Review of Linear Regression

Jerome Reboulleau[1]

[1] HEC, Lausanne

Quantitative Methods for Management

Outline

Simple linear regression

Outcome in simple linear analysis

- Calculate and interpret the correlation between 2 variables.
- Determine whether the correlation is significant.
- Determine whether a regression model is significant.
- Prediction.
- Confidence Intervals for the regression analysis.

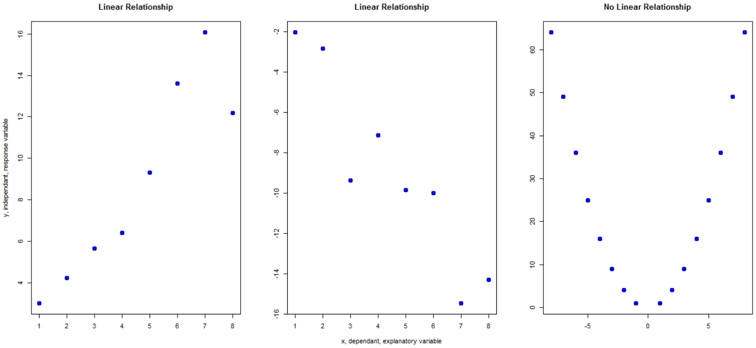
Preview: Multiple linear regression

Outcome in multiple linear analysis

- Understand the general concepts behind model building.
- Analyze the model output.
- Test hypotheses about the significance of a multiple regression model and independant variables.
- Understand the uses of stepwise regression.

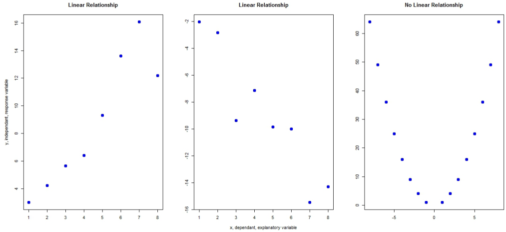
Correlation concept

- Calculate and interpret the correlation between 2 variables.



Correlation formula

- Calculate and interpret the correlation between 2 variables.



- Measure of the strength of the linear relationship is the sample correlation coefficient:

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{(\sum (x-\bar{x})^2)(\sum (y-\bar{y})^2)}} \tag{1}$$

Significance test for the correlation

- After computing the correlation, one would like to establish conclusions. Is this correlation significantly different from 0 ?
- To draw conclusions, we develop a statistical test in 3 steps: assumptions, statistic, p-value and significance level:
  1. Assumptions in the case of a two-sided test and where ρ represents the population correlation
  $$\begin{cases} H_0 : & \rho = 0 \\ H_1 : & \rho \neq 0 \end{cases}$$
  2. Test statistic for the correlation
  $$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \tag{2}$$
  3. Decision rule If $t > t_{n-2,\alpha/2}$ or If $t < t_{n-2,1-\alpha/2}$, we reject H_0. We can also conclude through the p-value: we reject if p-value < α

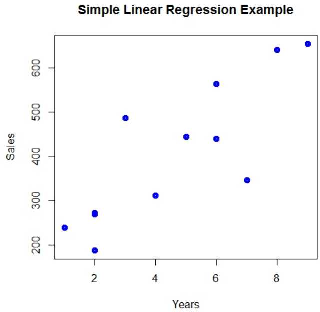
Example: analyzing correlation

- Suppose we analyze the sales of employees and the number of years of employment within this company

Sales Y	Years of employement X
487	3
445	5
272	2
641	8
187	2
440	6
346	7
238	1
312	4
269	2
655	9
563	6

Example: analyzing correlation

- Step 1: develop a scatterplot.



- The question of the linear relationship between Sales and Year seems to be a good question.

Example: analyzing correlation

- Step 2:: compute correlation

$$
\begin{aligned}
r &= \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{(\sum (x-\bar{x})^2)(\sum (y-\bar{y})^2)}} \\
&= \frac{3838.92}{\sqrt{76.92 \cdot 276434.92}} \\
&= 83\%
\end{aligned}
$$

- Due to the small sample size, one would like to confirm that this correlation is really different from 0.

