Poisson Regression

The Poisson Model

We use Poisson Regression to model count outcomes. The model uses the log link function:

$$log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$$

where $i = 1, ..., n$

What is λ? The mean and the variance of the outcome Y

Left side similar in that it uses the log function, but it's the log of the mean rather than log odds

Right side similar to logistic and linear regression (no index j for the intercept, slopes or both like we saw with multinomial and ordinal)

Assumptions

- Independent observations
- Linear relationship between $log(\lambda_i)$ and the predictors
- **Equidispersion:** assumption that the mean and the variance are equal (λ)

Interpretations

- If x_1 is continuous, $e^{\hat{\beta}_1}$ is the multiplicative increase/decrease in the **expected count of the outcome** when increasing x_1 by one unit
- If x_1 is categorical, $e^{\hat{\beta}_1}$ is the multiplicative increase/decrease in the **expected count of the outcome** for $x_1 = 1$ compared to the reference group

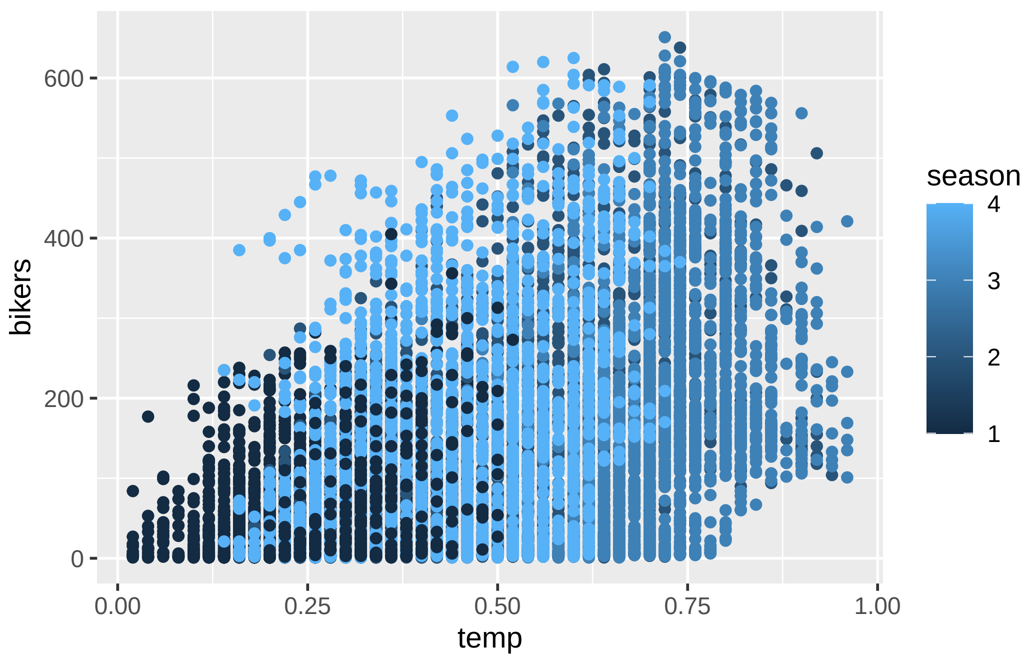
Example

Let's look at the Bikeshare data in the ISLR2 package:

```
library(MASS)
library(car)
library(tidyverse)
library(ISLR2)

data("Bikeshare")
# type ?Bikeshare in the console to see the data dictionary and description
```

Let's create a plot to visualize the relationship between `temp`, `season`, and `bikers`

