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
rpmodel<-glm(PAN.visits.06~competitive.district+marginality.06+PAN.
    governor.06, data=mexico_elections, family = poisson)
summary(rpmodel)
```

```
Call:
glm(formula = PAN.visits.06 ~ competitive.district + marginality.06 +
PAN.governor.06, family = poisson, data = mexico_elections)

Deviance Residuals:
Min       1Q   Median       3Q      Max
-2.1441  -0.3596  -0.1742  -0.0783  15.2935

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept)           -3.9304     0.1747 -22.503   <2e-16 ***
competitive.district  -0.4594     0.3276  -1.402    0.161
marginality.06        -2.0981     0.1210 -17.343   <2e-16 ***
PAN.governor.06       -0.2073     0.1660  -1.249    0.212
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 1433.83  on 2392  degrees of freedom
Residual deviance:  963.57  on 2389  degrees of freedom
(4 observations deleted due to missingness)
AIC: 1255.9

Number of Fisher Scoring iterations: 7
```

There is not evidence that PAN presidential candidates visit swing districts more. The test statistic is -1.402 and the p-value is 0.161, which is larger than the standard significance threshold of 0.05.

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

```r
exp(coef(rpmodel))
```

```
(Intercept) competitive.district       marginality.06      PAN.governor.06
0.0196349           0.6316508           0.1226841           0.8127638
```

On average, if we hold all other variables constant, a unit increase in poverty decreased candidate district visits by 0.1226841 units and having a governor decreased candidate district visits by 0.8127638 units. It is important to note that the p-value for PAN.governor.06 is not significant.

(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).

The estimated mean number of visits from the winning PAN presidential candidate for a hypothetical district that was competitive, had an average poverty level, and a PAN governor is 0.0101. This is found by exponentiating the following calculation: -3.9304+(-0.4594*1)+(-2.0981*0)+(-0.2073*1).

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.

```
pooledmodel<-lm(Reaction~Days, data=sleepstudy)
summary(pooledmodel)
plot(pooledmodel)
sleepstudy$pooledmodel <-fitted(pooledmodel)
```

```
Call:
lm(formula = Reaction ~ Days, data = sleepstudy)

Residuals:
Min       1Q   Median       3Q      Max
-110.848  -27.483    1.546   26.142  139.953

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  251.405      6.610  38.033  < 2e-16 ***
Days          10.467      1.238   8.454 9.89e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 47.71 on 178 degrees of freedom
Multiple R-squared:  0.2865,        Adjusted R-squared:  0.2825
F-statistic: 71.46 on 1 and 178 DF,  p-value: 9.894e-15
```

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

```r
unpooled1<-lm(Reaction~Days+factor(Subject)-1, data=sleepstudy)
summary(unpooled1)
sleepstudy$unpooled1 <-fitted(unpooled1)
```

```
Call:
lm(formula = Reaction ~ Days + factor(Subject) - 1, data = sleepstudy)

Residuals:
Min       1Q   Median       3Q      Max
-100.540  -16.389   -0.341   15.215  131.159

