PS3_Dongli Wu_21362988

Dongli Wu_21362988

28/3/2022

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 GitHub in .pdf form.
- This problem set is due before class on Monday March 28, 2022. No late assignments will be accepted.
- Total available points for this homework is 80.

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

Please answer the following questions:

Question 1.1 Construct and interpret an unordered multinomial logit with GDPWdiff as the output and "no change" as the reference category, including the estimated cutoff points and coefficients.

```
# import data
setwd(getwd())
gdp <- read.csv("gdpChange.csv")

# observe the data information
summary(gdp)
```

```
##        X           COUNTRY         CTYNAME              YEAR
##  Min.   :   1   Min.   :  1.00   Length:3721        Min.   :1954
##  1st Qu.: 931   1st Qu.: 39.00   Class :character   1st Qu.:1967
##  Median :1861   Median : 71.00   Mode  :character   Median :1976
##  Mean   :1861   Mean   : 70.42                      Mean   :1975
##  3rd Qu.:2791   3rd Qu.:103.00                      3rd Qu.:1983
```

```
##  Max.   :3721   Max.   :135.00                      Max.   :1990
##       GDPW            OIL              REG             EDT
##  Min.   :  509   Min.   :0.0000   Min.   :0.0000   Length:3721
##  1st Qu.: 2566   1st Qu.:0.0000   1st Qu.:0.0000   Class :character
##  Median : 6425   Median :0.0000   Median :0.0000   Mode  :character
##  Mean   : 9276   Mean   :0.1005   Mean   :0.4015
##  3rd Qu.:13470   3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :37903   Max.   :1.0000   Max.   :1.0000
##     GDPWlag         GDPWdiff      GDPWdifflag       GDPWdifflag2
##  Min.   :  509   Min.   :-9257   Min.   :-9257.0   Min.   :-9257.0
##  1st Qu.: 2533   1st Qu.:  -24   1st Qu.:  -20.0   1st Qu.:  -19.0
##  Median : 6245   Median :  111   Median :  117.0   Median :  116.0
##  Mean   : 9090   Mean   :  186   Mean   :  189.7   Mean   :  189.9
##  3rd Qu.:13167   3rd Qu.:  415   3rd Qu.:  415.0   3rd Qu.:  405.0
##  Max.   :37089   Max.   : 7867   Max.   : 7867.0   Max.   : 7867.0
```

```
# select the useful columns
gdp1 <- gdp[, c("GDPWdiff", "REG", "OIL")]

# check type of GDPWdiff
class(gdp1$GDPWdiff)
```

```
## [1] "integer"
```

```
# use cut() to make the levels to 3 groups. Because all the value of GDPWdiff are integer so we can use
gdp1$GDPWdiff <- cut(gdp1$GDPWdiff, breaks = c(min(gdp1$GDPWdiff), -0.1,0, max(gdp1$GDPWdiff)), labels =

# check and make sure the levels are right now
levels(gdp1$GDPWdiff)
```

```
## [1] "negative"  "no change" "positive"
```

```
# set reference level as no change and convert the factor to un-ordered
gdp1$GDPWdiff <- relevel(gdp1$GDPWdiff, ref = "no change", ordered = FALSE)

# check the data information that we can see 1 NA in the column of GDPWdiff
summary(gdp1)
```

```
##       GDPWdiff          REG              OIL
##  no change:  16   Min.   :0.0000   Min.   :0.0000
##  negative :1104   1st Qu.:0.0000   1st Qu.:0.0000
##  positive :2600   Median :0.0000   Median :0.0000
##  NA's     :   1   Mean   :0.4015   Mean   :0.1005
##                   3rd Qu.:1.0000   3rd Qu.:0.0000
##                   Max.   :1.0000   Max.   :1.0000
```

```
# remove NA
gdp1 <- gdp1[!is.na(gdp1$GDPWdiff),]
```

- After clean and organize the data as above, we now get an unordered variable GDPWdiff with 3 catagories("negative", "no change","positive")as the output and also set "no change" as the reference category
- Let's run multinomial logit regression now

```
# run multinomial logit regression
mult_log <- multinom(GDPWdiff ~ REG + OIL, data = gdp1)
```

```
## # weights:  12 (6 variable)
## initial  value 4086.837714
## iter  10 value 2339.091519
## final  value 2338.396147
## converged
```

```
stargazer(mult_log, type = "text")
```

