Machine learning and statistical learning

2.1 Linear Regression

Ewen Gallic
ewen.gallic@gmail.com

MASTER in Economics - Track EBDS - 2nd Year

2020-2021

This part presents some concepts of statistical learning, through the prism of regression.

1. Some context

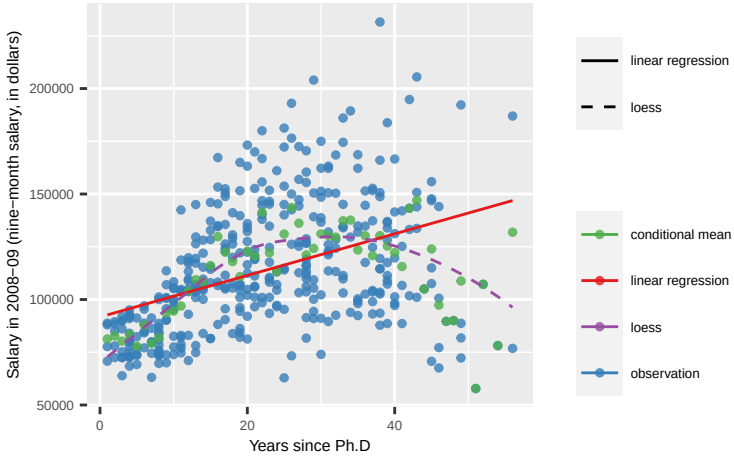
Model specification

In a regression problem, the aim is to understand how a **response variable** y varies, **conditionally** on the available information on some **predictors** $\mathbf{x}$.

Let us take an example, that of the salaries for Professors in the US in 2008-09.

The salary of a professor may be linked, among other things, to the number of years since he or she obtained their Ph.D.

Salary as a function of years since Ph.D



Salary as a function of years since Ph.D

Here, the linear regression suggests that on the average, the salary increases with the number of years since Ph.D:

- the slope of 985.3 indicates that each additional year since Ph.D leads to an increase of 985 dollars of 9-month salary.

But the relationship does not seem to be linear...

Salary as a function of years since Ph.D

It should be noted here that:

- the regression analysis does not depend on a **generative model** here (a model explaining how the data are generated)
- there is no **causal** claims regarding the way mean salary would change if the number of years since Ph.D is altered
- there is no statistical inference

We could **add some predictors** to the model to get a better story on what is going on with salary :

- some ommitted variables may play an important role in explaining the variations.

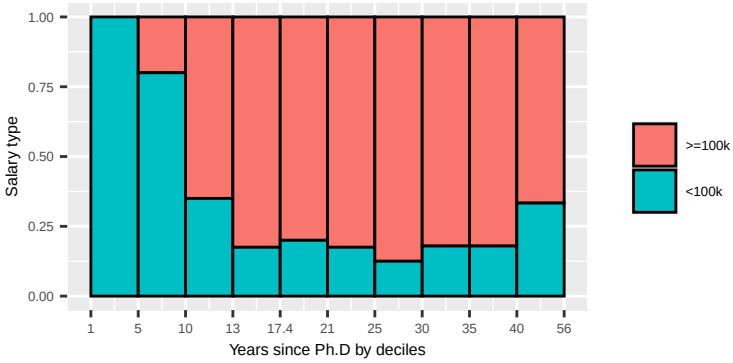
Salary as a function of years since Ph.D

We can also perform some regression analysis if the **response variable is categorical**.

Let us look at the salary in a different way: let us split it into two categories, either $< \$100k$ or $\geq \$100k$.

For each decile of years since Ph.D, we can plot the **conditional proportions**.

Salary as a function of years since Ph.D



Levels of regression analysis

Berk (2008) mentions **three levels of regression analysis**:

- Level I regression analysis:
  - aiming at **describing the data**
  - assumption free
  - should not be neglected
- Level II regression analysis:
  - based on **statistical inference**
  - uses results from level I regression analysis
  - use with real data may be challenging
  - allows to make predictions
- Level III regression analysis:
  - based on **causal inference**
  - uses level I analysis, sometimes coupled with level II
  - rely more on algorithmic methods rather than model-based methods.

2. The linear regression

Some references

- Berk (2008). Statistical learning from a regression perspective, volume 14. Springer.
- Cornillon and Matzner-Løber (2007). Régression: théorie et applications. Springer.
- James et al. (2013). An introduction to statistical learning, volume 112. Springer.

The linear regression

Linear regression combines level I and level II perspectives.

It is useful when one wants to **predict a quantitative response**.

A lot of newer statistical learning approaches can be seen as generalizations or extensions of linear regression, as reminded in James et al. (2013).

2.1 Simple linear regression

Principle

Let us consider first the case of **simple linear regression**.

We aim at **predicting a quantitative response variable** y using a single **predictor** $\mathbf{x}$ (or **regressor**).

- y is a $n \times 1$ numerical response variable, where n represents the numbr of observations
- $\mathbf{x}$ is a $n \times 1$ predictor.

We assume there exists a linear relationship between y and $\mathbf{x}$ such that:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where ε_i is an error term normally distributed with 0 mean and variance σ^2, *i.e*, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

Principle

In Eq. 1, the **coefficients** (or **parameters**) β_0 (*i.e.*, the constant) and β_1 (*i.e.*, the slope) are unknown parameters to be estimated.

These coefficients are **estimated** using a **training sample**.

The estimates of β_0 and β_1 are, respectively, $\hat{\beta}_0$ and $\hat{\beta}_1$.

Once they are estimated using a learning procedure (in this case using linear regression), they can used to **predict** values for y for some value x_0:

$$\hat{y_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0 \tag{2}$$

2.1.1 Estimating the coefficients

Estimating the coefficients

To estimate β_0 and β_1, we rely on a set of training examples $\{(x_1, y_1), \ldots, (x_n, y_n)\}$.

For example, let us go back to our data describing the 9 month salary of professors (the response variable) and look at the relationship between the salary and years since Ph.D ($\mathbf{x}$).

Estimating the coefficients

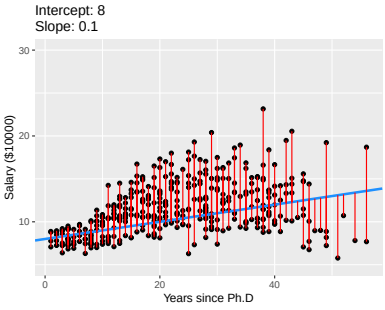


Figure 1: Varying the intercept.

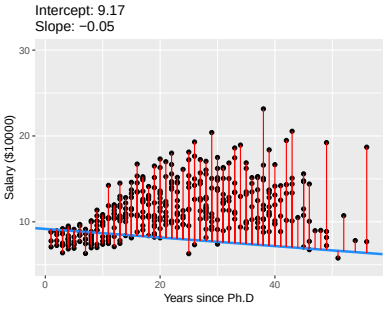


Figure 2: Varying the slope.

There is an infinity of possibles values that one can pick for $\hat{\beta}_0$ and $\hat{\beta}_1$.

However, we want to find an estimation that leads to a line being **as close as possible to the points**: but what does "close" mean?

Estimating the coefficients

The most common metric we want to minimize is known as the **least square criterion**.

The predictions $\hat{y}_i$ for each of the x_i, $i = 1, \ldots, n$ are given by $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

Let $e_i = y_i - \hat{y}_i$ the ith **residual**, *i.e.*, the difference between the osberves value and its prediction by the linear model.

