**PM592: Regression Analysis for Health Data Science**
**Lab 11 – Regression for Count Outcomes**
**Data Needed:** -

**Outline**

- Estimated Marginal Means

In today's lab we will be using the `emmeans` package, which is useful when presenting model results. With this package you can:

- Perform post-hoc pairwise comparisons among groups when you find a significant dummy variable set.
- Get the estimated value of Y for certain values of an X variable, averaged over the levels of all other X variables in the data set.
- Graphically display the results of your model, highlighting the effects of particular X variables of interest.

The following vignettes are useful for learning about the emmeans package:

- https://cran.r-project.org/web/packages/emmeans/vignettes/basics.html
- https://cran.r-project.org/web/packages/emmeans/vignettes/comparisons.html
- https://cran.r-project.org/web/packages/emmeans/vignettes/interactions.html

Note that the estimated marginal means approach can be used with <u>any</u> type of GLM (not just Poisson regression).

## Lab 11 Exercises

| Objective(s): | Use familiar regression techniques, applied to count outcomes |
|---|---|
| Datasets Required: | `hospitaler.dta` |

This data set contains temporal trends in LAC+USC hospital admission from the ER over a 3-year period. Suppose hospital administration asked us to determine if the rate of hospital admissions for those visiting the ER is changing over time. They want to know if, and how, this rate is changing.

| Variable | Description |
|---|---|
| `month` | Month number (0-35) |
| `mo4` | Month, grouped into 4-month aggregate bins |
| `er_visit` | # of individuals who are seen by the ER per month |
| `admit` | # of individuals who are seen by the ER per month, who subsequently are admitted to the hospital |
| `readmit` | # of individuals who are seen by the ER per month, who subsequently are admitted to the hospital, who had been discharged from the hospital in the past 30 days |
| `hosp_ed` | |

1) Perform exploratory data analysis.
   a) For each 4-month group, and overall, examine the mean and variance of `er_visit` and `admit`. Does the mean appear to be roughly equal to the variance? (We do this by 4-month grouping so we can get an estimate of the variance and thus evaluate the Poisson distribution assumption over time.)

```
> er %>%
+   group_by(mo4) %>%
+   summarise(
+     mean_er = mean(er_visit, na.rm=T),
+     var_er = var(er_visit, na.rm=T),
+     mean_ad = mean(admit, na.rm=T),
+     var_ad = var(admit, na.rm=T),
+     mean_read = mean(readmit, na.rm=T),
+     var_read = var(readmit, na.rm=T)
+   )
# A tibble: 9 x 7
    mo4 mean_er   var_er mean_ad var_ad mean_read var_read
  <dbl>   <dbl>    <dbl>   <dbl>  <dbl>     <dbl>    <dbl>
1     0  10538.   51396.   1593.  3093.      210.     160.
2     1  11064.   46858.   1652.  3185       228.     107
3     2  10680.  230161    1501.   372.      187.     332.
4     3  11298.   91968.   1600.   252.      186.     428.
5     4  11323    15553.   1472.  4114.      169      230
6     5  10970. 1042321.   1354   4743.      160.     240.
7     6  11742.   52497.   1416.   873.      163.     185.
8     7  11479.  156543.   1474.  2107       162      451.
9     8  10541.  597828.   1507.  7758.      187.     142.
```

