Lecture 6
Linear Regression

Lecturer: Alisa Melikyan, amelikyan@hse.ru, PhD,
Associate Professor of the School of Software Engineering

Regression analysis

Regression analysis is a set of statistical methods used to estimate relationships between a dependent variable and one or more independent variables. It depicts how dependent variable will change when one or more independent variable changes. The results of regression analysis can be used to forecast the values of the dependent variable. One point to keep in mind with regression analysis is that causal relationships among the variables cannot be determined. While the terminology is such that we say that X "predicts" Y, we cannot say that X "causes" Y. Wile forecasting we should also consider how external factors could affect the values of the dependent variable.

Linear regression

Linear regression analysis rests on the assumption that the dependent variable is continuous. The independent variables can be either continuous or dichotomous. Independent variables with more than two levels can also be used in regression analyses, but they first must be converted into variables that have only two levels. This is called dummy coding.

Simple linear regression

The linear relationship between the variables is expressed using an equation of a straight line. In simple regression we predict an outcome based on a single predictor.

$$Y = a + b * X$$

where:

- X is an independent variable (predictor);
- Y is the dependent variable;
- a, b are constant values (model parameters).

The model parameters are determined using the least squares method.

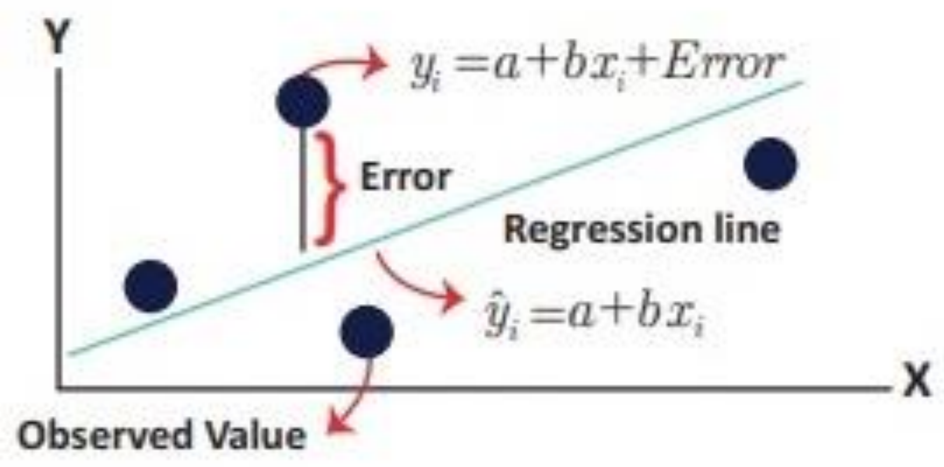
Method of least squares

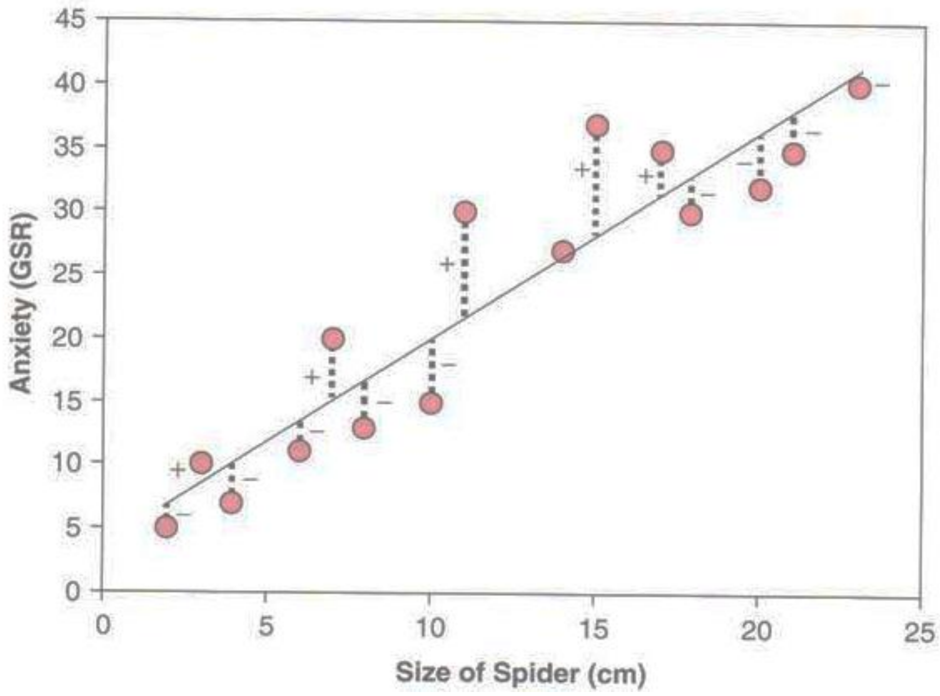
Method of least squares is a way of finding the line that best fits the data (i.e. the line that goes through, or is close to, as many of the data points as possible).