Coefficients:
Estimate Std. Error t value Pr(>|t|)
Days                 10.4673     0.8042   13.02   <2e-16 ***
factor(Subject)308 295.0310    10.4471   28.24   <2e-16 ***
factor(Subject)309 168.1302    10.4471   16.09   <2e-16 ***
factor(Subject)310 183.8985    10.4471   17.60   <2e-16 ***
factor(Subject)330 256.1186    10.4471   24.52   <2e-16 ***
factor(Subject)331 262.3333    10.4471   25.11   <2e-16 ***
factor(Subject)332 260.1993    10.4471   24.91   <2e-16 ***
factor(Subject)333 269.0555    10.4471   25.75   <2e-16 ***
factor(Subject)334 248.1993    10.4471   23.76   <2e-16 ***
factor(Subject)335 202.9673    10.4471   19.43   <2e-16 ***
factor(Subject)337 328.6182    10.4471   31.45   <2e-16 ***
factor(Subject)349 228.7317    10.4471   21.89   <2e-16 ***
factor(Subject)350 266.4999    10.4471   25.51   <2e-16 ***
factor(Subject)351 242.9950    10.4471   23.26   <2e-16 ***
factor(Subject)352 290.3188    10.4471   27.79   <2e-16 ***
factor(Subject)369 258.9319    10.4471   24.79   <2e-16 ***
factor(Subject)370 244.5990    10.4471   23.41   <2e-16 ***
factor(Subject)371 247.8813    10.4471   23.73   <2e-16 ***
factor(Subject)372 270.7833    10.4471   25.92   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 30.99 on 161 degrees of freedom
Multiple R-squared:  0.9907,        Adjusted R-squared:  0.9896
F-statistic: 901.6 on 19 and 161 DF,  p-value: < 2.2e-16
```

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
unpooled2<-lm(Reaction~Days:factor(Subject)-1, data=sleepstudy)
summary(unpooled2)
```

```
sleepstudy$unpooled2 <-fitted(unpooled2)
```

```
Call:
lm(formula = Reaction ~ Days:factor(Subject) - 1, data = sleepstudy)

Residuals:
Min      1Q  Median      3Q     Max
-207.75  -25.20   71.24  169.32  321.54

Coefficients:
Estimate Std. Error t value Pr(>|t|)
Days:factor(Subject)308   60.321      8.618   7.000 6.45e-11 ***
Days:factor(Subject)309   34.639      8.618   4.019 8.92e-05 ***
Days:factor(Subject)310   38.244      8.618   4.438 1.67e-05 ***
Days:factor(Subject)330   48.748      8.618   5.657 6.83e-08 ***
Days:factor(Subject)331   50.383      8.618   5.846 2.69e-08 ***
Days:factor(Subject)332   51.291      8.618   5.952 1.59e-08 ***
Days:factor(Subject)333   52.566      8.618   6.100 7.53e-09 ***
Days:factor(Subject)334   50.174      8.618   5.822 3.03e-08 ***
Days:factor(Subject)335   38.651      8.618   4.485 1.38e-05 ***
Days:factor(Subject)337   64.832      8.618   7.523 3.49e-12 ***
Days:factor(Subject)349   47.459      8.618   5.507 1.41e-07 ***
Days:factor(Subject)350   55.162      8.618   6.401 1.59e-09 ***
Days:factor(Subject)351   47.667      8.618   5.531 1.25e-07 ***
Days:factor(Subject)352   57.204      8.618   6.638 4.56e-10 ***
Days:factor(Subject)369   51.606      8.618   5.988 1.32e-08 ***
Days:factor(Subject)370   51.285      8.618   5.951 1.60e-08 ***
Days:factor(Subject)371   49.236      8.618   5.713 5.18e-08 ***
Days:factor(Subject)372   53.463      8.618   6.204 4.43e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 145.5 on 162 degrees of freedom
Multiple R-squared:  0.7935,        Adjusted R-squared:  0.7706
F-statistic: 34.59 on 18 and 162 DF,  p-value: < 2.2e-16
```

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.

```
unpooled2<-lm(Reaction~Days:factor(Subject)-1, data=sleepstudy)
summary(unpooled2)
sleepstudy$unpooled2 <-fitted(unpooled2)
```

```
Call:
lm(formula = Reaction ~ Days + factor(Subject) - 1, data = sleepstudy)

Residuals:
Min       1Q   Median       3Q      Max
-100.540  -16.389   -0.341   15.215  131.159

Coefficients:
Estimate Std. Error t value Pr(>|t|)
Days                 10.4673     0.8042   13.02   <2e-16 ***
factor(Subject)308 295.0310    10.4471   28.24   <2e-16 ***
factor(Subject)309 168.1302    10.4471   16.09   <2e-16 ***
factor(Subject)310 183.8985    10.4471   17.60   <2e-16 ***
factor(Subject)330 256.1186    10.4471   24.52   <2e-16 ***
factor(Subject)331 262.3333    10.4471   25.11   <2e-16 ***
factor(Subject)332 260.1993    10.4471   24.91   <2e-16 ***
factor(Subject)333 269.0555    10.4471   25.75   <2e-16 ***
factor(Subject)334 248.1993    10.4471   23.76   <2e-16 ***
factor(Subject)335 202.9673    10.4471   19.43   <2e-16 ***
factor(Subject)337 328.6182    10.4471   31.45   <2e-16 ***
factor(Subject)349 228.7317    10.4471   21.89   <2e-16 ***
factor(Subject)350 266.4999    10.4471   25.51   <2e-16 ***
factor(Subject)351 242.9950    10.4471   23.26   <2e-16 ***
factor(Subject)352 290.3188    10.4471   27.79   <2e-16 ***
factor(Subject)369 258.9319    10.4471   24.79   <2e-16 ***
factor(Subject)370 244.5990    10.4471   23.41   <2e-16 ***
factor(Subject)371 247.8813    10.4471   23.73   <2e-16 ***
factor(Subject)372 270.7833    10.4471   25.92   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 30.99 on 161 degrees of freedom
Multiple R-squared:  0.9907,        Adjusted R-squared:  0.9896
F-statistic: 901.6 on 19 and 161 DF,  p-value: < 2.2e-16
```

