**BUAN 6356.002**

Problem Set 3

Huyen Nguyen (Htn180001)

**Question 1:**

1. H0: β13= 0

H1: β13 ≠ 0

Other factors controlled

ln(salary) = 11.1295536 + 0.0584178 years + 0.0097670 gamesyr + 0.0004814 bavg

+ 0.0191459 hrunsyr + 0.0017875 rbisyr + 0.0118707 runsyr + 0.0002833 fldperc

+ 0.0063351 allstar - 0.1328008 frstbase - 0.1611010 scndbase + 0.0145271 thrdbase - 0.0605672 shrtstop + 0.2535592 catcher

se(β13-hat) = 0.1313128

t-statistic = 1.931

p-value = 0.05432

-> Reject the null hypothesis at 5% significant level

Controlling for the performance and experience variables, the estimated salary differential between catchers and outfielders = 100\*[exp(0.2535592) – 1] ≈ 28.9%, which is is considerable.

2. H0: β9= 0, β10= 0, β11= 0, β12= 0, β13= 0

Other factors controlled

F-statistic = 1.7774

p-value = 0.1168

-> Cannot reject null hypothesis at 10% significant level

3. The results in parts 1 and 2 are not consistent. We cannot reject the null hypothesis in part 2 because we are testing, along with the marginally significant catcher, several other insignificant variables (especially thrdbase and shrtstop, which has absolute t statistics below 1).

**Question 2:**

1. We expect the coefficients β3 < 0, β4 > 0, and β6 < 0 as the smaller the academic percentile in class, the better the student and the higher the SAT score; the more likely the student will have high GPA; and athletes may perform worse than other students in the class.

We're unsure about the coefficients of class size and whether males or females will have higher GPA.

2. colgpa = 1.241 - 5.685e-02 hsize + 4.675e-03 hsize^2 -1.321e-02 hsperc + 1.646e-03 sat + 1.549e-01 female + 1.693e-01 athlete

R-sq = 0.2925, df = 4130

An athelete is predicted to have 0.1693 points higher in GPA compared to nonathlete.

t-stat = 3.998 -> very significant

3. Drop sat from the model, we have:

colgpa = 3.0476980 - 0.0534038 hsize + 0.0053228 (hsize^2) - 0.0171365 hsperc + 0.0581231 female + 0.0054487 athlete

R-sq = 0.1885, df = 4131

Now, the athlete effect is much smaller (nearly 0). This happens because we didn't control the SAT score. When we control the SAT score, athletes do better than nonathletes. If we do not control for SAT score, there is not much difference.

4. We choose female nonathletes as the base group.

femath= female\* athlete

maleath=(1- female)\* athlete

malenonath=(1- female)\*(1- athlete)

colgpa = 1.396 - 5.680e-02 hsize + 4.670e-03 (hsize^2) - 1.321e-02 hsperc + 1.646e-03 sat + 1.751e-01 femath + 1.280e-02 maleath - 1.546e-01 malenonath

R-sq = 0.2925, df = 4129

-> Female athletes are predicted to have 0.1751 points in GPA compared to female nonathletes, other factors are fixed.

t-stat = 2.084, which is very significant at 5% confident level against 2-sided alternatives.

5. Add femsat=female\*sat to the equation, we have femsat coefficient = 0.0000512 and t-stat = 0.397 -> There's little evidence to say that the effect of sat on colgpa differ by gender.

**Question 3:**

1. If there is discrimination against minorities, and the appropriate factors have been controlled for, β1 > 0

2. approve = 0.70779 + 0.20060 white

R-sq = 0.04893, n = 1989

In the population of 1989 loan applicantions, an application submitted by a white applicant was 20.1% more likely to be approved than that of a nonwhite applicant.

t-stat = 10.11, and n = 1989

-> It's statistically significant and practically large

3. New coefficient on white = 0.128820, t-stat = 6.529 -> very significant

The coefficient has decrease by some margin as we controll other factors (much different from race) that can affect loan approval rates. However, there's still evidence of discrimination against nonwhites.

4. white\*obrate coefficient = 0.008088, t-stat = 3.531

The interaction term is significant. A white applicant is penalized less than a nonwhite applicant for having other obligations as a larger percent of income.

5. Obrat = 32

Race differential = coefficient on white = 0.112838, se = 0.020188

t-stat = 5.589, p value = 2.60e-08

95% confidence interval is 0.0732462753 to 0.152430088

This interval excludes zero, so at the average obrat there is evidence of discrimination.

**Question 4:**

1. price = -21.77 + 0.0020677 lotsize + 0.1227782 sqrft + 13.8525219 bdrms

(29.48) (0.00064) (0.013) (9.01) (\*) [37.1382106] [0.0012514] [0.0177253] [8.4786250] (\*\*)

(\*) usual SE

(\*\*) heteroskedasticity-robust SE

n = 88, R-sq = 0.6724

The most important difference is in the significance of lotsize. The robust standard error on lotsize is almost 2 times higher than the homoskedastic-only standard error, making lotsize much less significant.

The t-statistic on sqrft also falls, but it is still very significant.

The variable bdrms actually becomes somewhat more significant but is still not much significant.

2. log(price) = -1.29704 + 0.16797 log(lotsize) + 0.70023 log(sqrft) + 0.03696 bdrms

(0.65128) (0.03828) (0.09287) (0.02753) (\*) [0.781315] [0.041473] [0.103829] [0.030601] (\*\*)

n = 88, R-sq = 0.643

(\*) usual SE

(\*\*) heteroskedasticity-robust SE

The heteroscedasticity-robust standard error is slightly greater than the corresponding usual standard error, but the differences are relatively small. In particular, log(lotsize) and log(sqrft) still have very large t-statistics, and the t-statistic on bdrms is not significant at the 5% level against a one-sided alternative using either standard error.