The **residual sum of square** is defined as:

$$\text{RSS} = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_1 x_i - \hat{\beta}_0)^2. \tag{3}$$

We aim at **minimizing this metric**.

Least squares coefficient estimates

It can easily be shown that the minimization of the RSS leads to:

$$\begin{cases} \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \end{cases} \tag{4}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Least squares coefficient estimates

Here, the least squares coefficient estimates $\hat{\beta}_0$ and β_1 are, respectively, 9.1719 and 0.0985.

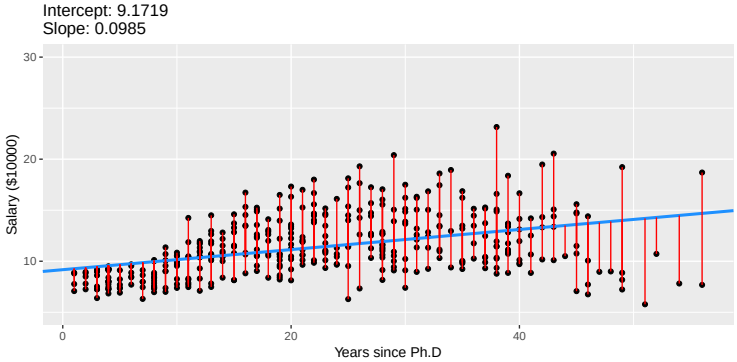


Figure 3: Fit of the Least Square for the regression of years since Ph.D onto the 9 months salaryof Professors.

Residual sum of squares

We can have a look at the RSS when we vary the values of $\hat{\beta}_0$ and $\hat{\beta}_1$:

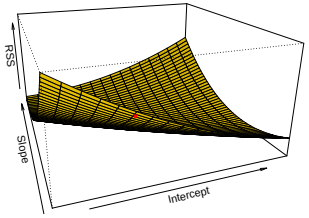


Figure 4: Surface plot of the RSS depending on the values of $\hat{\beta}_0$ and $\hat{\beta}_1$.

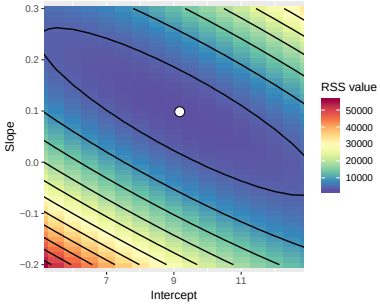


Figure 5: Contour plot of the RSS depending on the values of $\hat{\beta}_0$ and $\hat{\beta}_1$.

2.1.2 Accuracy of the coefficient estimates

Accuracy of the coefficient estimates

The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are point estimates.

When they are estimated by least squares, they are:

- **unbiased**
  - $\mathbb{E}(\hat{\beta}_0) = \beta_0$ and $\mathbb{E}(\hat{\beta}_1) = \beta_1$
- **efficient**
  - $\mathbb{V}(\hat{\beta}_0)$ and $\mathbb{V}(\hat{\beta}_1)$ are minimal
- **convergent**
  - $\lim_{n \to +\infty} \mathbb{V}(\hat{\beta}_0) = 0$ and $\lim_{n \to +\infty} \mathbb{V}(\hat{\beta}_1) = 0$

They are called **BLUE** (Best Linear Unbiased Estimator).

Accuracy of the coefficient estimates

It is easy to show that:

$$\begin{cases} \mathbb{V}(\hat{\beta}_0) & = \sigma^2 \left[\frac{1}{n} + \frac{\overline{x}}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right] \\ \mathbb{V}(\hat{\beta}_1) & = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \end{cases} \tag{5}$$

where σ^2 can be estimated:

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

Accuracy of the coefficient estimates

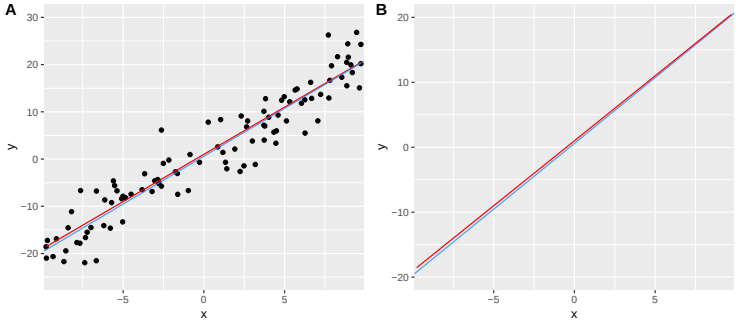


Figure 6: A: True relationship (in red), Observed values of y (points) and Least Squares line (in blue). B: True relationship (in red), Current Least Squares line (in blue), Previous Least Squares lines (in gray).

Accuracy of the coefficient estimates

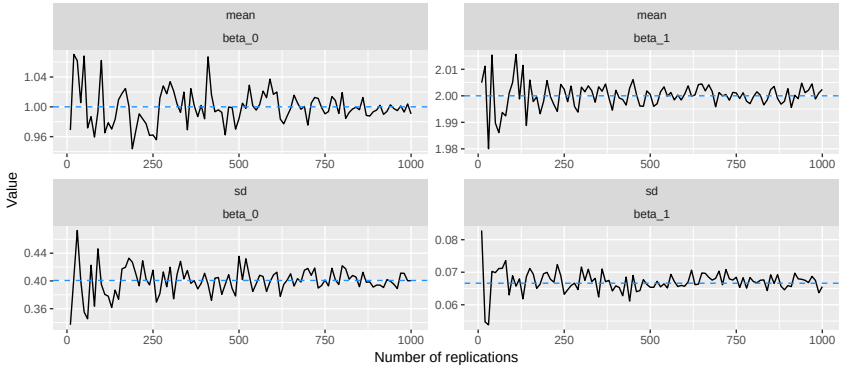


Figure 7: Mean of estimates of β_0 and β_1 depending on the number of resampling.

Hypothesis tests

We wish to test if a coefficient θ, $\theta \in \{\beta_0, \beta_1\}$ is equal to a specific value θ_0:

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{cases}$$

We know that $\hat{\theta} \sim \mathcal{N}\left(\theta, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$, so:

$$\frac{\hat{\theta} - \theta}{\sigma / \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \sim \mathcal{N}(0, 1).$$

Hypothesis tests

As $\frac{\sum_{i=1}^{n} \varepsilon_i^2}{\sigma^2} \sim \chi^2_{n-2}$, we can define a variable T as:

$$T = \frac{\frac{\hat{\theta}-\theta}{\sigma/\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}}{\sqrt{\frac{\sum_{i=1}^{n} \varepsilon_i^2}{\sigma_u^2}}/\sqrt{n-2}} \sim \mathcal{S}t(n-2)$$

We can show that the expression of T can be simplified to:

$$T = \frac{\hat{\theta} - \theta}{\hat{\sigma}_{\hat{\theta}}}$$

Hypothesis tests

It is thus possible to perform the following test:

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{cases}$$

knowing that $\frac{\hat{\theta}-\theta}{\hat{\sigma_\theta}} \sim \mathcal{S}t(n-2)$

Hypothesis tests

And we need to find the following probability:

$$\mathbb{P}\left(-t_{\alpha/2} < \frac{\hat{\theta} - \theta}{\hat{\sigma}_{\hat{\theta}}} < t_{\alpha/2}\right)$$

We therefore need to compute a t-statistic, that measures the number of standard deviations that $\hat{\theta}$ is away from θ_0:

$$t_{\text{obs.}} = \frac{\hat{\theta} - \theta_0}{\hat{\sigma}_{\hat{\theta}}}$$

- if $t_{\text{obs.}} \in \left[-t_{\alpha/2}, t_{\alpha/2}\right]$:
  - we **do not reject the null hypothesis** (H_0) with a first-order risk of $\alpha\%$
- if $t_{\text{obs.}} \notin \left[-t_{\alpha/2}, t_{\alpha/2}\right]$:
  - we **reject the null hypothesis** (H_0) with a first-order risk of $\alpha\%$

Hypothesis tests

Most of the time, we are interested in a specific case:

$$\begin{cases} H_0 : \alpha = 0 \\ H_1 : \alpha \neq 0, \end{cases}$$

In such a case, the t-statistic becomes:

$$T = \frac{\hat{\theta} - 0}{\hat{\sigma}_\theta} = \frac{\hat{\theta}}{\hat{\sigma}_{\hat{\theta}}}$$

The observed value is $t_{\text{obs.}} = \frac{\hat{\alpha}}{\hat{\sigma}_{\hat{\alpha}}}$.

