Applied Stats II - Problem Set 1

Imelda Finn (22334657)

Due: March 26, 2023

Code in PS3_ImeldaFinn.R

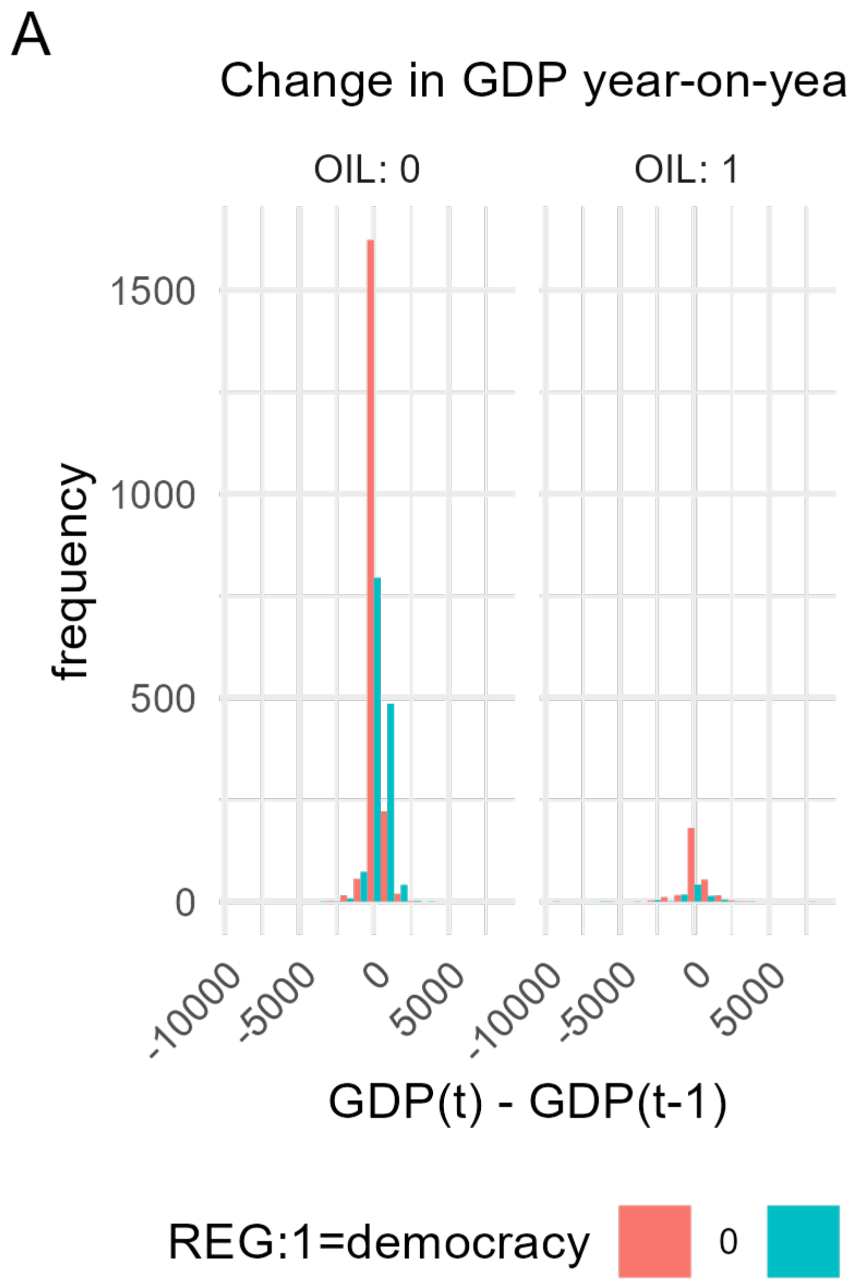
Question 1

We are interested in how governments' management of public resources impacts economic prosperity. Our data come from Alvarez, Cheibub, Limongi, and Przeworski (1996) and is labelled gdpChange.csv on GitHub. The dataset covers 135 countries observed between 1950 or the year of independence or the first year forwhich data on economic growth are available ("entry year"), and 1990 or the last year for which data on economic growth are available ("exit year"). The unit of analysis is a particular country during a particular year, for a total > 3,500 observations.

