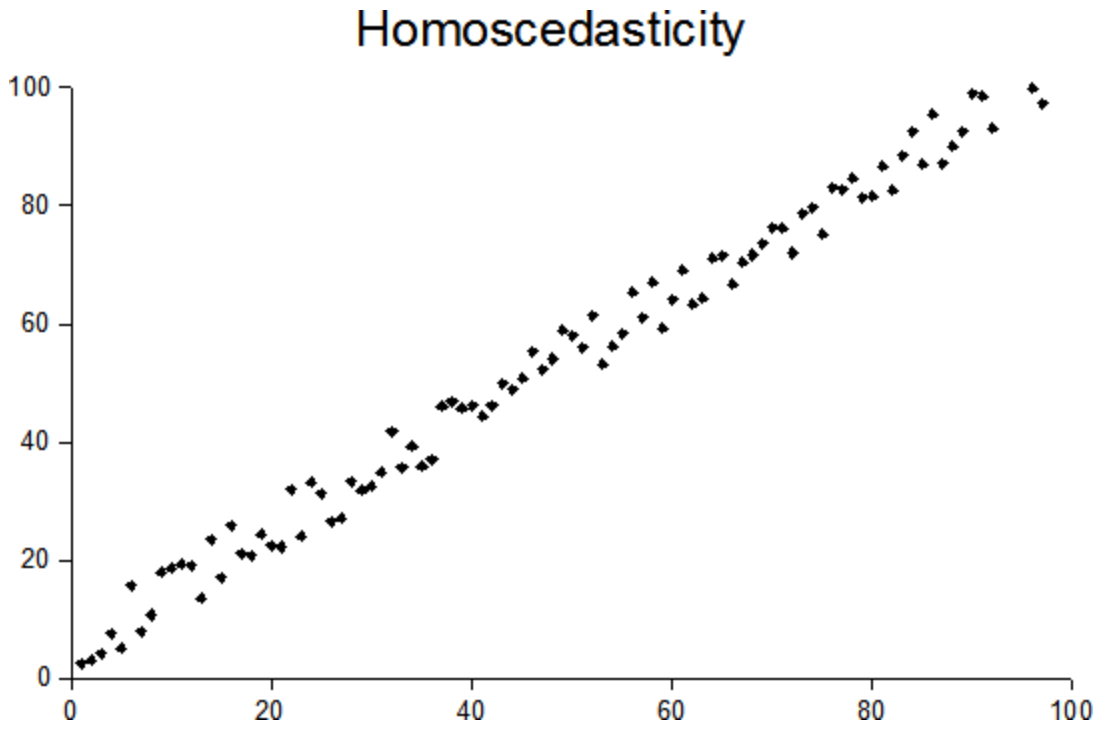
$$VIF\ x_i = \frac{1}{Tolerance} = \frac{1}{1 - R_i^2}$$

