Problem set 3

Jason Parker

Due: March 12th 2019 at 11:59pm

In this problem set, the data sets (except for Question 6) are available as tables in the `wooldridge2.db` file. Your solution should have 2 parts, the code you used (a `.R` file) and the text of your solution (any file format). Please format your code and text neatly in some way similar to the example that I have posted in box.

Question 1

A model that allows major league baseball player salary to differ by position is

$$\ln(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg + \beta_4 hrunsyr + \beta_5 rbisyr + \beta_6 runsyr + \beta_7 fldperc + \beta_8 allstar + \beta_9 frstbase + \beta_{10} scndbase + \beta_{11} thrdbase + \beta_{12} shrtstop + \beta_{13} catcher + u$$

where `outfield` is the base group.

1. State the null hypothesis that, controlling for other factors, catchers and outfielders earn, on average, the same amount. Test this hypothesis using the data in `mlb1` and comment on the size of the estimated salary differential.
2. State and test the null hypothesis that there is no difference in average salary across positions, once other factors have been controlled for.
3. Are the results from parts 1 and 2 consistent? If not, explain what is happening.

Question 2

Use the data in `gpa2` for this exercise.

1. Consider the equation

$$colgpa = \beta_0 + \beta_1 hsize + \beta_2 hsize^2 + \beta_3 hsperc + \beta_4 sat + \beta_5 female + \beta_6 athlete + u$$

   where `colgpa` is cumulative college grade point average, `hsize` is size of high school graduating class, in hundreds, `hsperc` is academic percentile in graduating class, `sat` is combined SAT score, `female` is a binary gender variable, and `athlete` is a binary variable, which is one for student-athletes. What are your expectations for the coefficients in this equation? Which ones are you unsure about?
2. Estimate the equation in part 1 and report the results in the usual form. What is the estimated GPA differential between athletes and nonathletes? Is it statistically significant?
3. Drop `sat` from the model and reestimate the equation. Now, what is the estimated effect of being an athlete? Discuss why the estimate is different than that obtained in part 2.
4. In the model from part 1, allow the effect of being an athlete to differ by gender and test the null hypothesis that there is no ceteris paribus difference between women athletes and women nonathletes.
5. Does the effect of sat on colgpa differ by gender? Justify your answer.

Question 3

Use the data in `loanapp` for this exercise. The binary variable to be explained is `approve`, which is equal to one if a mortgage loan to an individual was approved. The key explanatory variable is `white`, a dummy

variable equal to one if the applicant was white. The other applicants in the data set are black and Hispanic. To test for discrimination in the mortgage loan market, a linear probability model can be used:

$$approve = \beta_0 + \beta_1 white + other\ factors \ldots$$

1. If there is discrimination against minorities, and the appropriate factors have been controlled for, what is the sign of β_1?
2. Regress `approve` on `white` and report the results in the usual form. Interpret the coefficient on white. Is it statistically significant? Is it practically large?
3. As controls, add the variables `hrat`, `obrat`, `loanprc`, `unem`, `male`, `married`, `dep`, `sch`, `cosign`, `chist`, `pubrec`, `mortlat1`, `mortlat2`, and `vr`. What happens to the coefficient on `white`? Is there still evidence of discrimination against nonwhites?
4. Now, allow the effect of race to interact with the variable measuring other obligations as a percentage of income (`obrat`). Is the interaction term significant?
5. Using the model from part 4, what is the effect of being white on the probability of approval when $obrat = 32$, which is roughly the mean value in the sample? Obtain a 95% confidence interval for this effect.

Question 4

1. Use the data in `hprice1` to obtain the heteroskedasticity-robust standard errors for equation:

$$price = \beta_0 + \beta_1 lotsize + \beta_2 sqrft + \beta_3 bdrms + u$$

   Discuss any important differences with the usual standard errors.
2. Repeat part 1 for equation

$$\ln(price) = \beta_0 + \beta_1 \ln(lotsize) + \beta_2 \ln(sqrft) + \beta_3 bdrms + u$$

3. What does this example suggest about heteroskedasticity and the log transformation?

Question 5

Use the data set `gpa1` for this exercise.

1. Use OLS to estimate a model relating `colGPA` to `hsGPA`, `ACT`, `skipped`, and `PC`. Obtain the OLS residuals and fitted values.
2. In the regression of $\hat{u}_i$ on $\widehat{colGPA}$, $\widehat{colGPA}^2$, obtain the fitted values, say $\hat{h}_i$.
3. Verify that the fitted values from part 2 are all strictly positive. Then, obtain the weighted least squares estimates using weights $1/\hat{h}_i$. Compare the weighted least squares estimates for the effect of skipping lectures and the effect of PC ownership with the corresponding OLS estimates. What about their statistical significance?
4. In the WLS estimation from part 3, obtain heteroskedasticity-robust standard errors. In other words, allow for the fact that the variance function estimated in part 2 might be misspecified. Do the standard errors change much from part 3?

Question 6

Follow the steps as outlined here. These are not meant to be entirely straightforward. This is meant to be a learning experience with time series data. I've basically outlined the approach, but you have to follow the steps to create and model this data.

1 - 7 will be tested

1. Go online and search for daily bitcoin prices. Find and download a historical series of bitcoin prices going back to at least 2014.
2. Go to St Louis FRED. Find and download the S&P500 (`SP500`), the London bullion market price for gold in US dollars (`GOLDAMGBD228NLBM`), the US/Euro exchange rate (`DEXUSEU`), and the West Texas Intermediate spot price of oil (`DCOILWTICO`). These should all be available daily as well.
3. Merge all the data sets together (you can use either R or Excel or whatever).
4. Plot the series in R.
5. Use a naïve regression to find spurious correlations to the bitcoin price in the data set (e.g., regress the bitcoin price on the other series without any differencing to see if you find any interesting but total bullshit relationships).
6. Use the KPSS test to find how many differences each series takes to become stationary.
7. After taking differences, regress the bitcoin price on the other series. What relationships do you find now?
8. Remove all the data before 2017 where the bitcoin price starts to spike. Plot the new data. This is the data you are to use for the rest of the question.
9. Plot the ACF and PACF of the bitcoin price.
10. Fit various arima models to the bitcoin price. Which model fits best using the AIC?
11. Forecast the next 30 days of the bitcoin price and plot the forecast.
12. Plot the periodogram of the data. Do you see any seasonality in the data?
13. Fit a model where you regress the stationarity-transformed price on dummy variables for the different days of the week. Obtain the residuals from the model. Plot the periodogram of these residuals. Has the periodogram changed greatly? Do you think this transformation helps us to capture any seasonality in the data?
14. Using the AIC, select a VAR model which best captures the relationships between our 5 variables. What Granger causality relationships do you see between our prices?
15. Forecast the next 30 days of the prices using the VAR model. Compare your forecasts to one from the ARIMA model.

not on the test