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

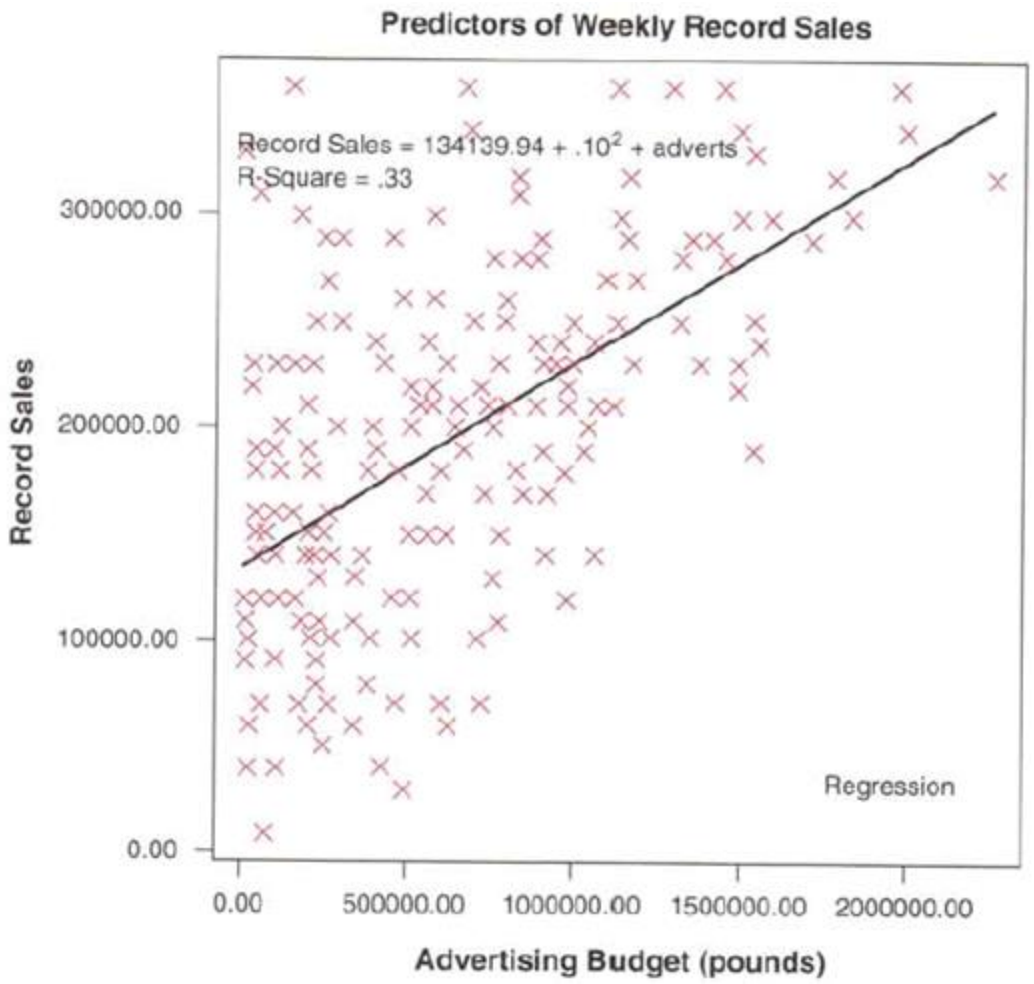
$$\text{i.e., } E(a,b)=\sum_{i=1}^{n}(y_i-a-bx_i)^2 .$$

Simple Linear Regression Model





This graph shows a scatterplot of some data with a line representing the general trend. The vertical lines (dashed) represent the differences (or residuals) between the line and the actual data. The method of least squares works by selecting the line that has the lowest sum of squared differences.



To evaluate how well a straight line expresses the relationship between variables, it is useful to examine the scatterplot to see if there is a linear relationship and to detect outliers that can significantly skew the result.

Sum of squares

Residual Sum of Squares (RSS)

