Practical Poisson Regression

Ricard Garcia Isern & Adam Koershuis

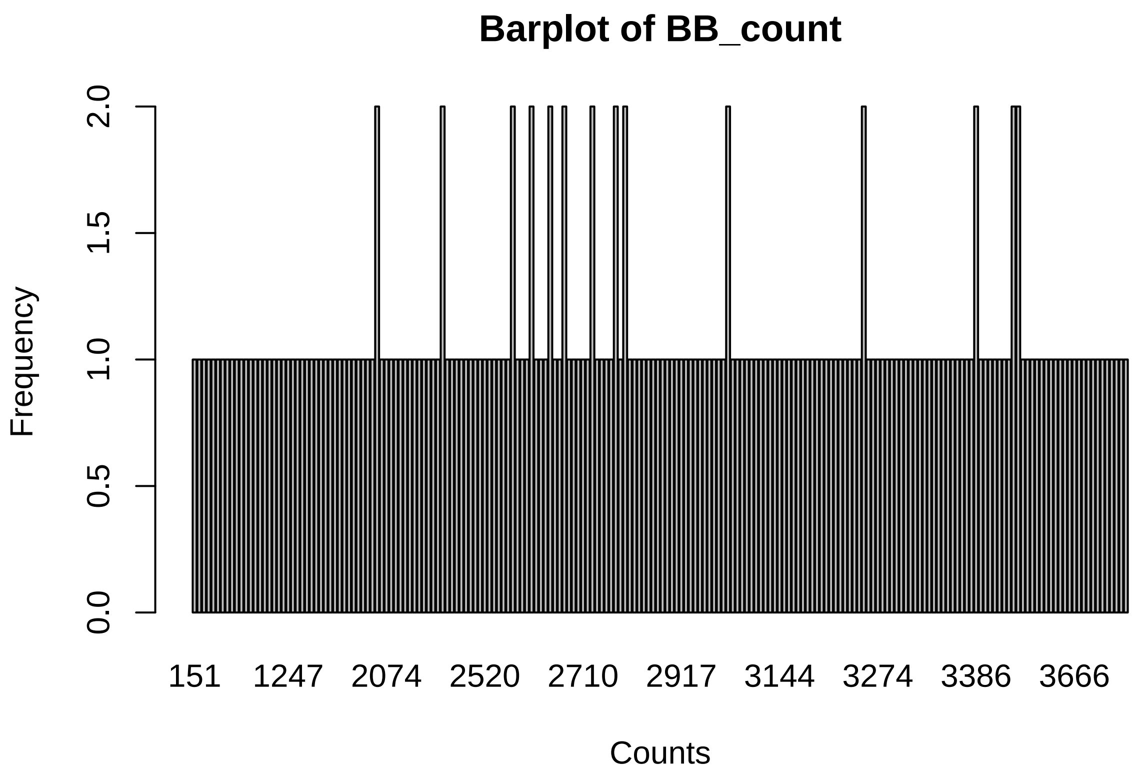
2023-11-08

```
library(readxl)
data<-read_excel("nyc_bb_bicyclist_counts_bis_clean.xls")
attach(data) #In order to making the variables directly accessible without
####specifying the data frame's name each time.
```

a) Read the dataset, you can use instruction attach the dataset for convenient access to the variables in the dataset.

```
barplot(table(BB_COUNT),
        main = "Barplot of BB_count",
        xlab = "Counts",
        ylab = "Frequency")
```

b) Make a barplot of the table of the possible outcomes of the response variable BB_COUNT. Calculate descriptive statistics of response BB_COUNT. Is there, at the exploratory level, evidence that the response does not follow a Poisson distribution?



We can also compute some descriptive statistics:

```
mean_response <- mean(BB_COUNT)
cat("The mean( ) is",mean_response)
```

```
## The mean( ) is 2680.042
```

```
variance_response <- var(BB_COUNT)
cat("The variance is",variance_response)
```

```
## The variance is 730530.7
```

```
std_dev_response <- sqrt(variance_response)
cat("The standard deviation is",std_dev_response)
```

```
## The standard deviation is 854.7109
```

We can manually observe that the plot does not follow a poisson distribution because the values that are more frequent should be and we grouped together in our plot, it is not the case. We can also state that the mean is not the same as the variance. Since the variance is higher, there is over dispersion.

```
poisson_model <- glm(BB_COUNT ~ HIGH_T, data = data, family = poisson(link="log"))
summary(poisson_model)
```