Hypothesis tests: confidence interval

We can also use the standard error of the coefficient estimates to construct a **confidence interval**:

$$\widehat{\text{I.C.}_\theta(1-\alpha)} = \left[\hat{\theta} \pm t_{\alpha/2} \times \hat{\sigma}_{\hat{\theta}}\right]. \tag{6}$$

If the intervals contain 0, then we can conclude that the coefficient θ is not statistically different from zero (at the $\alpha\%$ level of significance).

We can also compute the probability of observing any number equal to $\mid t \mid$ or larger while assuming $\theta = 0$ (this probability is known as **the p-value**).

Hypothesis tests

	Least squares
(Intercept)	9.17^{***}
	(0.28)
yrs.since.phd	0.10^{***}
	(0.01)
R^2	0.18
Adj. R^2	0.17
Num. obs.	397

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 1: Statistical models

2.1.3 Accuracy of the model

Accuracy of the model

Recall that the linear regression is a supervised learning method. Hence, we can compare the predictions we obtain with the observed values of the output variable.

We want to have an idea of the quality of the estimation, to know how well the model fits the data.

To that end, we usually use several **metrics**, among which:

- the root mean squared error (RMSE)
- the residual standard error (RSE)
- the R^2 statistic.

Accuracy of the model: RMSE

The **mean squared error** (MSE) is an estimate of the **average of the squares of the errors**:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{7}$$

The **root mean squared error** is the square root of the MSE:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} = \sqrt{\frac{\text{RSS}}{n}}, \tag{8}$$

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

Accuracy of the model: RMSE

The value of the RMSE is always non-negative.

A value of 0 indicates a perfect fit to the data.

Accuracy of the model: RSE

Recall that the linear model contains an error term (ε). Hence, we will not be able to perfectly predict the response variable.

The **Residual Standard Error** is the **average amount that the response will deviate from the true regression line**. It is an estimate of the standard deviation of ε:

$$\text{RSE} = \sqrt{\frac{1}{n-2}RSS} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \tag{9}$$

Accuracy of the model: RSE

In our example of the regression of salaries onto years since Ph.D, the value of the RSE is 2.7534.

This means that the actual salary can deviate from the true regression line by approximately 2.7534 thousand dollars, on average.

The mean salary in the data is $\$11.37065$ thousand dollars. Hence, the percentage error for any prediction, using our estimation would be $2.7534/11.37065 \approx 25\%$.

Accuracy of the model: R^2

Now, let us turn to the R^2 statistic, which provides another method to assess the quality of fit.

The R^2 measures the **proportion of variance explained**. It takes a value between 0 and 1.

Let us illustrate this.

Accuracy of the model: R^2

The variations of y are only partially explained by those of $\mathbf{x}$

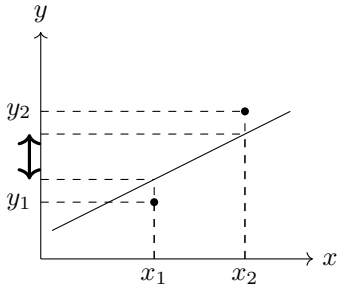


Figure 8: Variation from y_2 to y_1

Accuracy of the model: R^2

As shown in Figure 8, the variation from y_1 to y_2 is partially explained by the variation from x_1 to x_2.

The **quality** of fit at each point, as measured by the total variation, can therefore be broken down into two parts:

- the **explained variation**
- the **residual variation**

using the average point $(\overline{x}, \overline{y})$ as reference, *i.e.*:

$$\underbrace{y_i - \overline{y}}_{\text{total variation}} = \underbrace{\hat{y}_i - \overline{y}}_{\text{explained variation}} + \underbrace{y_i - \hat{y}_i}_{\text{residual variation}} .$$

Accuracy of the model: R^2

The closer $\hat{A}$ is to A, the stronger the explained variation is, relatively.

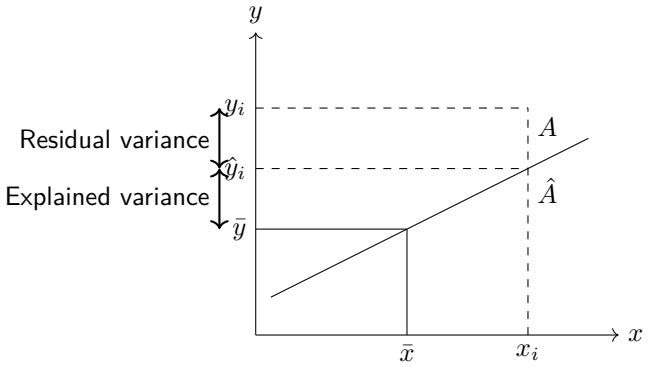


Figure 9: Decomposition of the variation.

Accuracy of the model: R^2

Thus, one way to assess the quality of the adjustment is to measure the following ratio:

$$\frac{\text{explained variance}}{\text{total variance}}$$

Or, for all observations:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\text{ESS}}{\text{TSS}} = \frac{\text{explained sum of squares}}{\text{total sum of squares}} \tag{10}$$

Accuracy of the model: R^2

We can write the R^2 differently, as we know that:

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

Thus:

$$\begin{aligned}R^2&=\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2-\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}\\&=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}=1-\frac{\text{RSS}}{\text{TSS}}\end{aligned}\tag{11}$$

Accuracy of the model: R^2

The value of the R^2 lies between 0 and 1:

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} \qquad \Rightarrow \qquad 0 \leq R^2 \leq 1.$$

- When the economic **theory** suggests that the relationship between the response and its predictor should be **linear**, we expect the value of the R^2 to be really close to one, otherwise, it suggests there might be something wrong with the generation of the data.
- In other situations, when the **linear relationship** can be at best a rough approximation of the real form, we expect to find low values of the R^2.

R^2 and correlation

It can be noted that in the case of simple linear regression, the R^2 is equal to the squared correlation coefficient.

