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.

```
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 no evidence that supports PAN presidential candidates visit swing districts more as the test statistic was came out to be -1.402 and the p value was 0.161 which is large.

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

```
(Intercept) competitive.district       marginality.06
-3.9304467           -0.4594186           -2.0981427
PAN.governor.06
-0.2073147
```

For states that are "safe seat" district without a PAN-affilated governor, by increasing the measure of poverty by 1, this increases the odds of visits from the winning PAN presidential candidate by the multiplication factor of 0.12.

For states that are "safe seat" district with a measure of poverty of 0, by having a PAN-affiliated governor increases the odds of visits from the winning PAN presidential candidate by 0.81.

(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 is 0.01.

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.

```
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.

Fitted values:

```
1        2        3        4        5        6        7        8
295.0310 305.4983 315.9656 326.4329 336.9002 347.3675 357.8348 368.3020
9       10       11       12       13       14       15       16
378.7693 389.2366 168.1302 178.5975 189.0648 199.5321 209.9993 220.4666
```

17	18	19	20	21	22	23	24
230.9339	241.4012	251.8685	262.3358	183.8985	194.3658	204.8331	215.3003
25	26	27	28	29	30	31	32
225.7676	236.2349	246.7022	257.1695	267.6368	278.1041	256.1186	266.5859
33	34	35	36	37	38	39	40
277.0532	287.5205	297.9878	308.4551	318.9223	329.3896	339.8569	350.3242
41	42	43	44	45	46	47	48
262.3333	272.8005	283.2678	293.7351	304.2024	314.6697	325.1370	335.6043
49	50	51	52	53	54	55	56
346.0716	356.5388	260.1993	270.6666	281.1339	291.6011	302.0684	312.5357
57	58	59	60	61	62	63	64
323.0030	333.4703	343.9376	354.4049	269.0555	279.5228	289.9901	300.4574
65	66	67	68	69	70	71	72
310.9247	321.3920	331.8592	342.3265	352.7938	363.2611	248.1993	258.6665
73	74	75	76	77	78	79	80
269.1338	279.6011	290.0684	300.5357	311.0030	321.4703	331.9376	342.4048
81	82	83	84	85	86	87	88
202.9673	213.4345	223.9018	234.3691	244.8364	255.3037	265.7710	276.2383
89	90	91	92	93	94	95	96
286.7055	297.1728	328.6182	339.0855	349.5528	360.0201	370.4874	380.9547
97	98	99	100	101	102	103	104
391.4219	401.8892	412.3565	422.8238	228.7317	239.1990	249.6663	260.1335
105	106	107	108	109	110	111	112
270.6008	281.0681	291.5354	302.0027	312.4700	322.9373	266.4999	276.9672
113	114	115	116	117	118	119	120
287.4345	297.9018	308.3690	318.8363	329.3036	339.7709	350.2382	360.7055
121	122	123	124	125	126	127	128
242.9950	253.4622	263.9295	274.3968	284.8641	295.3314	305.7987	316.2660
129	130	131	132	133	134	135	136
326.7333	337.2005	290.3188	300.7860	311.2533	321.7206	332.1879	342.6552
137	138	139	140	141	142	143	144
353.1225	363.5898	374.0570	384.5243	258.9319	269.3991	279.8664	290.3337
145	146	147	148	149	150	151	152
300.8010	311.2683	321.7356	332.2029	342.6701	353.1374	244.5990	255.0663
153	154	155	156	157	158	159	160
265.5336	276.0008	286.4681	296.9354	307.4027	317.8700	328.3373	338.8046
161	162	163	164	165	166	167	168
247.8813	258.3485	268.8158	279.2831	289.7504	300.2177	310.6850	321.1523
169	170	171	172	173	174	175	176
331.6195	342.0868	270.7833	281.2506	291.7179	302.1852	312.6525	323.1198
177	178	179	180				
333.5871	344.0543	354.5216	364.9889				