c) Perform Poisson regression of the number of BB_COUNT on HIGH_T, our first model. Report the regression equation. Is there evidence for association, and if so, what kind of association?

```
## 
## Call:
## glm(formula = BB_COUNT ~ HIGH_T, family = poisson(link = "log"), 
##     data = data)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -56.474   -5.965    0.673    7.652   34.407  
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept)  6.40688    0.04062 157.719  < 2e-16 ***
## HIGH_T48.00  0.70136    0.04968  14.117  < 2e-16 ***
## HIGH_T48.90 -0.27348    0.06180  -4.425 9.64e-06 ***
## HIGH_T51.10 -0.32695    0.05286  -6.185 6.21e-10 ***
## HIGH_T52.00  1.17637    0.04647  25.317  < 2e-16 ***
## HIGH_T53.10  0.85575    0.04849  17.647  < 2e-16 ***
## HIGH_T54.00  1.33045    0.04237  31.397  < 2e-16 ***
## HIGH_T55.00  0.85715    0.04848  17.680  < 2e-16 ***
## HIGH_T55.90  0.83377    0.04346  19.183  < 2e-16 ***
## HIGH_T57.00  1.50353    0.04282  35.112  < 2e-16 ***
## HIGH_T57.90  0.83568    0.04346  19.229  < 2e-16 ***
## HIGH_T59.00  0.80369    0.04355  18.455  < 2e-16 ***
## HIGH_T60.10  1.47449    0.04288  34.383  < 2e-16 ***
## HIGH_T61.00  1.21758    0.04258  28.596  < 2e-16 ***
## HIGH_T62.10  1.45780    0.04217  34.571  < 2e-16 ***
## HIGH_T63.00  1.48548    0.04176  35.575  < 2e-16 ***
## HIGH_T64.00  1.38071    0.04188  32.969  < 2e-16 ***
## HIGH_T64.90  1.30543    0.04117  31.709  < 2e-16 ***
## HIGH_T66.00  1.56386    0.04270  36.629  < 2e-16 ***
## HIGH_T66.90  1.45682    0.04121  35.352  < 2e-16 ***
## HIGH_T68.00  1.38594    0.04134  33.524  < 2e-16 ***
## HIGH_T69.10  0.92221    0.04259  21.652  < 2e-16 ***
## HIGH_T70.00  1.53477    0.04135  37.121  < 2e-16 ***
## HIGH_T71.10  1.43293    0.04131  34.688  < 2e-16 ***
## HIGH_T72.00  1.67577    0.04138  40.501  < 2e-16 ***
## HIGH_T73.00  1.26898    0.04156  30.531  < 2e-16 ***
## HIGH_T73.90  1.64955    0.04127  39.972  < 2e-16 ***
## HIGH_T75.00  1.63919    0.04127  39.715  < 2e-16 ***
## HIGH_T75.90  1.60392    0.04107  39.050  < 2e-16 ***
## HIGH_T77.00  1.53732    0.04134  37.184  < 2e-16 ***
## HIGH_T78.10  1.64437    0.04098  40.129  < 2e-16 ***
## HIGH_T79.00  1.49004    0.04212  35.376  < 2e-16 ***
## HIGH_T80.10  1.68864    0.04109  41.098  < 2e-16 ***
## HIGH_T81.00  1.57884    0.04104  38.472  < 2e-16 ***
## HIGH_T82.00  1.65647    0.04097  40.428  < 2e-16 ***
## HIGH_T82.90  1.57786    0.04104  38.447  < 2e-16 ***
## HIGH_T84.00  1.56265    0.04133  37.813  < 2e-16 ***
## HIGH_T84.90  1.66030    0.04101  40.489  < 2e-16 ***
## HIGH_T86.00  1.57622    0.04200  37.530  < 2e-16 ***
## HIGH_T87.10  1.51078    0.04136  36.525  < 2e-16 ***
## HIGH_T88.00  1.62472    0.04193  38.744  < 2e-16 ***
## HIGH_T89.10  1.44205    0.04517  31.925  < 2e-16 ***
## HIGH_T90.00  1.55649    0.04168  37.345  < 2e-16 ***
## HIGH_T91.00  1.57057    0.04166  37.696  < 2e-16 ***
## HIGH_T91.90  1.60679    0.04261  37.709  < 2e-16 ***
## HIGH_T93.00  1.54468    0.04475  34.520  < 2e-16 ***
## HIGH_T93.90  1.55936    0.04469  34.893  < 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: 70021  on 213  degrees of freedom
## Residual deviance: 38051  on 167  degrees of freedom
## AIC: 40209
##
## Number of Fisher Scoring iterations: 5
```

```
class(HIGH_T)
```

```
## [1] "character"
```

We can observe that the result, when applying the summary, is not familiar to us. This happens because HIGH_T is a character and not a numerical predictor. If we apply a transformation, we will get different results:

```
data$HIGH_T<-as.numeric(HIGH_T)
poisson_model <- glm(BB_COUNT ~ HIGH_T, data = data, family = poisson(link="log"))
summary(poisson_model)
```

```
##
## Call:
## glm(formula = BB_COUNT ~ HIGH_T, family = poisson(link = "log"),
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -57.269   -9.559    1.597   10.984   42.072
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 6.7117864  0.0101361   662.2   <2e-16 ***
## HIGH_T      0.0157516  0.0001325   118.9   <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: 70021  on 213  degrees of freedom
## Residual deviance: 55495  on 212  degrees of freedom
## AIC: 57563
##
## Number of Fisher Scoring iterations: 4
```

Now we can observe that the result is the correct one.