Indeed:

$$\begin{aligned} y_i - \hat{y}_i &= y_i - \bar{y} + \bar{y} - \hat{y}_i \\ &= (y_i - \bar{y}) - (\hat{y}_i - \bar{y}) \\ &= (y_i - \bar{y}) - \left(\hat{\beta}_1 x_i + \hat{\beta}_0 - \hat{\beta}_1 \bar{x} - \hat{\beta}_0\right) \\ &= (y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x}). \end{aligned}$$

Taking the squared value:

$$(y_i - \hat{y}_i)^2 = (y_i - \bar{y})^2 + \hat{\beta}_1^{\,2} (x_i - \bar{x})^2 - 2\hat{\beta}_1 (y_i - \bar{y})(x_i - \bar{x})$$

R^2 and correlation

Which leads to:

$$(y_i - \hat{y}_i)^2 = (y_i - \bar{y})^2 + \hat{\beta_1}^2(x_i - \bar{x})^2 - 2\hat{\beta_1}(y_i - \bar{y})(x_i - \bar{x})$$

Summing on all individuals:

$$\begin{aligned}
\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 &= \sum_{i=1}^{n}(y_i - \bar{y})^2 + \hat{\beta_1}^2\sum_{i=1}^{n}(x_i - \bar{x})^2 - 2\hat{\beta_1}\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) \\
&= \sum_{i=1}^{n}(y_i - \bar{y})^2 + \hat{\beta_1}^2\sum_{i=1}^{n}(x_i - \bar{x})^2 - 2\hat{\beta_1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \sum_{i=1}^{n}(y_i - \bar{y})^2 - \hat{\beta_1}^2\sum_{i=1}^{n}(x_i - \bar{x})^2
\end{aligned}$$

R^2 and correlation

In can indeed be shown that

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

We also have:

$$(\hat{y}_i - \bar{y}) = \hat{\beta}_1 x_i + \hat{\beta}_0 - \hat{\beta}_1 \bar{x} - \hat{\beta}_0 = \hat{\beta}_1 (x_i - \bar{x}).$$

By taking the squared value and summing for all individuals:

$$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \hat{\beta_1}^2 \sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{12}$$

R^2 and correlation

Then, introducing (12) in (10), we get:

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

$$= \underbrace{\left(\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)^2}_{\hat{\alpha}^2} \times \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

$$R^2 = \frac{\left(\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \right)^2}{\left(\sum_{i=1}^n (x_i - \bar{x})^2 \right) \left(\sum_{i=1}^n (y_i - \bar{y})^2 \right)} \tag{13}$$

$$= \frac{\left(\mathbb{C}ov(x, y) \right)^2}{\mathbb{V}(x) \times \mathbb{V}(y)} \tag{14}$$

2.2 Multiple linear regression

Principle

We have considered so far only one predictor in the design matrix $\mathbf{x}$. Let us now look at the case where we want to use **multiple predictors**: $\mathbf{x}$ becomes a $n \times p$ matrix, with n observations and p predictors.

We assume there exists a relationship between the response y and the predictors $\mathbf{x}$ such that:

$$y_i = \beta_0 + \beta_1 \mathbf{x}_{1i} + \ldots + \beta_p \mathbf{x}_{pi} + \varepsilon_i, \quad i = 1, \ldots, n, \tag{15}$$

where ε_i is an error term normally distributed with 0 mean and variance σ^2, *i.e*, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, and where $\mathbf{x}_{ji}$ represents the ith observation for the jth predictor, $j = 1, \ldots, p$.

Principle

In Eq. 22 the **coefficients** β_0 (*i.e.*, the constant) and β_j are unknown parameters to be estimated.

We interpret the coefficients β_j as the average effect on y of a one unit increase in $\mathbf{x}_j$, *ceteris paribus* (*i.e.*, holding all other predictors fixed).

2.2.1 Estimating the coefficients

Estimating the coefficients

The coefficients of the multiple linear regression, can once again be estimated so that they minimize the sum of squared residuals:

$$\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{16}$$

$$= \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 \mathbf{x}_{i1} - \ldots - \hat{\beta}_p \mathbf{x}_{ip}), \tag{17}$$

where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \mathbf{x}_{i1} + \ldots + \hat{\beta}_p \mathbf{x}_{ip}$

Estimating the coefficients

Using matrix algebra, it is easy to estimate the coefficients.

First, we can write:

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}, \tag{18}$$

$$\text{where } \hat{\boldsymbol{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} \mathbf{x}_{11} & \mathbf{x}_{12} & \cdots & \mathbf{x}_{1p} & 1 \\ \mathbf{x}_{21} & \mathbf{x}_{22} & \cdots & \mathbf{x}_{2p} & 1 \\ \vdots & \vdots & \ddots & \vdots & 1 \\ \mathbf{x}_{n1} & \mathbf{x}_{n2} & \cdots & \mathbf{x}_{np} & 1 \end{bmatrix}, \text{ and } \hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_p \\ \hat{\beta}_0 \end{bmatrix}.$$

Estimating the coefficients

$$\text{Let} \boldsymbol{y} - \hat{\boldsymbol{y}} \text{ denote the column vector } \boldsymbol{y} - \hat{\boldsymbol{y}} = \begin{bmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ \vdots \\ y_n - \hat{y}_n \end{bmatrix},$$

$$\text{and } \boldsymbol{y} - \hat{\boldsymbol{y}}^\top \text{ the vector column } \boldsymbol{y} - \hat{\boldsymbol{y}}\top = \begin{bmatrix} y_1 - \hat{y}_1 & y_2 - \hat{y}_2 & \cdots & y_n - \hat{y}_n \end{bmatrix}^\top.$$

By definition:

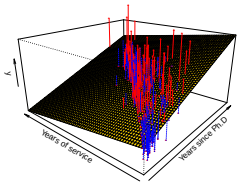
$$(\boldsymbol{y} - \hat{\boldsymbol{y}})\top(\boldsymbol{y} - \hat{\boldsymbol{y}}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 = \text{RSS}.$$

Estimating the coefficients

By replacing y by its expression given in Eq. 18:

$$\begin{aligned}\sum_{i=1}^{n} e_i^2 &= (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \\ &= (\boldsymbol{y}^\top - \hat{\boldsymbol{\beta}}^\top \boldsymbol{X}^\top)(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \\ &= \boldsymbol{y}^\top \boldsymbol{y} - \boldsymbol{y}^\top \boldsymbol{X}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^\top \boldsymbol{X}^\top \boldsymbol{y} + \hat{\boldsymbol{\beta}}^\top \boldsymbol{X}^\top \boldsymbol{X}\hat{\boldsymbol{\beta}}\end{aligned}$$

Estimating the coefficients



- Regression of salaries on years since Ph.D and years of service.
- The red dots represent the observed values. The plane minimizes the sum of squared distances represented by the red (overstimated values) and blue segments (underestimted values).

Estimating the coefficients

	Model 1	Model 2
(Intercept)	9.17^{***}	8.99^{***}
	(0.28)	(0.28)
yrs.since.phd	0.10^{***}	0.16^{***}
	(0.01)	(0.03)
yrs.service		-0.06^{*}
		(0.03)
R^2	0.18	0.19
Adj. R^2	0.17	0.18
Num. obs.	397	397

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 2: Statistical models

Bias of the coefficients

Let us look at the **bias of the coefficients**. First, we can write the estimated vector of coefficients $\hat{\boldsymbol{\beta}}$ as:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y} \\ &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top (\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\ &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1}(\boldsymbol{X}^\top \boldsymbol{X})\boldsymbol{\beta} + (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{\varepsilon} \\ &= \boldsymbol{\beta} + (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{\varepsilon}\end{aligned}$$

Bias of the coefficients

Hence the expected value of $\hat{\boldsymbol{\beta}}$ is given by:

$$\begin{aligned}\mathbb{E}(\hat{\boldsymbol{\beta}}) &= \mathbb{E}(\boldsymbol{\beta}) + \mathbb{E}\left[(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{\varepsilon}\right] \\ &= \boldsymbol{\beta} + (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \mathbb{E}(\boldsymbol{\varepsilon}) \\ &= \boldsymbol{\beta} \end{aligned} \tag{19}$$

since we assumed $\mathbb{E}(\varepsilon) = 0$.