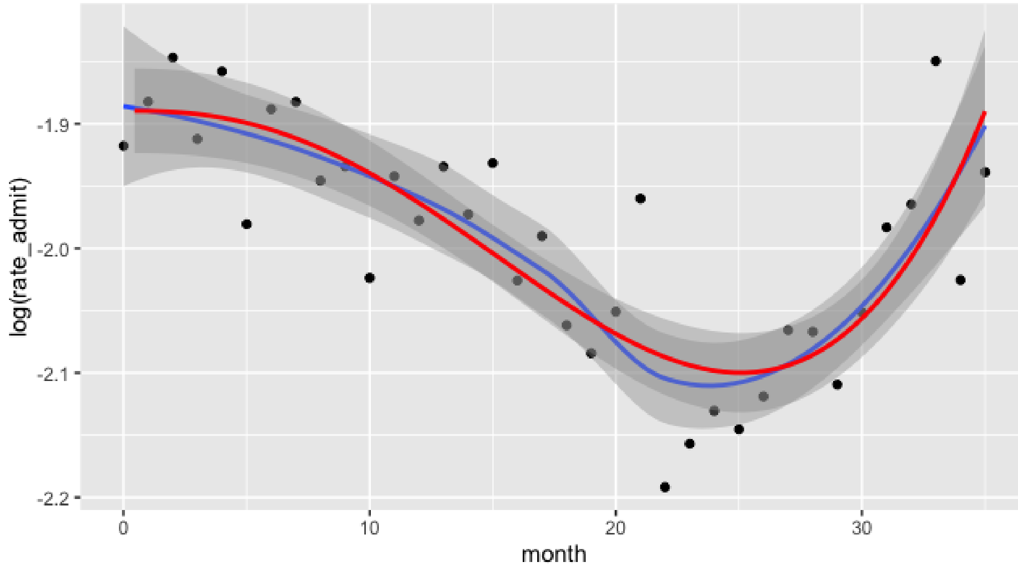
For ER visits and hospital admissions, the variance seems to be higher than the mean for most 4-month groups. For readmissions, the mean appears to be roughly equal to the variance.

b) Add a variable that indicates the rate at which individuals are admitted to the hospital from the ER.

```
> er <- er %>%
+   mutate(rate_admit = admit / er_visit)
```

c) Create a scatter plot of the (log) hospital admission rate vs. month. How does this relationship look in terms of linearity?

```
> ggplot(data = er, aes(x=month, y=log(rate_admit))) +
+   geom_point() +
+   geom_smooth() +
+   geom_smooth(method="lm", formula="y~I(x^3) + I(x^3*log(x))", col
or="red")
```




The relationship between (log) hospital admission rate and month does not look linear.

*Thoughts: we could either 1) create a very well-fit model using some higher-order polynomial terms or 2) group the data together and present the results from a dummy-variable predictor model. I'm going to proceed with the latter because it will be more interpretable when presenting the results to hospital administration.*

2) Fit the Poisson model for the effect of 4-month time period on <u>rate</u> of hospital admissions from the ER.

   a) Write the equation for the best fit model in terms of expected number of admissions.

```
Call:
glm(formula = admit ~ factor(mo4), family = poisson, data = er)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6686  -0.9877   0.1304   0.7198   2.7815

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   7.373217   0.012528 588.520  < 2e-16 ***
factor(mo4)1  0.036222   0.017560   2.063  0.03913 *
factor(mo4)2 -0.059164   0.017986  -3.289  0.00100 **
factor(mo4)3  0.004854   0.017696   0.274  0.78386
factor(mo4)4 -0.078500   0.018076  -4.343 1.41e-05 ***
factor(mo4)5 -0.162399   0.018482  -8.787  < 2e-16 ***
factor(mo4)6 -0.117450   0.018262  -6.432 1.26e-10 ***
factor(mo4)7 -0.077822   0.018073  -4.306 1.66e-05 ***
factor(mo4)8 -0.055507   0.017969  -3.089  0.00201 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 241.90  on 35  degrees of freedom
Residual deviance:  53.35  on 27  degrees of freedom
AIC: 400.9

Number of Fisher Scoring iterations: 3
```

The equation for the best fit model is: $\ln(\hat{\mu}_{Y|X}) = 7.37 + 0.036X_{mo1} - 0.059X_{mo2} + 0.0049X_{mo3} - 0.079X_{mo4} - 0.16X_{mo5} - 0.12X_{mo6} - 0.078X_{mo7} - 0.056X_{mo8}$

