Assignment 4
Poisson Regression

Jan Carreras & Jan Izquierdo

2023-11-6

We consider a data set reporting the daily total of bike counts conducted on the Brooklyn Bridge from 01 April 2017 to 31 October 2017 (Source: NYC Open Data: Bicycle Counts for East River Bridges). The variables included in the dataset are HIGH_T; LOW_T; PRECIP and LABOR_YESNO. The number of bikes (BB_COUNT) is used as response variable in a Poisson regression. Predictors (4 variables) we consider as potentially bearing a relationship with the number of bikes are the high and low temperature (HIGH_T; LOW_T;) the precipitation (PRECIP) and the if day of the week is a working day or not (LABOR_YESNO with values 1 and 0 to indicate yes or no, respectively). .

Setting libraries

```
#install.packages("AER")
library(readxl)
library(ggplot2)
library(AER)
```

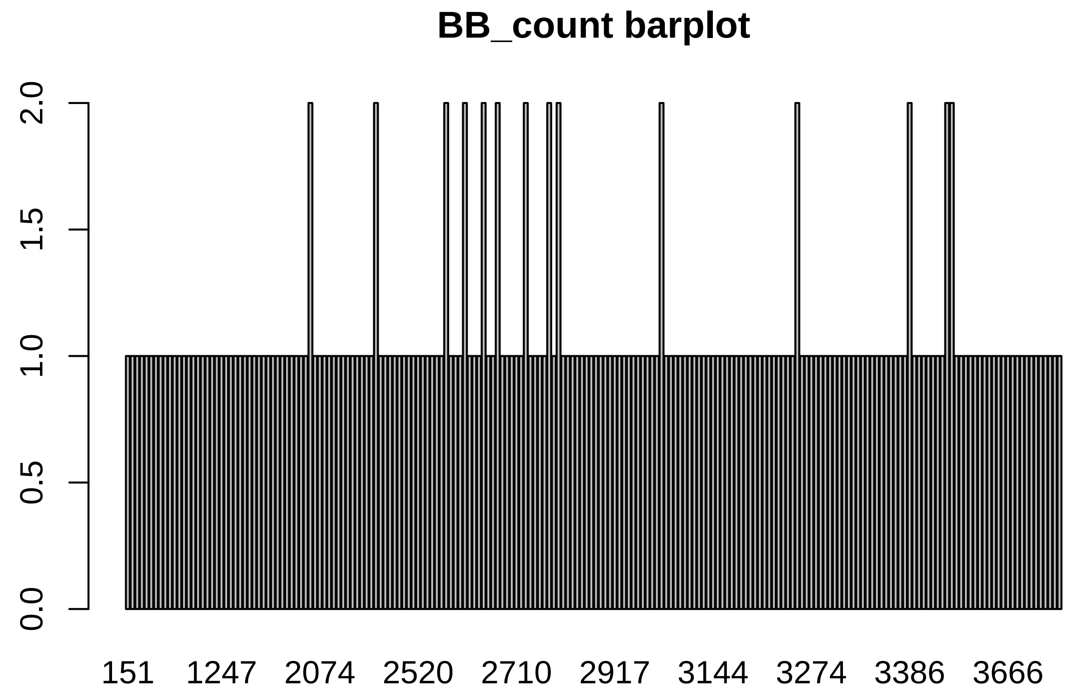
1.a) **Read the dataset, you can use instruction attach the dataset for convenient access to the variables in the dataset.** .

```
data<-read_excel("./assignment_data/bicyclist_data.xls")
n<-nrow(data)
n
```