As a consequence, $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of β, *i.e.*:

$$\mathcal{B}\left(\hat{\boldsymbol{\beta}}; \boldsymbol{\beta}\right) = \mathbb{E}\left(\hat{\boldsymbol{\beta}}\right) - \boldsymbol{\beta} = 0$$

Variance of the coefficients

The variance of the coefficients writes:

$$\begin{aligned}\mathbb{V}(\hat{\boldsymbol{\beta}}) &= \mathbb{E}\left[\hat{\boldsymbol{\beta}} - \mathbb{E}(\hat{\boldsymbol{\beta}})\right]^2 \\ &= \mathbb{E}\left[(\hat{\boldsymbol{\beta}} - \mathbb{E}(\boldsymbol{\beta}))(\hat{\boldsymbol{\beta}} - \mathbb{E}(\boldsymbol{\beta}))^\top\right] \\ &= \mathbb{E}\left[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top\right].\end{aligned}$$

Since:

$$\begin{aligned}\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon} \\ (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top &= \boldsymbol{\varepsilon}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}.\end{aligned}$$

Hence:

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}.$$

Variance of the coefficients

So in the end:

$$\begin{aligned}\mathbb{V}(\hat{\boldsymbol{\beta}}) &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \mathbb{E}(\boldsymbol{e}\boldsymbol{e}^\top) \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \\ &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \sigma_u^2 \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \\ &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} (\boldsymbol{X}^\top \boldsymbol{X}) (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \sigma_\varepsilon^2 \\ &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \sigma_\varepsilon^2. \end{aligned} \tag{20}$$

So the variance of $\hat{\boldsymbol{\beta}}$ is equal to the variance of ε multiplied by the ith term of the diagonal of $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$.

2.2.2 Accuracy of the estimation

Accuracy of the estimation

As in the simple linear regression case, we are interested in measuring the **accuracy of the estimation**.

In particular, we will look at the following aspects:

- is there a significant relationship between the response and the predictors?
- which predictors should be kept in the model, and which should be discarded?
- how well does the model fit to the data?

Relationship between y and $\mathbf{x}$

To infer whether there is a relationship between the response y and the predictors $\mathbf{x}$, a statitiscal test can be performes. The **null hypothesis** of this test writes:

$$\mathsf{H}_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$$

The alternative writes:

$$\mathsf{H}_1 : \text{ at least one } \beta_j \text{ is non-zero, } j = 1, \ldots, p$$

This test is based on the following **F-statistic**:

$$F = \frac{(\mathsf{TSS} - \mathsf{RSS})/p}{\mathsf{RSS}/(n-p-1)} \sim \mathcal{F}(p, n-p-1). \tag{21}$$

Relationship between y and $\mathbf{x}$

- If the linear model assumptions are correct:
  - $\mathbb{E}(\text{RSS}/(n-p-1)) = \sigma^2$
- and if H_0 is true:
  - $\mathbb{E}\left[(\text{TSS}-\text{RSS})/p\right] = \sigma^2$

As a consequence:

- when there is **no relationship** between the response and its predictors:
  - the value of the F-statistic should be close to **zero**
- when **H_1 is true**:
  - $\mathbb{E}\left[(\text{TSS}-\text{RSS})/p\right] > \sigma^2$, hence **F should be greater than one**.

Relationship between y and $\mathbf{x}$

In the example of the salary regressed on years since Ph.D and years of services, the value of the F statistic is 45.71 (2 and 394 degrees of freedom). The p-value associated to the test is lower than 2.2×10^{-16}.

Hence, we reject the null hypothesis in favor of the alternative at the 1% level: at least β_1 or β_2 is different from zero.

Variable selection

If at least one of the variables is related to the repsonse variable is, we would like to know which one is.

A first idea would be to **test for each variable if its associated coefficient is statistically different from zero**:

$$\begin{cases} \mathsf{H}_0 : \beta = 0 \\ \mathsf{H}_1 : \beta_j \neq 0 \end{cases} , j = 1, 2, \ldots, p.$$

The T statistic associated with this test writes:

$$T = \frac{\hat{\beta}_j - \beta_{j,H_0}}{\hat{\sigma}_{\hat{\beta}_j}} \sim \mathcal{S}t(n - p - 1,)$$

where β_{j,H_0} is the value of β_j under the null hypothesis.

Tests on the coefficients

To perform this bilateral test at an α level, we can get the quantiles $-t_{\alpha/2}$ and $t_{\alpha/2}$ from a Student distribution such as:

$$\mathbb{P}\left(-t_{\alpha/2} < \frac{\hat{\beta}_j - \beta_{j,H_0}}{\hat{\sigma}_{\hat{\beta}_j}} < t_{\alpha/2}\right) = 1 - \alpha.$$

From the estimates, we can compute the observed value of the T statistic as:

$$t_{j,\text{obs.}} = \frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}}.$$

Tests on the coefficients

The **decision** rule is:

- if $t_{j,\text{obs.}} \in [-t_{\alpha/2}, t_{\alpha/2}]$:
  - non-rejection region: we do not reject H_0 at the α level
  - β_j is not statistically different from zero
- if $t_{j,\text{obs.}} \notin [-t_{\alpha/2}, t_{\alpha/2}]$:
  - rejection region: we reject H_0 at the α level
  - β_j is statistically different from zero

Tests on the coefficients

When the **number p of predictors is larger than the number of observed values n**, it is not even possible to fit the multiple linear regression model using least squares:

- testing the coefficients one by one is therefore not possible
- performing the F-test is not possible either.

In that case, choosing which variables to keep in the model requires a different approach, such as:

- forward/backward/bi-directional selection
- reducing the dimension (see Sébastien Laurent's course).

Selecting variables

Most of the time, not all predictors are associated with the response.

It is then possible to select a model with a subset of predictors. But then, which one should we choose?

The basic idea is to use a metric to compare models with each other, *e.g.*:

- the Akaike Information Criterion (AIC)
- the Bayesian Information Criterion (BIC)
- the adjusted R^2
- Mallow's C_p
- ...

But, with p variables, there is a total of 2^p different models that can be estimated using a subset of p :

- **fitting all the possible subset is not to be considered**.

Selecting variables

Some **recursive algorithms** can be used to perform variable selection, without screening all the possible models:

- **forward selection**:
  - starting with a model with an intercept but not predictor
  - choosing the first variable to be included by fitting p regressions and selecting the one with the lowest RSS
  - finding another variable to be added by fitting $p-1$ regressions and selecting the one with the lowest RSS
  - and so on, util a stopping rule is satisfied
- **backward selection**:
  - starting with a model with an intercept and all predictors
  - removing the variable with the largest p-value
  - estimating the new model without that variable and remove the variable with the largest value
  - and so on, until a stoppig rule is satisfied
- **bidirectional elimination**:

Measuring the quality of fit

The quality of fit can be assessed using some metrics (RSE, R^2, ...), as in the simple linear case.

- While in the simple linear regression case, the R^2 is equal to the squared value of the correlation of the response and the variable..
- In the multiple linear regression, it can be shown that it is equal to the **square of the correlation between the response and the fitted linear model**

In the multiple linear regression case, the value of the R^2 increases with the number of predictor introduced in the model:

- adding another variable allows fitting better the training data

Prediction error

The prediction given by the linear model carries multiple sources of errors.