b) Write the equation for the best fit model in terms of expected rate of admissions.

```
Call:
glm(formula = admit ~ factor(mo4) + offset(log(er_visit)), family =
poisson,
    data = er)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.7752  -0.9834   0.1486   0.9512   4.8561

Coefficients:
             Estimate Std. Error  z value Pr(>|z|)
(Intercept)  -1.88950    0.01253 -150.817  < 2e-16 ***
factor(mo4)1 -0.01249    0.01756   -0.711 0.476971
factor(mo4)2 -0.07262    0.01799   -4.038  5.4e-05 ***
factor(mo4)3 -0.06483    0.01770   -3.663 0.000249 ***
factor(mo4)4 -0.15037    0.01808   -8.319  < 2e-16 ***
factor(mo4)5 -0.20258    0.01848  -10.961  < 2e-16 ***
factor(mo4)6 -0.22564    0.01826  -12.356  < 2e-16 ***
factor(mo4)7 -0.16336    0.01807   -9.039  < 2e-16 ***
factor(mo4)8 -0.05584    0.01797   -3.108 0.001886 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 450.98  on 35  degrees of freedom
Residual deviance: 125.66  on 27  degrees of freedom
AIC: 473.21

Number of Fisher Scoring iterations: 3
```

$\ln\left(\frac{\hat{\mu}_{Y|X}}{X_{ervisit}}\right) = -1.89 - 0.012X_{mo1} - 0.072X_{mo2} - 0.06X_{mo3} - 0.15X_{mo4} - 0.20X_{mo5} - 0.23X_{mo6} - 0.16X_{mo7} - 0.056X_{mo8}$

c) Explain how rate of hospital admissions has changed over time. Provide rate ratios, 95% confidence intervals, and p-values.

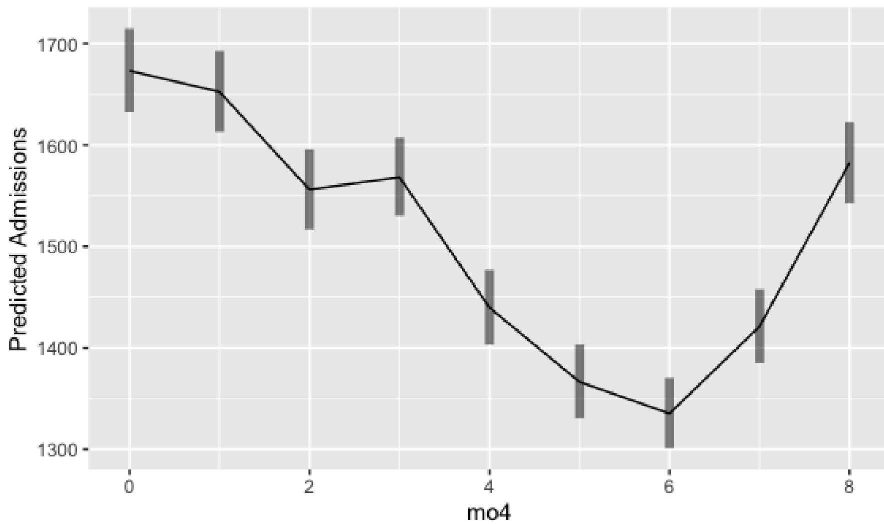
```
> tibble(parameter = names(er.m$coefficients),
+        rr = exp(er.m$coefficients),
+        as.data.frame.matrix(exp(confint.default(er.m)))
+        )
# A tibble: 9 x 4
  parameter       rr `2.5 %` `97.5 %`
  <chr>        <dbl>   <dbl>    <dbl>
1 (Intercept)  0.151   0.147    0.155
2 factor(mo4)1 0.988   0.954    1.02
3 factor(mo4)2 0.930   0.898    0.963
4 factor(mo4)3 0.937   0.905    0.970
5 factor(mo4)4 0.860   0.830    0.891
6 factor(mo4)5 0.817   0.788    0.847
7 factor(mo4)6 0.798   0.770    0.827
8 factor(mo4)7 0.849   0.820    0.880
9 factor(mo4)8 0.946   0.913    0.980
```

