Chapter 2: Simple Linear Regression

Option: DATA SCIENCE

September 2023

Dr. Abbas Rammal

Overview

1. Introduction
2. The simple regression model
3. Distribution of least squares estimators
    - Mathematical expectation of $\hat{\beta}_1$
    - Mathematical expectation of $\hat{\beta}_0$
    - Variance of $\hat{\beta}_1$
    - Variance of $\hat{\beta}_0$
4. Estimation of error variance
5. Inference on model parameters
6. Analysis of variance
7. Models with a single parameter
    - Model without explanatory variable
    - Model without constant

Introduction

- The model introduced in the first chapter is a model with a single explanatory variable. We are talking about simple linear regression.
- The regression line involves estimating from the data of a sample using the least squares method.
- Objective: The objective of this chapter is to generalize the inference on the regression line from these estimates.

The simple linear regression model

- The simple linear regression model is defined by the relationship:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

For $i = 1 \ldots n$

- The errors ε_i are unobservable random quantities.
- The y_i are also random variables (dependent of ε_i).
- The x_i are fixed numbers.

Assumptions on $\boldsymbol{\varepsilon_i}$

- The assumptions on ε_i are:

1. The ε_i errors are independent.
2. The ε_i errors are normally distributed.
3. All errors ε_i have zero expectation.
4. All errors ε_i have the same variance σ^2.

ε_i are random variables independent and identically distributed such that

$$\varepsilon_i \rightsquigarrow \mathcal{N}(0, \sigma^2)$$

- For the variable Y , we have:

$$\mathbb{E}(y_i) = \beta_0 + \beta_1 x_i + \mathbb{E}(\varepsilon_i) = \beta_0 + \beta_1 x_i.$$
$$\mathbb{V}(y_i) = \mathbb{V}(\beta_0 + \beta_1 x_i + \varepsilon_i) = \mathbb{V}(\varepsilon_i) = \sigma^2.$$

- The normality of ε_i implies the normality of y_i.
- The independence of the ε_i implies the independence of the y_i. In fact:

$$\begin{aligned} cov(y_i, y_j) &= cov(\beta_0 + \beta_1 x_i + \varepsilon_i; \beta_0 + \beta_1 x_j + \varepsilon_j) \\ &= cov(\varepsilon_i, \varepsilon_j) \\ &= 0 \end{aligned}$$

Distribution of least squares estimators

- The estimators obtained by the least squares method are given by:

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2} \quad \text{et} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

- These estimators are random variables because they depend on y_i.
- They are linear functions of the y_i.
- They are normally distributed.
- In order to know their distribution, it is necessary to calculate their expectation and their variance.

Mathematical expectation of $\hat{\beta}_1$

- We have

$$
\begin{aligned}
\mathbb{E}(\hat{\beta}_1) &= \mathbb{E}\left(\frac{\sum(x_i-\bar{x})y_i}{\sum(x_i-\bar{x})^2}\right) \\
&= \frac{\sum(x_i-\bar{x})\mathbb{E}(y_i)}{\sum(x_i-\bar{x})^2} \\
&= \frac{\sum(x_i-\bar{x})(\beta_0+\beta_1 x_i)}{\sum(x_i-\bar{x})^2} \\
&= \frac{\beta_0\sum(x_i-\bar{x})+\beta_1\sum(x_i-\bar{x})x_i}{\sum(x_i-\bar{x})^2} \\
&= \frac{0+\beta_1\sum(x_i-\bar{x})x_i}{\sum(x_i-\bar{x})^2} \\
&= \frac{\beta_1\sum(x_i-\bar{x})x_i}{\sum(x_i-\bar{x})x_i} \\
&= \beta_1
\end{aligned}
$$

Mathematical expectation of $\hat{\beta}_0$

- We have

$$\begin{aligned}\mathbb{E}(\hat{\beta}_0) &= \mathbb{E}(\bar{y} - \hat{\beta}_1\bar{x}) \\ &= \mathbb{E}(\bar{y}) - \bar{x}\mathbb{E}(\hat{\beta}_1) \\ &= \mathbb{E}(\bar{y}) - \bar{x}\beta_1\end{aligned}$$

$$\begin{aligned}\mathbb{E}(\bar{y}) &= \mathbb{E}\left(\frac{\sum y_i}{n}\right) = \frac{\sum \mathbb{E}(y_i)}{n} = \frac{\sum(\beta_0 + \beta_1 x_i)}{n} \\ &= \frac{n\beta_0 + \beta_1 \sum x_i}{n} = \beta_0 + \beta_1\bar{x}.\end{aligned}$$

$$\mathbb{E}(\hat{\beta}_0) = \beta_0 + \beta_1\bar{x} - \bar{x}\beta_1 = \beta_0.$$