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                        negative      positive
##                          (1)            (2)
## ------------------------------------------------
## REG                     1.380*        1.769**
##                        (0.769)        (0.767)
##
## OIL                     4.758          4.557
##                        (6.823)        (6.823)
##
## Constant               3.805***      4.534***
##                        (0.271)        (0.269)
##
## ------------------------------------------------
## Akaike Inf. Crit.   4,688.792      4,688.792
## ================================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

Check coefficents and cutoff point

- for checking cutoff points, we can have a look the odds with CI for the parameters in our model.

```
# get coefficents
exp <- exp(coef(mult_log))
stargazer(exp, type = "text")
```

```
##
## ==================================
##           (Intercept)  REG    OIL
## ----------------------------------
## negative   44.934    3.976 116.492
## positive   93.118    5.867 95.344
## ----------------------------------
```

```
# get cutoff points with odds and CI
cutoff_points <- exp(confint(mult_log, alpa = 0.05))
cutoff_points
```

```
## , , negative
##
##                    2.5 %       97.5 %
## (Intercept) 2.643434e+01 7.638092e+01
## REG         8.810600e-01 1.794355e+01
## OIL         1.813256e-04 7.483968e+07
```

```
##
## , , positive
##
##                    2.5 %       97.5 %
## (Intercept) 5.493982e+01 1.578259e+02
## REG         1.304209e+00 2.638918e+01
## OIL         1.484877e-04 6.122003e+07
```

Interpretion ("no change" as reference)::

-Given the odds of REG, if we hold OIL constant, the odds of a Democracy country tending to be negative on GDPdiff rather than no change on GDPdiff is 3.976 times higher than the odds for a Non-Democracy country. The probability of value 3.796 will fall between 0.881 and 17.94 is 95% (CI)

- While, Democracy country has 5.867 times the odds of being positive on GDPdiff as opposed to no change on GDPdiff when other variables are held constant. There is a 95% probability that the value 5.867 will fall between 1.304 and 26.389.

-Given the odds of OIL, if we hold REG constant, the odds of a county whose fuel exports in 1984-86 exceeded 50% tending to be negative on GDPdiff rather than no change on GDPdiff is 116.492 times higher opposed to the odds for a country whose fuel exports less than 50% in the same period of time. The probability of value 116.492 will fall between 0.0001813 and 7483968 is 95% (CI)

While, a county whose fuel exports in 1984-86 exceeded 50% has 95.344 times the odds of being positive on GDPdiff as opposed no change on GDPdiff when other variables are held constant.There is a 95% probability that the value 95.0344 will fall between 0.000148 and 6122003.

- check P-values

```
#  get P-values
z <- summary(mult_log)$coefficients/summary(mult_log)$standard.errors
p <- (1 - pnorm(abs(z), 0, 1)) * 2
stargazer(p, type = "text")
```

```
##
## ================================
##           (Intercept)  REG   OIL
## --------------------------------
## negative      0      0.073 0.486
## positive      0      0.021 0.504
## --------------------------------
```

```
# generate the table of coefficents and P-values
```

Check P-Values:

- the predictor variable REG is statistically significant on positive result of outcome (GDP difference between the year and the year before.) because its P-Value =0.021 < 0.05.
- There is not sufficient evidence to show other variables are statistically significant becasue their P-Values are greater than 0.05.

Question 1.2 Construct and interpret an ordered multinomial logit with GDPWdiff as the outcome variable, including the estimated cutoff points and coefficients.

```
# select the useful columns
gdp2 <- gdp[, c("GDPWdiff", "REG", "OIL")]

# check type of GDPWdiff
class(gdp2$GDPWdiff)

# use cut() to make the levels to 3 groups. Because all the value of GDPWdiff are integer so we can use
gdp2$GDPWdiff <- cut(gdp2$GDPWdiff, breaks = c(min(gdp2$GDPWdiff), -0.1,0, max(gdp2$GDPWdiff)), labels =

# check and make sure the levels are right now
levels(gdp2$GDPWdiff)

# set reference level and convert the factor to ordered
gdp2$GDPWdiff <- relevel(gdp2$GDPWdiff, ref = TRUE)
as.ordered(gdp2$GDPWdiff)
```

```
# check the data information that we can see 1 NA in the column of GDPWdiff
summary(gdp2)
```

```
##       GDPWdiff         REG              OIL
##  negative :1104   Min.   :0.0000   Min.   :0.0000
##  no change:  16   1st Qu.:0.0000   1st Qu.:0.0000
##  positive :2600   Median :0.0000   Median :0.0000
##  NA's     :   1   Mean   :0.4015   Mean   :0.1005
##                   3rd Qu.:1.0000   3rd Qu.:0.0000
##                   Max.   :1.0000   Max.   :1.0000
```

```
# remove NA
gdp2 <- gdp2[!is.na(gdp2$GDPWdiff),]
```

