Linear Regression

Fraida Fund

Contents

In this lecture

- Simple linear regression
- Regression performance metrics
- Multiple linear regression

Regression

The output variable y is continuously valued.

For each input $\mathbf{x_i}$, the model estimates

$$\hat{y}_i = y_i - \epsilon_i$$

where ϵ_i is an error term, also called the **residual**.

Simple linear regression

Assume a linear relationship between single feature x and target variable y:

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

$\boldsymbol{\beta} = (\beta_0, \beta_1)$, the intercept and slope, are model **parameters**.

Residual term

Actual relationship include variation due to factors other than x, includes **residual** term:

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

where $\epsilon = y - \hat{y}$.

Linear model with residual - illustration

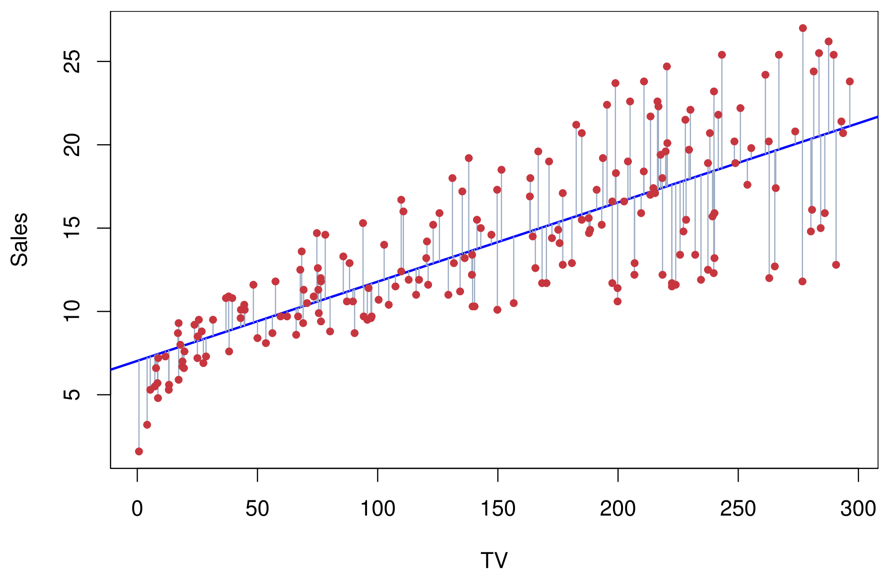

Figure 1: Example of linear fit with residuals shown as vertical deviation from regression line.

Interpretability of linear model

If slope β_1 is 0.0475 sales/dollar spent on TV advertising, we can say that a $1,000 increase in TV advertising budget is, on average, associated with an increase of about 47.5 in units sold.

However, note that:

- we can show a correlation, but can't say that the relationship is causative.
- the value for β_1 is only an *estimate* of the true relationship between TV ad dollars and sales.

"Recipe" for simple linear regression

- Choose a **model**: $\hat{y} = \beta_0 + \beta_1 x$
- Get **data** - for supervised learning, we need **labeled** examples: $(x_i, y_i), i = 1, 2, \cdots, N$
- Choose a **loss function** that will measure how well model fits data: **??**
- Find model **parameters** that minimize loss: find β_0 and β_1
- Use model to **predict** $\hat{y}$ for new, unlabeled samples

Least squares model fitting

Residual sum of squares:

$$RSS(\beta_0, \beta_1) := \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (\epsilon_i)^2$$

Least squares solution: find (β_0, β_1) to minimize RSS.

"Recipe" for simple linear regression

- Choose a **model**: $\hat{y} = \beta_0 + \beta_1 x$
- Get **data** - for supervised learning, we need **labeled** examples: $(x_i, y_i), i = 1, 2, \cdots, N$
- Choose a **loss function** that will measure how well model fits data: $RSS(\beta_0, \beta_1)$
- Find model **parameters** that minimize loss: find β_0 and β_1
- Use model to **predict** $\hat{y}$ for new, unlabeled samples

Minimizing RSS (1)

RSS is convex, so to minimize, we take

$$\frac{\partial RSS}{\partial \beta_0} = 0, \frac{\partial RSS}{\partial \beta_1} = 0$$

where

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Minimizing RSS (2)

First, the intercept:

$$\frac{\partial RSS}{\partial \beta_0} = \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i)(-1)$$

$$= -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0$$

using chain rule, power rule.