We can observe that the HIGH_T coefficient is significant when computing the poission regression, since the p-value is very low(smaller than 0.05).

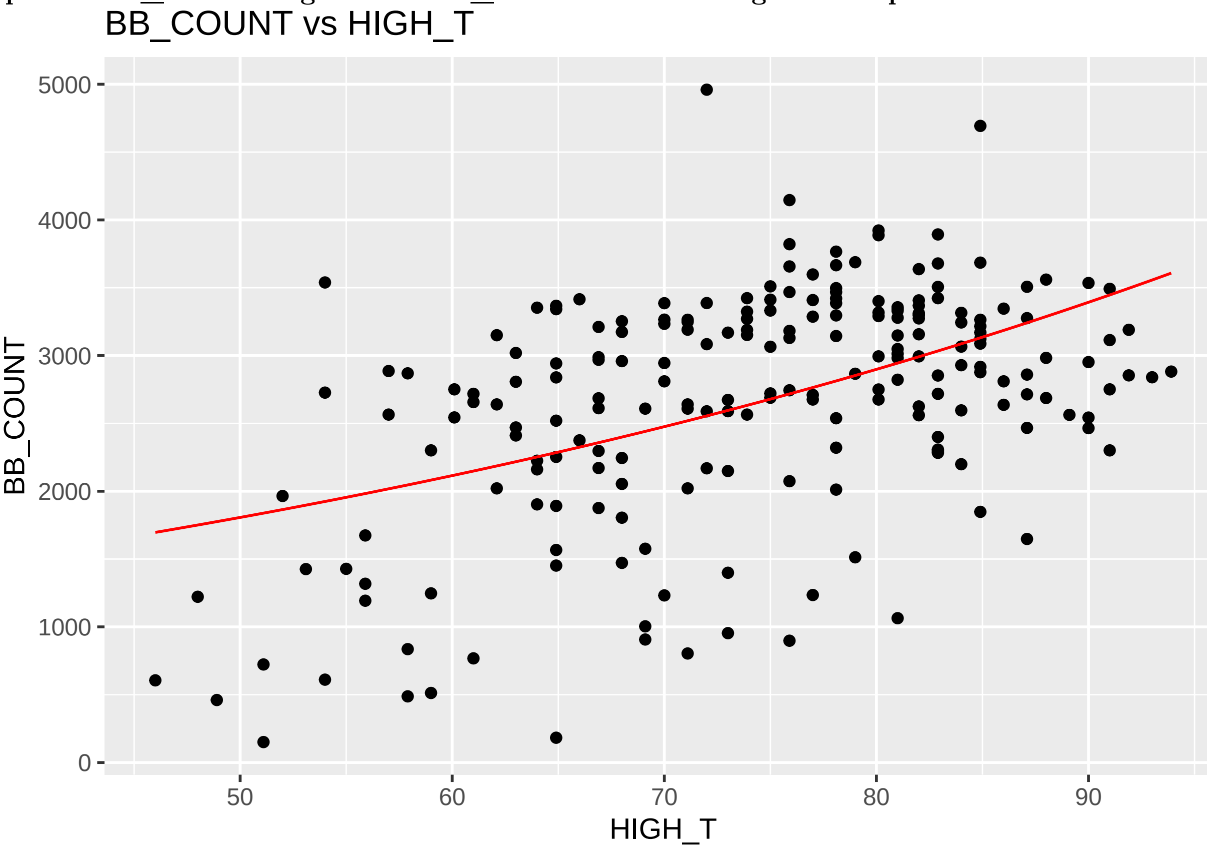
The formula is the following one: ln (ˆµ) = 6.71 + 0.01575*HIGH_T –> meaning that for every unit increase of HIGH_T the BB_COUNT(bike counts) increases 0.01575 units in the log-scale. If we are in the normal scale, it will increase 1.02 units (see exercise f)

```r
library(ggplot2)
#Create scatter plot
plot <- ggplot(data, aes(x = HIGH_T, y = BB_COUNT)) +
  geom_point() +
  labs(title = "BB_COUNT vs HIGH_T",
       x = "HIGH_T",
       y = "BB_COUNT")

#Calculate predicted values with the previous model
predicted_values <- data.frame(HIGH_T = seq(min(HIGH_T), max(HIGH_T), length.out=100))
predicted_values$BB_COUNT <- predict(poisson_model,newdata = predicted_values,type = "response")

#Add fitted regression curve
plot<-plot+geom_line(data = predicted_values,aes(x=HIGH_T, y=BB_COUNT), color="red")
plot
```

d) Make a scatter plot of BB_COUNT against HIGH_T. Add the fitted regression equation to the scatter plot.