- Response variable:
  - GDPWdiff: Difference in GDP between year t and $t-1$. Possible categories include: "positive", "negative", or "no change"
- Explanatory variables:
  - REG: 1=Democracy; 0=Non-Democracy
  - OIL: 1=if the average ratio of fuel exports to total exports in 1984-86 exceeded 50%; 0= otherwise

The data was read in and GDPWDiff was factored. The cutoff point was 0, ie values less than 0 were categorised as *negative*, values equal to 0 were categorised as *no change* and values above 0 were categorised as *positive*. *no change* was set as the reference category.

```
gdp <- read_csv("./data/gdpChange.csv")
gdp<- rename(gdp, indx = `...1`)
gdp$diff <- ifelse(gdp$GDPWdiff==0, "no change",
                   ifelse(gdp$GDPWdiff>0, "positive", "negative"))
gdp$diff2 <- relevel(factor(gdp$diff, ordered = FALSE), ref="no change")
```



OIL=1 : average ratio of fuel exports to total exports in 1984-86 exceeded 50%; REG=1 : democracy

variable	OIL=0		OIL=1	
	REG=0	REG=1	REG=0	REG=1
GDPWdiff				
Min / Max	-2506.0 / 2821.0	-3741.0 / 3722.0	-9257.0 / 7867.0	-5997.0 / 3555.0
Med [IQR]	50.0 [-30.0;215.0]	293.0 [13.5;644.8]	140.5 [-50.0;463.0]	39.0 [-527.2;421.8]
Mean (std)	106.0 (395.3)	319.4 (561.6)	141.2 (1147.3)	-46.5 (1228.4)
N (NA)	1939 (0)	1408 (0)	288 (0)	86 (0)

1. An unordered multinomial logit with `GDPWdiff` as the output was constructed as follows:

```
multinom_model <- multinom(diff2 ~ REG + OIL, data = gdp )
```

The results of the model are shown in Table 1. The predicted probabilities are shown in Table 2, Figure 2. All of the predicted classes are *positive*. This isn't unexpected as Figure 1 shows that the data is skewed towards positive values.

Table 1: Multinomial, unordered

	Dependent variable:	
	negative	positive
	(1)	(2)
REG	1.379	1.769
	t = 1.794	t = 2.306
	p = 0.073*	p = 0.022**
OIL	4.784	4.576
	t = 0.695	t = 0.665
	p = 0.488	p = 0.507
Constant	3.805	4.534
	t = 14.058	t = 16.842
	p = 0.000***	p = 0.000***
Akaike Inf. Crit.	4,690.770	4,690.770
Note:	*p<0.1; **p<0.05; ***p<0.01	

The baseline category is regime `REG` = 0 (non-democracy) and `OIL` = 0 (not a significant oil exporter). The predicted probability of having no change in GDP when in the baseline category is 0.7%.

```
    round(predict(multinom_model, newdata = data.frame(REG=0, OIL=0),
        type = "probs"),2)

    predict_data <- data.frame(REG = rep(c(0,1), each = 2),
                               OIL= rep(c(0,1), 2))
    cbind(predict_data, predict(multinom_model,
            newdata = predict_data, type = "class"))

```

```
  no change   negative   positive
     0.01       0.32       0.67
```

```

 REG OIL predict(multinom_model, newdata = predict_data, type = "class")
1    0   0                                                          positive
2    0   1                                                          positive
3    1   0                                                          positive
4    1   1                                                          positive

              no change  negative  positive   Sum
  no change           0         0        16    16
  negative            0         0      1105  1105
  positive            0         0      2600  2600
  Sum                 0         0      3721  3721

```

Table 2: Predicted results from unordered Multinomial

	REG	OIL	level	probability
1	0	0	no change	0.007
2	0	1	no change	0.0001
3	1	0	no change	0.001
4	1	1	no change	0.00001
5	0	0	negative	0.323
6	0	1	negative	0.373
7	1	0	negative	0.246
8	1	1	negative	0.287
9	0	0	positive	0.670
10	0	1	positive	0.627
11	1	0	positive	0.753
12	1	1	positive	0.713

Holding `OIL` constant:

- a change in REG from 0 to 1 increases the log-odds of `diff`= *positive* vs. `diff` = *no change* by 1.769
- a change in REG from 0 to 1 multiplies the odds of `diff`= *positive* vs. `diff` = *no change* by a factor of $e^{1.769} = 5.87$.