Variance of $\hat{\beta}_1$

- We have

$$\begin{aligned}
\mathbb{V}(\hat{\beta}_1) &= \mathbb{V}\left(\frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2}\right) \\
&= \frac{\sum(x_i - \bar{x})^2 \mathbb{V}(y_i)}{\left(\sum(x_i - \bar{x})^2\right)^2} \\
&= \frac{\sum(x_i - \bar{x})^2 \sigma^2}{\left(\sum(x_i - \bar{x})^2\right)^2} \\
&= \frac{\sigma^2}{\sum(x_i - \bar{x})^2}.
\end{aligned}$$

Variance of $\hat{\beta}_0$

- We have:

$$\begin{aligned}\mathbb{V}(\hat{\beta}_0) &= \mathbb{V}(\bar{y} - \hat{\beta}_1\bar{x}) \\ &= \mathbb{V}(\bar{y}) + \bar{x}^2\mathbb{V}(\hat{\beta}_1) - 2cov(\bar{y}, \hat{\beta}_1)\end{aligned}$$

Moreover, $cov(\bar{y}, \hat{\beta}_1) = 0$ (to be demonstrated) and

$$\mathbb{V}(\bar{y}) = \mathbb{V}\left(\frac{\sum y_i}{n}\right) = \frac{\sum \mathbb{V}(y_i)}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

$$\begin{aligned}\mathbb{V}(\hat{\beta}_0) &= \mathbb{V}(\bar{y}) + \bar{x}^2\mathbb{V}(\hat{\beta}_1) \\ &= \frac{\sigma^2}{n} + \bar{x}^2\frac{\sigma^2}{\sum(x_i - \bar{x})^2} = \frac{\sigma^2\sum x_i^2}{n\sum(x_i - \bar{x})^2}\end{aligned}$$

Covariance of $\hat{\beta}_1$ and $\hat{\beta}_0$

- We have

$$
\begin{aligned}
cov(\hat{\beta}_1, \hat{\beta}_0) &= cov(\hat{\beta}_1, \bar{y} - \hat{\beta}_1 \bar{x}) \\
&= cov(\hat{\beta}_1, \bar{y}) - \bar{x}\mathbb{V}(\hat{\beta}_1) \\
&= 0 - \frac{\bar{x}\ \sigma^2}{\sum (x_i - \bar{x})^2} \\
&= -\frac{\bar{x}\ \sigma^2}{\sum (x_i - \bar{x})^2}.
\end{aligned}
$$

Estimation of error variance

- The variance σ^2 appears in the previous formulas.
- σ^2 is however unknown $\Longrightarrow$ It must be estimated.

Unbiased estimator of $\boldsymbol{\sigma^2}$:

- If the errors ε_i could be observed, σ^2 would be estimated by the quantity:

$$\frac{\sum(\varepsilon_i - \bar{\varepsilon})^2}{n-1}$$

- But the ε_i are not observable $\Longrightarrow$ ε_i can be estimated by e_i such that

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

- σ^2 will be estimated using the sum of the squares of the residuals e_i as the estimator of the squares of the errors:

$$\sum (e_i - \bar{e})^2 = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2$$

- It was seen that:

$$\sum (y_i - \hat{y}_i)^2 = \sum y_i^2 - \sum \hat{y}_i^2.$$

- Moreover,

$$\begin{aligned} \mathbb{E}\left(\sum (y_i - \hat{y}_i)^2\right) &= \sum \mathbb{E}(y_i^2) - \sum \mathbb{E}(\hat{y}_i^2) \\ &= \sum \left(\mathbb{V}(y_i) + \mathbb{E}^2(y_i)\right) - \sum \left(\mathbb{V}(\hat{y}_i) + \mathbb{E}^2(\hat{y}_i)\right) \end{aligned}$$

- Besides

$$\begin{aligned}\mathbb{E}(\hat{y}_i) &= \mathbb{E}(\hat{\beta}_0 + \hat{\beta}_1 x_i) \\ &= \mathbb{E}(\hat{\beta}_0) + x_i \mathbb{E}(\hat{\beta}_1) \\ &= \beta_0 + x_i \beta_1 \\ &= \mathbb{E}(y_i)\end{aligned}$$

