Stat 20: Discussion at section

B. M. Bolstad, bolstad@stat.berkeley.edu

Nov 17, 2003

Inference for regression coefficients

Three types of sum of squares

SST: the total sum of squares (total variability)

$$SST = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2$$

SSE: the error sum of squares (unexplained variability)

$$SSE = \sum_{i=1}^{n} \left(Y_i - \hat{\mu}\left(\mathbf{x}_i\right)\right)^2$$

Note that the quantity $Y_i - \hat{\mu}\left(\mathbf{x}_i\right)$ is referred to as the residual.
SSM: the model sum of squares (explained variability)

$$SSM = \sum_{i=1}^{n} \left(\hat{\mu}\left(\mathbf{x}_i\right) - \bar{Y}\right)^2$$

These three sum of squares are related by $SST = SSM + SSE$. We commonly summarize these three sum of squares using an ANOVA table

Source	SS
Model	SSM
Error	SSE
Total	SST

There is a quantity R^2 called the squared multiple correlation. It is defined by

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

and gives the proportion of variation explained by the model. Note that

$$1 - R^2 = \frac{SSE}{SST}$$

is the proportion of variation unexplained by the model.

Estimating σ^2, $\text{SE}\left(\hat{\beta}_j\right)$

Our estimate of σ^2 for the regression model is given by

$$s^2 = \frac{SSE}{n-p-1}$$

and thus our estimate of σ is given by $s = \sqrt{s^2}$.

In general under the assumptions of the homoskedastic normal linear regression model $\hat{\beta}_j$ is an unbiased estimate of β_j and it is normally distributed with mean β_j and variance $\text{Var}\left(\hat{\beta}_j\right) = c_j\sigma^2$. Note that c_j is a constant depending on the model and the data. We would estimate the standard deviation of $\hat{\beta}_j$ using $\text{SE}\left(\hat{\beta}_j\right) = c_j s$. In general a computer is used to estimate the constants c_j.

If the basis is orthogonal then it can be shown that

$$\text{Var}\left(\hat{\beta}_j\right) = \frac{\sigma^2}{\sum_{i=1}^n g_j^2\left(\mathbf{x}_i\right)} \text{ for } j = 0, \ldots, p$$

and consequently we can estimate the standard deviation of $\hat{\beta}_j$ using

$$\text{SE}\left(\hat{\beta}_j\right) = \frac{s}{\sqrt{\sum_{i=1}^n g_j^2\left(\mathbf{x}_i\right)}} \text{ for } j = 0, \ldots, p$$

Note that random variable

$$\frac{\hat{\beta}_j - \beta_j}{\text{SE}\left(\hat{\beta}_j\right)}$$

has the t distribution with $n-p-1$ degrees of freedom.

Confidence Intervals and Hypothesis tests for β_j

A 100C% confidence interval for β_j is given by

$$\hat{\beta}_j \pm t^* \text{SE}\left(\hat{\beta}_j\right)$$

where t^* is given by the value where $P\left(-t^* < T < t^*\right) = C$ and T had the t distribution with $n-p-1$ degrees of freedom.

We use the test statistic

$$\frac{\hat{\beta}_j}{\text{SE}\left(\hat{\beta}_j\right)}$$

to test hypotheses about β_j.

We typically carry out one of the following hypothesis tests about β_j

$$H_0 : \beta_j = 0 \text{ vs } H_A : \beta_j \neq 0$$

$$H_0 : \beta_j \leq 0 \text{ vs } H_A : \beta_j > 0$$

$$H_0 : \beta_j \geq 0 \text{ vs } H_A : \beta_j < 0$$

Note that you use the t distribution with $n - p - 1$ degrees of freedom to look up your P-value.

Problem 1

Consider the following experimental data

y	x_1	x_2
21	-1	-1
14	0	-1
13	1	-1
23	-1	0
20	0	0
11	1	0
31	-1	1
20	0	1
27	1	1

We have previously shown that 1, x_1 and x_2 are an orthogonal basis for this dataset and when we previously fitted the regression model $\mu((x)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ our estimated model was $\hat{\mu}((x)) = 20 - 4x_1 + 5x_2$.

(a) Determine SST, SSE, the ANOVA table, R^2, s^2 and s.

(b) Compute SE $\left(\hat{\beta}_1\right)$ and SE $\left(\hat{\beta}_1\right)$?

(c) Compute 95% confidence intervals for β_1 and β_2.

(d) Carry out and interpret the hypothesis test

$$H_0 : \beta_1 = 0 \text{ vs } H_A : \beta_1 \neq 0$$

Make sure to compute the P-value then interpret the results in the context of the regression model.