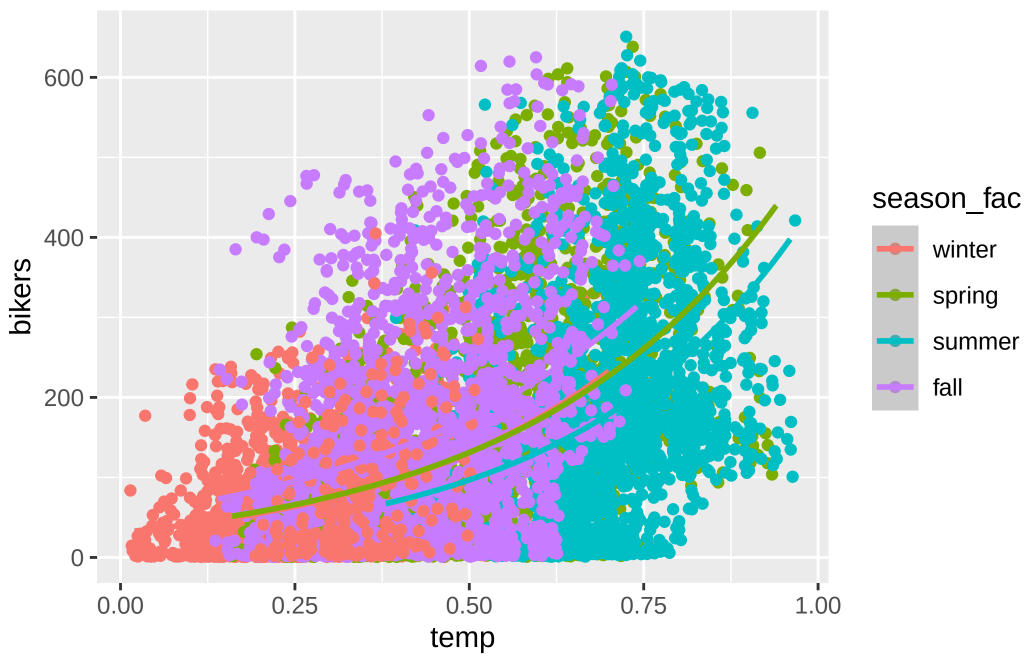
```
ggplot(Bikeshare, aes(x=temp, y=bikers, col=season))+
  geom_point()
```




```
#what's the issue here?

# need to factor season
Bikeshare <- Bikeshare |>
  mutate(season_fac = factor(season,
```

```r
                                levels=c(1,2,3,4),
                                labels=c("winter","spring","summer","fall")))

ggplot(Bikeshare, aes(x=temp, y=bikers, col=season_fac))+
  geom_jitter()+
  geom_smooth(method = "glm", method.args = list(family = "poisson"))
```

```
`geom_smooth()` using formula = 'y ~ x'
```



Fitting the model

```r
pois_mod <- glm(bikers ~ temp + season_fac,
                data=Bikeshare,
                family="poisson")
summary(pois_mod)
```

```
Call:
glm(formula = bikers ~ temp + season_fac, family = "poisson",
```

```
    data = Bikeshare)

Coefficients:
                  Estimate Std. Error  z value Pr(>|z|)    
(Intercept)       3.478226   0.003471 1002.064  < 2e-16 ***
temp              2.743482   0.007408  370.321  < 2e-16 ***
season_facspring  0.027581   0.003741    7.373 1.66e-13 ***
season_facsummer -0.202024   0.004285  -47.147  < 2e-16 ***
season_facfall    0.325015   0.003322   97.826  < 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: 1052921  on 8644  degrees of freedom
Residual deviance:  793484  on 8640  degrees of freedom
AIC: 846532

Number of Fisher Scoring iterations: 5
```

```r
(exp_coefs <- exp(cbind(est=coef(pois_mod),confint(pois_mod))))
```

```
Waiting for profiling to be done...
```

