

05 / 06. Linear Regression

Data Analysis for Networks - NDA'20
Anastasios Giovanidis

Sorbonne-LIP6



October 28, 2020

Bibliography

- B.1 Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. "An introduction to statistical learning: with applications in R". Springer Texts in Statistics.

👉 Chapter 3

ISBN 978-1-4614-7137-0 (DOI 10.1007/978-1-4614-7138-7)

- B.2 H. Pishro-Nik, "Introduction to probability, statistics, and random processes", available at <https://www.probabilitycourse.com>, Kappa Research LLC, 2014.

Intro

Linear Regression: A very simple approach for Supervised Learning:

- been around for a very long time...
- predicts a quantitative response.
- explains the relationship between two or more variables.

Many fancy statistical learning approaches can be seen as generalisation or extensions of linear regression.

This lecture: **Linear Regression, Least-Squares**

Some history

- ▶ (1875) [Sir Francis Galton](#) originally conceived modern notions of correlation and regression, by studying eugenics of sweet pea seeds.
- ▶ (1896) [Karl Pearson](#) publishes the first rigorous treatment of correlation and regression in the *Philosophical Transactions of the Royal Society of London*.
- ▶ (1981) [E.E. Ghiselli](#) presents a simple proof of optimality of regression related to the sum of squared errors.

Source

Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors, by Jeffrey M. Stanton (2017)

It all started like this...

A. Giovanidis 2020



Training data available:

$$D_n = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$$

Data come in pairs (\mathbf{x}_i, y_i) of

- ▶ $\mathbf{x}_i = (x_{i,1}, \dots, x_{i,K})$ **input data**, as a vector of size $K \geq 1$.
- ▶ y_i **output data** of size 1.

Goal: Given set of known I/O pairs, "guess" a function $f : \mathbb{R}^K \rightarrow \mathbb{R}$ that for any input $\mathbf{x}^{(o)}$ predicts its output $y^{(o)}$.

☞ We hope to get a "good" prediction. What is "good"?

We are looking for an expression

$$y = f(x_1, \dots, x_K) + \epsilon$$

which can express the output y as a deterministic function of the input $\mathbf{x} = (x_1, \dots, x_K)$ plus an additive random perturbation ϵ from some distribution (usually white noise $\mathcal{N}(0, \sigma^2)$).

☞ In this way we can do **predictions** of unknown output, given any input.

We will try to **estimate** the function $f(x_1, \dots, x_K)$ using the existing **training** data.

Linear Regression

Linear Regression: **Assumes** that the output is an **affine** function of the input

$$y = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_K x_K + \beta_0 + \epsilon. \quad (1)$$

Unknowns:

- $(\beta_1, \dots, \beta_K) \in \mathbb{R}^K$ is a vector of *parameters/coefficients*.
- β_0 is the *bias parameter or intercept*.
- ϵ is a **random error term**; independent of \mathbf{x} with mean 0.

☞ Using known pairs, find the Regression line = “line of best fit”.

Simple Linear Regression

Let us simplify for just 1-dimension, $x \in \mathbb{R}$.

$$y = \beta_1 x + \beta_0 + \epsilon \quad (2)$$

- ▶ β_1 is the **slope** (average increase in y for unit increase of x).
- ▶ β_0 is the **intercept term** (expected value of y when $x = 0$),
- ▶ ϵ is the **error term**, which summarises what we miss by using a simple linear model, e.g. non-linearities, other variables, or measurement errors.

Simple Linear Regression, cont'd

Simple Goal: Use training data to produce estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for the model coefficients. Then we can make the prediction \hat{y}_o for y_o , on the basis of input x_o :

$$\hat{y} = \hat{\beta}_1 x + \hat{\beta}_0, \quad (3)$$

☞ Draw a line in the $x - y$ plane that best "fits" our data points. This is called the **regression line** (without the error term).

