TA Notes - Week 9: Question 7.2 & 7.14

Conor Foley*

Econ 103 (Randall Rojas) Winter 2018

Contents

1 Overview

Topics Covered:

- Interpreting beta estimates (7.2A, 7.14C)
- Omitted Variable Bias (7.2B)
- Model Selection (7.2B, 7.2C)
- F-test (7.2C)
- Dummy Variables (7.14A)
- Binary Variables (7.14B)
- Fitted Values (7.14D, 7.14E, 7.14F)

*For errors or corrections, please email me at conor.teaches.econ@gmail.com.

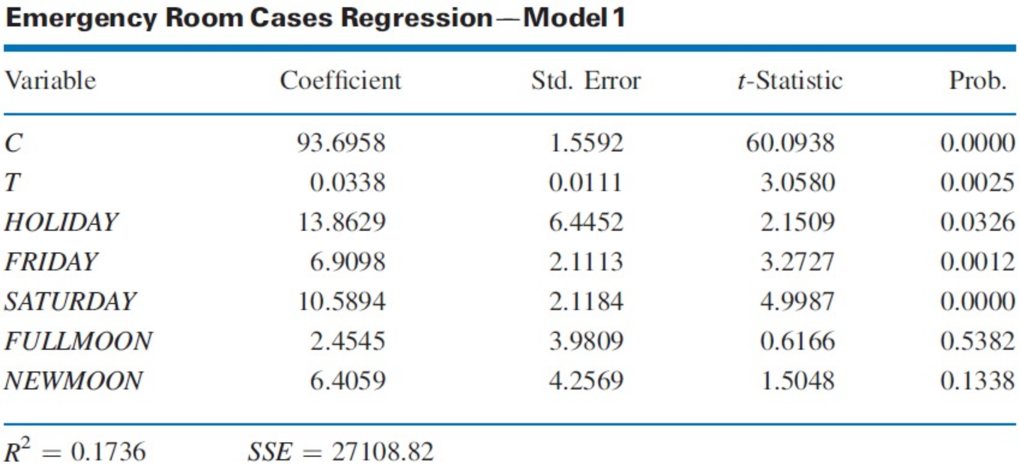
2 Question 7.2

2.1 7.2 Part A - Interpreting beta coefficients

You are given a regression with 229 observations, covering emergency room cases in a local hospital from January 1, 1998 until mid-August of that year. During that time, there were eight full moons and seven new moons.

The data include a time trend $T = 1, 2, 3, ..., 229$ and the rest are indicator variables. The indicators are: HOLIDAY, FRIDAY, SATURDAY, FULLMOON, and NEWMOON. All the indicators are 1 if the description for that day is valid, and 0 otherwise.

Using this data, we get the following results:



Emergency Room Cases Regression—Model 1

Variable	Coefficient	Std. Error	*t*-Statistic	Prob.
C	93.6958	1.5592	60.0938	0.0000
T	0.0338	0.0111	3.0580	0.0025
HOLIDAY	13.8629	6.4452	2.1509	0.0326
FRIDAY	6.9098	2.1113	3.2727	0.0012
SATURDAY	10.5894	2.1184	4.9987	0.0000
FULLMOON	2.4545	3.9809	0.6166	0.5382
NEWMOON	6.4059	4.2569	1.5048	0.1338

$R^2 = 0.1736$ $SSE = 27108.82$

Interpret these regression results. When should emergency rooms expect more calls?

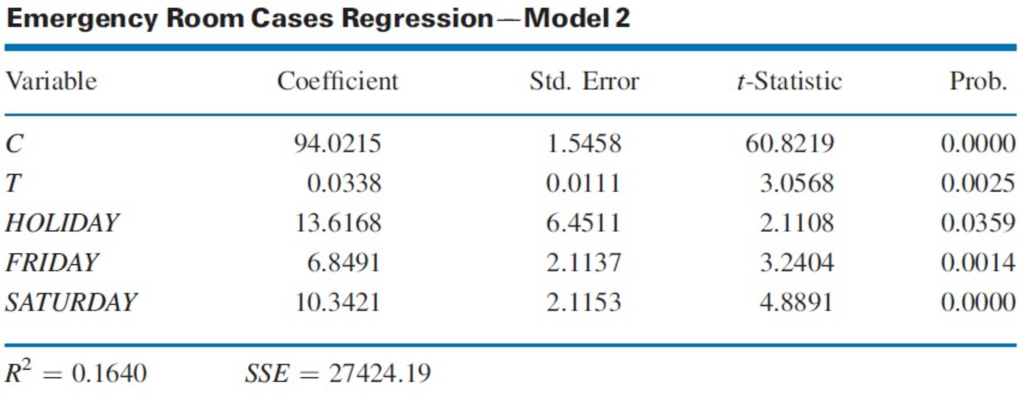
The time trend would control for whether the number of cases is increasing or decreasing over time in a smooth way. While the point estimate is statistically significant, it is not very meaningful because it only shifts the expected number of visits between the beginning of the year and August by 3 calls.

