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
model1<-glm(PAN.visits.06~competitive.district + marginality.06 + PAN.governor.06, family="poisson",data=mexico_elections)
summary(model1)
#There is little evidence that PAN candidates visit swing distrcits more. The p value of .161 would not be significant at any customary level of siginificance.
```

```
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
```

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

```
#On average, an increase of 1 in terms of poverty and having a PAN
    governor decreased the number of visits to a district by 2.1 and .2
    respectively. However the PAN governor variable was 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).

```
#This district would have a mean number of visits of -3.93 -.46 -.21 =
    -4.6
```

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.

```
model2<-plm(Reaction~Days,data=sleepstudy,index="Subject",model="pooling"
    )
summary(model2)
```

```
Balanced Panel: n = 18, T = 10, N = 180

Residuals:
    Min.  1st Qu.   Median  3rd Qu.     Max.
-110.848  -27.483    1.546   26.142  139.953

Coefficients:
            Estimate Std. Error t-value  Pr(>|t|)
(Intercept) 251.4051     6.6102 38.0332 < 2.2e-16 ***
Days         10.4673     1.2382  8.4537 9.894e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    567950
Residual Sum of Squares: 405250
R-Squared:      0.28647
Adj. R-Squared: 0.28246
F-statistic: 71.4644 on 1 and 178 DF, p-value: 9.8941e-15
```

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

```
model3<-lm(Reaction~Days + Subject,data=sleepstudy)
summary(model3)
```

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

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  295.0310    10.4471  28.240  < 2e-16 ***
Days          10.4673     0.8042  13.015  < 2e-16 ***
Subject309  -126.9008    13.8597  -9.156 2.35e-16 ***
Subject310  -111.1326    13.8597  -8.018 2.07e-13 ***
Subject330   -38.9124    13.8597  -2.808 0.005609 **
Subject331   -32.6978    13.8597  -2.359 0.019514 *
Subject332   -34.8318    13.8597  -2.513 0.012949 *
Subject333   -25.9755    13.8597  -1.874 0.062718 .
Subject334   -46.8318    13.8597  -3.379 0.000913 ***
Subject335   -92.0638    13.8597  -6.643 4.51e-10 ***
Subject337    33.5872    13.8597   2.423 0.016486 *
Subject349   -66.2994    13.8597  -4.784 3.87e-06 ***
Subject350   -28.5311    13.8597  -2.059 0.041147 *
Subject351   -52.0361    13.8597  -3.754 0.000242 ***
Subject352    -4.7123    13.8597  -0.340 0.734300
Subject369   -36.0992    13.8597  -2.605 0.010059 *
Subject370   -50.4321    13.8597  -3.639 0.000369 ***
Subject371   -47.1498    13.8597  -3.402 0.000844 ***
Subject372   -24.2477    13.8597  -1.750 0.082108 .
---
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:  0.7277,    Adjusted R-squared:  0.6973
F-statistic: 23.91 on 18 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.

```
model4<-lm(Reaction~Days:Subject,data=sleepstudy)
summary(model4)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-105.638  -16.324    1.741   16.870  133.186

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      251.405      4.020  62.539  < 2e-16 ***
Days:Subject308   20.626      1.832  11.257  < 2e-16 ***
Days:Subject309   -5.057      1.832  -2.760  0.00646 **
Days:Subject310   -1.452      1.832  -0.792  0.42941
Days:Subject330    9.052      1.832   4.940 1.94e-06 ***
Days:Subject331   10.687      1.832   5.833 2.91e-08 ***
Days:Subject332   11.595      1.832   6.328 2.36e-09 ***
Days:Subject333   12.871      1.832   7.024 5.74e-11 ***
Days:Subject334   10.478      1.832   5.718 5.10e-08 ***
Days:Subject335   -1.045      1.832  -0.570  0.56934
Days:Subject337   25.136      1.832  13.718  < 2e-16 ***
Days:Subject349    7.763      1.832   4.237 3.80e-05 ***
Days:Subject350   15.467      1.832   8.441 1.73e-14 ***
Days:Subject351    7.972      1.832   4.351 2.40e-05 ***
Days:Subject352   17.509      1.832   9.556  < 2e-16 ***
Days:Subject369   11.911      1.832   6.500 9.58e-10 ***
Days:Subject370   11.589      1.832   6.325 2.40e-09 ***
Days:Subject371    9.541      1.832   5.207 5.79e-07 ***
Days:Subject372   13.767      1.832   7.514 3.75e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 29.02 on 161 degrees of freedom
Multiple R-squared:  0.7613,    Adjusted R-squared:  0.7346
F-statistic: 28.53 on 18 and 161 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.

