1. Correlation and regression - baseball

(a) $\hat{\beta}_1 = 8624.2/62152.2 = 0.139$ and $\hat{\beta}_0 = (2429/30) - (0.139) \ast (3965/30) = 62.63$. The fitted regression line is wins $= 62.63 + 0.139$ pay.

(b) The s.e.($\hat{\beta}_1$) is $\sqrt{\sum_i(x_i - \bar{x})^2} = \sqrt{6134.3/28/62152.2} = .0593$ and the resulting $95\%$ CI is $0.139 \pm t_{28,.975}.0593 = (.017, .260)$.

(c) First note that the expected number of wins for a team with payroll $180$ is $62.63 + .139(180) = 87.6$ (recall this is an average over the “population” of teams with payroll $180$). The s.e. for this expected response is $\sqrt{\sigma^2 \ast (1/n + (x_i - \bar{x})^2)/\sum_i(x_i - \bar{x})^2} = \sqrt{(6134.3/28) \ast (1/30 + (180 - 132.167)^2/62152.2)} = 3.92$. Then the $95\%$ CI is $87.6 \pm t_{28,.975}.95$. If we want to make a prediction for a specific team with payroll $180$ we use the same estimate ($87.6$) but now the s.e. is $\sqrt{\sigma^2 \ast (1/n + (x_i - \bar{x})^2)/\sum_i(x_i - \bar{x})^2} = \sqrt{(6134.3/28) \ast (1 + 1/30 + (180 - 132.167)^2/62152.2)} = 15.31$. Notice how important that extra term is!! The $95\%$ PI for the number of wins for a team with payroll $180$ is $87.6 \pm t_{28,.975}.15.31 = (56.2, 119.0)$. 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. There is no reason to believe these are outliers. 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 .16 so about $84\%$ of the variability in team performance is explained by other factors.)

2. Correlation and regression - baseball

(a) $r = \frac{1}{n-1} \sum_i(x_i - \bar{x})(y_i - \bar{y})/s_x s_y = 8624.2/\sqrt{7331.0 \ast 62152.2} = .404$.

(b) Fisher’s $z$-transform of $r$ is $\frac{1}{2} \ln \frac{1 + r}{1 - r} = .428$. It has approximate standard error $\sqrt{\frac{1}{27}} = .192$. Thus a $95\%$ CI for the Fisher transform of $\rho$ is $.428 \pm 1.96 \ast .192 = (.051, .806)$. Doing the reverse transformation yields a $95\%$ confidence interval for $\rho$ of $(.05, .67)$.

3. Simple linear regression in R

(a) The correlation is $0.53$ which suggests a moderately strong linear pattern.

(b) The plot shows a relationship between stay and inf. It increases linearly at first but there are a couple of unusual values on the right-hand side of the plot that indicate the linear relationship doesn’t work through the entire range.

(c) The least squares regression line is inf $= .744 + .374$*stay.

(d) The residual plot does not suggest any major violations. The plot of residuals vs fitted values does not have any obvious pattern and the normal probability plot is reasonably linear. Note that the two unusual points do stand out in the regression plot. (More on them below.)

(e) Hospitals with one additional day of average length of stay have expected infection risk $.374$ infections per $100$ patients higher compared to their peers.

(f) The $95\%$ CI for the slope is $.374 \pm t_{111,.975}.056 = (.262, .486)$.

(g) Note that each observation in the regression is the average data for the hospital. There is no basis for applying this regression to predict what will happen to an individual patient.

(h) The code below shows how to obtain the confidence interval and the prediction interval. The CI is $(4.27, 4.70)$.

(i) The PI is $(2.22, 6.75)$. The previous part gives a range of plausible values for the mean infection rate in the population of hospitals with average stay $10$ days whereas this part gives a range of plausible values for the infection rate in an individual hospital. The latter is much more uncertain because there can be much variation among hospitals having the same average length of stay.