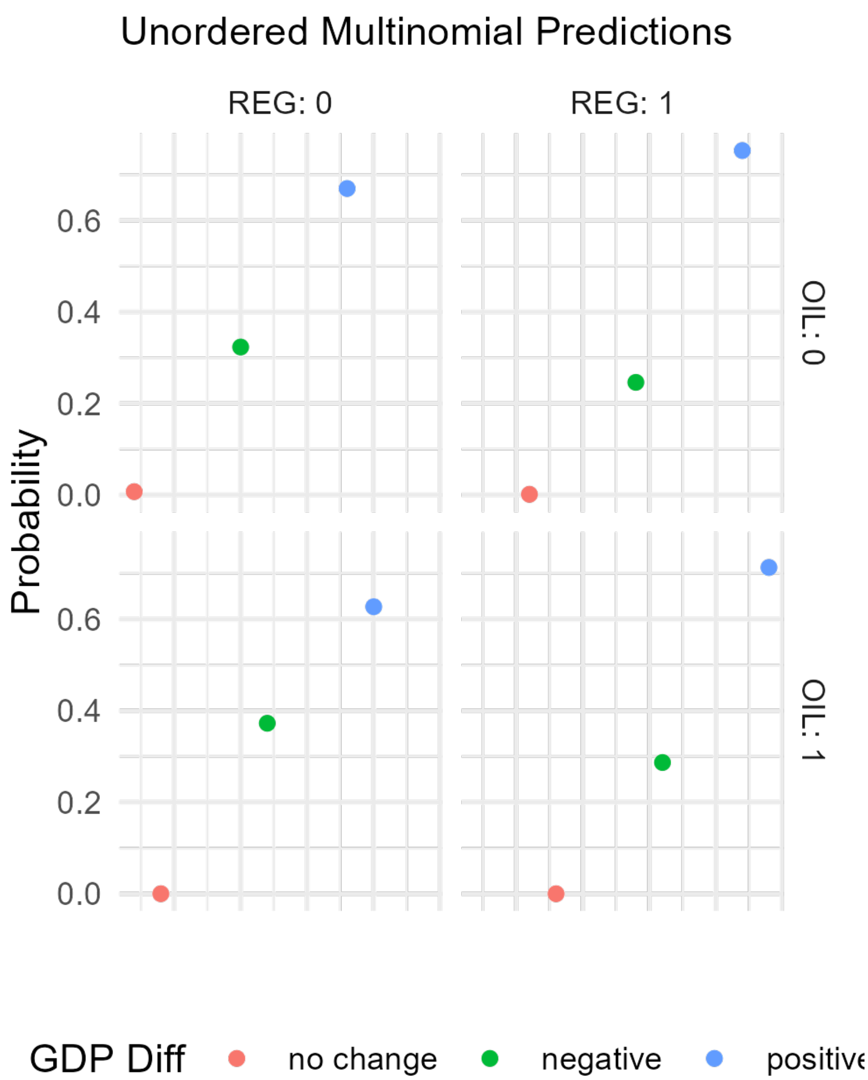


Figure 2: GDP Diff - unordered multinomial predictions

2. The factored `GDPWdiff` response variable (from 1) was ordered: (*negative* $<$ *no change* $<$ *positive*), i.e. the cutoff (0) was unchanged. An ordered multinomial logit model was created as follows:

```
# get an ordered factor for GDP difference
gdp$ordered_diff <- ordered(gdp$diff,
                             labels=c("negative", "no change", "positive"))
```

The results of the model are shown in Table 3. The baseline category is regime `REG` = 0 (non-democracy) and `OIL` = 0 (not a significant oil exporter). The predicted probability of having no change in GDP when in the baseline category is 0.5% (Table 4).