- We thus obtain,

$$\begin{aligned}\mathbb{E}(\sum (y_i - \hat{y}_i)^2) &= \sum \mathbb{V}(y_i) - \sum \mathbb{V}(\hat{y}_i) \\ &= n\sigma^2 - \sum \mathbb{V}(\hat{y}_i)\end{aligned}$$

- Regarding the calculation of $\mathbb{V}(\hat{y}_i)$:

$$\begin{aligned}\mathbb{V}(\hat{y}_i) &= \mathbb{V}(\hat{\beta}_0 + \hat{\beta}_1 x_i) \\ &= \mathbb{V}(\hat{\beta}_0) + x_i^2 \mathbb{V}(\hat{\beta}_1) + 2x_i cov(\hat{\beta}_0, \hat{\beta}_1)\end{aligned}$$

- By replacing all the quantities by their values we obtain:

$$\mathbb{V}(\hat{y}_i) = \sigma^2 \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right).$$

- We thus obtain:

$$\begin{aligned} \mathbb{E}(\sum (y_i - \hat{y}_i)^2) &= n\sigma^2 - \sigma^2 \sum \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right) \\ &= \sigma^2 (n-2) \end{aligned}$$

- We can thus define an unbiased estimator of σ^2 by setting:

$$s^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n-2} = \frac{SC_{\text{res}}}{n-2}$$

- **Estimator of σ^2**

$$s^2 = \frac{\sum(y_i - \hat{y}_i)^2}{n-2} = \frac{SC_{\text{res}}}{n-2}$$

- The estimator of $V(\hat{\beta}_1)$ is:

$$s^2(\hat{\beta}_1) = \frac{s^2}{\sum(x_i - \bar{x})^2}$$

- The estimator of $V(\hat{\beta}_0)$ is:

$$s^2(\hat{\beta}_0) = \frac{s^2 \sum x_i^2}{n\sum(x_i - \bar{x})^2}$$

Inference on model parameters

- **Objective**: To test hypotheses on the parameters β_1 and β_0 and construct confidence intervals.
- **The tests:**
- Test and confidence interval on slope β_1.
- Test and confidence interval on the intercept.

Test and confidence interval on slope β_1

- The estimator $\hat{\beta}_1$ is normally distributed such that.

$$\hat{\beta}_1 \rightsquigarrow \mathcal{N}(\mathbb{E}(\hat{\beta}_1), \mathbb{V}(\hat{\beta}_1)) \quad \Rightarrow \quad \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\mathbb{V}(\hat{\beta}_1)}} \rightsquigarrow \mathcal{N}(0,1).$$

- **In practice, $V(\hat{\beta}_1)$ is estimated by $S^2(\hat{\beta}_1)$**
- **The quantity $\frac{\hat{\beta}_1 - \beta_1}{s(\hat{\beta}_1)}$ then follows a Student's law with (n-2) degrees of freedom.**

- To test the hypothesis

$$\mathcal{H}_0 : \beta_1 = b_1 \text{ contre } \mathcal{H}_1 : \beta_1 \neq b_1,$$

we use the $t_c = \dfrac{\hat{\beta}_1 - \beta_1}{s(\hat{\beta}_1)}$ statistic to accept or reject H_0

- We accept H_0 at the significance level α if

$$-t_{(1-\alpha/2,(n-2))} \leq t_c \leq t_{(1-\alpha/2,(n-2))}$$

- An interesting hypothesis test is the test $\mathcal{H}_0 : \beta_1 = 0$
- If we accept this null hypothesis, this means that the y_i do not depend on the x_i and the simple regression model is inadequate.

- The confidence interval on slope β_1 is given by:

$$[\hat{\beta}_1 - t_{(1-\alpha/2,(n-2))} \times s(\hat{\beta}_1) \; ; \; \hat{\beta}_1 + t_{(1-\alpha/2,(n-2))} \times s(\hat{\beta}_1)]$$

Test and confidence interval on slope β_0

- The estimator $\hat{\beta}_0$ is normally distributed such that.

$$\hat{\beta}_0 \rightsquigarrow \mathcal{N}(\mathbb{E}(\hat{\beta}_0), \mathbb{V}(\hat{\beta}_0)) \quad \Rightarrow \quad \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\mathbb{V}(\hat{\beta}_0)}} \rightsquigarrow \mathcal{N}(0,1).$$

- **In practice, $V(\hat{\beta}_0)$ is estimated by $S^2(\hat{\beta}_0)$**
- **The quantity $\frac{\hat{\beta}_0 - \beta_0}{s(\hat{\beta}_0)}$ then follows a Student's law with (n-2) degrees of freedom.**

