Quantitative Methods II
Assignment 9
October 23, 2011
Solutions

Problem: Which Coin is Fair? [[10]]

This problem simply has you fit the data (bivariate) with a binary dependent variable model with your choice of a link function. The default is the logit link and is (perhaps) the best in this case. Thus, the research model is

$$\texttt{head} \sim \texttt{trial}$$

In fitting this with the logit link, you will get raw results similar to those at the bottom of the page.

⋆ ⋆ ⋆

At this point, there are a couple methods for finding the answer. For those who are mathematically inclined, I offer the following:

A fair coin has a $\mathbb{P}\left[Head\right] = 0.500$. The $\text{logit}(0.500) = 0$. Thus, we solve the regression equation for zero:

$$0 = -2.2929 + 0.0345 \times c$$

This gives us $c = 66.46$, which indicates that Coin 66 is closest to being fair.

	Estimate	Std. Error	z-value	p-value
Constant Term	-2.2929	0.5384	-4.26	$\ll 0.0001$
Trial Number	0.0345	0.0087	3.97	0.0001

Alternatively, for those who like the `predict()` function (as do I), we can run the following line of code

```
predict(m1, newdata=data.frame(trial=seq(1,100)))
```

This lists out the predicted success probability for each of the 100 coins. Then, you just select the one closest to having a percent of 0.500. This also gives Coin 66.

Note. *Had you chosen a different link function, you would have gotten a different answer. For instance, using the complementary log-log link gives you Coin 85 as the fair coin.*

Since I manufactured this data, I know the real answer. The coin that was actually fair was Coin 71. The lesson? This is all statistics, we cannot know the truth, we can only approximate it as best we can.

Problem: It Probably will Pass, Right? ⟦15⟧

Homework has a variety of purposes. Some homework is assigned to give practice in applying what we did in class this week. Other homework is assigned to have you think and explain, more than do and write. This problem was a "think and explain" problem. In this problem, you were to use the `ssm3` data file to predict the probability of the SSM ballot measure passing in Washington.

Note that this question is the same as the question from Problem 5. However, the method to arrive at an answer is quite different. In Assignment 5, you predicted the vote proportion in favor of the bill, then used monte Carlo techniques to estimate the probability that the measure will pass. Here, your dependent variable was whether the bill passed. Thus, if we use logistic regression, we are directly predicting the probability that the ballot measure will pass. But, there was a hitch in the data:

There was almost no variation in the dependent variable.

Recall from Chapter 4, when we were first exposed to the assumptions of OLS. One of those assumptions was that there was variation in the independent variable. Why was that assumption important? What were the effects of violating it? It was important so that we could determine the effects of each of the independent variables. If that assumption was violated, then there would be no way to determine the effect of the x-variable on the y-variable. (Why?)

The same assumption is required for the dependent variable. (Why?) This data set demonstrates what your results look like when there is little (if any) variation in the dependent variable: standard errors are large and statistical significance is missing — for each of your independent variables.

When an x-variable had little variation, we could not estimate the effect of *that* variable — others could be estimated. When there is little variation in the dependent variable, the *entire* model suffers.