11. SIMPLE LINEAR REGRESSION

Study time: 60 minutes

Aim - you will be able to

- explain a general linear model notion
- explain a linear regression model principle
- use regression analysis results
- verify a regression model by determination index

Explication

11.1. Introduction

Mathematical formulation of statistical models

Symbolically, the basic additive formulation of statistical models can be expressed as

$$\underline{Y} = f(X) + \zeta(\varepsilon)$$

where Y is the observed value, $f(X)$ is the systematic component and $\zeta(\varepsilon)$ is the random component. This schematic model explicitly identifies three type of variables.

Y – Response, Criterion, dependent Variable (observed value of primary interest)
X– Predictor, Stimulus, Independent Variable (those factors to which the value of the systematic component may be attributed)
ε - random error

Only Y and X are observable. Random error is always unobservable.
$\zeta(\varepsilon)$ is always estimated as the residua difference between the estimated systematic component and the observed response, Y.

$$\overline{\zeta(\varepsilon)} = Y - \overline{f(X)}$$

Therefore the estimated split of the observed response into its systematic and random components is as much a consequence of the choice of models, f and ζ, and the method of estimation as it is of the observed stimulus and response, X and Y.

11.2. General linear model

The general linear statistical model is a special simple case of the schematic statistical models discussed above. The so-called linear statistical model stipulates that the systematic component is a linear combination of the systematic factors or variables, and the random component is the identity function of random error.

- **random component:** $\zeta(\varepsilon) = \varepsilon$
- **systematic component**: $f(X) = \beta_0 + \sum_{i=1}^{p} \beta_i X_i$

Why use a Linear Systematic Function?

Linear systematic components have three fundamental properties which are desirable for statistical models – *simplicity, estimability* and *stability*.

Linear functions represent or give algebraic expression to the simplest kind of relationship. linear functions postulate either:

- stimulus and response tend to increase and decrease together
- response decreases as stimulus increases

For the simple linear model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

if $\beta_1 < 0$; the relation is negative => Y decreases as X increases
if $\beta_1 > 0$; the relation is positive => Y and X increase together

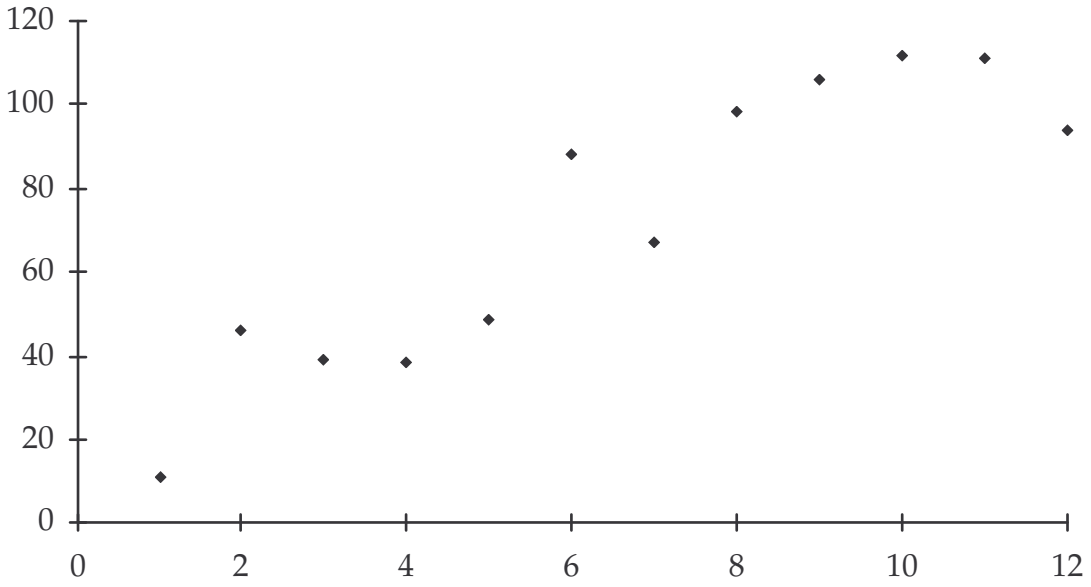
Assumptions about the random component

In decreasing order of impact on results and interpretation, the following three assumptions about the behavior of the random component of a linear statistical model are widely adopted.

