Homework_3

Joshua Ingram

9/17/2020

Problem 1

1.

Note: I am using λ instead of μ for my notation in this homework. I've been doing other work with the poisson distribution and am using λ for notation for that. It's just easier to stay consistent all around. Let me know if I need to stick with μ after this.

$$Y_i \sim_{ind.} Pois(\lambda),\ i = 1, 2, ..., 248,\ \text{where } \lambda \text{is the average number of interlocks}$$
$$log(\lambda_i) = \beta_0 + \beta_1(assets_i) + \beta_2 I_{nationOTH,i} + \beta_3 I_{nationUK,i} + \beta_4 I_{nationUS,i} +$$
$$\beta_5(I_{nationOTH,i} * assets_i) + \beta_6(I_{nationUK,i} * assets_i) + \beta_7(I_{nationUS,i} * assets_i)$$
$$\text{CAN is baseline category, } I_{nationOTH} \in \{0 = not\ other\ foreign, 1 = other\ foreign\}$$
$$I_{nationUK} \in \{0 = not\ UK, 1 = UK\},\ I_{nationUS} \in \{0 = not\ US, 1 = US\}$$

2.

```r
fit_1 <- glm(interlocks ~ assets + nation + nation:assets, family = poisson, data = ornstein)
summary(fit_1)
```

```
##
## Call:
## glm(formula = interlocks ~ assets + nation + nation:assets, family = poisson,
##     data = ornstein)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.8166  -2.7387  -0.9006   1.9493   9.1197
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       2.724e+00  2.430e-02 112.095  < 2e-16 ***
## assets            1.490e-05  4.426e-07  33.672  < 2e-16 ***
## nationOTH        -2.042e-01  9.576e-02  -2.132    0.033 *
## nationUK         -1.272e+00  1.610e-01  -7.902 2.73e-15 ***
## nationUS         -1.072e+00  5.444e-02 -19.697  < 2e-16 ***
## assets:nationOTH  3.353e-05  2.310e-05   1.451    0.147
```

```
## assets:nationUK   4.131e-04  6.937e-05   5.955 2.60e-09 ***
## assets:nationUS   6.157e-05  5.673e-06  10.854  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 3737.0  on 247  degrees of freedom
## Residual deviance: 2116.2  on 240  degrees of freedom
## AIC: 3030.2
##
## Number of Fisher Scoring iterations: 5
```

Test for Interaction:

$H_0: \ \beta_5 = \beta_6 = \beta_7 = 0$

$H_A: \ \{\exists \ \beta_j \neq 0 \mid j = 5, 6, 7\}$

$\alpha = 0.05$

Null Model: $Y_i \sim_{ind.} Pois(\lambda), \ i = 1, 2, ..., 248, \ log(\lambda_i) = \beta_0 + \beta_1(assets_i) + \beta_2 I_{nationOTH,i} + \beta_3 I_{nationUK,i} + \beta_4 I_{nationUS,i}$

LRT statistic $= 2log(\frac{L_1}{L_0}) = G_0^2 \sim \chi_3^2$

p-value: $P(\chi_3^2 \geq G_0^2)$

```r
fit_null <-  glm(interlocks ~ assets + nation, family = poisson, data = ornstein)
anova(fit_null, fit_1, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: interlocks ~ assets + nation
## Model 2: interlocks ~ assets + nation + nation:assets
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       243     2248.9
## 2       240     2116.2  3   132.65 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After performing the likelihood ratio test, we get an extremely small p-value (basically 0). We have significant evidence to reject the null hypothesis and our interaction term is statisically significant.

3.

```r
1- summary(fit_1)$deviance/summary(fit_1)$null.deviance
```

```
## [1] 0.433715
```

$R^2 = 0.433715$

43.4% of the variation in our response, the number of interlocks, is explained by our model.

4.

For a firm with the U.S. as the nation of control, per 1 million dollar increase in assets, the average number of interlocks will increase by a factor of $e^{7.647e-7}$. If the U.S. controlled firm has 0 dollars in assets, the average number of interlocks will be $e^{1.072e+00}$ time LESS than a Canadian controlled firm.

I'm not specifying "ceteris paribus" since nation and assets are our only two variables in the model and these are being explicitly addressed in the interpretation.

Problem 2

1.

