Biomedical Engineering Degree

6. Linear Regression

Felipe Alonso Atienza

✉felipe.alonso@urjc.es

@FelipeURJC

Escuela Técnica Superior de Ingeniería de Telecomunicación
Universidad Rey Juan Carlos

References

1. R. Bernard. *Fundamentals of Biostatistics*. Ed.: Thompson. Chapter 11
2. D. Díez, M Cetinkaya-Rundel and CD Barr. *OpenIntro Statistics*. Chapter 8, 9.
3. J. James, D. Witten, T Hastie and R. Tibshirani. *An Introduction to Statistical Learning*. Chapter 3.

Introduction

- A **regression model** is a model that allows us to describe an effect of a variable X on a variable Y

$$Y = f(X)$$

  - X: **independent** or **explanatory** or **exogenous** or **predictor** variable
  - Y: **dependent** or **response** or **endogenous** or **target** variable

- Examples:
  - Estimate the price of an apartment depending on its size
  - Estimate the weight of individuals depending on their height
  - Estimate the voltage depending on the current through a real resistor

- The objective is to obtain reasonable estimates of Y for X based on a sample of n **observations** $(x_1, y_1), \ldots, (x_n, y_n)$

Outline

Deterministic

- Given a value of X, the value of Y can be perfectly identified

$$Y = f(X)$$

  - Example: Ohm's law relationship for Voltage and Current through a Resistor:

$$V = I \cdot R$$

    notice that in the real practice, this relationship expressed in the Ohm's law **physical model** will not be a perfect due to the resistor tolerance

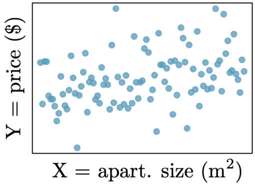
Nondeterministic (random/stochastic)

- Given a value of X, the value of Y **cannot** be perfectly identified

$$Y = f(X) + U$$

  where U is an unknown (random) perturbation (random variable).

  - Example: Estimate the price of an apartment depending on its size



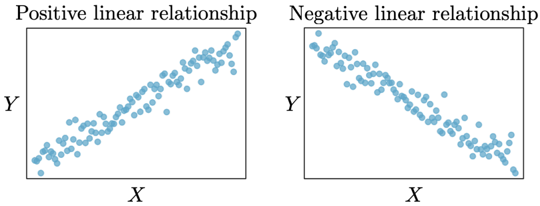
notice that there is a linear pattern, but not perfect. This means that the *apartment size* is not enough to linearly model the price. We might want to add more exogenous/predictor variables to reduce the uncertainty.

Linear

- When the function $f(X)$ is linear, then

$$f(X) = \beta_0 + \beta_1 X$$

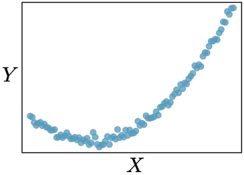
  - if $\beta_1 > 0$ there is a positive linear relationship
  - if $\beta_1 < 0$ there is a negative linear relationship



Nonlinear

- When the function $f(X)$ is nonlinear. For example:

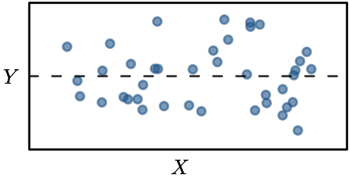
$$
\begin{aligned}
f(X) &= \beta_0 + \beta_1 X + \beta_2 X^2 \\
f(X) &= \log\left(X^2\right) \\
&\ldots
\end{aligned}
$$



  - Notice that in the first case, $f(x)$ is nonlinear with respect to the exogenous variable, but it is linear with respect to the β's!

Lack of relationship

- When $f(X) = 0$

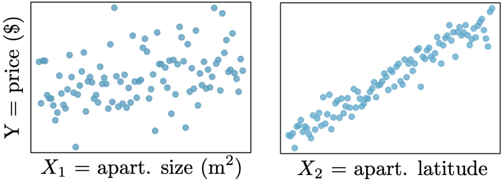


Multiple linear

- When the function $f(\cdot)$ depends on **two or more variables**: $X_1, X_2, \ldots$

$$f(X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

  - Example: Estimate the price of an apartment depending on its size (X_1) and location (X_2).



