Problem Set 7

QTM 200: Applied Regression Analysis

Due: May 1, 2020

Instructions

- Please show your work! You may lose points by simply writing in the answer. If the problem requires you to execute commands in `R`, please include the code you used to get your answers. Please also include the `.R` file that contains your code. If you are not sure if work needs to be shown for a particular problem, please ask.
- Your homework should be submitted electronically on the course GitHub page in `.pdf` form.
- This problem set is due before midnight on Friday, May 1, 2020. No late assignments will be accepted.
- Total available points for this homework is 100.

Question 1 (50 points): Political Science

Consider the data set `MexicoMuniData.csv`, which includes municipal-level information from Mexico. The outcome of interest is the number of times the winning PAN presidential candidate in 2006 (`PAN.visits.06`) visited a district leading up to the 2009 federal elections, which is a count. Our main predictor of interest is whether the district was highly contested, or whether it was not (the PAN or their opponents have electoral security) in the previous federal elections during 2000 (`competitive.district`), which is binary (1=close/swing district, 0="safe seat"). We also include `marginality.06` (a measure of poverty) and `PAN.governor.06` (a dummy for whether the state has a PAN-affiliated governor) as additional control variables.

(a) Run a Poisson regression because the outcome is a count variable. Is there evidence that PAN presidential candidates visit swing districts more? Provide a test statistic and p-value.

```r
###Part 1 Question 1###
PANdata = MexicoMuniData
summary(PANdata)

pan_model2 <- lm(PAN.visits.06 ~ competitive.district + PAN.governor.06 +
    marginality.06, data = PANdata, family=poisson(link = "log"))
summary(pan_model2)
```

The intercept is 0.101 where the competitive district slope value is -.048. The p-value for this test is also significant at 7.98e-08. Due to the fact that the slope is negative, there seems to be the case that candidates visit the swing districts less than they do when they have a safe seat in the district.

(b) Interpret the `marginality.06` and `PAN.governor.06` coefficients. Both the PAN.governor.06 and marginality.06 have negative values, illustrating the negative slope or effect on the whole data. As marginality illustrates a measure of poverty, the more a district is in poverty, or mariginality = 1, then the candidate is less likely to go visit than if they were less impoverished. Similarly, if a district already had a PAN-affiliated governor, PAN.governor = 1, then the candidate would also visit the district less, most likely do to the security of the vote from the same PAN affiliation.

(c) Provide the estimated mean number of visits from the winning PAN presidential candidate for a hypothetical district that was competitive (`competitive.district`=1), had an average poverty level (`marginality.06` = 0), and a PAN governor (`PAN.governor.06`=1).

$$EstimatedVisits = 0.10076 - 0.04827(competitive.district) \tag{1}$$

$$- 0.03918(PAN.governor.06) - 0.12046(marginality.06) \tag{2}$$

Estimated visits in those circumstances would be = 0.01331 through that equation. Could be translated into a percentage like the rest of the dataset.

Question 2 (50 points): Biology

We'll be using data from a longitudinal sleep study of under 20 undergraduate students (n=18), which took place over the course of 10 days to see if sleep deprivation has any effect on participants' reaction time. Load the data through the `lmer` package.

1. Create a "pooled" linear model where you regress `Days` on the outcome `Reaction`. Make sure to run regression diagnostics to check if the variance around the regression line is equal for every year.

```r
###Part 2 Question 1###
library(lme4)
str(sleepstudy)

model1 <- lm(Reaction ~ Days, data = sleepstudy)
summary(model1)
ggplot(data = sleepstudy, aes(Days, Reaction))+
  geom_point() +
  geom_smooth(method = "lm")

sleepstudy$PooledPredictions <- fitted(model1)
plot(model1) #checking residuals and variance. Seems like there are a few
    points that are view more as extreme such as point 57, 99, and 10.
```

2. Fit an "un-pooled" regression model with varying intercepts for patient (include an additive factor for patient) and save the fitted values.

```r
###Part 2 Question 2###
model2 <- lm(Reaction ~ Days + Subject, data = sleepstudy)

sleepstudy$WithSubject <- fitted(model2)
```

3. Fit a "un-pooled" regression model with varying slopes of time (days) for patients (include only the interaction `Days:Subject`) and save the fitted values.

```r
###Part 2 Question 3###
model3 <- lm(Reaction ~ Days:Subject, data = sleepstudy)

sleepstudy$Varying <- fitted(model3)
```

4. Fit an "un-pooled" regression model with varying intercepts for patients with varying slopes of time (days) by patient (include the interaction and constituent terms of `Days` and `Subject`, `Days + Subject + Days:Subject`) and save the fitted values.

```r
###Part 2 Question 4###
interaction_model <- lm(Reaction ~ Days + Subject + Days:Subject, data=
    sleepstudy)
sleepstudy$Interaction <- fitted(interaction_model)
```

5. Fit a "semi-pooled" multi-level model with varying-intercept for subject and varying-slope of day by subject. Is it worthwhile for us to run a multi-level model with varying

effects of time by subject? Why? Compare your model from part 5 to the other completely "pooled" or "un-pooled models".

```
###Part 2 Question 5 ###

plot_sleep <- ggplot(data=sleepstudy, aes(x=Days, y = Reaction, group = Subject))+
  geom_line(aes(y = PooledPredictions), color = "red")+
  geom_line(aes(y=WithSubject), color = "blue")+
  #geom_line(aes(y=Varying), color = "green") +
  #geom_line(aes(y = Interaction), color = "orange")+
  geom_point(alpha = 0.2, size = 2) +
  facet_wrap(~Subject)
plot_sleep

plot_sleep2 <- ggplot(data=sleepstudy, aes(x=Days, y = Reaction, group = Subject))+
  geom_line(aes(y = PooledPredictions), color = "red")+
  #geom_line(aes(y=WithSubject), color = "blue")+
  geom_line(aes(y=Varying), color = "green") +
  #geom_line(aes(y = Interaction), color = "orange")+
  geom_point(alpha = 0.2, size = 2) +
  facet_wrap(~Subject)
plot_sleep2

plot_sleep3 <- ggplot(data=sleepstudy, aes(x=Days, y = Reaction, group = Subject))+
  geom_line(aes(y = PooledPredictions), color = "red")+
  #geom_line(aes(y=WithSubject), color = "blue")+
  #geom_line(aes(y=Varying), color = "green") +
  geom_line(aes(y = Interaction), color = "orange")+
  geom_point(alpha = 0.2, size = 2) +
  facet_wrap(~Subject)
plot_sleep3
```

It definitely could be worth looking at it, because each variable seemingly interacts differently when applied in different situations. It's difficult to tell based on each subject exactly what changes, but the variations do exist especially visible through the slopes of the graphs and their interactions.