1. *Independence* – the random errors ε_i and ε_j are independent for all pairs of observations i and j
2. *Equal Variance* – the random errors ε_i all have the same variance σ^2 for all observations
3. *Normality* – the random errors ε_i are normally distributed

11.3. Estimation of parameters for the simple linear regression model

The following scatter plot illustrates the type of data which is typically described by a simple linear model.



From the formulation of the general linear model, the special case of the simple linear model in which the systematic component is a linear function of a single variable, that is a straight line, may be expressed as:

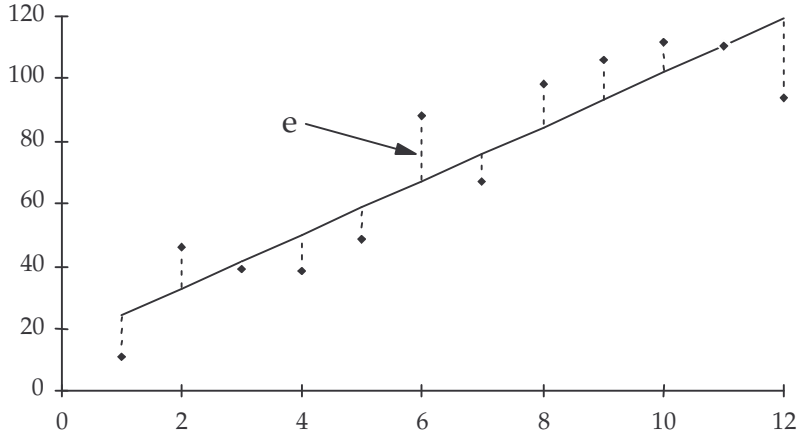
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$
$$\varepsilon_i \rightarrow N\left(0, \sigma^2\right)$$

and all ε_i are mutually independent.

For any estimates of the parameters, β_0 and β_1, say b_0 and b_1, the residual errors of estimation are:

$$e_i = Y_i - b_0 - b_1 X_i$$

as illustrated below.



The least squares parameter estimates are those values of b_0 and b_1 which minimize the sum of squared residual errors.

$$S(b_0, b_1) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

To find the parameter estimates which minimize the sum of squared residuals, we compute the derivatives with respect to b_0 and b_1 and equate them to zero.

$$\frac{\partial S(b_0, b_1)}{\partial b_0} = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = \sum_{i=1}^{n} e_i = 0$$
$$\frac{\partial S(b_0, b_1)}{\partial b_0} = \sum_{i=1}^{n} x_i (y_i - b_0 - b_1 x_i) = \sum_{i=1}^{n} x_i e_i = 0$$

The solutions to the above equations are the least squares parameter estimates. Notice that the first equation insures that the residuals for the least squares estimates of β_0 and β_1 always sum to zero.
Because the least squares estimates are also maximum likelihood estimates under the assumption of normally distributed errors, they are usually denoted by the symbols β_0 and β_1. The solutions to the least squares equations are:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = r_{xy} \frac{s_y}{s_x}$$

The intercept parameter, $\hat{\beta}_0$, merely places the vertical position of the line at the point where the residual errors sum to zero. The operative parameter is the slope estimate, $\hat{\beta}_1$, which has a particularly simple form in terms of the correlation and relative standard deviations of the response *Y* and the explanatory variable *X*.

$$\hat{\beta}_1 = \overbrace{r_{xy}}^{\substack{\text{Relation} \\ \text{Between} \\ X \text{ and } Y}} \quad \overbrace{\frac{s_y}{s_x}}^{\substack{\text{Scale} \\ \text{Factor}}}$$

The residual sum of squares for the simple regression model is

$$S(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = (1 - r_{xy}^2) \sum_{i=1}^{n} (y_i - \bar{y})^2 = (1 - r_{xy}^2)(n-1) s_y^2$$

which like the least squares slope estimate, $\hat{\beta}_1$, has a simple expression in terms of the correlation between *X* and *Y* and the variance of *Y*.