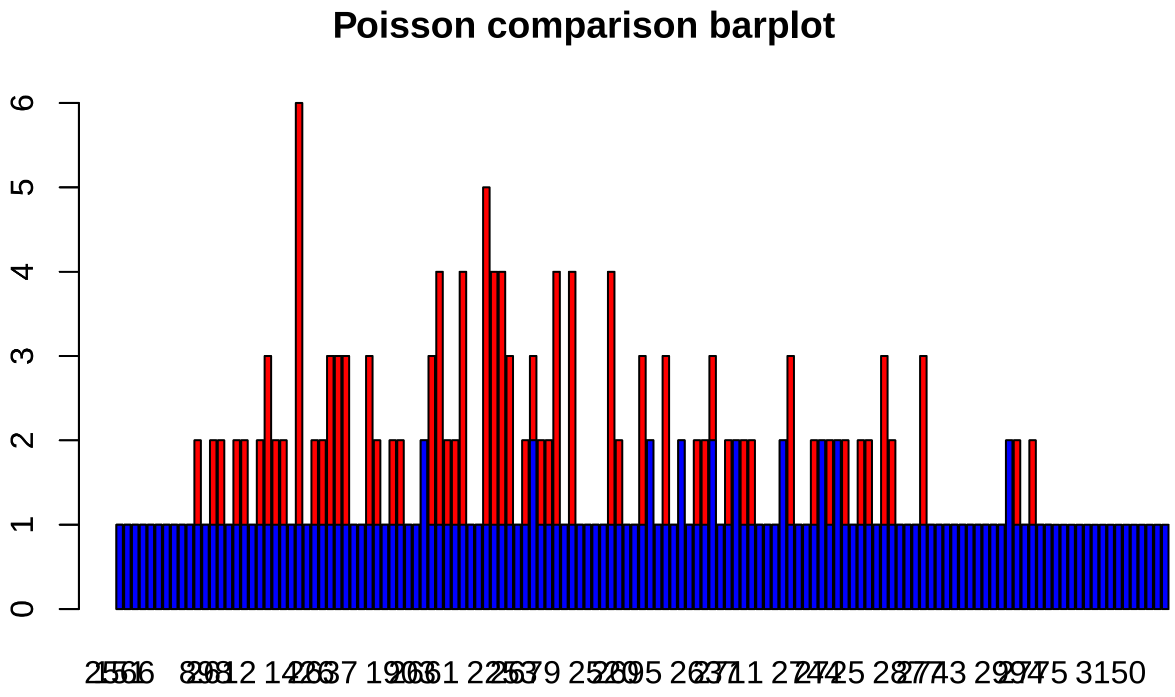
```
## [1] 214
```

1.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?** .

```
bb_table <- table(data$BB_COUNT)
barplot(bb_table, main="BB_count barplot")
```



```
bb_sum<-summary(data$BB_COUNT)
pois_dis<-rpois(214, mean(data$BB_COUNT))
pois_table<-table(pois_dis)
barplot(pois_table, col="red", main="Poisson comparison barplot");barplot(bb_table,col="blue",add=T)
```



```
## The descriptive statistics of BB_count is

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     151    2298    2857    2680    3285    4960

## In the BB_count barplot we can see that, the majority of BB_COUNT values have a frequency of 1, with
## only a few occurrences having a frequency of 2.

## In the barplot above we created data in a poisson distribution arrangement(red) and compare it with
## the distribution of our data(blue), we can see that the distributions are very different so we can
## concludethat BB_count does not follow a Poisson regression

## A feature of the Poisson distribution is that the mean equal to the variance. However, in our
## scenario, the variance significantly exceeds the mean, indicating that our data does not follow
## the distribution. To assess it, we compute the overdispersion parameter for our model.
```

```
model <- glm(BB_COUNT ~ HIGH_T + LOW_T + PRECIP + LABOR_YESNO,data=data, family = poisson(link = "log"))
# Get the residual deviance and the degrees of freedom
residual_dev <- model$deviance
df <- model$df.residual

# Calculate the overdispersion parameter
overdispersion <- residual_dev/df
```

```
## The overdispersion parameter is: 54.22276
```

1.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?** .