Examples

A. Giovanidis 2020

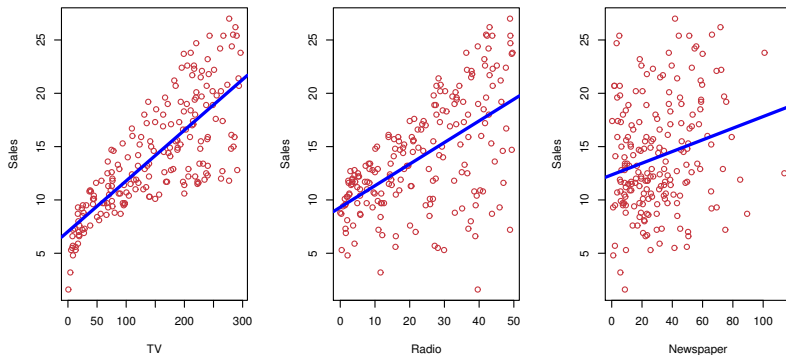


Figure: Advertisement Related Example. ¹

¹Figure from [B.1] with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani

How to draw the line?

We need to draw a line that "fits well" the **known data**. How?

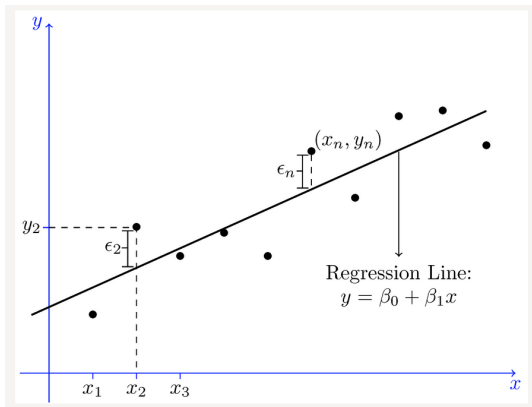
- ▶ The available known data are the pairs (x_i, y_i) , $i = 1, \dots, n$.
- ▶ The regression line gives $\hat{y}_i = \hat{\beta}_1 x_i + \hat{\beta}_0$, for input x_i .

Definition: the following quantities are called **residuals**

$$\epsilon_i := y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \quad . \quad (4)$$

Residuals

A. Giovanidis 2020



²Source [B.2]

Sum of Squares

The *Residual Sum of Squares (RSS)* is a function of $\hat{\beta}_0, \hat{\beta}_1$,

$$RSS(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^n \epsilon_i^2 := \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

It is always non-negative $RSS \geq 0$.

👉 Find and use the coefficients that **minimize** the RSS! This should provide a good fit to the data.

Least Squares

The optimal *least squares coefficient estimates* are found by:

$$\min \text{RSS}(\hat{\beta}_0, \hat{\beta}_1).$$

Solution:

$$\frac{\partial \text{RSS}}{\partial \hat{\beta}_0} = 2(-1) \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0,$$

$$\frac{\partial \text{RSS}}{\partial \hat{\beta}_1} = 2(-1) \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0.$$

Coefficient Estimates

Given $n \geq 3$ observations $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, we estimate β_0 and β_1 as

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad (5)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \quad (6)$$

where $\bar{y} := \frac{1}{n} \sum_{i=1}^n y_i$ and $\bar{x} := \frac{1}{n} \sum_{i=1}^n x_i$ are the **sample means**.

Alternative Method

Consider again the linear model with random errors.

Error ϵ has zero mean $\mathbb{E}[\epsilon] = 0$ and is independent of x .

$$y = \beta_1 x + \beta_0 + \epsilon$$

👉 Applying expectation on both sides, we get

$$\begin{aligned}\mathbb{E}[y] &= \beta_1 \mathbb{E}[x] + \beta_0 + \mathbb{E}[\epsilon] \\ &= \beta_1 \mathbb{E}[x] + \beta_0.\end{aligned}$$

Alternative Method cont'd

A. Giovanidis 2020

☞ We take **covariance** between x and $y = \beta_0 + \beta_1 x + \epsilon$.

$$\begin{aligned}\text{Cov}(x, y) &:= \mathbb{E}[(x - \mathbb{E}[x])(y - \mathbb{E}[y])] \\ &= \mathbb{E}[(x - \mathbb{E}[x])(\beta_0 + \beta_1 x + \epsilon - \beta_0 - \beta_1 \mathbb{E}[x] - \mathbb{E}[\epsilon])] \\ &= \mathbb{E}[(x - \mathbb{E}[x])(\epsilon + \beta_1(x - \mathbb{E}[x]))] \\ &= \beta_1 \mathbb{E}[(x - \mathbb{E}[x])^2] =: \beta_1 \cdot \text{Var}(x)\end{aligned}$$