- After clean and organize the data as above, we now get an ordered variable GDPWdiff with 3 catagories("negative", "no change","positive") as the output and the level" "negative" becomes reference category by default.
- Let's run multinomial logit regression now

```
# run multinomial logit regression
mult_log2 <- multinom(GDPWdiff ~ REG + OIL, data = gdp2)
```

```
## # weights:  12 (6 variable)
## initial  value 4086.837714
## iter  10 value 2338.391122
## final  value 2338.374498
## converged
```

```
stargazer(mult_log2, type = "text")
```

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                       no change      positive
##                          (1)            (2)
## ------------------------------------------------
```

```
## REG                  -1.352*       0.389***
##                      (0.758)        (0.076)
##
## OIL                   -7.919        -0.200*
##                      (33.018)       (0.116)
##
## Constant            -3.801***      0.729***
##                      (0.270)        (0.048)
##
## ------------------------------------------------
## Akaike Inf. Crit.   4,688.749      4,688.749
## ================================================
## Note:              *p<0.1; **p<0.05; ***p<0.01
```

Check coefficents and cutoff point

- for checking cutoff points, we can have a look the odds with CI for the parameters in our model.

```
# get coefficents
exp2 <- exp(coef(mult_log2))
stargazer(exp2, type = "text")
```

```
##
## ==================================
##           (Intercept)  REG   OIL
## ----------------------------------
## no change    0.022    0.259 0.0004
## positive     2.072    1.476 0.818
## ----------------------------------
```

```
# get cutoff points with odds and CI
cutoff_points2 <- exp(confint(mult_log2, alpa = 0.05))
cutoff_points2
```

```
## , , no change
##
##                     2.5 %       97.5 %
## (Intercept) 1.315957e-02 3.794583e-02
## REG         5.850537e-02 1.143338e+00
## OIL         2.860570e-32 4.630330e+24
##
## , , positive
##
##                 2.5 %   97.5 %
## (Intercept) 1.8866713 2.276374
## REG         1.2726786 1.711212
## OIL         0.6519229 1.027473
```

Interpretion ("negative" as reference):

-Given the odds of REG, if we hold OIL constant, the odds of a Democracy country tending to be no change on GDPdiff rather than negative on GDPdiff is 0.259 times higher than the odds for a Non-Democracy country. The probability of value 0.259 will fall between 0.0585 and 1.143 is 95% (CI).

- While, Democracy country has 1.476 times the odds of being positive on GDPdiff as opposed to negative on GDPdiff when other variables are held constant. There is a 95% probability that the value 1.476 will fall between 1.273 and 1.711.

-Given the odds of OIL, if we hold REG constant, the odds of a county whose fuel exports in 1984-86 exceeded 50% tending to be no change on GDPdiff rather than negative on GDPdiff is 0.0004 times higher opposed to the odds for a country whose fuel exports less than 50% in the same period of time. The probability of value 0.0004 will fall between 2.860570e-32 and 4.630330e+24 is 95% (CI)

- While, a county whose fuel exports in 1984-86 exceeded 50% has 0.818 times the odds of being positive on GDPdiff as opposed negative on GDPdiff when other variables are held constant.There is a 95% probability that the value 90.818 will fall between 0.652 and 1.027.

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 can- didate 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 previ- ous 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 ad- ditional 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.

```
# import the data
dt <- read.csv("MexicoMuniData.csv")

# select useful variables
dt <- dt[, c("PAN.visits.06", "competitive.district", "marginality.06", "PAN.governor.06")]

summary(dt)
```

```
##  PAN.visits.06      competitive.district marginality.06      PAN.governor.06
##  Min.   : 0.00000   Min.   :0.0000       Min.   :-2.270000   Min.   :0.0000
##  1st Qu.: 0.00000   1st Qu.:1.0000       1st Qu.:-0.746000   1st Qu.:0.0000
##  Median : 0.00000   Median :1.0000       Median :-0.051000   Median :0.0000
##  Mean   : 0.09182   Mean   :0.8214       Mean   :-0.001373   Mean   :0.2152
##  3rd Qu.: 0.00000   3rd Qu.:1.0000       3rd Qu.: 0.628500   3rd Qu.:0.0000
##  Max.   :35.00000   Max.   :1.0000       Max.   : 3.355000   Max.   :1.0000
```