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

```
##
## Call:
## glm(formula = BB_COUNT ~ HIGH_T, family = "poisson", 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
```

```
coefficients <- coef(mdl_pois_reg)

intercept <- coefficients[1]
pendent <- coefficients[2]
```

```
## Regression equation:


## ln(BB_COUNT) = 6.711786 + 0.01575161 * HIGH_T


## BB_COUNT = e^( 6.711786 ) * e^( 0.01575161 * HIGH_T)


## There is proof of a connection, the p_value(<2e-16) is lower than critical value 0.05, so they are
## significant between each other. Also, according to this equation, a 1-unit elevation in HIGH_T
## leads to a BB_COUNT increase of 0.01575161 . This demonstrates a positive linear correlation between
## the variables, denoting a positive association. The positive sign of the slope signifies that as
## HIGH_T increases, BB_COUNT also experiences an increase.
```

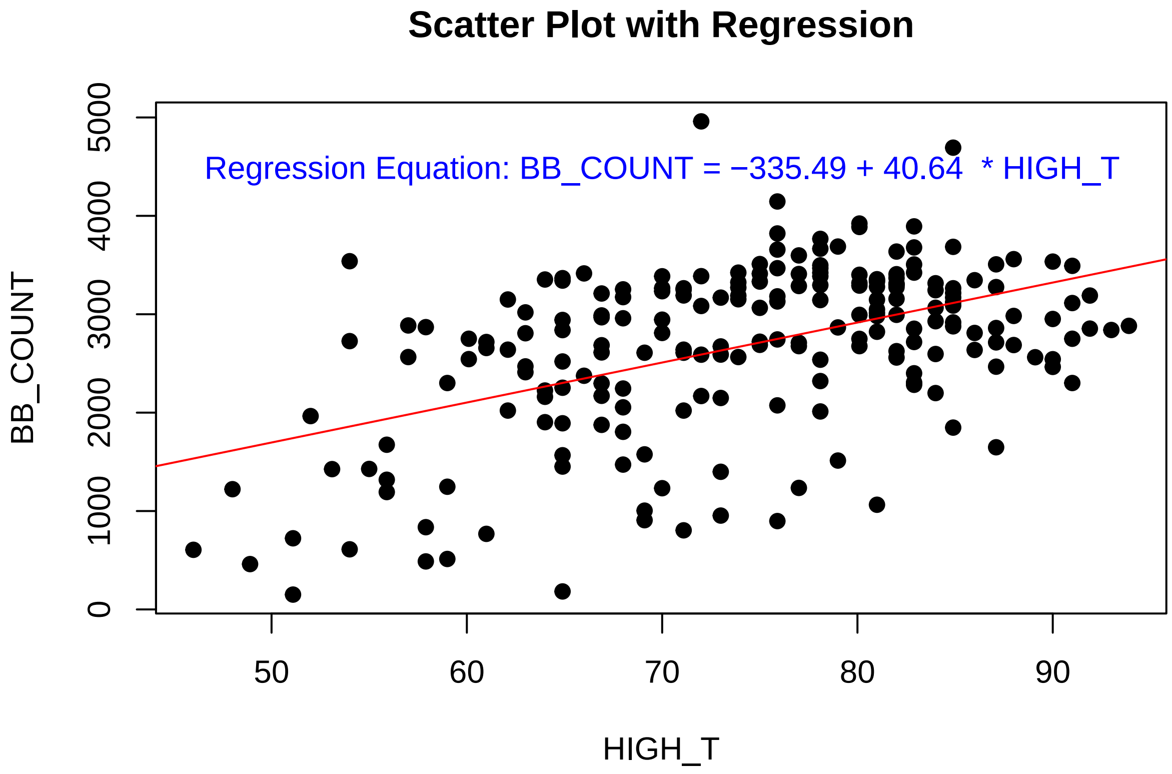
1.d) **Make a scatter plot of BB_COUNT against HIGH_T. Add the fitted regression equation to the scatter plot.** .

```
#Fit the linear regression model with standardized variables
model <- glm(BB_COUNT ~ HIGH_T, data = data)
#Create the scatter plot with standardized variables
plot(data$HIGH_T, data$BB_COUNT,
     main = "Scatter Plot with Regression",
     xlab = "HIGH_T", ylab = "BB_COUNT", pch=19)

#Add the regression line to the plot
abline(model, col = "red")

#Get the coefficients of the regression equation
coefficients <- coef(model)
intercept <- coefficients[1]
slope <- coefficients[2]

#Add the regression equation to the plot
equation <- paste("Regression Equation: BB_COUNT =",round(intercept, 2), "+", round(slope, 2), " * HIGH_
text(x=70, y=4460,
     equation,
     col="blue")
```