The residual sum of squares for a regression model measures how well the model fits the data. A smaller residual sum of squares indicates a better fit. Because a higher squared correlation between X and Y is associated with a smaller residual sum of squares as a proportion of the variance of Y, the squared correlation between X and Y is usually used as a measure of the goodness of fit of the regression model. When $r_{xy} = \pm 1$, the sample observations of X and Y all lie on a straight line and the residual sum of squares is zero. When $r_{xy} = 0$, X and Y are independent and the residual sum of squares will equal the sum of squared deviations of Y about its mean.

If the residual sum of squares measures the size of the random component of the regression model, then the remainder, the difference between the original sum of squared deviations of Y about its mean and the residual sum of squares of Y about the regression line must represent the systematic component of the model. To better understand what this systematic component measures, let the point on the regression line or predicted value of Y for the i^{th} observation of X be

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

Firstly, note that the least squares estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ insure that the mean of the predicted value of Y will always equal the mean of the original observations of Y. That is,

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

Then as was the case in the analysis of variance, the total sum of squared deviations of Y from its mean

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

can be partitioned into the sum of squared residual errors,

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

and the sum of squared deviations of the predicted values of Y from their mean.

$$SS_{\mathrm{Re}\,gression} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

We see that

$$SS_{Total} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$
$$= SS_{\mathrm{Re}\,gression} + SS_{Error}$$

The sum of squares due to regression is often called the explained variation and conversely the sum of squared residual errors, the unexplained variation. The partitioning of the total variation of *Y* into these two components is due to the fact that the least squares estimates must satisfy the condition,

$$\sum_{i=1}^{n} x_i e_i = 0$$

That is, the residual errors must be orthogonal to the predictor variable.

The partitioning of the total sum of squared deviations of the response, *Y*, about its mean into the systematic component, explained variation, and the random component, sum of squared residuals is frequently presented as an Analysis of Variance table. The F-test computed by this ANOVA Table tests the null hypothesis that the systematic component of the model is zero.

Source	Degrees of Freedom	Sum of Squares	Mean Squares	F-ratio
Total	*n*-1	$(n-1)s_y^2$		
Regression	1	$r_{xy}^2(n-1)s_y^2$	$r_{xy}^2(n-1)s_y^2$	$\frac{(n-2)r_{xy}^2}{(1-r_{xy}^2)}$
Error	*n*-2	$(1-r_{xy}^2)(n-1)s_y^2$	$\frac{(1-r_{xy}^2)(n-1)s_y^2}{(n-2)}$	

Thus, the F-test for testing the significance of the regression model depends only on the correlation between response and explanatory variables and on the sample size. In practice, the null hypothesis of no regression effect is almost always rejected, but even if rejected does not imply that the regression model will provide satisfactory predictions.

As in the case of analysis of variance for factorial models, the estimate of the error variance, σ^2, is the mean squared error.

$$\hat{\sigma}^2 = {SS_{Error}}/{n-2} = (1-r_{xy}^2)\left(\frac{n-1}{n-2}\right)s_y^2$$

This estimated error variance for the regression line is also called the conditional variance of *Y* given *X*, that is, the variance of *Y* remaining after the effect of *X* has been removed.

$$s_{y|x}^2 = \hat{\sigma}^2 = (1-r_{xy}^2)\left(\frac{n-1}{n-2}\right)s_y^2$$

A second consequence of least squares estimates of β_0 and β_1 is that the least squares line will always pass through the point of means $(\overline{X}, \overline{Y})$. In fact the z-value of the prediction for *Y* is simply the correlation between *X* and *Y* times the corresponding z-value for *X*. That is,

$$\hat{y}_i = \overline{y} - \left(r\frac{s_y}{s_x}\right)\overline{x} + \left(r\frac{s_y}{s_x}\right)x_i$$

$$(\hat{y}_i - \bar{y}) = \left(r \frac{s_y}{s_x} \right)(x_i - \bar{x})$$

$$\left(\frac{\hat{y}_i - \bar{y}}{s_y} \right) = r \left(\frac{x_i - \bar{x}}{s_x} \right).$$

Clearly when $x_i = \bar{x}$, then $\hat{y}_i = \bar{y}$.