The expected rate of hospital admissions for month group 1 is 0.988 times the rate of hospital admissions in month group 0 (CI = (0.954, 1.02), p=0.48). The expected rate of hospital admissions for month group 2 is 0.930 times the rate of hospital admissions in month group 0 (CI = (0.898, 0.963), p<0.01). The expected rate of hospital admissions for month group 3 is 0.937 times the rate of hospital admissions in month group 0 (CI = (0.905, 0.970), p=0.00025). The expected rate of hospital admissions for month group 4 is 0.860 times the rate of hospital admissions in month group 0 (CI = (0.830, 0.891), p<0.01). The expected rate of hospital admissions for month group 5 is 0.817 times the rate of hospital admissions in month group 0 (CI = (0.788, 0.830), p=0.902). The expected rate of hospital admissions for month group 6 is 0.798 times the rate of hospital admissions in month group 0 (CI = (0.770, 0.827), p<0.01). The expected rate of hospital admissions for month group 7 is 0.849 times the rate of hospital admissions in month group 0 (CI = (0.820, 0.880), p<0.01). The expected rate of hospital admissions for month group 8 is 0.946 times the rate of hospital admissions in month group 0 (CI = (0.913, 0.980), p=0.0019).

3) Compute the predicted number of hospital admissions and the predicted rate of hospital admissions. You can use either the approach from the class notes or the emmeans package.

a) Compute the predicted number of hospital admissions in each 4-month period. Plot the predicted number of admissions, with 95% confidence intervals, in each group.

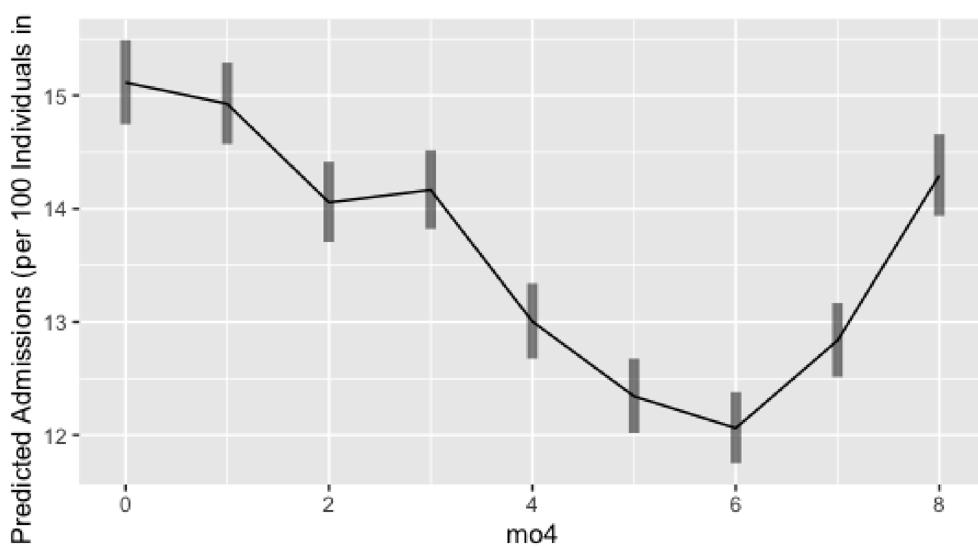
```
> emmeans(er.m, "mo4", type = "response")
 mo4 rate   SE  df asymp.LCL asymp.UCL
   0 1673 21.0 Inf      1633      1715
   1 1653 20.3 Inf      1613      1693
   2 1556 20.1 Inf      1517      1596
   3 1568 19.6 Inf      1530      1607
   4 1440 18.8 Inf      1403      1477
   5 1366 18.6 Inf      1331      1403
   6 1335 17.7 Inf      1301      1371
   7 1421 18.5 Inf      1385      1458
   8 1582 20.4 Inf      1543      1623
```




*Note: If you use emmeans, the predictions will be based on the average value of the offset across all observations. Thus this procedure provides an "adjusted" expected number of admissions holding the offset constant across months.*

b) Compute the predicted rate of hospital admissions per 100 individuals seen in the ER in each 4-month period by specifying the offset to be log(100). Plot the predicted rate of admissions (per 100), with 95% confidence intervals, in each group.

```
> emmeans(er.m, "mo4", type = "response", offset = (log(100)))
 mo4 rate    SE  df asymp.LCL asymp.UCL
   0 15.1 0.189 Inf      14.7      15.5
   1 14.9 0.184 Inf      14.6      15.3
   2 14.1 0.181 Inf      13.7      14.4
   3 14.2 0.177 Inf      13.8      14.5
   4 13.0 0.169 Inf      12.7      13.3
   5 12.3 0.168 Inf      12.0      12.7
   6 12.1 0.160 Inf      11.8      12.4
   7 12.8 0.167 Inf      12.5      13.2
   8 14.3 0.184 Inf      13.9      14.7
```




4) Assess the model goodness of fit.

