Linear Models and Estimation by Least Squares

Regression

Radu T. Trîmbiţaş

April 5, 2016

1 Linear Models

Linear Statistical Models

Definition 1. A *linear statistical model* relating a random response Y to a set of independent variables $x_1, x_2, \ldots, x_k$ is of the form

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

where $\beta_0, \beta_1, \ldots, \beta_k$ are unknown parameters, ε is a random variable, and the variables $x_1, x_2, \ldots, x_k$ assume known values. We will assume that $E(\varepsilon) = 0$ and hence

$$M(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k.$$

If $k = 1$ we call the model *simple*.

Physical interpretation: Y is equal to an expected value, $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$ (a function of the independent variables $x_1, x_2, \ldots, x_k$), plus a random error ε. From a practical point of view, ε acknowledges our inability to provide an exact model for nature. In repeated experimentation, Y varies about $E(Y)$ in a random manner because we have failed to include in our model all of the many variables that may affect Y. Fortunately, many times the net effect of these unmeasured, and most often unknown, variables is to cause Y to vary in a manner that may be adequately approximated by an assumption of random behavior.

2 The Method of Least Squares

The Method of Least Squares

- simple regression

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

  ε is a RV such that $E(\varepsilon) = 0$.

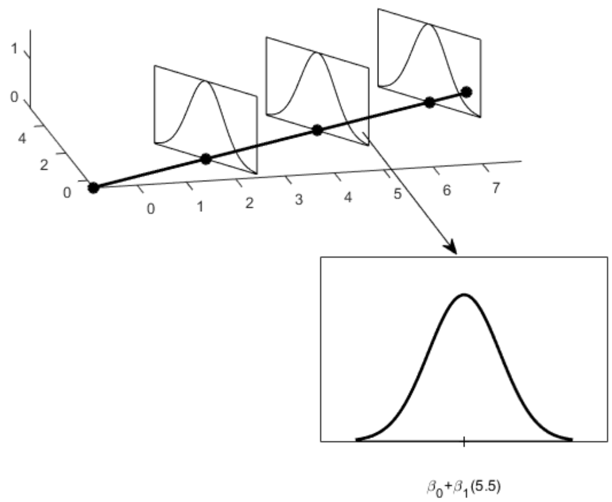

Figure 1: A linear statistical model

- if $\hat{\beta}_0$, $\hat{\beta}_1$ estimators for β_0 and β_1 then $\hat{Y} = \widehat{\beta}_0 + \hat{\beta}_1 x$ estimator for $M(Y)$.
- Prediction for Y when $x = x_i$

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

- The deviation of the observed value of y_i from $\widehat{y_i} = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$ called error

$$error = y_i - \hat{y};$$

- We'll find βs which minimize *sum of squares for error*

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y})^2 = \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2.$$

- Solve

$$\frac{\partial SSE}{\partial \hat{\beta}_0} = 0 \text{ and } \frac{\partial SSE}{\partial \hat{\beta}_1} = 0;$$

- Normal equations

$$\begin{aligned}\frac{\partial SSE}{\partial \hat{\beta}_0} &= -\sum_{i=1}^{n} 2[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i] \\ &= -2\left(\sum y_i - n\beta_0 + \hat{\beta_1}\sum x_i\right) = 0 \\ \frac{\partial SSE}{\partial \hat{\beta}_1} &= -\sum 2[y_i - (\hat{\beta}_0 + \beta_1 x_i)]x_i \\ &= -2\left(\sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2\right) = 0\end{aligned}$$

- Solutions

$$\begin{aligned}&\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2} \\ &\beta_0 = \overline{y} - \hat{\beta}_1 \overline{x}.\end{aligned}$$

  The Hessian matrix is positive definite

- Introducing

$$S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \text{ and } S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2$$

  the solution is

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \quad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}.$$

Example

Example 2. Use the method of least squares to fit a straight line to the n = 5 data points given below

x	-2	-1	0	1	2
y	0	0	1	1	3

Find the value for $x = 3$.

See the file `ex11_1WMS.pdf`.

3 Properties of the Least-Squares Estimators

Properties of the Least-Squares Estimators

- Model

$$Y = \beta_0 + \beta_1 x + \varepsilon,$$

- Assumptions ε is a RV such that $E(\varepsilon) = 0$, $V(\varepsilon) = \sigma^2$ (independent of x). Notice that $V(Y) = V(\varepsilon) = \sigma^2$.

Properties of the Least-Squares Estimators

Theorem 3. *1. $\hat{\beta}_0$ and $\hat{\beta}_1$ unbiased estimators i.e.*

$$E(\hat{\beta}_i) = \beta_i, \quad i = 0, 1.$$

2. $V(\hat{\beta}_0) = c_{00}\sigma^2$ *where* $c_{00} = \dfrac{\sum x_i^2}{nS_{xx}}$.

3. $V(\hat{\beta}_1) = c_{11}\sigma^2$, *where* $c_{11} = \dfrac{1}{S_{xx}}$.

4. $Cov(\hat{\beta}_0, \hat{\beta}_1) = c_{01}\sigma^2$, *where* $c_{01} = \dfrac{-\overline{x}}{S_{xx}}$.

5. $S^2 = SSE/(n-2)$, *where* $SSE = S_{yy} - \hat{\beta}_1 S_{xy}$ *and* $S_{yy} = \sum(y_i - \overline{y})^2$, *is an unbiased estimator for* σ^2.