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

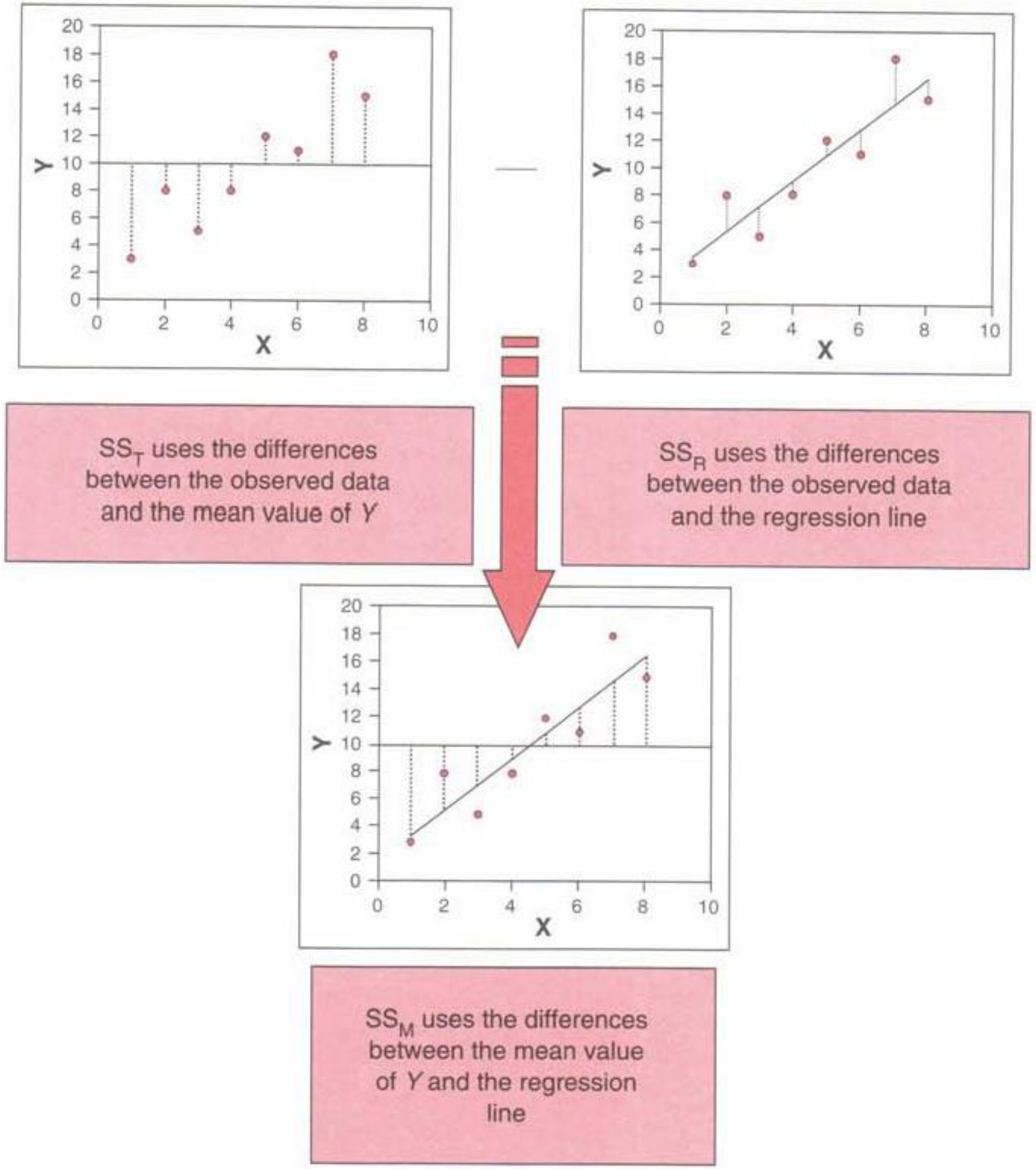
Total Sum of Squares (TSS)

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

Explained Sum of Squares (ESS)

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

TSS = RSS + ESS

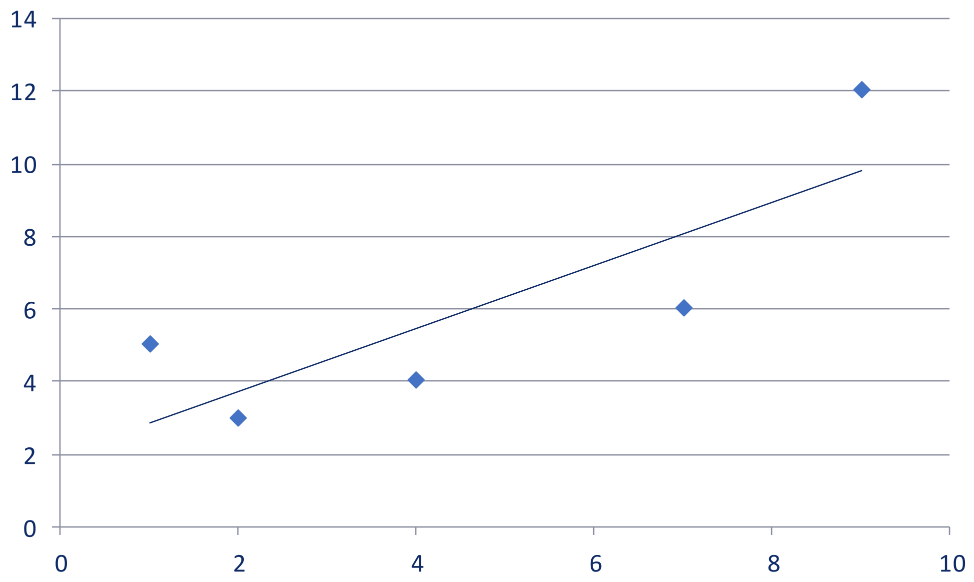

SS$_T$ uses the differences between the observed data and the mean value of Y
SS$_R$ uses the differences between the observed data and the regression line
SS$_M$ uses the differences between the mean value of Y and the regression line
X
Y

Simple linear regression: example

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

Multiple Linear Regression

The multiple regression model supposes that we have more than one predictor. Each predictor variable has its own coefficient, and the outcome variable is predicted from a combination of all the variables multiplied by their respective coefficients plus a residual term.

$$Y_i = (b_0 + b_1X_1 + b_2X_2 + \dots + b_nX_n) + e_i$$

Regression Equation

$$Y_i = (b_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n) + e_i$$

In the equation:

- b_1 is the coefficient of the first predictor (X_1);
- b_n is the coefficient of the n^{th} predictor (X_n);
- e_i is the difference between the predicted and the observed value of Y for the i^{th} participant.

Selection of predictors

The predictors included and the way in which they are entered into the model can have an impact on the result. It is recommended to select predictors based on past research (but it’s necessary to be sure that the past research was done appropriately). The selection could be also based of the theoretical framework of the research.

Adding categorical predictors into the model

If the variable is dichotomous, it could be added into the model without any preliminary transformation.

If the categorical variable has more than two values, it should be recoded into several dummy variables. Each group of cases, defined by the values of the categorical variable, should contain not less then 15% of cases.

