

1. Introduction to regression I - baseball

- (a) $\hat{\beta}_1 = 5273/33147.37 = 0.159$ and $\hat{\beta}_0 = (2430/30) - (0.159) * (2657/30) = 66.92$. The fitted regression line is wins = 66.92 + .159*payroll. (Interestingly I did this several years ago and got very similar regression coefficients!)
- (b) The $s.e.(\hat{\beta}_1) = \hat{\sigma}/\sqrt{\sum_i(x_i - \bar{x})^2} = \sqrt{2947.18/28}/\sqrt{33147.37} = .056$. Then we test the hypothesis that the slope is zero with $t = .159/.056 = 2.82$. When compared to the t_{28} distribution (one-sided) we find $p = .0044$. This test tells us that there is a significant association between team payroll and the expected number of wins.
- (c) First note that the expected number of wins for a team with payroll 100 is $66.92 + .159(100) = 82.8$ (recall this is an average over the “population” of teams with payroll 100). The s.e. for this expected response is $\sqrt{\hat{\sigma}^2 * (1/n + (x_i - \bar{x})^2 / \sum_i(x_i - \bar{x})^2)} = \sqrt{(2947.18/28) * (1/30 + (100 - 88.5667)^2/33147.37)} = 1.98$. Then the 95% CI is $82.8 \pm t_{28,.975}1.98 = (78.7, 86.9)$. If we want to make a prediction for a specific team with payroll 100 we use the same estimate (82.8) but now the s.e. is $\sqrt{\hat{\sigma}^2 * (1 + 1/n + (x_i - \bar{x})^2 / \sum_i(x_i - \bar{x})^2)} = \sqrt{(2947.18/28) * (1 + 1/30 + (100 - 88.5667)^2/33147.37)} = 10.45$. Notice how important that extra term is!! The 95% PI for the number of wins for a team with payroll 100 is $82.8 \pm t_{28,.975}10.45 = (61.4, 104.2)$. The latter interval is trying to predict the outcome for a single team which will be affected by many other factors. The CI is for an average of teams which will be less variable. The PI is so wide as to be useless – almost all teams have wins in that interval!
- (d) We saw above that payroll is a significant predictor so the commissioner is wrong. His argument picks two teams that appear to go against the general trend – it’s always possible to find such teams. Of course the large s.e. for a prediction does support the commissioner in as much as it points out that other factors play a large role too! (R^2 here is about .22 so almost 80% of the variability in team performance is explained by other factors.)

2. Basics of regression II - temperature and latitude

- (a) This is a bit of an unusual problem. There is enough information provided but it is not in the usual form. First we learn that the slope must be -1.6 because a 5-degree change in latitude leads to an 8-degree drop in average April temperature. Second, by noting the prediction for LA we find the intercept is $\hat{\beta}_0 = 63 - 34 * (-1.6) = 117.4$. Thus the regression equation is $\text{avgtemp} = 117.4 - 1.6 * \text{latitude}$. The final sentence provides some information about σ but you first must recognize that it is giving you a value (the ± 6) that corresponds to $\pm 2\sqrt{\hat{\sigma}^2(1 + 1/n + (x_i - \bar{x})^2 / \sum_i(x_i - \bar{x})^2)}$. The last term varies for different observations but it should be about $1/n$ on average. So we can get $\hat{\sigma}^2 \approx 6^2/2^2/(1.1) = 8.182$ and $\hat{\sigma} \approx 2.86$.
- (b) The information in part (b) is telling us something about the total variation in the response (avgtemp). The s.d. is 9.5 and the variance is 9.5^2 . Then our approximation for R^2 is $R^2 = 1 - (2.86^2/9.5^2) = .91$. Many people found this by noting that $SST = (n - 1)9.5^2$, $SSE = (n - 2)2.86^2$, and then $R^2 = 1 - SSE/SST$.

