

Handed out: Wednesday November 25, 2009

Due: Friday December 4, 2009

Reading:	Nov. 25	Multiple regression: non-normality (Ch. 11.3, 14.1, 14.14)
	Nov. 30	Multiple regression: model selection (Chapter 9)
	Dec. 2	Multiple regression: polynomials, interactions, indicators (Chapter 8)
	Dec. 4	Multiple regression and ANOVA, Factorial experiments (overview) (Ch. 19-20, 23-24)

NOTE: **Final Exam** – The final exam has two components, a data analysis and a classroom exam. Last year’s final exam (both components) is posted on the website along with solutions for the classroom exam and some brief comments on the data analysis. I will hand out (and post) the data analysis question on Friday Dec 4. This will be due on Friday Dec 11 at 8am, the start of the classroom exam. The data analysis is NOT, REPEAT NOT, intended to be a one-week project – it involves the analysis of a single data and the production of a short (2-5 page) summary report. The classroom exam will be held in the usual classroom on Friday December 11 from 8am-10am. I am happy to make myself available for a review session prior to the exam. We will schedule this after Thanksgiving.

1. **Data analysis.** The US Navy attempts to develop equations to estimate manpower requirements for different types of installations. The file navydorm.txt / navydorm.xls / navdorm.csv contains eight variables for 25 existing officers dormitory buildings.

col 1 = average daily occupancy  
 col 2 = average monthly check-ins  
 col 3 = weekly hours of service desk operations  
 col 4 = square feet of common area  
 col 5 = number of building wings  
 col 6 = operational berthing capacity  
 col 7 = number of rooms  
 col 8 = monthly manpower required (in man-hours)

- Regress manpower requirements (col 8) on the seven available predictors (cols 1-7) that describe the different dorms. Plot the (internally) studentized residuals versus the predicted values and versus the number of monthly check-ins. Do you see any patterns?
- One possible conclusion from the residual plots is that there is nonconstant variance. Consider the square root and logarithmic transformations of the response. Which transformation yields the best value of  $R^2$ ? Do the transformations improve the residual plot?
- Examine the case diagnostics using the **untransformed** model: are there any possible outliers, high leverage, or influential points? If so briefly discuss those cases – look at the original data to see what may be causing the diagnostic statistics you see. (SAS: Remember that you get case diagnostics by including the options “r” and “influence” on the model command in PROC REG; see the sample SAS multiple regression program on the course website.)
- One possible conclusion after looking at the case diagnostics is that there are problems caused by building 23. Perhaps this building is different from the others (maybe it is long-term housing while the others are short-term). Rerun the (untransformed) regression without this observation (recall that you can delete a case by adding “if \_n\_ = 23 then delete;” to the data statement in SAS). How does the residual plot look now? Which variables are significant here.
- Summarize your findings in a paragraph. How would you recommend the Navy estimate manpower requirements for officers dormitory buildings (i.e., which model would you use and why)? What additional studies/analyses (if any) would you suggest that the Navy carry out?

## 2. Model assumptions (theory):

- Incorrect model specification:** Suppose the “true” model generating responses  $Y$  is  $Y = X\beta + Z\alpha + \epsilon$  with  $\epsilon \sim N(0, \sigma^2 I)$ . We are unaware of this and regress  $Y$  on  $X$  (but not  $Z$ ).
  - The sum of squared residuals from our fitted regression model is  $Y^T(I - H)Y$  where  $H = X(X^T X)^{-1}X^T$ . Show that the expected value of the mean square error ( $MSE = SSE/(n - (k + 1))$  where  $k + 1$  is the rank of  $X$ ) is equal to  $\sigma^2$  PLUS a term that depends on  $Z$  and  $\alpha$ .
  - Suppose that you have a good idea about what the value of  $\sigma^2$  should be – perhaps from previous regression studies of  $Y$ . Explain how this information could be used to assess whether we have the correct model specification.

(b) **Non-constant variance:** Consider a simple linear regression model where we believe  $Y = \beta_0 + \beta_1 X + \epsilon$  FOR INDIVIDUALS. Suppose however that data are only available on families; each observation consists of the average response and the average predictor for the members of the family. Let  $Y_i$  denote the average response for family  $i$ ,  $X_i$  denote the average predictor, and  $n_i$  denote the number of individuals in the family.

