Biostatistics
Week X

Ege Ülgen, MD, PhD

8 December 2022

ACIBADEM
MEHMET ALİ AYDINLAR
ÜNİVERSİTESİ

Regression Analysis

- Regression analysis is used primarily to **model causality** and **provide prediction**
- Predict the values of a **dependent** (response) variable based on values of at least one **independent** (explanatory) variable
- Explain the **effect** of the independent variables on the dependent variable

Regression Analysis

- Regression can be used to
  - Understand the relationship between variables
  - Predict the value of one variable based on other variables
- Examples:
  - Quantifying the relative impacts of age, gender, and diet on BMI
  - Predicting whether the treatment will be successful or not

Regression Analysis

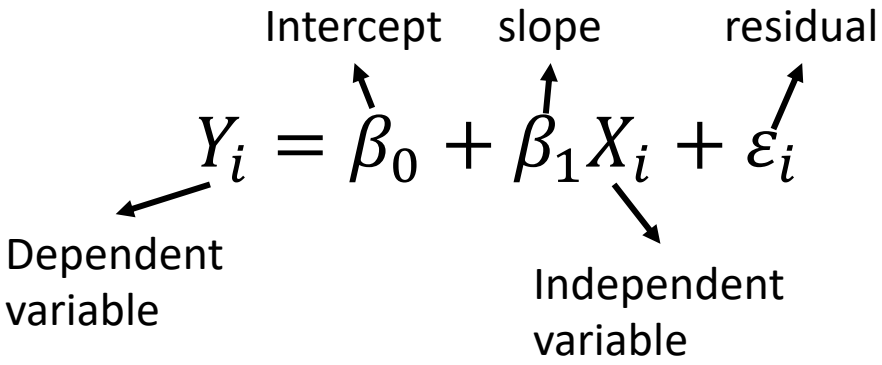
- The variable to be predicted is called the **dependent variable**
  - Also called the **response variable**
- The value of this variable depends on the value of the **independent variable(s)**
  - Also called the **explanatory** or **predictor variable(s)**

Simple Linear Regression

E.g., quantifying the impact of age on BMI

- Linear regression is a method for estimating the **linear relationship** between the dependent and independent variables

- Relationship between variables is described by a linear function



- The coefficients are estimated by minimizing the sum of the squared errors/residuals (Least squares)

James G, Witten D, Hastie T, Tibshirani R. Linear regression. In: James G, Witten D, Hastie T, Tibshirani R, editors. An Introduction to Statistical Learning: with Applications in R [Internet]. New York, NY: Springer; 2013 [cited 2021 May 7]. p. 59–126. (Springer Texts in Statistics). Available from: https://doi.org/10.1007/978-1-4614-7138-7_3

