Statistics II
Lesson 4. Simple linear regression

Academic Year 2010/11

Lesson 4. Simple linear regression

Contents

Lesson 4. Simple linear regression

Learning objectives

- Know how to construct a simple linear regression model that describes how a variable X influences another variable Y
- Know now to obtain point estimations of the parameters of this model
- Know to construct confidence intervals and perform tests about the parameters of the model
- Know to estimate the mean value of Y for a specified value of X
- Know to predict future values for the dependent (response) variable Y

Lesson 4. Simple linear regression

Recommended bibliography

- Meyer, P. "Probabilidad y aplicaciones estadísticas" (1992)
  - Chapter
- Newbold, P. "Estadística para los negocios y la economía" (1997)
  - Chapter 10
- Peña, D. "Regresión y análisis de experimentos"(2005)
  - Chapter 5

Introduction

A regression model is a model that describes how a variable X influences the value of another variable Y.

- X: Independent or explanatory or exogenous variable
- Y: Dependent or response o endogenous variable

The aim is to obtain reasonable estimations of Y for different values of X from a sample of n pairs of values $(x_1, y_1), \ldots, (x_n, y_n)$.

Introduction

Examples

- Study how the parents' height may influence their children's height
- Estimate the price of a house depending on its surface
- Predict the unemployment level for different ages
- Approximate the grades attained in a subject as a function of the number of study hours per week
- Forecast the execution time of a program depending on the speed of the processor

Introduction

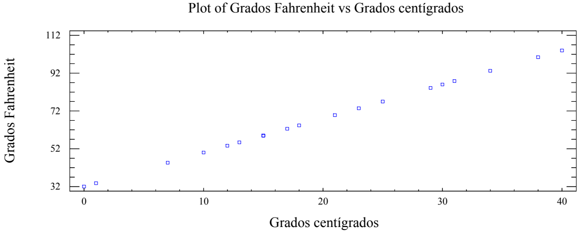
Types of relation

- Deterministic: Given the value of X, the value of Y is perfectly established. They are of the form:

$$Y = f(X)$$

Example: The relationship between the temperature measured in degrees Celsius (X) and the equivalent measure in degrees Fahrenheit (Y) is given by:

$$Y = 1.8X + 32$$



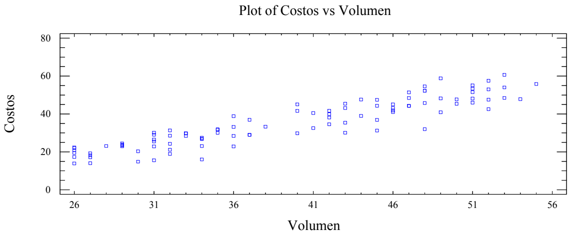
Introduction

Types of relation

- Non deterministic: Given the value of X, the value of Y is not completely determined. They are of the form:

$$y = f(x) + u$$

where u is an unknown perturbation (random variable).
Example: A sample is taken regarding the volume of production (X) and the total cost (Y) associated with a product in a corporate group.



A relation exists but it is not exact

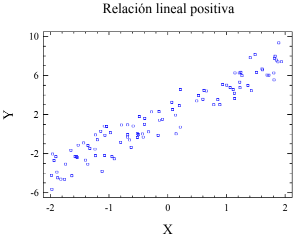
Introduction

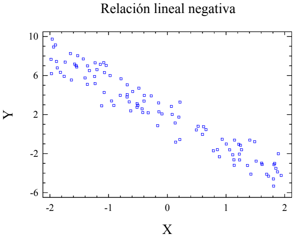
Types of relation

- Linear: When the function $f(x)$ is linear,

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

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



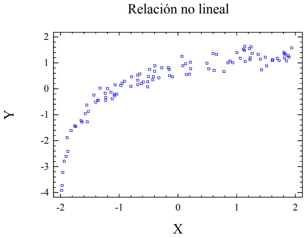


The data show a linear pattern

Introduction

Types of relation

- Nonlinear: When the function $f(x)$ is nonlinear. For exmaple, $f(x) = log(x)$, $f(x) = x^2 + 3, \ldots$

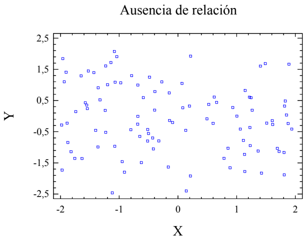


The data do not show linear patterns

Introduction

Types of relation

- Absence of relation: Whenever $f(x) = c$, that is, whenever $f(x)$ does not depend on x