Table 3: Multinomial Logit, ordered

	Dependent variable:
	ordered_diff
REG	0.398
	(0.075)
	t = 5.300
	p = 0.00000***
OIL	−0.199
	(0.116)
	t = −1.717
	p = 0.086*
negative|no change	−0.731
	(0.048)
	t = −15.360
	p = 0.000***
no change|positive	−0.710
	(0.048)
	t = −14.955
	p = 0.000***
Observations	3,721
Note:	*p<0.1; **p<0.05; ***p<0.01

The coefficients and their confidence intervals are:

```
    cbind(logOdds = coef(ord_model), confint(ord_model))
    #        logOdds       2.5 %      97.5 %
```

```
    #REG   0.3984834   0.2516548  0.54643410
    #OIL  -0.1987177  -0.4237548  0.03019571

```

Holding `REG` constant:

- a change in OIL from 0 to 1 changes the log-odds of `diff`= *no change* vs. `diff` = *negative* by -0.199
- a change in OIL from 0 to 1 multiplies the odds of `diff`= *no change* vs. `diff` = *negative* by a factor of $e^{-0.199} = 0.82$.

The odds of having no change in GDP growth for a country that has oil, are .18% lower compared to a country that doesn't have oil, holding regime status constant.

```
 round(predict(ord_model, newdata = data.frame(REG=0, OIL=0), type = "
   probs"),2)
 predict(ord_model, newdata = data.frame(REG=0, OIL=0), type = "class")
 cbind(predict_data, predict(ord_model, predict_data, type="class"))

```

```
 negative no change  positive
     0.32       0.00       0.67

[1] positive
Levels: negative no change positive

  REG OIL predict(ord_model, predict_data, type = "class")
1   0   0                                         positive
2   0   1                                         positive
3   1   0                                         positive
4   1   1                                         positive

              negative no change positive  Sum
  negative           0         0     1105 1105
  no change          0         0       16   16
  positive           0         0     2600 2600
  Sum                0         0     3721 3721
```

The predicted probabilities are shown in Table 4, Figure 3.

```
 pred_ord <- melt(cbind(predict_data,
                        predict(ord_model, predict_data, type="probs")),
                  id.vars=c("REG", "OIL"),
                  variable.name="level", value.name="probability")

```

The model predictions for the individual GDP difference categories are give in Table 5 The proportional-odds assumption does not appear to hold for this regression i.e. the coefficients are not consistent (coef for `OIL` goes from -0.04 to 0.05 to -0.01).