Alternative Method cont'd II

$$\beta_1 = \frac{\text{Cov}(x, y)}{\text{Var}(x)}, \quad (7)$$

$$\beta_0 = \mathbb{E}[y] - \beta_1 \mathbb{E}[x]. \quad (8)$$

☞ With the observed pairs $(x_1, y_1), \dots, (x_n, y_n)$ we get the estimates $\hat{\beta}_0, \hat{\beta}_1$,

$$\mathbb{E}[x] \approx \frac{1}{n} \sum_{i=1}^n x_i =: \bar{x}, \quad (9)$$

$$\mathbb{E}[y] \approx \frac{1}{n} \sum_{i=1}^n y_i =: \bar{y}, \quad (10)$$

$$\text{Var}(x) \approx \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 =: s_{xx} \quad (11)$$

$$\text{Cov}(x, y) \approx \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) := s_{xy} \quad (12)$$

Remark: For s_{xx} and s_{xy} , division by n or $n-1$ does not affect the coefficients. But the same choice should be applied to both estimators.

Numerical Example

Consider a data set of $n = 4$ known (x, y) samples

$$D_4 = \{(1, 3), (2, 4), (3, 8), (4, 9)\}$$

Find:

- ▶ The coefficients $\hat{\beta}_0, \hat{\beta}_1$ for the simple linear regression.
- ▶ The residuals ϵ_i from the estimated values.

Solution

A. Giovanidis 2020

$$\bar{x} = \frac{1 + 2 + 3 + 4}{4} = 2.5 \quad , \quad \bar{y} = \frac{3 + 4 + 8 + 9}{4} = 6.$$

$$s_{xx} = \frac{1}{4 - 1} [(1 - 2.5)^2 + (2 - 2.5)^2 + (3 - 2.5)^2 + (4 - 2.5)^2] = \frac{5}{3}$$

$$s_{xy} = \frac{1}{4 - 1} [(1 - 2.5)(3 - 6) + (2 - 2.5)(4 - 6) + \\ + (3 - 2.5)(8 - 6) + (4 - 2.5)(9 - 6)] = \frac{11}{3}$$

Solution cont'd

$$\hat{\beta}_1 = \frac{s_{xy}}{s_{xx}} = \frac{11}{5} = 2.2$$

$$\hat{\beta}_0 = 6 - (2.2)(2.5) = 0.5$$

Regression-line:

$$\hat{y}_i = 0.5 + 2.2x_i,$$

Residuals:

$$\hat{y}_1 = 2.7, \hat{y}_2 = 4.9, \hat{y}_3 = 7.1 \hat{y}_4 = 9.3$$

$$\hat{\epsilon}_1 = 0.3, \hat{\epsilon}_2 = -0.9, \hat{\epsilon}_3 = 0.9 \hat{\epsilon}_4 = -0.3$$

👉 **Verification:** $\hat{\epsilon}_1 + \hat{\epsilon}_2 + \hat{\epsilon}_3 + \hat{\epsilon}_4 = 0$. (Why?)

Accuracy of the model

- ▶ *Mean Squared Error (MSE).*
- ▶ *Residual Standard Error (RSE).*
- ▶ R^2 statistic or *Coefficient of determination.*
- ▶ Confidence intervals.
- ▶ p-value.

Mean Squared Error

One can use the **Mean Squared Error (MSE)**, defined as

$$MSE := \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (13)$$

$$= \frac{1}{n} RSS \quad (14)$$

It is a measure of *lack of fit* for the model.

Residual Standard Error

The **residual standard error** provides the average amount that the response y deviates from the true regression line $\beta_0 + \beta_1 x$.

$$RSE = \sqrt{\frac{1}{n-2} RSS} = \sqrt{\frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2}. \quad (15)$$

It is a measure of *lack of fit* for the model, having also the nice interpretation,

$$\text{Var}(\epsilon) \approx RSE^2. \quad (16)$$

R^2 statistic

✎ An alternative name is **Coefficient of determination**.

$R^2 \in [0, 1]$ is the proportion of variability in y (dependent) that can be explained / predicted by knowing the variability of x (independent).