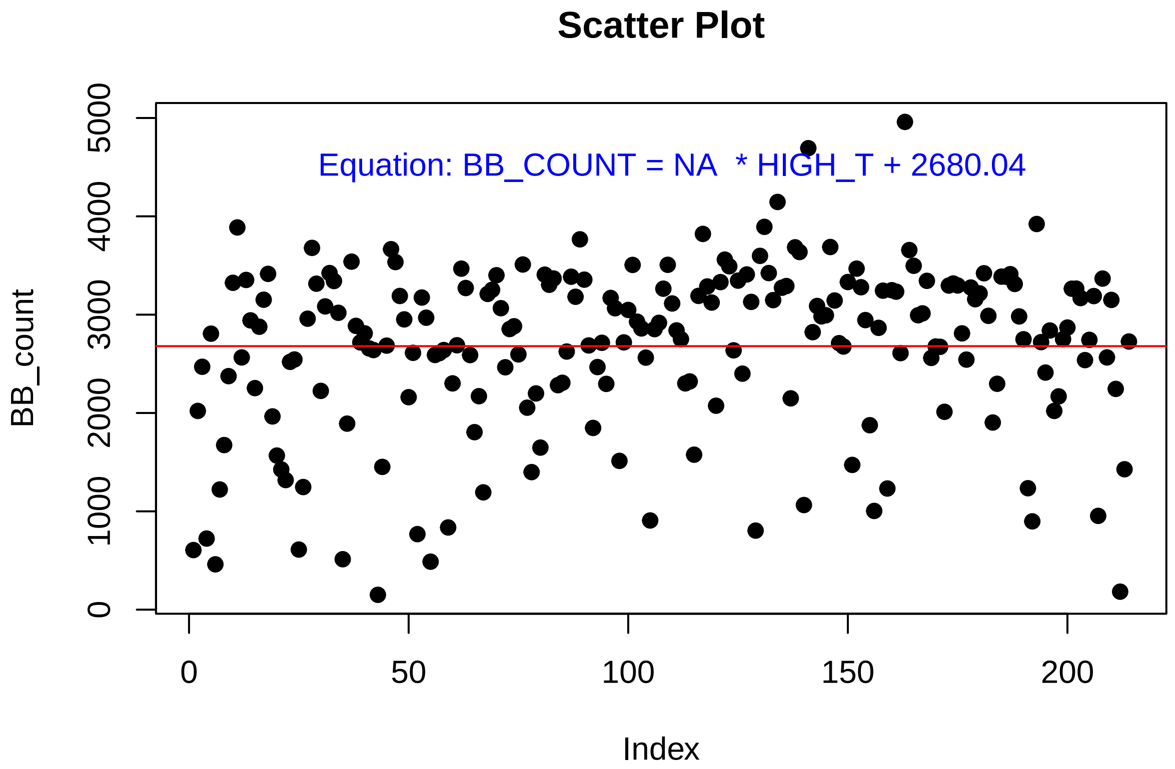
```
## We can see that there exists a positive linear correlation between these variables.
```

1.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?** .

```
null_model<- glm(BB_COUNT ~ 1, data = data)
null_plot<-plot(data$BB_COUNT, ylab= "BB_count",
     main = "Scatter Plot", pch=19)
abline(null_model, col = "red")

coefficients <- coef(null_model)
intercept <- coefficients[1]
slope <- coefficients[2]

null_equation <- paste("Equation: BB_COUNT =", round(slope, 2), " * HIGH_T +", round(intercept, 2))
text(x=110, y=4500,
     null_equation,
     col="blue")
```



```
## This model is characterized by an equation that remains unaffected by any variable, essentially
## representing the null model. Consequently, it produces a horizontal line, which stays consistent
## along the y-axis at the mean value. Hence, the null model is depicted as a steady line that aligns
## with the mean.
```