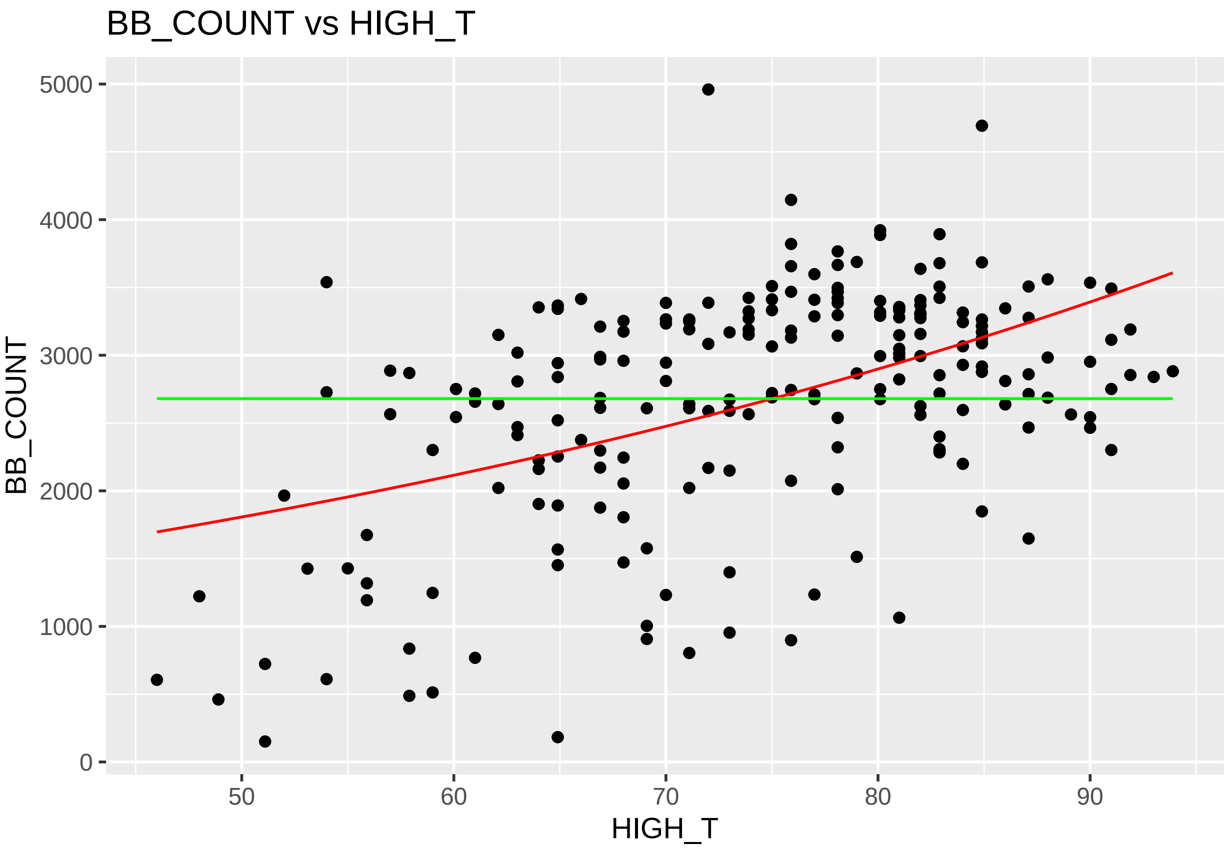
```
null_model <- glm(BB_COUNT~1,data=data,family = "poisson")
summary(null_model)
```

e) Estimate the null model without predictors (BB_COUNT 1), and also plot the equation according to this model to the plot. What do you observe?

```
##
## Call:
## glm(formula = BB_COUNT ~ 1, family = "poisson", data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -64.726   -7.566    3.382   11.283   39.326
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  7.89359    0.00132    5978   <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: 70021  on 213  degrees of freedom
## Residual deviance: 70021  on 213  degrees of freedom
## AIC: 72087
##
## Number of Fisher Scoring iterations: 4
```

We can observe that the INTERCEPT is just the log of the mean BB_COUNT. We can do null model in order to have a basis to compare it with the models including predictors. In this case, the prediction is just the mean for all levels of HIGH_T, not a good predictor.

```
#Calculate the predicted values with the null model
predicted_null_values <- data.frame(HIGH_T = seq(min(HIGH_T), max(HIGH_T),
length.out = 100))
predicted_null_values$BB_COUNT <- predict(null_model, newdata = predicted_null_values, type =
"response")
#Add  null model equation to the existing plot
plot <- plot +
geom_line(data = predicted_null_values, aes(x = HIGH_T, y = BB_COUNT), color = "green")
plot
```