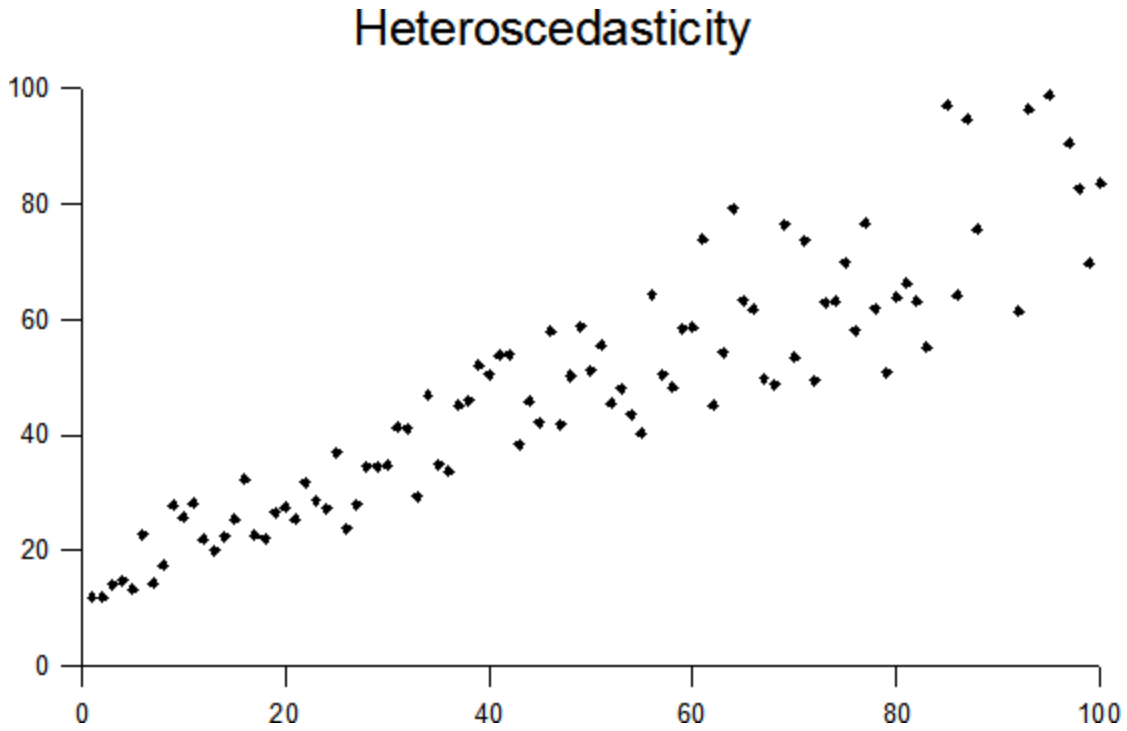
Гетероскедастичность

Непостоянство дисперсии остатков. Дисперсия остатков не случайна по отношению к зависимой переменной.

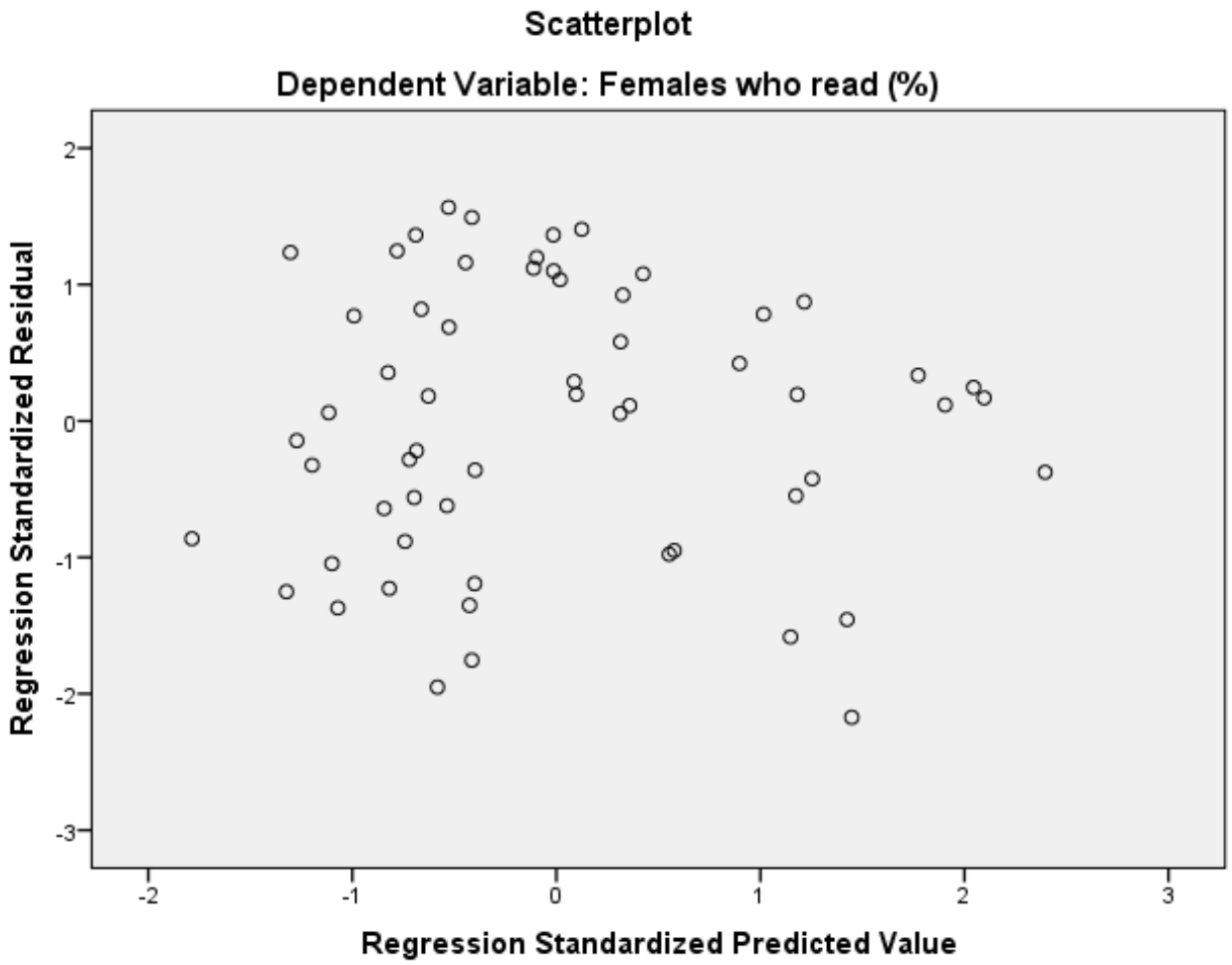
Можно выявить на диаграмме рассеяния между стандартизированными остатками и предсказанными значениями зависимой переменной. Графическая зависимость будет указывать на наличие гетероскедастичности.