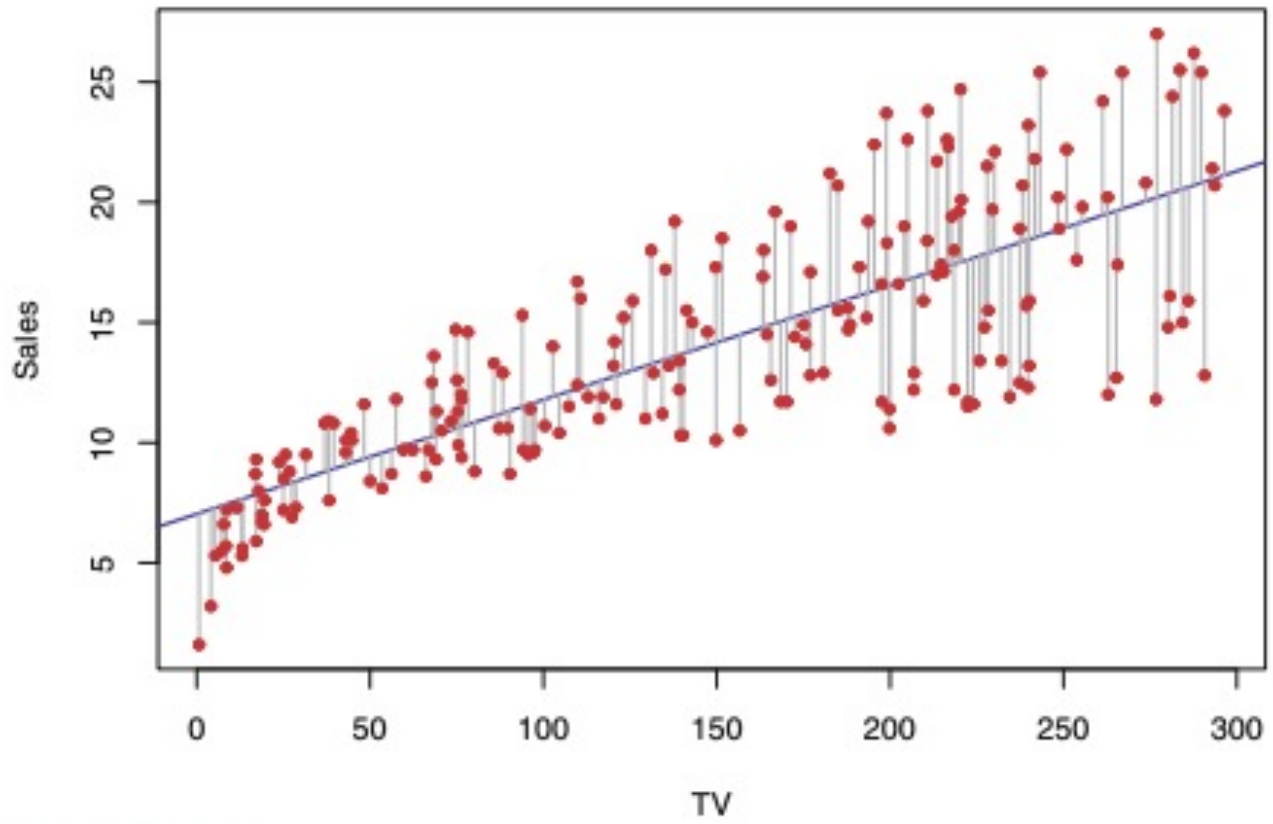
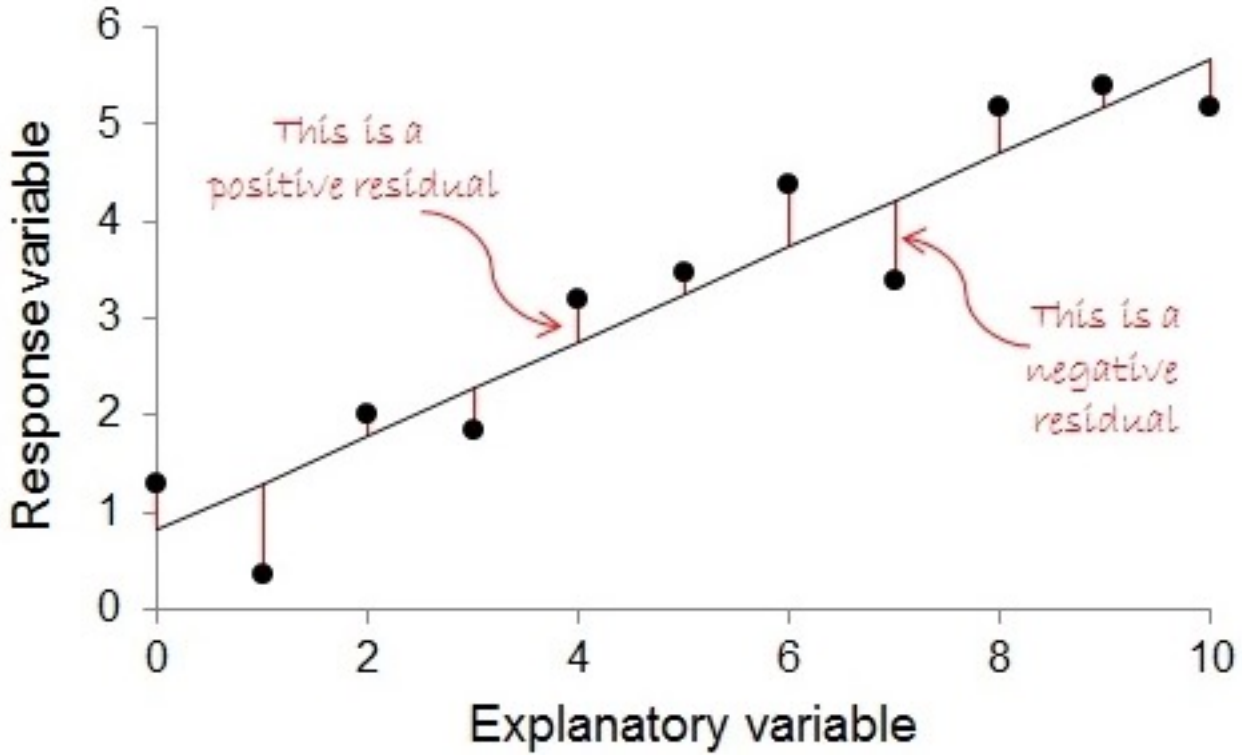