Minimizing RSS (3)

This is equivalent to setting sum of residuals to zero:

$$\sum_{i=1}^{n} \epsilon_i = 0$$

Minimizing RSS (4)

Now, the slope:

$$\frac{\partial RSS}{\partial \beta_1} = \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i)(-x_i)$$

$$= -2\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) = 0$$

Minimizing RSS (5)

This is equivalent to:

$$\sum_{i=1}^{n} x_i \epsilon_i = 0$$

Minimizing RSS (6)

Two conditions,

$$\sum_{i=1}^{n} \epsilon_i = 0, \sum_{i=1}^{n} x_i \epsilon_i = 0$$

where

$$\epsilon_i = y_i - \beta_0 - \beta_1 x_i$$

Minimizing RSS (7)

Which we expand into

$$\sum_{i=1}^{n} y_i = n\beta_0 + \sum_{i=1}^{n} x_i\beta_1$$

$$\sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} x_i\beta_0 + \sum_{i=1}^{n} x_i^2\beta_1$$

Minimizing RSS (8)

Divide

$$\sum_{i=1}^{n} y_i = n\beta_0 + \sum_{i=1}^{n} x_i\beta_1$$

by n, we find the intercept

$$\beta_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - \beta_1\frac{1}{n}\sum_{i=1}^{n} x_i$$

Minimizing RSS (9)

$$\beta_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - \beta_1\frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\beta_0 = \bar{y} - \beta_1\bar{x}$$

where sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Minimizing RSS (10)

To solve for β_1: Multiply

$$\sum_{i=1}^{n} y_i = n\beta_0 + \sum_{i=1}^{n} x_i\beta_1$$

by $\sum x_i$, and multiply

$$\sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} x_i\beta_0 + \sum_{i=1}^{n} x_i^2\beta_1$$

by n.

Minimizing RSS (11)

$$\sum_{i=1}^n x_i \sum_{i=1}^n y_i = n \sum_{i=1}^n x_i \beta_0 + (\sum_{i=1}^n x_i)^2 \beta_1$$

$$n \sum_{i=1}^n x_i y_i = n \sum_{i=1}^n x_i \beta_0 + n \sum_{i=1}^n x_i^2 \beta_1$$

Subtract the first equation from the second to get…

Minimizing RSS (12)

$$n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i = n \sum_{i=1}^n x_i^2 \beta_1 - (\sum_{i=1}^n x_i)^2 \beta_1$$

$$= \beta_1 \left(n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2 \right)$$

Minimizing RSS (13)

Solve for β_1:

$$\beta_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Minimizing RSS (14)

which is:

$$\frac{s_{xy}}{s_x^2}$$

- sample covariance $s_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$
- sample variance $s_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$

Minimizing RSS (15)

Also express as

$$\frac{r_{xy} s_y}{s_x}$$

where sample correlation coefficient $r_{xy} = \frac{s_{xy}}{s_x s_y}$.

(Note: from Cauchy-Schwartz law, $|s_{xy}| < s_x s_y$, we know $r_{xy} \in [-1, 1]$)

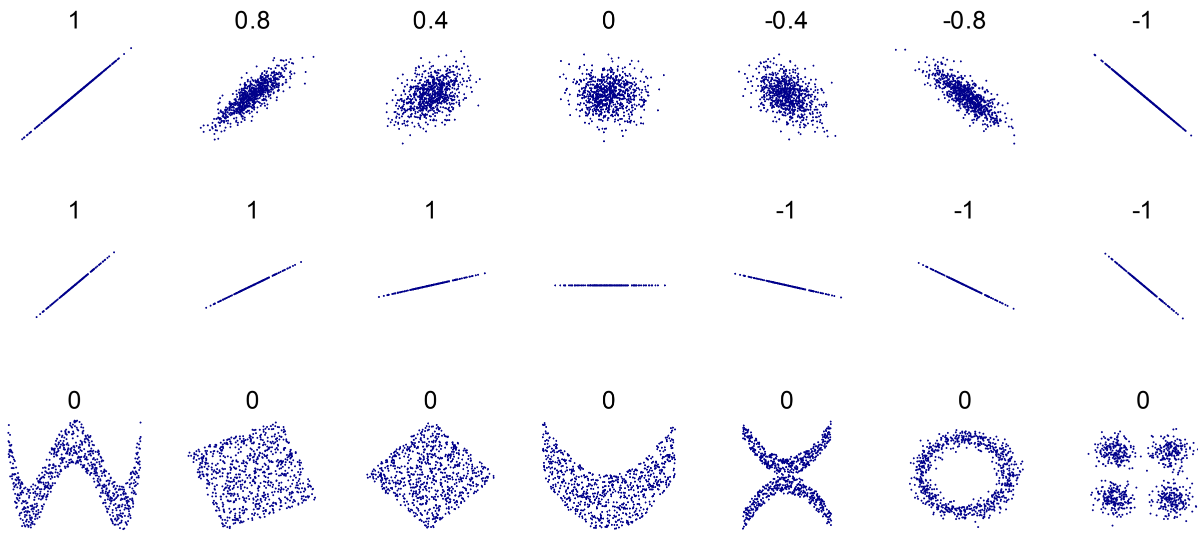

Figure 2: Several sets of (x, y) points, with r_{xy} for each. Image via Wikipedia.