Гетероскедастичность на графике

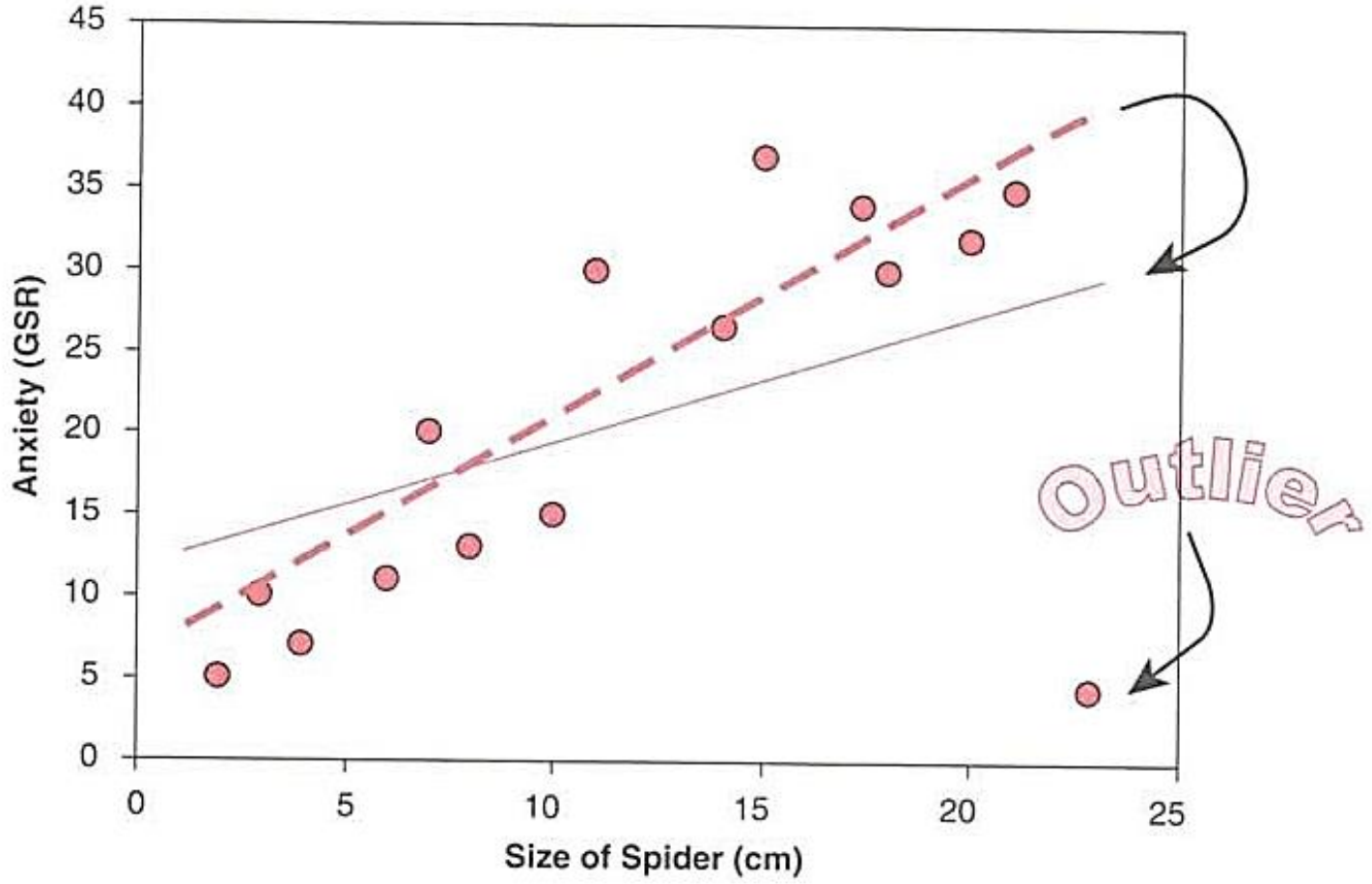




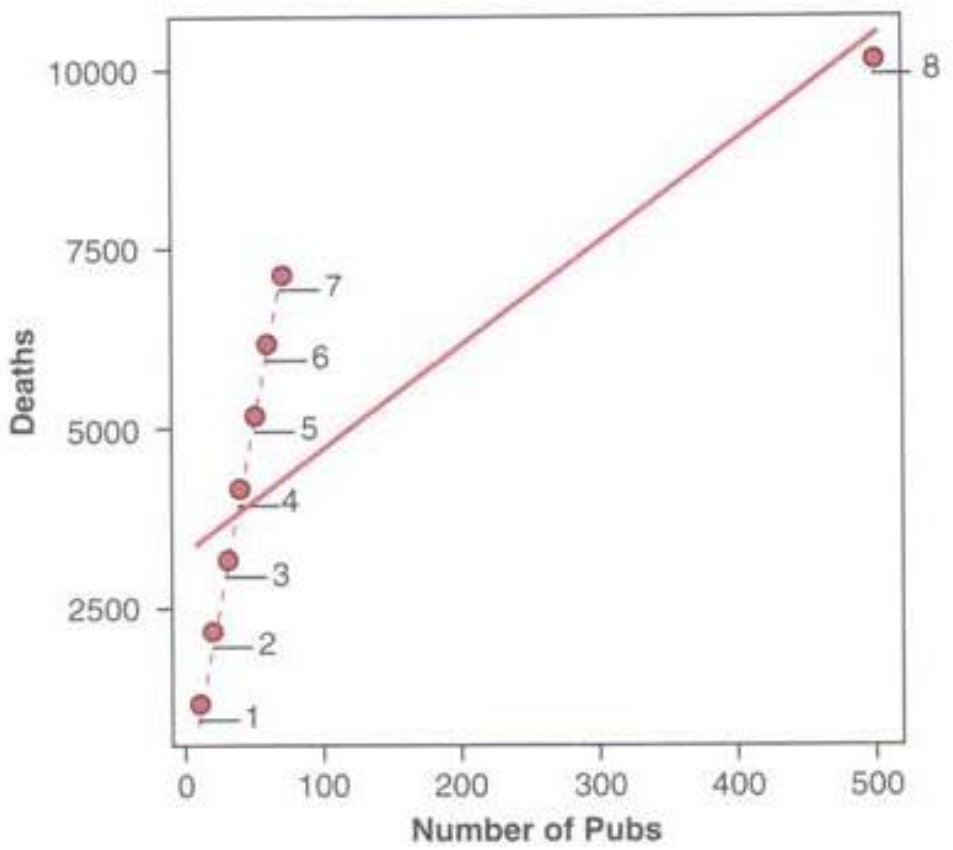
Гетероскедастичность на графике



Диагностика модели: выбросы



Диагностика модели: влияющие наблюдения



Константа – значение зависимой переменной, когда все предикторы равны нулю

OLS Regression Results

Dep. Variable:	SalePrice	R-squared:	0.681
Model:	OLS	Adj. R-squared:	0.681
Method:	Least Squares	F-statistic:	1037.
Date:	Fri, 25 Feb 2022	Prob (F-statistic):	0.00
	15:31:21	Log-Likelihood:	-17709.
	1460	AIC:	3.543e+04
	1456	BIC:	3.545e+04
	3		
	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-3.132e+04	3992.921	-7.844	0.000	-3.92e+04	-2.35e+04
GrLivArea	65.6106	2.667	24.600	0.000	60.379	70.842
GarageCars	3.365e+04	1854.413	18.146	0.000	3e+04	3.73e+04
TotalBsmtSF	50.4508	3.136	16.087	0.000	44.299	56.603

Omnibus:	520.280	Durbin-Watson:	1.978
Prob(Omnibus):	0.000	Jarque-Bera (JB):	31276.464
Skew:	-0.822	Prob(JB):	0.00
Kurtosis:	25.615	Cond. No.	6.64e+03

Коэффициент при переменной означает, что увеличение её значения на единицу приведёт к росту значения зависимой переменной на 65,6 единиц

OLS Regression Results

Dep. Variable:	SalePrice	R-squared:	0.681
Model:	OLS	Adj. R-squared:	0.681
Method:	Least Squares	F-statistic:	1037.
Date:	Fri, 25 Feb 2022	Prob (F-statistic):	0.00
Time:	15:31:21	Log-Likelihood:	-17709.
	1460	AIC:	3.543e+04
	1456	BIC:	3.545e+04
	3		
	obust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-3.132e+04	3992.921	-7.844	0.000	-3.92e+04	-2.35e+04
GrLivArea	65.6106	2.667	24.600	0.000	60.379	70.842
GarageCars	3.365e+04	1854.413	18.146	0.000	3e+04	3.73e+04
TotalBsmtSF	50.4508	3.136	16.087	0.000	44.299	56.603