Example: analyzing correlation

- Step 3: Correlation test
- Assumptions:

$$\begin{cases} H_0: & \rho = 0 \\ H_1: & \rho \neq 0 \end{cases}$$

- Test statistic

$$\begin{aligned} t &= \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \\ &= \frac{83\%}{\sqrt{\frac{1-83\%^2}{12-2}}} \\ &= 4.752 \end{aligned}$$

Example: analyzing correlation

- Decision rule at significance level $\alpha = 5\%$:
- The rejection region is defined by $\pm t_{n-2,\alpha/2} = \pm t_{12-2,2.5\%} = \pm 2.228$
- Our statistic t belongs to the rejection region and we reject H_0. The correlation is significantly different from 0.
- Another way to see thing is through the p-value. Here p-value=0.00077 (see R result).
- When p-value < α, we reject H_0, leading to the same conclusion as previously.

Simple Linear Regression Model

- The method of simple regression analysis when a single independent variable x is used to predict the dependent variable x
- We represent the relationship between x and y through a straight line described as

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

where

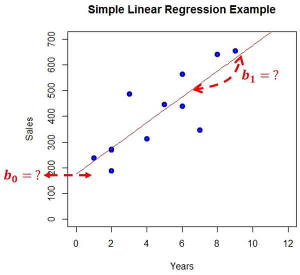
$$\begin{aligned} y &= \text{Value of the dependent variable} \\ x &= \text{Value of the independent variable} \\ \beta_0 &= \text{Intercept} \\ \beta_1 &= \text{Slope} \\ \epsilon &= \text{Random error term} \end{aligned}$$

Simple Linear Regression Assumptions

1. The error terms ϵ are statistically independent of one another.
2. The distribution of ϵ is normal.
3. For all values of x, the ϵ have equal variance

Model fitting: method idea

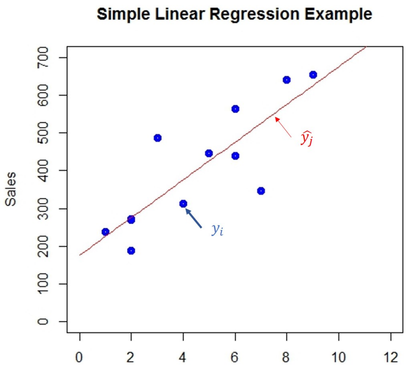
- If the linear relationship seems to be satisfying, the next question is to determine the "best" model coefficient



where b_0 and b_1 are estimates of the population model coefficients β_0 and β_1.

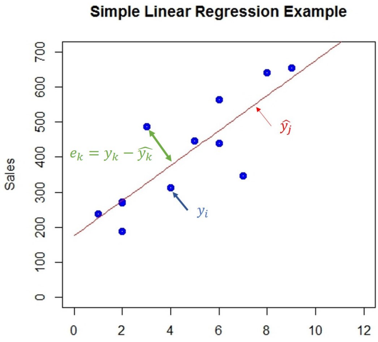
Model fitting: method idea

- The main idea is to make as close as possible the data y_i to the model value $\hat{y}_i = b_0 + b_1 \cdot x_i$,



Model fitting: OLS

- The method is called the **Ordinary Least Squares Criterion (OLS)**.
- It aims to determine coefficients that minimizes the overall "prediction error", i.e. $\sum_k e_k$, also called the **residuals**.



Model fitting: Sum of Squared Errors, SSE

- However, $\sum_k e_k = \sum_k (y_k - \hat{y}_k) = 0$, so that it does not help to find b_0 and b_1.
- Instead we have to minimize the **Sum of Squared Residuals (Errors)**:

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

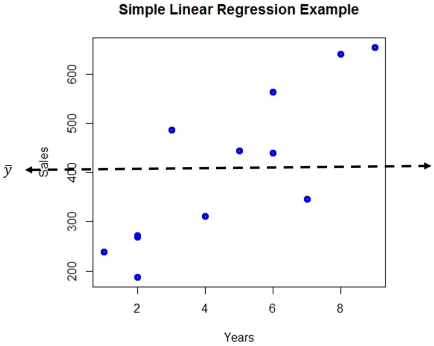
Model fitting: OLS Model coefficient

- Such minimization gives us the regression's coefficient:

$$\begin{aligned} b_1 &= \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y}_i)}{\sum_i (x_i - \bar{x})^2} \\ b_0 &= \bar{y} - b_1 \cdot \bar{x} \end{aligned}$$