Outline

Covariance

- It is used to quantify the relationship between two random variables

$$\text{Cov}(X,Y) = \mathsf{E}\left[(X-\mu_X)(Y-\mu_Y)\right] = \mathsf{E}\left[XY\right] - \mu_X\mu_Y$$

which can be calculated using the observations $(x_1, y_1), \ldots, (x_n, y_n)$ as

$$\text{cov}(x,y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{n-1}$$

  - If there is a **positive linear** relationship, Cov > 0
  - If there is a **negative linear** relationship, Cov < 0
  - If there is no relationship or the relationship is nonlinear, Cov ≈ 0: uncorrelated variables

- Covariance depends on the units of X and Y

Correlation coefficient

- Normalize the covariance by the standard deviation

$$\rho = r(X,Y) = \mathrm{Cor}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

which can be calculated using the observations $(x_1, y_1), \ldots, (x_n, y_n)$ as

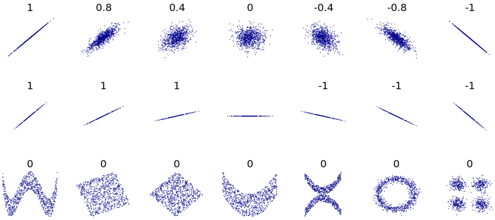
$$\mathrm{cor}(x,y) = \frac{\mathrm{cov}(x,y)}{s_x s_y}$$

where

$$s_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \quad \text{and} \quad s_y^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}$$

Correlation coefficient

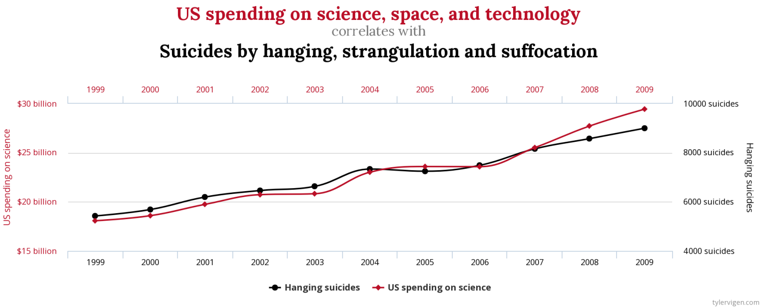
- Also known as Pearson correlation coefficient
- The correlation coefficient is unitless
- Symmetric: $\text{cor}(x, y) = \text{cor}(y, x)$
- $-1 \leq \text{cor}(x, y) \leq 1$
  - $\text{cor}(x, y) = 1$, perfect (positive) linear relationship
  - $\text{cor}(x, y) = -1$, perfect (negative) linear relationship
  - $\text{cor}(x, y) = 0$, no linear relationship



Correlation coefficient

- If X and Y are statistically independent then they are uncorrelated
  - Independent variables: $\mathsf{E}[XY] = \mathsf{E}[X]\mathsf{E}[Y]$
  - Uncorrelated variables $\mathsf{Cov}(X, Y) = \mathsf{E}[XY] - \mathsf{E}[X]\mathsf{E}[Y] = 0$

- **Correlation does not imply causation**[1]!