Properties of the Least-Squares Estimators

Theorem 4. *6. Moreover, if individual errors ε_i are normally distributed then*

a) *$\hat{\beta}_0$ and $\hat{\beta}_1$ are normally distributed*

b) *the RV $\dfrac{(n-2)S^2}{\sigma^2}$ has a χ^2 with $n-2$ dfs.*

c) *Statistic S^2 is independent of both $\hat{\beta}_0$ and $\hat{\beta}_1$.*

Proof

Assume that n independent observations are to be made on this model so that before sampling we have n independent random variables of the form

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

But,

$$\begin{aligned}
\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} &= \frac{\sum(x_i - \overline{x})(Y_i - \overline{Y})}{\sum(x_i - \overline{x})^2} \\
&= \frac{\sum(x_i - \overline{x})Y_i - \overline{Y}\overbrace{\sum(x_i - \overline{x})}^{0}}{S_{xx}} \\
&= \frac{\sum(x_i - \overline{x})Y_i}{S_{xx}}
\end{aligned}$$

and

$$\begin{aligned}
E(\hat{\beta}_1) &= \frac{\sum(x_i - x)M(Y_i)}{S_{xx}} = \frac{\sum(x_i - \overline{x})(\beta_0 + \beta_1 x_i)}{S_{xx}} \\
&= \beta_0 \frac{\overbrace{\sum(x_i - \overline{x})}^{0}}{S_{xx}} + \beta_1 \frac{\sum(x_i - \overline{x})x}{S_{xx}} \\
&= \beta_1 \frac{\sum(x_i - \overline{x})^2}{S_{xx}} = \beta_1,
\end{aligned}$$

that is $\hat{\beta}_1$ is an unbiased estimator of β_1. Variance of $\hat{\beta}_1$:

$$\begin{aligned}
V(\hat{\beta}_1) &= \left[\frac{1}{S_{xx}}\right]^2 \sum V[(x_i - \overline{x})Y_i] \\
&= \left[\frac{1}{S_{xx}}\right]^2 \sum(x_i - \overline{x})V(Y_i) = \frac{\sigma^2}{S_{xx}}
\end{aligned}$$

The expected value and variance of $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{x}$

$$V(\hat{\beta}_0) = V(\overline{Y}) + \overline{x}^2 V(\beta_1) - 2xCov(\overline{Y}, \beta_1)$$

We need $V(\overline{Y})$ and $Cov(\overline{Y}, \hat{\beta}_1)$ to obtain $V(\hat{\beta}_0)$.Since $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, we see that

$$\overline{Y} = \frac{1}{n}\sum Y_i = \beta_0 + \beta_1 \overline{x} + \overline{\varepsilon}$$

Thus,

$$E(\overline{Y}) = \beta_0 + \beta_1 \overline{x} + M(\overline{\varepsilon}) = \beta_0 + \beta_1 \overline{x}$$

and

$$V(\overline{Y}) = V(\varepsilon) = \frac{1}{n} V(\varepsilon_i) = \frac{\sigma^2}{n}$$

To find $Cov(\overline{Y}, \hat{\beta}_1)$, rewrite the expression of $\hat{\beta}_1$ as

$$\hat{\beta}_1 = \sum c_i y_i$$

where

$$c_i = \frac{x_i - \overline{x}}{S_{xx}}.$$

(Notice that $\sum c_i = 0$.) Then,

$$Cov(\overline{Y}, \hat{\beta}_1) = Cov\left[\sum\left(\frac{1}{n}\right) Y_i, \sum c_i Y_i\right]$$

$$= \sum\left(\frac{c_i}{n}\right) V(Y_i) + \sum_{i<j}\sum\left(\frac{c_j}{n}\right) Cov(Y_i, Y_j).$$

Because Y_i and Y_j, where $i \neq j$, are independent, $Cov(Y_i, Y_j) = 0$. Also, $V(Y_i) = \sigma^2$, and hence

$$Cov(\overline{Y}, \beta_1) = \frac{\sigma^2}{n} \sum c_i = 0.$$

Returning to our original task of finding the expected value and variance of

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{x}$$

from mean value properties

$$E(\hat{\beta}_0) = E(\overline{Y}) - E(\hat{\beta}_1)\overline{x} = \beta_0 + \beta_1 \overline{x} - \beta_1 \overline{x} = \beta_0.$$

Since $V(\overline{Y})$, $V(\hat{\beta}_1)$, and $Cov(\overline{Y}, \hat{\beta}_1)$ were already derived

$$\begin{aligned} V(\hat{\beta}_0) &= V(\overline{Y}) + x^2 V(\hat{\beta}_1) - 2x Cov(\overline{Y}, \beta_1) \\ &= \frac{\sigma^2}{n} + \overline{x}^2 \left[\frac{\sigma^2}{S_{xx}} \right] = \sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}} \right] = \frac{\sigma^2 \sum x_i^2}{n S_{xx}}. \end{aligned}$$

Further

$$Cov(\hat{\beta}_0, \beta_1) = Cov(\overline{Y} - \hat{\beta}_1 x, \beta_1) = \underbrace{Cov(\overline{Y}, \beta_1)}_{0} - \overline{x} Cov(\beta_1, \beta_1)$$

$$= -\overline{x} V(\beta_1) = \frac{-\overline{x}\sigma^2}{S_{xx}}$$

So, $\hat{\beta}_0$ and $\hat{\beta}_1$ are correlated (and therefore dependent), unless $\overline{x} = 0$. The variances of estimators depends on unknown quantity $\sigma^2 = V(\varepsilon)$. Will show that

$$S^2 = \frac{1}{n-2} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \frac{1}{n-2} SSE$$

is an unbiased estimator of σ^2. Notice that the 2 occurring in the denominator of S^2 corresponds to the number of β parameters estimated in the model.