Holidays and Saturdays both see the largest spikes in calls. Holidays see 13.86 more calls than otherwise similar days, while Saturdays sees 10.59more calls than other days of the week. Friday is also somewhat higher, with 6.91 more calls than other days of the week.

Fullmoon and Newmoon days are also estimated to more calls than other days, but the estimates have lower significance than those previously discussed.

2.2 Part B - Remove FULLMOON and NEWMOON

The model was reestimated omitting the variables FULLMOON and NEWMOON, as shown below. Comment on any changes you observe.



Emergency Room Cases Regression—Model 2

Variable	Coefficient	Std. Error	*t*-Statistic	Prob.
C	94.0215	1.5458	60.8219	0.0000
T	0.0338	0.0111	3.0568	0.0025
HOLIDAY	13.6168	6.4511	2.1108	0.0359
FRIDAY	6.8491	2.1137	3.2404	0.0014
SATURDAY	10.3421	2.1153	4.8891	0.0000

$R^2 = 0.1640$ $\quad$ $SSE = 27424.19$

The point estimates and standard errors for the non-FULLMOON and non-NEWMOON estimates are largely unchanged. Besides the constant, the magnitude of shift is on the over of 0.03 calls. The shift in the standard errors is also modest, with some moving up slightly and other moving down slightly. This suggests that FULLMOON and NEWMOON are only slightly correlated with the other indicators variables.

2.3 Part C - F-test for FULLMOON and NEWMOON

Test the joint significant of FULLMOON and NEWMOON. State the null and alternative hypotheses and indicate the test statistic you use. What do you conclude?

Let's refer to the estimated coefficients for FULLMOON and NEWMOON as b_f and b_m, respectively. In this case, the null hypothesis is:

$$b_f = b_m = 0$$

The alternative is that at least one of these restrictions is violated. To list out the alternatives, we have:

$$\begin{array}{ll} b_f = 0 & b_m \neq 0 \\ b_f \neq 0 & b_m = 0 \\ b_f \neq 0 & b_m \neq 0 \end{array}$$

The test statistic we want to use is the F-statistic. We have two restrictions so we have $J = 2$ numerator degrees of freedom. The "unrestricted" model is the model including FULLMOON and NEWMOON, meaning the $N-K$ degrees of freedom are $229-7 = 222$.

We have two versions of the F-stat that are equivalent:

$$F = \frac{(SSE_r - SSE_u)/J}{SSE_u/(N-K)}$$

Plugging in the reported SSE values and our degrees of freedom, we have:

$$\begin{aligned} F &= \frac{(27424.19 - 27108.82)/2}{27108.82/222} \\ &= 1.291317 \end{aligned}$$

Next, using the tables in the textbook, we have 2 numerator (column) degrees of freedom and 222 denominator (row) degrees of freedom. The tables in the textbook only go up to 120 denominator (row) degrees of freedom, so we can compare the 120 and the ∞ values in the 2 column. The tables show us that we need an F-stat of about 3.00 to 3.07 to reject the null at the $\alpha = 5\%$ level and about 4.61 to 4.79 to reject at the $\alpha = 1\%$ level. Clearly, our F statistic is too small to reject at these levels.

We can also use the relationship between the $F(d_1, d_2)$ distribution and the $\chi^2(d_1)$ distribution to try to approximate the $\alpha = 10\%$ critical value. Specifically, the inverse CDF of an F is related to the inverse CDF of the χ^2 distribution as follows:

$$CDF_F^{-1}(d_1, \infty, p) = \frac{1}{d_1} CDF_{\chi^2}^{-1}(d_1, p)$$

Using this expression and the χ^2 tables in the textbook, we can calculate the $\alpha = 10\%$ critical value as being approximately $4.605/2 = 2.3025$. Once again, our F stat is too small and we fail to reject at the $\alpha = 10\%$ level.

We conclude that we cannot reject the null that $b_f = b_m = 0$. Our p-value is larger than 0.1 (using STATA, we can see that the p-value is 0.2770).

