Coursework MAP501 2022

You will submit your coursework in the form of a single R notebook (i.e. `.Rmd` file) which can be rendered (“knitted”) to an `.pdf` document. Specifically, submit on Learn:

- your R notebook (i.e. the `.Rmd` file),
- the rendered `.pdf` version of your notebook. You might find it easier to knit to html, then print the html file to a pdf.

The coursework will be marked on the basis of correctness of code, interpretation of outputs and commentary as indicated. Therefore, please ensure that all code and outputs are visible in the knit document.

Preamble

```
library(rio)
library(dplyr)
library(janitor)
library(tidyverse)
library(here)
library(lindia)
library(tidyr)
library(magrittr)
library(ggplot2)
library(pROC)
library(car)
library(nnet)
library(caret)
library(lme4)
library(AmesHousing)
```

```
Ames<-make_ames()
```

1. Data Preparation

a. Import the soccer.csv dataset as “footballer_data”. (2 points)

```
#importing the soccer dataset as footballer_data
footballer_data <- read_csv(here("Data", "soccer.csv"))
```

b. Ensure all character variables are treated as factors and where variable names have a space, rename the variables without these. (3 points)

```
footballer_data <- clean_names(footballer_data) #cleaning and renaming variable names
footballer_data <- footballer_data %>%
  mutate(across(c(full_name, position, current_club, nationality), as.factor)) #converting charact
er variables to factors
```

c. Remove the columns birthday and birthday_GMT. (2 points)

```
#creating a new dataset named footballer_data2
footballer_data2 <- footballer_data %>%
  select(!birthday & !birthday_gmt) #removing columns birthday and birthday_gmt
```

d. Remove the cases with age<=15 and age>40. (2 points)

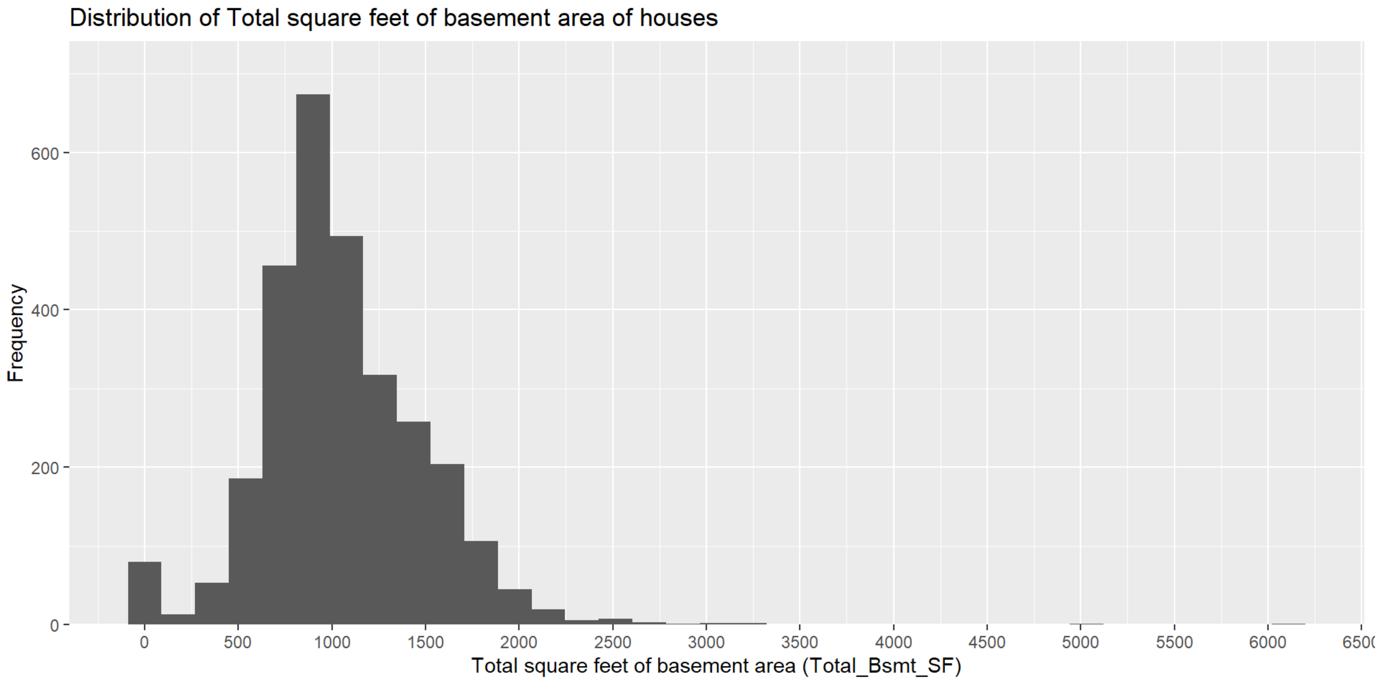
```
#observations with age less than or equal 15 and age greater 40 are removed
footballer_data2 <- footballer_data %>%
  filter(age >15 & age <= 40)
```