$$\begin{aligned} E(SSE) &= E\left[\sum (Y_i - \hat{Y}_i)^2 \right] = E\left[\sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \right] \\ &= E\left[\sum (Y_i - \overline{Y} + \hat{\beta}_1 \overline{x} - \hat{\beta}_1 x_i)^2 \right] \\ &= E\left[\sum [(Y_i - \overline{Y}) - \hat{\beta}_1 (x_i - \overline{x})]^2 \right] \\ &= E\left[\sum (Y_i - \overline{Y})^2 + \hat{\beta}_1^2 \sum (x_i - \overline{x}) - 2\hat{\beta}_1 \sum (x_i - \overline{x})(Y_i - \overline{Y}) \right] \end{aligned}$$

Because $\sum (x_i - \overline{x})(Y_i - \overline{Y}) = \sum (x_i - \overline{x})^2 \hat{\beta}_1$, the last two terms in the expectation combine to give $-\hat{\beta}_1^2 \sum (x_i - \overline{x})^2$. Also,

$$\sum (Y_i - \overline{Y})^2 = \sum Y_i^2 - n\overline{Y}^2,$$

and therefore

$$E\left[\sum(Y_i-\hat{Y}_i)^2\right] = E\left[\sum Y_i^2 - n\overline{Y}^2 - \hat{\beta}_1^2 S_{xx}\right]$$
$$= \sum E(Y_i^2) - nE(\overline{Y})^2 - S_{xx}E(\hat{\beta}_1^2).$$

Noting that, for any random variable U, $E(U^2) = V(U) + [E(U)]^2$, we see that

$$E\left[\sum(Y_i-\hat{Y}_i)^2\right] = \sum\{V(Y_i)+[E(Y_i)]^2\} - n\{V(\overline{Y})+[E(\overline{Y})]^2\}$$
$$- S_{xx}\{V(\hat{\beta}_1)+[E(\beta_1)]^2\}$$
$$= n\sigma^2 + \sum(\beta_0+\beta_1 x_i)^2 - n\left[\frac{\sigma^2}{n} + (\beta_0+\beta_1\overline{x})^2\right]$$
$$- S_{xx}\left[\frac{\sigma^2}{S_{xx}} + \beta_1^2\right]$$

This expression simplifies to $(n-2)\sigma^2$. Thus, we find that an unbiased estimator of σ^2 is given by

$$S^2 = \left(\frac{1}{n-2}\right)\sum(Y_i-\hat{Y}_i)^2 = \frac{1}{n-2}SSE$$

A simple way to compute SSE is given by

$$SSE = \sum(y_i-\overline{y})^2 - \hat{\beta}_1\sum(x_i-\overline{x})(y_i-\overline{y}) = S_{yy} - \overline{\beta}_1 S_{xy},$$

where $S_{yy} = \sum_{i=1}^{n}(y_i-\overline{y})^2$. Thus far, the only assumptions that we have made about the error term ε in the model $Y = \beta_0+\beta_1 x+\varepsilon$ were $E(\varepsilon)=0$ and $V(\varepsilon) = \sigma^2$ (independent of x). It is natural to assume $\varepsilon \in N(0,\sigma^2)$. It follows that Y_i is normally distributed with with mean $\beta_0+\beta_1 x_2$ and variance σ^2.Because both $\hat{\beta}_0$ and $\hat{\beta}_1$ are *linear functions* of $Y_1, Y_2, \ldots, Y_n$, the estimators are normally distributed , with means and variances as previously derived. Further, if the assumption of normality is warranted, it follows that

$$\frac{(n-2)S^2}{\sigma^2} = \frac{SSE}{\sigma^2}$$

has a χ^2 distribution with $n-2$ dfs.

4 Inferences Concerning the Parameters β_i

Inferences concerning the parameters β_i

- If ε is normally distributed $\widehat{\beta}_i$, $i = 0, 1$ are normal and unbiased estimators of β_i, $i = 0, 1$.

$$V(\hat{\beta}_0) = c_{00}\sigma^2, \text{ where } c_{00} = \frac{\sum x_i^2}{nS_{xx}} \tag{1}$$

$$V(\hat{\beta}_1) = c_{11}\sigma^2, \text{ where } c_{11} = \frac{1}{S_{xx}} \tag{2}$$

- To test $H_0: \beta_i = \beta_{i0}$ (β_{i0} given) use

$$Z = \frac{\hat{\beta}_i - \beta_{i0}}{\sigma\sqrt{c_{ii}}}$$

with c_{ii} given by (1) and (2)

- σ or a good estimation ($n \geq 30$) is not available, we estimate σ by

$$S = \sqrt{\frac{SSE}{n-2}}$$

- The statistic

$$T = \frac{\hat{\beta}_i - \beta_{i0}}{S\sqrt{c_{ii}}} \tag{3}$$

has a Student distribution with $n - 2$ dfs.