3. SAS regression

- (a) It’s really hard to tell if there is a linear relationship between cost and duration. The range of the response cost is so large that many observations with low cost are down near the horizontal axis. (This is a first clue that we may need a transformation.)
- (b) The regression line is $\text{cost} = 1205.55 + 9.72 * \text{dur}$. It’s about \$1200 plus \$10 per day. Note that the slope is significant. Thus if we don’t look at the residuals we might just accept this result.
- (c) The normal probability plot shows the residuals are definitely not normally distributed. The plot of residuals vs predicted cost is kind of unusual. There’s a line along the bottom. This is another manifestation of the non-normality – there are some big positive costs with big positive residuals but no big negative residuals.
- (d) Taking the logarithm seems to work wonders. The normal probability plot now shows that the residuals are approximately normal. The residual plot vs fitted values looks like a nice random spread around zero. There seems to be a large number of points at the same predicted value at the left of the plot; these all have duration zero. (SAS Note: Several of you discovered (as I have) that if you use the same SAS session and same basic SAS program to run the log analysis you don’t get updated residual plots. You need to save the data from the log analysis to a different data set.)
- (e) The fitted regression line is $\log(\text{cost}) = 5.30 + .00614 * \text{dur}$. The 95% confidence interval for the slope is $.00614 \pm 1.96 * .0005 = (.0051, .0071)$. The slope indicates that every extra day adds .006 to the expected logarithm of the cost. It’s not exactly right but if we exponentiate then we can say that the cost increases by a multiple of $e^{.006} = 1.006$ for each day (i.e., by 0.6%).
- (f) The predicted $\log(\text{cost})$ is 6.42 and the 95% prediction interval for $\log(\text{cost})$ is (3.05, 9.80). If we exponentiate then you get a prediction of \$614 and a prediction interval of \$21 and \$18,034. Note that when we are predicting an individual value (as here) then we don’t have to worry that the mean of the log is not the same as the log of the mean. If we have a predicted log cost then we can exponentiate to get a predicted cost (i.e., it’s not a mean so we don’t have a problem). I’d also note that it is important to give results as requested; I asked for a prediction interval for cost, not $\log(\text{cost})$.

- (g) Paragraph: A health insurance company decided to explore the relationship of cost of care for heart patients and the duration of the patient's treatment. A regression analysis suggests an exponential relationship with each extra day associated with a 0.6% increase in cost. A key limitation is that there are many other variables that effect health care costs. This is most evident in the large residual standard error (1.7 on the log scale) which means that prediction intervals for patients are very imprecise, with lower bounds near zero and upper bounds in the tens of thousands of dollars. NOTE: Paragraphs are getting much better. One problem remains though. Most of you don't give any sense of quantitative results – what is the relationship between duration and cost? You should say something.