5. 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
unpooled3<-lm(Reaction~Days+Subject+Days:factor(Subject)-1, data=
    sleepstudy)
summary(unpooled3)
sleepstudy$unpooled3 <-fitted(unpooled3)
```

```
Call:
```

```
lm(formula = Reaction ~ Days + Subject + Days:factor(Subject) -
1, data = sleepstudy)

Residuals:
Min       1Q   Median       3Q      Max
-106.397  -10.692   -0.177   11.417  132.510

Coefficients:
Estimate Std. Error t value Pr(>|t|)
Days                       21.765      2.818   7.725 1.74e-12 ***
Subject308                244.193     15.042  16.234  < 2e-16 ***
Subject309                205.055     15.042  13.632  < 2e-16 ***
Subject310                203.484     15.042  13.528  < 2e-16 ***
Subject330                289.685     15.042  19.259  < 2e-16 ***
Subject331                285.739     15.042  18.996  < 2e-16 ***
Subject332                264.252     15.042  17.568  < 2e-16 ***
Subject333                275.019     15.042  18.284  < 2e-16 ***
Subject334                240.163     15.042  15.966  < 2e-16 ***
Subject335                263.035     15.042  17.487  < 2e-16 ***
Subject337                290.104     15.042  19.287  < 2e-16 ***
Subject349                215.112     15.042  14.301  < 2e-16 ***
Subject350                225.835     15.042  15.014  < 2e-16 ***
Subject351                261.147     15.042  17.362  < 2e-16 ***
Subject352                276.372     15.042  18.374  < 2e-16 ***
Subject369                254.968     15.042  16.951  < 2e-16 ***
Subject370                210.449     15.042  13.991  < 2e-16 ***
Subject371                253.636     15.042  16.862  < 2e-16 ***
Subject372                267.045     15.042  17.754  < 2e-16 ***
Days:factor(Subject)309   -19.503      3.985  -4.895 2.61e-06 ***
Days:factor(Subject)310   -15.650      3.985  -3.928 0.000133 ***
Days:factor(Subject)330   -18.757      3.985  -4.707 5.84e-06 ***
Days:factor(Subject)331   -16.499      3.985  -4.141 5.88e-05 ***
Days:factor(Subject)332   -12.198      3.985  -3.061 0.002630 **
Days:factor(Subject)333   -12.623      3.985  -3.168 0.001876 **
Days:factor(Subject)334    -9.512      3.985  -2.387 0.018282 *
Days:factor(Subject)335   -24.646      3.985  -6.185 6.07e-09 ***
Days:factor(Subject)337    -2.739      3.985  -0.687 0.492986
Days:factor(Subject)349    -8.271      3.985  -2.076 0.039704 *
Days:factor(Subject)350    -2.261      3.985  -0.567 0.571360
Days:factor(Subject)351   -15.331      3.985  -3.848 0.000179 ***
Days:factor(Subject)352    -8.198      3.985  -2.057 0.041448 *
Days:factor(Subject)369   -10.417      3.985  -2.614 0.009895 **
```

```
Days:factor(Subject)370   -3.709      3.985  -0.931 0.353560
Days:factor(Subject)371  -12.576      3.985  -3.156 0.001947 **
Days:factor(Subject)372  -10.467      3.985  -2.627 0.009554 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 25.59 on 144 degrees of freedom
Multiple R-squared:  0.9943,        Adjusted R-squared:  0.9929
F-statistic: 700.4 on 36 and 144 DF,  p-value: < 2.2e-16
```

6. 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".

```
semipooled<-lmer(Reaction~Days+(1+Days|Subject), data=sleepstudy)
summary(semipooled)
sleepstudy$semipooled <-fitted(semipooled)
plot(sleepstudy$Days,sleepstudy$semipooled)
plot(sleepstudy$Days,sleepstudy$pooledmodel)
plot(sleepstudy$Days,sleepstudy$unpooled1)
plot(sleepstudy$Days,sleepstudy$unpooled2)
plot(sleepstudy$Days,sleepstudy$unpooled3)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: Reaction ~ Days + (1 + Days | Subject)
Data: sleepstudy

REML criterion at convergence: 1743.6

Scaled residuals:
Min      1Q  Median      3Q     Max
-3.9536 -0.4634  0.0231  0.4633  5.1793

Random effects:
Groups   Name        Variance Std.Dev. Corr
Subject  (Intercept) 611.90   24.737
Days         35.08    5.923   0.07
Residual             654.94   25.592
Number of obs: 180, groups:  Subject, 18

Fixed effects:
Estimate Std. Error t value
(Intercept)  251.405      6.824  36.843
Days          10.467      1.546   6.771
```

```
Correlation of Fixed Effects:
(Intr)
Days -0.138
```