FIGURE 3.1. *For the* `Advertising` *data, the least squares fit for the regression of* `sales` *onto* `TV` *is shown. The fit is found by minimizing the sum of squared errors. Each grey line segment represents an error, and the fit makes a compromise by averaging their squares. In this case a linear fit captures the essence of the relationship, although it is somewhat deficient in the left of the plot.*

Error/residual = Actual value – Predicted value



Kumar A. Linear regression explained with python examples [Internet]. Data Analytics. 2020 [cited 2021 Oct 9]. Available from: https://vitalflux.com/linear-regression-explained-python-sklearn-examples/

Estimators

The coefficients are estimated by the **Least-squares regression line** (LSRL) is the line that minimizes the sum of squared errors (SSE)

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x$$

$$SSE = (y - \hat{y})^2$$

$$\widehat{\beta_1} = r \frac{s_y}{s_x}$$

$$\widehat{\beta_0} = \bar{y} - \beta_0 \bar{x}$$

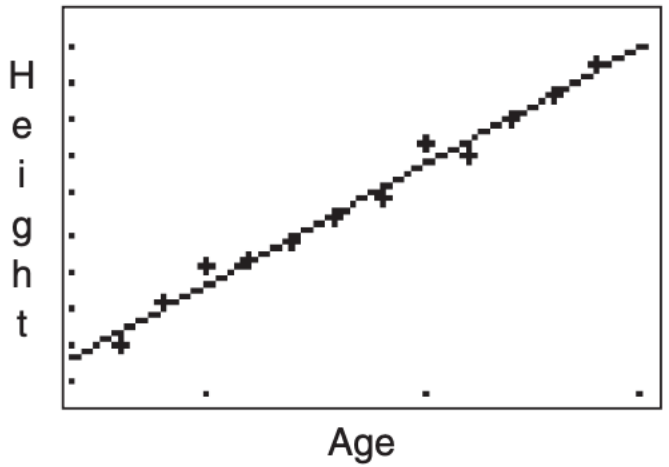
example: The data given below show the height (in cm) at various ages (in months) for a group of children.

(a) Does a line seem to be a good model for the data? Explain.

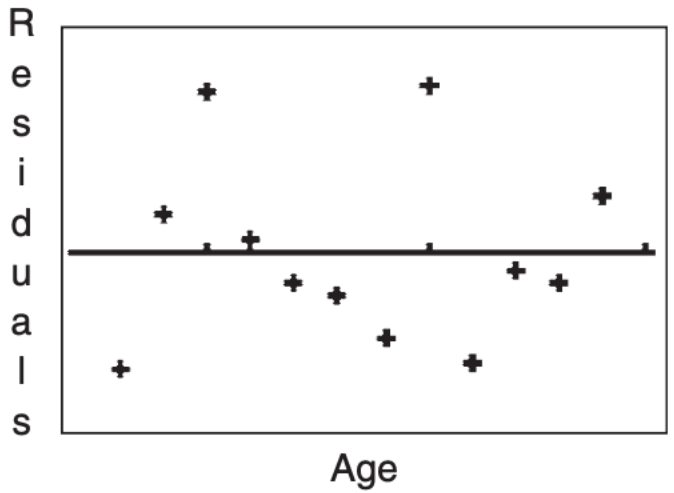
(b) What is the value of the residual for a child of 19 months?

Age	18	19	20	21	22	23	24	25	26	27	28	29
Height	76	77.1	78.1	78.3	78.8	79.4	79.9	81.3	81.1	82.0	82.6	83.5

(a)



Scatter plot and LSRL for predicting height from age.



Scatter plot of residuals vs. age.

example: The data given below show the height (in cm) at various ages (in months) for a group of children.