Omnibus:	520.280	Durbin-Watson:	1.978
Prob(Omnibus):	0.000	Jarque-Bera (JB):	31276.464
Skew:	-0.822	Prob(JB):	0.00
Kurtosis:	25.615	Cond. No.	6.64e+03

OLS Regression Results

Dep. Variable:	SalePrice	R-squared:	0.681
Model:	OLS	Adj. R-squared:	0.681
Method:	Least Squares	F-statistic:	1037.
	ri, 25 Feb 2022	Prob (F-statistic):	0.00
	15:31:21	Log-Likelihood:	-17709.
Covariance Type:	nonr		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-3.132e+04	3992.921	-7.844	0.000	-3.92e+04	-2.35e+04
GrLivArea	65.6106	2.667	24.600	0.000	60.379	70.842
GarageCars	3.365e+04	1854.413	18.146	0.000	3e+04	3.73e+04
TotalBsmtSF	50.4508	3.136	16.087	0.000	44.299	56.603

Omnibus:	520.280	Durbin-Watson:	1.978
Prob(Omnibus):	0.000	Jarque-Bera (JB):	31276.464
Skew:	-0.822	Prob(JB):	0.00
Kurtosis:	25.615	Cond. No.	6.64e+03

Коэффициенты регрессионной модели, на их основе формируется уравнение регрессии.

Уравнение регрессии: SalePrice = -31320 + 65,6*GrLivArea + 33650*GarageCars + 50,5*TotalBsmtSF

OLS Regression Results

Dep. Variable:	SalePrice	R-squared:	0.681
Model:	OLS	Adj. R-squared:	0.681
Method:	Least Squares	F-statistic:	1037.
Date:	Fri, 25 Feb 2022	Prob (F-statistic):	0.00
Time:	15:31:21	Log-Likelihood:	-17709.
No. C	1460		543e+04
Df Re	1456		545e+04
Df Mo	3		
Covar	nonrobust		

Стандартные ошибки коэффициентов регрессии

$$s(b_1) = \sqrt{\frac{1}{n-2} * \frac{\sum (y_i - \hat{y}_i)^2}{\sum (x_i - \bar{x})^2}}$$

	coef	std err	t	P>|t|	[0.025	0.975]
const	-3.132e+04	3992.921	-7.844	0.000	-3.92e+04	-2.35e+04
GrLivArea	65.6106	2.667	24.600	0.000	60.379	70.842
GarageCars	3.365e+04	1854.413	18.146	0.000	3e+04	3.73e+04
TotalBsmtSF	50.4508	3.136	16.087	0.000	44.299	56.603

Omnibus:	520.280	Durbin-Watson:	1.978
Prob(Omnibus):	0.000	Jarque-Bera (JB):	31276.464
Skew:	-0.822	Prob(JB):	0.00
Kurtosis:	25.615	Cond. No.	6.64e+03

OLS Regression Results

Dep. Variable:	SalePrice	R-squared:	0.681
Model:	OLS	Adj. R-squared:	0.681
Method:	Least Squares	F-statistic:	1037.
Date:	Fri, 25 Feb 2022	Prob (F-statistic):	0.00
Time:	15:31:21	Log-Likelihood:	-17709.
No. Observ	1460	AIC:	3.543e+04
Df Residua	1456	BIC:	3.545e+04
Df Model:	3		
Covariance	robust		

$$t = \frac{coef}{std\ err}$$

	coef	std err	t	P>|t|	[0.025	0.975]
const	-3.132e+04	3992.921	-7.844	0.000	-3.92e+04	-2.35e+04
GrLivArea	65.6106	2.667	24.600	0.000	60.379	70.842
GarageCars	3.365e+04	1854.413	18.146	0.000	3e+04	3.73e+04
TotalBsmtSF	50.4508	3.136	16.087	0.000	44.299	56.603

Omnibus:	520.280	Durbin-Watson:	1.978
Prob(Omnibus):	0.000	Jarque-Bera (JB):	31276.464
Skew:	-0.822	Prob(JB):	0.00
Kurtosis:	25.615	Cond. No.	6.64e+03

