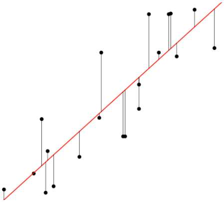

A 10 minute-ish introduction to linear regression

Eduardo García-Portugués

University of Copenhagen

26th February 2016

Motivating example

- A production line where we measure the time (in minutes) it takes to produce a number items
- Two rv's $Y =$ "time required to produce an order" and $X =$ "number of items in the order"
- A dataset like this

X	Y
195	175
215	189
243	344
⋮	⋮

- We are interested in describing the **relation** between Y (response, dependent variable) and X (predictor, independent) and **predict** Y from X

Simple linear regression

- The **conditional expectation** of Y given $X = x$ can be seen as a function, called the **regression function**

$$m(x) := \mathbb{E}[Y|X = x] = \int y\, f_{Y|X}(y|x)\, \mathrm{d}y = \int y\, \frac{f_{XY}(x, y)}{f_X(x)}\, \mathrm{d}y$$

- We can **always** write $Y = m(X) + \varepsilon$, with ε (**error, noise**) st $\mathbb{E}[\varepsilon|X] = 0$
- Linear regression assumes that ***m* is linear** (or at least close to it) for some **unknown parameters** (β_0, β_1):

$$m(x) = \beta_0 + \beta_1 x$$

- How to estimate (β_0, β_1) from a sample $\{(X_i, Y_i)\}_{i=1}^n$?

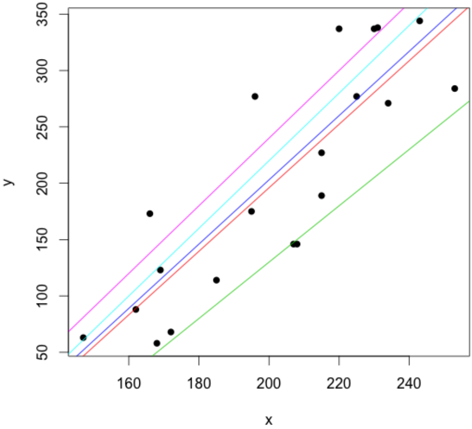

Figure: Scatterplot of the time required to produce an order (Y) versus the number of items in the order (X). Which of the linear fits is "better"?

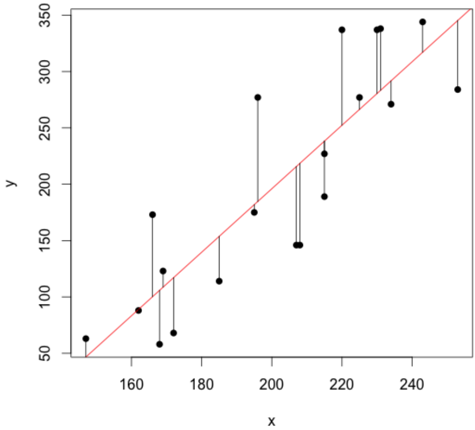

Figure: Scatterplot of the time required to produce an order (Y) versus the number of items in the order (X). Which of the linear fits is "better"?

Least squares fitting

- Let's denote $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$, $\mathbf{X} = \begin{pmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix}$ and $\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$
- We seek to minimize the **residual sum of squares**:

$$\mathrm{RSS}(\boldsymbol{\beta}) = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2 = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

- RSS is a quadratic function in $\boldsymbol{\beta}$, so

$$\frac{\partial \mathrm{RSS}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = 0, \quad \frac{\partial^2 \mathrm{RSS}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = 2\mathbf{X}^T\mathbf{X} > 0$$

- This gives $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ (if $\mathbf{X}^T\mathbf{X} > 0$) as the minimizer of the RSS
- Recall we have not required any **statistical assumption** for obtaining $\hat{\boldsymbol{\beta}}$

Model assumptions

- We assume the next hypothesis on the linear model $Y = \beta_0 + \beta_1 X + \varepsilon$:
  - A1 **Homocedasticity**: $\mathbb{V}\mathrm{ar}[\varepsilon | X = x] = \mathbb{E}\left[\varepsilon^2 | X = x\right] = \sigma^2$
  - A2 **Normality**: the error is normal, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$
  - A3 **Independence**: the rv's $\varepsilon_i = Y_i - \beta_0 - \beta_1 X_i$, $i = 1, \ldots, n$ are independent

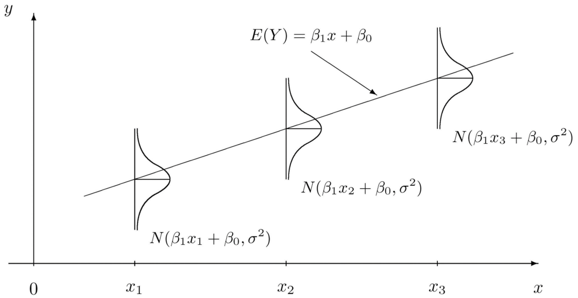

Figure: Sketch of the linear model assumptions

- Two possible frameworks for the predictor:
  - A4 **Fixed design** (assumed): the values of X are deterministic
  - A5 **Random design**: both the predictor and the response Y are random

Properties of estimators and prediction

- The **Maximum Likelihood Estimator** of β under A1-A3 **coincides** with $\hat{\beta}$
- From Fisher's theorem and linear transformation of normals we have:

Theorem

Under A1-A4, $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ and $\hat{\sigma}^2 := \frac{1}{n-2}\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$ are unbiased, independent and st

$$\hat{\beta} \sim \mathcal{N}(\beta, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2) \text{ and } \hat{\sigma}^2 \sim \frac{\sigma^2}{(n-2)}\chi^2_{n-2} \quad (1)$$

- We can compute **confidence intervals for** β_j based on (1)
- This allows the testing of $H_0 : \beta_j = b$ vs $H_1 : \beta_j \neq b$
- **Prediction** is done by the conditional mean:

Prediction

For a given x_0, the corresponding Y_0 is predicted as $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$

The determination coefficient R^2

- Percentage of **variability of Y explained by the model**

Variability of Y	Sum of squares
Explained by model (signal)	$\sum_{i=1}^n \left(\hat{Y}_i - \bar{Y}\right)^2$
Unexplained (noise)	$\sum_{i=1}^n \left(Y_i - \hat{Y}_i\right)^2 =: \text{RSS}$
Total	$\sum_{i=1}^n \left(Y_i - \bar{Y}\right)^2 =: \text{TSS}$

- $R^2 := 1 - \frac{\text{RSS}}{\text{TSS}}$ is the square of the Pearson correlation coefficient
- **If the model assumptions hold**, the larger R^2, the better fit

Caution!

$$\text{Validity of linear model} \not\Leftrightarrow \text{large } R^2$$

- The model is **valid** if the assumptions hold
- The model is **useful** if, *in addition*, the R^2 is large

Implementation

- Code in `https://egarpor.shinyapps.io/lin-reg/`
- We illustrate the following concepts:
  - Least squares estimator
  - Significances of coefficients
  - Prediction
  - Coefficient of determination