BIOS663 Midterm Spring 2013
Thursday, March 7, 2013

Instructions: Please be as rigorous as possible in all of your answers and show all your work.

Please sign the honor code pledge and submit it with your report. Violation of the honor code below will be prosecuted (penalties may include failure of the course and expulsion from the university).

Honor Code Pledge: On my honor, I have neither given nor received aid on this examination.

Name:

Signature:

Date:

1. *(20 points total)* MULTIPLE CHOICE QUESTIONS (Please circle the best answer).

   - *(5 points)* Which choice is not an appropriate description of $\hat{y}$ in a regression model?
     A. Estimated response
     B. Predicted response
     C. Estimated average response
     D. Observed response

     Solution: D

   - *(5 points)* Which of the following is the best way to determine whether or not there is a statistically significant linear relationship between two variables?
     A. Compute a regression line from a sample and see if the sample slope is 0.
     B. Compute the correlation coefficient and see if it is greater than 0.5 or less than 0.5.
     C. Conduct a test of the null hypothesis that the population slope is 0.
     D. Conduct a test of the null hypothesis that the population intercept is 0.

     Solution: C

   - *(5 points)* Which of the following case diagnostic measures is based on Y values only (and not X values)?
     A. Cooks distance
     B. Studentized residual
     C. Leverage
     D. None of the above

     Solution: D

   - *(5 points)* Which of the following is NOT true for the linear regression model $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i (i = 1, \cdots, 100)$ where all 5 model assumptions hold.
     A. $\hat{\beta_1} + \beta_2$ is a statistic
     B. $\hat{y}_i$ can be uniquely predicted from the above model
     C. β_1 may not be always estimable
     D. the residuals from the model are summed to 0

     Solution: A

2. *(40 points total)* You are working on a statistical consulting lab. One day, a client came with a gas consumption data. In this study, the client is interested in modeling the fuel efficiency of automobiles. A typical measure of fuel efficiency used by EPA and car manufactures is "gallons/100 miles". The client collected data on 100 cars. He measured two explanatory variables, x1=weight (in unit of 1000lb); and x2=number of cylinders. He also measured the fuel efficiency of each car (in "gallons/100 miles"). Let $\mathbf{X} = (\mathbf{J}_n, \mathbf{x}_1, \mathbf{x}_2)$ and the linear regression model considered is

$$y = \beta_0 + \beta_1 x1 + \beta_2 x2 + error.$$

Potentially helpful results:

$$(X'X)^{-1} = \begin{bmatrix} 0.308 & -0.06 & -0.017 \\ -0.06 & 0.025 & -0.004 \\ -0.017 & -0.004 & 0.006 \end{bmatrix} \text{ and } X'y = \begin{bmatrix} 405 \\ 1402 \\ 2350 \end{bmatrix}.$$

(a) *(8 points)* A partial ANOVA table for testing the association of the three covariates with the response y is given below. Complete the table.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	???	79	???	???	<0.001
Error	???	11	???		
Corrected Total	???	???			

Solution:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	79	39.5	349.6	<0.001
Error	97	11	0.113		
Corrected Total	99	90			

(b) *(6 points)* Fill in the cells with ??? in the following table.

Parameter	Estimate	Standard Error	t Value	Pr > |t|
(Intercept)	???	???	???	0.009
x1	???	???	???	<0.001
x2	???	???	???	<0.001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
(Intercept)	0.67	0.187	3.58	0.009
x1	1.35	0.053	25.47	<0.001
x2	1.61	0.026	61.81	<0.001

(c) *(6 points)* Test the following hypothesis: $H_0 : \beta_1 = 1$.

Solution:

$t - test = \frac{\hat{\beta}_1 - 1}{SE(\hat{\beta}_1)} = \frac{1.35-1}{0.053} = 6.6 \sim t_{97}$ which is greater than 1.96, so reject the null hypothesis at $\alpha = 0.05$.

(d) *(6 points)* Test the following hypothesis: $H_0 : \beta_1 = \beta_2 = 1$.

Solution:

Let $\mathbf{C} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ & & \end{bmatrix}$ and $\boldsymbol{\theta}_0 = \begin{bmatrix} 1 \\ 1 \\ \end{bmatrix}$, then $H_0 : C\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

So $M = \mathbf{C}(X'X)^{-1}\mathbf{C}' = \begin{bmatrix} 0.025 & -0.004 \\ -0.004 & 0.006 \\ & \end{bmatrix}$ with $M^{-1} = \begin{bmatrix} 44.78 & 29.85 \\ 29.85 & 186.57 \\ & \end{bmatrix}$ and $\hat{\boldsymbol{\theta}} = (1.35, 1.61)'$.