11.4. Distribution of least squares parameter estimates

If the predictor or explanatory variable *X* is assumed to be a fixed constant rather than a random variable, then both $\hat{\beta}_0$ and $\hat{\beta}_1$ are linear combinations of the normally distributed criterion or response variable, *Y*, and hence are normally distributed themselves. The mean and variance of the slope parameter estimate are

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

These results can readily be established by noting that the least squares estimate of β_1 may be expressed as the following linear combination of the observations of *Y*.

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

Then the expected value of the slope parameter estimate is

$$E(\hat{\beta}_1) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{(n-1)s_x^2} \mathrm{E}(y_i) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(\beta_0 + \beta_1 x_i)}{(n-1)s_x^2}$$

and the variance of the slope estimate is

$$V(\hat{\beta}_1) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{(n-1)s_x^2} \mathrm{V}(y_i) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{(n-1)^2 s_x^4} \sigma^2 .$$

The significance of these results is that the least squares estimate of the slope parameter is unbiased and its variance becomes smaller as the sample size increases. In addition, the variance of the estimate becomes smaller when the variance or range of *X* becomes larger.

By substitution of the mean squared error estimate of σ^2 into the expression of the variance of the slope parameter estimate, the following estimated variance of the slope parameter is obtained

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{SS_{Error}}{(n-1)(n-2)s_x^2} = \frac{\left(1-r_{xy}^2\right)s_y^2}{(n-2)s_x^2}$$

Because $\hat{\beta}_1$ is unbiased, substitution into the least squares determining equation for $\hat{\beta}_0$ readily shows that the least squares intercept estimate, $\hat{\beta}_0$, is also unbiased.

$$E\left[\hat{\beta}_0\right] = \beta_0$$

The variance of $\hat{\beta}_0$ is obtained by again noting that from the least squares determining equation, $\hat{\beta}_0$ can be expressed as the following linear combination of the observations of *Y*.

$$\hat{\beta}_0 = \sum_{i=1}^{n}\left[\frac{1}{n} - \frac{\bar{x}\left(x_i-\bar{x}\right)}{(n-1)s_x^2}\right]y_i$$

By squaring and summing constant terms in this linear combination, the variance is found to be

$$V\left[\hat{\beta}_0\right] = \frac{\sigma^2\sum_{i=1}^{n}x_i^2}{n\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2} = \sigma^2\left[\frac{1}{n} + \frac{n\,\bar{x}^2}{(n-1)s_x^2}\right]$$

This expression for the variance of the least squares estimate of the intercept consists of two parts, the reciprocal of the sample size, *n*, which is the usual factor for the variance of a mean, and the ratio of the square of the mean of *X* to its variance. As for $\hat{\beta}_1$, an estimate of the variance of $\hat{\beta}_0$ can be obtained by substituting the mean squared error estimate of σ^2.

$$\hat{\sigma}^2_{\hat{\beta}_0} = \frac{\mathrm{SS_{Error}}}{\mathrm{n-2}}\left[\frac{1}{n} + \frac{n\,\bar{x}^2}{(n-1)s_x^2}\right] = \frac{\left(1-\mathrm{r_{xy}^2}\right)(n-1)s_y^2}{\mathrm{n-2}}\left[\frac{1}{n} + \frac{n\,\bar{x}^2}{(n-1)s_x^2}\right]$$