Measures of linear dependency

Covariance

A measure of linear dependency is the covariance:

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

- If there is a positive linear relation, the covariance will be positive and large
- If there is a negative linear relation, the covariance will be negative and large in absolute value.
- If there is no relation between the variables or the relation is significantly linear, the covariance will be close to zero.

but the covariance depends on the units of measurement of the variables

Measures of linear dependency

The correlation coefficient

A measure of linear dependency that doesn't depend on the units of measurement is the correlation coefficient:

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

where

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

- $-1 \leq cor\,(x, y) \leq 1$
- $cor\,(x, y) = cor\,(y, x)$
- $cor\,(ax + b, cy + d) = sign(a)\, sign(c)\, cor\,(x, y)$
  for any values a, b, c, d

The 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 represents the value of the dependent variable for the i-th observation
- x_i represents the value of the independent variable for the i-th observation
- u_i represents the error for the i-th observation, which we will assume to be normal,
$$u_i \sim N(0, \sigma)$$
- β_0 and β_1 are the regression coefficients:
  - β_0 : intercept
  - β_1 : slope

The parameters to estimate are: β_0, β_1 and σ

The simple linear regression model

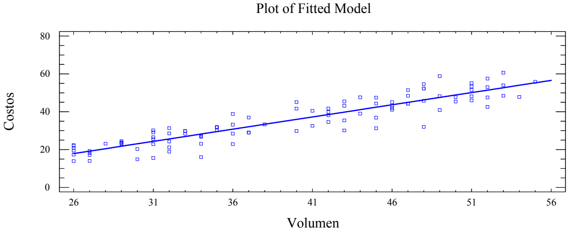
Our goal is to obtain estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for β_0 and β_1 to define the regression line

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

that provides the best fit for the data

Example: Assume that the regression line of the previous example is:

$$\text{Cost } = -15.65 + 1.29 \text{ Volume}$$



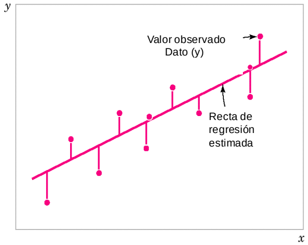
From this model it is estimated that a company that produces 25 thousand units will have a cost given by:

$$\text{Cost } = -15.65 + 1.29 \times 25 = 16.6 \text{ thousand euros}$$

The simple linear regression model

The difference between each value y_i of the response variable and its estimation $\hat{y}_i$ is called a residual:

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



Example (cont.): Undoubtedly, a certain company that has produced exactly 25 thousand units is not going to have a cost exactly equal to 16.6 thousand euros. The difference between the estimated cost and the real one is the error. If for example the real cost of the company is 18 thousand euros, the residual is:

$$e_i = 18 - 16.6 = 1.4 \text{ thousand euros}$$

Hypotheses of the simple linear regression model

- Linearity: The existing relation between X and Y is linear,

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

- Homogeneity: The mean value of the error is zero,

$$E[u_i] = 0$$

- Homoscedasticity: The variance of the errors is constant,

$$Var(u_i) = \sigma^2$$

- Independence: The observations are independent,

$$E[u_i u_j] = 0$$

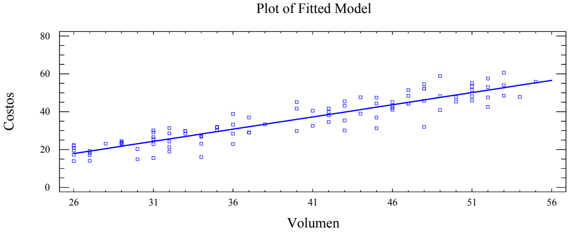
- Normality: The errors follow a normal distribution,

$$u_i \sim N(0, \sigma)$$

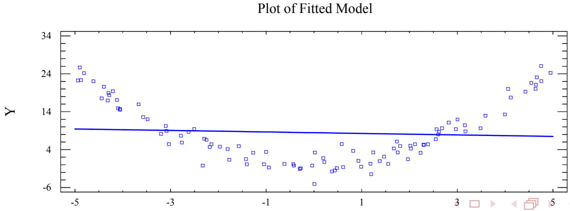
Hypotheses of the simple linear regression model

