Assignments handed in at the beginning of lecture on Wednesday, November 26 will be returned at the final lecture on December 3. Assignments handed in by 2:00 p.m. on Friday, November 28 will not be considered late. Assignments received on Monday, December 1 will be considered 2 days late. The deduction for late assignments is 10% of the total marks for the assignment for each day late. Poorly presented solutions will not receive full marks.

1. (Source: Ramsey and Schafer)

The U.S. Presidential Election of November 7, 2000 was one of the closest in history. As returns were counted on election night, it became clear that the outcome in the state of Florida would determine the next president. At one point in the evening, television networks projected that the state was won by Al Gore, but a retraction of the projection followed a few hours later. Then, early in the morning of November 8, the networks projected that George W. Bush had carried Florida and won the presidency. Gore called Bush to concede. While on route to his concession speech, the Florida count changed rapidly in his favour. The networks once again reversed their projection, and Gore called Bush to retract his concession. When the roughly 6 million Florida votes had been counted, Bush was shown to be leading by only 1,738, and the narrow margin triggered an automatic recount. The recount, completed in the evening of November 9, showed Bush’s lead to be less than 400.

Meanwhile, angry Democratic voters in Palm Beach County in Florida complained that a confusing “butterfly” lay-out ballot caused them to accidentally vote for Pat Buchanan instead of Gore. Two pieces of evidence supported this claim: Buchanan had an unusually high percentage of the vote in that county, and an unusually large number of ballots (19,000) were discarded because voters had marked two circles (possibly by inadvertently voting for Buchanan and then trying to correct the mistake by then voting for Gore).

The data file on the course web site for this question has the number of votes for Bush and the number of votes for Buchanan for each of the 67 counties in Florida. The counties are coded 1-67; Palm Beach county is 50.

(a) Construct a scatterplot with the number of votes for Buchanan on the vertical axis and the number of votes for Bush on the horizontal axis? What evidence is there in the plot that Buchanan received more votes than expected in Palm Beach County?

(b) Analyze the data without Palm Beach County (clearly an outlier that is probably also influential) to obtain an equation predicting Buchanan votes from Bush votes. Make sure your model satisfies the assumptions of simple linear regression. If it does not, carry out appropriate remedial measures. (One way to remove Palm Beach from the analysis is to set the value for Buchanan’s vote in Palm Beach to missing. The SAS missing value code is a period. So you could put the line

```
if county=50 then buchanan=.;
```

in the data step.)

(c) Obtain a 95% prediction interval for the number of Buchanan votes in Palm Beach county from your model in (b), assuming that the relationship is the same in this county as in the others. How many standard errors of prediction is the actual value from the predicted value? If it is assumed that Buchanan’s actual count contains a number of votes intended for Gore, what can be said about the likely size of this number from the prediction interval? (Make sure you know why a prediction interval is the appropriate interval for this question but don’t include this as part of your solution.)
There have been many analyses of the vote from the 2000 election, most focussing on whether or not the result in Palm Beach county is an anomaly. One compelling analysis, in addition to looking at the relationship between the number of votes for Buchanan and the number of votes for Bush, considered several demographic variables for the Florida counties that might be related to how the vote in that county is split amongst the candidates. The data file is on the course web site. It contains the following variables:
- County Number (1-67) (Palm Beach is 50)
- Latitude (degrees north)
- Longitude (degress west)
- Population (1997)
- Percentage of whites (1996)
- Percentage of blacks (1996)
- Percentage of Hispanics (1996)
- Percentage of population over 65 years
- Percentage completing high school education
- Mean income per person
- Votes for Bush
- Votes for Gore
- Votes for Buchanan

An initial examination of the data indicated that the analysis should be carried out with the following transformations of the predictor variables: log of the percentage of blacks, log of the percentage of Hispanics, and log of the population. (No other predictor variables should be transformed.)

(a) Fit a multiple regression model with a suitable transformation for the votes for Buchanan (as determined in (a)) for the dependent variable, and all other variables except latitude and longitude as the independent variables. Using 0.05 as a cutoff for whether individual tests are statistically significant, you will notice that many of the predictor variables have coefficients which are not statistically significantly different from 0. Your goal will be to construct a model where all variables have a statistically significant relationship with the predictor variable. One possible method is to use a backwards elimination technique. First remove the predictor variable which contributes least to the model (has the largest p-value) and re-fit a regression model. Repeat until all remaining variables have coefficients which are statistically significantly different from 0. What is the resulting fitted regression equation? Outline the steps you took. You should not include Palm Beach county in this analysis.

(b) Your final model in (a) should not include both the Bush and Gore votes. Is it surprising that one is a statistically significant predictor of the Buchanan vote but the other is not? Why or why not?

(c) Obtain a 95% prediction interval for the number of Buchanan votes in Palm Beach county from your final model in (a), assuming that the relationship is the same in this county as in the others. How many standard errors of prediction is the actual value from the predicted value?

(d) Is there a geographic pattern to the data? Construct plots of the residuals and all other variables in your final model in (a) versus latitude and longitude. What do these plots tell you? (Omit the outlier of Palm Beach county.)
3. We have seen the use of transformations as a remedial tool when the assumptions of the linear regression model are violated. Here is an example where the use of transformations allows us to use linear regression to estimate the parameters of a model that is non-linear.

In economics, a production function is a function that relates various inputs such as capital investment and labour to the production level of an industry or a national economy. The Cobb-Douglas model for production was developed in 1928, based on U.S. economic data collected annually from 1899 to 1922. 1899 is used as a base year, for which the units of labour, capital and production were all set at 100. All other measurements were normalised to that standard. The data are available on the course web site.

Based on some economic assumptions, Cobb and Douglas surmised that production can be modeled by the function

$$P_t = \alpha K_t^{\beta_1} L_t^{\beta_2} \eta_t$$

where the subscript $t$ indicates year, $P_t$ is production, $K_t$ is capital, $L_t$ is labour, and $\eta_t$ is an error term. (Note that the error is multiplicative rather than additive.) We will assume that $\text{E}[\log(\eta_t)] = 0$, $\text{Var}[\log(\eta_t)]$ is a constant, and the errors are independent. We must determine which constants $\alpha$, $\beta_1$, and $\beta_2$ best fit the data.

(a) Use the data for 1899 to 1922 to estimate the constant parameters in the model. Make sure your answer is an estimate of each parameter, and not a transformation of it. (Don’t worry about the error variance.)

(b) If $\beta_1 + \beta_2 = 1$ we say that production has “constant returns to scale”. What happens to production if both labour and capital are doubled for the model with constant returns to scale?

(c) If $\beta_1 + \beta_2 < 1$ we say that production has “decreasing returns to scale”, and if $\beta_1 + \beta_2 > 1$ we say that production has “increasing returns to scale”. If you were an economist, would you assume that production has constant, increasing or decreasing returns to scale based on your answer to (a)? (You don’t need to do any statistical analysis to answer this question.)

(d) Suppose you believe a priori that production has constant returns to scale. Estimate $\beta_1$ and $\beta_2$ under this constraint.

(e) Sometimes the model

$$P_t = \alpha \gamma^t K_t^{\beta_1} L_t^{\beta_2} \eta_t$$

is considered where $\gamma^t$ is assumed to account for technological development. Estimate the constant parameters in this model. Assume constant returns to scale.

(f) Look at the $p$-values for the tests for whether the estimated parameters in (d) and (e) are zero. Are there any apparent contradictions? If so, why did they happen?