[1]Figure extracted from *Spurious correlations*

Other correlation measures

- *Spearman's* rank correlation: assesses monotonic relationships (whether linear or not)
- *Kendall's tau* (rank) coefficient: measures the similarity of the orderings of the data when ranked by each of the quantities

Outline

Simple linear regression model

- The simple linear regression model assumes that

$$\boxed{y_i = \beta_0 + \beta_1 x_i + u_i}$$

  where

  - y_i is the $i-$th observation of the dependent variable Y when the random variable X takes the observation value x_i
  - x_i is the $i-$th observation of the exogenous variable X
  - u_i is the error term, which is a **random variable** that accounts for the uncertainty on the observations, an it is assumed to be normal with a mean 0 and an unknown variance σ^2, that is

$$U \sim \mathcal{N}(0, \sigma^2)$$

  - β_0 and β_1 are the regression (population) coefficients:
    - ★ β_0 is the intercept
    - ★ β_1 is the slope

- The (population) parameters need to be estimated: β_0, β_1 and σ^2

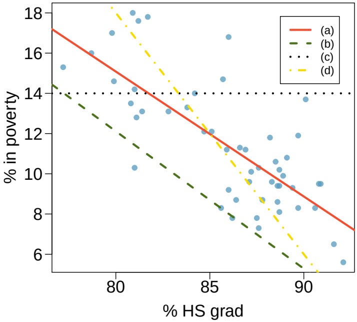
Simple linear regression model

- Based on the observed data $(x_1, y_1), \ldots, (x_n, y_n)$, we want to find the estimates $\hat{\beta}_0$, $\hat{\beta}_1$, in order to obtain the regression line

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

which is the best fit to the data with a linear pattern

- Example: High School (HS) graduate rate vs % of residents who live below the poverty line

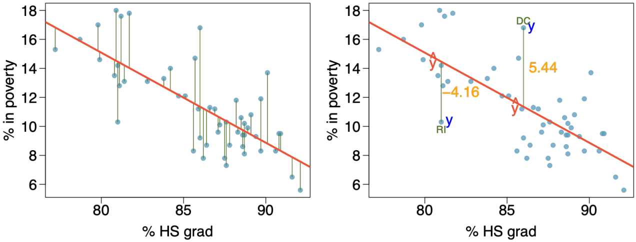


Simple linear regression model

- **Residual** is the difference between the observed y_i and predicted $\hat{y}_i$

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

- Example: High School (HS) graduate rate vs % of residents who live below the poverty line



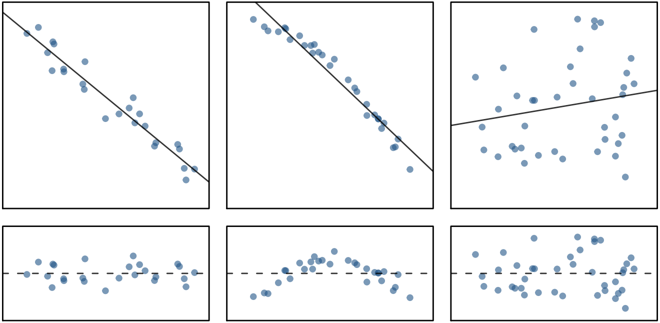
  - %living in poverty in DC is 5.44 % more than predicted.
  - % living in poverty in RI is 4.16 % less than predicted.

Simple linear regression model: model assumptions

1. **Linearity**: the relationship between the explanatory and the response variable should be linear.
2. **Homogeneity**: the errors have zero mean
$$\mathsf{E}[u_i] = 0$$
3. **Homoscedasticity**: the variance of the errors is constant
$$\mathsf{Var}(u_i) = \sigma^2$$
4. **Independence**: the errors are independent
$$\mathsf{E}[u_i u_j] = 0$$
5. **Normality**: the errors should be nearly normal
$$u_i \sim \mathcal{N}(0, \sigma^2)$$

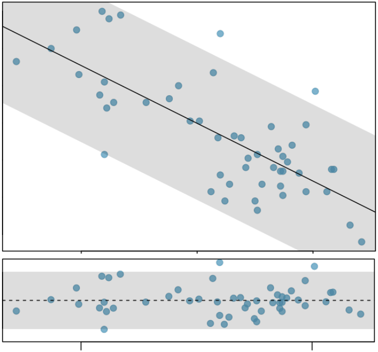
Model assumptions: linearity

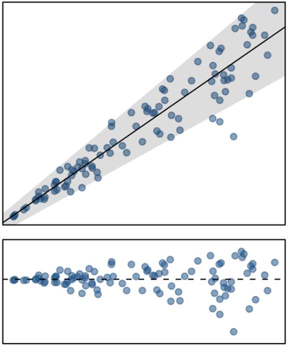
- The relationship between the explanatory and the response variable should be linear.
- Check using a scatterplot of the data, or a **residuals plot**.