- We can test hypotheses on $\hat{\beta}_i$ or to derive CIs based on T given by (3)

- $H_0: \beta_i = \beta_{i0}$

- H_a : $\beta_i > \beta_{i0}$ upper-tail test
  $\beta_i < \beta_{i0}$ lower-tail test
  $\beta_i \neq \beta_{i0}$ two-tailed test

- Test statistic

$$T = \frac{\hat{\beta}_i - \beta_{i0}}{S\sqrt{c_{ii}}}$$

- Rejection region

$$t > t_\alpha$$

$$t < -t_\alpha$$

$$|t| > t_{\alpha/2}$$

- $n - 2$ dfs. $1 - \alpha$ CIs for β_i

$$\beta_i = \hat{\beta}_i \pm t_{n-2,\frac{\alpha}{2}} S\sqrt{c_{ii}}$$

5 Inferences Concerning Linear Functions of the Model Parameters: Simple Linear Regression

Inferences Concerning Linear Functions of the Model Parameters

- Consider
$$\theta = a_0\beta_0 + a_1\beta_1,\ a_0, a_1 \in \mathbb{R}$$

-
$$\hat{\theta} = a_0\hat{\beta}_0 + a_1\hat{\beta}_1$$
is an unbiased estimator of θ.

- Its variance is
$$V(\hat{\theta}) = a_0^2 V(\hat{\beta}_0) + a_1^2 V(\hat{\beta}_1) + 2a_0a_1 Cov(\hat{\beta}_1, \hat{\beta}_1)$$
that using Theorem 3 yields
$$V(\hat{\theta}) = \frac{a_0^2 \frac{\sum x_i^2}{n} + a_1^2 - 2a_0a_1\overline{x}}{S_{xx}}\sigma^2. \tag{4}$$

- Since $\hat{\beta}_0$ and $\hat{\beta}_1$ are normally distributed, $\hat{\theta}$ is normally distributed and
$$Z = \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} \sim N(0, 1)$$

- $1 - \alpha$ CI for θ: $\hat{\theta} \pm z_{\alpha/2}\sigma_{\hat{\theta}}$.
- If σ^2 is not available replace it by $S^2 \longrightarrow$ a $T(n-2)$ distribution
- Let $\theta_0 = a_0\beta_0 + a_1\beta_1$ be a specified value of θ
- **Test for** $\theta = a_0\beta_0 + a_1\beta_1$

$$H_0 : \ \theta = \theta_0$$
$$H_a : \begin{cases} \theta > \theta_0 \\ \theta < \theta_0 \\ \theta \neq \theta_0 \end{cases}$$

- Test statistic
$$T = \frac{\hat{\theta} - \theta_0}{S\sqrt{\left(\frac{a_0^2 \frac{\sum x_i^2}{n} + a_1^2 - 2a_0a_1\overline{x}}{S_{xx}}\right)}} \sim T(n-2)$$

- Rejection region

$$\begin{cases} t > t_{n-2,1-\alpha} \\ t < t_{n-2,\alpha} \\ |t| > t_{\alpha/2} \end{cases}$$

- $1-\alpha$ CI for $\theta = a_0\beta_0 + a_1\beta_1$

$$\hat{\theta} \pm t_{n-2,\frac{\alpha}{2}} S \sqrt{\left(\frac{a_0^2 \frac{\sum x_i^2}{n} + a_1^2 - 2a_0 a_1 \overline{x}}{S_{xx}} \right)}$$

- To estimate $E(Y)$ for a given value $x = x^*$, i. e.

$$E(Y) = \beta_0 + \beta_1 x^*$$

  we chose in $a_0\beta_0 + a_1\beta_1$ $a_0 = 1$ and $a_1 = x^*$.

- Using (4) for variance, we obtain

$$\frac{a_0^2 \frac{\sum x_i^2}{n} + a_1^2 - 2a_0 a_1 \overline{x}}{S_{xx}} = \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}$$

- It results a $(1-\alpha)$-CI for $E(Y)$ when $x = x^*$

$$\hat{\beta}_0 + \hat{\beta}_1 x \pm t_{n-2,\frac{\alpha}{2}} S \sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}}$$

6 Predicting a Particular Value of Y by Using Simple Linear Regression

Predicting a Particular Value of Y by Using Simple Linear Regression

- We shall estimate $Y^* = \beta_0 + \beta_1 x + \varepsilon$ for $x = x^*$ by $\hat{Y}^* = \hat{\beta}_0 + \hat{\beta}_1 Y^*$.

$$err = Y^* - \hat{Y}^*$$

- Y^*, $\hat{Y}^*$ normally distributed, so *err* is normally distributed

$$\begin{aligned} E(err) &= E(Y^* - \hat{Y}^*) = E(Y^*) - E(\hat{Y}^*) \\ &= \beta_0 + \beta_1 x^* + E(\varepsilon) - \beta_0 - \beta_1 x^* = 0 \end{aligned}$$

- Also

$$V(err) = V(Y^* - \hat{Y}^*) = V(Y^*) - V(\hat{Y}^*) - 2Cov(Y^*, \hat{Y}^*)$$

- Since we predict a future value Y^* not involved in computation of $\hat{Y}^*$, Y^* and $\hat{Y}^*$ independent and $Cov(Y^*, \hat{Y}^*) = 0$. Then

$$\begin{aligned} V(err) &= V(Y^*) + V(\hat{Y}^*) = \sigma^2 + V(\hat{\beta}_0 + \hat{\beta}_1 x^*) \\ &= \sigma^2 + \left[\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}\right]\sigma^2 = \sigma^2\left[1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}\right] \end{aligned}$$