1.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.** .

```
# Calculate 95% confidence interval for the coefficient of 'High_T'
summary(mdl_pois_reg)
```

```
## 
## Call:
## glm(formula = BB_COUNT ~ HIGH_T, family = "poisson", 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
```

```
pendent <- summary(mdl_pois_reg)$coefficients[2, 1]
standard_error <- summary(mdl_pois_reg)$coefficients[2, 2]
z <- qnorm(0.975)

upper <- pendent + z*standard_error
lower <- pendent - z*standard_error
```

```
## The 95% confidence interval is: [ 0.01549186 , 0.01601136 ]
```

```
## Within this interval, it can be inferred that a rise of 1 unit in HIGH_T is associated with a
## BB_COUNT change falling between 0.01549186 and 0.01601136
```

***1.g)* Is the value 0 inside the interval you obtained? Is the value 1 inside the interval? What is the relevance of this?** .

```
## Neither 1 nor 0 are within this interval [ 0.01549186 , 0.01601136 ]. We have a relationship
## between the variables under study. A 0 within our interval would indicate otherwise, as it
## would signify the multiplication of our variable by 0 when HIGH_T increases by one unit.
## A 1 within our interval would mean there would be no change, so it makes sense that neither 0
## nor 1 are in our interval.
```

***1.h)* Is there any indication that overdispersion is a problem for you model? Justify your answer.** .

```
#Overdispersion in mdl_pois_reg
summary(mdl_pois_reg)
```

```
##
## Call:
## glm(formula = BB_COUNT ~ HIGH_T, family = "poisson", 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
```

```
# Residual deviance and the degrees of freedom
res_dev_1 <- mdl_pois_reg$deviance
df_1 <- mdl_pois_reg$df.residual

# Calculate the overdispersion parameter
overdispersion_1 <- res_dev_1/df_1
```

```
## The overdispersion parameter is: 261.77
```

```
## We can observe that this overdispersion is nearly twice the size of the previous one. It is too
## large for us to model our data using Poisson distribution, we can say that overdispersion poses
## problem for our model.
```

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

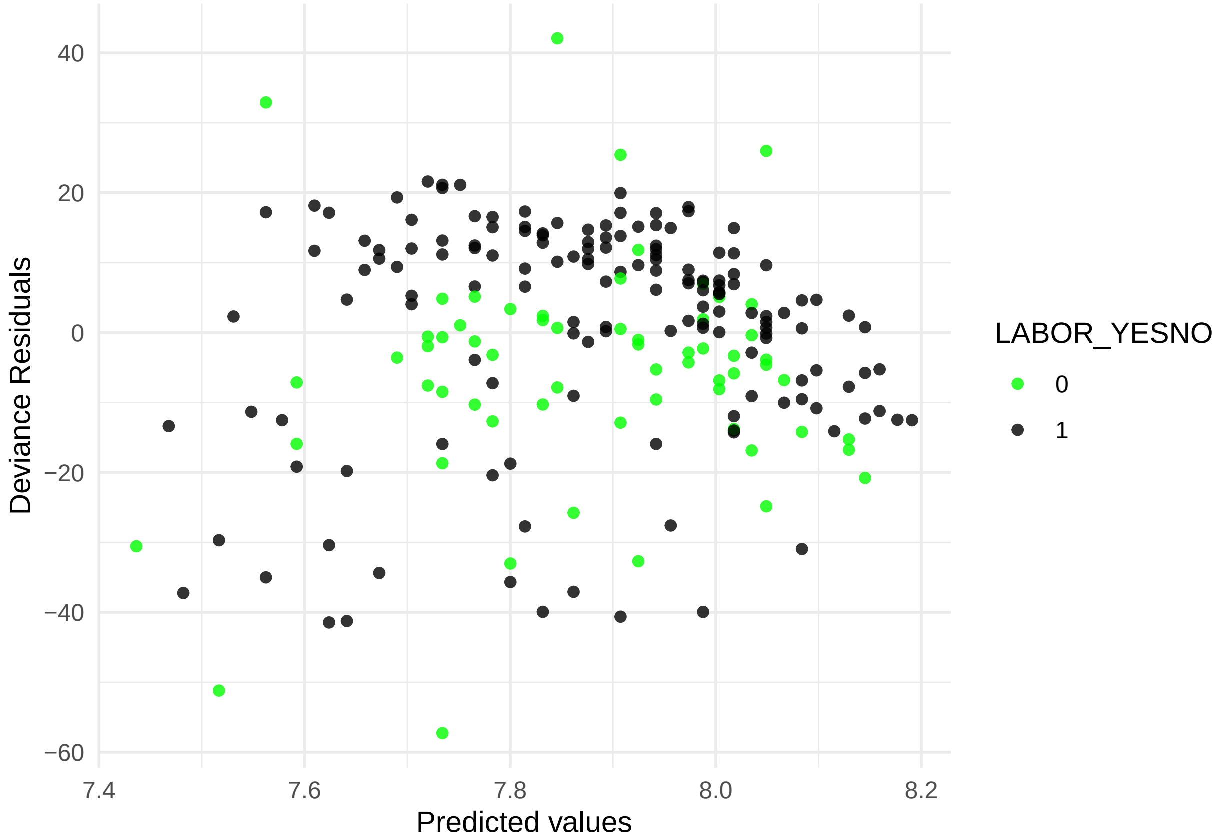
```
dispersion_test<-dispersiontest(mdl_pois_reg)
dispersion_test
```

```
##
##  Overdispersion test
##
## data:  mdl_pois_reg
## z = 9.6905, p-value < 2.2e-16
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
## dispersion
##   228.6938
## The results from our test strongly support our previous calculation: there is overdispersion. This
## indicates that the variability in the data is higher than what would be expected from a Poisson
## distribution.
```