Carry out and interpret the hypothesis test

$$H_0 : \beta_j = 0 \text{ vs } H_A : \beta_2 \neq 0$$

Make sure to compute the P-value then interpret the results in the context of the regression model.

Answer:

(a) The sum of squares total (SST) is given by

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

Using the data we can compute $\bar{y} = 20$ and $\sum_{i=1}^{9} y_i^2 = 3946$. Therefore $SST = 3946 - 9 * 20^2 = 346$.

The error sum of squares (SSE) is given by

$$SSE = \sum_{i=1}^{n} \left(Y_i - \hat{\mu}\left(\mathbf{x}_i\right)\right)^2$$

to compute it, lets augment our original data matrix with some additional columns.

y	x_1	x_2	$\hat{y} = 20 - 4x_1 + 5x_2$	$y - \hat{y}$
21	-1	-1	19	2
14	0	-1	15	-1
13	1	-1	11	2
23	-1	0	24	-1
20	0	0	20	0
11	1	0	16	-5
31	-1	1	29	2
20	0	1	25	-5
27	1	1	25	6

and so $SSE = 2^2 + (-1)^2 + \cdots + 6^2 = 100$. Finally we may compute $SSM = SST - SSE = 346 - 100 = 246$. Therefore the ANOVA table will be

Source	SS
Model	246
Error	100
Total	346

The proportion of variability explained by the regression model is $R^2 = 246/300 = 0.7109$. Our estimate of σ^2 is given by $s^2 = \frac{SSE}{n-p-1} = \frac{100}{9-2-1} = \frac{100}{6} = 16.6666$ and so the estimate of σ is given by $s = \sqrt{s^2} = 4.0825$.

(b) Since 1, x_1 and x_2 are an orthogonal basis we can estimate the standard error of $\hat{\beta}_1$ and $\hat{\beta}_2$ using the formula

$$\text{SE}\left(\hat{\beta}_j\right) = \frac{s}{\sqrt{\sum_{i=1}^{n} g_j^2\left(\mathbf{x}_i\right)}} \text{ for } j = 0, \ldots, p$$

In particular,

$$\text{SE}\left(\hat{\beta}_1\right) = \frac{s}{\sqrt{\sum_{i=1}^{n} x_{i1}^2}} = \frac{4.0825}{\sqrt{6}} = 1.67$$

and

$$\text{SE}\left(\hat{\beta}_2\right) = \frac{s}{\sqrt{\sum_{i=1}^{n} x_{i2}^2}} = \frac{4.0825}{\sqrt{6}} = 1.67$$

(c) The degrees of freedom are $9-2-1 = 6$ and so $t^* = 2.447$ therefore the 95% confidence interval for β_1 is given by

$$-4 \pm 2.447\,(1.67)$$

which corresponds to $(-8.078, 0.0781)$.

The 95% confidence interval for β_2 is given by

$$5 \pm 2.447\,(1.67)$$

which corresponds to $(0.921, 9.08)$.

(d) First we carry out the test:

$$H_0 : \beta_1 = 0 \text{ vs } H_A : \beta_1 \neq 0$$

the test statistic is given by

$$t = \frac{-4}{1.67} = -2.40$$

Since we are testing against the two sided alternative the P-value is given by $2P\,(T > |-2.40|) = 2P\,(T > 2.40)$. We will put bounds on the P-value. Going to the t-table we find that

$$1.943 < 2.40 < 2.447$$

converting to upper tail probability values

$$0.05 > P(T > 2.40) > 0.025$$

multiply through by 2 to get

$$0.1 > 2P(T > 2.40) > 0.05$$

and so we see that $0.05 <$ P-value < 0.1. Since the P-value is not very small we would not reject the null hypothesis. This means that it is possible that $\beta_1 = 0$ and therefore x_1 has no explanatory value for y.

Next we carry out the test:

$$H_0 : \beta_2 = 0 \text{ vs } H_A : \beta_2 \neq 0$$

the test statistic is given by

$$t = \frac{5}{1.67} = 3.00$$

Since we are testing against the two sided alternative the P-value is given by $2P\left(T > 3.00\right)$. We will put bounds on the P-value. Going to the t-table we find that

$$2.612 < 3.00 < 3.143$$

converting to upper tail probability values

$$0.02 > P(T > 3.00) > 0.01$$

multiply through by 2 to get

$$0.04 > 2P(T > 3.00) > 0.02$$

and so we see that $0.02 <$ P-value < 0.04. Since the P-value small we would reject the null hypothesis. This means that $\beta_2 \neq 0$ and therefore x_2 does have explanatory value for y.