```
str(dt)
```

```
## 'data.frame':    2407 obs. of  4 variables:
##  $ PAN.visits.06       : int  5 0 0 0 0 0 0 0 0 0 ...
##  $ competitive.district: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ marginality.06      : num  -1.831 -0.62 -0.875 -0.747 -1.234 ...
##  $ PAN.governor.06     : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
# run Poisson Regression Model
dt_p <- glm (PAN.visits.06 ~ competitive.district + marginality.06 + PAN.governor.06, data=dt, family=p
stargazer(dt_p, type = "text")
```

```
##
```

```
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                             PAN.visits.06
## ------------------------------------------------
## 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
```

- P-Value of competitive.district is 0.6336, much greater than 0.05, so it doesn't have a statistally significant impact on our outcome PAN presidential candidates visit swing districts.

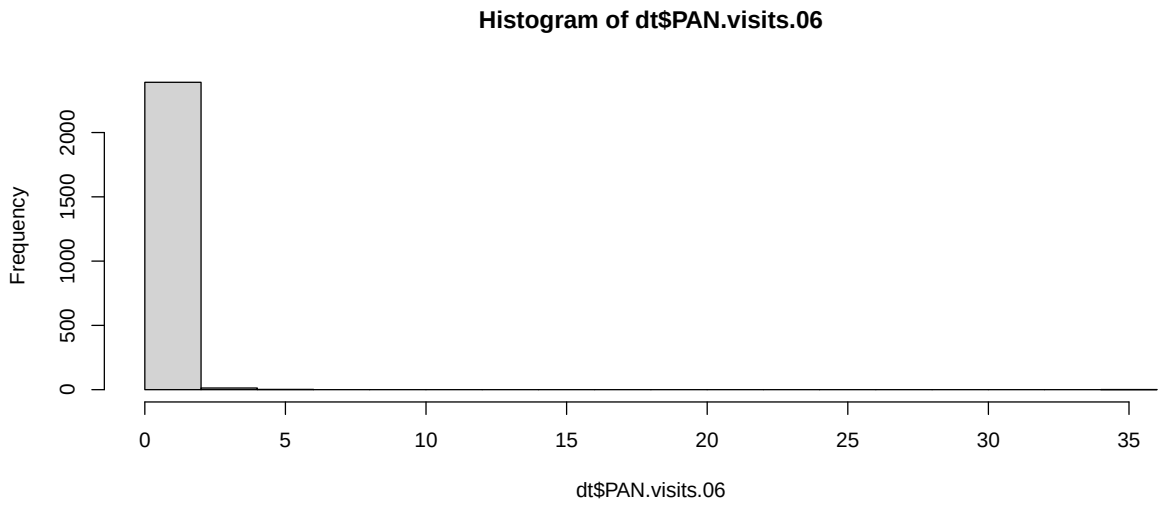
```
summary(dt_p)
```

```
##
## Call:
## glm(formula = PAN.visits.06 ~ competitive.district + marginality.06 +
##     PAN.governor.06, family = poisson, data = dt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2309  -0.3748  -0.1804  -0.0804  15.2669
##
## 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 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1473.87  on 2406  degrees of freedom
## Residual deviance:  991.25  on 2403  degrees of freedom
## AIC: 1299.2
##
## Number of Fisher Scoring iterations: 7
```

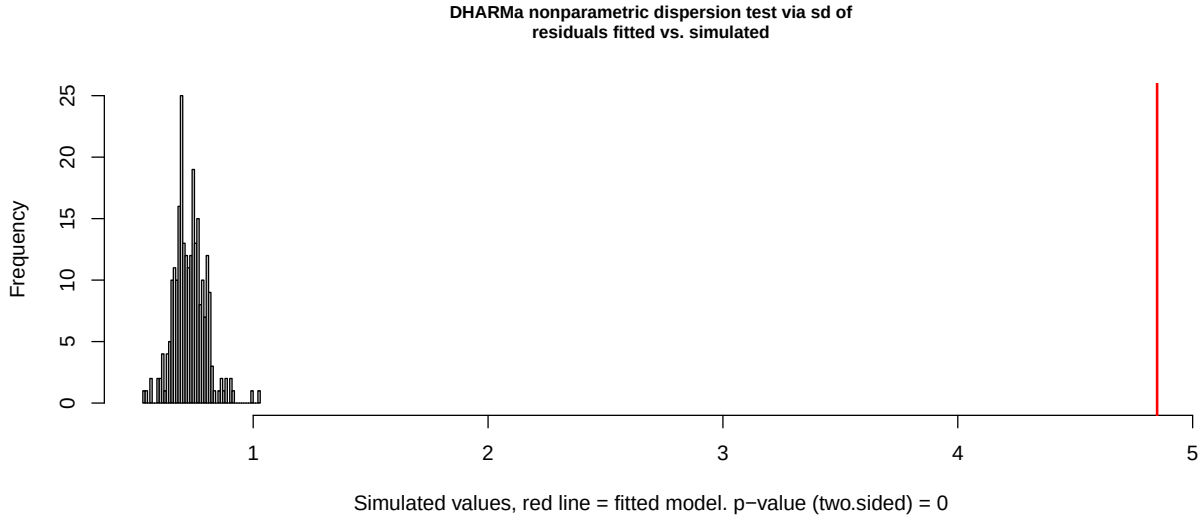
- as we can see from summary(dt), many 0 values in variable, let's have a look if Zero-Inflated model is

better, and if it is over dispersion