Note, if we did not have information on the SSE we could also calculate the F stat using the R^2 values. While the values match exactly when we do the calculation in STATA, using the values provided in the textbook will give a slightly different answer due to rounding. The R^2 calculation (based on the values in the textbook tables) gives us:

$$\begin{aligned} F &= \left[\frac{1 - R_r^2}{1 - R_u^2} - 1\right] \frac{N-K}{J} \\ &= \left[\frac{1 - 0.1640}{1 - 0.1736} - 1\right] \frac{222}{2} \\ &= 1.289448 \end{aligned}$$

3 Question 7.14

3.1 7.14A - Interpret Dummy (=0 or =1) Variables

Part A - Consider the regression model:

$$\begin{aligned}\mathbf{VOTE} = \boldsymbol{\beta_1} + \boldsymbol{\beta_2}\mathbf{GROWTH} + \boldsymbol{\beta_3}\mathbf{INFLATION} + \boldsymbol{\beta_4}\mathbf{GOODNEWS}\\ + \boldsymbol{\beta_5}\mathbf{PERSON} + \boldsymbol{\beta_6}\mathbf{DURATION} + \boldsymbol{\beta_7}\mathbf{PARTY} + \boldsymbol{\beta_8}\mathbf{WAR} + \boldsymbol{e}\end{aligned}$$

Discuss the anticipated effects of the dummy variables PERSON and WAR.

The definition of PERSON is that PERSON = 1 if a sitting president is running in the election. Given that incumbents tend to do better than other candidates, I would expect β_5 to be positive.

The definition of WAR is that WAR = 1 if the election is around the world wars (1920, 1944, and 1948). We would expect β_8 to be positive if these periods saw a strong "rally around the flag" effect in support of the incumbent party. On the other hand, it would be negative if we think there is war weariness and people want a change in leadership.

3.2 7.14B - Interpret Binary (=1 or =-1) Variable

Part B - The binary variable PARTY is somewhat different from the dummy variables we have considered. Write out the regression function E(VOTE) for the two values of PARTY. Discuss the effects of this specification.

Rather than being 0 or 1, the PARTY variable can be 1 (if the incumbent party is Democratic) or -1 (if the incumbent party is Republican). The effect of this is that we ADD β_7 when the incumbent party is Democratic and we SUBTRACT β_7 when the incumbent party is Republican.

Writing out the full equation, we have:

$$\begin{aligned}\mathbb{E}\left[\text{VOTE} \mid \text{Democrat}\right] &= \beta_1 + \beta_2\mathbf{GROWTH} + \beta_3\mathbf{INFLATION} + \beta_4\mathbf{GOODNEWS}\\ &\quad + \beta_5\mathbf{PERSON} + \beta_6\mathbf{DURATION} + \beta_7 + \beta_8\mathbf{WAR}\\ \mathbb{E}\left[\text{VOTE} \mid \text{Republican}\right] &= \beta_1 + \beta_2\mathbf{GROWTH} + \beta_3\mathbf{INFLATION} + \beta_4\mathbf{GOODNEWS}\\ &\quad + \beta_5\mathbf{PERSON} + \beta_6\mathbf{DURATION} - \beta_7 + \beta_8\mathbf{WAR}\end{aligned}$$

Note that bar notation indicates a conditional expectation. For example, the expression $\mathbb{E}\left[\text{VOTE} \mid \text{Democrat}\right]$ can be read as "the expectation of VOTE conditional on the incumbent party being the Democrats."

3.3 7.14C - Run Model, Evaluate Output

Part C - Use the data for the period 1916-2004 to estimate the proposed model. Discuss the estimate results. Are the signs as expected? Are the estimates sta-

tistically significant? How well does the model fit the data?

All coefficients are significant at the 10% level, while all coefficients except WAR are significant at the 5% level. The variables GROWTH, GOODNEWS, and PARTY are significant at the 1% level. Overall, the model has a very high R^2 of 0.9052, suggesting it does a good job of explaining the data.

All the betas have reasonable estimates. In particular, growth in the election year (GROWTH) and growth in the preceding 3-4years (GOODNEWS) benefit the incumbent party, while inflation in the preceding 3-4 years (INFLATION) hurts the incumbent party.

The negative coefficient on DURATION indicates that parties do worse when they have been in power for a long time. However, the interpretation of this variable is slightly complicated since it is coded in an odd way. A value of 1 indicates that the incumbent party has been in power for 2 consecutive terms (e.g. Bush Sr. in 1998, Al Gore in 2000, or McCain in 2008), and 1.25 if the party has been in power for 3 consecutive terms (e.g. Bush Sr. in 1992).

