MSG500/MVE190 Linear Models Final 2011-04-26

Open book and notes are allowed, but no computers or cell-phones.

The exam consist of 6 questions - all equally contributing to the final grade. Make sure to give detailed and specific answers. Avoid yes/no answers. You should also provide a motivation. Good Luck!

Question 1

Suppose that there are n regression data (x_{i1}, x_{i2}, y_i) satisfying

$$y_i = \alpha + \beta_2 x_{i1} + \beta_2 x_{i2}$$

a) What is the least squares estimate of α, β_1, β_2 under the assumption that $\beta_1 = \beta_2$? Write down the solution.
b) What about the case $\beta_1 = 2\beta_2$ - write down the solution.
c) Is it possible to write down the solution explicitly for the case $\beta_1 > \beta_2$? Why/why not?
d) Let's say you do least squares estimation without any particular assumptions on β_1 and β_2. Explain how you would test the assumptions in a)-c) using the results from the standard least squares fit.

Question 2

Suppose you have n observations with one dependent variable and several explanatory variables.
a) What is the largest number of explanatories you can include in a linear regression model and estimate the coefficients using least squres if the explanatories are numerical variables?
b) What if the explanatories are categorical, each with 2 levels?
c) What if the explanatories are categorical, each with 3 levels?

Question 3

In the scatter plot I have identified 3 observations, A,B and C. The mean of x is marked with a vertical dashed line and the least squares fit with a solid line.
a) Which of these observations has the largest residual?
b) Which of these observations has the largest leverage?
c) Which has the largest Cook's distance?
d) Draw in the plot what you think the least squares fit would be if each of these observations were dropped (3 different least squares fits, one for each of A, B or C being dropped).
e) Recommend an analysis for this data set, i.e. suggest which, if any, observations should be dropped and what you think the resulting fit would be in terms of R^2, significance of estimated coefficients, etc.

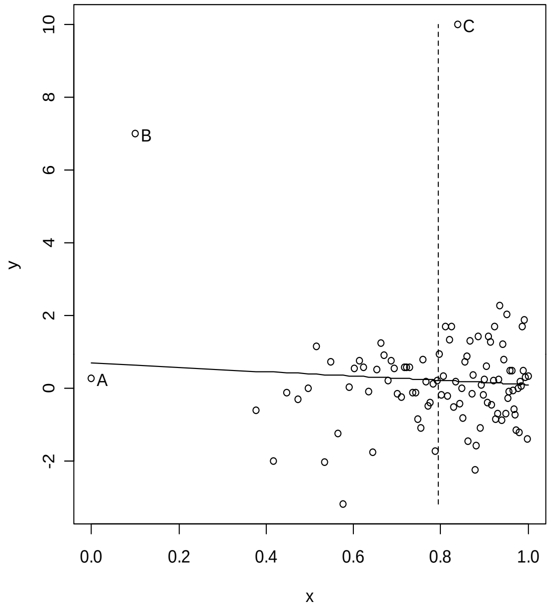

Figure 1: Scatterplot - Question 3

Question 4

Below I show you the summary statistics from two different fits.
Model 1 - Linear regression $y = \alpha + \beta x + \epsilon$

```
            Estimate Std. Error t value Pr(>|t|)
Intercept   0.03818    0.20675   0.185    0.854
x           0.61672    0.12623   4.886   <0.001

R-Squared: 0.3321; Adjusted R-squared: 0.3182
F-statistic: 23.87 on 1 and 48 DF, p-value: < 0.001
```

Model 2 - Quadratic regression $y = \alpha + \beta_1 x + \beta_2 x^2 + \epsilon$

```
             Estimate Std. Error t value Pr(>|t|)
Intercept   -0.008506   0.208413  -0.041    0.968
x            0.360560   0.233740   1.543    0.130
x^2            0.130633   0.100613   1.298    0.200

R-Squared: 0.3553; Adjusted R-squared: 0.3278
F-statistic: 12.95 on 2 and 47 DF, p-value: < 0.001
```

a) Explain which of these models you would choose to describe the data.
b) Is there any evidence of collinearity in the fit for Model 2. From which results can you

deduce this? Explain the source of the collinearity.
c) Discuss some approaches one could use to handle the collinearity problem.