- Statistic

$$Z = \frac{Y^* - \hat{Y}^*}{\sigma\sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}}} \sim N(0, 1)$$

- Estimating σ by S the statistic

$$T = \frac{Y^* - \hat{Y}^*}{S\sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}}} \sim T(n-2)$$

- $1 - \alpha$ CI for Y^*.

$$P(t_{n-2,\frac{\alpha}{2}} < T < t_{n-2,1-\frac{\alpha}{2}}) = 1 - \alpha \Leftrightarrow$$

$$P\left[-t_{n-2,1-\frac{\alpha}{2}} < \frac{Y^* - \hat{Y}^*}{S\sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}}} < t_{n-2,1-\frac{\alpha}{2}}\right] = 1 - \alpha$$

Finally,

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{n-2,\frac{\alpha}{2}} S\sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}}$$

Example

Example 5. Suppose that the experiment that generated the data of Example 2 is to be run again with $x = 2$. Predict the particular value of Y with $1 - \alpha = .90$.

See `ex11_7WMS.pdf`

7 Correlation

Correlation

- Let (X, Y) be a random vector. We wish to test if X and Y are independent.

- If (X, Y) has a bivariate normal distribution, then independence $\Longleftrightarrow \rho = 0$

- Let $(X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n)$ be the selection vars. MLE for ρ is the sample correlation coefficient

$$r=\frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2\sum_{i=1}^{n}(Y_i-\overline{Y})^2}}$$

- We can rewrite it as

$$r=\frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}=\beta_1\sqrt{\frac{S_{xx}}{S_{yy}}}$$

r and $\hat{\beta}_1$ have the same sign.

- If (X,Y) has a bivariate normal distribution, then

$$E(Y|X=x)=\beta_0+\beta_1 x \text{ where } \beta_1=\frac{\sigma_y}{\sigma_x}\rho$$

- Testing $H_0:\ \rho=0$ with respect to alternative $H_1:\ \rho>0 \Longleftrightarrow H_0:\ \beta_1=0$ w.r.t. $H_1:\ \beta_1>0$ and analogous . We may use

$$T=\frac{\hat{\beta}_1-0}{\frac{S}{\sqrt{S_{xx}}}}\tilde{\ }T(n-2)$$

- We rewrite T as

$$T=\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

- The distribution of r is difficult to obtain, but
- $\frac{1}{2}\ln\frac{1+r}{1-r}$ is approximately normally distributed with mean $\frac{1}{2}\ln\frac{1+\rho}{1-\rho}$ and variance $\frac{1}{n-3}$.
- To test $H_0:\ \rho=\rho_0$ we may use a z-test with

$$Z=\frac{\frac{1}{2}\ln\frac{1+r}{1-r}-\frac{1}{2}\ln\frac{1+\rho_0}{1-\rho_0}}{\frac{1}{\sqrt{n-3}}}$$

- The statistic R^2 is called the *coefficient of determination* and has an interesting and useful interpretation.

$$R^2=\left(\frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}\right)^2=1-\frac{SSE}{S_{yy}}.$$

- Thus, R^2 can be interpreted as the proportion of the total variation in the y_i's that is explained by the variable x in a simple linear regression model.

Example

Example 6. The data given below represent a sample of mathematics achievement test scores and calculus grades for ten independently selected college freshmen. From this evidence, would you say that the achievement test scores and calculus grades are independent? Use $\alpha = .05$. Identify the corresponding attained significance level.

Student	1	2	3	4	5	6	7	8	9	10
Math	39	43	21	64	57	47	28	75	34	52
Final	65	78	52	82	92	89	73	98	56	75

See `ex11_8WMS.pdf`

8 Fitting the Linear Model by Using Matrices

Fitting the Linear Model by Using Matrices

- Suppose that we have the linear model

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$$

  and we make n independent observations, y_1, y_2, ..., y_n, on Y. We can write the observation y_i as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{ik} + \varepsilon_i$$

  where x_{ij} is the setting of the jth independent variable for the ith observation, $i = 1, 2, \ldots, n$.

- We define the matrices

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \; X = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix},$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \; \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_k \end{bmatrix}$$

- The n equations representing y_i as a function of the x's, β's, and ε's can be simultaneously written as

$$Y = X\beta + \varepsilon$$

- Suppose $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ IRV with $E(\varepsilon_i) = 0$ and $V(\varepsilon_i) = \sigma^2$. Then the least-squares estimators are given by

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

  provided that $(X^T X)^{-1}$ exists.

Example

Example 7. Fit a parabola to the data of Example 2, using the model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon.$$

Solution. $Y = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 3 \end{bmatrix}, X = \begin{bmatrix} 1 & -2 & 4 \\ 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \end{bmatrix}$. $\beta = (X^T X)^{-1} X^T y = \begin{bmatrix} 0.571\,43 \\ 0.7 \\ 0.214\,29 \end{bmatrix}$

$$\left(X^T X\right)^{-1} = \begin{bmatrix} 0.485\,71 & 0 & -0.142\,86 \\ 0 & 0.1 & 0 \\ -0.142\,86 & 0 & 7.142\,9 \times 10^{-2} \end{bmatrix}$$

A se vedea `ex11_13aWMS.pdf`

□

9 Properties of the Least-Squares Estimators: Multiple Linear Regression

Properties of the Least-Squares Estimators

This is a multivariate analogous of Theorem 3.