Creating dummy variables

If the categorical variable has n values, we will enter into the model n-1 dichotomous variables.

One category is selected as a “basic” or “reference” and the other categories will be compared with this “basic” category. It could be a category, which contains the biggest number of cases.

Creating dummy variables

The initial variable "Educational level" takes the following values: no (1), secondary (2), higher (3), postgraduate (4). We can choose "higher" as a reference group.

Educational level
higher
no
secondary
postgraduate
higher
secondary

Ed_no	Ed_secondary	Ed_postgraduate
0	0	0
1	0	0
0	1	0
0	0	1
0	0	0
0	1	0

Coefficient of determination (R-squared)

R-squared is the proportion of variation in the dependent variable that is accounted for by the model. Value of the coefficient varies from 0 to 1. An R-squared of 1 indicates that the regression predictions perfectly fit the data.

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

R-squared

Regression models with low R-squared values can be perfectly good models. Some fields of study have an inherently greater amount of unexplainable variation. In these areas R^2 values are bound to be lower. For example, studies that try to explain human behavior generally have R^2 values less than 50%. People are just harder to predict than things like physical processes. If we have a low R-squared value but the independent variables are statistically significant, we can still draw important conclusions about the relationships between the variables. There is a scenario where small R-squared values can cause problems. If we need to generate predictions that are relatively precise, a low R^2 can be a showstopper.

Adjusted R-squared

Adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases when the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected. Adjusted R-squared could be used to compare models that have a different number of variables.

$$\text{adjusted } R^2 = 1 - \left[\left(\frac{n-1}{n-k-1}\right)\left(\frac{n-2}{n-k-2}\right)\left(\frac{n+1}{n}\right)\right](1 - R^2)$$

where n — number of cases, k — number of predictors

F-statistics

The F-test compares the current model with zero predictor variables (the intercept only model) and decides whether the added coefficients significantly improved the model. If we get a significant result, then whatever coefficients we included in the model improved the model's fit.

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

Mean Absolute Error (MAE)

In the fields of statistics and machine learning, the Mean Absolute Error (MAE) is a frequently employed metric. It’s a measurement of the typical absolute discrepancies between a dataset’s actual values and projected values.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |x_i - y_i|$$

Where:

- x_i represents the actual or observed values for the i-th data point.
- y_i represents the predicted value for the i-th data point.

Mean Squared Error (MSE)

A popular metric in statistics and machine learning is the Mean Squared Error (MSE). It measures the square root of the average discrepancies between a dataset's actual values and projected values. MSE is frequently utilized in regression issues and is used to assess how well predictive models work.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2$$

where:

- x_i represents the actual or observed value for the i-th data point.
- y_i represents the predicted value for the i-th data point.

Root Mean Squared Error (RMSE)

RMSE is a usually used metric in regression analysis and machine learning to measure the accuracy or goodness of fit of a predictive model, especially when the predictions are continuous numerical values. The RMSE quantifies how well the predicted values from a model align with the actual observed values in the dataset.

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

Where:

- RMSE is the Root Mean Squared Error.
- x_i represents the actual or observed value for the i-th data point.
- y_i represents the predicted value for the i-th data point.

Evaluating the accuracy of the regression model

When we have produced a model based on a sample of data two important questions should be considered:

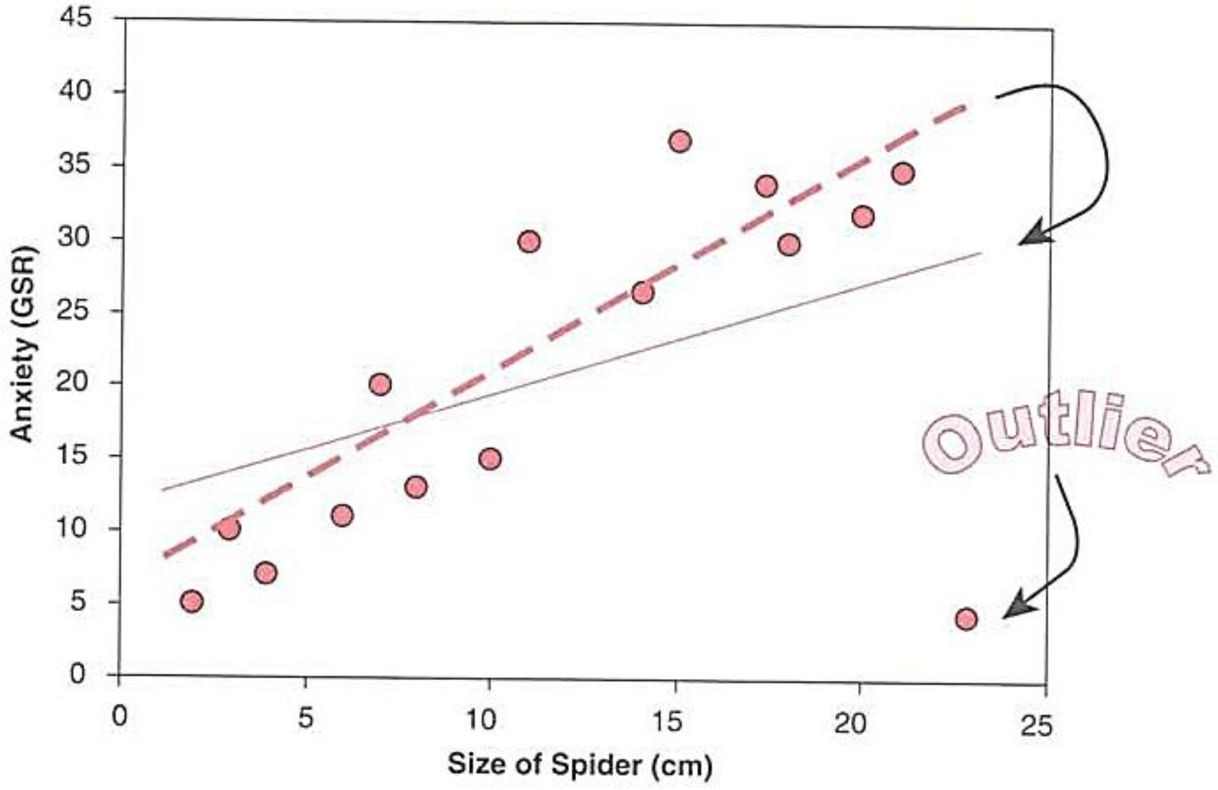
1. Does the model fit the observed data well, or it’s influenced by a small number of cases?
2. Can the model generalize to other samples?