Correlation coefficient: visual

Minimizing RSS - final solution

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

$$\beta_1 = \frac{s_{xy}}{s_x^2} = \frac{r_{xy} s_y}{s_x}$$

Minimum RSS

$$\min_{\beta_0, \beta_1} RSS(\beta_0, \beta_1) = N(1 - r_{xy}^2) s_y^2$$

- **coefficient of determination**: $R^2 = r_{xy}^2$, explains the portion of variance in y explained by x.
- s_y^2 is variance in target y
- $(1 - R^2) s_y^2$ is the residual sum of squares after accounting for x.

Visual example (1)

Visual example (2)

Visual example (3)

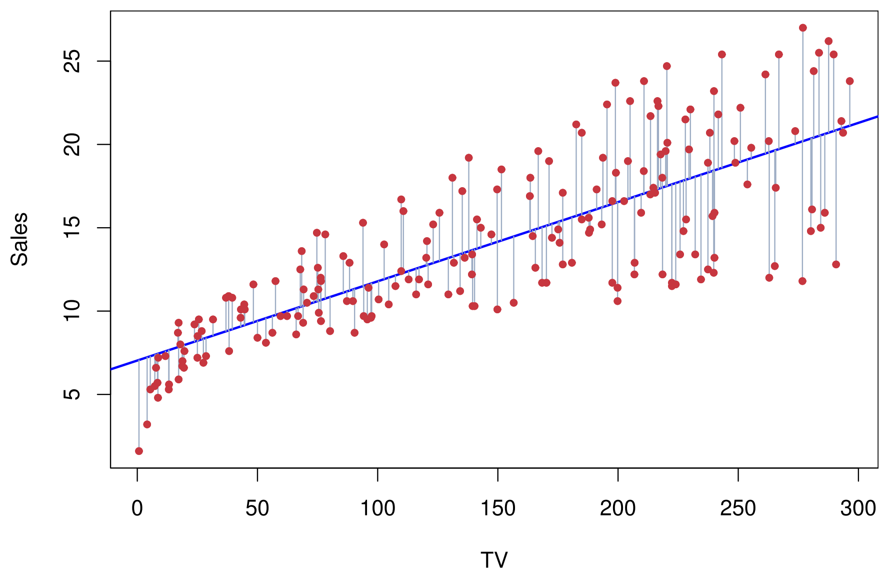

Figure 3: Example of linear fit with residuals shown as vertical deviation from regression line.

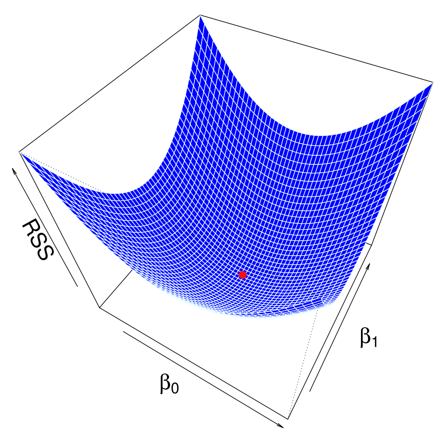

Figure 4: Regression parameters - 3D plot.

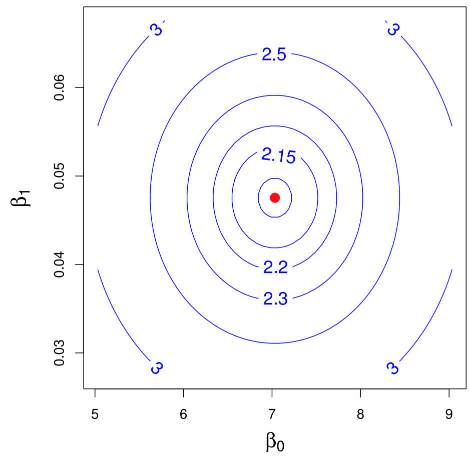

Figure 5: Regression parameters - contour plot.