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);
proc means mean stddev min max n;
  var cost logcost dur;
proc gplot;
  plot cost*dur;
proc reg;
  model cost = dur;
  output out = resids p = yhat r = resid student = studres;
proc rank normal=blom data = resids out = norm;
  var resid;
  ranks nrm;
proc gplot data=norm;
  plot resid*nrm resid*yhat studres*nrm studres*yhat;
proc reg;
  model logcost = dur / p r clm cli;
  output out = resids2 p = yhat2 r = resid2 student = studres2;
proc rank normal=blom data = resids2 out = norm2;
  var resid2;
  ranks nrm2;
proc gplot data=norm2;
  plot resid2*nrm2 resid2*yhat2 studres2*nrm2 studres2*yhat2;
run;
```

4. Regression/ANOVA/model checking

(a) Regression

- i. The regression coefficient of -0.324 indicates that the expected or mean chemical concentration drops by .324 for each additional hour the solution sits.
- ii. The t -statistics for the slope is $t = -.324/.043 = 7.53$. With 13 d.f and a two-sided alternative it turns out that $P < .0001$ (in fact much less). Thus we are confident the slope is not zero and time is a significant predictor of concentration.

- (b) ANOVA: The F test for a difference between group means yields $F = (15.36/4)/(0.16/10) = 240.1$ which yields $P < .0001$ when compared to the F-distribution with 4 and 10 d.f. There are clearly significant differences among the mean concentrations under the five conditions. The extreme F-value is almost certainly a sign of a made up data set which yields an extremely small MSE.

- (c) The ANOVA SS(Error) describes variation due to factors other than time since it pools variance estimates for different solutions taken under the same time condition. This is what we usually mean when we think of “pure error” or “individual variation”. By contrast the regression SS(Error) will also include any failure of the linear model (e.g., a quadratic pattern). You can see this analytically by noting that $SSE(Reg) = \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - \bar{y}_i)^2 + \sum_i (\bar{y}_i - \hat{y}_i)^2$ where we have introduced \bar{y}_i to denote the mean of all the responses that share the same covariate value x_i . This last sum is $SSE(ANOVA)$ + a term that measures differences between means and regression fits (a lack-of-fit term).

- (d) $F_{lack-of-fit} = ((SSE(Reg) - SSE(ANOVA))/3)/MSE(ANOVA) = (2.76/3)/0.016 = 57.5$. This can be compared to the F-distribution with 3 d.f. in the numerator and 10 d.f. in the denominator which yields $P < .0001$. Once again there is strong evidence to reject H_o (no lack of fit) and hence we conclude that there is a lack of fit.

5. Theory behind residual analysis

- (a) To minimize we take derivatives with respect to β_o and β_1 . The first yields $-2 \sum_i (Y_i - \beta_o - \beta_1 X_i) = 0$ which means that at the solution $\sum_i e_i = 0$. The second yields $-2 \sum_i X_i (Y_i - \beta_o - \beta_1 X_i) = 0$ which means that at the solution $\sum_i e_i X_i = 0$.

- (b) Clearly then $\sum_i e_i \hat{Y}_i = \sum_i e_i (\hat{\beta}_0 + \hat{\beta}_1 X_i) = \hat{\beta}_0 \sum_i e_i + \hat{\beta}_1 \sum_i e_i X_i = 0$
- (c) This one is a bit trickier. Start by noting that $\sum_i e_i (Y_i - \bar{Y}) = \sum_i e_i Y_i - \bar{Y} \sum_i e_i = \sum_i e_i Y_i$ as in the hint. Then $\sum_i e_i = \sum_i e_i (Y_i - \bar{Y}) = \sum_i (Y_i - \bar{Y})(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = \sum_i (Y_i - \bar{Y})(Y_i - \bar{Y} - \hat{\beta}_1 (X_i - \bar{X}))$ where the last equality uses the definition of $\hat{\beta}_0$ as suggested by the hint. Now expanding the last expression gives $\sum_i (Y_i - \bar{Y})^2 - \hat{\beta}_1 \sum_i (Y_i - \bar{Y})(X_i - \bar{X})$. If we introduce notation $s_{yy} = \sum_i (Y_i - \bar{Y})^2$ (and similarly for s_{xx}) and $s_{xy} = \sum_i (Y_i - \bar{Y})(X_i - \bar{X})$, then $\sum_i e_i Y_i = s_{yy} - (s_{xy}/s_{xx}) * s_{xy} = s_{yy} - (s_{xy}^2/s_{xx}s_{yy}) * s_{yy} = (1 - r^2)s_{yy}$.
- (d) The interpretation here is important. The results of (a) and (b) show that under the assumptions of our model the residuals e would be uncorrelated with the predicted values. Thus we don't expect a pattern. The result of (c) shows that the residuals would be correlated with the Y 's and so even if the model is true there will be a pattern in this plot. This is actually not very surprising; you will tend to have big residuals when Y is large! That's what part (c) showed.

6. Regression theory - no intercept

- (a) If Y is income and X is hours per week of employment than we might expect there to be no intercept (0 hours of work implies 0 income).
- (b) We want to minimize $\sum_i (y_i - \beta x_i)^2$. Taking the derivative with respect to β and setting equal to zero yields $-2 \sum_i x_i (y_i - \beta x_i) = 0$ which leads to $\hat{\beta} = \sum_i x_i y_i / \sum_i x_i^2$.
- (c) $E(\hat{\beta}) = \sum_i x_i E(y_i) / \sum_i x_i^2 = \sum_i \beta x_i^2 / \sum_i x_i^2 = \beta$ so this estimator is unbiased.
 $Var(\hat{\beta}) = \frac{1}{(\sum_i x_i^2)^2} \sum_i x_i^2 Var(y_i) = \sigma^2 / \sum_i x_i^2$.

7. Correlation

- (a) The correlation of payroll and wins, using the summary statistics from Problem 1 is $r = 5273 / \sqrt{33147.37 * 3786} = 0.47$.
- (b) Fisher's Z -transformation is $Z_r = 0.5 \log(1.47/0.53) = 0.51$. In statistics we always use the natural logarithm! The 95% CI for Z_ρ is $0.51 \pm 1.96 * \sqrt{1/27} = (0.13, 0.89)$. Back transforming yields CI for ρ equal to $(0.13, 0.71)$.
- (c) Let Z_{r_1} and Z_{r_2} denote the Fisher Z -transformations of the two sample correlations. Then $Z_{r_1} - Z_{r_2} \sim N(Z_{\rho_1} - Z_{\rho_2}, 1/(n_1 - 3) + 1/(n_2 - 3))$ and this result can be used to derive a test or CI for the difference in two Fisher Z -transformed correlations. Some people treated this as a one-sample question; they compared the observed $r = 0.47$ with $H_o : \rho = 0.7$ but this ignores the fact that the 0.7 value comes from a sample. There are two populations and samples. The null hypothesis is that $\rho_1 = \rho_2$ or equivalently that $Z_{\rho_1} = Z_{\rho_2}$ (the null hypothesis is not about r_1 and r_2 as these are observed quantities!). For the test we calculate $Z = (Z_{r_1} - Z_{r_2}) / \sqrt{1/(n_1 - 3) + 1/(n_2 - 3)} = (0.51 - 0.87) / \sqrt{2/27} = 1.32$. Then using the normal distribution (two-tailed) we find $p \approx .19$ and conclude there is no significant difference. Note that 0.70 is just inside the CI for the current data so the test result is not a surprise. You might be surprised that the p -value is so big (i.e., not even close to significance) but this happens because each of the values (.47 this year, .70 previously) are sample correlations and are uncertain measurements of the true population correlations. When we adjust for the variability in both we see that there is little evidence of a difference.