We can observe that the green line,representing the null model predicted values, is constant, as it is the mean. We can say that the red line fits better the data than the green line. Therefore, we can state that there is an evidence that HIGH_T predictor can be significantly associated with the response variable BB_COUNT.

```
coefficient <- 0.0157516
std_error <- 0.0001325
critical_value <- qnorm(0.975)  # Critical value for a 95% confidence interval

# Calculate the confidence interval
lower_bound <- coefficient - critical_value * std_error
upper_bound <- coefficient + critical_value * std_error

cat("Confidence Interval is [", lower_bound,":",upper_bound,"]")
```

f) Interpret the first model by quantifying the effect of the predictor on the average of the response. Give a 95% confidence interval for the parameter representing that effect.

```
## Confidence Interval is [ 0.0154919 : 0.0160113 ]
```

This means that for each unit increase in HIGH_T the expected BB_COUNT will increase between 0.0154919 and 0.0160113 units.

g) Is the value 0 inside the interval you obtained? Is the value 1 inside the interval? What is the relevance of this? Value 0 is not in the interval [0.0154919 : 0.0160113], 1 is also not in this interval.

This means that there is statistical evidence to suggest that the corresponding predictor variable has a significant effect on the response variable. If it does not include 1, it suggests that the effect is not only significant but also multiplicative in nature. In your case, a confidence interval for the HIGH_T coefficient that excludes both 0 and 1 implies that a change in HIGH_T has a statistically significant and multiplicative impact on the expected count of BB_COUNT.

h) Is there any indication that over dispersion is a problem for you model? Justify your answer. We can compute the dispersion parameter as follows: Dispersion Parameter = (Degrees-of-Freedom-Residual)/(Deviance)

```
overdispersion_value <- (deviance(poisson_model))/(df.residual(poisson_model))
cat("Over dispersion ratio is", overdispersion_value)
```

```
## Over dispersion ratio is 261.77
```

Since the over dispersion value is greater than 1, we have over dispersion. Over dispersion occurs when the variance of the response variable is larger than what is expected in a Poisson distribution. In such cases, you may want to consider using a different modeling approach to account for the over dispersion in our data.

```


library(AER)
```

i) Formally test for over dispersion using the function dispersion test of the AER package. What is your conclusion?

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
## Loading required package: lmtest
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
## Loading required package: sandwich
```

```
## Loading required package: survival
```

```
dispersion <- dispersiontest(poisson_model)
dispersion
```

```
##
##  Overdispersion test
##
## data:  poisson_model
## z = 9.6905, p-value < 2.2e-16
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
## dispersion
##   228.6938
```

Ho –> dispersion parameter = 1 (no over dispersion), Ha -> dispersion parameter not equal 1 (over disperion).

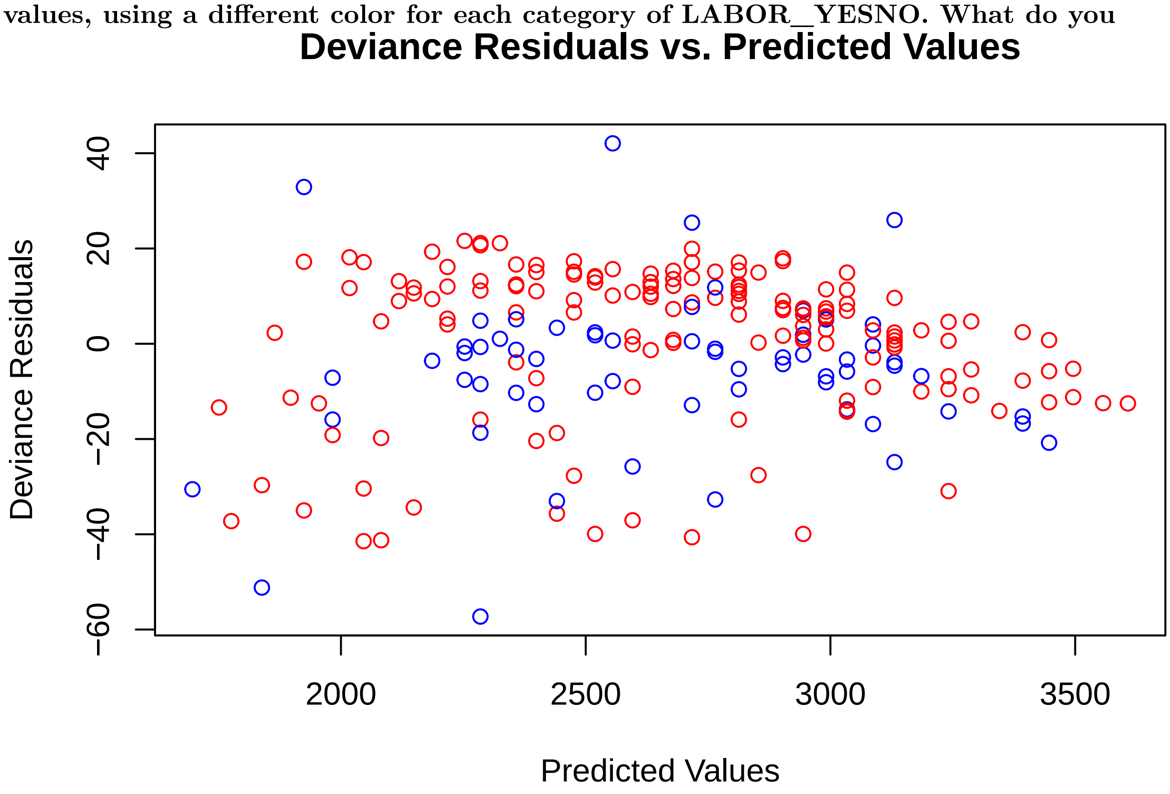
Since the p-value is very small, lower than 0.05, we can reject the null hypothesis and state that there is over dispersion, we should use a different method to fit the data.