```r
fit_2 <- glm(visits ~ chronic + age + gender + income + insurance, family = poisson, data = nmes)
summary(fit_2)
```

```
##
## Call:
## glm(formula = visits ~ chronic + age + gender + income + insurance,
##     family = poisson, data = nmes)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -6.0349  -2.0695  -0.7102   0.7390  17.6511
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.491e+00  7.728e-02  19.296  < 2e-16 ***
## chronic       2.038e-01  4.113e-03  49.562  < 2e-16 ***
## age          -3.278e-02  1.009e-02  -3.249  0.00116 **
## gendermale   -1.154e-01  1.304e-02  -8.849  < 2e-16 ***
## income       -5.927e-05  2.163e-03  -0.027  0.97814
## insuranceyes  2.464e-01  1.620e-02  15.210  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 26943  on 4405  degrees of freedom
## Residual deviance: 24438  on 4400  degrees of freedom
## AIC: 37225
##
## Number of Fisher Scoring iterations: 5
```

a.

Yes, there is evidence of overdispersion because the residual deviance is much greater than the residual degrees of freedom. This means that our variance is greater than expected, which under a poisson model, should be the same as the mean.

b.

We should use a Quasi-Poisson model.

```
fit_quasi <- glm(visits ~ chronic + age + gender + income + insurance, family = quasipoisson, data = nme
summary(fit_quasi)
```

```
##
## Call:
## glm(formula = visits ~ chronic + age + gender + income + insurance,
##     family = quasipoisson, data = nmes)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -6.0349  -2.0695  -0.7102   0.7390  17.6511
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.491e+00  2.091e-01   7.130 1.17e-12 ***
## chronic       2.038e-01  1.113e-02  18.313  < 2e-16 ***
## age          -3.278e-02  2.731e-02  -1.201  0.22996
## gendermale   -1.154e-01  3.528e-02  -3.270  0.00108 **
## income       -5.927e-05  5.853e-03  -0.010  0.99192
## insuranceyes  2.464e-01  4.385e-02   5.620 2.03e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 7.324197)
##
##     Null deviance: 26943  on 4405  degrees of freedom
## Residual deviance: 24438  on 4400  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

Differences:

The standard errors are greater for the quasi-poisson, thus affecting the t-statistics and p-values given for each beta. The inverse is true for the regular poisson model.

Similarities:

Both the quasi-poisson model and the poisson model have the same exact estimates for the betas and the Null/Residual deviances are the same (and degrees of freedom).

c.

```
Anova(fit_2)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: visits
```

```
##           LR Chisq Df Pr(>Chisq)
## chronic    2255.50  1  < 2.2e-16 ***
## age          10.62  1   0.001119 **
## gender       78.97  1  < 2.2e-16 ***
## income        0.00  1   0.978132
## insurance   242.47  1  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
Anova(fit_quasi)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: visits
##           LR Chisq Df Pr(>Chisq)
## chronic    307.952  1  < 2.2e-16 ***
## age          1.450  1   0.228535
## gender      10.782  1   0.001025 **
## income       0.000  1   0.991919
## insurance   33.105  1  8.731e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With the quasi-poisson model I would drop the age and incoe predictors, but for the regular poisson model I would drop only the income predictor.

2.

a.

By looking at the distribution of counts of interlocks, we may be running into an issue of having too many zero-counts of interlocks. It may be more appropriate to fit a zero-inflated poisson model to our data.

b.

Full GLM Model Formulation:

$$p_i = P(Y_i \in \text{"}no\ visists)$$
$$log(\frac{p_i}{1-p_i}) = \gamma_0 + \gamma_1 chronic_i + \gamma_2 age_i + \gamma_3 I_{gender,i} + \gamma_4 income_i + \gamma_5 I_{insurance,i}$$
$$Y_i \sim_{ind.} Pois(\lambda_i),\ \text{where } \lambda \text{ average number of physician visits}$$
$$log(\lambda_i)\beta_0 + \beta_1 chronic_i + \beta_2 age_i + \beta_3 I_{gender,i} + \beta_4 income_i + \beta_5 I_{insurance,i}$$
$$I_{gender} \in \{0 = female, 1 = male\},\ I_{insurance} \in \{0 = no, 1 = yes\}$$

c.