Since the predicted value, $\hat{y}_i$, is also a linear combination of the least squares parameter estimates, it too will be normally distributed.

$$\hat{y}_i = \sum_{k=1}^{n}\left[\frac{1}{n} + \frac{\left(x_i-\bar{x}\right)\left(x_k-\bar{x}\right)}{\left(n-1\right)s_x^2}\right]y_k$$

The expected value of $\hat{y}_i$ is obtained by direct substitution into the linear prediction equation.

$$E[\hat{y}_i] = E[\hat{\beta}_0] + E[\hat{\beta}_1] x_i = \beta_0 + \beta_1 x_i$$

As for $\hat{\beta}_0$ and $\hat{\beta}_1$, the variance of $\hat{y}_i$ is derived by squaring and summing the constant terms in the expression of $\hat{y}_i$ as a linear combination of the observations of *Y*.

$$V[\hat{y}_i] = \sigma^2 \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{k=1}^{n} (x_k - \bar{x})^2} \right) = \sigma^2 \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{(n-1)s_x^2} \right)$$

Again notice that the expression for the variance of the predicted value for the i^{th} observation of *Y* consists of two components. The first component, the reciprocal of the sample size, is the usual factor for the variance of a mean. The second component is a normalized squared distance of x_i, the i^{th} observation of the explanatory variable, from its mean. Thus the variances of predictions of *Y* for values of x_i near its mean will be close to the variance of an ordinary sample mean. But for values of x_i far from its mean, the variances of the predictions will increase linearly with the squared normalized distance from the mean.

$$V[\hat{y}_i] = \sigma^2 \{ \overbrace{\frac{1}{n}}^{\text{Variance of Mean of } Y} + \overbrace{\frac{(x_i - \bar{x})^2}{(n-1)s_x^2}}^{\text{Normed Distance From Mean of } X} \}$$

The foregoing expression for the variance of $\hat{y}_i$ is the variance of the estimate of the regression line, which is the conditional mean of *Y* given *X*. But the variance of a prediction for single observation of *Y* at *X* will be much greater. This prediction error will be the original variance of *Y*, σ^2, plus the variance of error due to estimation of the regression line. Therefore, the estimated variance of a single observation or prediction at X_i is

$$\hat{\sigma}^2_{\hat{y}_i} = \hat{\sigma}^2 \{ \overbrace{1}^{\text{Single Observation}} + \overbrace{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{(n-1)s_x^2}}^{\text{Regression Line}} \}$$

Both estimates of the regression line and predictions from the regression line will be more accurate for values of *X* near the mean.

11.5. Inference for the regression line

It is often of interest to test hypotheses about the parameters of the regression model or to construct confidence intervals for various quantities associated with the model. There may be theoretically prescribed values for the parameters. Confidence intervals for predictions from the regression model are frequently required. Inferential procedures follow in a natural way from the fact that least squares parameter estimates and hence the estimated regression line is all linear combination of the response variable, *Y*, and like *Y* will be normally distributed. In addition, the estimated variances of these parameters are derived from the sum of squared deviations of *Y* and hence will have a χ^2 distribution. Therefore, the following statistics all have Student's t distributions with (*n*-2) degrees of freedom.

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma} \Big/ \sqrt{(n-1)}s_x} \Rightarrow t_{n-2}$$

$$\frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_{\hat{\beta}_0}} = \frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}\sqrt{\left[\frac{1}{n} + \frac{n\,\bar{x}^2}{(n-1)s_x^2}\right]}} \Rightarrow t_{n-2}$$

$$\frac{\hat{y}_i - \beta_0 - \beta_1 x_i}{\hat{\sigma}_{\hat{y}_i}} = \frac{y_i - \beta_0 - \beta_1 x_i}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{(n-1)s_x^2}}} \Rightarrow t_{n-2}$$

where β_0 and β_1 are the true or hypothesized values of the regression parameters. The most commonly tested null hypothesis is that the slope and intercept equal zero. This test is the t-test produced by most regression software. For the slope parameter, this test has a particularly interesting interpretation.