(a) Does a line seem to be a good model for the data? Explain.

(b) What is the value of the residual for a child of 19 months?

Age	18	19	20	21	22	23	24	25	26	27	28	29
Height	76	77.1	78.1	78.3	78.8	79.4	79.9	81.3	81.1	82.0	82.6	83.5

(b)

LSRL:

$\hat{y} = 64.94 + 0.634(x)$

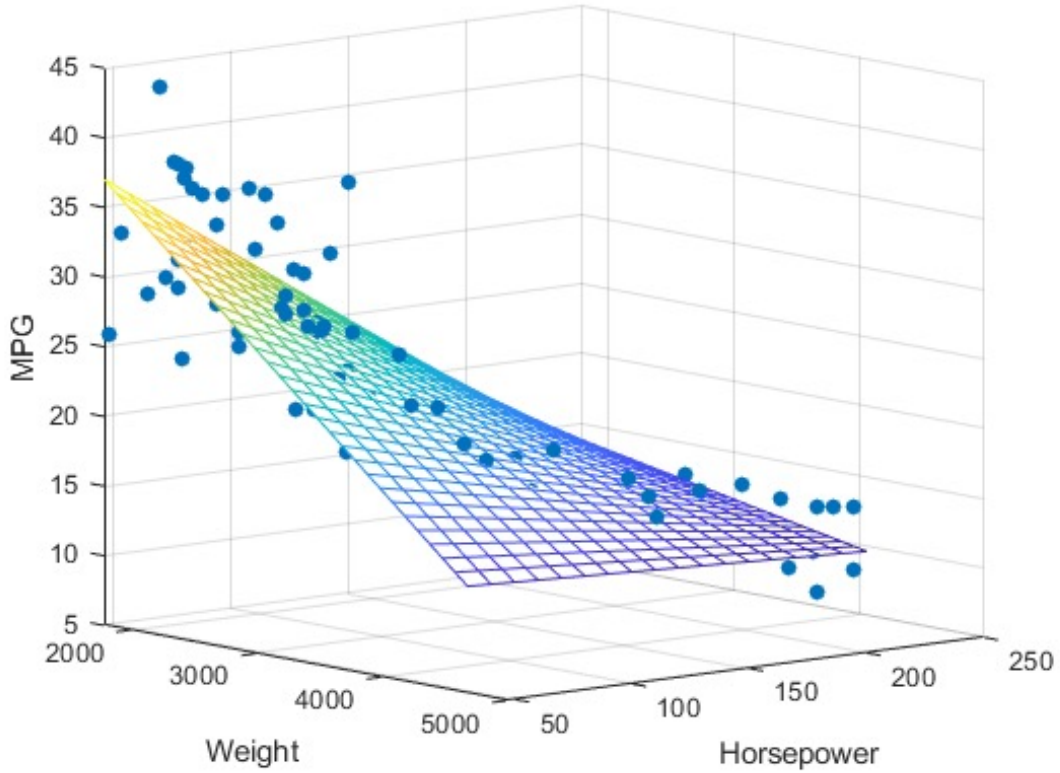
Predicted height = 64.9 + 0.634 * 19 = 76.946 cm

Multiple Linear Regression

E.g., quantifying the relative impacts of age, gender, and diet on BMI

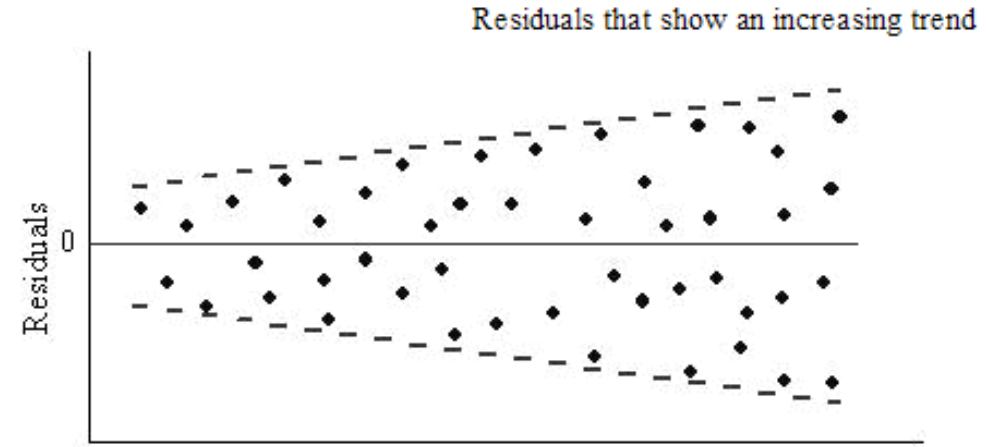
$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