Linearity

The data have to look reasonably straight



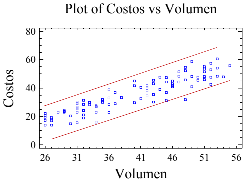
Otherwise, the regression line doesn't represent the structure of the data



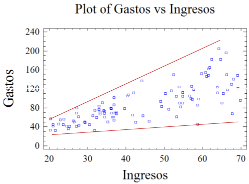
Hypotheses of the simple linear regression model

Homoscedasticity

The dispersion of the data must be constant for the data to be homoscedastic



If this condition does not hold, the data are said to be heteroscedastic



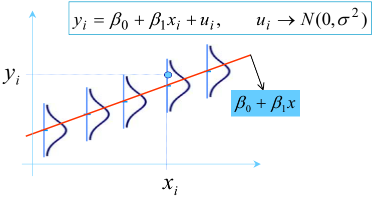
Hypotheses of the simple linear regression model

Independence

- The data must be independent
- An observation must not give information about the rest of the observations
- Usually, it is known from the type of the data if they are adequate or not for this analysis
- In general, time series do not satisfy the independence hypothesis

Normality

- We will assumed that the data are a priori normal



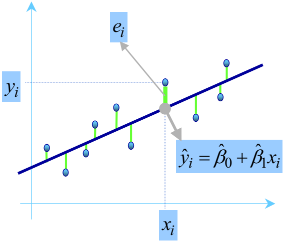
Least squares estimators

Gauss proposed in 1809 the method of least squares for obtaining the values $\hat{\beta}_0$ and $\hat{\beta}_1$ that best fit the data:

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

The method consists of minimizing the sum of the squares of the vertical distances between the data and the estimations, that is, minimize the sum of the squared residuals

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

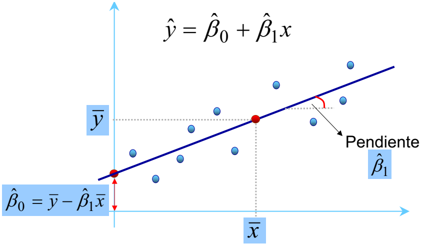


Least squares estimators

The results are given by

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

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



Least squares estimators

Exercise 4.1

The data regarding the production of wheat in tons (X) and the price of the kilo of flour in pesetas (Y) in the decade of the 80's in Spain were:

Wheat production	30	28	32	25	25	25	22	24	35	40
Flour price	25	30	27	40	42	40	50	45	30	25

Fit the regression line using the method of least squares

Results

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{10} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{10} x_i^2 - n\bar{x}^2} = \frac{9734 - 10 \times 28.6 \times 35.4}{8468 - 10 \times 28.6^2} = -1.3537$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 35.4 + 1.3537 \times 28.6 = 74.116$$

The regression line is:

$$\hat{y} = 74.116 - 1.3537x$$

Least squares estimators

Exercise 4.1

The data regarding the production of wheat in tons (X) and the price of the kilo of flour in pesetas (Y) in the decade of the 80's in Spain were:

Wheat production	30	28	32	25	25	25	22	24	35	40
Flour price	25	30	27	40	42	40	50	45	30	25

Fit the regression line using the method of least squares

Results

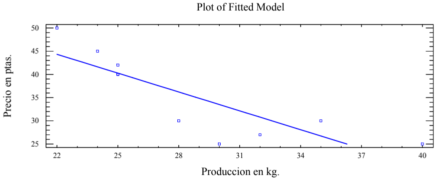
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{10} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{10} x_i^2 - n\bar{x}^2} = \frac{9734 - 10 \times 28.6 \times 35.4}{8468 - 10 \times 28.6^2} = -1.3537$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 35.4 + 1.3537 \times 28.6 = 74.116$$

The regression line is:

$$\hat{y} = 74.116 - 1.3537x$$

Least squares estimators




```
Regression Analysis - Linear model: Y = a + b*X
-----------------------------------------------------------------------------
Dependent variable: Precio en ptas.
Independent variable: Produccion en kg.
-----------------------------------------------------------------------------
                                 Standard          T
Parameter       Estimate          Error        Statistic        P-Value
-----------------------------------------------------------------------------
Intercept        74,1151         8,73577         8,4841          0,0000
Slope           -1,35368          0,3002        -4,50924         0,0020
-----------------------------------------------------------------------------


                           Analysis of Variance
-----------------------------------------------------------------------------
Source                Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                        528,475      1      528,475      20,33       0,0020
Residual                     207,925      8      25,9906
-----------------------------------------------------------------------------
Total (Corr.)                  736,4      9

Correlation Coefficient = -0,84714
R-squared = 71,7647 percent
Standard Error of Est. = 5,0981
```