2. Linear Regression

In this problem, you are going to investigate the response variable Total_Bsmt_SF in "Ames" dataset through linear regression.

a. By adjusting x axis range and number of bars, create a useful histogram of Total_Bsmt_SF on the full dataset. Ensure that plot titles and axis labels are clear. (4 points)

```
#plotting a histogram of Total_Bsmt_SF
Ames %>%
  ggplot(mapping = (aes(x = Total_Bsmt_SF))) +
  geom_histogram(bins = 35) +
  scale_x_continuous(breaks = seq(0, 6500, 500)) +
  scale_y_continuous(breaks = seq(0, 1000, 200)) +
  scale_y_continuous(expand = expansion(mult = c(0, 0.1))) +
  labs(x = "Total square feet of basement area (Total_Bsmt_SF)",
       y = "Frequency") +
  ggtitle(label = "Distribution of Total square feet of basement area of houses")
```



The variable Total square feet of basement area (Total_Bsmt_SF) looks normally distributed with some outliers to the right.

b. Using "Ames" dataset to create a new dataset called "Ames2" in which you remove all cases corresponding to:

i. MS_Zoning categories of A_agr (agricultural), C_all (commercial) and I_all (industrial),
ii. BsmtFin_Type_1 category of "No_Basement".

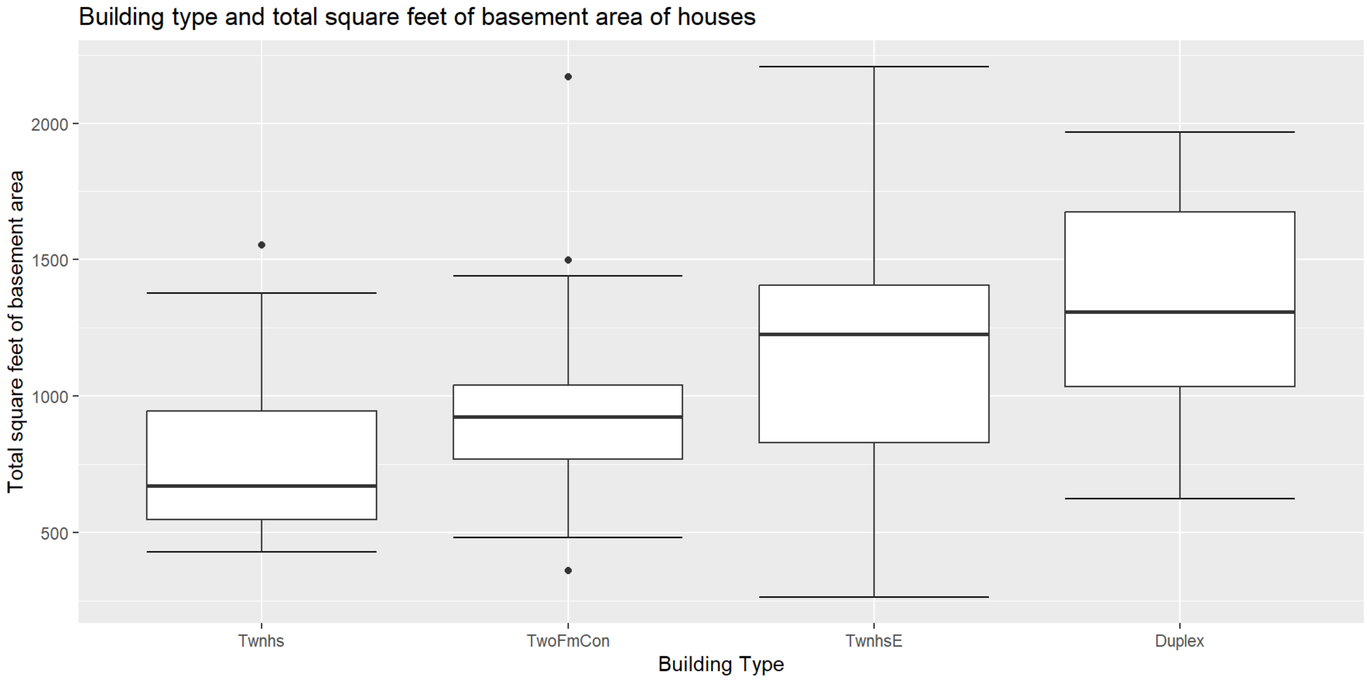
iii. Bldg_Type category of “OneFam”