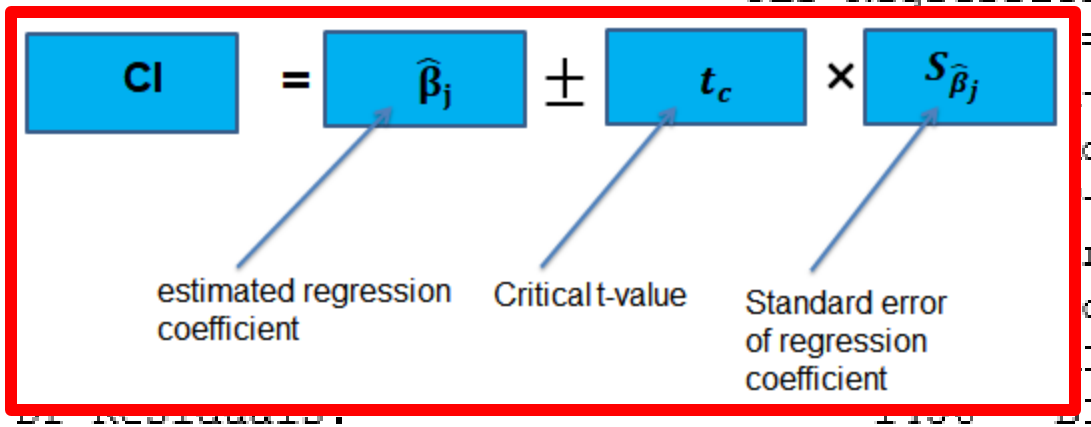
OLS Regression Results

Dep. Variable:	SalePrice	R-squared:	0.681
Model:	OLS	Adj. R-squared:	0.681
Method:	Least Squares	F-statistic:	1037.
Date:	Fri, 25 Feb 2022	Prob (F-statistic):	0.00
Time:	15:31:21	Log-Likelihood:	-17709.
No. Observations:			3.543e+04
Df Residuals:			3.545e+04
Df Model:			
Covariance Type:			

Статистическая значимость коэффициентов регрессии

	coef	std err	t	P>|t|	[0.025	0.975]
const	-3.132e+04	3992.921	-7.844	0.000	-3.92e+04	-2.35e+04
GrLivArea	65.6106	2.667	24.600	0.000	60.379	70.842
GarageCars	3.365e+04	1854.413	18.146	0.000	3e+04	3.73e+04
TotalBsmtSF	50.4508	3.136	16.087	0.000	44.299	56.603

Omnibus:	520.280	Durbin-Watson:	1.978
Prob(Omnibus):	0.000	Jarque-Bera (JB):	31276.464
Skew:	-0.822	Prob(JB):	0.00
Kurtosis:	25.615	Cond. No.	6.64e+03



OLS Regression Results

-squared:			0.681
dj. R-squared:			0.681
-statistic:			1037.
rob (F-statistic):			0.00
og-Likelihood:			-17709.
IC:			3.543e+04
IC:			3.545e+04
Df Model:			3
Covariance Type:			nonrobust

95% доверительный интервал

	coef	std err	t	P>|t|	[0.025	0.975]
const	-3.132e+04	3992.921	-7.844	0.000	-3.92e+04	-2.35e+04
GrLivArea	65.6106	2.667	24.600	0.000	60.379	70.842
GarageCars	3.365e+04	1854.413	18.146	0.000	3e+04	3.73e+04
TotalBsmtSF	50.4508	3.136	16.087	0.000	44.299	56.603

Omnibus:	520.280	Durbin-Watson:	1.978
Prob(Omnibus):	0.000	Jarque-Bera (JB):	31276.464
Skew:	-0.822	Prob(JB):	0.00
Kurtosis:	25.615	Cond. No.	6.64e+03

OLS Regression Results

Dep. Variable:	SalePrice	R-squared:	0.681
Model:	OLS	Adj. R-squared:	0.681
Method:	Least Squares	F-statistic:	1037.
Date:	Fri, 25 Feb 2022	Prob (F-statistic):	0.00
Time:	15:31:21	Log-L	-17709.
No. Observations:	1460	AIC:	3.543e+04
Df Residuals:	1456	BIC:	3.545e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-3.132e+04	3992.921	-7.844	0.000	-3.92e+04	-2.35e+04
GrLivArea	65.6106	2.667	24.600	0.000	60.379	70.842
GarageCars	3.365e+04	1854.413	18.146	0.000	3e+04	3.73e+04
TotalBsmtSF	50.4508	3.136	16.087	0.000	44.299	56.603

Omnibus:	520.280	Durbin-Watson:	1.978
Prob(Omnibus):	0.000	Jarque-Bera (JB):	31276.464
Skew:	-0.822	Prob(JB):	0.00
Kurtosis:	25.615	Cond. No.	6.64e+03