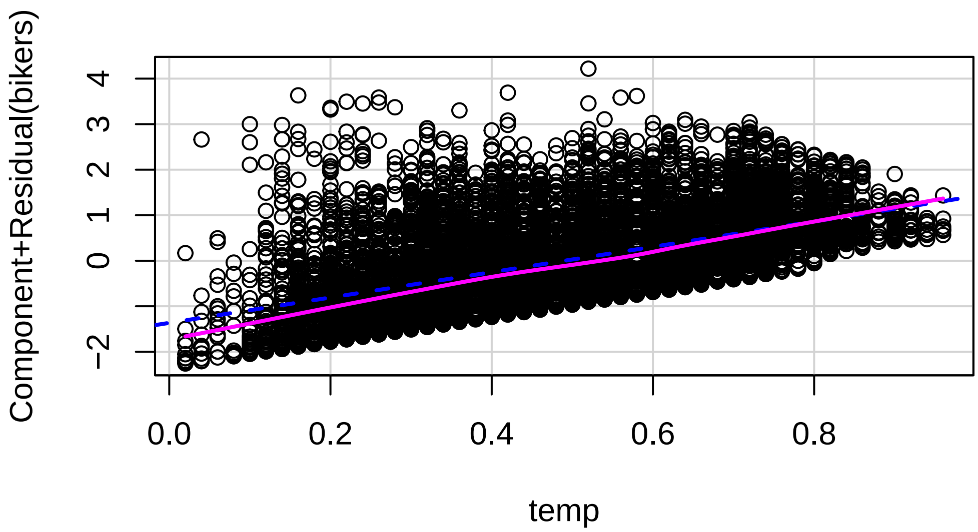
```
                       est      2.5 %     97.5 %
(Intercept)      32.4021861 32.1823949 32.6232722
temp             15.5410025 15.3169841 15.7683178
season_facspring  1.0279653  1.0204576  1.0355310
season_facsummer  0.8170749  0.8102428  0.8239672
season_facfall    1.3840512  1.3750707  1.3930960
```

- The expected number of bike rentals is 1.03 times higher in the spring compared to winter ($p < 0.01$, 95% CI: [1.02,1.04])
- For each unit increase in normalized temperature, the expected number of bike rentals increases 15.5 times ($p < 0.01$, 95% CI: [15.3, 15.8]) (may want to interpret in terms of tenth of a unit increase instead of the entire range zero to one)

Assessment

Partial residual plot:

```r
car::crPlot(pois_mod, "temp")
```



Looks pretty good! (Assessing linearity with partial residual plot)

Can compare to negative binomial model to assess equidispersion:

```r
nb_mod <- MASS::glm.nb(bikers ~ temp+factor(season), data=Bikeshare)

AIC(pois_mod, nb_mod)
```

```
         df      AIC
pois_mod  5 846532.0
nb_mod    6 101249.1
```

```r
summary(nb_mod)
```

```
Call:
MASS::glm.nb(formula = bikers ~ temp + factor(season), data = Bikeshare,
    init.theta = 1.052252403, link = log)
```

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)    
(Intercept)      3.43151    0.03205 107.058  < 2e-16 ***
temp             2.89122    0.08581  33.693  < 2e-16 ***
factor(season)2 -0.01410    0.03737  -0.377    0.706    
factor(season)3 -0.26865    0.04720  -5.692 1.25e-08 ***
factor(season)4  0.32110    0.03291   9.757  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(1.0523) family taken to be 1)

    Null deviance: 11983  on 8644  degrees of freedom
Residual deviance:  9876  on 8640  degrees of freedom
AIC: 101249

Number of Fisher Scoring iterations: 1


              Theta:  1.0523 
          Std. Err.:  0.0146 

 2 x log-likelihood:  -101237.1120 
```

```r
(exp_coefs_nb <- exp(cbind(est=coef(nb_mod),confint(nb_mod))))
```

```
Waiting for profiling to be done...

