

Stat 210 - HW 7 Solutions/Comments (Fall 2009)

1. Manpower data analysis.

- As hinted in part (b), the residual plot does show a pattern in that the variance of the residuals appears to increase as the fitted value (or the number of monthly check-ins) increases.
- Transformations - The square root transformation yields a better residual plot and higher $R^2 (= 0.9333)$ than the log transformation ($R^2 = 0.8407$). Note that R^2 in the original untransformed scale is higher (0.9613) but the residual plot indicates the assumptions are not satisfied on the original scale.
- Here we go back to the original, untransformed regression. For those data, observations 22, 23, 24 are highly influential (Cook's $D > 1$) and have high leverages. Observation 23 has an especially high Cook's D . The studentized residuals for observations 23, 24 are also high. It is a good idea to examine the data. In doing so, we find that 22, 23, 24 are all large buildings and that for building 23 the avg occupancy is great than capacity (this seems odd!).
- Without case 23 the residual plot is improved. There is still some slight evidence of increasing variance but it is minor. Notice that observations 22, 23 (which used to be 22, 24) are still high leverage but not nearly so influential. Also note that avg occupancy and checkin are significant which makes good sense.
- The writeups should focus on the question that motivated the study. Short write-ups should describe the problem under study, the key results (including sign and magnitude of important relationships), and any limitations or future directions. Here's a possible paragraph for these data – This study attempts to develop a formula or approach for predicting US Navy dormitory manpower requirements. The analysis suggests that it is critical for the Navy to determine the nature of the buildings under consideration. In particular one observation in the current data set does not appear to be like the others and plays a key role in the analysis. If observation 23 has been misclassified or the data recorded in error, then it should be removed from the data set. In that case, there is a fairly natural model forecasting expected manpower (in man-hours required per month) as $21 \times$ average daily occupancy plus $1.4 \times$ average monthly checkin. If on the other hand we must find a model that appears to work across this full set of dormitories, then it might be best to rely on a model which explains square root of manpower in terms of capacity, number of rooms, average monthly checkins, and weekly hours of desk operation.

SAT program

```
filename regdata 'h://HAL/Courses/Stat210/navydorm.txt';
data regress;
  infile regdata firstobs=2;
  input avgocc avgchkin hours sqft wings capacity rooms manpower;
  logmanp = log(manpower);
  sqrtmanp = sqrt(manpower);
proc reg;
  model manpower = avgocc avgchkin hours sqft wings capacity rooms / r influence;
  output out=resids p=yhat r=resid;
proc rank normal=blom out=residnrm data=resids;
  var resid;
  ranks residnrm;
proc gplot data=residnrm;
  plot resid*residnrm resid*yhat;
NOTE: For (b) repeat last 3 procedures with logmanp as response and then sqrtmanp as response.
NOTE: For (d) add 'if _n_=23 then delete;' to data statement and rerun original regression.
```

2. Models assumptions (theory questions)

- Incorrect model specification
 - Let's start by calculating the SSE. It turns out that $SSE = Y^T(I-H)Y = (X\beta + Z\alpha + \epsilon)(I-H)(X\beta + Z\alpha + \epsilon) = (X\beta + Z\alpha + \epsilon)(I-H)X\beta + \beta^T X^T(I-H)Z\alpha + \beta^T X^T(I-H)\epsilon + \alpha^T Z^T(I-H)\epsilon + \epsilon^T(I-H)Z\alpha + \epsilon^T(I-H)\epsilon = \alpha^T Z^T(I-H)Z\alpha + 2\alpha^T Z^T(I-H)\epsilon + \epsilon^T(I-H)\epsilon$. The last equality is obtained by noticing that $(I-H)X = X^T(I-H) = 0$ (see HW 6) which makes the first three terms zero. The first term is a constant and the second term has mean zero since ϵ has mean zero. Thus $E(SSE) = \alpha^T Z^T(I-H)Z\alpha + E(\epsilon^T(I-H)\epsilon)$. The final expected value can be calculated by multiplying this quadratic form out and taking term by term expectations or by noting that $E(\epsilon^T(I-H)\epsilon) = E(\text{tr}(\epsilon^T(I-H)\epsilon)) = E(\text{tr}((I-H)\epsilon\epsilon^T)) = \text{tr}((I-H)E(\epsilon\epsilon^T)) = \text{tr}(\sigma^2(I-H)I) = \sigma^2 \text{tr}(I-H) = \sigma^2(n - (k+1))$. The latter argument uses properties of the "trace" which is the sum of the diagonal elements of a matrix and recalls that $\sum_i h_{ii} = (k+1)$. Putting everything together yields $E(MSE) = E(SSE)/(n - (k+1)) = \sigma^2 + \frac{1}{n-(k+1)}\alpha^T Z^T(I-H)Z\alpha$ as advertised.
 - If you have some knowledge about σ^2 then you can compare your regression MSE to the expected value. If the MSE is greater than the value you were expecting then this is evidence that there are some variables effecting

Y that are not included in your model (these are the Z's). Note that this argument is similar to the one used in the lack of fit test on HW 5.