3. Using the logarithmic transformation of the dependent variable often mitigates, if not

entirely eliminates, heteroskedasticity.

In this case, there's no important conclusions in the model for log(price) depend on the choice of standard error.

**Question 5:**

1. colGPA = 1.35651 + 0.41295 hsGPA + 0.01334 ACT - 0.07103 skipped + 0.12444 PC

R-sq = 0.2593, Adjusted R-sq = 0.2375, n = 141

2. usq = u^2

usq = 4.44160 - 2.85617 colGPA+ 0.46311 (colGPA^2)

3. min(h) = 0.03800

max(h) = 0.42674

-> The fitted values from part 2 are all strictly positive.

WLS regression:

colGPA = 2.294580 + 0.167340 hsGPA + 0.008570 ACT - 0.043827skipped + 0.046924 PC

R-sq = 0.09464, n = 141

There is little difference in the estimated coefficient on skipped and PC, and the OLS t statistic and WLS t statistic are also very close.

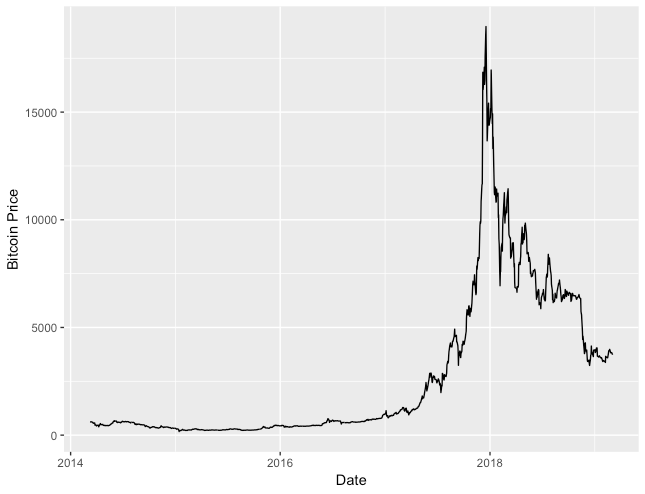
Both skipped and PC is statistically significant in OLS model but only skipped is significant in WLS model at 5% level.

4. The robust standard errors do not differ much from those in part 3; the explanatory variables that were statistically significant before are still statistically significant.

**Question 6:**

1.2.3. Downloaded files and used Excel to merge all data

4.

A picture containing wall, indoor

Description automatically generatedA picture containing wall

Description automatically generatedA picture containing wall, photo

Description automatically generated

A picture containing wall

Description automatically generated

5.

Close = 1591.299 + 18.305 DCOILWTICO

R-sq = 0.009382, n = 1217, p-value = 0.000716 (significant)

Close = - 3691 + 5464 DEXUSEU

R-sq = 0.01762, n = 1217, p-value = 3.38e-06 (significant)

Close = -19985.070 + 18.245 GOLDAMGBD228NLBM

R-sq = 0.1308, n = 1217, p-value = <2e-16 (significant)

Close = -18520 + 9.273 SP500

R-sq = 0.6841, n = 1217, p-value = <2e-16 (significant)

All models have significant results, but SP500 is the most important one and has the highest R-sq.

6. All series just need first-differencing to become stationary.

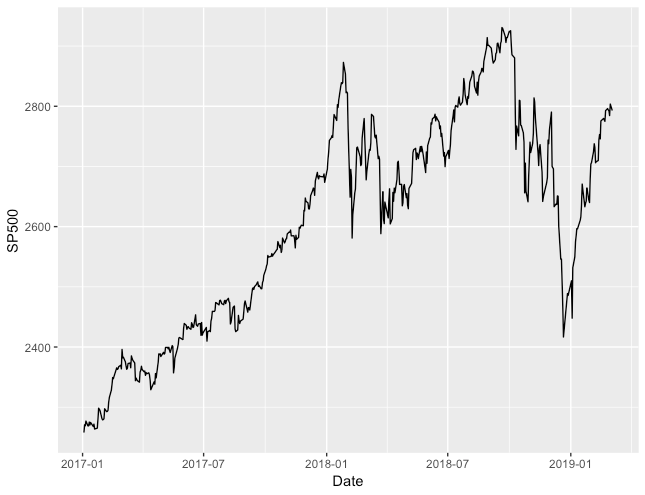
7. After taking differences, regress the bitcoin price on the other series, only SP500 is significant.

Close = 1.8814 + 0.9071 SP500

R-sq = 0.003993, n = 1216, p-value = 0.0276 (significant)

8. Used Excel to create new data (NewMergeData.csv)

A picture containing wall, text

Description automatically generated

A picture containing indoor, wall

Description automatically generated A screenshot of text

Description automatically generated

A picture containing wall

Description automatically generated

9.

A picture containing sky, water, object

Description automatically generated

10. Arima(6,1,5) is the best model with the lowest AIC = - 4018.15

11.

A screenshot of a cell phone

Description automatically generated

12. There're no seasonality in the data.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

13.

A screenshot of a cell phone

Description automatically generated

The periodogram has changed greatly, but this transformation didn't helps us to capture any seasonality in the data.

14. The best VAR model is model V4 with the highest R-sq = 0.06744

V4 = V1.l1 + V2.l1 + V3.l1 + V4.l1 + V5.l1 + const + trend

2 Granger causality relationships that're significant:

diff(log(GOLDAMGBD228NLBM)) ~ diff(log(DEXUSEU))

diff(log(SP500)) ~ diff(log(DEXUSEU)

15.

A picture containing screenshot

Description automatically generated

Compared to ARIMA model in part 10, the ARIMA model gives us a better view about the future trend of bitcoin prices than VAR model.