1.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?** .

```
#Calculate residues of first model
res<-resid(mdl_pois_reg)
```

```
ggplot(data, aes(x=mdl_pois_reg[["linear.predictors"]], y= res,
  color =as.factor(LABOR_YESNO))) +
  xlab("Predicted values") + ylab("Deviance Residuals") +
  labs(color = "LABOR_YESNO") +
  geom_point(alpha = .8, shape = 19) +
  scale_color_manual(values = c("green", "black"))+theme_minimal()
```



```
## This plot is highly scattered and lacks linearity. Which indicates that the data still doesn't fit
## the model.
```

1.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.** .

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

```
##
## 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
```

```
coef(mdl_pois_reg_laboral )
```

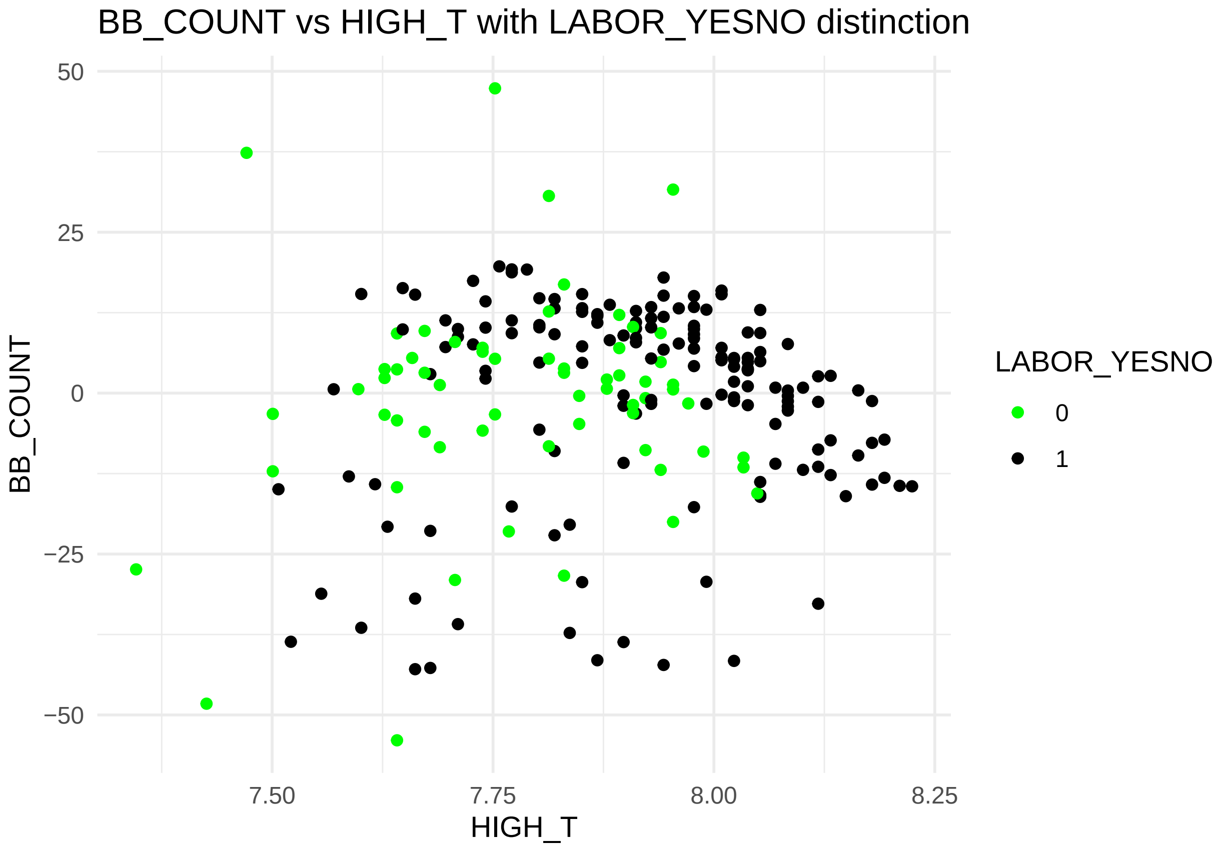
```
## (Intercept)      HIGH_T LABOR_YESNO
##  6.62737271  0.01562384  0.12985492
```