One is related to the reducible error.

Recall that the least squares plane is given by

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}},$$

which is an estimation for the **true population regression plane**:

$$f(\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}$$

Providing a confidence interval to the prediction allows us to determine how close the prediction is to the true value.

Prediction error

The model itself carries a reducible error: we assume a linear model, which is usually an approximation for the true form of the relationship between the response and the predictors.

Finally, the prediction contains an irreductible error coming from the error term ε of the model. By using prediction intervals, we can account for this error term.

2.3 Qualitative Predictors

Qualitative predictors

We have so far used two predictors in our example: years since Ph.D and years of service. These two variabls were considered as real-valued.

Now, we will consider another type of predictor: the qualitative predictors.

In the example of the salaries, some information regarding the gender of the professor is provided (Female/Male), the rank (Professor, Associate Professor, Assistant Professor), and the discipline (Theoretical/Applied departmens).

2.3.1 Two levels

Two levels

Let us first focus on qualitative predictors with only two levels. This is the case of the gender variable in the data.

It is a factor variable, created as a **dummy variable**:

$$x_i = \begin{cases} 1 & \text{if the } i\text{th person is female} \\ 0 & \text{if the } i\text{th person is male} \end{cases}, \quad i = 1, \dots, n$$

The model thus writes:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{if the } i\text{th person is female} \\ \beta_0 + \varepsilon_i & \text{if the } i\text{th person is male} \end{cases}$$

Two levels

The interpretation of the constant β_0 therefore changes. It should now be viewed as the average salary for male professors. The average salary among female professors is equal to $\beta_0 + \beta_1$.

	Least squares
(Intercept)	11.51^{***}
	(0.16)
genderFemale	-1.41^{**}
	(0.51)
R^2	0.02
Adj. R^2	0.02
Num. obs.	397

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 3: Statistical models

Two levels

The average salary for male professor is therefore $\$11.509$ thousand dollars for 9 months, while it is only $\$11.509 - 1.409 = 10.1$ for women. This difference is significative at the 5% level.

Coding "Female"" as 0 and "Male" as 1 does not change the regression fit, but it changes the interpretation:

	Least squares
(Intercept)	10.10^{***}
	(0.48)
genderMale	1.41^{**}
	(0.51)
R^2	0.02
Adj. R^2	0.02
Num. obs.	397

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 4: Statistical models

2.3.2 More than two levels

More than two levels

Now, let us consider an example with a qualitative predictor with more than two levels: the rank (Professor, Assistant Professor, Associate Professor).

In this situation, we can create an additional **dummy variable**:

The first one would be, let us say:

$$x_{1i} = \begin{cases} 1 & \text{if the } i\text{th person is professor} \\ 0 & \text{if the } i\text{th person is not professor} \end{cases}, \quad i = 1, \ldots, n$$

And the second:

$$x_{2i} = \begin{cases} 1 & \text{if the } i\text{th person is associate professor} \\ 0 & \text{if the } i\text{th person is not associate professor} \end{cases}, \quad i = 1, \ldots, n$$

More than two levels

The model then writes:

$$
\begin{aligned}
y_i &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i \\
&= \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{if the } i\text{th person is professor} \\ \beta_0 + \beta_2 + \varepsilon_i & \text{if the } i\text{th person is associate professor} \\ \beta_0 + \varepsilon_i & \text{if the } i\text{th person is assistant professor} \end{cases}
\end{aligned}
$$

- β_0 is the average 9 months salary for assistant professor
- β_1 is the difference in the average 0 months salary between the assistant professor and professor categories
- $\beta_0 + \beta_1$ is the average 9 months salary for professor
- β_2 is the difference in the average 0 months salary between the assistant professor and associate professor categories
- $\beta_0 + \beta_2$ is the average 9 months salary for associate professor

More than two levels

	Least squares
(Intercept)	8.08^{***}
	(0.29)
rankProf	4.60^{***}
	(0.32)
rankAssocProf	1.31^{**}
	(0.41)
R^2	0.39
Adj. R^2	0.39
Num. obs.	397

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 5: Statistical models

2.4 Interaction terms

Interaction terms

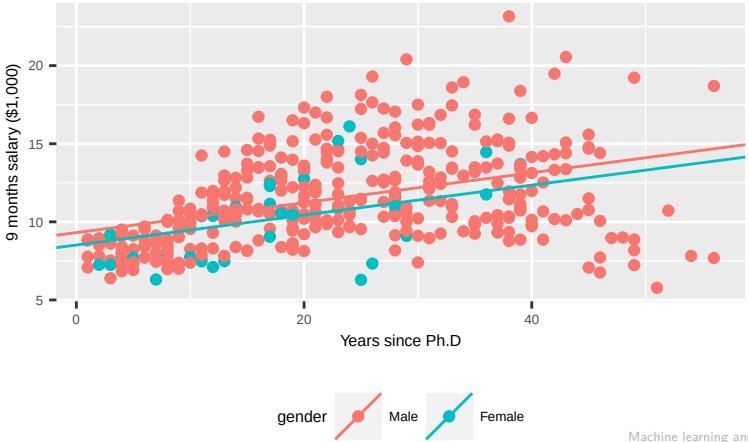
Let us consider adding some **interaction terms** to the model. In previous estimations, we have assumed that the effect on the response of changing the value of one predictor is independant of the values of the other predictors. This assumption is known as the **additive assumption**.

Let us suppose that the salary of professors depends on the number of years since Ph.D and on the gender of the individual. The model writes:

$$\begin{aligned}\text{salary}_i &= \beta_0 + \beta_1 \text{Years since Ph.D}_i + \beta_2 \text{Gender}_i + \varepsilon_i \\ &= \begin{cases} (\beta_0 + \beta_2) + \beta_1 \text{Years since Ph.D}_i + \varepsilon_i & \text{if the } i\text{th person is female} \\ \beta_0 + \beta_1 \text{Years since Ph.D}_i + \varepsilon_i & \text{if the } i\text{th person is male} \end{cases}\end{aligned}$$

Interaction terms

This corresponds to fitting two slopes: one for the females and another for males.



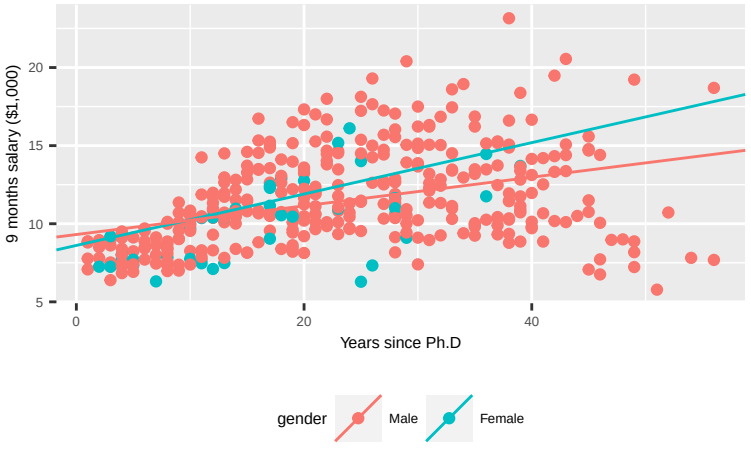
Interaction terms

Now, let us consider that the the effect of a unit increase in the number of years since Ph.D may be different depending on the gender of the professor.

