Stat 20 Fall 2003, Quiz 4 Answers

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

Dec 5, 2003

Question 1. (12 points)

A technician is interested in tuning a particular machine in a factory. He wants to tune the machine so that it is producing the smallest possible y values. To control the machine he can use a dial which sets the x value. He performs an experiment by setting the dial at different settings of x and then recording the y value. Consider the following data as the results of his experiment

y	x
2.80	-2
1.46	-1
1.99	0
8.75	1
17.47	2
3.25	-2
1.35	-1
2.73	0
7.20	1
16.40	2

You may assume that 1, x, $x^2 - 2$ is an orthogonal basis. The technician believes that the regression model $\mu(x) = \beta_0 + \beta_1 x + \beta_2(x^2 - 2)$ is appropriate for the this data. The coefficient estimates are $\hat{\beta}_0 = 6.34$, $\hat{\beta}_1 = 3.439$ and $\hat{\beta}_2 = 1.844$.

(a) Compute the ANOVA table and state R^2.

Answer:

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

From the data

$$\sum_{i=1}^{n} y_i = 2.80 + 1.46 + 1.99 + 8.75 + 17.47 + 3.25 + 1.35 + 2.73 + 7.20 + 16.40 = 63.4$$

which implies $\bar{y} = 63.4/10 = 6.34$ and

$$\sum_{i=1}^{n} y_i = 2.80^2 + 1.46^2 + 1.99^2 + 8.75^2 + 17.47^2 + 3.25^2 + 1.35^2 + 2.73^2 + 7.20^2 + 16.40^2 = 736.33$$

Therefore $SST = 736.3 - 10(6.34)^2 = 334.38$.

To work out SSE, augment the data matrix

y	x	$\hat{y} = 6.34 + 3.439x + 1.844(x^2 - 2)$	$y - \hat{y}$
2.80	-2	3.15	-0.35
1.46	-1	1.06	0.40
1.99	0	2.65	-0.66
8.75	1	7.94	0.82
17.47	2	16.91	0.56
3.25	-2	3.15	0.10
1.35	-1	1.06	0.29
2.73	0	2.65	0.08
7.20	1	7.94	-0.74
16.40	2	16.91	-0.51

and since

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = -0.35^2 + \cdots + -0.51^2 = 2.60$$

Since $SSE + SSM = SST$ we find that $SSM = 334.38 - 2.60 = 331.78$. The ANOVA table is given by

Source	SS
Model	331.78
Error	2.60
Total	334.38

and so $R^2 = \frac{SSM}{SST} = \frac{331.78}{334.38} = 0.992$

(b) Compute the standard error estimates for all three regression parameters.

Answer: Since the basis is orthogonal we may use the formula

$$\mathrm{SE}\left(\hat{\beta}_j\right) = \frac{s}{\sqrt{\sum_{i=1}^{n} g_j^2\left(\mathbf{x}_i\right)}}$$

First we compute

$$s = \sqrt{\frac{SSE}{n - p - 1}} = \sqrt{\frac{2.60}{10 - 2 - 1}} = 0.6094$$

and

$$\sum_{i=1}^{10} 1^2 = 10$$

$$\sum_{i=1}^{10} x^2 = 20$$

$$\sum_{i=1}^{10}(x^2-2)^2 = 28$$

Therefore

$$\text{SE}\left(\hat{\beta}_0\right) = \frac{0.609}{\sqrt{10}} = 0.19$$

$$\text{SE}\left(\hat{\beta}_1\right) = \frac{0.609}{\sqrt{20}} = 0.14$$

$$\text{SE}\left(\hat{\beta}_2\right) = \frac{0.609}{\sqrt{28}} = 0.12$$

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

Answer: The confidence intervals are of the form

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

where the $df = n - p - 1$. Therefore the 95% confidence interval for β_1 is given by

$$3.439 \pm 2.365(0.14)$$

which is $(3.11, 3.77)$. The 95% confidence interval for β_2 is given by

$$1.844 \pm 2.365(0.12)$$

which is $(1.56, 2.13)$.