Theorem 8.

1. *$E(\hat{\beta}_i) = \beta_i$, $i = \overline{0,k}$*
2. *$D^2(\beta_i) = c_{ii}\sigma^2$, where c_{ij} are elements of $(X^T X)^{-1}$. (numbering starts from 0)*
3. *$Cov(\hat{\beta}_i, \hat{\beta}_j) = c_{ij}\sigma^2$.*
4. *An unbiased estimator of σ^2 is $S^2 = SSE/[n-(k+1)]$, where $SSE = Y^T Y - \hat{\beta}^T X^T Y$.*
   If** ε_i, $i = \overline{1,n}$ **are normally distributed
5. *$\hat{\beta}_i$, $i = \overline{0,k}$ is normally distributed.*
6. *The RV*
   $$\frac{[n-(k+1)]S^2}{\sigma^2}$$
   has a χ^2 distribution with $n-(k+1)$ dfs.
7. *Statistics S^2 and $\hat{\beta}_i$, $i = \overline{1,k}$ are independent.*

10 Inferences Concerning Linear Functions of the Model Parameters: Multiple Linear Regression

Inferences Concerning Linear Functions of the Model Parameters

- Suppose we wish to make inferences on linear function

$$a_0\beta_0 + a_1\beta_1 + a_2\beta_2 + \cdots + a_k\beta_k \tag{5}$$

  where $a_0, a_1, \ldots, a_k$ are real constants.

- If $a = [a_0\ a_1 \ldots a_k]^T$ we can rewrite (5) as

$$a^T\beta = a_0\beta_0 + \cdots + a_k\beta_k.$$

- $a^T\hat{\beta}$ is an unbiased estimator of $a^T\beta$ since

$$\begin{aligned} E(a^T\hat{\beta}) &= E(a_0\hat{\beta}_0 + \cdots + a_k\hat{\beta}_k) \\ &= a_0\beta_0 + \cdots + a_k\beta_k = a^T\beta. \end{aligned}$$

- For its variance we obtain

$$\begin{aligned} V(a^T\hat{\beta}) = V(a_0\hat{\beta}_0 + \cdots + a_k\hat{\beta}_k) &= a_0^2 V(\hat{\beta}_0) + \ldots \\ &+ a_k^2 V(\hat{\beta}_k) + 2a_0a_1 Cov(\hat{\beta}_0, \hat{\beta}_1) + \ldots \\ &\quad + 2a_1a_2 Cov(\hat{\beta}_1, \hat{\beta}_2) + \cdots + 2a_{k-1}a_k Cov(\hat{\beta}_{k-1}, \hat{\beta}_k) \end{aligned}$$

  where $V(\hat{\beta}_i) = c_{ii}\sigma^2$ şi $Cov(\hat{\beta}_i, \hat{\beta}_j) = c_{ij}\sigma^2$. It is easy to check that

$$V(a^T\hat{\beta}) = [a^T(X^TX)^{-1}a]\sigma^2.$$

- Since $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ are normally distributed, $a^T\hat{\beta}$ is normal with mean $a^T\beta$ and $V(a^T\hat{\beta}) = [a^T(X^TX)^{-1}a]\sigma^2$, and we conclude

$$Z = \frac{a^T\hat{\beta} - a^T\beta}{\sqrt{D^2(a^T\beta)}} = \frac{a^T\hat{\beta} - a^T\beta}{\sigma\sqrt{a^T(X^TX)^{-1}a}} \sim N(0,1).$$

- We could use it to test

$$H_0:\ a^T\beta = (a^T\beta)_0$$

  where $(a^T\beta)_0$ is a given value. The $1-\alpha$ CI for $a^T\beta$ is

$$a^T\beta \pm z_{\alpha/2}\sigma\sqrt{a^T(X^TX)^{-1}a}.$$

- If we estimate σ by S, RV

$$T = \frac{a^T\hat{\beta} - a^T\beta}{S\sqrt{a^T(X^TX)^{-1}a}} \sim T[n-(k+1)]$$

- Test
$$H_0:\ a^T\beta = (a^T\beta)_0$$
$$H_1:\ \begin{cases} a^T\beta > (a^T\beta)_0 \\ a^T\beta < (a^T\beta)_0 \\ a^T\beta \neq (a^T\beta)_0 \end{cases}$$
Test statistic
$$T = \frac{a^T\hat{\beta} - (a^T\beta)}{S\sqrt{a^T(X^TX)^{-1}a}}$$
Rejection region
$$\begin{cases} t > t_{n-(k+1),\alpha} \\ t < -t_{n-(k+1),\alpha} \\ |t| > t_{n-(k+1),\frac{\alpha}{2}} \end{cases}$$

- $(1-\alpha)$ CI for $a^T\beta$ is given by
$$\alpha^T\hat{\beta} \pm t_{n-(k+1),\frac{\alpha}{2}} S\sqrt{a^T(X^TX)^{-1}a}.$$

- For inferences on individual parameters $\hat{\beta}_i$ we choose a with components
$$a_j = \begin{cases} 1, & \text{dacă} \quad j = i \\ 0, & \text{dacă} \quad j \neq i \end{cases}$$