✎ The closer to 1, the better the fit.

$$\begin{aligned}
 R^2 &= \frac{\text{Explained Variation}}{\text{Total Variation}} = 1 - \frac{RSS}{TSS} \\
 &= \frac{\beta_1^2 \text{Var}(x)}{\text{Var}(y)} \stackrel{(7)}{=} \frac{[\text{Cov}(x, y)]^2}{\text{Var}(x)\text{Var}(y)} \approx \frac{s_{xy}^2}{s_{xx}s_{yy}} =: \rho^2. \quad (17)
 \end{aligned}$$

✎ In this case equal to ρ^2 the “sample correlation coefficient”.

$TSS = \sum_{i=1}^n (y_i - \bar{y})^2$ is the *Total Sum of Squares*.

R^2 statistic cont'd

A. Giovanidis 2020

To better understand, let us take a look at the variance of $y = \beta_0 + \beta_1 x + \epsilon$,

$$\text{Var}(y) \stackrel{\text{indep. } x, \epsilon}{=} \beta_1^2 \text{Var}(x) + \text{Var}(\epsilon).$$

1. Variance due to variation of x : $\beta_1^2 \text{Var}(x)$
2. Variance of error: $\text{Var}(\epsilon)$. (Variance left in y after we know x)

☞ If $\text{Var}(\epsilon)$ is small, then y will be close to $\beta_0 + \beta_1 x$, so that our linear regression model will successfully estimate y .

R^2 example

A. Giovanidis 2020

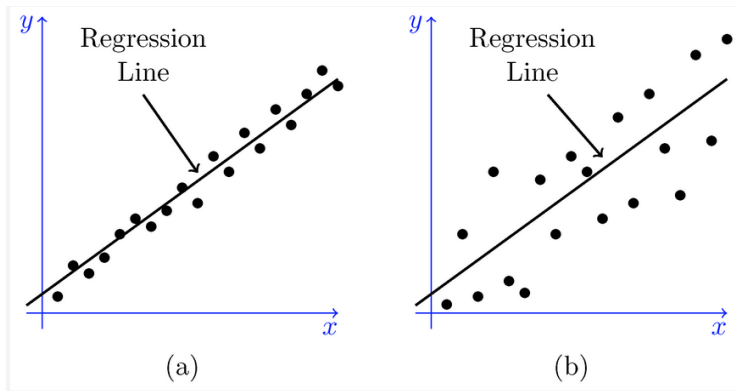


Figure: left: high value of R^2 , right: low value of R^2 .³

³Source [B.2]

Confidence intervals

The 95% confidence intervals for β_0 and β_1 are

$$\hat{\beta}_0 \pm 2 \cdot SE(\hat{\beta}_0), \quad \text{and} \quad \hat{\beta}_1 \pm 2 \cdot SE(\hat{\beta}_1), \quad (18)$$

where SE is the *Standard Error*.

$$SE(\hat{\beta}_0)^2 = \text{Var}(\epsilon) \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right], \quad (19)$$

$$SE(\hat{\beta}_1)^2 = \frac{\text{Var}(\epsilon)}{\sum_{i=1}^n (x_i - \bar{x})^2}. \quad (20)$$

To estimate $\text{Var}(\epsilon)$ we use the *Residual Standard Error (RSE)*

$$\text{Var}(\epsilon) \approx RSE^2. \quad (21)$$

p-value: Are x and y related?

$$H_0 : \beta_1 = 0$$

versus

$$H_1 : \beta_1 \neq 0$$

If $\beta_1 = 0$ then the model reduces to $Y = \beta_0 + \epsilon$, and x is not associated with y .

Q: Given $\hat{\beta}_1 > 0$ what is the probability of this being a false alarm?

👉 To answer, we compute the **p-value**. This relates to the probability that $\hat{\beta}_1$ is close to 0, related to the standard error.

If the **p-value** is very small (1% – 5%) we reject the null hypothesis, i.e. a relationship does exist between x and y .

Numerical Example cont'd

$$\blacktriangleright \text{RSS} = \hat{\epsilon}_1^2 + \hat{\epsilon}_2^2 + \hat{\epsilon}_3^2 + \hat{\epsilon}_4^2 = 0.3^2 + 0.9^2 + 0.9^2 + 0.3^2 = 1.8$$

$$\blacktriangleright \text{MSE} = \frac{\text{RSS}}{n} = \frac{1.8}{4} = 0.45.$$

$$\blacktriangleright \text{RSE} = \sqrt{\frac{\text{RSS}}{n-2}} = \sqrt{\frac{1.8}{4-2}} \approx 0.949.$$

$$\blacktriangleright R^2 = \frac{s_{xy}^2}{s_{xx}s_{yy}} = \frac{(11/3)^2}{(5/3) \cdot s_{yy}} = \frac{(11/3)^2}{(5/3) \cdot (26/3)} \approx 0.93.$$

$$\blacktriangleright s_{yy} = \frac{1}{4-1} [(3-6)^2 + (4-6)^2 + (8-6)^2 + (9-6)^2] = \frac{26}{3}.$$