Model fitting: Total Sum of Squares, SST

- The target of our model is to efficiently represents the data variations, i.e. SST



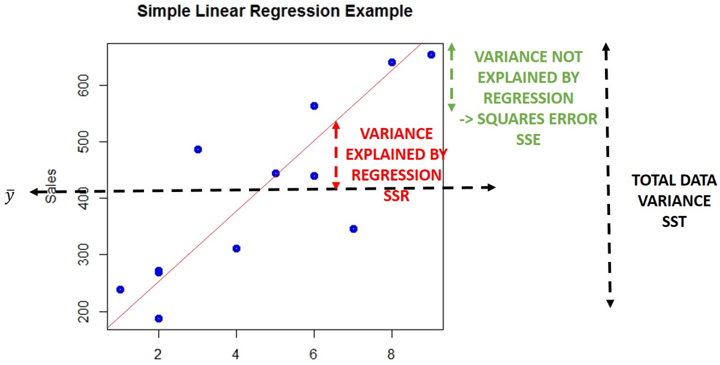
SST is the total sum of squares.
For a sample of size n

TOTAL DATA VARIANCE SST

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Model fitting: Variance Decomposition

- According to our model, this variance can be divided into 2 parts: one coming from the regression, the second one not.



Significance Test in Regression Analysis

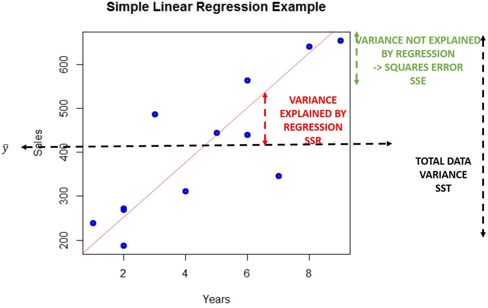
- SST = SSR + SSE

SSE is the sum of squares error.

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

SSR is the sum of squares regression.

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$



Significance Test in Regression Analysis

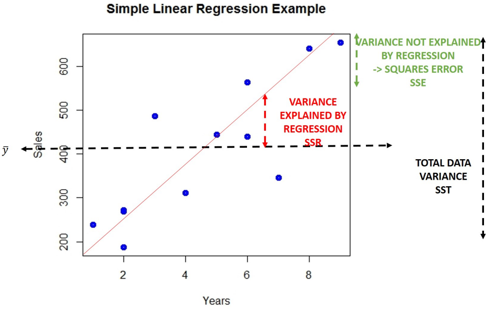
- Of course, the idea is having SSR as large as possible and SSE as small as possible:

$$SST = SSR + SSE$$

This gives us a first measure of the quality of the regression: the coefficient of determination R^2

$$R^2 = \frac{SSR}{SST}$$

R^2 will have values between 0 and 1 and we wish R^2 to be as large as possible. In simple linear regression $R^2 = r^2$ where r is the correlation coefficient.



Test statistic for Significance of the R^2

- After computing the R^2, one would like to establish a conclusion. Is this coefficient significantly different from 0 ?
- To draw conclusions, we develop a statistical test
  1. Assumptions in the case of a two-sided test and where ρ represents the population correlation
  $$\begin{cases} H_0 : & \rho^2 = 0 \\ H_1 : & \rho^2 \neq 0 \end{cases}$$
  2. Test statistic for the R^2
  $$F = \frac{\frac{SSR}{1}}{\frac{SSE}{n-2}} \tag{3}$$
  3. Decision rule: We compare this statistic to the F-distribution. The F-distribution is defined by 2 parameters (degrees of freedom):
     - The first one is the number of explanatory variable, i.e. here 1.
     - The second one is the number of data n minus 2, $(n-2)$.
     - If $F > F_{1,n-2,\alpha}$, we reject H_0.
     - Of course, we can also conclude through the p-value.