```
deviance_residuals <- residuals(poisson_model, type = "deviance")

colors <- ifelse(data$LABOR_YESNO == 0, "blue", "red")

plot(predict(poisson_model, type = "response"), deviance_residuals, col = colors,
     xlab = "Predicted Values", ylab = "Deviance Residuals",
     main = "Deviance Residuals vs. Predicted Values")
```

j) Calculate deviance residuals according to the first model and plot these as a function of the predicted values, using a different color for each category of LABOR_YESNO. What do you observe?



```
poisson_model <- glm(BB_COUNT ~ HIGH_T + LABOR_YESNO, data = data, family = poisson(link="log"))
summary(poisson_model)
```

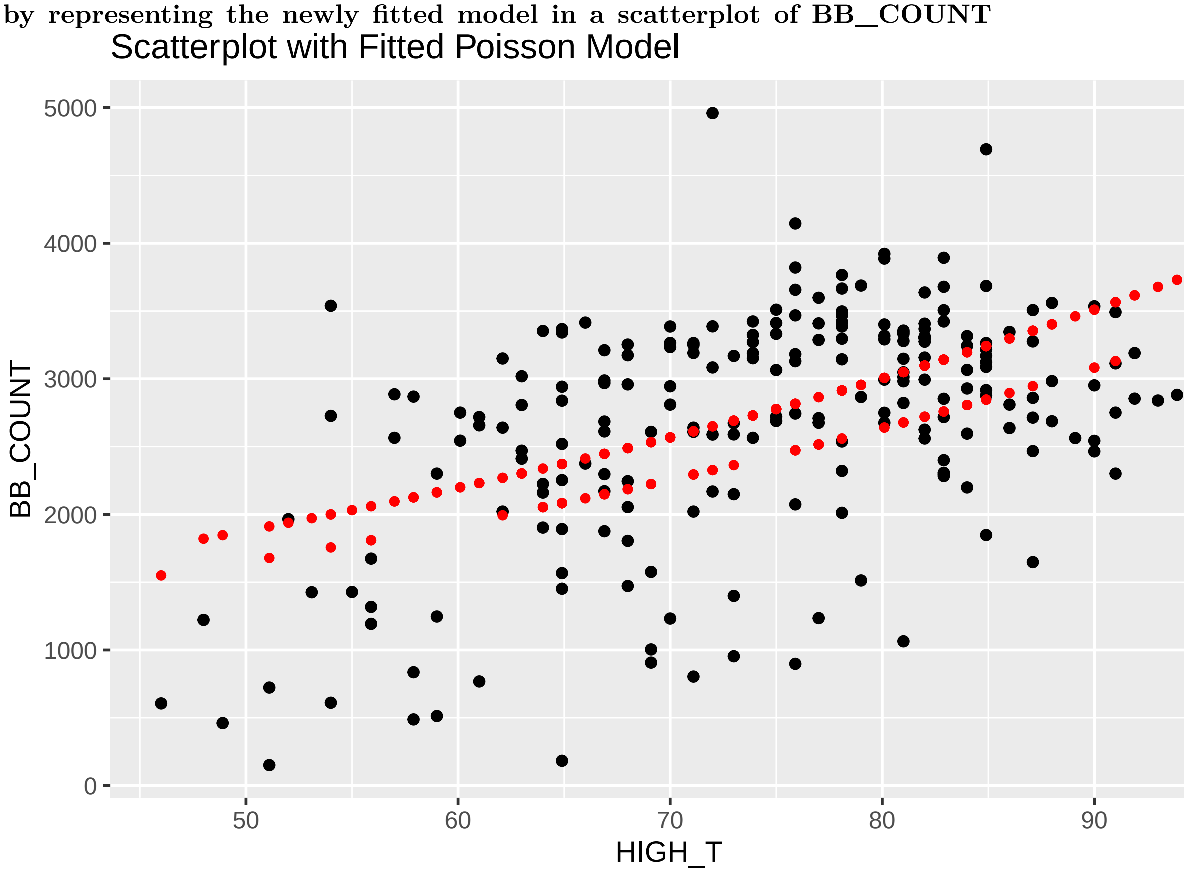
k) Do a Poisson regression of BB_COUNT on HIGH_T and LABOR_YESNO. Report the fitted equation. Is there evidence for any effect of the variable LABOR_YESNO? Justify your answer.