Diagnostics of the model: outliers and influential cases

An **outlier** is a case which differs substantially from the main trend of the data. Outliers can cause the model to be biased because they affect the values of the estimated regression coefficients.

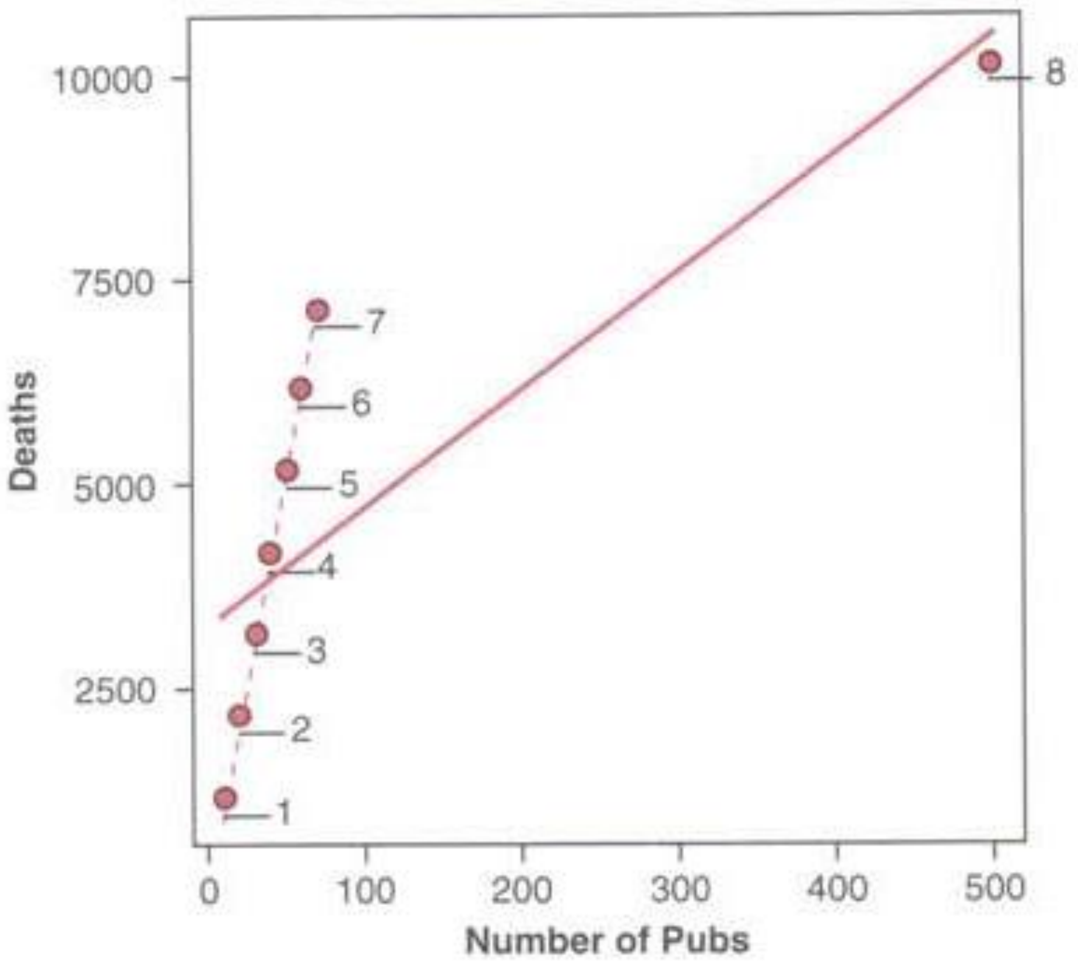
An **influential case** is a case which has serious influence over the parameters of the model. If we delete it the regression coefficients will change.

Outliers



The change in one point had a dramatic effect on the regression model: the gradient reduced (the line becomes flatter) and the intercept increases (the line crosses the Y-axis at a higher point).