Test statistic for Significance of the slope coefficient β_1

- Another intuitive way to analyze the regression is a test related to the slope coefficient
- We will develop the following test
  1. Assumptions of a two-sided test related to the slope coefficient

$$\begin{cases} H_0 : & \beta_1 = 0 \\ H_1 : & \beta_1 \neq 0 \end{cases}$$

  2. Of course if H_0 is rejected, it will mean that the slope coefficient is different from 0, so that the variable x will be of interest to explain varations of y.
  3. Before presenting the test, we need to present some definitions

Test statistic for Significance of the slope coefficient β_1

- Population standard error of the estimate σ_ϵ
- Estimator s_ϵ of the standard error σ_ϵ (estimate of the deviation of y around the regression line)

$$s_\epsilon = \sqrt{\frac{SSE}{n-2}}$$

- Standard error of the slope coefficient

$$\sigma_{b_1} = \frac{\sigma_\epsilon}{\sqrt{\sum (x_i - \bar{x})^2}}$$

- Estimator of the standard error of the slope coefficient

$$s_{b_1} = \frac{s_\epsilon}{\sqrt{\sum (x_i - \bar{x})^2}}$$

Test statistic for Significance of the slope coefficient β_1

- We now have everything is develop our test
  1. Assumptions of a two-sided test related to the slope coefficient
$$\begin{cases} H_0: & \beta_1 = 0 \\ H_1: & \beta_1 \neq 0 \end{cases}$$
  2. Statistic for test of the significance of the slope
$$t = \frac{b_1 - \beta_1}{s_{b_1}}$$
  3. Decision rule:
     - We compare our statistic with $\pm t_{n-2,\alpha/2}$.
     - We reject if $t > t_{n-2,\alpha/2}$ or $t < t_{n-2,1-\alpha/2}$.
     - We can also draw conclusions through the p-value.

Conclusions

In the simple linear case, we have 3 methods to test for the significance of the regression:

- Correlation test: t-test.
- R^2 test: F-test.
- Slope coefficient test: t-test.

Example... continued

We first apply the F-test. We need to compute SSE, SSR and first we need $\hat{y}$, so b_0 and b_1

- $\bar{y} = 404.85$, $\bar{x} = 4.58$
- $\sum (x_i - \bar{x})(y_i - \bar{y}) = 3838.92$
- $\sum (x_i - \bar{x})^2 = 76.92$
- $$\begin{aligned} b_1 &= \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y}_i)}{\sum_i (x_i - \bar{x})^2} \\ &= \frac{3838.92}{76.92} \\ &= 49.91 \end{aligned}$$
- $$\begin{aligned} b_0 &= \bar{y} - b_1 \cdot \bar{x} \\ &= 404.85 - 49.91 \cdot 4.58 \\ &= 175.83 \end{aligned}$$

Example... continued

Let's start the F-test

1. $$SSE = \sum_{i=1}^{12} (y_i - \hat{y}_i)^2 = 84834.29$$
2. $$SSR = \sum_{i=1}^{12} (\hat{y}_i - \bar{y})^2 = 191600$$
3. $$F = \frac{\frac{SSR}{1}}{\frac{SSE}{n-2}} = \frac{\frac{191600}{1}}{\frac{84834}{12-2}} = 22.58$$
4. $F_{1;n-2;\alpha} = F_{1;10;5\%} = 4.965$
5. The F statistic belongs to the rejection region. We reject H_0.
6. The p-value is $0.0008 < \alpha$. We reject H_0.

Example...continued

Let's start the Slope test

1. $$s_\epsilon = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{84834.92}{12-2}} = 92.10$$
2. $$s_{b_1} = \frac{s_\epsilon}{\sqrt{(x_i - \bar{x})^2}} = \frac{92.10}{\sqrt{76.92}} = 10.50$$
3. $$t = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{49.91 - 0}{10.50} = 4.75$$
4. $\pm t_{n-2;\alpha/2} = \pm t_{10;2.5\%} = \pm 2.228$
5. The t statistic belongs to the rejection region. We reject H_0.
6. The p-value is $0.0008 < \alpha$. We reject H_0.

Confidence Interval

We develop confidence interval in the following cases:

- Slope coefficient
- Prediction interval

Confidence Interval: Slope Coefficient

The confidence interval for the slope coefficient is very simple:

$$\text{Parameter estimate} \pm \text{quantile} \cdot \text{standard error of the estimate}$$

$$\Leftrightarrow \quad b_1 \pm t_{n-2,\alpha/2} \cdot s_{b_1}$$

In our example, we get:

$$49.91 \pm t_{12-2,\alpha/2} \cdot 10.50 = 49.91 \pm 2.228 \cdot 10.50 = [25.97; 73.85]$$

