Statistics 203
Introduction to Regression and Analysis of Variance
Take Home Final
Due Friday, March 18

Prof. J. Taylor

Q. 1) The director of admissions of a small college adminstered a newly designed entrance test to 20 students selected at random from the new freshman class in a study to determine whether a student's grade point average (GPA) at the end of the freshman year (Y) can be predicted from the entrance test score (X). The results of the study can be found at

http://www-stat.stanford.edu/~jtaylo/courses/stats203/data/GPA.table

(a) Fit a simple linear regression to the model.

(b) Plot the estimated regression function along with the data in `R`. Does the line appear to fit well?

(c) Fit a robust regression model to the data? Does this improve the fit? Plot the fitted values on the same plot.

(d) Subsequently, a recording error shows that the 6th test score should read 6.2 rather than 3.2. Refit the OLS model. Compare the estimated coefficients and standard errors in the robust and "corrected" OLS model. Plot the new fitted values on this plot as well.

(e) Obtain an approximate confidence interval for the mean freshman GPA of students with entrance test score $X = 5.0$. Which model do you prefer to use? Why?

Q. 2) A psychologist conducted a study to examine the nature of the relation, if any, between an employee's emotional stability X and the employee's ability to perform in a task group Y. Emotional stability was measured by a written test for which the higher the score, the greater is the emotional stability. Ability to perform in a task group ($Y = 1$ if able, $Y = 0$ if unable) was evaluated by the supervisor. The results can be found in

http://www-stat.stanford.edu/~jtaylo/courses/stats203/data/stability.table

(a) Fit a logistic regression model to the data, state the fitted response function.

(b) Obtain $\exp(\widehat{\beta}_1)$ and interpret this number (β_1 corresponds to stability).

(c) What is the estimated probability that employees with an emotional stability test score of 550 will be able to perform in a task group?

(d) Estimate the emotional stability test score for which 70 % of the employees with this test score are expected to be able to perform in a task group.

(e) Plot the standard diagnostics of the model – do any observations seem to be highly influenftial? outliers? What about the `qqplot`?

(f) Obtain a 95% confidence interval for $\exp(\beta_1)$.

(g) Obtain joint confidence intervals for the mean response π_h for persons with emotional stability test scores $X_h = 550$ and 625, respectively, with an approximate 90% (joint) confidence coefficient. Interpret your intervals.

Q. 3) The carbonation level of a soft drink beverage is affected by the temperature of the product and the filler operating pressure. Twelve observations were obtained and the resulting data are shown below. The data can be found at

http://www-stat.stanford.edu/~jtaylo/courses/stats203/data/carbonation.table

(a) Fit a second order polynomial regression model to the data, including quadratic effects for each main effect as well as an interaction.

(b) Compare the full model to one with just an intercept at level $\alpha = 0.05$.

(c) Does the interaction term contribute significantly to the model?

(d) What about the quadratic terms?

Q. 4) This question studies the Sattherwaite approximation for the distribution of a weighted sum of χ^2 random variables. Suppose we are in a two-sample setting:

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \qquad 1 \leq i \leq 2, 1 \leq j \leq n_i$$

where the errors $\varepsilon_{ij} \sim N(0, \sigma_i^2)$ are independent but may not have equal variance.

(a) Consider a T-test to test whether $\mu_1 = \mu_2$. What appears in the numerator?

(b) What is the variance of the numerator? Propose an unbiased estimate of this variance, to be used in the denominator.

(c) Use Sattherwaite's approximation, which we covered in

http://www-stat.stanford.edu/ jtaylo/courses/stats203/notes/fixed+random.pdf

to approximate the distribution of the denominator.

(d) Compute the T-statistic to test whether $\mu_1 = \mu_2$ using the data in

http://www-stat.stanford.edu/~jtaylo/courses/stats203/data/ttest.table

where `group` is the grouping variable.

(e) Verify that your results agree with `t.test` when using the `var.equal=F` option.

Q. 5) This question studies binomial regression, a generalization of binary regression. The data

http://www-stat.stanford.edu/~jtaylo/courses/stats203/data/florida2000.table

contains a subset of the Florida election results of 2000. They were taken from a subset of the entire national county by county data that can be found at

http://wand.stanford.edu/elections/us/fl/

The columns have the following meaning

- `buchanan` – the number of votes for Pat Buchanan by county;
- `total` – the total number of votes cast by county ;
- `prep96` – the number of votes for the Republican candidate in 1996 by county;
- `pperot96` – the number of votes for Ross Perot candidate in 1996 by county;
- `demographics` – a summary of demographic information from the 1999 census by county.

We will logistic build a model for the number of votes for Pat Buchanan, which was the subject of much debate due to the "butterfly ballot" (see paper in the above website).

(a) Given that N_i votes were cast in the i-th county, propose an additive logistic model for the distribution of

$$Y_i = B_i/N_i,$$

the proportion of votes cast for Pat Buchanan in this county. Include variables for the Perot and Republican votes as well as the demographic variables.

(b) What is the link function of the model?

(c) What is the variance function of the model, is it the same as a binary regression model? How is it different?

(d) Using `glm`, fit the above model. (HINT: YOU MAY NEED TO USE THE `weights` ARGUMENT).

(e) Plot the standard diagnostics. Are there any particularly noteworthy residuals, in terms of influence and absolute value?

Q. 6) Consider the Latin square ANOVA model discussed in the class on experimental design with r treatments and two blocking variables each with r categories:

$$Y_{ijk} = \mu_{...} + \rho_i + \kappa_j + \tau_k + \varepsilon_{ijk} \qquad 1 \leq i, j, k \leq r.$$

Remember, there are only r^2 observations. That is, there is only one observation for each pair (i, j) of blocking variables. The ANOVA table can be found in the notes

http://www-stat.stanford.edu/~jtaylo/courses/stats203/notes/design.pdf

Of interest is to determine whether the treatment has any main effect, i.e. whether

$$H_0 : \tau_1 = \cdots = \tau_r = 0$$

is true or not.

(a) Compute the power of the appropriate F-test to test H_0, as a function of

$$\phi = \frac{r \sum_{k=1}^{r} \tau_k^2}{\sigma^2}$$

which, up to a factor of σ^2 is the non-centrality parameter in $SSTR$

(b) Write an `R` function to compute the power as a function of ϕ, r and α, the level at which the test is carried out.

(c) It may be that the treatments are such that ϕ is quite small, in which case having only one observation per "cell" (i, j) in the Latin square will not have much power. An alternative is to consider n replications within each cell

$$Y_{ijkm} = \mu_{...} + \rho_i + \kappa_j + \tau_k + \varepsilon_{ijkm} \qquad 1 \leq i, j, k \leq r, 1 \leq m \leq n.$$

Repeat (a)-(b) incorporating n into the power and the `R` function.

(d) Using your power function, write an `R` function that gives you the approximate sample size needed to reach power of $1 - \beta$ for some pre-specified (β, ϕ, r, α).