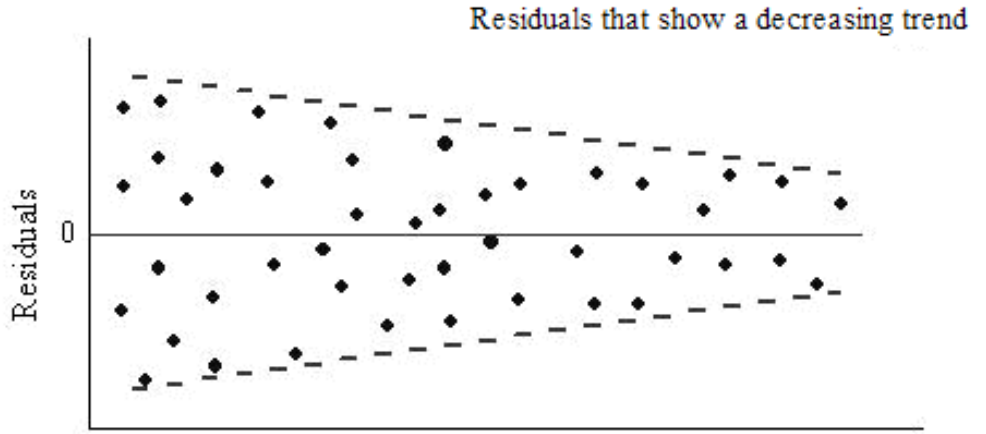
where *Y* is the dependent variable, X_1 to X_p are *p* independent variables, β_0 to β_p are the coefficients, and ε is the error term

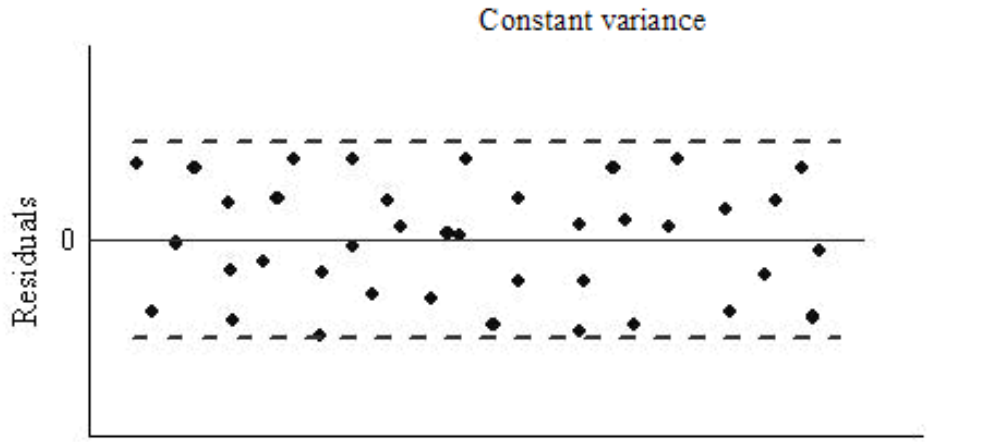


Linear Regression Assumptions

- There is a **linear relationship** between the independent and dependent variables
- **Normality** – (Q-Q plot / Shapiro-Wilk test)
  - Y values are normally distributed for each X
  - Residuals are normally distributed
- Homoscedasticity (**constant variance**) of the residuals
- **Independence of observations**