$$\frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{r_{xy}\dfrac{s_y}{s_x}}{\sqrt{\dfrac{(1-r_{xy}^2)}{(n-2)}}\dfrac{s_y}{s_x}} = \frac{r_{xy}}{\sqrt{\dfrac{(1-r_{xy}^2)}{(n-2)}}}$$

which is simply the square root of the F-test for the regression model derived earlier. Thus, testing whether the systematic component is zero is equivalent to testing whether the slope of the regression line is zero. If $\beta_1 = 0$, then the regression line will be horizontal at the mean of *Y*, that is, the mean of *Y* will be predicted at every value of *X* and *X* will have no effect on predictions of *Y*.

Confidence intervals for the intercept, slope, regression line, and predictions from the regression line are calculated in the usual manner.

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2} \hat{\sigma} \Big/ \sqrt{(n-1)}s_x$$

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

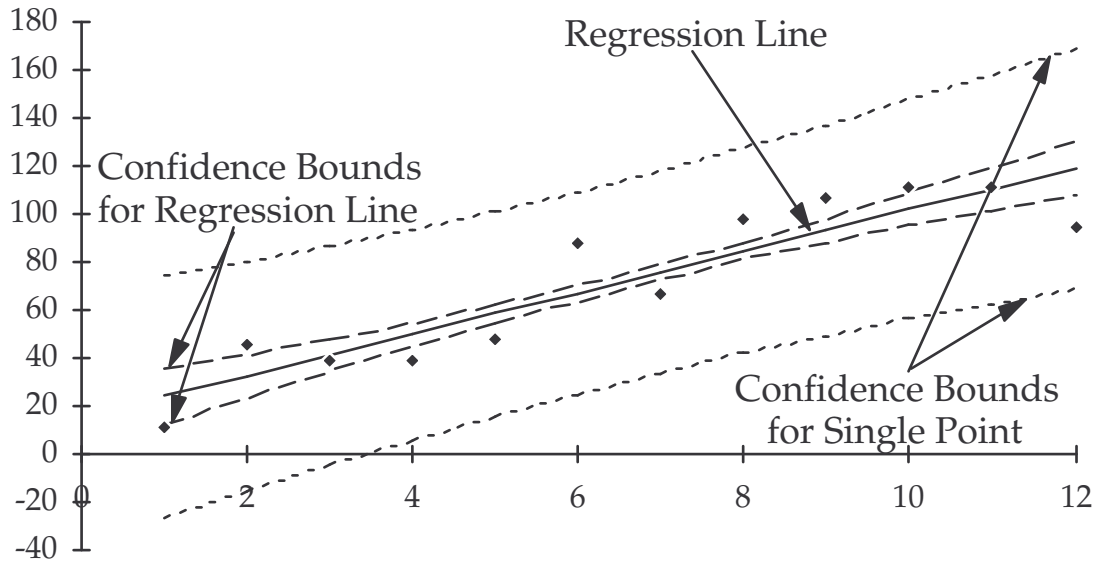
The confidence interval for the regression line is

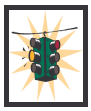
$$\hat{\beta}_0 + \hat{\beta}_1 x_i \pm t_{\alpha/2,n-2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{(n-1)s_x^2}}$$

And the confidence interval for predictions from the regression line is

$$\hat{\beta}_0 + \hat{\beta}_1 x_i \pm t_{\alpha/2,n-2}\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_i-\bar{x})^2}{(n-1)s_x^2}}$$

The following chart displays 95% confidence intervals for both the regression line and individual predictions from the regression line. Notice that the confidence bounds for the regression line are very narrow and include very few of the original data points. This is because the correlation between predictor and criterion variables is high and the fit of the regression line is good. On the other hand, all original observations are included within the confidence bounds for single points.





Solved example

The company repairs the desktop calculators and cashes. The data from 18 repairs are written in the table. Each repair has 2 important data. The former is a number of repaired calculators (X) and the latter is a total repair time (Y).