👉 R^2 -statistic is 0.93, hence close to 1, and the fit is good!
(Cannot tell just by RSE)

Multiple Linear Regression Model

A. Giovanidis 2020

Very often, a response depends on several ($K > 1$) features.

- ▶ **Naive approach:** Run one simple linear regression per feature.
But! In this way the estimates ignore all other features left outside:
Not always good, due to correlation among features.
- ▶ **Correct approach:** Extend the simple linear model to accommodate multiple predictors. Give to each predictor a **separate slope coefficient** in a single model,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \epsilon. \quad (22)$$

Coefficient Estimation

Again, the regression coefficients are unknown and must be estimated.

We use the formula:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_k x_k. \quad (23)$$

We choose the parameters that minimise again the RSS

$$\begin{aligned} RSS &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i,1} - \hat{\beta}_2 x_{i,2} - \dots - \hat{\beta}_k x_{i,k})^2 \end{aligned} \quad (24)$$

☞ Complicated expressions, better use existing software packages.

Least-squares plane

A. Giovanidis 2020

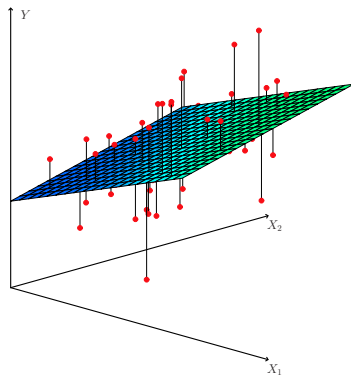


Figure: Regression plane for two features.⁴

⁴ Source [B.1]

Hypothesis test (again)

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$$

versus

$$H_a : \text{at least one } \beta_j \text{ non-zero.}$$

☞ To answer, we compute the *F-statistic*:

$$F = \frac{(TSS - RSS)/k}{RSS/(n - k - 1)}, \quad (25)$$

$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$, and $TSS = \sum_{i=1}^n (y_i - \bar{y})^2$ is the Total Sum of Squares.

If the *F-statistic* is very close to 1 then we expect no relationship between the response and the predictors. On the other hand, if H_a is true, we expect $F > 1$.

Variable selection

How do we decide on the important variables that influence the y ?

☞ use the F -statistic and the individual p -value.

- ▶ Use different combinations of features to derive the F -statistic. If for a specific combination among these the value of the statistic drops considerably, then this is an indicator that among the included features, some are unrelated to the response.
- ▶ From the individual p -values, the one with the highest p -value is a candidate to be removed from the model.
- ▶ The R^2 value determines how much of the data variance is explained by the model. If it is low then the feature set used does not contain much info.
- ▶ other selection methods (forward and backward selection...)

Potential Problems of Linear Regression

- ▶ Non-linearity of the response-predictor relationships.
- ▶ Correlation of error terms.
- ▶ Non-constant variance of error terms ([heteroscedasticity](#))
- ▶ Outliers (e.g. incorrect data collection)
- ▶ High leverage points.

Use [Residual plots](#) to detect problems:

$(y_i - \hat{y}_i, x_i)$ or $(y_i - \hat{y}_i, \hat{y}_i)$.

Non-linearity

A. Giovanidis 2020

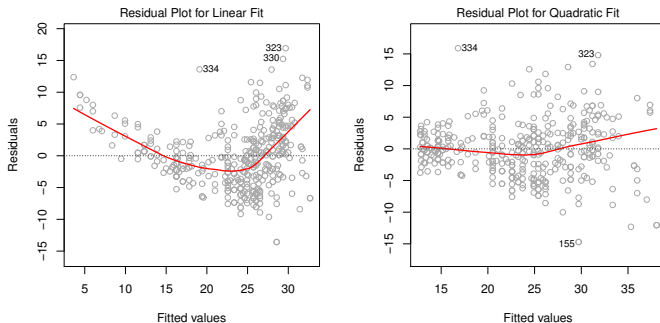


Figure: Residual plots for Linear and Polynomial fit.⁵

⁵Source [B.1]

