Stat 20: Discussion at section

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

Dec 1, 2003

Problem 1

Consider data where we have a dependent variable y and potential explanatory variables x_1,x_2,x_3,x_4 and x_5. We have 209 data values for which the summary statistics are

Variable	Sample Mean	Sample SD
y	201.287	33.890
x_1	4.488	2.275
x_2	7.804	4.605
x_3	0.517	0.501
x_4	0.033	0.180
x_5	0.344	0.476

In addition the correlation matrix is

	y	x_1	x_2	x_3	x_4	x_5
y	1.000	- 0.243	0.312	0.579	-0.078	-0.005
x_1	- 0.243	1.000	-0.061	0.031	-0.017	0.088
x_2	0.312	-0.061	1.000	0.094	-0.073	-0.144
x_3	0.579	0.031	0.094	1.000	0.180	-0.004
x_4	-0.078	-0.017	-0.073	0.180	1.000	0.033
x_5	-0.005	0.088	-0.144	-0.004	0.033	1.000

(a) Pick the variable that is most explanatory for y.

Answer: we choose the variable which has the highest $|r|$ value with y. In this case that is x_3 where r is 0.579.

(b) Determine R^2 when you use the variable chosen in (a) to predict y.

Answer:

In the case of regression with a single variable $R^2 = r^2$ and so $R^2 = r^2 = 0.579^2 = 0.335$.

(c) Determine SST, SSE and the ANOVA table

Answer:

Since $SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$ and $s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}}$ it follows that $SST = (n-1)s_y^2$ and so $SST = 208\,(33.890)^2 = 238894.7$.

Now $R^2 = \frac{SSM}{SST}$ so $SSTR^2 = SSM$ which means $SSM = 0.335\,(238894.7) = 80029.7$. Similarly $1 - R^2 = \frac{SSE}{SST}$ and so $SSE = (1-R^2)\,SST = 158865.0$. Therefore the ANOVA table is

Source	SS
Model	80029.7
Error	158865.0
Total	238894.7

(d) Estimate σ^2 and σ.

Answer: we use s^2 to estimate σ^2.

$$s^2 = \frac{SSE}{n-2} = \frac{158865.0}{207} = 767.46$$

and so

$$s = \sqrt{s^2} = 27.70$$

(e) Determine the least squares fit for $\beta_0 + \beta_1 x_3$

Answer:

$$\hat{\beta}_1 = r_{x_3 y}\frac{s_y}{s_x} = 0.579\frac{33.890}{0.501} = 39.16$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_3 = 201.287 - 39.16\,(0.517) = 181.0413$$

and so the fitted model is $181.0413 + 39.16x$.

(f) Determine a 95% confidence interval for the slope. (Hint: $\text{Var}\left(\hat{\beta}_1\right) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$ for the simple linear regression model $\beta_0 + \beta_1 x$).

Answer:

The SE is an estimate of the SD and is thus given by

$$\begin{aligned}\text{SE}\left(\hat{\beta}_1\right) &= \frac{s}{\sqrt{\sum_{i=1}^{n}(x_{i3} - \bar{x}_3)^2}} \\ &= \frac{s}{\sqrt{n-1 s_{x_3}}} \\ &= \frac{27.703}{\sqrt{2080.501}} \\ &= 3.83\end{aligned}$$

and so the 95% confidence interval is given by

$$39.16 \pm 1.984(3.83)$$

which is $(31.56, 46.76)$.

Problem 2

Consider the same data as in Problem 1 and consider the multiple linear regression model:

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$$

A computer is used to fit the model and the following coefficient table is produced:

Term	Coef	SE	t	P-value
1	183.570	5.221	35.162	0.000
x_1	-3.808	0.748	-5.089	0.000
x_2	1.741	0.375	4.644	0.000
x_3	40.325	3.456	11.667	0.000
x_4	-32.716	9.581	-3.415	0.001
x_5	4.279	3.602	1.188	0.236

with $R^2 = 0.494$.

(a) Determine the ANOVA table.

Answer:

From question 1 we find that $SST = 238894.7$. And so

$$SSE = \left(1 - R^2\right) SST = 120880.7$$

and

$$SSM = R^2 SST = 118014.0$$

Therefore the ANOVA table is

Source	SS
Model	118014.0
Error	120880.7
Total	238894.7

(b) What do you conclude from the coefficient table about x_5 and its predictive value for y?

Answer:

From the coefficient table $\hat{\beta}_5 = 4279$ is the change in y if we increase x_5 by one unit holding all the other variables fixed. For the test

$$H_0 : \beta_5 = 0 \text{ vs } H_A : \beta_5 \neq 0$$

the corresponding t statistic is 1.188 and the P-value is 0.236. SO we cannot reject the null hypothesis. So we would would conclude that x_5 is not explanatory for y and could be removed from the model.

(c) What effect do we have on y if we change $x_3 = 0$ and $x_4 = 0$ to $x_3 = 1$ and $x_4 = 1$?

Answer:

$$\hat{y}_{\text{init}} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_5 x_5$$

$$\hat{y}_{\text{final}} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 + \hat{\beta}_4 + \hat{\beta}_5 x_5$$

and so

$$\Delta y = \hat{y}_{\text{final}} - \hat{y}_{\text{init}} = \hat{\beta}_3 + \hat{\beta}_4 = 40.325 - 32.716 = 7.609$$

Problem 3

Consider the following regression coefficient table

Term	Coef	SE	t	P-value
1	2573.050	284.502	9.044	0.000
Age	-3.650	9.618	-0.379	0.705
Lwt	4.354	1.735	2.509	0.013
White	357.051	114.729	3.112	0.002
Black	-132.390	159.360	-0.831	0.407
Smoke	-350.618	106.454	-3.294	0.001
Ptl	-48.839	101.950	-0.479	0.632
Ht	-5.92.812	202.279	-2.931	0.004
Ui	-514.928	138.857	-3.708	0.000
Ftv	-14.072	46.458	-0.303	0.762

(a) Based on the above regression model. Which explanatory variable should we drop (if any from the model)?

Answer: We use the P-value column to judge which variables, if any to remove from the model. The higher the P-value the more likely it is that we fail to reject the null hypothesis. Since “Ftv” has the highest P-value we should remove it first.

(b) Why is it better to drop variables from the model one at a time rather than all at once?

Answer:

Because if some explanatory variables are nearly linear combinations of other explanatory variables, then it is possible that the P-value could change substantially when we remove another variable. ie removing one variable could cause a variable that previously had a high P-value to now have low P-value.

(c) Suppose that the correlation is small between any pair of explanatory variables. Which terms would you then drop from the model?

Answer:

If the correlations are all small then it is not likely that the explanatory variables are linear combinations of each other. This means that when we remove variables from the model, the coefficients and corresponding P-values would not change much. So we should remove "Ftv", "Age", "Ptl" and "Black" since they have high P-values and keep the remaining parameters because they have small P-values.