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.

```
      1        2        3        4        5        6        7        8        9       10
244.1927 265.9574 287.7221 309.4868 331.2515 353.0162 374.7809 396.5456 418.3103 44
     13       14       15       16       17       18       19       20       21       22
209.5785 211.8403 214.1021 216.3639 218.6257 220.8874 223.1492 225.4110 203.4842 20
     25       26       27       28       29       30       31       32       33       34
227.9438 234.0587 240.1736 246.2885 252.4034 258.5183 289.6851 292.6932 295.7012 29
     37       38       39       40       41       42       43       44       45       46
307.7335 310.7416 313.7497 316.7577 285.7390 291.0050 296.2710 301.5370 306.8030 31
     49       50       51       52       53       54       55       56       57       58
327.8671 333.1331 264.2516 273.8184 283.3852 292.9519 302.5187 312.0855 321.6522 33
     61       62       63       64       65       66       67       68       69       70
275.0191 284.1612 293.3032 302.4452 311.5873 320.7293 329.8714 339.0134 348.1555 35
     73       74       75       76       77       78       79       80       81       82
264.6692 276.9223 289.1755 301.4286 313.6818 325.9349 338.1880 350.4412 263.0347 26
     85       86       87       88       89       90       91       92       93       94
251.5106 248.6295 245.7485 242.8675 239.9864 237.1054 290.1041 309.1301 328.1561 34
     97       98       99      100      101      102      103      104      105      106
404.2600 423.2859 442.3119 461.3379 215.1118 228.6057 242.0996 255.5936 269.0875 28
    109      110      111      112      113      114      115      116      117       11
323.0632 336.5572 225.8346 245.3386 264.8426 284.3467 303.8507 323.3547 342.8587 36
    121      122      123      124      125      126      127      128      129       13
261.1470 267.5805 274.0140 280.4475 286.8810 293.3145 299.7480 306.1815 312.6150 31
    133      134      135      136      137      138      139      140      141       14
303.5052 317.0717 330.6383 344.2048 357.7714 371.3379 384.9045 398.4710 254.9681 26
    145      146      147      148      149      150      151      152      153       15
300.3606 311.7087 323.0568 334.4049 345.7530 357.1011 210.4491 228.5052 246.5614 26
    157      158      159      160      161      162      163      164      165       16
318.7860 336.8421 354.8983 372.9544 253.6360 262.8245 272.0129 281.2014 290.3898 29
    169      170      171      172      173      174      175      176      177       17
327.1436 336.3320 267.0448 278.3429 289.6409 300.9390 312.2371 323.5352 334.8332 34
```

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.

```
      1        2        3        4        5        6        7        8        9
244.1927 265.9574 287.7221 309.4868 331.2515 353.0162 374.7809 396.5456 418.3103 44
     13       14       15       16       17       18       19       20       21       22
209.5785 211.8403 214.1021 216.3639 218.6257 220.8874 223.1492 225.4110 203.4842 20
```

```
25       26       27       28       29       30       31       32       33       34
227.9438 234.0587 240.1736 246.2885 252.4034 258.5183 289.6851 292.6932 295.7012 29
37       38       39       40       41       42       43       44       45       46
307.7335 310.7416 313.7497 316.7577 285.7390 291.0050 296.2710 301.5370 306.8030 31
49       50       51       52       53       54       55       56       57       58
327.8671 333.1331 264.2516 273.8184 283.3852 292.9519 302.5187 312.0855 321.6522 33
61       62       63       64       65       66       67       68       69       70
275.0191 284.1612 293.3032 302.4452 311.5873 320.7293 329.8714 339.0134 348.1555 35
73       74       75       76       77       78       79       80       81       82
264.6692 276.9223 289.1755 301.4286 313.6818 325.9349 338.1880 350.4412 263.0347 26
85       86       87       88       89       90       91       92       93       94
251.5106 248.6295 245.7485 242.8675 239.9864 237.1054 290.1041 309.1301 328.1561 34
97       98       99      100      101      102      103      104      105      106
404.2600 423.2859 442.3119 461.3379 215.1118 228.6057 242.0996 255.5936 269.0875 28
109      110      111      112      113      114      115      116      117      11
323.0632 336.5572 225.8346 245.3386 264.8426 284.3467 303.8507 323.3547 342.8587 36
121      122      123      124      125      126      127      128      129      13
261.1470 267.5805 274.0140 280.4475 286.8810 293.3145 299.7480 306.1815 312.6150 31
133      134      135      136      137      138      139      140      141      14
303.5052 317.0717 330.6383 344.2048 357.7714 371.3379 384.9045 398.4710 254.9681 26
145      146      147      148      149      150      151      152      153      15
300.3606 311.7087 323.0568 334.4049 345.7530 357.1011 210.4491 228.5052 246.5614 26
157      158      159      160      161      162      163      164      165      16
318.7860 336.8421 354.8983 372.9544 253.6360 262.8245 272.0129 281.2014 290.3898 29
169      170      171      172      173      174      175      176      177      17
327.1436 336.3320 267.0448 278.3429 289.6409 300.9390 312.2371 323.5352 334.8332 34
```

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 is worthwhile for us to run a multi-level model as in the pooled regression, the p values were very small, illustrating that the variable has an effect, however, in the unpooled regression, we can see that only certain answers of the variable has an effect of time by subject.