- To test the hypothesis

$$\mathcal{H}_0 : \beta_0 = b_0 \text{ contre } \mathcal{H}_1 : \beta_0 \neq b_0,$$

we use the $t_c = \dfrac{\hat{\beta}_0 - \beta_0}{s(\hat{\beta}_0)}$ statistic to accept or reject H_0

- We accept H_0 at the significance level α if

$$-t_{(1-\alpha/2,(n-2))} \leq t_c \leq t_{(1-\alpha/2,(n-2))}$$

- An interesting hypothesis test is the test $\mathcal{H}_0 : \beta_1 = 0$
- If we accept this null hypothesis this means that the ordinate of the regression line passes through the origin.

- The confidence interval on the slope β_0 is given by:

$$[\hat{\beta}_0 - t_{(1-\alpha/2,(n-2))} \times s(\hat{\beta}_0) \; ; \; \hat{\beta}_0 + t_{(1-\alpha/2,(n-2))} \times s(\hat{\beta}_0)]$$

Confidence interval for $\mu_Y(\mathrm{X})$

- Objective: Find a confidence interval for the ordinate of the abscissa point x which lies on the regression line. i.e. finding a confidence interval for

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

- The estimator of $\mu_Y(\mathrm{X})$ is given by the regression line:

$$\hat{\mu}_Y(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

- This estimator is an unbiased and normally distributed estimator of variance:

$$\mathbb{V}(\hat{\mu}_Y(x)) = \sigma^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right)$$

- This variance is all the greater the further $\bar{x} \Longrightarrow$ do not use the regression line to estimate $\mu_Y(x)$ for values x that are too far from $\bar{x}$.
- $V(\hat{\mu}_Y(x))$ is unknown, it is estimated by:

$$s^2(\hat{\mu}_Y(x)) = s^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)$$

- The quantity:

$$\frac{\hat{\mu}_Y(x) - \mu_Y(x)}{s(\hat{\mu}_Y(x))} \rightsquigarrow t(n-2)$$

- The confidence interval of $\mu_Y(x)$ is:

$$\left[\hat{\mu}_Y(x) \pm t_{(1-\alpha/2,(n-2))} s(\hat{\mu}_Y(x))\right]$$

Variance analysis

- Objective: Test the hypothesis $\mathcal{H}_0 : \beta_1 = 0$ using the three sums of squares SCtot , SCreg and SCres .

- If the hypothesis H_0 is verified, the expectations of the three sums of squares are:

$$\begin{aligned} \mathbb{E}(SC_{tot}) &= (n-1)\sigma^2 \\ \mathbb{E}(SC_{reg}) &= \sigma^2 \\ \mathbb{E}(SC_{res}) &= (n-2)\sigma^2 \end{aligned}$$

Variance analysis

- The following quantities are unbiased estimators of σ^2 called mean squares:

$$MC_{tot} = \frac{SC_{tot}}{n-1}$$
$$MC_{reg} = \frac{SC_{reg}}{1}$$
$$MC_{res} = \frac{SC_{res}}{n-2}$$

- If H_0 is not verified, only MCres is always an unbiased estimator of σ^2.

$$s^2 = MC_{res} = \frac{SC_{res}}{n-2}$$

- The denominators of the three estimators are called the degrees of freedom (df).
  - n - 1 is the df associated with SCtot
  - 1 is the df associated with SCreg
  - n - 2 is the df associated with SCres
- These numbers correspond to the number of linearly independent terms involved in each of these sums.
- These degrees of freedom are based on the n independent observations $(y_1, \ldots, y_n)$ and on the number of parameters to be estimated.

- The total sum of squares requires (n - 1) independent terms since $\sum(y_i - \bar{y}) = 0.$

- The sum of squares of the regression is equal to 1 since it can be calculated using a single function of $y_1, \ldots, y_n$ be $\sum(\hat{y}_i - \bar{y})^2 = \hat{\beta}_1^2 \sum(x_i - \bar{x})^2.$

- The residual sum of squares equals n - 2 since the subtraction of (n -1) -1 gives the df of SCres .

- When the hypothesis H0 is verified, we have

$$\frac{SC_{tot}}{\sigma^2} \rightsquigarrow \mathcal{X}^2(n-1)$$

$$\frac{SC_{reg}}{\sigma^2} \rightsquigarrow \mathcal{X}^2(1)$$

$$\frac{SC_{res}}{\sigma^2} \rightsquigarrow \mathcal{X}^2(n-2)$$

- The $\frac{SC_{reg}}{\sigma^2}$ and $\frac{SC_{res}}{\sigma^2}$ quantities are independent.

