

Stat 210 - HW 6 Solutions/Comments (Fall 2009)

1. SAT regression output

- (a) The numerical codes assigned to regions don't have any meaning. If we include this variable in the regression, then we are implicitly assuming that the difference between S and NE (2-1) is the same as the difference between MW and S (3-2) and W and MW (4-3). This assumption is almost certainly crazy. As we saw on Dec 2 a categorical variable like this should be replaced by a series of indicators for the different regions.
- (b) Model 1
 - i. Using a two-sided t test with 40 d.f. the cutoff for significance at the .05 level is about 2.02 and the cutoff for significance at the .01 level is about 2.70. Rank and South both have p -values less than .01; no other variable has p -value less than .05.
 - ii. The coefficient of rank means that for each additional 1 point improvement in the rank of students taking the exam in a state with all other factors held fixed there is an expected increase of 9.02 points in state mean SAT score.
 - iii. The square root of the MSE is an estimate of the standard deviation of the regression errors or regression variation. Here we are confident that state mean SAT scores are within 42 points of the regression line (this is two standard deviations).
- (c) To test the hypothesis of five coefficients being simultaneously zero we must use information from model 1 and model 2. Our test statistic is $F = ((SSE(2) - SSE(1))/5)/MSE(1) = (10730/5)/437.1367 = 4.91$. This should be compared to the $F_{5,40}$ distribution which yields a p -value of .0014. This being small suggests that we reject the null hypothesis; at least one of these variables is helping the regression.
- (d) Expenditures must be correlated with region. It would be better to say that expenditures vary a great deal across regions in the same way that SAT scores do. (Since you can't really "correlate with region") This is another example of the difficulty with regression – the regression of Model 1 indicates that expenditures don't matter if you know what region a state is in but Model 3 indicates that expenditures are important information if you are not given the regions.

2. Heart disease cost regression in SAS - The quality of the SAS work this week was generally quite poor. You should present results in a way that makes them easy to find – many folks just scribbled answers on output.

- (a) The intent of this part of the question was to reinforce that it is important to think about the relationships among your variables in advance. You should comment on each variable. In this case interv is the most highly correlated with logcost (around 0.7) so this is the most important predictor (more interventions/procedures means higher cost). Other medical variables were also significantly positively correlated with logcost including ervisits , drugs , complications , comorbidities and duration (correlations between .15 and .4). These correlations are weaker so the plots do not show as obvious a trend. Also, if you look carefully at the plots you can see that the more strongly related variables don't seem to have a linear relationship – the impact of the covariates seems to flatten out for larger values.
- (b) Most of the variables are highly coefficients that are significantly different than zero ($p < .0001$) with the obvious signs (i.e., positive relationships for ervisits , interventions , complications , comorbidities , and duration). Age and gender are not significant predictors. The only surprise result was that drugs had a negative sign (it had a positive correlation) but the coefficient is not significantly different than zero so this is not a big issue.
- (c) The normal probability plot seems fairly linear suggesting that the log transformation (of cost) leads to roughly Gaussian errors. The residual plot shows a non-linear pattern – perhaps a quadratic pattern. We can't tell from here which variables are involved but it's fairly clear that there are primarily negative residuals for small and large predicted values and mainly positive residuals in the middle. Several people identified non-constant variance in this plot as well. I see some signs of that but it's not too dramatic – there is less variance on the right-hand-side of the plot but there are fewer points here as well so it may be that we haven't seen as many large residuals yet out there.
- (d) Apologies ... taking logarithms didn't work very well because the logarithm of 0 is not defined and there are many zero values for some variables. One idea is to not transform variables like gender or complications where there were many zeros. Another approach in such instances is to take the $\log(x+1)$ rather than the $\log(x)$. The $\log(x+1)$ is not quite so intuitive a transformation though. The quadratic idea seemed to work out better for everyone although it doesn't make any sense to include a quadratic term for gender since this is a dichotomous variable. Several of the medical variables have significant quadratic terms and significant linear terms (interv , dur , comorb). The quadratic terms make the regression a bit harder to interpret but they definitely lead to an improved regression. The residual plot looks much better (still a touch of non-linearity but much improved), R^2 has increased from .59 to .67, and the $\hat{\sigma}_e$ has decreased from 1.23 to 1.085. The latter means prediction intervals are more precise.

- (e) Remember to always start with a description of the scientific problem when writing up your results and to emphasize scientific results rather than statistical steps. For example: The aim of this study is to understand the key factors associated with the cost of treatment for heart disease based on a database for 788 patients of a large insurance company. Previous work suggested that the linear regression model with normal errors was more appropriate for analyzing logarithm of cost, so this is the variable used here. Note that this impacts our interpretation of coefficients. Key medical variables include the number of emergency room visits, number of interventions performed, complications recorded, drugs ordered, comorbid conditions, and the duration of hospital stays. Gender and age are also available but did not seem to have a significant impact on cost. All of the medical variables seem to be related to cost although a multiple regression suggests that drugs is not a useful predictor after controlling for the other variables. An initial analysis suggested that the relationship was still non-linear, even after transforming the response. As a result a model was fit that included quadratic effects for each variable. Number of interventions, number of comorbidities, and duration of stay all demonstrate quadratic relationships with cost; cost increases as these variables increase but the quadratic coefficients are negative suggesting that the rate of increase slows with more interventions (or comorbidities or duration). Another significant finding is that even after transformation a considerable amount of patient-to-patient variation is unexplained. For a “typical” patient having average values on the predictors (female, age 60, with 5 interventions, 0 drugs, 3 er visits, 0 complications, 4 comorbidities, and 160 days of hospital stay) the predicted (median) cost is \$880 and a 95% prediction interval ranges from \$100 to \$7500.