Question 5

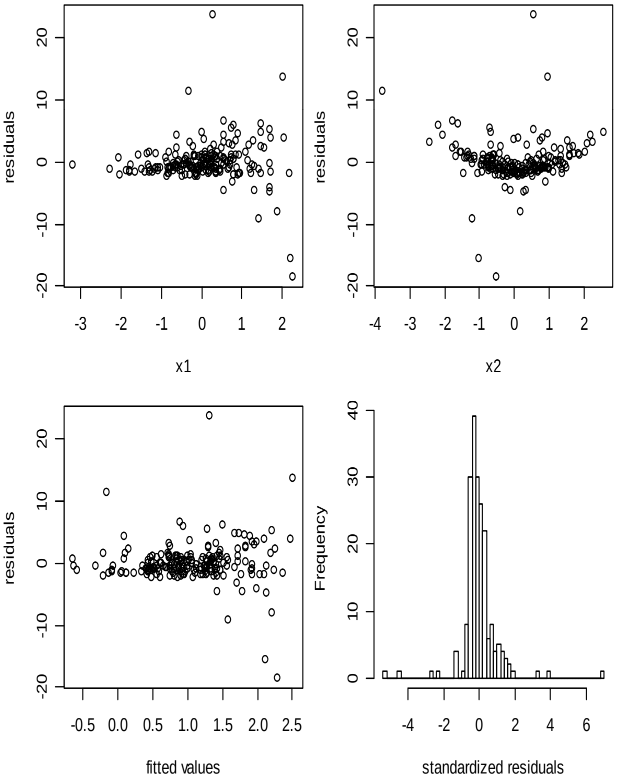

Figure 2: Residual plots - Question 5

In the figure I show some residual plot from a fit of model $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \epsilon$ to data. Identify three potential problems and suggest solutions to each of the three problems.

Question 6

The scatterplot shows y =(bedload deposition) by x =(flow rate), for several samples from the Little Granite Creek near Jackson, Wyoming. The bedload describes particles in the flowing water that are transported along the bed (bottom of the creek).
a) Discuss how you would go about modeling this data set with a linear model. Discuss problems with the data and how you would resolve them.

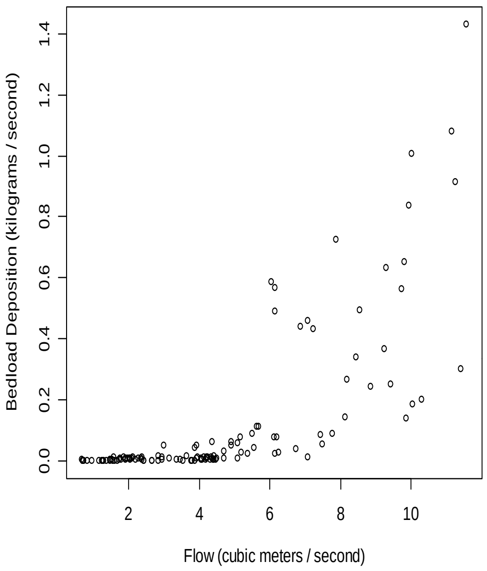

Figure 3: Scatterplot - Question 6

b) An alternative to fitting a linear model is to fit a piecewise linear model. I suggest fit a piecewise linear model with breakpoint at flow-rate $x = 5$: that is, the regression slope is one value up to flow rate $x = 5$ and a different value after that point, and I want the regression lines to be connected at $x = 5$. Explain how you would fit this kind of model to the data using a linear model package like `lm()` in R.
c) Explain how you would use this model to test the hypothesis that the slope of the regression line is constant for all flow.