and drop the unused levels from the dataset “Ames2”. (4 points)

```
#creating a new dataset called Ames2
Ames2 <- Ames %>%
  filter(!MS_Zoning %in% c("A_agr", "I_all","C_all"),
         Bldg_Type != "OneFam",
         BsmtFin_Type_1 != "No_Basement") %>% # removing A_agr, C_all and I_all categories of MS_Z
oning, OneFam from from Bldg_Type and No_Basement from BsmtFin_Type_1
  droplevels() #dropping all unused levels from the variables that were filtered
```

c. Choose an appropriate plot to investigate the relationship between Bldg_Type and Total_Bsmt_SF in Ames2. (2 points)

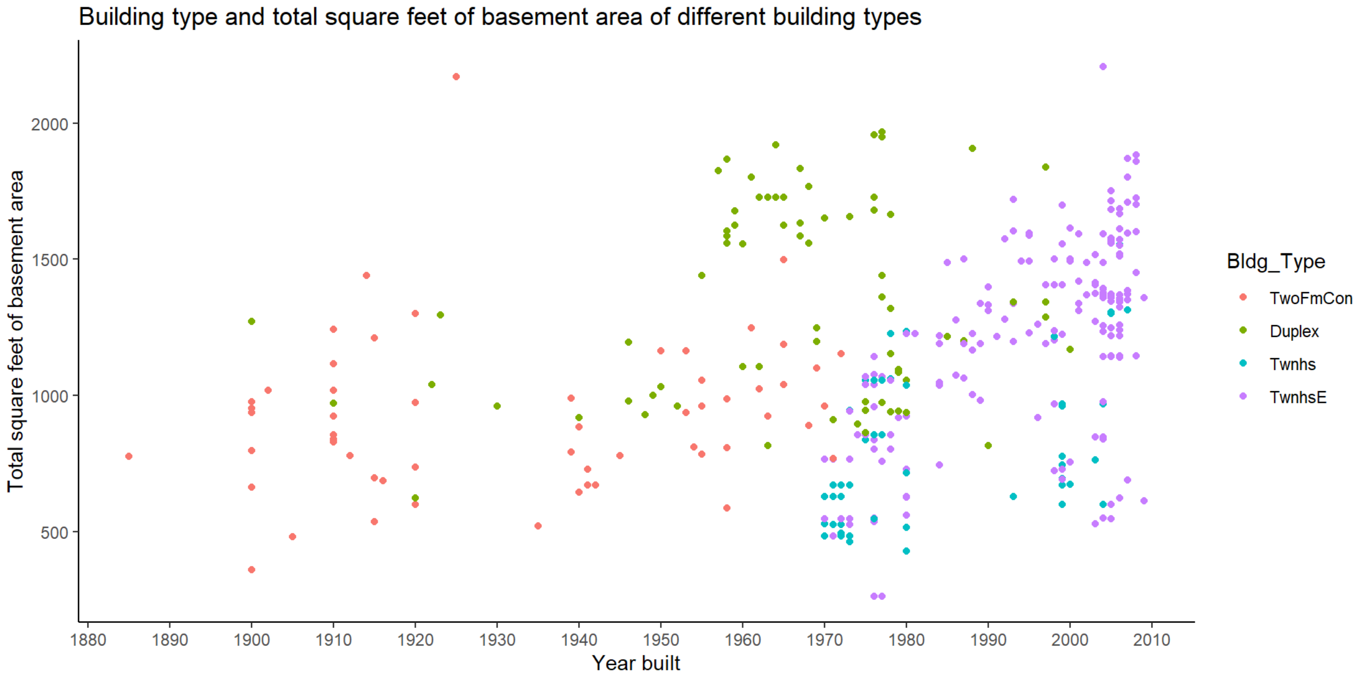
```
#creating a boxplot to investigate the relationship between Bldg_type and Total_Bsmt_SF
Ames2 %>%
  ggplot(mapping = (aes(x = reorder(Bldg_Type, Total_Bsmt_SF), y = Total_Bsmt_SF))) + # ordered th
e plot in ascending order
  stat_boxplot(geom = "errorbar") +
  geom_boxplot() +
  labs(x = "Building Type", y = "Total square feet of basement area") +
  ggtitle(label = "Building type and total square feet of basement area of houses")
```