Model assumptions: homoscedasticity

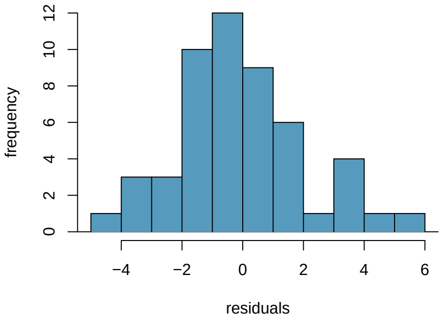
- The variability of points around the (prediction) line should be roughly constant.
- This implies that the variability of residuals around the 0 line should be roughly constant as well.
- If that's not the case, **heteroscedasticity** is present





Model assumptions: normality

- The residuals should be nearly normal
- This condition may not be satisfied when there are unusual observations that don't follow the trend of the rest of the data.
- Check using a histogram



Model assumptions: independence

- The observations should be independent
- One observation doesn't imply any information about another.
- In general, time series fail this assumption.

Outline

Fitting the regression line

- Our aim is to obtain the estimator $\hat{\beta}_0$ and $\hat{\beta}_1$ that provide **the best fit**

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \hat{x}_i$$

- There are several criteria to obtain the best fit:
  1. Minimize the sum of magnitudes (absolute values) of residuals

$$|e_1| + |e_2| + \ldots + |e_n| = \sum_{i=1}^{n} |e_i|$$

  2. Minimize the **residual sum of squares** (RSS), also known as least squares

$$\text{RSS} = e_1^2 + e_2^2 + \ldots + e_n^2 = \sum_{i=1}^{n} e_i^2$$

- The least squares method was proposed by Gauss in 1809

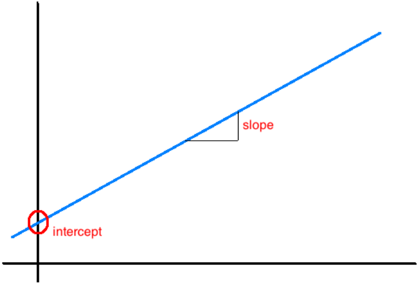
$$\text{RSS} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 \hat{x}_i \right) \right)^2$$

Least squares estimators

- The resulting estimators are

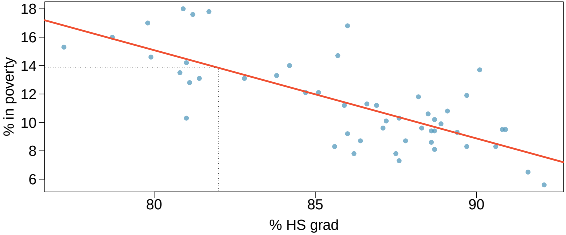
$$\hat{\beta}_1 = \frac{\text{cov}(x,y)}{s_x^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$



Prediction

- Using the linear model to predict the value of the response variable for a given value of the explanatory variable is called prediction, simply by plugging in the value of x in the linear model equation
- There will be some uncertainty associated with the predicted value.



Goodness-of-fit of the model

- The quality of a linear regression fit is typically assessed using two related quantities:

  1. The residual standard error (RSE): an (unbiased) estimate of σ, the standard deviation of u_i

  $$\mathsf{RSE} = \sqrt{\frac{\mathsf{RSS}}{n-2}}$$

  notice that RSE it is measured in the units of Y, thus it is not always clear what constitutes a good RSE

  2. The R^2 statistic, which is calculated as

  $$\mathsf{R}^2 = 1 - \frac{\mathsf{RSS}}{\mathsf{TSS}}$$

  where $\mathsf{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the **total sum of squares**