Thus

$F - test = \frac{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' M^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)/2}{\hat{\sigma}^2} = \frac{43.83}{0.113} = 387.8 \sim F_{2,97}$

(e) *(6 points)* Find a 95% confidence interval of $\beta_1 + \beta_2$.

Solution:

Let $\theta = \beta_1 + \beta_2$ then $\hat{\theta} = \hat{\beta}_1 + \hat{\beta}_2 = 2.96$.

$\hat{var}(\hat{\theta}) = (0,1,1)\hat{var}(\hat{\boldsymbol{\beta}})(0,1,1)' = 0.0026$ and $SE(\hat{\theta}) = \sqrt{0.0026} = 0.051$

95 % CI of θ is $2.96 \pm 1.96 * 0.051 = [2.86, 3.06]$

(f) *(8 points)* Now you decide to transform x1 and x2 to z1=x1 - 2 and z2=x2-4 where 2 and 4 refer the population minimal car weight and minimal number of cylinders. Refit data with the following linear model $y = \beta_0^* + \beta_1^* z1 + \beta_2^* z2 + error$. Please describe the meaning of β_0^* and fill in the following table:

		Standard		
Parameter	Estimate	Error	t Value	Pr > |t|
(Intercept)	???	???	???	-
z1	???	???	???	???
z2	???	???	???	???

		Standard		
Parameter	Estimate	Error	t Value	Pr > |t|
(Intercept)	9.8	0.085	115.3	-
21	1.35	0.053	25.47	<0.001
22	1.61	0.026	61.81	<0.001

3. *(40 points total)* Consider the set of hypothetical data below $\mathbf{y}_{5\times1} = \mathbf{X}_{5\times3}\boldsymbol{\beta}_{3\times1} + \boldsymbol{\varepsilon}_{5\times1}$, where

$$\mathbf{y} = \begin{bmatrix} 0 \\ 2 \\ 4 \\ 6 \\ 10 \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & 6 & 11 \\ 1 & 7 & 13 \\ 1 & 8 & 15 \\ 1 & 9 & 17 \\ 1 & 11 & 21 \end{bmatrix}, \quad \text{and} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

with $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^{\mathbf{2}}\mathbf{I})$.

(a) *(5 points)* Is there a problem of multicollinearity in this regression? Prove or disprove that there exists no multicollinearity problem.

Solution: yes since the rank of the desing matrix is 2 instead of 3.

(b) *(5 points)* Can you compute OLS estimates of the three parameters and explain why.

Solution: No since the design matrix is not full rank.

(c) *(5 points)* Throwing out any redundant columns of the X matrix if necessary and re-express the model as $\mathbf{y} = \mathbf{X}^*\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$ where $\mathbf{X}^*$ is full rank. Express $\boldsymbol{\beta}^*$ in terms of $\boldsymbol{\beta}$.

One solution is to let $\mathbf{X}^* = \begin{bmatrix} 1 & 6 \\ 1 & 7 \\ 1 & 8 \\ 1 & 9 \\ 1 & 11 \end{bmatrix}$ and $\boldsymbol{\beta}^* = \begin{bmatrix} \beta_0 - \beta_2 \\ \beta_1 + 2\beta_2 \end{bmatrix}$.

(d) *(5 points)* Suppose that there are two students in the Bios663 class whose names are Jim and Chris. Suppose further that they estimated the parameters β_0, β_1 and β_2 by trial and error. As a result, they got different answers, i.e., (-6, -10, 6) and (-10, -2, 2) respectively. And each of them argues that his answer is better. What do you think about these two answers? Which answer fits better to the data?

Solution: the two solutions are the same since they give the same estimated regression model.

(e) *(8 points)* Compute a 95% confidence interval for the mean response of individuals with $x_1 = 1$ and $x_2 = 1$. Do you think the model provides a good estimate for this mean response? Why?

Solution: SSE from the model is 0 and also we can check that the mean respose for $x_1 = 1$ and $x_2 = 1$ is estimable, so the 95% CI is $[-10, -10]$. The model does not provide a good estimate for this mean response since $x_1 = 1$ is far outside the range of the observed x_1 values.

(f) *(6 points)* Show as rigorously as possible whether $H_0 : \beta_0 - \beta_2 = 0 \ \&\beta_1 + 2\beta_2 = 2 \ \&2\beta_0 + \beta_1 = 2$ is testable. If not, can it be reduced to an equivalent testable hypothesis? If yes, present an equivalent testable hypothesis.

Solution: it is not testable but can be reduced to a testable hypothesis, such as $H_0 : \beta_0 - \beta_2 = 0\ \&\beta_1 + 2\beta_2 = 2$.

(g) *(6 points)* Show as rigorously as possible whether $H_0 : \beta_0 + \beta_1 = 0$ is testable. If so, report your test.

Solution: it is not testable.