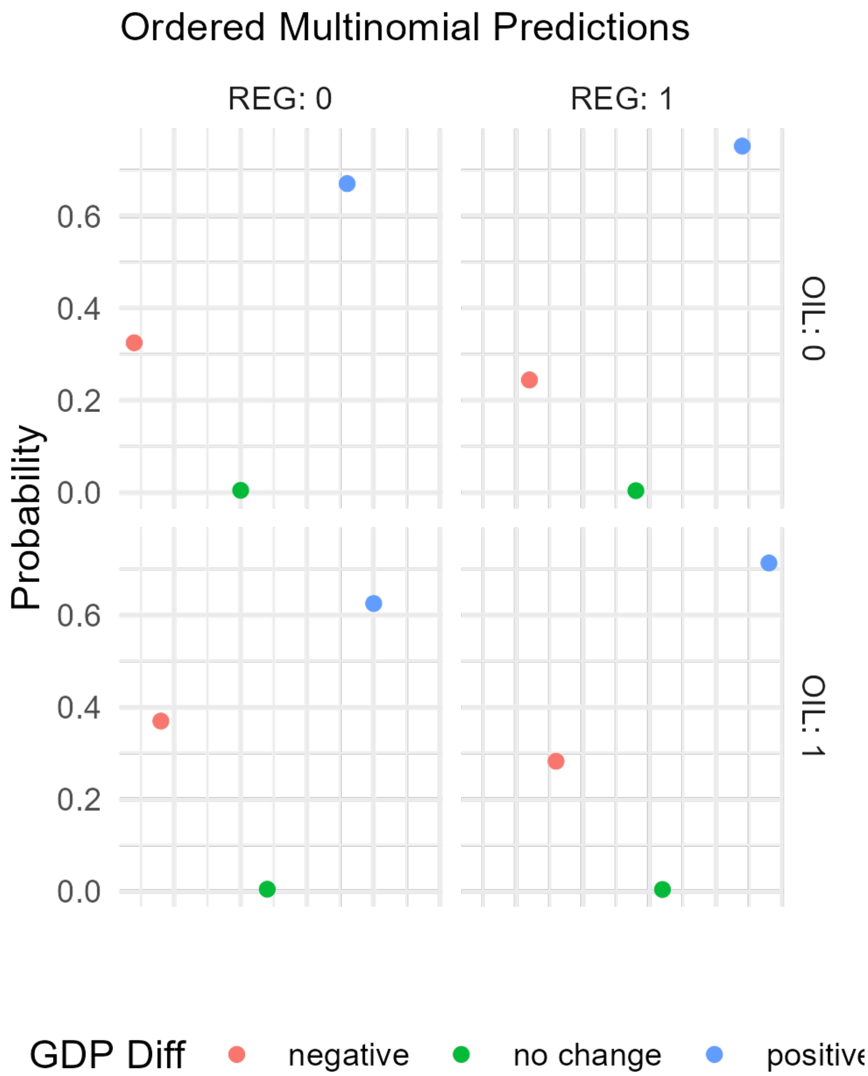


Figure 3: GDP Diff - ordered multinomial predictions

Table 4: Predicted results from Ordered Model

	REG	OIL	level	probability
1	0	0	negative	0.325
2	0	1	negative	0.370
3	1	0	negative	0.244
4	1	1	negative	0.283
5	0	0	no change	0.005
6	0	1	no change	0.005
7	1	0	no change	0.004
8	1	1	no change	0.004
9	0	0	positive	0.671
10	0	1	positive	0.625
11	1	0	positive	0.752
12	1	1	positive	0.713

```

id.vars=c("REG", "OIL")
for ( i in 1:length(unique(gdp$ordered_diff))){
  assign(paste("logit_model", i, sep=""),
         glm(ifelse(ordered_diff==unique(gdp$ordered_diff)[i],
                    1 , 0) ~ REG + OIL, data = gdp),
         envir = globalenv())
```

T he cutoff point affects the results. For example, changing the cutoff to split the data into 3 equal-length sections changes the coefficients and the confusion matrices. The resulting models aren't more accurate, because they reduce the number of true positive predictions for *positive* without increasing the true predictions for the other 2 categories enough to compensate. They also have higher deviance. They could be refined to better categorise/predict the data.

```
    cutoffs:  <=14, 14-283,   >=283
------------------------------------------------------------------
Multinomial model
                          Dependent variable:
                     ---------------------------
                        negative        positive
                          (1)             (2)
--------------------------------------------------
REG                     0.184**         1.365***
                        (0.088)          (0.087)

OIL                     0.428***        0.641***
                        (0.139)          (0.145)

```

Table 5: Comparison of models for GDP Diff categories

	Dependent variable:		
	ordered_diff)[i], 1, 0)		
	negative	no change	positive
REG	0.0832***	−0.0779***	−0.0054**
	(0.0154)	(0.0153)	(0.0022)
OIL	−0.0416*	0.0474*	−0.0058
	(0.0251)	(0.0250)	(0.0036)
Constant	0.6695***	0.3235***	0.0070***
	(0.0102)	(0.0102)	(0.0015)
Observations	3,721	3,721	3,721
Log Likelihood	−2,364.5140	−2,350.4150	4,869.1730
Akaike Inf. Crit.	4,735.0280	4,706.8310	−9,732.3450
Note:			*p<0.1; **p<0.05; ***p<0.01

```
Constant                  -0.091*          -0.649***
                          (0.051)          (0.059)

deviance: 7850.215

confusion matrix:
pred_m_all  no change  negative  positive
  no change       799       747       393
  negative         82        99       107
  positive        353       395       746

-------------------------------------------------------------------
Ordered Model
                  Dependent variable:
             ------------------------------
                         odiff
---------------------------------------
REG                       0.930***
                          (0.064)

OIL                       0.120
                          (0.105)

deviance: 7691.950

```

```
confusion matrix:
pred_o_all  negative no change  positive
  negative       846       881       500
  no change        0         0         0
  positive       395       353       746
```

Question 2

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.