R^2 coefficient

- R^2 measures the proportion of variability in Y that can be explained using X
  - RSS measures the amount of **variability that is left unexplained** after performing the regression
  - TSS measures the total variance in the response Y, and can be thought of as the **amount of variability inherent in the response before the regression is performed**

- $R^2 \in [0, 1]$
  - An $R^2 \approx 1$ indicates that a large proportion of the variability in the response has been explained by the regression (good fit).
  - An $R^2 \approx 0$ indicates that the regression did not explain much of the variability in the response (bad fit)
  - In the real practice, less than 0.6, not so good

Outline

Inference in simple linear regression

- Up to this point we only talked about **point estimation**
- With **confidence intervals** for model parameters, we can obtain information about the estimation error.
- And **hypothesis tests** will help us to decide if a given parameter is statistically significant.
- We will use a $t-$test in inference for regression, and remember

$$\text{test statistic} = \frac{\text{point estimate} - \text{null value}}{\text{SE}}$$

Inference for the slope β_1

- It can be demonstrated that

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{(n-1)s_x^2}\right)$$

- In general, σ^2 is not known, but can be estimated from the RSE ($\sigma^2 = \text{RSE}^2$). So, we can approximate the SE of $\hat{\beta}_1$ by

$$\text{SE}_{\hat{\beta}_1} = \sqrt{\frac{\text{RSE}}{(n-1)s_x^2}}$$

- Since we use the *sample* SE, thus

$$t = \frac{\hat{\beta}_1 - \beta_1}{\text{SE}_{\hat{\beta}_1}} \sim t_{n-2}$$

- Notice that we lose 1 df for each parameter we estimate, and in simple linear regression we estimate 2 parameters, β_0 and β_1.

Confidence interval for the slope β_1

- $1-\alpha$ confidence interval

$$\hat{\beta}_1 \pm t_{n-2,1-\alpha/2} \times \underbrace{\sqrt{\frac{\mathsf{RSE}}{(n-1)s_x^2}}}_{\mathsf{SE}_{\hat{\beta}_1}}$$

- The length of the interval decreases if
  - The sample size increases
  - The variance of X increases
  - The RSE decreases

Hypothesis test on β_1

H_0 : There is no relationship between X and Y

H_1 : There is some relationship between X and Y

- Mathematically, this corresponds to testing

$$H_0 : \beta_1 = 0 \quad \text{vs} \quad H_1 : \beta_1 \neq 0$$

- Thus, we calculate our statistic

$$t = \frac{\hat{\beta}_1 - 0}{\mathsf{SE}_{\hat{\beta}_1}} \sim t_{n-2}$$

- Then, for a α significance level
  - We reject the null hypothesis if $|t| > t_{n-2,1-\alpha/2}$

Inference for the intercept β_0

- In this case

$$\hat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}\right)\right)$$

- In general, σ^2 is not known, but can be estimated from the RSE ($\sigma^2 = \text{RSE}^2$). So, we can approximate the SE of $\hat{\beta}_0$ by

$$\text{SE}_{\hat{\beta}_1} = \sqrt{\text{RSE}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}\right)}$$

- Since we use the *sample* SE, thus

$$t = \frac{\hat{\beta}_0 - \beta_0}{\text{SE}_{\hat{\beta}_0}} \sim t_{n-2}$$

Inference for the intercept β_0

- $1-\alpha$ confidence interval

$$\hat{\beta}_0 \pm t_{n-2,1-\alpha/2} \times \underbrace{\sqrt{\mathsf{RSE}^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}\right)}}_{\mathsf{SE}_{\hat{\beta}_0}}$$

- Hypothesis testing: if the true value of β_0 is 0, it means that the population regression line goes through the origin.

$$H_0 : \beta_0 = 0 \quad \text{vs} \quad H_1 : \beta_0 \neq 0$$

  using the statistic

$$t = \frac{\hat{\beta}_0 - 0}{\mathsf{SE}_{\hat{\beta}_0}} \sim t_{n-2}$$