Error Correlation

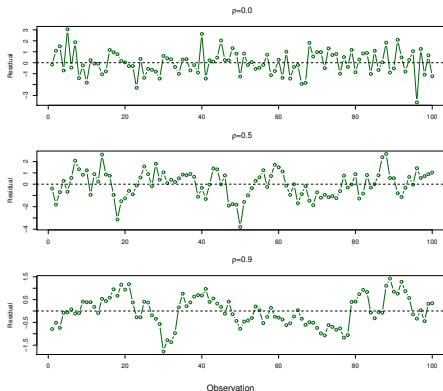


Figure: Residual plots for Time-series.⁶

⁶Source [B.1]

Heteroscedasticity

A. Giovanidis 2020

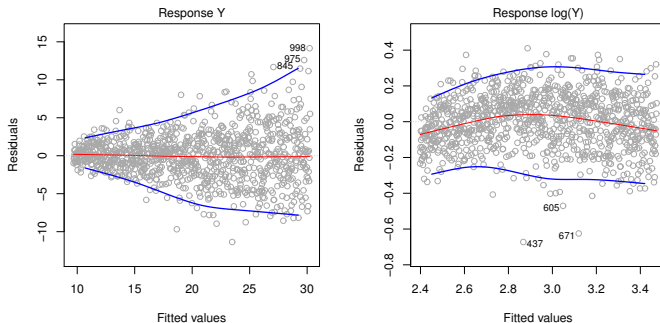


Figure: Residual plots for non-constant variance.⁷

⁷ Source [B.1]

Outliers

A. Giovanidis 2020

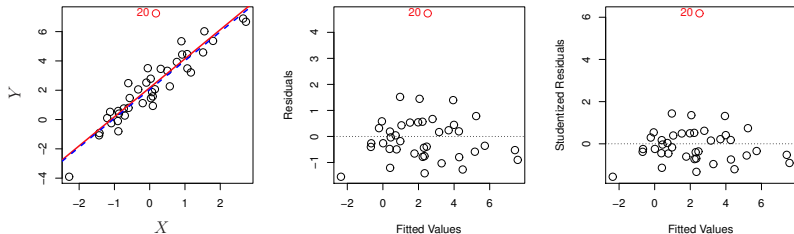


Figure: Effect of an outlier in least-squares fit.⁸

Outliers influences RSE, confidence intervals and p-values.

⁸Source [B.1]

High Leverage Points

A. Giovanidis 2020

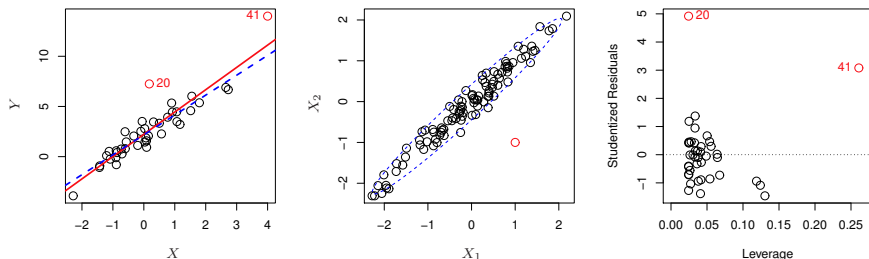


Figure: Effect of a high-leverage point in least-squares fit.⁹

⁹Source [B.1]

Collinearity

A. Giovanidis 2020

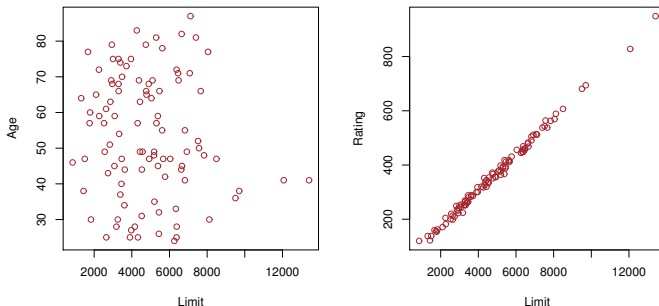


Figure: Effect of collinearity in least-squares fit.¹⁰

☞ Difficult to separate out the individual effects of collinear variables on the response.

¹⁰Source [B.1]

END