```
# check the histogram of PAN. visits.06
hist(dt$PAN.visits.06)
```




```
#  Let's check if it is over-dispersion
testDispersion(dt_p)
```




```
##
##  DHARMa nonparametric dispersion test via sd of residuals fitted vs.
##  simulated
##
## data:  simulationOutput
## dispersion = 6.6474, p-value < 2.2e-16
## alternative hypothesis: two.sided
```

- Seems Zero-Inflated Model is better than Poisson becasue there are more 0 values would show up with the increasement of visit.
- dispersion = 6.607, p-value < 2.2e-16, so we know it is true over dispension. ### Therefore, Zero-

Inflated Model should be considered.

```
dt_zip <- zeroinfl(PAN.visits.06 ~ competitive.district + marginality.06 + PAN.governor.06, data=dt, di
summary(dt_zip)
```

```
##
## Call:
## zeroinfl(formula = PAN.visits.06 ~ competitive.district + marginality.06 +
##     PAN.governor.06, data = dt, dist = "poisson")
##
## Pearson residuals:
##      Min       1Q   Median       3Q      Max
## -0.95323 -0.24006 -0.12842 -0.06045 37.56114
##
## Count model coefficients (poisson with log link):
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -1.9145     0.4982  -3.843 0.000122 ***
## competitive.district   0.4024     0.3119   1.290 0.197028
## marginality.06        -1.2398     0.2610  -4.750 2.03e-06 ***
## PAN.governor.06       -0.4703     0.2707  -1.737 0.082341 .
##
## Zero-inflation model coefficients (binomial with logit link):
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)            1.2719     0.6753   1.883  0.05966 .
## competitive.district   0.9000     0.5106   1.763  0.07794 .
## marginality.06         0.8716     0.3021   2.885  0.00392 **
## PAN.governor.06       -0.1749     0.4119  -0.425  0.67106
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 16
## Log-likelihood: -600.4 on 8 Df
```

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

```
coeff <- coefficients(dt_p)
coeff_ecp <- exp(coeff)
stargazer(coeff, type = "text")
```

```
##
## ===================================================================
## (Intercept) competitive.district marginality.06 PAN.governor.06
## -------------------------------------------------------------------
## -3.810             -0.081            -2.080         -0.312
## -------------------------------------------------------------------
```

```
stargazer(coeff_ecp, type = "text")
```

```
##
## ===================================================================
## (Intercept) competitive.district marginality.06 PAN.governor.06
## -------------------------------------------------------------------
## 0.022              0.922             0.125           0.732
## -------------------------------------------------------------------
```

Interpretion:

- In Poisson model, marginality P-value is much smaller than 0.05, so it is statistically significant impact visit in negative way. While PAN governor P-value is 0.0617, so it is not sufficient to say governor is a significant predictor of visit.

-let's have a look at their parameters

- We can see from the table above, when marginality increases by 1 unit, the visit value decrease by 0.125, as the exponent value is -2.08.
- when the state changes from without a PAN-affilated governor to have a PAN-affiliated governor, the visit value decrease by 0.732, as the exponent value is -0.312.

(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).

```
# calculate the exp
exp(coeff[1] + coeff[2]*1 + coeff[3]*0 + coeff[4]*1)
```

```
## (Intercept)
##  0.01494818
```

- The estimated mean number of visits is 0.0149