(b) Non-constant variance

- i. The problem specifies that we believe a linear model is appropriate for individuals. Let's use the subscript ij to denote the j th individual in family i . Then we believe $Y_{ij} = \beta_o + \beta_1 X_{ij} + \epsilon_i$ with $\epsilon_i \sim N(0, \sigma^2)$. We only have the family average values $Y_i = \frac{1}{n_i} \sum_j Y_{ij}$ and $X_i = \frac{1}{n_i} \sum_j X_{ij}$ where n_i is the number of members in the family. By averaging our model we find that $Y_i = \beta_o + \beta_1 X_i + \eta_i$ where $\eta_i = \frac{1}{n_i} \sum_j \epsilon_{ij}$ is $N(0, \sigma^2/n_i)$. It follows from this last piece that the regression on household averages will have non-constant variance with bigger households providing more precise information.
- ii. As we discussed in class the least squares regression estimate of the slope is still unbiased when there is constant variance. Thus the estimate is still useful though no longer optimal.
- iii. In this case we actually have a good idea about the cause of the non-constant variance and therefore how to adjust for it. We should do weighted least squares with the weight for family i equal to the number of individuals in the family n_i .
- iv. This last part of the problem was just provided as a caveat about being sure what your model is saying about the data. If you run a regression on households, then the conclusions hold for households and not individuals. The hypothetical here provides a case where Y and X are positively related WITHIN each family (individuals with more X are associated with higher values of Y) and negatively related BETWEEN families (families with more X are associated with lower values of Y).

(c) Non-normality

- i. Inference about the regression coefficients is based on the distribution of $\hat{\beta}$ which we have shown is multivariate normal $N(\beta, \sigma^2(X^T X)^{-1})$. This result assumes normal errors. It turns out however that we can write $\hat{\beta} = \beta + (X^T X)^{-1} X^T \epsilon$ which shows us that $\hat{\beta}$ is a weighted average of the elements of ϵ (more specifically $\hat{\beta}_j$ is obtained as $\sum_k a_{jk} \epsilon_k$ where a_{jk} is the relevant element of $(X^T X)^{-1} X^T$). This weighted average of the ϵ 's will follow a normal distribution even if the individual ϵ 's do not. This means we can trust our regression inferences in large samples even if the normal assumption is not correct. (This is true as long as the a_{jk} elements are mainly all the same order of magnitude.) This is basically the central limit theorem in action (but now for weighted averages).
- ii. This question was a bit vague. To make a prediction we don't require normality, our prediction for an individual with covariates x_* is $x_*^T \hat{\beta}$. However our prediction interval DOES require normality. The prediction interval $x_*^T \hat{\beta} \pm t_{n-(k+1), 1-\alpha/2} \sqrt{MSE(1 + x_*^T (X^T X)^{-1} x_*)}$ relies on the assumption that $Y_* = x_* \beta + \epsilon_*$ is normal which of course relies on the assumption that the errors are normal.

3. Model selection

- (a) For a given size model all of the criteria are equivalent to comparing the SSE (SS(Error)) and so all will agree on the ordering of the models. The criteria differ in how they trade off SSE vs. size of model.
- (b) Rank and takers are both effective predictors but the model that includes both is not much better than the model with just rank. This could happen if the two variables are highly correlated.
- (c) C_p would choose the model with "yrs, public, expend, rank" which has $C_p = 3.58$. Note that this is lower than $p + 1 = 5$ suggesting some caution is required. SBC chooses the smaller model without "public".
- (d) This question was trying to raise the issue of model validity. We would hope that whatever model we end up with is accurate for making predictions on new data (from the same population). If we do a great deal of work on the model (deleting outliers, trying many transformations, using model selection to decide on variables), then we open up the possibility that we have overfit to the nuances of our data set. If so, we may not predict as well. To put it another way we analyze data with the view that the observed data Y is determined by the model AND random individual variation. We must be careful that we don't fit our model to the random individual variation because if we do then our conclusions won't be valid outside the current data set.