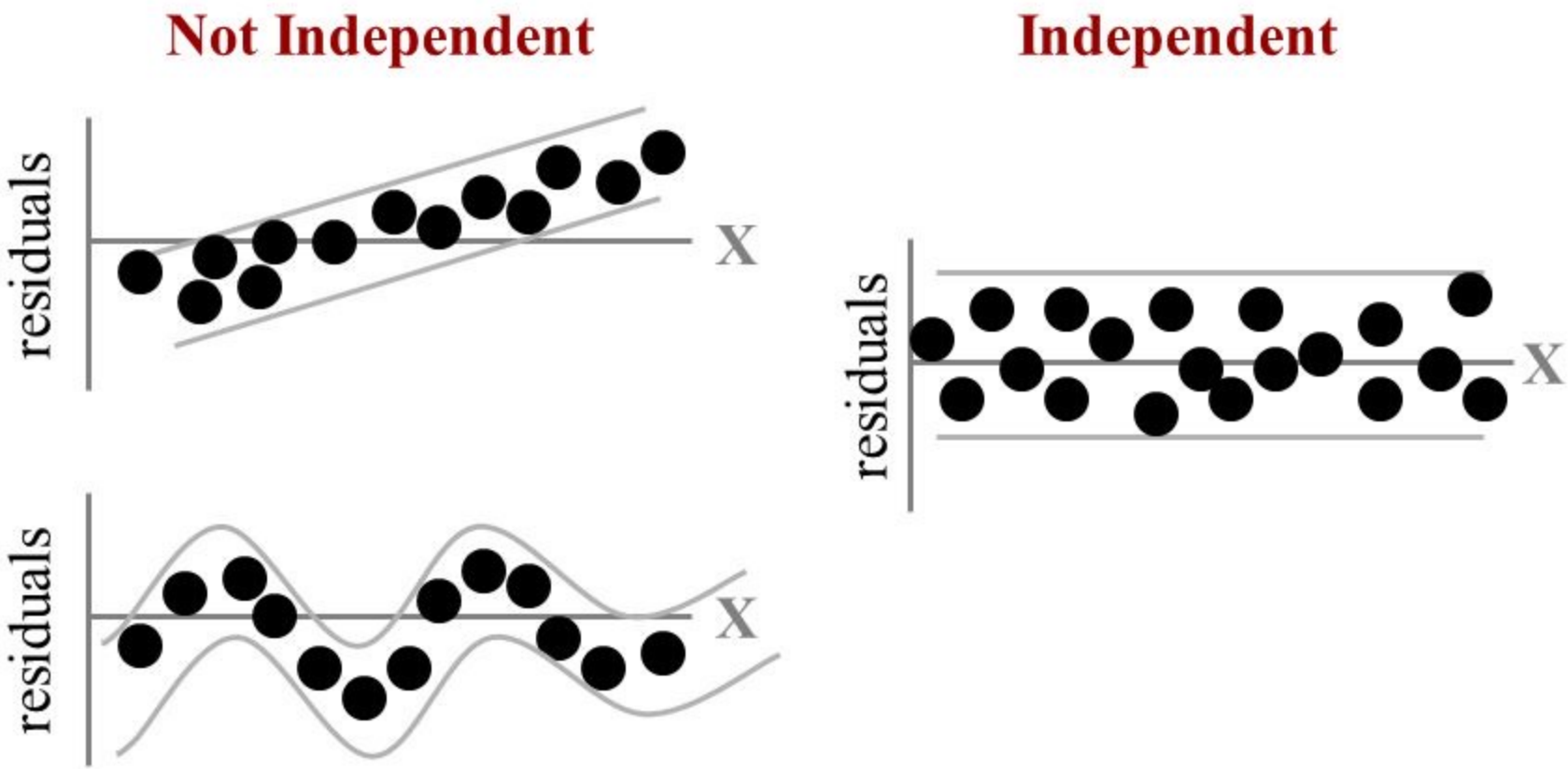






ww.originlab.com/doc/Origin-Help/Residual-Plot-Analysis

Residual Analysis for Independence



Linear Regression - Example

Prognostic factors for body fat

- Number of observed individuals: n = 241
- Dependent variable: body fat = percental body fat
- We are interested in the influence of three independent variables:
  - BMI in kg/m2
  - Waist circumference (abdomen) in cm.
  - Waist/hip-ratio

https://www.ebpi.uzh.ch/dam/jcr:ffffffff-c1f2-5119-ffff-ffffacc68d53/slides-04.pdf

Prognostic factors for body fat - Simple Linear Regression Models

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-27.617	2.939	-9.398	0.000
bmi	1.844	0.116	15.957	0.000

BMI: $R^2 = 0.516$, $R^2_{adj} = 0.514$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-42.621	2.869	-14.855	0.000
abdomen	0.668	0.031	21.570	0.000