11 Predicting a Particular Value of Y by Using Multiple Regression

Predicting a Particular Value of Y by Using Multiple Regression

- Consider the linear model
$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$
we wish to predict the value of Y^* for $x = x_1^*$, $x_2 = x_2^*, \ldots, x_k = x^*$; we use formula
$$\hat{Y}^* = \hat{\beta}_0 + \hat{\beta}_1 x_1^* + \cdots + \hat{\beta}_k x_k^* = a^T\hat{\beta}$$

- The error is
$$error = Y^* - \hat{Y}^*$$
It is normally distributed (Y^* and $\hat{Y}^*$ are normal) with
$$E(error) = 0 \text{ and } V(error) = \sigma^2[1 + a^T(x^TX)^{-1}a]$$

- RV
$$Z = \frac{Y^* - \hat{Y}^*}{\sigma\sqrt{1 + a^T(X^TX)^{-1}a}} \sim N(0,1)$$

- If σ is estimated by S

$$T = \frac{Y^* - \hat{Y}^*}{S\sqrt{1 + a^T(X^TX)^{-1}a}} \sim T\,[n - (k+1)]$$

- $1 - \alpha$ CI for Y

$$a^T\hat{\beta} \pm t_{n-(k+1),\frac{\alpha}{2}} S\sqrt{1 + a^T(X^TX)^{-1}a}$$

where $x_1 = x_1^*, x_2 = x_2^*, \ldots, x_k = x_k^*$ and $a^T = [1, x_1^*, x_2^*, \ldots, x_k^*]$.

Example

A response Y is a function of three independent variables x_1, x_2, and x_3 that are related as follows:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon.$$

(a) Fit this model to the $n = 7$ data points shown in the accompanying table.

y	1	0	0	1	2	3	3
x_1	−3	−2	−1	0	1	2	3
x_2	5	0	−3	−4	−3	0	5
x_3	−1	1	1	0	−1	−1	1

(b) Predict Y when $x_1 = 1$, $x_2 = -3$, $x_3 = -1$. Compare with the observed response in the original data. Why are these two not equal?

(c) Do the data present sufficient evidence to indicate that x_3 contributes information for the prediction of Y ? (Test the hypothesis $H_0 : \beta_3 = 0$, using $\alpha = .05$.)

(d) Find a 95% confidence interval for the expected value of Y, given $x_1 = 1$, $x_2 = -3$, and $x_3 = -1$.

(e) Find a 95% prediction interval for Y, given $x_1 = 1$, $x_2 = -3$, and $x_3 = -1$.

See `prob3lab.pdf`.

Coefficient of determination

- It is also useful in multiple regression
- Formula

$$R^2 = 1 - \frac{\sum(y_i - \widehat{y}_i)^2}{\sum(y_i - \overline{y})^2}$$

- The coefficient of determination is influenced by the number of regressors. For a given sample size n, the R^2 value will increase by adding more regressors into the linear model. The value of R^2 may therefore be high even if possibly irrelevant regressors are included.

- An adjusted coefficient of determination for p regressors and a constant intercept ($p+1$ parameters) is

$$R^2_{adj} = R^2 - \frac{p\left(1-R^2\right)}{n-p+1}.$$

12 Testing hypothesis $H_0 : \beta_{g+1} = \beta_{g+2} = \cdots = \beta_k = 0$

Model comparison

- Suppose,we wish to compare a *reduced model* of the form

$$\text{model } R : Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_g x_g + \varepsilon$$

to the linear model with all candidate independent variables present (the *complete model*):

$$\begin{aligned} \text{model } C : Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_g x_g + \\ \beta_{g+1} x_{g+1} + \cdots + \beta_k x_k + \varepsilon \end{aligned}$$

- $SSE_C < SSE_R$ (why?)
- null hypothesis

$$H_0 : \beta_{g+1} = \beta_{g+2} = \cdots = \beta_k = 0. \tag{6}$$

- $SSE_R - SSE_C$ is called the *sum of squares associated with the variables* x_{g+1}, $x_{g+2}, \ldots, x_k$, adjusted for the variables $x_1, x_2, \ldots, x_g$.
- Notice that

$$SSE_R = SSE_C + (SSE_R - SSE_C).$$

In other words, we have partitioned SSE_R into two parts: SSE_C and the difference $(SSE_R - SSE_C)$.

- If H_0 is true, then (proof left to the reader)

$$\begin{aligned} \chi^2_3 &= \frac{SSE_R}{\sigma^2} \sim \chi^2(n - [g+1]) \\ \chi^2_2 &= \frac{SSE_C}{\sigma^2} \sim \chi^2(n - [k+1]) \\ \chi^2_1 &= \frac{SSE_R - SSE_C}{\sigma^2} \sim (k-g). \end{aligned}$$

- Further, it can be shown that χ^2_2 and χ^2_1 are statistically independent.

- Consider the ratio

$$F = \frac{\frac{\chi_1^2}{k-g}}{\frac{\chi_2^2}{n-(k+1)}} = \frac{\frac{SSE_R - SSE_C}{k-g}}{\frac{SSE_C}{n-(k+1)}}.$$

  If $H_0 : \beta_{g+1} = \beta_{g+2} = \cdots = \beta_k = 0$ is true, then F possesses an F distribution with $\nu_1 = k - g$ numerator degrees of freedom and $\nu 2 = n - (k+1)$ denominator degrees of freedom.