Influential cases

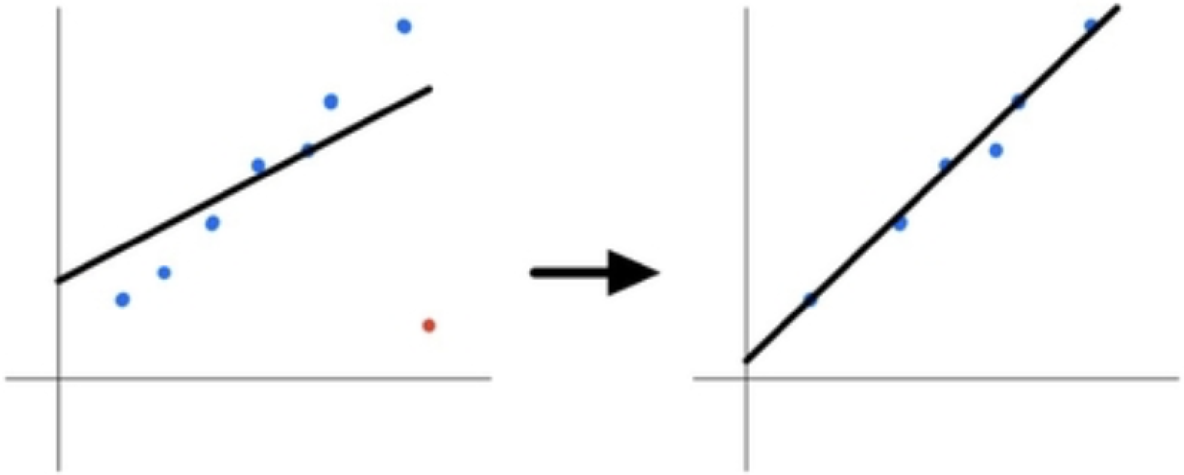


An influential case is any case that significantly alters the value of a regression coefficient whenever it is deleted from an analysis. If the deletion of particular cases in an analysis alters the parameters of the regression equation significantly, then these cases represent influential cases.

How to identify an influential case?

Influential statistics: DfBeta(s) shows how the regression coefficients change if the influential case is excluded.

The case could be an influential if the values of these statistics are greater than 1.



Diagnostics of the model: residuals

Residuals are the differences between the values of the outcome observed in the sample and the values of the outcome predicted by the model. If the model fits the sample data well then all residuals will be small and their distribution will be not different from normal. If a particular case has a large residual, then it could be an outlier. Potential outlier is a case with standardized residual greater than 3 or less than -3.

Multicollinearity

Multicollinearity is the occurrence of high intercorrelations among two or more independent variables in a multiple regression model. Multicollinearity can lead to skewed or misleading results when a researcher or analyst attempts to determine how well each independent variable can be used most effectively to predict or understand the dependent variable in a statistical model. It is a situation when there is a strong correlation (r > 0.7) between two or more predictors in a regression model.

VIF (variance-inflation factor) indicates whether a predictor has a strong linear relationship with the other predictor(s). If VIF is greater than 5 there could be a multicollinearity.

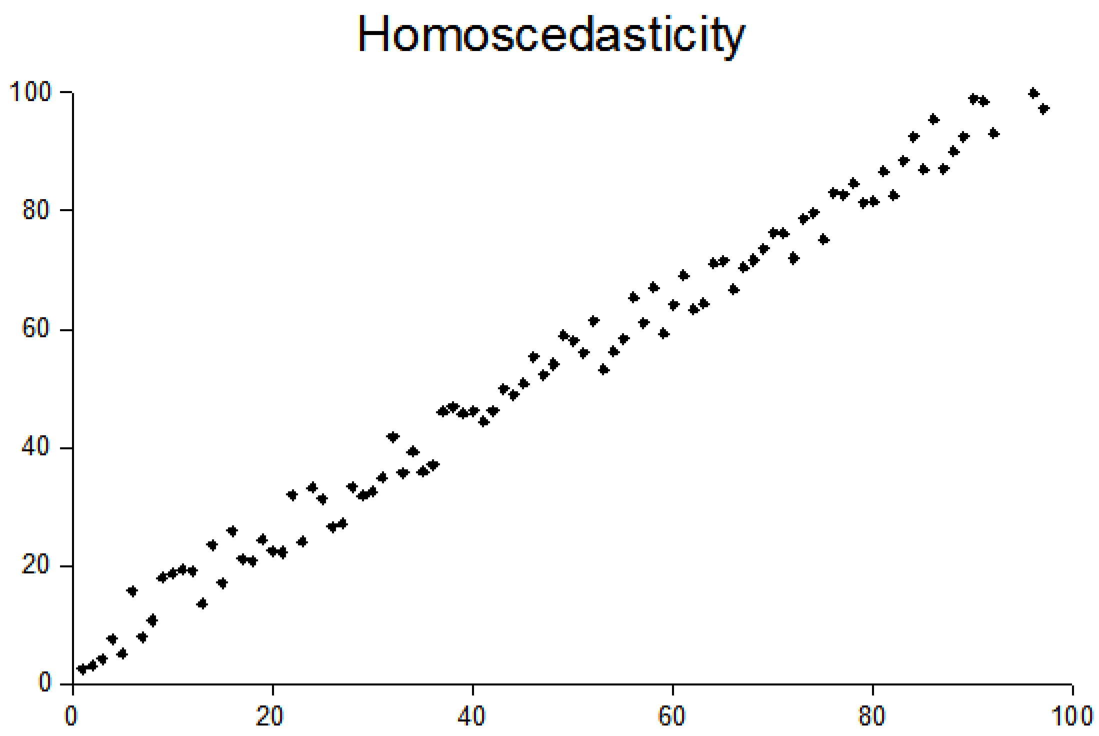
$$VIF\ x_i = \frac{1}{Tolerance} = \frac{1}{1 - R_i^2}$$