```r
fit_3 <- zeroinfl(visits ~ chronic + age + gender + income + insurance, data = nmes)
summary(fit_3)
```

```
##
## Call:
## zeroinfl(formula = visits ~ chronic + age + gender + income + insurance,
##     data = nmes)
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -3.8761 -1.1927 -0.5008  0.5634 24.5517
##
## Count model coefficients (poisson with log link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.929459   0.079075  24.400  < 2e-16 ***
## chronic       0.153552   0.004298  35.728  < 2e-16 ***
## age          -0.046298   0.010301  -4.494 6.98e-06 ***
## gendermale   -0.054136   0.013140  -4.120 3.79e-05 ***
## income       -0.003732   0.002222  -1.680    0.093 .
## insuranceyes  0.107787   0.016391   6.576 4.84e-11 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.37284    0.54343   0.686   0.4927
## chronic      -0.55759    0.04332 -12.872  < 2e-16 ***
## age          -0.11425    0.07153  -1.597   0.1102
## gendermale    0.42225    0.08892   4.749 2.05e-06 ***
## income       -0.03572    0.01909  -1.872   0.0613 .
## insuranceyes -0.88248    0.09706  -9.092  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 17
## Log-likelihood: -1.668e+04 on 12 Df
```

```r
Anova(fit_3)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: visits
##           Df     Chisq Pr(>Chisq)
## chronic    1 1276.4630  < 2.2e-16 ***
## age        1   20.1992  6.978e-06 ***
## gender     1   16.9743  3.789e-05 ***
## income     1    2.8218    0.09299 .
## insurance  1   43.2411  4.839e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We use the Anova() function to test the predictor as a whole. Although the name was not explicitly given in class, the test statistic follows a chi-square distribution.

For age:

$H_0 : \gamma_2 = \beta_2 = 0$

$H_0 : \{\exists\ \beta_2\ or\ \gamma_2 \neq 0\}$

$\alpha = 0.05$

income is statistically insignificant based on the output from Anova().

d.

```
fit_4 <- zeroinfl(visits ~ chronic + age + gender + insurance, data = nmes)
summary(fit_4)
```

```
##
## Call:
## zeroinfl(formula = visits ~ chronic + age + gender + insurance, data = nmes)
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -3.8765 -1.1910 -0.5018  0.5649 24.5887
##
## Count model coefficients (poisson with log link):
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.917709   0.078744  24.354  < 2e-16 ***
## chronic       0.153913   0.004293  35.852  < 2e-16 ***
## age          -0.045515   0.010287  -4.424 9.68e-06 ***
## gendermale   -0.056966   0.013036  -4.370 1.24e-05 ***
## insuranceyes  0.103952   0.016237   6.402 1.53e-10 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.24524    0.53818   0.456    0.649
## chronic      -0.55392    0.04321 -12.821  < 2e-16 ***
## age          -0.10454    0.07119  -1.468    0.142
## gendermale    0.40163    0.08823   4.552 5.31e-06 ***
## insuranceyes -0.91887    0.09537  -9.635  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 15
## Log-likelihood: -1.669e+04 on 10 Df
```

Chronic:

logit - For every one additional chronic conditions, the odds of an individual not having any physician visits decrease by a factor of $e^{0.554}$, ceteris paribus.

poisson - For every one additional chronic condition, the average number of physician visits will increase by a factor of $e^{0.154}$, ceteris paribus.

Insurance:

logit - For a person with insurance, the odds of them having no physician visits is $e^{0.92}$ times lower than those that have no insurance, ceteris paribus.

poisson - For a person with insurance, the average number of physician visits is $e^{0.11}$ times greater than those without insurance, ceteris paribus.