Regression performance metrics

R^2: coefficient of determination

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

- For linear regression: What proportion of the variance in y is "explained" by our model?
- $R^2 \approx 1$ - model "explains" all the variance in y
- $R^2 \approx 0$ - model doesn't "explain" any of the variance in y
- Depends on the sample variance of y - can't be compared across datasets

RSS

Definition: **Residual sum of squares** (RSS), also called **sum of squared residuals** (SSR) and **sum of squared errors** (SSE):

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

RSS increases with n (with more data).

Relative forms of RSS (1)

- RSS per sample

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

Relative forms of RSS (2)

- Normalized RSS (divide RSS per sample, by sample variance of y), the ratio of *average error of your model* to *average error of prediction by mean*.

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

Multiple linear regression

Matrix representation of data

Represent data as a **matrix**, with n samples and k features; one sample per row and one feature per column:

$$X = \begin{bmatrix} x_{1,1} & \cdots & x_{1,k} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,k} \end{bmatrix}, y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

$x_{i,j}$ is jth feature of ith sample.

Linear model

Assume a linear relationship between feature vector $x = [x_1, \cdots, x_k]$ and target variable y:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k + x_k$$

Model has $p = k + 1$ terms.

Matrix representation of linear regression (1)

Samples are $(\mathbf{x_i}, y_i), i = 1, 2, \cdots, n$

Each sample has a feature vector $\mathbf{x_i} = [x_i, 1, \cdots, x_i, k]$ and scalar target y_i

Predicted value for ith sample will be $\hat{y_i} = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_k x_{i,k}$

Matrix representation of linear regression (2)

Define **feature matrix** and **regression vector**:

$$A = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,k} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$$

Then, $\hat{\mathbf{y}} = A\boldsymbol{\beta}$, and given a new sample with feature vector $\mathbf{x}$, predicted value is $\hat{y} = [1, \mathbf{x}^T]\boldsymbol{\beta}$.

Least squares model fitting

Problem: learn the best coefficients $\boldsymbol{\beta} = [\beta_0, \beta_1, \cdots, \beta_k]$ from the labeled training data.

$$RSS(\boldsymbol{\beta}) := \sum_{i=1}^{n} (y_i - \hat{y_i})^2$$

Least squares solution: Find $\boldsymbol{\beta}$ to minimize RSS.

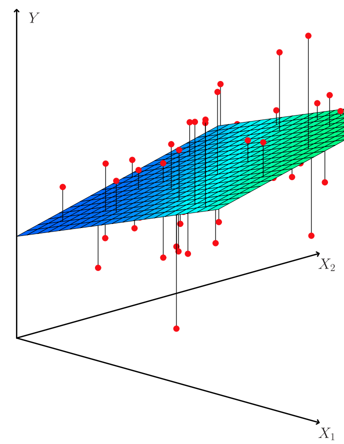

Figure 6: The least squares regression is now a plane, chosen to minimize sum of squared distance to each observation.

Illustration - two features

Supervised learning recipe for linear regression

- Linear model: $\hat{y} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$
- Data: $(\mathbf{x_i}, y_i), i = 1, 2, \cdots, n$
- Loss function:

$$RSS(\beta_0, \beta_1, \cdots, \beta_k) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- Find parameters: Select $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_k)$ to minimize $RSS(\boldsymbol{\beta})$

Setup: $\ell 2$ norm

Definition: Euclidian norm or $\ell 2$ norm of a vector $\mathbf{x} = (x_1, \cdots, x_n)$:

$$||\mathbf{x}|| = \sqrt{x_1^2 + \cdots + x_n^2}$$

Intuitively, it is the “length” of a vector. We will want to minimize the norm of the residual.

Setup: Finding maxima/minima

For $f(x)$, can find local maxima and minima by finding where the derivative with respect to x is zero.

For a multivariate function $f(\mathbf{x}) = f(x_1, \cdots, x_n)$, we find places where the **gradient** - vector of partial derivatives - is zero, i.e. each entry must be zero:

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix}$$

If function is convex, there is a single global minimum.

Setup: RSS as vector norm

$$RSS = ||\mathbf{y} - \hat{\mathbf{y}}||^2$$

$$RSS = ||\mathbf{y} - \mathbf{A}\boldsymbol{\beta}||^2$$

Least squares solution (1)

RSS is convex, so there is a single global minimum

Cost function (remember, $p = k + 1$):

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \hat{y}_i = \sum_{j=0}^{p} A_{i,j}\beta_j$$

Least squares solution (2)

In matrix form (note: $||Ax - b|| = ||b - Ax||$):

$$RSS = ||A\boldsymbol{\beta} - \mathbf{y}||^2$$

Compute gradient via chain rule, power rule:

$$\nabla RSS = 2A^T(A\boldsymbol{\beta} - \mathbf{y})$$

Least squares solution (3)