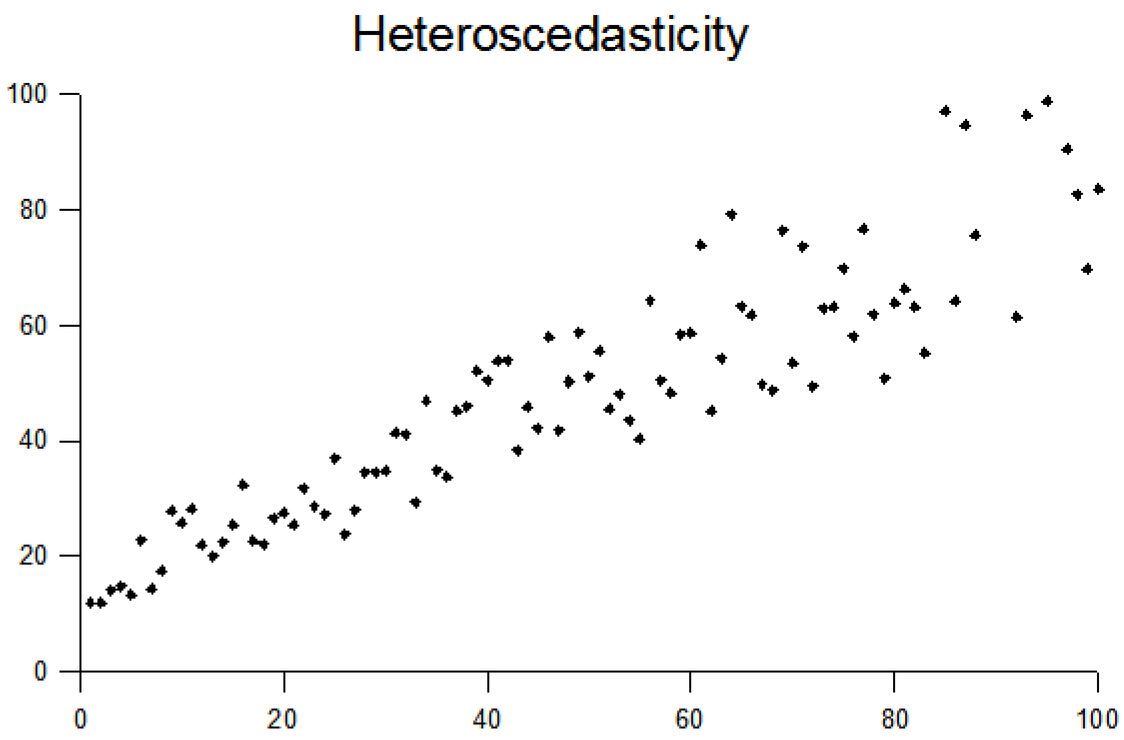
Heteroscedasticity

It is supposed that in good regression models the variance of the residuals is homogeneous across levels of the predicted values (homoscedasticity). If the model is well-fitted, there should be no pattern to the residuals plotted against the fitted values. If the variance of the residuals is non-constant, then there is a heteroscedasticity.

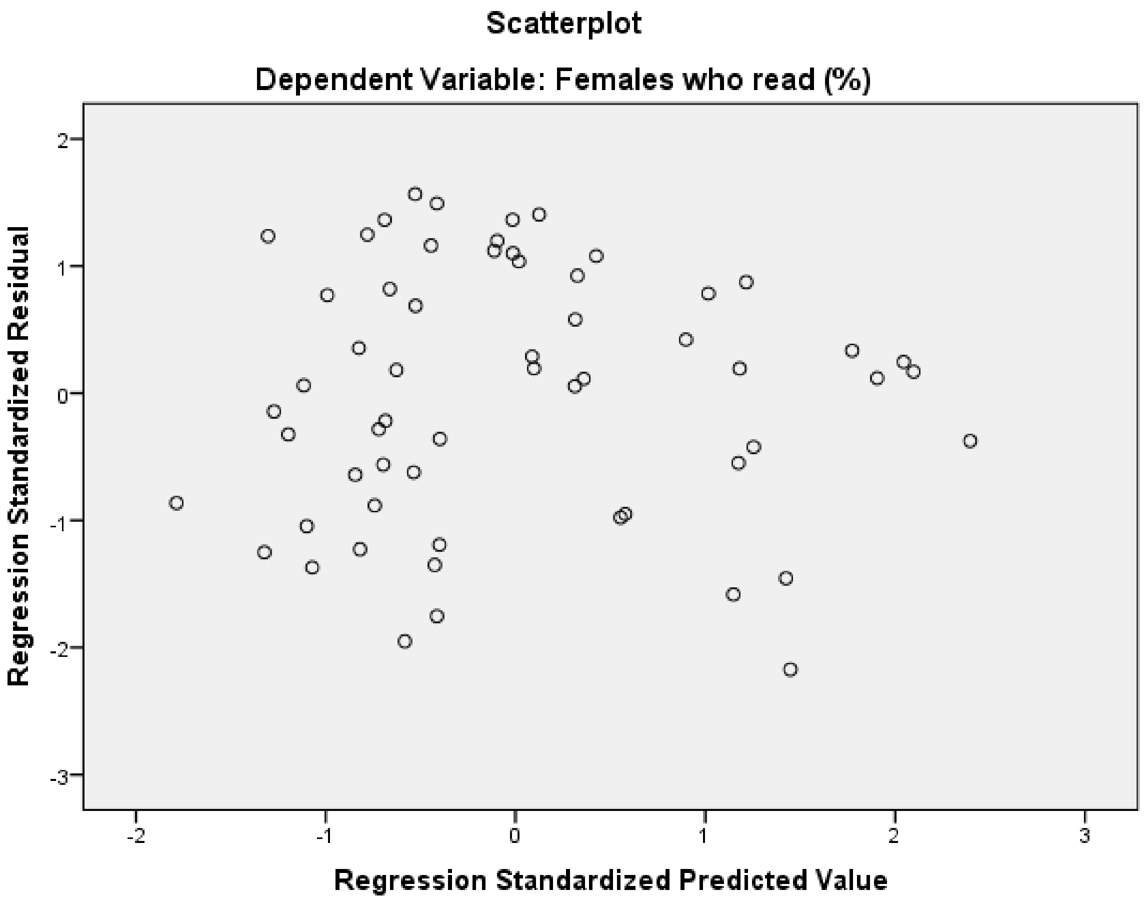
The graphical analysis could detect heteroscedasticity. A commonly used graphical method is to use the plot to show the residuals versus fitted (predicted) values.