```
##
## Call:
## glm(formula = BB_COUNT ~ HIGH_T + LABOR_YESNO, family = poisson(link = "log"),
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -53.936   -8.833    2.650    9.405   47.349
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 6.6273727  0.0103142  642.55   <2e-16 ***
## HIGH_T      0.0156238  0.0001323  118.07   <2e-16 ***
## LABOR_YESNO 0.1298549  0.0030017   43.26   <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: 70021  on 213  degrees of freedom
## Residual deviance: 53586  on 211  degrees of freedom
## AIC: 55656
##
## Number of Fisher Scoring iterations: 4
```

After making the poisson regression we can say that there's a strong evidence for the effect of the variable LABOR_YESNO as it's p-value is very low (<2e-16). It's significant.

```
library(ggplot2)
plot_data <- data.frame(HIGH_T = HIGH_T, BB_COUNT = BB_COUNT, Fitted_Values = predict(poisson_model, ty
ggplot(plot_data, aes(x = data$HIGH_T, y = BB_COUNT)) +
  geom_point(color = "black") +
  geom_point(aes(y = Fitted_Values), color = "red", shape = 16) +
  labs(x = "HIGH_T", y = "BB_COUNT", title = "Scatterplot with Fitted Poisson Model")
```

l) Make a graphic by representing the newly fitted model in a scatterplot of BB_COUNT against HIGH_T.



The fitted values are the red ones.

```
poisson_model_interaction <- glm(BB_COUNT ~ HIGH_T * LABOR_YESNO, data = data, family = poisson(link =
summary(poisson_model_interaction)
```

m) Is there evidence for interaction between the variables LABOR_YESNO and HIGH_T? Justify your answer. Try to make a graphical representation of the fitted model with interaction in a scatterplot of BB_COUNT against HIGH_T.

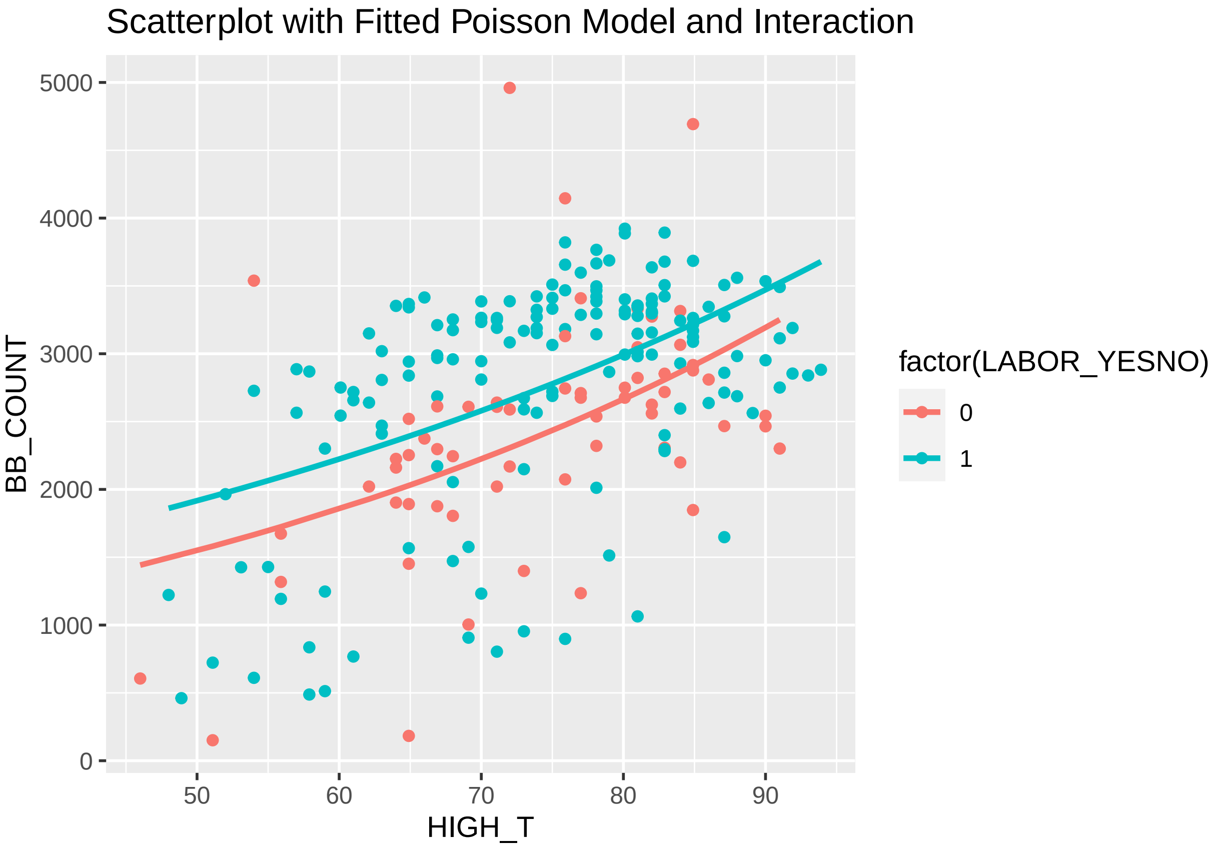
```
##
## Call:
## glm(formula = BB_COUNT ~ HIGH_T * LABOR_YESNO, family = poisson(link = "log"),
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -53.020   -8.869    2.652    9.425   47.836
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)        6.4423435  0.0206797  311.53   <2e-16 ***
## HIGH_T             0.0180719  0.0002711   66.67   <2e-16 ***
## LABOR_YESNO        0.3736235  0.0237164   15.75   <2e-16 ***
## HIGH_T:LABOR_YESNO -0.0032218  0.0003106  -10.37   <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: 70021  on 213  degrees of freedom
## Residual deviance: 53478  on 210  degrees of freedom
## AIC: 55550
##
## Number of Fisher Scoring iterations: 4
```

Yes, there's a strong evidence for interaction between the variables LABOR_YESSNO and HIGH_T. We can affirm that thanks to the poisson model results obtained, the p-value of the interaction is much lower than 0.05. This means that the interaction is strongly significant.