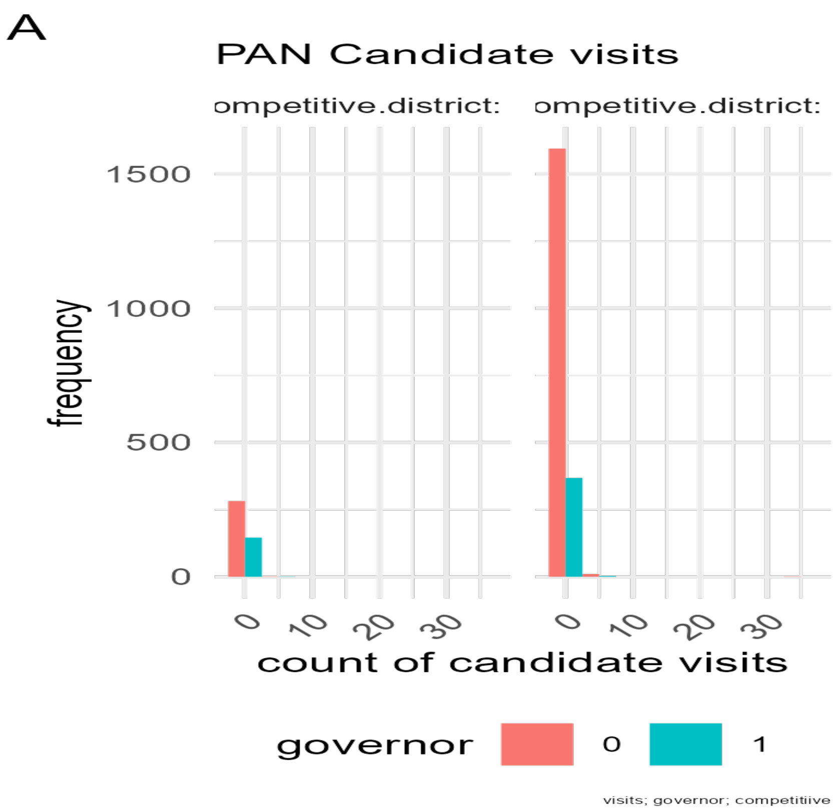


Figure 4: Presidential candidate visits

```
mexico <- read_csv("./data/MexicoMuniData.csv")
```

(a) A Poisson regression model was run because the outcome is a count variable, to consider whether there is evidence that PAN presidential candidates visit swing districts more? The model output is in Table 6. `competitive.district` coefficient is -0.081, but it is not a significant predictor for number of visits.

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

A poisson model was run to get the regression coefficients.

Table 6: Poisson Model of candidate visit counts

	Dependent variable:
	PAN.visits.06
	Poisson
competitive.district	−0.081
	(0.171)
marginality.06	−2.080***
	(0.117)
PAN.governor.06	−0.312*
	(0.167)
Constant	−3.810***
	(0.222)
Observations	2,407
Log Likelihood	−645.606
Akaike Inf. Crit.	1,299.213
Note:	*p<0.1; **p<0.05; ***p<0.01

Hypothesis Test

The summary results for the poisson model for the coefficients are:

```
 Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
 (Intercept)          -3.81023    0.22209 -17.156   <2e-16 ***
 competitive.district -0.08135    0.17069  -0.477   0.6336
 marginality.06       -2.08014    0.11734 -17.728   <2e-16 ***
PAN.governor.06       -0.31158    0.16673  -1.869   0.0617 .
```

1. H_0 PAN presidential candidate visits swing districts less than other districts ($E(\lambda|competitive.district = 1) < E(\lambda|competitive.district = 0)$)
2. H_a candidates visit swing districts at least as many times other districts