Prediction Interval: Prediction for average value

We develop the confidence interval for the average value, i.e. $E[y]$ given the value dependent value x_p:

$$\hat{y} \pm t_{n-2,\alpha/2} \cdot s_\epsilon \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x - \bar{x})^2}}$$

For $x_p = 3$,

$$\hat{y} = 175.83 + 49.91 \cdot 3 = 325.56$$

and we get:

$$325.56 \pm 2.228 \cdot 92.10 \cdot \sqrt{\frac{1}{12} + \frac{(3 - 4.583)^2}{76.92}} = [255.7; 395.4]$$

Prediction Interval: Prediction for a particular y

We develop the confidence interval for a particular value y, given the value dependent value x_p:

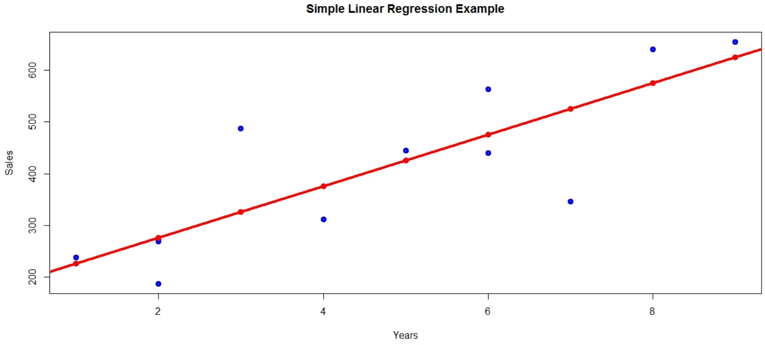
$$\hat{y} \pm t_{n-2,\alpha/2} \cdot s_\epsilon \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x - \bar{x})^2}}$$

For the same $x_p = 3$,

$$325.56 \pm 2.28 \cdot 92.10 \cdot \sqrt{1 + \frac{1}{12} + \frac{(3 - 4.583)^2}{76.92}} = [108.77; 542.35]$$

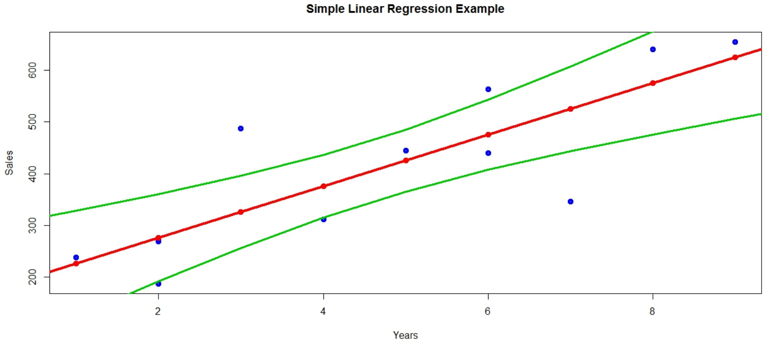
Prediction without interval

Without prediction interval, one would get the following prediction line:



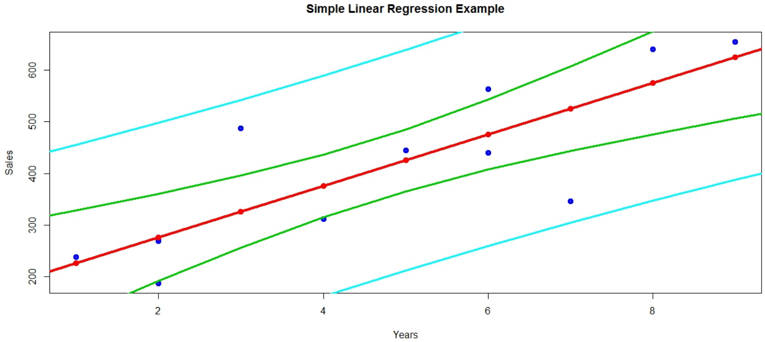
Prediction for the average value $E[y]$

Including the prediction interval for the average, we get



Prediction for a particular y

Including finally the prediction for a particular y



Model diagnostics

The model diagnostics will be discussed in the Multiple Linear case (next Chapter) as it is exactly the same analysis

- Normality assumptions
- Equal Variance assumptions

Executive Summary

We have reviewed the Simple Linear Regression covering

- Correlation testing
- Model outlook and assumptions
- Model fitting
- Variance structure and Model testing
- Confidence Interval for slope and prediction