```
model5<-lm(Reaction~Days + Subject + Days:Subject,data=sleepstudy)
summary(model5)
```

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

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       244.193     15.042  16.234  < 2e-16 ***
Days               21.765      2.818   7.725 1.74e-12 ***
Subject309        -39.138     21.272  -1.840 0.067848 .
Subject310        -40.708     21.272  -1.914 0.057643 .
Subject330         45.492     21.272   2.139 0.034156 *
Subject331         41.546     21.272   1.953 0.052749 .
Subject332         20.059     21.272   0.943 0.347277
Subject333         30.826     21.272   1.449 0.149471
Subject334         -4.030     21.272  -0.189 0.850016
Subject335         18.842     21.272   0.886 0.377224
Subject337         45.911     21.272   2.158 0.032563 *
Subject349        -29.081     21.272  -1.367 0.173728
Subject350        -18.358     21.272  -0.863 0.389568
Subject351         16.954     21.272   0.797 0.426751
Subject352         32.179     21.272   1.513 0.132535
Subject369         10.775     21.272   0.507 0.613243
Subject370        -33.744     21.272  -1.586 0.114870
Subject371          9.443     21.272   0.444 0.657759
Subject372         22.852     21.272   1.074 0.284497
Days:Subject309   -19.503      3.985  -4.895 2.61e-06 ***
Days:Subject310   -15.650      3.985  -3.928 0.000133 ***
Days:Subject330   -18.757      3.985  -4.707 5.84e-06 ***
Days:Subject331   -16.499      3.985  -4.141 5.88e-05 ***
Days:Subject332   -12.198      3.985  -3.061 0.002630 **
Days:Subject333   -12.623      3.985  -3.168 0.001876 **
Days:Subject334    -9.512      3.985  -2.387 0.018282 *
Days:Subject335   -24.646      3.985  -6.185 6.07e-09 ***
Days:Subject337    -2.739      3.985  -0.687 0.492986
Days:Subject349    -8.271      3.985  -2.076 0.039704 *
Days:Subject350    -2.261      3.985  -0.567 0.571360
Days:Subject351   -15.331      3.985  -3.848 0.000179 ***
Days:Subject352    -8.198      3.985  -2.057 0.041448 *
Days:Subject369   -10.417      3.985  -2.614 0.009895 **
Days:Subject370    -3.709      3.985  -0.931 0.353560
Days:Subject371   -12.576      3.985  -3.156 0.001947 **
Days:Subject372   -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:  0.8339,    Adjusted R-squared:  0.7936
F-statistic: 20.66 on 35 and 144 DF,  p-value: < 2.2e-16
```

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".

```
model6<-lm(Reaction~Days + Subject + Subject:Days,data=sleepstudy)
summary(model6)
#It is worthwhile to examine varying effects of time by subject because
    other factors that are not included in the model could make it so that
     time affects some subjects more than others. Time could also just
    affect subjects' sleep differently based on their biology. This is
    evident from the fact that in this model the various subjects are
    mostly not significant while the interaction is really what is
    significant.
```

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

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       244.193     15.042  16.234  < 2e-16 ***
Days               21.765      2.818   7.725 1.74e-12 ***
Subject309        -39.138     21.272  -1.840 0.067848 .
Subject310        -40.708     21.272  -1.914 0.057643 .
Subject330         45.492     21.272   2.139 0.034156 *
Subject331         41.546     21.272   1.953 0.052749 .
Subject332         20.059     21.272   0.943 0.347277
Subject333         30.826     21.272   1.449 0.149471
Subject334         -4.030     21.272  -0.189 0.850016
Subject335         18.842     21.272   0.886 0.377224
Subject337         45.911     21.272   2.158 0.032563 *
Subject349        -29.081     21.272  -1.367 0.173728
Subject350        -18.358     21.272  -0.863 0.389568
Subject351         16.954     21.272   0.797 0.426751
Subject352         32.179     21.272   1.513 0.132535
Subject369         10.775     21.272   0.507 0.613243
Subject370        -33.744     21.272  -1.586 0.114870
Subject371          9.443     21.272   0.444 0.657759
Subject372         22.852     21.272   1.074 0.284497
Days:Subject309   -19.503      3.985  -4.895 2.61e-06 ***
Days:Subject310   -15.650      3.985  -3.928 0.000133 ***
Days:Subject330   -18.757      3.985  -4.707 5.84e-06 ***
Days:Subject331   -16.499      3.985  -4.141 5.88e-05 ***
Days:Subject332   -12.198      3.985  -3.061 0.002630 **
Days:Subject333   -12.623      3.985  -3.168 0.001876 **
Days:Subject334    -9.512      3.985  -2.387 0.018282 *
Days:Subject335   -24.646      3.985  -6.185 6.07e-09 ***
Days:Subject337    -2.739      3.985  -0.687 0.492986
Days:Subject349    -8.271      3.985  -2.076 0.039704 *
Days:Subject350    -2.261      3.985  -0.567 0.571360
Days:Subject351   -15.331      3.985  -3.848 0.000179 ***
Days:Subject352    -8.198      3.985  -2.057 0.041448 *
Days:Subject369   -10.417      3.985  -2.614 0.009895 **
Days:Subject370    -3.709      3.985  -0.931 0.353560
Days:Subject371   -12.576      3.985  -3.156 0.001947 **
Days:Subject372   -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:  0.8339,    Adjusted R-squared:  0.7936
F-statistic: 20.66 on 35 and 144 DF,  p-value: < 2.2e-16
```