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
x	7	6	5	1	5	4	7	3	4	2	8	5	2	5	7	1	4	5
Y	97	86	78	10	75	62	101	39	53	33	118	65	25	71	105	17	49	68

a) Find parameter estimates of the regression line.
b) Draw data and regression function.
c) Use t-tests for the values of all parameters of regression function.

Solution

– we can use STATGRAPHIC software:

Linear regression – Repair time vs. Number

Regression Analysis - Linear model: Y = b_0 + b_1 *x

Dependent variable: Repair Time
Independent variable: Number

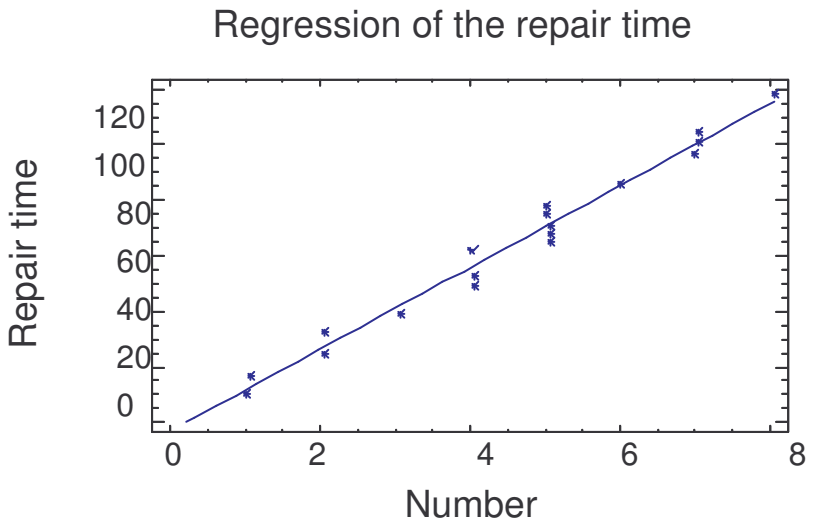
Parameter	Estimate	Standard Error	T Statistic	P-Value
b_0 - Intercept	-2,32215	2,56435	-0,905549	0,3786
b_1 - Slope	14,7383	0,519257	28,3834	0,0000

---

b_0 = *Intercept*, b_1 = *Slope*, the results of these values may be found in the second column. The following function introduces an equation for the estimate of predicted value:

Repair Time = -2,32215 + 14,7383 . Number

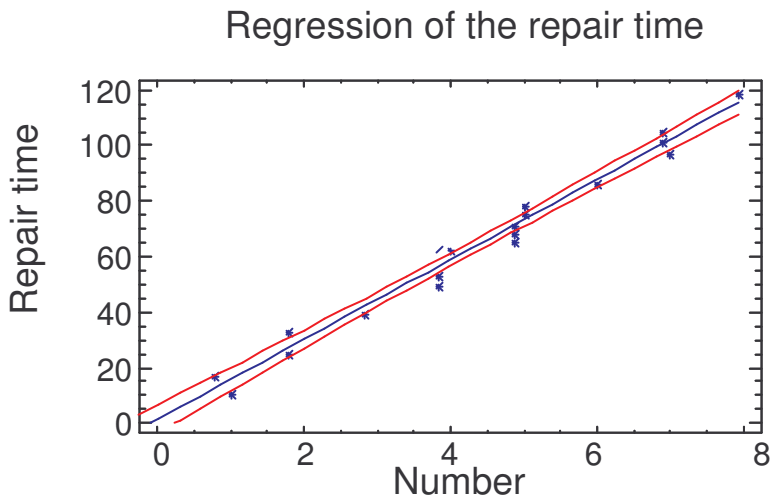
The observed values of the t-tests are shown in the fourth column (T Statistic) and corresponding p-values are displayed in the last one. It is obvious that hypothesis H_0: $\beta_0=0$ will not be rejected considering the important value in p-value column. Based on this, we can say that regression line passes through the beginning what is a logical conclusion, considering the data nature. The second of particular test says that Slope is a value that significantly differs from zero since we have rejected H_0 hypothesis H_0: $\beta_1=0$.



d) Let's find the 95% confidential interval for the repair time in dependence on the number of calculators.

e) Let's find point and interval estimation for an expected repair time for 5 calculators.