                       est      2.5 %     97.5 %
(Intercept)     30.9232887 29.0233835 32.9623531
temp            18.0152451 15.1698728 21.3950445
factor(season)2  0.9859986  0.9156194  1.0618637
factor(season)3  0.7644078  0.6957800  0.8397262
factor(season)4  1.3786397  1.2927553  1.4702142
```

Exercise

1. Compare the results we obtained above to a linear model:

- Fit a linear model regressing the number of bikers on normalized temperature and season. Show the summary table and write an interpretation for the temperature predictor.

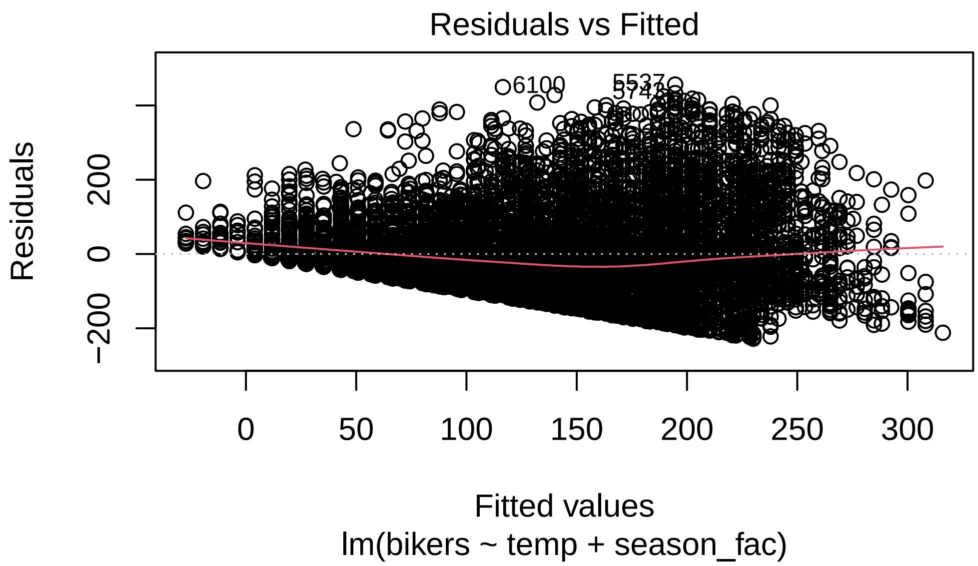
```r
linear_mod <- lm(bikers ~ temp + season_fac,
                 data=Bikeshare)

summary(linear_mod)
```

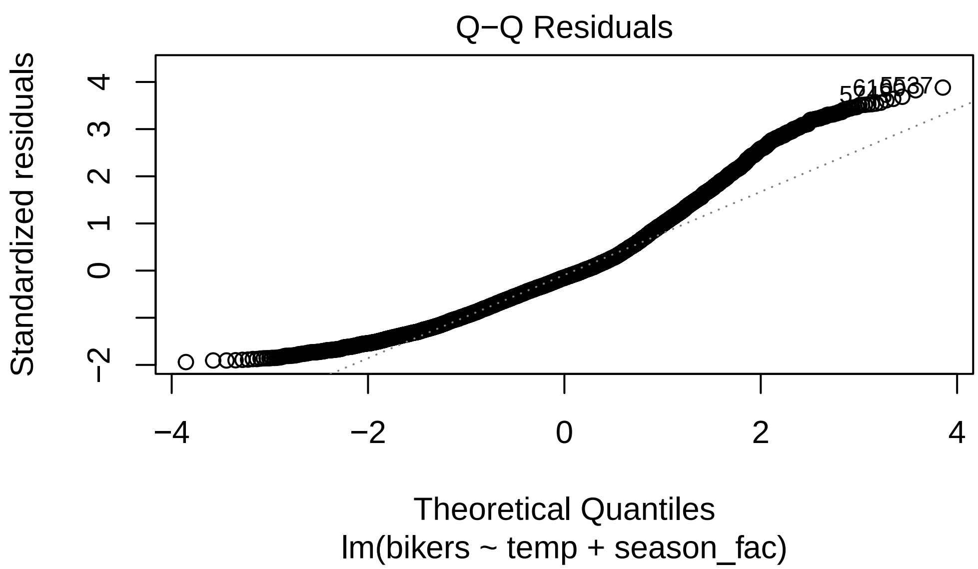
```

Call:
lm(formula = bikers ~ temp + season_fac, data = Bikeshare)