Question 2. (13 points)

An automobile designer is interested in what makes a vehicle fuel efficient. She gathers data on 82 different models from a number of manufacturers. In particular the following variables were measured for each vehicle

1. **Vol** Cubic feet of passenger space
2. **HP** Engine horsepower
3. **SP** Top speed (mph)
4. **WT** Vehicle weight (100 lb)
5. **MPG** $\log_2$ average miles per gallon

She computes the following summary statistics for her dataset

Variable	Sample Mean	Sample Standard Deviation
Vol	98.80	22.16
HP	117.13	56.84
SP	112.41	14.04
WT	30.91	8.14
MPG	5.01	0.44

and the correlation matrix for this data is

	VOL	HP	Speed	Weight	MPG
VOL	1.0000	0.0765	-0.0431	0.3850	-0.3355
HP	0.0765	1.0000	0.9665	0.8322	-0.8570
Speed	-0.0431	0.9665	1.0000	0.6785	-0.7407
Weight	0.3850	0.8322	0.6785	1.0000	-0.9490
MPG	-0.3355	-0.8570	-0.7407	-0.9490	1.0000

(a) Which variable should she use if wants to predict MPG using a single variable? Explain why. What would R^2 be in this case?

Answer: The variable that should be used is Weight because it has the highest r^2 where r is the pairwise correlation between MPG and each of the other variables. For regression on a single variable $R^2 = r^2 = (-0.949)^2 = 0.90$. Note that if we choose any other variable R^2 would be smaller and thus since $1 - R^2 = \frac{SSE}{SST}$ other variables would have higher SSE (more unexplained variation).

(b) Complete the following coefficient table:

Answer: Note that the $df = 82 - 5 = 77$. To be conservative we will use the $df = 60$ line of the t table.

Term	Coef	Std. Error	t value	P-Value
1	8.3280	0.8147	*10.22*	P-value << .001
VOL	-0.0004	0.0008	*-0.5*	0.5 < P-value < 1.0
HP	0.0043	0.0028	*1.54*	0.1 < P-value < 0.2
Speed	-0.0188	0.0085	*-2.21*	0.02 < P-value < 0.04
Weight	-0.0534	0.0074	*-7.22*	P-value << 0.001

(c) Which if any terms would you remove from the model in part (b)? Make sure to explain the approach you would take.

Answer: We should remove VOL from the menu because it has a high P-value. Note that this P-value corresponds to the test $H_0 : \beta_1 = 0$ vs $H_A : \beta_1 \neq 0$. In this case with such a high P-value we would have no evidence to reject the null hypothesis, thus we can remove VOL from the model.

It would probably also be safe to remove HP from the model, this is because it has low correlation with VOL which implies that the parameter estimates we would get if we refitted the model without VOL would not differ significantly and would likely also have a high P-value. Note that in general to be safe we would refit the model after removing VOL and assess the P-values for the new model to see if there were any other terms that could safely be removed from the model.

Question 3. (OPTIONAL: worth up to 5 bonus points)

Outline the assumptions of the general homoskedastic normal linear regression model.

Explain the difference between regression for experimental data and for observational data.

Answer:

We assume that the random variable Y is normally distributed with mean $\mu(\mathbf{x})$ and standard deviation σ (which does not depend on $\mathbf{x}$ in the homoskedastic model). Here $\mathbf{x} = (x_1, \ldots, x_m)$ are factor variables. We further assume that this mean can be written in the form $\mu(\mathbf{x}) = \beta_0 g_0(\mathbf{x}) + \cdots + \beta_p g_p(\mathbf{x})$. Note that the $g_j(\mathbf{x})$ for $j = 0, \ldots, p$ are referred to as the basis functions. Sometimes we choose to write the response variable in the form

$$Y = \mu(\mathbf{x}) + \epsilon = \mu(\mathbf{x}) = \beta_0 g_0(\mathbf{x}) + \cdots + \beta_p g_p(\mathbf{x}) + \epsilon$$

where ϵ is assumed to be normally distributed with mean 0 and standard deviation σ. Furthermore we assume that the Y_i for $i = 1, \ldots, n$ are independent.

In the context of experimental data $\mathbf{x} = (x_1, \ldots, x_m)$ are fixed values controlled by the experimenter. This allows us to make causal interpretations about the effect changing an x_i variable has on Y. In the context of observational data we now have random variables $\mathbf{X} = (X_1, \ldots, X_m)$ which are not controlled, but merely observed. We may still fit regression models as above, but now we should be wary about making causal conclusions.