Abdomen: $R^2 = 0.661$, $R^2_{adj} = 0.659$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-78.066	5.318	-14.680	0.000
waist_hip_ratio	104.976	5.744	18.275	0.000

Waist/hip-ratio: $R^2 = 0.583$, $R^2_{adj} = 0.581$

Prognostic factors for body fat - Multiple Linear Regression

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-60.045	5.365	-11.192	0.000
bmi	0.123	0.236	0.519	0.605
abdomen	0.438	0.105	4.183	0.000
waist_hip_ratio	38.468	10.262	3.749	0.000

$R^2 = 0.681,\ R^2_{\text{adj}} = 0.677$

https://www.ebpi.uzh.ch/dam/jcr:ffffffff-c1f2-5119-ffff-ffffacc68d53/slides-04.pdfc

Example - Prognostic factors for body fat - Multiple Linear Regression

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-60.045	5.365	-11.192	0.000
bmi	0.123	0.236	0.519	0.605
abdomen	0.438	0.105	4.183	0.000
waist_hip_ratio	38.468	10.262	3.749	0.000

$R^2 = 0.681, \boxed{R^2_{\text{adj}} = 0.677}$

the proportion of the variation in the dependent variable that is predictable from the independent variable

$$Estimated\ Body\ Fat = -60.045 + 0.123 * bmi + 0.438 * abdomen + 38.468 * waist\_hip\_ratio$$

Example - Prognostic factors for body fat - Multiple Linear Regression

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-60.045	5.365	-11.192	0.000
bmi	0.123	0.236	0.519	0.605
abdomen	0.438	0.105	4.183	0.000
waist_hip_ratio	38.468	10.262	3.749	0.000

- For a person with bmi = 0, abdomen = 0, waist_hip_ratio = 0, the boy fat is estimated to be -60.045 ($p < 0.001$)
- (Keeping all other variables the same) with one unit increase in bmi, body fat increases by 0.123 (not significant since $p > 0.05$)
- With 95% confidence, it can be stated that with one unit increase in abdomen, body fat increases by 0.438 ($p < 0.001$)
- With one unit increase in waist_hip_ratio, body fat increases by 38.468 ($p < 0.001$)

Example II

- We'll analyze the prostate cancer dataset
- The main aim of collecting this data set was to inspect the associations between **prostate-specific antigen (PSA)** and **prognostic clinical measurements** in men advanced prostate cancer
- Data were collected on 97 men who were about to undergo radical prostectomies

**PSA was transformed to logPSA for "normalization"*

Example II – Model 1

$$logPSA = 1.8 + 0.07 * \boldsymbol{vol} + 0.77 * I(\boldsymbol{invasion} = \mathbf{1})$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.8035	0.1141	15.81	<0.001
vol	0.0725	0.0133	5.43	<0.001
invasion1	0.7755	0.2541	3.05	0.003

Adjusted R-squared: 0.472

Example II – Model 2

$$logPSA = 1.67 + 0.1021 * \boldsymbol{vol} + 1.326 * I(\boldsymbol{invasion} = \mathbf{1}) - 0.056 * I(\boldsymbol{invasion} = \mathbf{1}) * \boldsymbol{vol}$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.6673	0.1289	12.94	<0.001
vol	0.1021	0.0191	5.35	<0.001
invasion1	1.326	0.3588	3.7	<0.001
vol:invasion1	-0.056	0.0262	-2.13	0.0354

Adjusted R-squared: 0.491

For a patient with invasion, there is an additional -0.056 change in PSA when vol changes one unit
= For a patient with invasion, one unit change in volume results in (0.1021 – 0.056) change in PSA

Example II – Model 3

$$logPSA = 1.55 + 0.076 * \boldsymbol{vol} + 0.45 * I(\boldsymbol{Gleason = 7}) + 0.9 * I(\boldsymbol{Gleason = 8})$$

(compared to **Gleason = 6**)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.5523	0.1548	10.02	< 2e-16
vol	0.0758	0.0131	5.79	9.30E-08
Gleason7	0.4521	0.1928	2.34	0.0212
Gleason8	0.9043	0.2747	3.29	0.0014

Adjusted R-squared: 0.48

Brief Summary

- Regression
  - Understand the relationship between variables
  - Predict the value of one variable based on other variables
- Linear regression is a method for estimating the linear relationship between the dependent and independent variables