Set derivative to zero:

$$2A^T(A\boldsymbol{\beta} - \mathbf{y}) = 0 \rightarrow A^T A\boldsymbol{\beta} = A^T\mathbf{y}$$

then

$$\boldsymbol{\beta} = (A^T A)^{-1} A^T \mathbf{y}$$

Least squares solution (4)

Minimum RSS:

$$RSS = \mathbf{y}^T [I - A(A^T A)^{-1} A^T] \mathbf{y}$$

Interpretation using autocorrelation (1)

Each sample has feature vector

$$A_i = (A_{i0}, \cdots, A_{ik}) = (1, x_{i1}, \cdots, x_{ik})$$

Interpretation using autocorrelation (2)

Define:

- Sample autocorrelation matrix: $R_{AA} = \frac{1}{n}A^T A, R_{AA}(l,m) = \frac{1}{n}\sum_{i=1}^n A_{il}A_{im}$ (correlation of feature l and feature m)
- Sample cross-correlation vector: $R_{Ay} = \frac{1}{n}A^T y, R_{yA}(l) = \frac{1}{n}\sum_{i=1}^n A_{il}y_i$ (correlation of feature l and target)

Interpretation using autocorrelation (3)

Least squares solution:

$$\boldsymbol{\beta} = R_{AA}^{-1}R_{Ay}$$

Categorical feature?

Can use **one hot encoding**:

- For a categorical variable x with values $1, \cdots, M$
- Represent with M binary features: $\phi_1, \phi_2, \cdots, \phi_m$
- Model as $y = \beta_0 + \beta_1\phi_1 + \cdots + \beta_M\phi_M$

Linear regression - what can go wrong?

- Relationship may not actually be linear (may be addressed by non-linear transformation - future lecture)
- Violation of additive assumption (need interaction terms)
- "Tracking" in residuals (e.g. time series)
- Outliers - may be difficult to spot - may have outsize effect on regression line and/or R^2
- Collinearity

Residuals plot

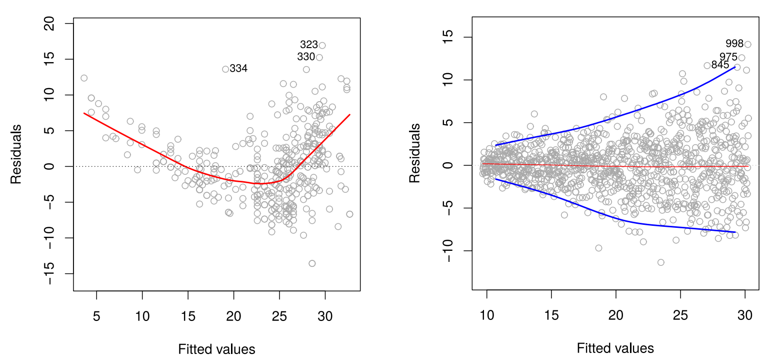

Figure 7: Residuals plot

Dealing with outliers

References

- Figures in this presentation are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R.

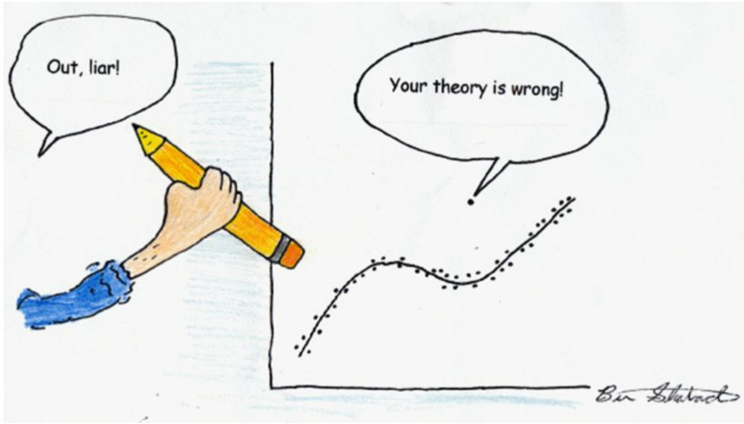

Figure 8: “Remove outliers” is not a strategy for dealing with outliers.

Tibshirani.

- For more detail on the derivation of the least squares solution to the multiple linear regression, refer to Chapter 12 in “Introduction to Applied Linear Algebra”, Boyd and Vandenberghe.
- For more detail on the statistical aspects of linear regression (outside the scope of the ML course), please refer to chapter 3 of: “An Introduction to Statistical Learning with Applications in R”, G. James, D. Witten, T. Hastie and R. Tibshirani.