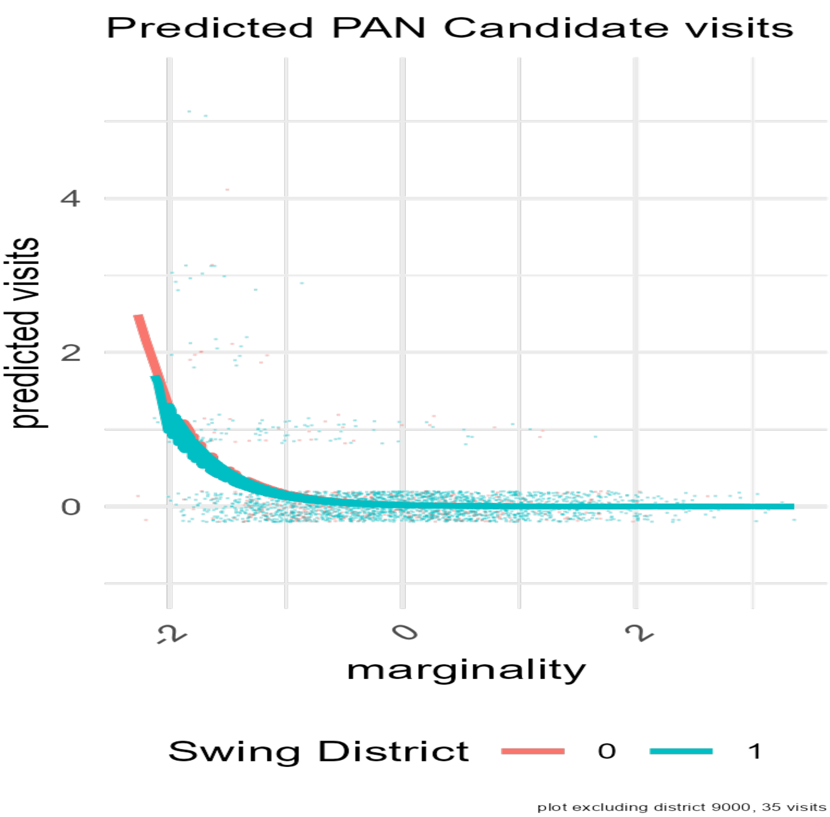


Figure 5: predicted presidential candidate visits

3. the test statistic $= \beta_{competitive}/se_{competitive} = -0.08135/0.17069 = -0.477$ ($\sim N(0,1)$)
4. the α value is 0.05, one-sided, left-tailed z-test
5. the `pvalue` $p = 0.6833189$ [1]
6. as `pvalue` is greater than α, we cannot reject the null hypothesis that closely contested districts receive fewer visits.

   Using R's `poisson.test`, the p-value is 0.8544, so we also reject the null hypothesis, i.e. there is not evidence to support the theory that presidential candidates visit swing districts more.

```
cd1 <-mexico$PAN.visits.06[mexico$competitive.district==1]
cd0 <-mexico$PAN.visits.06[mexico$competitive.district==0]

poisson.test(x=c(sum(cd1), sum(cd0)), T=c(length(cd1), length(cd0)),
             alternative="greater", conf.level=0.95)

Comparison of Poisson rates

data:  c(sum(cd1), sum(cd0)) time base: c(length(cd1), length(cd0))
count1 = 176, expected count1 = 181.52, p-value = 0.8544
alternative hypothesis: true rate ratio is greater than 1
95 percent confidence interval:
 0.6411102       Inf
sample estimates:
rate ratio
0.8506716

```

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

```
> lo_cis<-cbind(logOdds = coef(mexico_poisson), confint(mexico_poisson))
Waiting for profiling to be done...

# Coefficients and confidence intervals
                           logOdds      2.5 %       97.5 %
(Intercept)            -3.81023498 -4.2606981 -3.389583340
competitive.district   -0.08135181 -0.4063661  0.264275617
marginality.06         -2.08014361 -2.3151624 -1.855053854
PAN.governor.06        -0.31157887 -0.6484827  0.006518468

> exp(lo_cis)
                      exp(beta) 2.5 % 97.5 %
(Intercept)               0.022 0.014  0.034
competitive.district      0.922 0.666  1.302
marginality.06            0.125 0.099  0.156
PAN.governor.06           0.732 0.523  1.007

```

[1] pnorm(0.477)

`marginality.06` is the only coefficient which is significant at $\alpha = 0.01$; `PAN.governor.06` is significant at $\alpha = 0.1$.

The coefficient for `marginality.06` is -2.08 ($CI_{0.05} = -2.315, -1.855$). This means that, keeping all else constant, we expect a decrease of 2.08 in log count for a one-unit increase in `marginality.06`, i.e. if `marginality.06` increases by 1, we expect the estimated mean number of visits to decrease by 87.5% (multiply previous expected count by $e^{-2.08} = 0.125$). Districts with higher marginality receive fewer visits.

The coefficient for `PAN.governor.06` is -0.312, which means that if `PAN.governor.06` switches from 0 to 1, keeping all other variables constant, we expect an decrease in log count of 0.312. If `PAN.governor.06` changes from 0 to 1, we expect the estimated mean number of visits to decrease by 26.8% (ie multiply previous expected count by 0.732). Districts with a PAN governor are expected to receive fewer visits from PAN candidates. Note: $CI_{0.05} = -0.648, 0.007$, which includes 0. The test results suggest that having a PAN governor is not a significant predictor for the number of candidate visits.