Heteroscedasticity in a simple regression model





Heteroscedasticity in a multiple regression model



Gauss Markov Theorem

The **Gauss Markov theorem** tells us that if a certain set of assumptions are met, the ordinary least squares estimate for regression coefficients gives you the best linear unbiased estimate possible.

There are five Gauss Markov assumptions:

- **Linearity**: the parameters we are estimating using the OLS method must be themselves linear.
- **Random**: our data must have been randomly sampled from the population.
- **Non-Collinearity**: the regressors being calculated aren’t perfectly correlated with each other.
- **Exogeneity**: the regressors aren’t correlated with the error term.
- **Homoscedasticity**: no matter what the values of our regressors might be, the error of the variance is constant.

Akaike Information Criterion (AIC)

A common way to compare models is by using the so-called information criterion. It is a way to balance bias and variance or accuracy (fit) and simplicity (parsimony).

$$AIC_p = n * ln(\frac{SSE_p}{n}) + 2 * p$$

p is the number of estimated parameters (including the constant),

n is the number of observations,

SSE is the residual sum of squares.

The smaller the AIC the better. A model is going to be better when the sample size is larger, the unexplained variance is lower and we use the fewer parameters. AIC is a relative measure that compares one model to another to choose the one that loses less information. It's NOT a measure of how good a model is.

```
                            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
```

Constant – value of the dependent variable if all the predictors are equal to zero

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.
	460	AIC:	3.543e+04
	456	BIC:	3.545e+04
	3		
	ust		

Regression coefficient for a certain variable means that one-unit increase in variable value will lead to an increase in the value of the dependent variable by 65.6 units

	coef	std err	t	P>|t|	[0.025	0.975]
const	-[illegible].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:
Covariance Type:            nonr
==============================================================================
                 coef    std er
------------------------------------------------------------------------------
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
```

Regression coefficients

Regression Equation: 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. Observations:               1460                                 .543e+04
Df Re                           1456                                 .545e+04
Df Mo                              3
Covar                      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
```

Standard errors of regression coefficients

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

```
                            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. Observ                    1460      AIC:                         3.543e+04
Df Residua                    1456      BIC:                         3.545e+04
Df Model:                        3
Covariance                  robust
==============================================================================
                 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
```

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

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              SalePrice   R-squared:                       0.681
Model:                            OLS   Adj. R-squared:                  0.681
Method:                                                                    37.
Date:                                                                     .00
Time:                                                                     09.
No. Observations:                                                    .545e+04
Df Residuals:                                                        3.545e+04
Df Model:
Covariance Type:
================================================================================
                 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
```

If a variable significantly predicts an outcome, then it should have a b-value significantly different from zero. This hypothesis is tested using a t-test.

Statistical significance of regression coefficients

CI = $\hat{\beta}_j \pm t_c \times S_{\hat{\beta}_j}$

- $\hat{\beta}_j$ — estimated regression coefficient
- t_c — Critical t-value
- $S_{\hat{\beta}_j}$ — Standard error of regression coefficient

95% confidence interval

```
                            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
Df Residuals:                           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
```

```
                            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
```

R-squared, the model explains 68% of variation of the dependent variable’s values

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

The model is statistically significant

```
                            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
```

Result of normality test for residuals: residuals are not normally distributed

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			-7.844	0.000	-3.92e+04	-2.35e+04
GrLiv			24.600	0.000	60.379	70.842
Garag			18.146	0.000	3e+04	3.73e+04
Total			16.087	0.000	44.299	56.603

Skewness and Kurtosis calculated for residuals

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:                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
```

Homoscedasticity test. If the values vary from 1 to 2 there is a homoscedasticity.

```
                            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                               +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
```

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

Normality test for the residuals

Useful links

- https://dss.princeton.edu/online_help/analysis/regression_intro.htm
- https://datatofish.com/statsmodels-linear-regression/
- https://mlu-explain.github.io/linear-regression/

Thank you for your attention!