$\hat{\beta}_0$

$\hat{\beta}_1$

Estimation of the variance

To estimate the variance of the errors, σ^2, we can use,

$$\hat{\sigma}^2 = \frac{\displaystyle\sum_{i=1}^{n} e_i^2}{n}$$

which is the maximum likelihood estimator of σ^2, but it is a biased estimator.

An unbiased estimator of σ^2 is the residual variance,

$$\boxed{s_R^2 = \frac{\displaystyle\sum_{i=1}^{n} e_i^2}{n-2}}$$

Estimation of the variance

Exercise 4.2

Compute the residual variance in exercise 4.1

Results

We first compute the residuals, e_i, from the regression line,

$$\hat{y}_i = 74.116 - 1.3537x_i$$

x_i	30	28	32	25	25	25	22	24	35	40
y_i	25	30	27	40	42	40	50	45	30	25
$\hat{y}_i$	33.5	36.21	30.79	40.27	40.27	40.27	44.33	41.62	26.73	19.96
e_i	-8.50	-6.21	-3.79	-0.27	1.72	-0.27	5.66	3.37	3.26	5.03

The residual variance is:

$$s_R^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{207.92}{8} = 25.99$$

Estimation of the variance

Exercise 4.2

Compute the residual variance in exercise 4.1

Results

We first compute the residuals, e_i, from the regression line,

$$\hat{y}_i = 74.116 - 1.3537 x_i$$

x_i	30	28	32	25	25	25	22	24	35	40
y_i	25	30	27	40	42	40	50	45	30	25
$\hat{y}_i$	33.5	36.21	30.79	40.27	40.27	40.27	44.33	41.62	26.73	19.96
e_i	-8.50	-6.21	-3.79	-0.27	1.72	-0.27	5.66	3.37	3.26	5.03

The residual variance is:

$$s_R^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{207.92}{8} = 25.99$$

Estimation of the variance

```
Regression Analysis - Linear model: Y = a + b*X
-----------------------------------------------------------------------------
Dependent variable: Precio en ptas.
Independent variable: Produccion en kg.
-----------------------------------------------------------------------------
                                Standard          T
Parameter       Estimate          Error       Statistic         P-Value
-----------------------------------------------------------------------------
Intercept        74,1151        8,73577         8,4841          0,0000
Slope           -1,35368         0,3002       -4,50924          0,0020
-----------------------------------------------------------------------------


                           Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                    528,475       1      528,475      20,33       0,0020
Residual                 207,925       8      25,9906
-----------------------------------------------------------------------------
Total (Corr.)              736,4       9

Correlation Coefficient = -0,84714
R-squared = 71,7647 percent
Standard Error of Est. = 5,0981
```

$\hat{S}_R^2$

Inference on the regression model

- Up to this point we have obtained only point estimates for the regression coefficients
- Using confidence intervals we can obtain a measure of the precision of the above mentioned estimates
- Using hypothesis testing we can verify if a certain value can be the true value of the parameter

Inference about the slope

The estimator $\hat{\beta}_1$ follows a normal distribution because it is a linear combination of normals,

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

where $y_i = \beta_0 + \beta_1 x_i + u_i$, satisfying $y_i \sim N\left(\beta_0 + \beta_1 x_i, \sigma^2\right)$.
Additionally, $\hat{\beta}_1$ is an unbiased estimator for β_1,

$$E\left[\hat{\beta}_1\right] = \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{(n-1)s_X^2} E\left[y_i\right] = \beta_1$$

and its variance is given by

$$Var\left[\hat{\beta}_1\right] = \sum_{i=1}^{n} \left(\frac{(x_i - \bar{x})}{(n-1)s_X^2}\right)^2 Var\left[y_i\right] = \frac{\sigma^2}{(n-1)s_X^2}$$

Thus,

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

Confidence intervals for the slope

We want to obtain a confidence interval for β_1 at a $1?\alpha$ level. Since σ^2 is unknown, it will be estimated using s_R^2. The basic result when the variance is unknown is:

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\dfrac{s_R^2}{(n-1)s_X^2}}} \sim t_{n-2}$$

that allows us to obtain a confidence interval for β_1:

$$\hat{\beta}_1 \pm t_{n-2,\alpha/2}\sqrt{\frac{s_R^2}{(n-1)s_X^2}}$$

The length of this interval will decrease if:

- The sample size increases
- The variance of the independent observations x_i increases
- The residual variance decreases

Hypothesis testing on the slope

Using the previous result we can perform hypothesis testing for β_1. In particular, if the true value of β_1 is zero then Y does not depend linearly on X. Therefore, the following contrast is of special interest:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

The rejection region for the null hypothesis is:

$$\left| \frac{\hat{\beta}_1}{\sqrt{s_R^2/(n-1)s_X^2}} \right| > t_{n-2,\alpha/2}$$

Equivalently, if the value zero is outside of the confidence interval for β_1 at a $1?\alpha$ level, we reject the null hypothesis at this level. The p-value of the test is:

$$p\text{-value} = 2\Pr\left(t_{n-2} > \left| \frac{\hat{\beta}_1}{\sqrt{s_R^2/(n-1)s_X^2}} \right| \right)$$

Inference for the slope

Exercise 4.3

1. Compute a 95% confidence interval for the slope of the regression line obtained in exercise 4.1
2. Test the hypothesis that the price of flour depends linearly on the production of wheat, using a 0.05 significance level

Results

1. $t_{n-2,\alpha/2} = t_{8,0.025} = 2.306$

$$-2.306 \leq \frac{-1.3537 - \beta_1}{\sqrt{\frac{25.99}{9 \times 32.04}}} \leq 2.306$$

$$-2.046 \leq \beta_1 \leq -0.661$$

2. As the interval does not contain the value zero, we reject $\beta_1 = 0$ at the 0.05 level. In fact:

$$\left| \frac{\hat{\beta}_1}{\sqrt{s_R^2 / (n-1)\, s_X^2}} \right| = \left| \frac{-1.3537}{\sqrt{\frac{25.99}{9 \times 32.04}}} \right| = 4.509 > 2.306$$

p-value $= 2\Pr(t_8 > 4.509) = 0.002$

Inference for the slope

Exercise 4.3

1. Compute a 95% confidence interval for the slope of the regression line obtained in exercise 4.1
2. Test the hypothesis that the price of flour depends linearly on the production of wheat, using a 0.05 significance level

Results

1. $t_{n-2,\alpha/2} = t_{8,0.025} = 2.306$

$$-2.306 \leq \frac{-1.3537 - \beta_1}{\sqrt{\frac{25.99}{9\times 32.04}}} \leq 2.306$$

$$-2.046 \leq \beta_1 \leq -0.661$$

2. As the interval does not contain the value zero, we reject $\beta_1 = 0$ at the 0.05 level. In fact:

$$\left| \frac{\hat{\beta}_1}{\sqrt{s_R^2 / (n-1)\, s_X^2}} \right| = \left| \frac{-1.3537}{\sqrt{\frac{25.99}{9\times 32.04}}} \right| = 4.509 > 2.306$$

p-value $= 2\,\Pr(t_8 > 4.509) = 0.002$

Inference for the slope

$$\sqrt{\frac{s_R^2}{(n-1)s_X^2}} \qquad\qquad \frac{\hat{\beta}_1}{\sqrt{s_R^2/(n-1)s_X^2}}$$

```
Regression Analysis - Linear model: Y = a + b*X
-----------------------------------------------------------------------------
Dependent variable: Precio en ptas.
Independent variable: Produccion en kg.
-----------------------------------------------------------------------------
                                Standard            T
Parameter        Estimate         Error         Statistic        P-Value
-----------------------------------------------------------------------------
Intercept         74,1151        8,73577          8,4841          0,0000
Slope            -1,35368         0,3002        -4,50924          0,0020
-----------------------------------------------------------------------------


                           Analysis of Variance
-----------------------------------------------------------------------------
Source              Sum of Squares     Df  Mean Square     F-Ratio     P-Value
-----------------------------------------------------------------------------
Model                      528,475      1      528,475       20,33      0,0020
Residual                   207,925      8      25,9906
-----------------------------------------------------------------------------
Total (Corr.)                736,4      9

Correlation Coefficient = -0,84714
R-squared = 71,7647 percent
Standard Error of Est. = 5,0981
```

