STATS 500 - Info and Practice Exam 1 Questions

October 14, 2015

Information: The First Exam will be Thursday, October 22 from 6:00-8:00pm in **1528 CCL**. It will cover all the material in the Lectures through Wednesday/Oct 14 class – chapters 1-3, 6 and first part of chapter 7 of textbook as covered in the Lectures. The exam is open textbook (only that book and it must be paper) and open notes. Below is a set of practice Exam Problems. Solutions for the problems will be posted on Monday, October 19. Other example problems to work on are the (undone) problems/exercises presented in class.

Instructions: Attempt all questions. **Give short and specific answers.** Show intermediate steps if there are any. Give only one answer to each question – **if you give multiple answers, the worst answer will be graded**.

Note: The exact number of questions on the actual exam may be different.

The questions concern data on the energy content of municipal waste. This information is useful for the design and operation of municipal waste incinerators. The variables are `Energy` — energy content (kcal/kg), `Plastic` — % plastic composition by weight, `Paper` — % paper composition by weight, `Garbage` — % garbage composition by weight, `Water` — % moisture by weight.

A number of samples of municipal waste were obtained, a regression to predict the `Energy` was computed and the following output obtained:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2245.09     177.89   12.62  2.4e-12
Plastics       28.92       2.82   10.24  2.0e-10
Paper           7.64       2.31    3.30   0.0029
Garbage         4.30       1.92    2.24   0.0340
Water         -37.36       1.83  -20.37  < 2e-16

Residual standard error: 31.5 on 25 degrees of freedom
Multiple R-Squared: 0.964,   Adjusted R-squared: 0.958
F-statistic:  168 on 4 and 25 DF,  p-value: <2e-16
```

1. What are the values of n, p, and $\hat{\sigma}$?
2. Suppose sample A has 10% more plastic (in absolute terms) than sample B. What is the predicted energy content of sample A compared to sample B?
3. What energy content would this model predict for a sample that was purely water? Comment on the reliability of this prediction. What is the technical term for predictions like this?
4. It is difficult to measure these predictors precisely, so here we may expect some error in the predictors. If it had been possible to record the predictors without error, would the R^2 tend to be larger, smaller, or about the same?
5. In the regression summary above, the significance of the predictor `Garbage` is on the borderline. If we remove this variable from the model, would the residual sum of squares (RSS) increase, decrease, stay the same or there is no way of knowing without seeing the output from the refitted model?
6. Compute the 95% confidence interval for β_{paper}. (You may use $t_{25}^{0.025} = 2$).

7. The model
`Energy ~ Plastics + Paper + Water`

   was fit to the data. An F test was computed comparing this model to the original model above. What was the value of the F-statistic for this test?

8. For the model of the previous question, the residual standard error is 33.8 which is larger than in the original model. Does removing a predictor from a model always increase the standard error?

9. The 95% confidence ellipse for $(\beta_{Paper}, \beta_{Garbage})$ is shown in the first panel of Figure 1. What does this suggest concerning the correlation between the predictors `Paper` and `Garbage`? Is this correlation likely to be positive or negative?

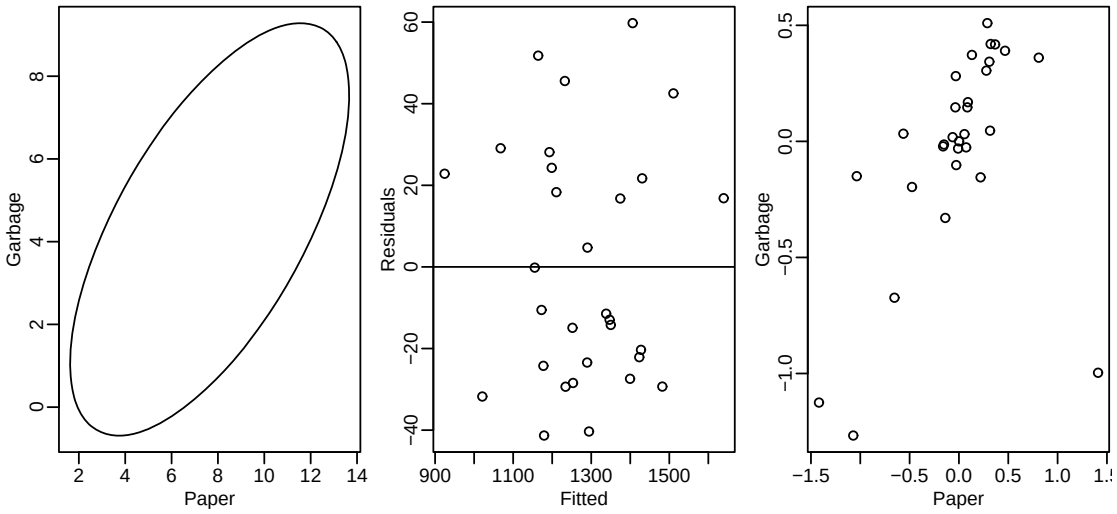

Figure 1: Plots for Garbage Energy Data

10. What would be the conclusion of a test of the hypothesis $H_0 : \beta_{Paper} = \beta_{Garbage} = 0$ in the original model at the 5% level?

11. A diagnostic plot for the original model is shown in the second panel of Figure 1. Does this plot suggest any violation of the standard linear model assumptions?

12. The largest externally studentized residual for the original model was 2.12. The area under the corresponding t-curve to the right of 2.12 is 0.021. Should this point be considered an outlier?

13. The third panel of Figure 1 shows a plot of $\hat{\beta} - \hat{\beta}_{(i)}$ for `Paper` and `Garbage` for the original model. For any of the models fit to the data less one case, is it possible that either of the two predictors, `Paper` and `Garbage`, would not be statistically significant?

The following questions are general and do not relate to the data above.

14. Using $\hat{\beta} = (X^T X)^{-1} X^T y$, derive the formula for $\text{Var}(\hat{\beta})$.

15. Suppose on a partial residual plot you see the data broken up in two non-overlapping clusters. Is this a problem? Why? How would you check?

16. Explain what "regression effect" is, in the context of simple regression.

17. Consider the linear regression model with errors in the (observed) predictors – the model for the output data and the observed input data is given by

$$\begin{aligned} y_i &= \beta_o + \beta_1 x_i^A + \epsilon_i \\ x_i^O &= x_i^A + \delta_i \end{aligned}$$

where the errors $\{\epsilon_i\}$, $\{\delta_i\}$ are independent and have mean 0. Suppose that someone argued that the new model is given by

$$y_i = \beta_o + \beta_1 x_i^O + \epsilon_i - \beta_1 \delta_i$$

and the errors $\{\epsilon_i - \beta_1 \delta_i\}$ are again independent with mean 0, implying that the standard least squares estimates of β_o, β_1 (based on observed input data) are again unbiased. State whether this is true or false and justify your answer.