The model now writes:

$$\begin{aligned}\text{salary}_i &= \beta_0 + \beta_1 \text{Years since Ph.D}_i + \beta_2 \text{Gender}_i + \beta_3 \text{Years since Ph.D}_i \times \text{Gender}_i + \varepsilon_i \\ &= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3)\text{Years since Ph.D}_i + \varepsilon_i & \text{(female)} \\ \beta_0 + \beta_1 \text{Years since Ph.D}_i + \varepsilon_i & \text{(male)} \end{cases}\end{aligned}$$

Interaction terms



Interaction terms

As the slope of the line for women is larger than that of men, this suggests that the effect on salary of an additional year since Ph.D is larger for women than it is for men.

This result may sound odd, as we would have (unfortunately) expected the contrary.

Why do we observe such a result?

Interaction terms

Two reasons may explain that:

1. we can look at the coefficient of the interaction term between the number of years since Ph.D and gender (next slide): it is not significant ;
2. contrary to men, there is no observations for women who got their Ph.D more than 39 years ago. We saw that the relationship between salary and the number of years since Ph.D does not seem linear and shows a hill-shaped effect. Hence, the non-linearity not well accounted for in the estimation lowers the slope for men due to values corresponding to large number of years since Ph.D. This is not observed for women, since such values are not in the data.

Interaction terms

	Without interaction	With interaction
(Intercept)	9.31^{***}	9.41^{***}
	(0.29)	(0.29)
yrs.since.phd	0.10^{***}	0.09^{***}
	(0.01)	(0.01)
genderFemale	-0.79	-2.02^{*}
	(0.47)	(0.92)
yrs.since.phd:genderFemale		0.07
		(0.05)
R^2	0.18	0.19
Adj. R^2	0.18	0.18
Num. obs.	397	397

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 6: Statistical models

Accounting for non-linear effects

The relationship between salary and years since Ph.D. does not seem to be linear. It may be a good idead to try to look at a quadratic effect instead, by introducing the squared value of number of years since Ph.D:

$$\text{salary}_i = \beta_0 + \beta_1 \text{Years since Ph.D}_i + \beta_2 \text{Years since Ph.D}_i^2 + \varepsilon_i$$

	Polynomial regression
(Intercept)	6.51^{***}
	(0.39)
yrs.since.phd	0.41^{***}
	(0.04)
yrs.since.phd_squared	-0.01^{***}
	(0.00)
R^2	0.31
Adj. R^2	0.31
Num. obs.	397

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 7: Statistical models

2.5 Working with wrong models

Working with wrong models

Among the problems that may occur when we fit a linear regression model (James et al., 2013):

- non-linearity of the relationship between y and $\mathbf{x}$
- correlation of error terms
- non-constant variance of error terms
- outliers
- high-leverage points
- collinearity

When facing models that are wrong, Berk (2008) recalls that two approaches can be used:

1. Patching up models that are mispecified
2. Working with misspecified models

Let us begin by talking about the last point, using some illustration.

Illustration: linear regression

- red lines: true conditional means (nature's response surface)
- vertical black dotted lines: distribution of y values around each conditional mean (also from nature), assuming the same variance for each conditional distribution
- The relationship between y and $\mathbf{x}$ seems approximatively quadratic
- Red circle: an observed value, realization of y

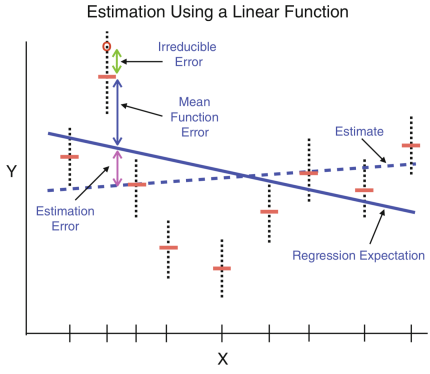


Figure 10: Estimation of a nonlinear response surface under the true linear model perspective. (Source: Berk 2008).

Illustration: linear regression

- Let us assume a linear model: $y_i = \beta_0 + \beta_1 \mathbf{x}_i + \varepsilon_i$ for which we obtain the estimates $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\sigma}^2$
- Dashed blue line: estimated mean function
- Solid blue line: expectation of the mean function

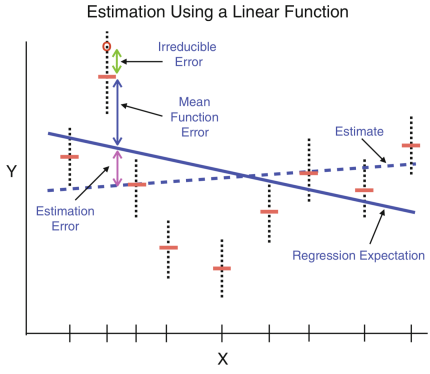


Figure 11: Estimation of a nonlinear response surface under the true linear model perspective. (Source: Berk 2008).

Illustration: linear regression

- Blue arrow: bias at a value $\mathbf{x}_i$ (Bias $\left(\hat{f}(\mathbf{x}_0)\right)$)
- Magenta arrow: random variation ($\mathbb{V}ar\left(\hat{f}(\mathbf{x}_0)\right)$)
- Green arrow: irreducible error ($\mathbb{V}ar(\varepsilon)$)

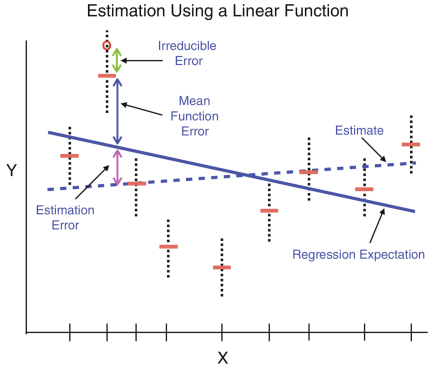


Figure 12: Estimation of a nonlinear response surface under the true linear model perspective. (Source: Berk 2008).

Illustration: non-linear function

- The three sources of error remains when using a nonlinear function
- Still not possible to know the bias...

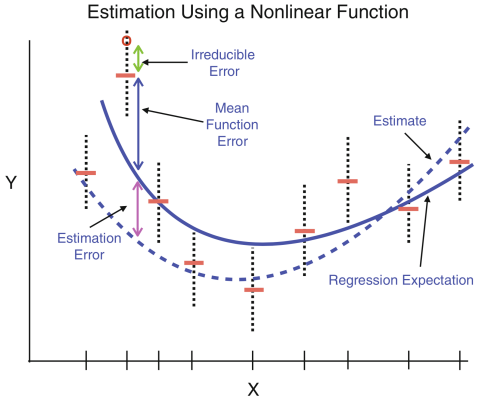


Figure 13: Estimation a nonlinear response surface under the true nonlinear model perspective. (Source: Berk, 2008).

2.5.1 Correlation of the error terms

Correlation of the error terms

Recall the linear model:

$$y_i = \beta_0 + \beta_1 \mathbf{x}_{1i} + \ldots + \beta_p \mathbf{x}_{pi} + \varepsilon_i, \quad i = 1, \ldots, n, \tag{22}$$

where ε_i is an error term normally distributed with 0 mean and variance σ^2, *i.e*, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, and where $\mathbf{x}_{ji}$ represents the ith observation for the jth predictor, $j = 1, \ldots, p$.

We **assume** that the error terms are uncorrelated, *i.e.*, $\varepsilon \sim \mathcal{N}\left(0, \sigma^2 I_n\right)$.