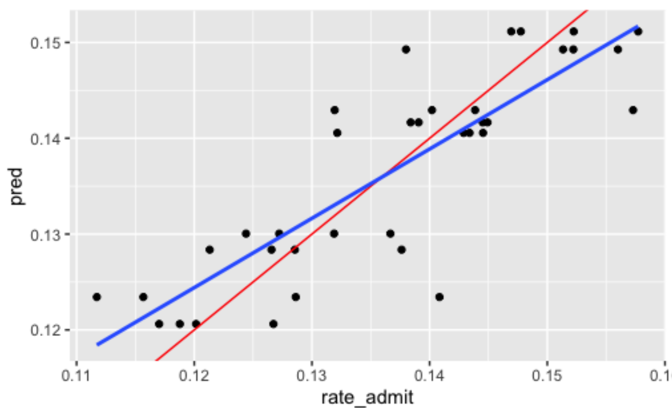
a) Provide the Pearson GOF test statistic and p-value.

```
> pois_pearson_gof(er.m)
$pval
[1] 8.771802e-15

$df
[1] 27
```

b) Provide the Deviance GOF test statistic and p-value.

```
> pois_dev_gof(er.m)
$pval
[1] 1.125712e-14

$df
[1] 27
```

c) Plot the expected number of hospital admissions vs. the observed number of hospital admissions. What are your impressions?




The fit is not that good, the line of expected rate of admissions is different from the line of the observed rate of admissions.

d) Check for evidence of model overdispersion.

```
> AER::dispersiontest(er.m)

        Overdispersion test

data:  er.m
z = 2.8587, p-value = 0.002127
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
   3.50631
```

There is evidence to suggest that overdispersion parameter is not equal to (greater than) 1. The dispersion parameter is 3.5.

5) **Extra Practice 1.** Upon finding evidence of model overdispersion, re-fit this model as a negative binomial model.

```
> MASS::glm.nb(admit ~ factor(mo4) + offset(log(er_visit)), data =
r)

Call:  MASS::glm.nb(formula = admit ~ factor(mo4) + offset(log(er_v
sit)),
    data = er, init.theta = 597.5546774, link = log)

Coefficients:
 (Intercept)  factor(mo4)1  factor(mo4)2  factor(mo4)3
    -1.88941      -0.01201      -0.07175      -0.06458
factor(mo4)4  factor(mo4)5  factor(mo4)6  factor(mo4)7
    -0.15047      -0.19824      -0.22542      -0.16265
factor(mo4)8
    -0.05396

Degrees of Freedom: 35 Total (i.e. Null);  27 Residual
Null Deviance:      127.5
Residual Deviance: 36.25        AIC: 431.1
```

6) **Extra Practice 2.** The hospital administration was also interested in whether the rate of individuals readmitted to the hospital is changing over time. Fit another Poisson model for the number of readmissions, with number of admissions as the offset.

```
Call:
glm(formula = readmit ~ factor(mo4) + offset(log(admit)), family = poiss
n,
    data = er)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.54926  -0.77429   0.00815   0.61915   1.81063

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.02849    0.03454 -58.721  < 2e-16 ***
factor(mo4)1  0.04620    0.04788   0.965 0.334511
factor(mo4)2 -0.05579    0.05032  -1.109 0.267557
factor(mo4)3 -0.12652    0.05041  -2.510 0.012075 *
factor(mo4)4 -0.13632    0.05170  -2.637 0.008365 **
factor(mo4)5 -0.11028    0.05254  -2.099 0.035827 *
factor(mo4)6 -0.13506    0.05224  -2.585 0.009733 **
factor(mo4)7 -0.17931    0.05231  -3.428 0.000609 ***
factor(mo4)8 -0.05677    0.05028  -1.129 0.258881
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 61.957  on 35  degrees of freedom
Residual deviance: 28.982  on 27  degrees of freedom
AIC: 300.42

Number of Fisher Scoring iterations: 3
```