Inference for the intercept

The estimator $\hat{\beta}_0$ follows a normal distribution, as it can be written as a linear combination of normal random varaibles,

$$\hat{\beta}_0 = \sum_{i=1}^{n} \left(\frac{1}{n} - \bar{x} w_i \right) y_i$$

where $w_i = (x_i - \bar{x}) / n s_X^2$ and $y_i = \beta_0 + \beta_1 x_i + u_i$, satisfying $y_i \sim N\left(\beta_0 + \beta_1 x_i, \sigma^2\right)$. Additionally, $\hat{\beta}_0$ is an unbiased estimator for β_0,

$$E\left[\hat{\beta}_0\right] = \sum_{i=1}^{n} \left(\frac{1}{n} - \bar{x} w_i \right) E\left[y_i\right] = \beta_0$$

and its variance is

$$Var\left[\hat{\beta}_0\right] = \sum_{i=1}^{n} \left(\frac{1}{n} - \bar{x} w_i \right)^2 Var\left[y_i\right] = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1) s_X^2} \right)$$

implying

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

Confidence interval for the intercept

We wish to obtain a confidence interval for β_0 at a $1?\alpha$ level. Since σ^2 is unknown, this value will be estimated using s_R^2. The basic result when the variance is unknown is:

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{s_R^2\left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_X^2}\right)}} \sim t_{n-2}$$

From it we can obtain a confidence interval for β_0:

$$\hat{\beta}_0 \pm t_{n-2,\alpha/2}\sqrt{s_R^2\left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_X^2}\right)}$$

The length of the confidence interval decreases if:

- The sample size increases
- The variance of the independent observations x_i increases
- The residual variance decreases
- The mean of the independent observations decreases

Hypothesis testing on the intercept

Using the previous result we can perform hypothesis testing for β_0. In particular, if the true value of β_0 is zero then the regression line passes through the origin. Therefore, the following test is of special interest:

$$H_0 : \beta_0 = 0$$
$$H_1 : \beta_0 \neq 0$$

The critical region for this test is:

$$\left| \frac{\hat{\beta}_0}{\sqrt{s_R^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_X^2} \right)}} \right| > t_{n-2,\alpha/2}$$

Equivalently, if zero lies outside the confidence interval for β_0 at a level $1-\alpha$, we reject the null hypothesis at that level. The p-value is given by:

$$p\text{-value} = 2\Pr\left(t_{n-2} > \left| \frac{\hat{\beta}_0}{\sqrt{s_R^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_X^2} \right)}} \right| \right)$$

Inference for the intercept

Exercise 4.4

1. Compute a 95% confidence interval for the intercept of the regression line obtained in exercise 4.1
2. Test the hypothesis that the regression line passes through the origin, using a 0.05 significance level

Results

1. $t_{n-2,\alpha/2} = t_{8,0.025} = 2.306$

$$-2.306 \leq \frac{74.1151 - \beta_0}{\sqrt{25.99\left(\frac{1}{10} + \frac{28.6^2}{9 \times 32.04}\right)}} \leq 2.306 \Leftrightarrow 53.969 \leq \beta_0 \leq 94.261$$

2. As the computed interval does not contain the value zero, we reject that $\beta_0 = 0$ at the level 0.05. In fact,

$$\left| \frac{\hat{\beta}_0}{\sqrt{s_R^2\left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_X^2}\right)}} \right| = \left| \frac{74.1151}{\sqrt{25.99\left(\frac{1}{10} + \frac{28.6^2}{9 \times 32.04}\right)}} \right| = 8.484 > 2.306$$

p-value $= 2\Pr(t_8 > 8.483) = 0.000$

Inference for the intercept

Exercise 4.4

1. Compute a 95% confidence interval for the intercept of the regression line obtained in exercise 4.1
2. Test the hypothesis that the regression line passes through the origin, using a 0.05 significance level

Results

1. $t_{n-2,\alpha/2} = t_{8,0.025} = 2.306$

$$-2.306 \leq \frac{74.1151 - \beta_0}{\sqrt{25.99\left(\frac{1}{10} + \frac{28.6^2}{9 \times 32.04}\right)}} \leq 2.306 \Leftrightarrow 53.969 \leq \beta_0 \leq 94.261$$