(c) 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).

```
    mex_pred_data <- data.frame(competitive.district = 1,
                                marginality.06=0,
                                PAN.governor.06=1)
    pred_mex <- cbind(predict(mexico_poisson,
                             mex_pred_data,
                             type="response", se.fit =TRUE),
                  mex_pred_data)
    # create lower and upper bounds for CIs
    pred_mex$lowerBound <- pred_mex$fit - 1.96 * pred_mex$se.fit
    pred_mex$upperBound <- pred_mex$fit + 1.96 * pred_mex$se.fit

    round(pred_mex,3)


      fit se.fit residual.scale competitive.district marginality.06
1 0.015  0.003              1                    1              0
  PAN.governor.06 lowerBound upperBound
1               1      0.009      0.021



```

$\lambda = e^{\beta_0 + \beta_{competitive} \times competitive + \beta_{marginality} \times marginality + \beta_{governor} \times governor}$

$= e^{-3.810 - 0.081 \times 1 + -2.080 \times 0 + -0.312 \times 1} = e^{-4.203} = 0.015$ (The mean visits is 0.092, the median is 0.)

The estimated mean number of visits, in the time frame, by the winning PAN presidential candidate to a district which was a swing state, with average poverty (=0) and a PAN governor is 0.015.

Validation There are 2,272 zero count values in our dataset.

```
dispersiontest(mexico_poisson)

  Overdispersion test

data:  mexico_poisson
z = 1.0668, p-value = 0.143
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
   2.09834
```

A zero-inflated poisson model was run for comparison, the only coefficient with a significant deviance is the `marginality.06`, where an increase of 1 unit in marginality corresponds to a increase in log-odds of a 0 count value of 0.872(Table 8). The zero-inflated poisson changes the value of the coefficients, but doesn't change their statistical significance.

Table 7: Zero-inflation model

	Dependent variable:
	PAN.visits.06
	zero-inflated count data
competitive.district	0.900*
	(0.511)
marginality.06	0.872***
	(0.302)
PAN.governor.06	−0.175
	(0.412)
Constant	1.272*
	(0.675)
Observations	2,407
Log Likelihood	−600.386
Note:	*p<0.1; **p<0.05; ***p<0.01

```
anova(mexico_poisson, zeroinfl_poisson, test = "Chi")
Analysis of Deviance Table

```

Table 8: Zero Infl Poisson vs Poisson Model

	Dependent variable:	
	PAN.visits.06	
	zero-inflated count data	*Poisson*
competitive.district	0.402	−0.081
	(0.312)	(0.171)
marginality.06	-1.240^{***}	-2.080^{***}
	(0.261)	(0.117)
PAN.governor.06	-0.470^{*}	-0.312^{*}
	(0.271)	(0.167)
Constant	-1.914^{***}	-3.810^{***}
	(0.498)	(0.222)
Observations	2,407	2,407
Log Likelihood	−600.386	−645.606
Akaike Inf. Crit.		1,299.213
Note:	*p<0.1; **p<0.05; ***p<0.01	

```
Model: poisson, link: log

Response: PAN.visits.06

Terms added sequentially (first to last)


                      Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                                   2406    1473.87
competitive.district   1     0.91      2405    1472.96  0.34078
marginality.06         1   478.03      2404     994.93  < 2e-16 ***
PAN.governor.06        1     3.68      2403     991.25  0.05502
```

There is one outlier with visits = 35; predicted visits based on the model are 0.467. Excluding that district (ie assuming it's bad data) would lead to changes in the model, but would only change the prediction in c) from 0.015 to 0.016.

```
                                PAN.visits.06
                         outlier removed      default
------------------------------------------------------
competitive.district         −0.261            −0.081
                             (0.176)           (0.171)
marginality.06              −1.954***         −2.080***
                             (0.124)           (0.117)
PAN.governor.06              −0.129            −0.312*
                             (0.172)           (0.167)
Constant                    −3.727***         −3.810***
                             (0.227)           (0.222)
------------------------------------------------------
Observations                  2,406             2,407
Log Likelihood              −521.651          −645.606
Akaike Inf. Crit.          1,051.301         1,299.213

```