- Large values of F favor rejection of H_0; rejection region

$$F > F_{\nu_1,\nu_2,\alpha}$$

Examples

Example 9. Do the data of Example 7 provide sufficient evidence to indicate that the second order model

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

contributes information for the prediction of Y? That is, test the hypothesis $H_0 : \beta_1 = \beta_2 = 0$ against the alternative hypothesis H_a : at least one of the parameters β_1, β_2, differs from 0. Use $\alpha = .05$. Give bounds for the attained significance level.

Solution. See ex11_18.R and `ex11_18.pdf`. □

Examples

It is desired to relate abrasion resistance of rubber (Y) to the amount of silica filler x_1' and the amount of coupling agent x_2'. Fine-particle silica fibers are added to rubber to increase strength and resistance to abrasion. The coupling agent chemically bonds the filler to the rubber polymer chains and thus increases the efficiency of the filler. The unit of measurement for x_1' and x_2' is parts per 100 parts of rubber, which is denoted phr. For computational simplicity, the actual amounts of silica filler and coupling agent are rescaled by the equations

$$x_1 = \frac{x_1' - 50}{6.7}, \qquad x_2 = \frac{x_2' - 4}{2}.$$

The data[1] are given in Table 1. Notice that five levels of both x_1 and x_2 are used, with the $(x_1 = 0, x_2 = 0)$ point repeated three times. Let us fit the second-order model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \varepsilon$$

[1]**Source**: Ronald Suich and G. C. Derringer, *Technometrics* 19(2) (1977): 214.

y	x_1	x_2
83	1	-1
113	1	1
92	-1	1
82	-1	-1
100	0	0
96	0	0
98	0	0
95	0	1.5
80	0	-1.5
100	1.5	0
92	-1.5	0

Table 1: Data for Example 11.19

to these data. This model represents a conic surface over the (x_1, x_2) plane. Fit the second-order model and test $H_0 : \beta_3 = \beta_4 = \beta_5 = 0$. (We are testing that the surface is actually a plane versus the alternative that it is a conic surface.) Give bounds for the attained significance level and indicate the proper conclusion if we choose $\alpha = .05$.

Solution. See file ex11_19WMS.R and `ex11_19WMS.pdf`

□

13 Statistical Models in R

Statistical Models in R

- The operaror ˜ is used to define a model formula in R.
- The form of an ordinary linear model is `response~op_1 term1 op2 term2 op_3 term_3`

`response` vector or matrix or expresion evaluated to vector or matrix defining the response variable(s)

`op_i` operator, either + or -, implying the inclusion or exclusion of a term in the model (the first is optional)

`term_i` is either

- a vector or matrix, or 1,
- a factor, or
- a formula expression consisting of factors, vectors or matrices connected by formula operators

- In all cases each term define a collection of columns to be added or removed from the model matrix. A 1 stands for an intercept column and it is by default included in the model matrix unless explicitely removed.
- The formula operators are similar in effects to the Wilkinson and Rogers notation [4], with . changed to :, since . is a valid name character in R.
- The notation is sumarized below [5]
  - `Y~M` `Y` is modeled as `M`
  - `M_1+M_2` Include `M_1` and `M_2`
  - `M_1-M_2` Include `M_1` and exclude `M_2`
  - `M_1:M_2` Tensor product of `M_1` and `M_2`
  - `M_1 %in% M_2` Similar to `M_1:M_2`, but with different coding
  - `M_1*M_2` or `M_1+M_2+M_1:M_2` or `M_1/M_2` or `M_1+M_2%in%M_1` these are equivalent, include `M_1` and `M_2` and product of `M_1` and `M_2`
  - `M^n` all terms in `M` together with interactions up to order `n`
  - `I(M)` identity, insulate `M`.
- to fit a linear model
  `fitted.model<-lm(formula, data=data.frame)`
- Generic functions for extracting model information
  - `anova(`*ob_1*`,`*obj_2*`)` compare a submodel with an outer model and produce an ANOVA table
  - `coef(`*obj*`)`extract the regression coefficient (matrix)
  - `deviance(`*obj*`)` residual sum of squares, weighted if appropriate
  - `formula(`*obj*`)`
  - `plot(`*obj*`)` produce four plots, showing residuals, fitted values and some diagnostics
  - `predict(`*obj*`,newdata=`*data.frame*`)`The data frame supplid must have variables specified with the same labels as the original. The value is a vector or matrix of predicted values corresponding to the determining variable values in *data.frame*.
  - `print(`*obj*`)` print a concise version of object
  - `residuals(`*obj*`)` extract the residuals, weighted as appropriate
  - `step(`*obj*`)` select a suitable model by adding or droping terms and preserving hierarchies. The model with the smallest AIC (Akaike's An Information Criterion) discovered in the stepwise search is returned
  - `summary(`*obj*`)`print a comprehensive summary of the regression analysis

14 Bibliography

References

References

[1] W. N. Venables, D. M. Smith and the R Development Core Team, *An Introduction to R*, 2015

[2] Peter Dalgaard, *Introductory Statistics with R*, 2nd ed., Springer Verlag, 2008.

[3] Norman Matloff, *THE ART OF R PROGRAMMING. A Tour of Statistical Software Design*, No Starch Press, San Francisco, 2011

[4] Wilkinson, G. N., and C. E. Rogers, Symbolic description of factorial models for analysis of variance. *J. Royal Statistics Society* 22, pp. 392–399, 1973

[5] Chambers, M., T. Hastie, *Statistical Models in S*, Chapman & Hall, 1992