SAS program

```
filename heart 'h:\HAL\Courses\Stat210\Kutner5th\ischemic.txt';
data psa;
  infile heart firstobs=1;
  input id cost age gender interv drugs ervisit comp comorb dur;
  logcost = log(cost);
  age2 = age*age;
  interv2 = interv*interv;
  drugs2 = drugs*drugs;
  ervisit2 = ervisit*ervisit;
  comp2 = comp*comp;
  comorb2 = comorb*comorb;
  dur2 = dur*dur;
proc corr;
  var logcost age gender interv drugs ervisit comp comorb dur;
proc gplot;
  plot logcost*age logcost*gender logcost*interv logcost*drugs logcost*ervisit logcost*comp
       logcost*comorb logcost*dur;
proc reg;
  model logcost = age gender interv drugs ervisit comp comorb dur;
  output out = resids p = yhat r = resid student = studres;
proc rank normal=blom data = resids out = norm;
  var studres;
  ranks nrm;
proc gplot data=norm;
  plot studres*nrm studres*yhat;
proc reg;
  model logcost = age age2 gender interv interv2 drugs drugs2 ervisit ervisit2 comp comp2
               comorb comorb2 dur dur2;
  output out = resids2 p = yhat2 r = resid2 student = studres2;
proc rank normal=blom data = resids2 out = norm2;
  var studres2;
  ranks nrm2;
proc gplot data=norm2;
  plot studres2
```

3. Interpreting coefficients

- (a) Standardized coefficients
- i. People argued that the standardized coefficient is unitless in a variety of ways but the clearest is to take explicit notice of the units for different quantities. The regression coefficients are measured in units of Y divided by units of X (insurance policies per fire). The s.d. of X is in the same units as X and the s.d. of Y is in the same units as Y thus in forming the standardized coefficient all of the units cancel and that is why it is a unitless quantity.

- ii. The standardized coefficient is the least squares regression coefficient multiplied by the s.d. of the predictor and divided by the s.d. of the response. For these predictors we get: pctmin $.0091 * 32.59 / .63 = .47$; fires $.57$; thefts $-.36$; pctold $.30$; income $.09$. Thus fires has the largest standardized coefficient (as well as the most significant t-statistic).
 - iii. In a simple linear regression the standardized slope (this is what I was referring to) turns out to be equal to the simple Pearson correlation of X and Y . This can help us interpret standardized coefficients – they are something like a correlation (but not precisely so because of the impact of other variables).
- (b) As discussed in great detail in discussion last week, the Type I SS is the sequential sum of squares. It measures the change in SSE (or SS(Model)) when a variable is added to a regression model that contains all preceding variables in the sequence. The Type II SS is the partial sum of squares. It measures the change in SSE (or SS(Model)) when a variable is added to a regression model containing ALL other predictor variables. Naturally these two definitions agree for the final variable in our sequence.
- (c) The positive correlation reflects the estimated negative direct effect as well as some positive indirect effects. If you apply the formula given in class, then you can get some insight into the indirect effects. Specifically, we find thefts is highly positively correlated with fires and pctmin which both have important positive effects. Here I'd argue that the positive correlation matches our intuition more than the negative coefficient – the latter is tricky to interpret because it involves imagining that we can keep all variables the same while increasing thefts (but of course its not really possible to do this). Many of you just cited the formula – I was looking for some specific information about what other variables might explain the unexpected sign in the context of this example.