The positive coefficient on PERSON indicates that when the sitting president runs they do better than other candidates in similar circumstances.

The coefficient on WAR indicates that the 1920, 1944, and 1948 elections saw strong performances for the incumbent (Democratic) party in those years (controlling for other factors). However, the coefficient on WAR has the lowest significance of any of the explanatory variables and has a borderline p-value of 0.054.

The coefficient on PARTY indicates that Democrats tend to run 5.35 points (2 × 2.64) lower than a Republican would in similar circumstances.

3.4 7.14D - Predict 2008 Vote

Part D - Predict the outcome of the 2008 election using the given 2008 data for values of the explanatory variables. Based on the prediction, would you have picked the outcome of the election correctly?

Note that in STATA, we keep the estimates from the 1916-2004 regression to calculate the fitted value for 2008. Thus, we can say that the 2008 estimate is an “out-of-sample” estimate because our regression estimate did not include 2008 data. If you are focused on the forecast performance of the model, doing successive “out-of-sample” forecasts is generally a best-practice because it simulates how the model would have performed if we used it in 2008 (when we wanted to predict the outcome).

Based on estimates calculated in STATA, we would have predicted an incumbent vote share of 48.1%. This would have led us to guess that the incumbent (Republican) party would lose the election, which turned out to be the actual outcome. However, we had a modest miss on the actual vote which actually turned out to be 46.6% in the actual 2008 election.

3.5 7.14E - Calculate Confidence Interval for 2008 Fitted Value

Part E - Construct a 95% confidence interval for the outcome of the 2008 election.

We let STATA calculate the STDF for us. Then using the formula:

$$\text{Confidence Interval} = \widehat{\text{VOTE}} \pm t^c s.e.(f)$$

There are 23 observations - every 4th year from 1916 to 2004, inclusive, or $(2004 - 1916)/4 + 1$. We also estimated 8 beta coefficients, giving us 23-8 = 15 degrees of freedom. Since we are looking for the 95% confidence interval ($\alpha = 0.05$), and a confidence interval is equivalent to a 2-sided test, we want to look at the $t_{0.975}$ column ($0.975 = 1 - (0.05/2)$). Given the degrees of freedom and our confidence level, we have: $t^c = 2.131$.

Using the $\widehat{\text{VOTE}} = 48.1$ from earlier and the $s.e.(f) = 2.814966$ calculated by STATA, the confidence interval is:

$$[42.09,\ 54.09]$$

3.6 7.14F - Estimate a 2012 Fitted Value (pick x values)

Part F - Using data values of your choice (you must explain them), predict the outcome of the 2012 election.

We have 7 values to assign. Of these, there are 4 variables that are fixed characteristics of the election (that we would know as early as 2008):

- PARTY = 1 because the incumbent (Obama) is a Democrat
- PERSON = 1 because the incumbent (Obama) is running for reelection
- WAR = 0 because it is not World War I or World War II
- DURATION = 0 matching what is coded for the previous re-election runs (e.g. 1984, 1996, 2004)[1]

The remaining 3 variables have to do with economic conditions: GOODNEWS, INFLATION, and GROWTH. These are up to your discretion, or the scenarios that you are interested in evaluating.[2] Suppose we are sitting in 2008 and we forecast what growth and

[1] The variable label for DURATION says that it is the number of terms that the incumbent party has been in power. However, during the first reelection run for incumbent presidents, such as in years 1984, 1996, and 2004, the DURATION variable is coded as 0. The answer to this question given in the textbook sets this value equal to 0.

[2] I have tried to compare the data in the file to currently available and 2008 vintage BEA data on US GDP but I was not able to match what is in the data set. Thus, I just propose a few other values to test.

inflation will be like during the next four years. Presumably there will be low average inflation as we emerge from the recession, but there will be higher-than-average growth. During the year of 2012 we would expect growth to be back to normal, or about 2%. Let's use the following values:

- GOODNEWS = 4 (say 4 out of 15 quarters with high growth due to the recovery)
- INFLATION = 1.5 (inflation moves from negative back closer to 2% over the 1st term)
- GROWTH = 2 (by 2012, expect growth to return to close-to-normal level)

Finally, we re-estimate the model this time including the 2008 observation since we are imagining we are forecasting for 2012 and know the 2008 information.

With these values, the estimated vote share for Obama in 2012 is 52.5 with a confidence interval of 46.5 to 58.5.