```
intercept <- coef(mdl_pois_reg_laboral )[1]
slope <- coef(mdl_pois_reg_laboral )[2]
labor<-coef(mdl_pois_reg_laboral )[3]
#Create the fitted
fitted_equation <- paste("Regression Equation: BB_COUNT =",round(intercept, 2),"+",round(slope, 2),
" * HIGH_T ","+", round(labor, 2), "* LABOR_YESNO ")
```

```
## [1] "Regression Equation: BB_COUNT = 6.63 + 0.02  * HIGH_T  + 0.13 * LABOR_YESNO "
```

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

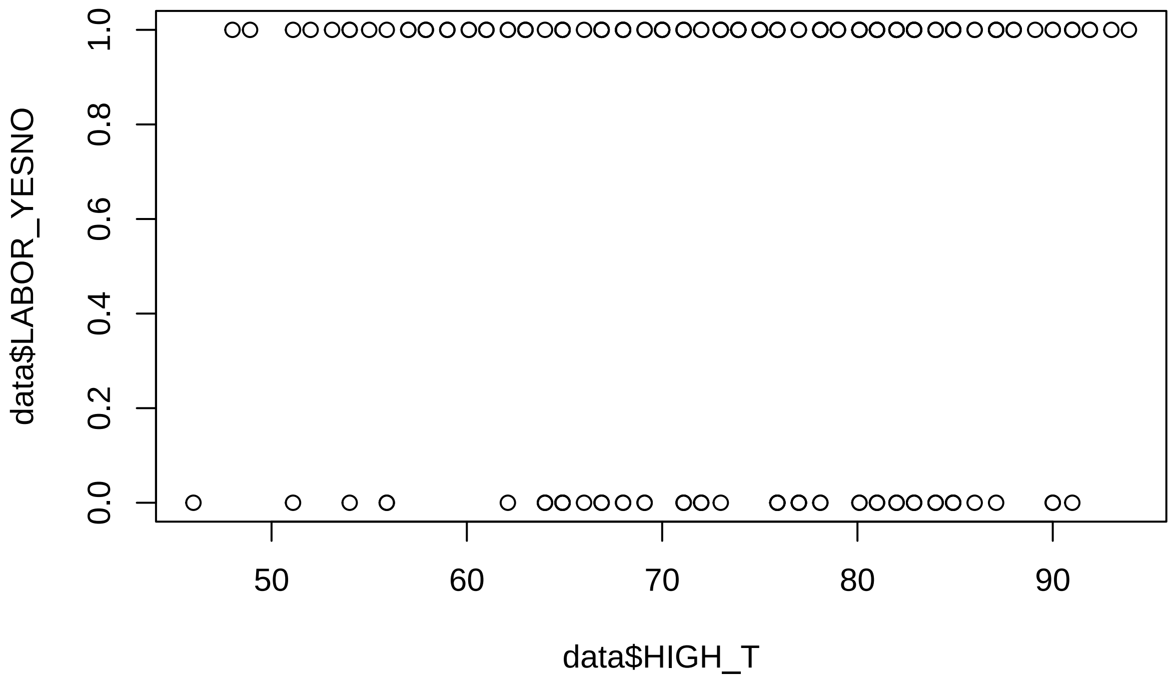
```
res_lab<-resid(mdl_pois_reg_laboral)
ggplot(data, aes(x=mdl_pois_reg_laboral[["linear.predictors"]], y= res_lab,
  color =as.factor(LABOR_YESNO))) + geom_point()+
  xlab("HIGH_T") + ylab("BB_COUNT") + labs(color = "LABOR_YESNO",
  title="BB_COUNT vs HIGH_T with LABOR_YESNO distinction")+
  scale_color_manual(values = c("green", "black"))+theme_minimal()
```



