Problem Set 7

QTM 200: Applied Regression Analysis

Due: May 6, 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 Wednesday, May 6, 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
model <- glm(PAN.visits.06 ~ competitive.district + marginality.06 + PAN.governor.06, family = "poisson", data=mexico_elections)
model
```

```
Coefficients:
(Intercept)  competitive.district        marginality.06
-3.9304                -0.4594                -2.0981
PAN.governor.06
-0.2073
```

```r
anova(model, test = "Chisq")
```

```
                     Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                                  2392    1433.83
competitive.district  1     3.04      2391    1430.79  0.08122 .
marginality.06        1   465.60      2390     965.19  < 2e-16 ***
PAN.governor.06       1     1.62      2389     963.57  0.20367
```

Since the p-value for competitive district is greater than 0.05, there is no enough evidence to support that whether the district is highly contested or not has a statistically significant effect on the number of times PAN presideential candidates visited.

(b) Interpret the `marginality.06` and `PAN.governor.06` coefficients.

The coefficient for marginality.06 is -2.098 means that keeping other variables constant, the expected log count for a one-unit increase in marginality.06 is -2.098.
In other words, the district being a safe seat would decrease the log odds of marginality.06 by 2.098.
The coefficient for PAN.governor.06 is -0.207 means that keeping other variables constant, the expected log count for a one-unit increase in PAN.governor.06 is -0.207.

(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).

```r
Y = -3.9304 - 0.4594*competitive.district - 2.0981* marginality.06 - 0.2073*PAN.governor.06
```

When competitive.district = 1, marginality.06 = 0, and PAN.governor.06 = 1, Y would be -4.60.

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
pooled_1 <- lm(Reaction ~ Days, data=lme4)
```

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

```r
unpooled_2 <- lm(Subject, data=lme4)
```

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
unpooled_3 <- lm(Reaction ~ Subject:Days, data=lme4)
```

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
unpooled_4 <- lm(Reaction ~ Subject + Days + Subject:Days, data=lme4)
```

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".

   It's not worthy to run a multi-level model with varying effects of time by subject.