Коэффициент детерминации – модель объясняет 68% вариации значений зависимой переменной

OLS Regression Results

Dep. Variable:	SalePrice	R-squared:	0.681
Model:	OLS	Adj. R-squared:	0.681
Method:	Least Squares	F-statistic:	1037.
Date:	Fri, 25 Feb 2022	Prob (F-statistic):	0.00
Time:	15:31:21	Log-Likelihood:	-17709.
No. Observations:	1460	AIC:	3.543e+04
Df Residuals:	1456	BIC:	3.545e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|	.975]
const	-3.132e+04	3992.921	-7.844	0.000	-3.92e+04 -2.35e+04
GrLivArea	65.6106	2.667	24.600	0.000	60.379 70.842
GarageCars	3.365e+04	1854.413	18.146	0.000	3e+04 3.73e+04
TotalBsmtSF	50.4508	3.136	16.087	0.000	44.299 56.603

Omnibus:	520.280	Durbin-Watson:	1.978
Prob(Omnibus):	0.000	Jarque-Bera (JB):	31276.464
Skew:	-0.822	Prob(JB):	0.00
Kurtosis:	25.615	Cond. No.	6.64e+03

Модель статистически значима

OLS Regression Results

Dep. Variable:	SalePrice	R-squared:	0.681
Model:	OLS	Adj. R-squared:	0.681
Method:	Least Squares	F-statistic:	1037.
Date:	Fri, 25 Feb 2022	Prob (F-statistic):	0.00
Time:	15:31:21	Log-Likelihood:	-17709.
No. Observations:	1460	AIC:	3.543e+04
Df Residuals:	1456	BIC:	3.545e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|		.975]
const	-3.132e+04	3992.921	-7.844	0.0		6e+04
GrLivArea	65.6106	2.667	24.600	0.000	60.379	70.842
GarageCars	3.365e+04	1854.413	18.146	0.000	3e+04	3.73e+04
TotalBsmtSF	50.4508	3.136	16.087	0.000	44.299	56.603

Omnibus:	520.280	Durbin-Watson:	1.978
Prob(Omnibus):	0.000	Jarque-Bera (JB):	31276.464
Skew:	-0.822	Prob(JB):	0.00
Kurtosis:	25.615	Cond. No.	6.64e+03

Информационный критерий Акаике

OLS Regression Results

Dep. Variable:	SalePrice	R-squared:	0.681
Model:	OLS	Adj. R-squared:	0.681
Method:	Least Squares	F-statistic:	1037.
Date:	Fri, 25 Feb 2022	Prob (F-statistic):	0.00
Time:	15:31:21	Log-Likelihood:	-17709.
No. Observations:	1460	AIC:	3.543e+04
Df Residuals:	1456	BIC:	3.545e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err			[0.025	0.975]
const	-3.132e+04	3992.921	-7		92e+04	-2.35e+04
GrLivArea	65.6106	2.667	24		60.379	70.842
GarageCars	3.365e+04	1854.413	18.146	0.000	3e+04	3.73e+04
TotalBsmtSF	50.4508	3.136	16.087	0.000	44.299	56.603

Omnibus:	520.280	Durbin-Watson:	1.978
Prob(Omnibus):	0.000	Jarque-Bera (JB):	31276.464
Skew:	-0.822	Prob(JB):	0.00
Kurtosis:	25.615	Cond. No.	6.64e+03

Принудительное включение переменных в модель

OLS Regression Results

Dep. Variable:	SalePrice	R-squared:	0.681
Model:	OLS	Adj. R-squared:	0.681
Method:	Least Squares	F-statistic:	1037.
Date:	Fri, 25 Feb 2022	Prob (F-statistic):	0.00
Time:	15:31:21	Log-Likelihood:	-17709.
No. Observations:	1460	AIC:	3.543e+04
Df Residuals:	1456	BIC:	3.545e+04
Df Model:	3		
Covariance Type:	nonrobust		

			t	P>|t|	[0.025	0.975]
const			844	0.000	-3.92e+04	-2.35e+04
GrLiv			600	0.000	60.379	70.842
GarageCars	3.365e+04	1854.413	18.146	0.000	3e+04	3.73e+04
TotalBsmtSF	50.4508	3.136	16.087	0.000	44.299	56.603

Omnibus:	520.280	Durbin-Watson:	1.978
Prob(Omnibus):	0.000	Jarque-Bera (JB):	31276.464
Skew:	-0.822	Prob(JB):	0.00
Kurtosis:	25.615	Cond. No.	6.64e+03