Residuals:
    Min      1Q  Median      3Q     Max 
-228.08  -80.83  -17.66   59.09  456.37 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)       -35.036      3.839  -9.128  < 2e-16 ***
temp              390.670     10.302  37.923  < 2e-16 ***
season_facspring  -16.163      4.483  -3.605 0.000314 ***
season_facsummer  -51.612      5.668  -9.107  < 2e-16 ***
season_facfall     21.302      3.948   5.395 7.03e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 117.6 on 8640 degrees of freedom
Multiple R-squared:  0.2277,	Adjusted R-squared:  0.2274 
F-statistic: 636.9 on 4 and 8640 DF,  p-value: < 2.2e-16
```

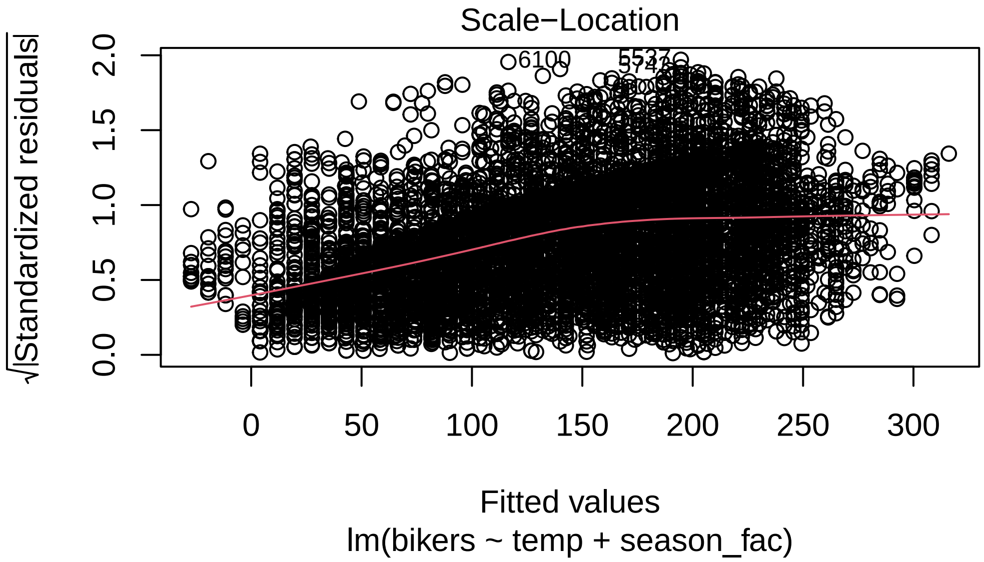
Per unit increase in normalized temperature, bike rentals increase by 390.6, on average, controlling for season.

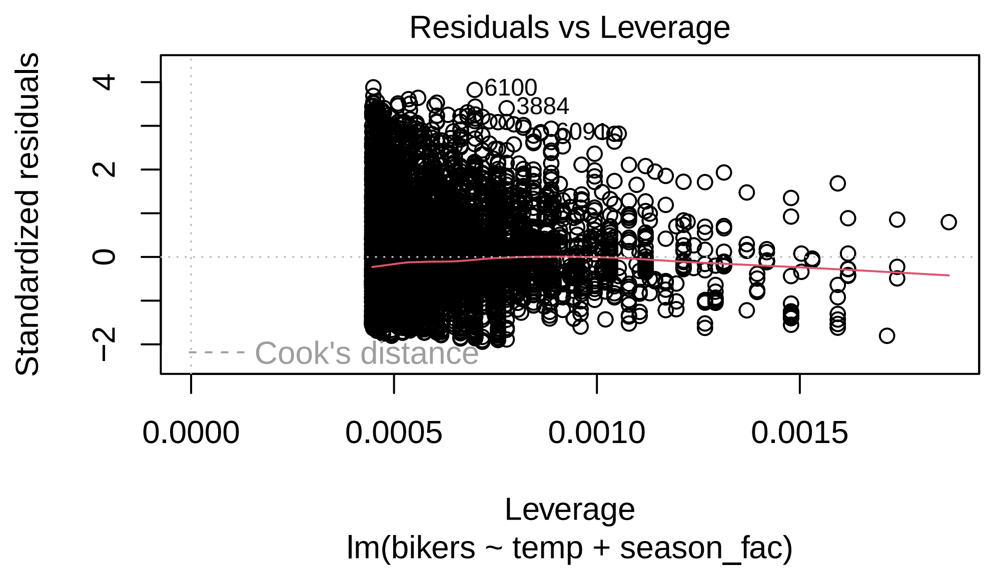
- Generate the diagnostic plots for the linear model. Comment on what you observe, particularly in the Residuals vs. Fitted plot. Which assumption(s) appears to be violated? How would a Poisson model address this/these assumption(s)?