2. As the computed interval does not contain the value zero, we reject that $\beta_0 = 0$ at the level 0.05. In fact,

$$\left| \frac{\hat{\beta}_0}{\sqrt{s_R^2\left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_X^2}\right)}} \right| = \left| \frac{74.1151}{\sqrt{25.99\left(\frac{1}{10} + \frac{28.6^2}{9 \times 32.04}\right)}} \right| = 8.484 > 2.306$$

p-value $= 2\,\Pr(t_8 > 8.483) = 0.000$

Inference for the intercept

$$\sqrt{s_R^2\left(\frac{1}{n}+\frac{\bar{x}^2}{(n-1)s_X^2}\right)}$$

$$\frac{\hat{\beta}_0}{\sqrt{s_R^2\left(\frac{1}{n}+\frac{\bar{x}^2}{(n-1)s_X^2}\right)}}$$

```
Regression Analysis - Linear model: Y = a + b*X
-----------------------------------------------------------------------------
Dependent variable: Precio en ptas.
Independent variable: Produccion en kg.
-----------------------------------------------------------------------------
                                Standard          T
Parameter       Estimate          Error       Statistic        P-Value
-----------------------------------------------------------------------------
Intercept        74,1151         8,73577        8,4841          0,0000
Slope           -1,35368          0,3002      -4,50924          0,0020
-----------------------------------------------------------------------------


                           Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                     528,475      1      528,475      20,33       0,0020
Residual                  207,925      8      25,9906
-----------------------------------------------------------------------------
Total (Corr.)               736,4      9

Correlation Coefficient = -0,84714
R-squared = 71,7647 percent
Standard Error of Est. = 5,0981
```

Inference for the variance

The basic result is:

$$\frac{(n-2)\, s_R^2}{\sigma^2} \sim \chi^2_{n-2}$$

Using this result we can:

- Compute a confidence interval for the variance:

$$\frac{(n-2)\, s_R^2}{\chi^2_{n-2,\alpha/2}} \leq \sigma^2 \leq \frac{(n-2)\, s_R^2}{\chi^2_{n-2,1-\alpha/2}}$$

- Perform hypothesis testing of the form:

$$H_0 : \sigma^2 = \sigma_0^2$$
$$H_1 : \sigma^2 \neq \sigma_0^2$$

Estimation of the mean response and prediction of a new response

We consider two types of problems:

1. Estimate the mean value of the variable Y corresponding to a given value $X = x_0$
2. Predict a future value of the variable Y for a given value $X = x_0$

For example, in exercise 4.1 we might be interested in the following questions:

1. What will be the mean price of a Kg of flour in those years where wheat production equals 30 tons?
2. If in a given year wheat production is 30 tons, which will be the price of a Kg of flour?

In both cases the estimation is:

$$\begin{aligned}\hat{y}_0 &= \hat{\beta}_0 + \hat{\beta}_1 x_0 \\ &= \bar{y} + \hat{\beta}_1 (x_0 - \bar{x})\end{aligned}$$

but the precision in the estimations is different

Estimation of a mean response

Taking into account that:

$$Var\left(\hat{y}_0\right) = Var\left(\bar{y}\right) + \left(x_0 - \bar{x}\right)^2 Var\left(\hat{\beta}_1\right)$$
$$= \sigma^2 \left(\frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{\left(n-1\right) s_X^2}\right)$$

the confidence interval for the mean response is:

$$\hat{y}_0 \pm t_{n-2,\alpha/2} \sqrt{s_R^2 \left(\frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{\left(n-1\right) s_X^2}\right)}$$

Prediction of a new response

The variance of the prediction of a new response is the mean squared error of the prediction:

$$\begin{aligned} E\left[(y_0 - \hat{y}_0)^2\right] &= Var\,(y_0) + Var\,(\hat{y}_0) \\ &= \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)\, s_X^2}\right) \end{aligned}$$

The confidence interval for the prediction of a new response is:

$$\hat{y}_0 \pm t_{n-2,\alpha/2} \sqrt{s_R^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)\, s_X^2}\right)}$$

The length of this interval is larger than the one for the preceding case (we have less precision), as we are not estimating a mean value but a specific one.

Estimation of the mean response and prediction of a new response

The intervals for the estimated means are shown in red in the figure below, while those for the predictions are drawn in pink
It is readily apparent that the size of the latter ones is considerably larger