Тест на нормальность остатков, их распределение отличается от нормального

OLS Regression Results

Dep. Variable:	SalePrice	R-squared:	0.681
Model:	OLS	Adj. R-squared:	0.681
Method:	Least Squares	F-statistic:	1037.
Date:	Fri, 25 Feb 2022	Prob (F-statistic):	0.00
Time:	15:31:21	Log-Likelihood:	-17709.
No. Observations:	1460	AIC:	3.543e+04
Df Residuals:	1456	BIC:	3.545e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const			844	0.000	-3.92e+04	-2.35e+04
GrLiv			600	0.000	60.379	70.842
Garag			146	0.000	3e+04	3.73e+04
Total			087	0.000	44.299	56.603

Omnibus:	520.280	Durbin-Watson:	1.978
Prob(Omnibus):	0.000	Jarque-Bera (JB):	31276.464
Skew:	-0.822	Prob(JB):	0.00
Kurtosis:	25.615	Cond. No.	6.64e+03

Симметричность и пологость распределения остатков

OLS Regression Results

Dep. Variable:	SalePrice	R-squared:	0.681
Model:	OLS	Adj. R-squared:	0.681
Method:	Least Squares	F-statistic:	1037.
Date:	Fri, 25 Feb 2022	Prob (F-statistic):	0.00
Time:	15:31:21	Log-Likelihood:	-17709.
No. Observations:	1460	AIC:	3.543e+04
Df Residuals:	1456	BIC:	3.545e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t			
const	-3.132e+04	3992.921	-7.844			
GrLivArea	65.6106	2.667	24.600			
GarageCars	3.365e+04	1854.413	18.146	0.000	3e+04	3.73e+04
TotalBsmtSF	50.4508	3.136	16.087	0.000	44.29	56.603

Omnibus:	520.280	Durbin-Watson:	1.978
Prob(Omnibus):	0.000	Jarque-Bera (JB):	31276.464
Skew:	-0.822	Prob(JB):	0.00
Kurtosis:	25.615	Cond. No.	6.64e+03

Тест на гомоскедастичность. Значения должны быть в диапазоне от 1 до 2.

OLS Regression Results

Dep. Variable:	SalePrice	R-squared:	0.681
Model:	OLS	Adj. R-squared:	0.681
Method:	Least Squares	F-statistic:	1037.
Date:	Fri, 25 Feb 2022	Prob (F-statistic):	0.00
Time:	15:31:21	Log-Likelihood:	-17709.
No. Observations:	1460	AIC:	3.543e+04
Df Residuals:	1456	BIC:	3.545e+04
Df Model:	3		
Covariance Type:	nonrobust		

$$\mathcal{H}_0: S = 0,\ K = 3$$

	coef	std err	t	P>|t|	[0.025	0.975]
const	-3.132e+04	3992.921	-7.844			+04
GrLivArea	65.6106	2.667	24.600			842
GarageCars	3.365e+04	1854.413	18.146			+04
TotalBsmtSF	50.4508	3.136	16.087	0.000	44.299	56.603

Omnibus:	520.280	Durbin-Watson:	1.978
Prob(Omnibus):	0.000	Jarque-Bera (JB):	31276.464
Skew:	-0.822	Prob(JB):	0.00
Kurtosis:	25.615	Cond. No.	6.64e+03

Тест на оценку нормальности остатков.

OLS Regression Results

Dep. Variable:	SalePrice	R-squared:	0.681
Model:	OLS	Adj. R-squared:	0.681
Method:	Least Squares	F-statistic:	1037.
Date:	Fri, 25 Feb 2022	Prob (F-statistic):	0.00
Time:	15:31:21	Log-Likelihood:	-17709.
No. Observations:	1460	AIC:	3.543e+04
Df Residuals:	1456	BIC:	3.545e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-3.132e+04	3992.921	-7.844	0.000	-3.92e+04	-2.35e+04
GrLivArea	65.6106	2.667	24.600			
GarageCars	3.365e+04	1854.413	18.146			
TotalBsmtSF	50.4508	3.136	16.087			

Omnibus:	520.280	Durbin-Watson:	1.978
Prob(Omnibus):	0.000	Jarque-Bera (JB):	31276.464
Skew:	-0.822	Prob(JB):	0.00
Kurtosis:	25.615	Cond. No.	6.64e+03

Число обусловленности, может использоваться для диагностики мультиколлинеарности.

Полезные ссылки

- https://www.k2analytics.co.in/multicollinearity-and-variance-inflation-factor/
- https://medium.com/swlh/interpreting-linear-regression-through-statsmodels-summary-4796d359035a

Спасибо за внимание!