```r
plot(linear_mod)
```


Residuals vs Fitted
Residuals
200
0
−200
0
50
100
150
200
250
300
Fitted values
lm(bikers ~ temp + season_fac)


Q−Q Residuals
Standardized residuals
4
3
2
1
0
−2
−4
−2
0
2
4
Theoretical Quantiles
lm(bikers ~ temp + season_fac)





Normality and homoscedasticity seem to be violated. In a poisson model, the variance will increase with the mean, which is inherently heteroscedastic.

2. Fit a poisson model regressing bikers on temperature and season and include an in-

teraction term for temperature and season. Use the summary output to calculate the coefficient estimates for temperature **for each season**. Then, exponentiate those calculated estimates. Write interpretations for these values.

```r
pois_mod2 <- glm(bikers ~ temp*season_fac,
                 data=Bikeshare,
                 family="poisson")

summary(pois_mod2)
```

```

Call:
glm(formula = bikers ~ temp * season_fac, family = "poisson",
    data = Bikeshare)

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)            3.436185   0.007118 482.724  < 2e-16 ***
temp                   2.877312   0.021049 136.693  < 2e-16 ***
season_facspring       0.066642   0.010029   6.645 3.04e-11 ***
season_facsummer      -0.396280   0.014193 -27.921  < 2e-16 ***
season_facfall         0.512060   0.009903  51.707  < 2e-16 ***
temp:season_facspring -0.128844   0.023972  -5.375 7.67e-08 ***
temp:season_facsummer  0.191463   0.026893   7.120 1.08e-12 ***
temp:season_facfall   -0.445632   0.025498 -17.477  < 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: 1052921  on 8644  degrees of freedom
Residual deviance:  792597  on 8637  degrees of freedom
AIC: 845651

Number of Fisher Scoring iterations: 5
```

```r
temp_coefs <- data.frame(temp_winter=coef(pois_mod2)[2], #winter
                temp_spring=(coef(pois_mod2)[2] + coef(pois_mod2)[6]),
                temp_summer=coef(pois_mod2)[2] + coef(pois_mod2)[7],
                temp_fall=coef(pois_mod2)[2] + coef(pois_mod2)[8])

exp(temp_coefs)
```

```
     temp_winter temp_spring temp_summer temp_fall
```

```
temp    17.76646    15.61868    21.51553  11.37798
```

Per unit increase in normalized temperature in the winter, expected bike rentals increase 17.8 times.

Per unit increase in normalized temperature in the spring, expected bike rentals increase 15.6 times.

Per unit increase in normalized temperature in the summer, expected bike rentals increase 21.5 times.

Per unit increase in normalized temperature in the fall, expected bike rentals increase 11.3 times.