Correlation of the error terms

If there is **correlation** between the ε_i :

- the estimation of the **standard errors** underestimate the true standard errors
- **confidence/prediction intervals** are therefore narrower than they should be
- **p-values** associated with the model will be lower than they should be

2.5.2 Non-constant variance of error terms

Non-constant variance of error terms

In the linear model, we also assume that the variance of the error term is constant: $\mathbb{V}ar(\varepsilon_i) = \sigma^2$ for all $i = 1, \ldots, n$.

Once again, if it is not the case (if there is **heteroscedasticity**), this has consequences on the estimation of the standard errors, on the confidence and prediction intervals and also on the p-values associated with the model.

In presence of **heteroscedasticity**, a way to tackle the issue is to transform the data using a concave function (such as the log function).

Another way of getting around the problem is to estimate the model by **weighted least squares**, where the weights are proportional to the inverse variance.

2.5.3 Outliers

Outliers

The prediction of some points may be relatively far from the observed value. These points are called **outliers**.

They can be the result of an incorrect recording, or the observation can come from a sub-population.

To detect such points, Cornillon and Matzner-Løber (2007) suggests using **standardized residuals**.

Outliers

Normalized residuals are given by:

$$r_i = \frac{e_i}{\sigma\sqrt{1-h_{ii}}}, \tag{23}$$

where h_{ij} is the (i,j)th element of the matrix $\boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top$

Replacing σ by its estimate $\hat{\sigma}$ gives the **standardized residuals**:

$$t_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}, \tag{24}$$

Outliers

The standardized residuals are not independant by construction (the residual variance $\hat{\sigma}^2$ was estimated with all data):

- they cannot be representative of an absence or a presence of autocorrelation
- but they have the same variance unit and can therefore be used to detect residuals with high variance.

However, Cornillon and Matzner-Løber (2007) suggest that we should use **studentized residual** (obtained by cross validation) instead of the standardized residuals:

$$t_i^\star = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}}, \tag{25}$$

where $\hat{\sigma}_{(i)}$ is the estimation of σ by least squares based on all observed valued except for the ith.

Outliers

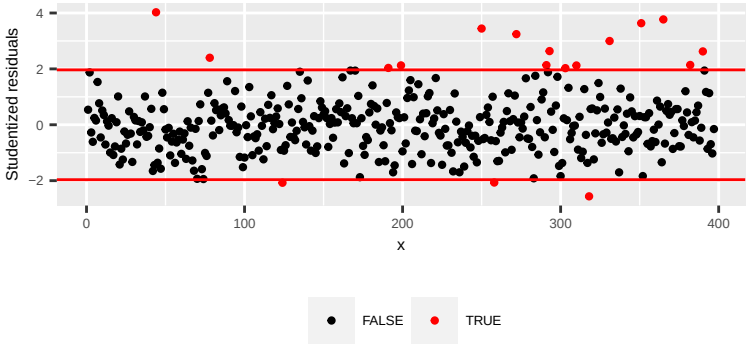
It can be shown, assuming the residuals are normally distributed, that $t_i^\star \sim \mathcal{S}t(n-p-1)$.

Using these studentized residual, we can define an **outlier** as a point $(\mathbf{x}_i, y_i)$ for which the value associated with $t_i^\star$ is high, compared to the threshold given by a Student distribution, *i.e.*:

$$| t_i^\star | > t_{n-p-1}\left(1-\alpha/2\right)$$

Outliers

As an illustration, let us consider the cas in which we regress the salary of professors on the number of years since their Ph.D, the same value squared, the gender and the discipline (*i.e.*, a total of 5 regressors).



2.5.4 High-leverage points

High-leverage points

While **outliers** are observations for which the response y_i is unusual given the predictors, **high leverage points** are observations which have unusual value for x_i.

Let us recall that:

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

For the ith observation:

$$\hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j = h_{ii} y_i + \sum_{j \neq i} h_{ij} y_j,$$

where h_{ij} is the (i,j)th element of the matrix $\boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top$.

This allows us to know the weight of the observation on its prediction, through h_{ii}.

High-leverage points

This leads to the definition of a **leverage point**, provided by Cornillon and Matzner-Løber (2007):

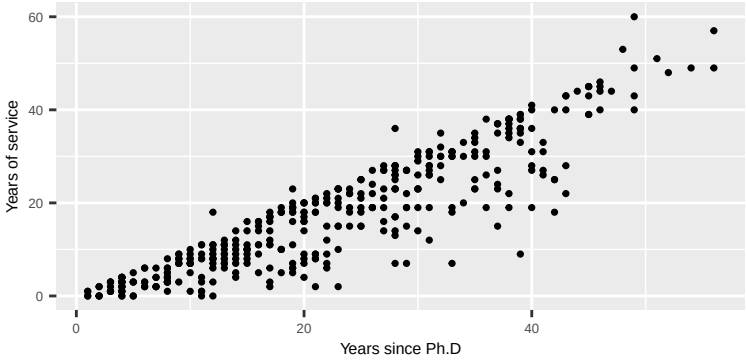
- A point is a leverage point if the values h_{ii} of the projection matrix $\boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top$ are greater than:
  - $h_{ii} > 2p/n$ according to Hoaglin and Welsch (1978)
  - $h_{ii} > 3p/n$ for $p > 6$ and $n - p > 12$ according to Velleman and Welsch (1981)
  - $h_{ii} > 0.5$ according to Huber (1981)

2.5.5 Collinearity

Collinearity

When two or more predictors are closely related, we face a phenomenon known as **collinearity**.

This is the case of the predictors "Years since Ph.D" and "Years of service":



Collinearity

The presence of collinearity may be the source of problems when estimating a linear model:

- it can then become difficult to disentangle the individual effects of collinear variables on the response
- the variance of at leat one of the estimated coefficients $\hat{\beta}_j$ tends to be inflated

As a consequence, since the t-statistic for each predictor uses the estimated variance of the coefficient, it can lead to a p-value lower tan it should be.

Looking at the correlation matrix of the predictors may help identifying possible problems of collinearity.

But collinearity can exist between three or more variables. In that case, known as **multicollinearity**, looking at the correlation matrix does not help.

Multicollinearity

There are multiple ways of detecting the presence of multicollinearity. One of those consists in computing the **variance inflation factor** (VIF):

$$\text{VIF}\left(\hat{\beta}_j\right) = \frac{1}{1 - R^2_{\mathbf{x}_j|\mathbf{x}_{-j}}}, \tag{26}$$

where $R^2_{\mathbf{x}_j|\mathbf{x}_{-j}}$ is the R^2 obtained from a regression of $\mathbf{x}_j$ onto all the other predictors $\mathbf{x}_{-j}$.

- The smallest value for VIF is 1: complete absence of collinearity
- When the value is high (> 5 or > 10): we can suspect the presence of multicollinearity, due to the predictor $\mathbf{x}_j$

When facing multicollinearity, a simple solution consists in dropping one of the problematic variables.

References I

Berk, R. A. (2008). *Statistical learning from a regression perspective*, volume 14. Springer.

Cornillon, P.-A. and Matzner-Løber, É. (2007). *Régression: théorie et applications*. Springer.

Hoaglin, D. C. and Welsch, R. E. (1978). The hat matrix in regression and anova. *The American Statistician*, 32(1):17–22.

Huber, P. J. (1981). *Robust Statistics*. John Wiley & Sons, Inc.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Velleman, P. F. and Welsch, R. E. (1981). Efficient computing of regression diagnostics. *The American Statistician*, 35(4):234–242.