1.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.** .

```
## It seems that when HIGH_T 8 BB_COUNT is hgher, but there seems to be no relation between
## LABOR_YESNO and HIGH_T, it would make sense if there was no relation, as they are the temperature
## andif the day is laborable or not, wich are things we know are not related. Nevertheless, we will
## prove it
```

```
plot(data$HIGH_T, data$LABOR_YESNO)
```



```
model_H_lab <- lm(formula =HIGH_T ~ LABOR_YESNO, data = data)
anova(model_H_lab, test="Chisq")
```

```
## Analysis of Variance Table
##
## Response: HIGH_T
##              Df  Sum Sq Mean Sq F value Pr(>F)
## LABOR_YESNO   1     4.7   4.719  0.0435  0.835
## Residuals   212 22991.0 108.448
```

```
## In the plot we can observe that they are not related, we confirm it through an anova on a lm test
## of LABOR_YESNO on HIGH_T, where we can see that they are not significant on each other, which means
## that they are not related
```

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

.

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

```
##
## 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
```

```
## PRECIP is a significant predictor for BB_COUNT, along with LABOR_YESNO and HIGH_T as they all are
## significant (p_value<0.05)
```

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

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

```
##
## 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
```

```
## LOW_T is a significant predictor for BB_COUNT, along with LABOR_YESNO, HIGH_T and PRECIP as they
## all are significant (p_value<0.05)
```

1.p) **What would be your final model for the data? Justify your answer.**

```
## A model predicting BB_COUNT with all the other variables, to see if they are all significant in
## predicting the outcome variable
```

```
mdl_pois_reg_laboral <- glm(formula = BB_COUNT ~ HIGH_T + LABOR_YESNO + as.double(PRECIP) + as.double(L
summary(mdl_pois_reg_laboral)
```

```
##
## Call:
## glm(formula = BB_COUNT ~ HIGH_T + LABOR_YESNO + as.double(PRECIP) +
##     as.double(LOW_T) + Date, family = poisson(link = "log"),
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -33.039   -6.437    0.099    5.866   38.610
##
## Coefficients:
##                     Estimate Std. Error  z value Pr(>|z|)
## (Intercept)       -6.825e-01  3.915e-01   -1.743   0.0813 .
## HIGH_T             2.432e-02  2.983e-04   81.541   <2e-16 ***
## LABOR_YESNO        1.209e-01  3.012e-03   40.149   <2e-16 ***
## as.double(PRECIP) -7.647e-01  6.816e-03 -112.195   <2e-16 ***
## as.double(LOW_T)  -1.516e-02  3.255e-04  -46.568   <2e-16 ***
## Date               5.117e-09  2.622e-10   19.517   <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: 28087  on 208  degrees of freedom
## AIC: 30163
##
## Number of Fisher Scoring iterations: 4
```

1.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 outcomes would be:
##     Number of traffic accidents based on weather conditions.
##     Number of visits to E.R. from different causes of injury.
##     The number of calls a call center receives in a fixed period of time.
##     The number of hits a website receives in a given period of time
```