```
plot_data_interaction <- data.frame(HIGH_T = data$HIGH_T, BB_COUNT = data$BB_COUNT, LABOR_YESNO = data$L

ggplot(plot_data_interaction, aes(x = HIGH_T, y = BB_COUNT, color = factor(LABOR_YESNO))) +
  geom_point() +
  geom_line(aes(y = Fitted_Values), size = 1) +
  labs(x = "HIGH_T", y = "BB_COUNT", title = "Scatterplot with Fitted Poisson Model and Interaction")
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
```



```
data$PRECIP <- as.numeric(PRECIP)
poisson_model <- glm(BB_COUNT ~ HIGH_T + LABOR_YESNO + PRECIP, data = data, family = poisson(link="log"
summary(poisson_model)
```

n) Add the variable PRECIP to the model. Is it a significant predictor? Justify your answer

```
##
## Call:
## glm(formula = BB_COUNT ~ HIGH_T + LABOR_YESNO + PRECIP, family = poisson(link = "log"),
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -35.485   -6.402    0.914    6.667   41.863
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  6.9585590  0.0105923  656.95   <2e-16 ***
## HIGH_T       0.0123747  0.0001353   91.46   <2e-16 ***
## LABOR_YESNO  0.1105169  0.0030027   36.81   <2e-16 ***
## PRECIP      -0.8393595  0.0067742 -123.91   <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: 70021  on 213  degrees of freedom
## Residual deviance: 30380  on 210  degrees of freedom
## AIC: 32452
##
## Number of Fisher Scoring iterations: 5
```

Since the p-value of the variable PRECIP is lower than 0.05 (<2e-16) we can affirm that is a significant predictor, just as the other variables HIGH_T and LABOR_YESNO.

```
data$LOW_T <- as.numeric(LOW_T)
poisson_model <- glm(BB_COUNT ~ HIGH_T + LABOR_YESNO + PRECIP + LOW_T, data = data, family = poisson(li
summary(poisson_model)
```

o) Add the variable LOW_T to the model. Is it a significant predictor? Justify your answer

```
##
## Call:
## glm(formula = BB_COUNT ~ HIGH_T + LABOR_YESNO + PRECIP + LOW_T,
##     family = poisson(link = "log"), data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -34.186   -6.887   -0.026    6.427   40.232
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  6.9543297  0.0105798  657.32   <2e-16 ***
## HIGH_T       0.0239511  0.0002970   80.64   <2e-16 ***
## LABOR_YESNO  0.1211231  0.0030127   40.20   <2e-16 ***
## PRECIP      -0.7734866  0.0068100 -113.58   <2e-16 ***
## LOW_T       -0.0140505  0.0003202  -43.88   <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: 70021  on 213  degrees of freedom
## Residual deviance: 28467  on 209  degrees of freedom
## AIC: 30541
##
## Number of Fisher Scoring iterations: 4
```

Again, we can conclude that LOW_T is a significant predictor as it have a p-value of <2e-16 (lower than 0.05) in our poisson regression. As all the other variables.

p) What would be your final model for the data? Justify your answer. My final model for the data would be the last one from section o). That's because is the only one that includes all the variables, that are also all significant since we tested them with the poisson regression. This means that this is the only model that include all the variables that have an impact in BB_COUNT.

q) Give examples of outcomes that can be modelled using a Poisson regression, such as the number of goals in a handball match. Some examples of data that can be modelled using a Poisson regression:** - Number of tackles for a team in a season.

- Number of drinks consumed in a dinner.

- Number of hours slept in the last month.