- When H_0 holds, the statistic:

$$F_c = \frac{\frac{SC_{reg}}{\sigma^2}}{\frac{SC_{res}}{(n-2)\sigma^2}} = \frac{MC_{reg}}{MC_{res}} \rightsquigarrow \text{Fisher}(1,(n-2))$$

The Fc statistic is used to decide the acceptance or rejection of H_0.

- **Reject region**

We reject H_0 at the significance level α if $F_c > f_\alpha(1, n-2)$

ANOVA Table

Source of variation	df	Sum of squares	Mean of squares	Fc
Regression	1	SC_{reg}	MC_{reg}	$\frac{MC_{reg}}{MC_{res}}$
Residual	n-2	SC_{res}	MC_{res}	
Total	n-1	SC_{tot}		

Models with a single parameter

- If the hypotheses $\beta_0 = 0$ or $\beta_1 = 0$, the model will be a model with a single parameter.
  - Model without explanatory variable
  - Model without constant

Model without explanatory variable

- The model is written in the form:

$$y_i = \beta_0 + \varepsilon_i$$

- The y_i are independent and normally distributed with

$$\mathbb{E}(y_i) = \beta_0 \text{ et } \mathbb{V}(y_i) = \sigma^2$$

- The estimated values:

$$\hat{y}_i = \hat{\beta}_0$$

and the residues:

$$e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0$$

- The estimator of $\hat{\beta}_0$ is given by:

$$\hat{\beta}_0 = \frac{\sum y_i}{n} = \bar{y}$$

- So we have:

$$SC_{res} = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - \bar{y})^2 = SC_{tot}$$

$$SC_{reg} = \sum (\hat{y}_i - \bar{y})^2 = \sum (\hat{y}_i - \hat{y}_i)^2 = 0$$

- Moreover:

$$\mathbb{E}(\hat{\beta}_0) = \beta_0$$

$$\mathbb{V}(\hat{\beta}_0) = \frac{\sigma^2}{n}$$

- An unbiased estimator of σ^2 is given by:

$$s^2 = \frac{\sum(y_i - \bar{y})^2}{n-1}$$

- An unbiased estimator of $V(\hat{\beta}_0)$ is given by

$$\mathbb{V}(\hat{\beta}_0) = \frac{s^2}{n}$$

- The quantity

$$\frac{\hat{\beta}_0 - \beta_0}{s(\hat{\beta}_0)} \rightsquigarrow t(n-1)$$

Model without constant

- The model is written in the form:

$$y_i = \beta_1 x_i + \varepsilon_i$$

- The y_i are independent and normally distributed with

$$\mathbb{E}(y_i) = \beta_1 x_i \text{ et } \mathbb{V}(y_i) = \sigma^2$$

- The estimated values:

$$\hat{y}_i = \hat{\beta}_1 x_i$$

and the residues:

$$e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_1 x_i$$

- The estimator of $\hat{\beta}_1$ is given by:

$$\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2}$$

- The regression line for this model does not pass through the point $(\bar{x}, \bar{y})$.
- We have $\sum y_i \neq \sum \hat{y}_i \Rightarrow \sum e_i \neq 0$
- We also don't have equality $SC_{tot} = SC_{reg} + SC_{res}$

- On the other hand, we always have

$$\sum \hat{y}_i^2 = \sum \hat{y}_i y_i$$

- Moreover, we have

$$\sum (y_i - \hat{y}_i)^2 = \sum y_i^2 - \sum \hat{y}_i^2$$

- So we have

$$SC_{tot} = \sum y_i^2$$

$$SC_{reg} = \sum \hat{y}_i^2$$

- Moreover:

$$\mathbb{E}(\hat{\beta}_1) = \beta_1$$

$$\mathbb{V}(\hat{\beta}_1) = \frac{\sigma^2}{\sum x_i^2}$$

- An unbiased estimator of σ^2 is given by:

$$s^2 = \frac{\sum(y_i - \hat{y}_i)^2}{n-1}$$

- An unbiased estimator of $V(\hat{\beta}_1)$ is given by

$$\mathbb{V}(\hat{\beta}_0) = \frac{s^2}{\sum x_i^2}$$

- The quantity

$$\frac{\hat{\beta}_1 - \beta_1}{s(\hat{\beta}_1)} \rightsquigarrow t(n-1)$$