- i. A regression analysis using  $Y_i$  and  $X_i$  will have non-constant variance. Explain.
- ii. Is the estimate of  $\beta_1$  still useful despite the non-constant variance? Explain. (But see also part (iv).)
- iii. Tell how you could use weighted least squares to address the non-constant variance.
- iv. This situation can lead to more severe problems than non-constant variance. Suppose family 1 has 5 people with  $Y = (11, 12, 13, 14, 15)$  and  $X = (1, 2, 3, 4, 5)$  and family 2 has 5 people with  $Y = (1, 2, 3, 4, 5)$  and  $X = (11, 12, 13, 14, 15)$ . What is the relationship of  $X$  and  $Y$  for the individuals in each family? What is the relationship of  $X$  and  $Y$  based on the average data for each family (there are only two observations)?

NOTE: This problem demonstrates what is known as ecological inference (trying to draw conclusions about individuals based on aggregate data) and the problem with such inferences (we can't tell for sure that the aggregate pattern will hold for individuals). It is also related to Simpson's paradox. Here's a real example: A 1950 regression of literacy rate in each U.S. state on the percentage of immigrants in each state showed a significant POSITIVE relationship so that states with more immigrants had higher literacy rates BUT the data clearly show that at the individual level immigrants are less likely to be literate than non-immigrants.

(c) **Non-normality:**

- i. Explain why inference about regression coefficients is NOT sensitive to non-normality of the errors if the sample size is large.
- ii. Explain why prediction IS sensitive to non-normality of the errors even if the sample size is large.

3. **Model selection:** The output below is a partial summary of an all-subsets regression analysis of the state SAT data. There are  $2^6 - 1 = 63$  models in all but we only report all single variable models and the top 5 models (by  $R^2$ ) for other sizes. Recall that there are  $n = 50$  observations in the data set.

- (a) Note that ALL of the methods agree on the ranking of models for a given fixed size. Explain why this is so.
- (b) Notice that rank and takers are the top two variables individually, yet they do not appear to work well together. Explain how this can happen.
- (c) Which model would be selected by C(p)? by SBC?
- (d) Explain why one might be concerned about predictions made from a model chosen via extensive model adjustment and model selection.

# of	Pred	R-Sq	Adj R-Sq	C(p)	MSE	SBC	Variables in Model
1	0.7742	0.7695	34.03	1157.1	358.5	rank	
1	0.7358	0.7303	47.64	1353.9	366.3	takers	
1	0.3416	0.3279	187.39	3374.4	412.0	income	
1	0.1095	0.0910	269.65	4563.8	427.1	yrs	
1	0.0065	-.0142	306.19	5092.1	432.6	public	
1	0.0040	-.0168	307.08	5104.0	432.7	expend	
-----							
2	0.8471	0.8405	10.22	800.5	342.9	yrs rank	
2	0.8054	0.7971	24.97	1018.5	354.9	expend rank	
2	0.7959	0.7872	28.34	1068.2	357.3	takers yrs	
2	0.7929	0.7841	29.40	1083.8	358.1	income rank	
2	0.7814	0.7721	33.49	1144.2	360.8	takers rank	
-----							
3	0.8711	0.8627	3.69	689.3	338.3	yrs expend rank	
3	0.8584	0.8491	8.21	757.5	343.0	income yrs rank	
3	0.8502	0.8405	11.09	801.0	345.8	takers yrs rank	
3	0.8472	0.8372	12.16	817.1	346.8	yrs public rank	
3	0.8411	0.8308	14.31	849.6	348.7	public expend rank	
-----							
4	0.8771	0.8661	3.58	672.1	339.8	yrs public expend rank	
4	0.8767	0.8657	3.72	674.2	340.0	takers yrs expend rank	
4	0.8746	0.8634	4.47	685.8	340.8	income yrs expend rank	
4	0.8593	0.8468	9.87	769.1	346.6	income yrs public rank	
4	0.8584	0.8458	10.20	774.2	346.9	takers income yrs rank	
-----							
5	0.8787	0.8649	5.00	678.3	343.1	takers yrs public expend rank	
5	0.8773	0.8634	5.48	685.8	343.6	income yrs public expend rank	
5	0.8769	0.8629	5.64	688.4	343.8	takers income yrs expend rank	
5	0.8594	0.8434	11.84	786.1	350.4	takers income yrs public rank	
5	0.8425	0.8246	17.82	880.5	356.1	takers income public expend rank	
-----							
6	0.8787	0.8618	7.00	694.0	347.0	takers income yrs public expend rank	