Solution



For value $x=5$:

$$\hat{Y}(x) = b_0 + b_1 x = \sum_{i=1}^{n} \left(\frac{1}{n} + \frac{(x_i - \bar{x}) \cdot (x - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) \cdot Y_i = 71.3691$$

$$\mathrm{E}(\mathrm{Y}|\,\mathrm{x}) = \beta_0 + \beta_1 x \in < \hat{Y}(x) - S_{\hat{Y}} \cdot t_{1-\frac{\alpha}{2}}(n-2), \hat{Y}(x) + S_{\hat{Y}} \cdot t_{1-\frac{\alpha}{2}}(n-2) >= < 69.063,\ 73.6752 >$$

f) Consider the quality of examined model of linear regression for the repair time in dependence on a number of calculators using a coefficient of determination

Solution

$$SS_Y = SS_{\hat{Y}} + SS_R$$

Source		Sum of Squares
Regression	$SS_{\hat{Y}}$	16182,6
Error	SS_R	321,396
Total	SS_Y	16504,0

$$I^2 = \frac{SS_{\hat{Y}}}{SS_Y} = 98.0526\ \%$$

Summary of notion

Regression model is a special case of general linear model. The basic assumptions are independence, homoscedasticity and normality.

Dependent variable is the variable of a regression model that is random and we try explaining its behavior and describing by mathematical curve.

Independent (explanatory) variables are the variables in the regression model whose behavior explains the behavior of the dependent variable.

Linear regression model with one explanatory variable is a basic model and it is based on the **Least-Squares Method**. By this method we can determine model parameters. The sum of squared deviations of the real values from modeled values is called the residual sum of squares.

We can obtain interval estimation for the expected value of the dependent variable. These interval bounds form **confidential interval** of the regression line.

Question

1. Describe and explain equation of linear regression.
2. What means p-value in the ANOVA table for linear regression?
3. What property describes a coefficient of determination?

Problems

Example 1: During control measurements of industrial components size we randomly chosen 8 components showing mostly positive divergences from normal values in the length and height*:*

length divergence [mm]	3	4	4	5	8	10	6	3
height divergence [mm]	4	6	5	6	7	13	9	4

Let's find the linear regression model of dependency between the length divergence and height divergence.

{*Answer:* Use a suitable software package.}

Example 2: In the years 1931-1961, water flow in profile of Šance and Morávka water reservoirs were measured. Averages per year (m^3/s) are given by the following table:

year	Šance	Morávka
1931	4,130	2,476
1932	2,386	1,352
1933	2,576	1,238
1934	2,466	1,725
1935	3,576	1,820
1936	2,822	1,913
1937	3,863	2,354
1938	3,706	2,268
1939	3,710	2,534
1940	4,049	2,308
1941	4,466	2,517
1942	2,584	1,726
1943	2,318	1,631
1944	3,721	2,028
1945	3,290	2,423

year	Šance	Morávka
1946	2,608	1,374
1947	2,045	1,194
1948	3,543	1,799
1949	4,055	2,402
1950	2,224	1,019
1951	2,740	1,552
1952	3,792	1,929
1953	3,087	1,488
1954	1,677	0,803
1955	2,862	1,878
1956	3,802	1,241
1957	2,509	1,165
1958	3,656	1,872
1959	2,447	1,381
1960	2,717	1,679

Let's assume that in one of following years, the average value of whole year water flow of Morávka reservoir is missing. In this year, the average water flow for Šance reservoir was 2,910 m^3/s. Based on linear regression, try to determine the average water flow in Morávka reservoir.

{*Answer:* Use a suitable software package.}