Problem Set 7

Vanessa Wong

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.

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

- p = 0.161, z = -1.402

- p >0.05, therefore swing district status is not a statiscally reliable predictor of the number of visits the winning PAN candidate made in 2006. thus there is no evidence that PAN presidential candidates' visits swing districts more.

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

```
exp(coef(p_model))
```

- marginality coef. (exp) = 0.1226841

- marginality coef (non-exp) = -2.0981

- interpetation: holding all other variables constant, a one unit increase in poverty is associated with an average decrease in district visits by a multiplicative factor of 0.1227.

- PAN.governor coef. (exp) = 0.8127638

- PAN.governor coef (non-exp) = -0.2073

- interpretation: holding all other variables constant, a district's having a PAN-affiliated governor is associated with an average decrease in district visits by a multiplicative factor of 0.8128.

(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).

```
lambda <- exp(-3.9304 - .4594 - .2073)
lambda
```

- the estimated mean number of visits from winning PAN presidential candidates under these conditions is 0.01008 (approximately 0 visits).

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.

```
pooled <- lm(Reaction ~ Days, data=sleepstudy)
summary(pooled)
```

- model: y = 251.405 + 10.467x

```
par(mfrow=c(2,2)); plot(pooled)
```

- based on residuals vs. fitted values plot, residuals (variance) appears to be constant based on normal Q-Q plot, normality assumption appears to be met as all points are clustered very tightly around the QQ line.

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

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

- fitted values:

```
#########################################################################
#fitted values
#                        Estimate   Std. Error  t value  Pr(>|t|)

# 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
#########################################################################
```

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.

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

- fitted values:

```
#########################################################################
#fitted values
# Days:factor(Subject)308    60.321       8.618    7.000
```

```
# Days:factor(Subject)309    34.639      8.618    4.019
# Days:factor(Subject)310    38.244      8.618    4.438
# Days:factor(Subject)330    48.748      8.618    5.657
# Days:factor(Subject)331    50.383      8.618    5.846
# Days:factor(Subject)332    51.291      8.618    5.952
# Days:factor(Subject)333    52.566      8.618    6.100
# Days:factor(Subject)334    50.174      8.618    5.822
# Days:factor(Subject)335    38.651      8.618    4.485
# Days:factor(Subject)337    64.832      8.618    7.523
# Days:factor(Subject)349    47.459      8.618    5.507
# Days:factor(Subject)350    55.162      8.618    6.401
# Days:factor(Subject)351    47.667      8.618    5.531
# Days:factor(Subject)352    57.204      8.618    6.638
# Days:factor(Subject)369    51.606      8.618    5.988
# Days:factor(Subject)370    51.285      8.618    5.951
# Days:factor(Subject)371    49.236      8.618    5.713
# Days:factor(Subject)372    53.463      8.618    6.204
##############################################################################
```

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.

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

- fitted values:

```
##############################################################################
#fitted values
# factor(Subject)308        244.193     15.042   16.234
# factor(Subject)309        205.055     15.042   13.632
# factor(Subject)310        203.484     15.042   13.528
# factor(Subject)330        289.685     15.042   19.259
# factor(Subject)331        285.739     15.042   18.996
# factor(Subject)332        264.252     15.042   17.568
# factor(Subject)333        275.019     15.042   18.284
# factor(Subject)334        240.163     15.042   15.966
# factor(Subject)335        263.035     15.042   17.487
# factor(Subject)337        290.104     15.042   19.287
# factor(Subject)349        215.112     15.042   14.301
# factor(Subject)350        225.835     15.042   15.014
# factor(Subject)351        261.147     15.042   17.362
# factor(Subject)352        276.372     15.042   18.374
# factor(Subject)369        254.968     15.042   16.951
# factor(Subject)370        210.449     15.042   13.991
# factor(Subject)371        253.636     15.042   16.862
# factor(Subject)372        267.045     15.042   17.754
# Days:factor(Subject)309   -19.503      3.985   -4.895
# Days:factor(Subject)310   -15.650      3.985   -3.928
```

```
# Days:factor(Subject)330   -18.757      3.985   -4.707
# Days:factor(Subject)331   -16.499      3.985   -4.141
# Days:factor(Subject)332   -12.198      3.985   -3.061
# Days:factor(Subject)333   -12.623      3.985   -3.168
# Days:factor(Subject)334    -9.512      3.985   -2.387
# Days:factor(Subject)335   -24.646      3.985   -6.185
# Days:factor(Subject)337    -2.739      3.985   -0.687
# Days:factor(Subject)349    -8.271      3.985   -2.076
# Days:factor(Subject)350    -2.261      3.985   -0.567
# Days:factor(Subject)351   -15.331      3.985   -3.848
# Days:factor(Subject)352    -8.198      3.985   -2.057
# Days:factor(Subject)369   -10.417      3.985   -2.614
# Days:factor(Subject)370    -3.709      3.985   -0.931
# Days:factor(Subject)371   -12.576      3.985   -3.156
# Days:factor(Subject)372   -10.467      3.985   -2.627
##############################################################
```

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".

```
semipool <- lmer(Reaction ~ Days + (1 + Days|Subject), data=sleepstudy)
summary(semipool)
sleepstudy$pooled_new <- fitted(pooled)
sleepstudy$unpooled_new <- fitted(unpooled)
sleepstudy$semipooled_new <- fitted(semipool)

plot(sleepstudy$Days, sleepstudy$pooled_new)
plot(sleepstudy$Days, sleepstudy$unpooled_new)
plot(sleepstudy$Days, sleepstudy$semipooled_new)
```

- No, because the semipooled multilevel model appears to be pretty similar to the pooled and unpooled models, especially the unpooled model.