(j) When you give a summary of the study intended for non-statisticians you should try to avoid statistical terminology such as ”confidence interval”, ”slope”, ”test”. Here’s a possible summary: A study of $113$ hospitals was used to assess factors associated with in-hospital infection risk. This analysis focuses on the effect of length of stay. Average length of patient stay in a hospital is positively associated with risk of infection in the hospital. For every additional day in the hospital’s average length of stay, the expected risk of infection increases by between $0.26$ and $0.49$ infections per $100$ patients. There is considerable variation across hospitals however as the residual standard error is $1.1$ infections per $100$ patients. The data contain two unusual hospitals with average lengths of stay considerably larger than the other $111$ hospitals; it would be valuable to determine if these two hospitals are really part of the same population as the others.
4. Regression and the t-test

(a) First change the definition of \( m \) so that we use \( m' = 2 - m \) (\( m' \) is 1 if there is a medical school and 0 if there is not). The least squares regression line is \( \inf = 4.224 + .870 m' \).

(b) The mean infection rate in hospitals without a medical school is 4.224 per 100 patients. The mean infection rate in hospitals with a medical school is 5.094 per 100 patients. The \( t \)-statistic is -2.525 which yields a two-sided p-value of .013. A reminder that the \texttt{t.test()} function in R does the unpooled procedure by default; if you want the pooled procedure you need to use the \texttt{var.equal=TRUE} option.

(c) Note that the t-statistics are equal to each other.

(d) If the predictor is a 0/1 variable (no med school / med school) then we can let

\[
e = \begin{cases} 1 & \text{if med school} \\ 0 & \text{otherwise} \end{cases}
\]

\( n \) of .013. A reminder that the \texttt{t.test()} function in R does the unpooled procedure by default; if you want the pooled procedure you need to use the \texttt{var.equal=TRUE} option.

\[n \] and we can see that the regression analysis replicates the usual t-test.

5. 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 confidence interval is \(-.324 \pm t_{13.975}.043 = (-.417, -.231)\) which suggests that we are very confident there is a decrease in concentration over time.

(b) The ANOVA SS(\text{Error}) describes individual sample variation among repeated observations with the same time period (regardless of whether the linear model is correct or some other function of time would be better). This is what we usually mean when we think of “pure error” or “individual variation”. By contrast the regression SS(\text{Error}) will also include any failure of the linear model (e.g., a quadratic pattern). You can see this analytically by noting that

\[ \text{SSE(Reg)} = \sum_{i}(y_i - \hat{y})^2 = \sum_{i}(y_i - \bar{y})^2 + \sum_{i}(\bar{y} - \bar{y})^2 \]

where we have introduced \( \bar{y} \) to denote the mean of all the responses that share the same covariate value \( x_i \). This last sum is equal to the \( \text{SSE(ANOVA)} \) (first term) plus a second term that measures differences between means and regression fits (a lack-of-fit term).

(c) \( F_{\text{Lack-of-fit}} = (\text{SSE(Reg)} - \text{SSE(ANOVA)})/3)/MSE(\text{ANOVA}) = (2.76/3)/0.016 = 57.5. \) This can be compared to the F-distribution with 3 d.f. in the numerator and 10 d.f. in the denominator which yields \( P < .0001 \). There is strong evidence to reject \( H_0 \) (no lack of fit) and hence we conclude that there is a lack of fit. Note that you can also see this if you plot the data (or examine the data). The concentration drops very quickly at first and then slows down so it is not a linear pattern throughout the range of times.

6. Theory behind residual analysis

(a) To minimize we take derivatives with respect to \( \beta_0 \) and \( \beta_1 \). The first yields

\[-2 \sum_i (Y_i - \beta_0 - \beta_1 X_i) = 0 \]

which means that at the solution \( \sum_i = 0 \). The second yields

\[-2 \sum_i X_i (Y_i - \beta_0 - \beta_1 X_i) = 0 \]

which means that at the solution \( \sum_i = 0 \).
(b) Clearly then \( \sum_i e_i \hat{Y}_i = \sum_i e_i (\hat{\beta}_o + \hat{\beta}_1 X_i) = \hat{\beta}_o \sum_i e_i + \hat{\beta}_1 \sum_i e_i X_i = 0 \)

(c) 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 and any pattern we do say is an indication of a problem. The result given at the start of this part 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! This makes the latter plot (residuals vs \( Y \)) not very helpful.