4. Case diagnostics

- (a) Case 1 - The residual is easiest to obtain from $Y_i - \hat{Y}_i = 1088 - 1057 = 31$. This is not the most precise because the predicted value if rounded off but it is sufficient. Many of you used $e_i = r_i \sqrt{MSE(1 - h_{ii})} = 1.250 * \sqrt{694 * (1 - .1161)} = 30.96$.
- Case 2 - The studentized residual is $r_i = e_i / \sqrt{MSE(1 - h_{ii})} = 37.3739 / \sqrt{694 * (1 - .1693)} = 1.56$.
- Case 3 - The externally studentized residual is $t_i = e_i \sqrt{\frac{n - (k + 1) - 1}{SSE(1 - h_{ii}) - e_i^2}}$. Here $t_i = 26.2569 * \sqrt{\frac{42}{29842 * (1 - .1092) - 26.2569^2}} = 1.057$.
- Case 4 - Leverage can be easily found from studentized residual formula $r_i = e_i / \sqrt{MSE(1 - h_{ii})}$. This yields $(1 - h_{ii}) = 24.0234^2 / .928^2 (694) = .9656$ so the leverage is $.0344$.
- (b) The average leverage is $(k + 1) / n = .14$. I apologize – I initially thought k was 5 and marked everyone wrong; I should have known better as the chances of everyone else making a mistake are very small. Fortunately I corrected this before returning the papers. We can use two or three times the average leverage as a criterion. Cases LA (.36) and AK (.58) are high leverage cases. These have unusual X 's and **may** have a big impact on the regression. A review of the data shows that LA is high on mean income (this surprises me) and low on percentage of public school students, while AK is high on income, high on percentage of public school students, and especially high on expenditures per student.
- (c) The traditional terminology is that “outliers” refers to unusual values of the response for the given values of the predictors. It is most common to assess this by looking at studentized or externally studentized residuals. There are 3 studentized residuals out of 50 that are larger than 2 which is not terribly unusual. One of these (Alaska) is larger than 3 which is quite unusual. You can also use the externally studentized residuals which tell the same basic story (except one more inches above 2). Note that using the Bonferroni correction for examining 50 externally studentized residuals suggests a critical value of 3.5 (which none of the points surpass).
- (d) Not surprisingly Alaska shows up as being influential. It has high leverage and a large residual. The single best measure of influence is Cook's Distance. It would certainly be worthwhile to re-run the regression without Alaska and see how things change (I did this – the same variables remain significant but coefficients change quite a bit with rank going from 8.47 to 9.78, expend going from 2.2 to 3.7 and years going from 22.6 to 16.5).

5. Basic multivariate regression theory

- (a) β_1 = change in the expected value of Y for a one-unit change in X_1 with all other variables being held fixed.
- (b) The simple correlation coefficient does not tell you what sign the multiple regression coefficient will have. The multiple regression coefficient is also effected by the correlation of X_3 with other variables, and the correlation of those variables with Y .
- (c) The intuition here is pretty easy. The effect on the expected value of Y for a one unit change in X_j should be the same before and after re-scaling (since Y is not being re-scaled). Let's use $\beta_{j,old}$ and $\beta_{j,new}$ to identify the two coefficients. A one unit change in X_j leads to a $\beta_{j,old}$ change in the expected value of Y before re-scaling. After rescaling THE SAME CHANGE is a $1/c_j$ change in the new variable which leads to a $\beta_{j,new}/c_j$ change in the

expected value of Y . These will be the same if $\beta_{j,new} = c_j \beta_{j,old}$. If we go from inches to feet, then the X values get smaller so the β should get larger. You can also prove this from the definition of $\hat{\beta} = (X^T X)^{-1} (X^T Y)$ but that is not done here.

(d) H matrix

- i. $H^2 = (X(X^T X)^{-1} X^T)(X(X^T X)^{-1} X^T) = X(X^T X)^{-1} X^T = H$
- ii. $HX = X(X^T X)^{-1} X^T X = X$. Recall that HW calculates the closest point to the vector W in the space spanned by X . H finds the regression coefficients β_w such that $X\beta_w$ is as close as possible to W . If we apply H to any column in X then the closest point in the space spanned by X provides a perfect match (just choose regression coefficient 1 for the column in question and 0 for the other columns).

6. Regression theory

(a) First note that $W = (X^{*T} X^*) = \begin{pmatrix} X^T X & X^T Z \\ Z^T X & Z^T Z \end{pmatrix}$ and that the coefficient of Z will be obtained by seeing what

happens when the bottom row of W^{-1} is multiplied by $(X^{*T} Y) = \begin{pmatrix} X^T Y \\ Z^T Y \end{pmatrix}$. We can identify $D = Z^T Z$, $C = Z^T X$, $B = X^T Z$, and $A = X^T X$ and then calculate $E = Z^T Z - Z^T X (X^T X)^{-1} X^T Z = Z^T (I - H_x) Z$ where $H_x = X (X^T X)^{-1} X^T$. Finally $\hat{\beta}_z = -E^{-1} C A^{-1} X^T Y + E^{-1} Z^T Y = -[Z^T (I - H_x) Z]^{-1} Z^T X (X^T X)^{-1} X^T Y + [Z^T (I - H_x) Z]^{-1} Z^T Y = [Z^T (I - H_x) Z]^{-1} [Z^T (I - H_x) Y]$.

(b) $e_y = Y - \hat{Y} = Y - H_x Y = (I - H_x) Y$.

(c) Recall from earlier homework that if you regress Y on X without an intercept then the slope estimate is $\sum_i x_i y_i / \sum_i x_i^2 = X^T Y / X^T X$. Applying this with e_y and e_z yields $\hat{\beta}_z = e_z^T e_y / e_z^T e_z = [Z^T (I - H_x)^T (I - H_x) Y] / [Z^T (I - H_x)^T (I - H_x) Z]$. It turns out that $(I - H_x)^T (I - H_x) = (I - H_x)$ and therefore the two expressions ((a) and (b)) are the same. The interpretation of adding a variable is that the coefficient of the new variable is the same as would be obtained by regressing residuals of Y on X (what is not yet explained) on the residuals of Z on X (what new information does Z contain).