From the boxplot above, Twnhs has the lowest median total square feet of basement area (Total_Bsmt_SF). Twnhse and Duplex have similar interquartile range. Twnhse also has the highest median and the most varaiablity in total square feet of basement area and is potentially left skewed. TwoFmCon has the most number of outliers (The Open University, no date).

d. Choose an appropriate plot to investigate the relationship between Year_Built and Total_Bsmt_SF in Ames2. Color points according to the factor Bldg_Type. Ensure your plot has a clear title, axis labels and legend. What do you notice about how Basement size has changed over time? Were there any slowdowns in construction over this period? When? Can you think why? (4 points)

```
Ames2 %>%
  ggplot(mapping = (aes(x = Year_Built,
                        y = Total_Bsmt_SF,
                        colour = Bldg_Type)))+
  geom_point() +
  scale_x_continuous(breaks = seq(1800, 2010, 10)) +
  labs(x = "Year built", y = "Total square feet of basement area")+
  ggtitle(label = "Building type and total square feet of basement area of different building type
s") +
  theme_classic()
```



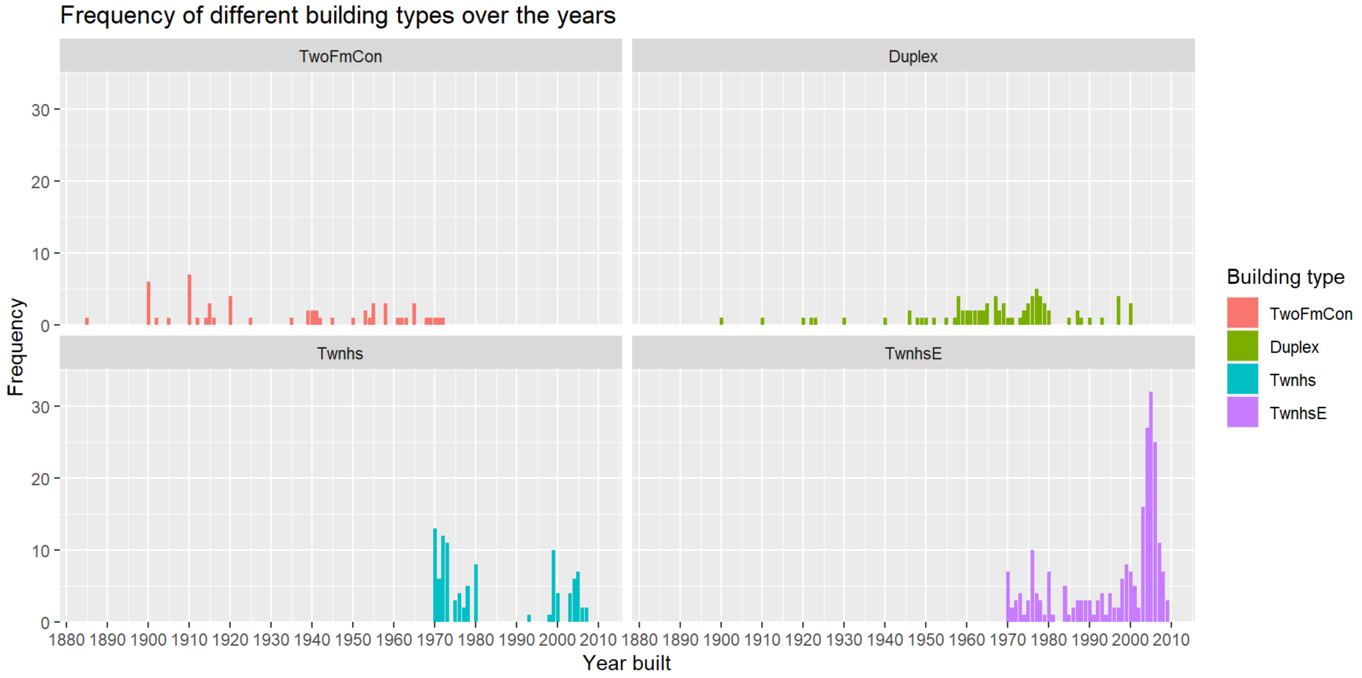
From the plot, it is observed that for most of the building types, the total basement area increased over time. It is also observed that building construction slowed down between year 1920 and 1945. This is probably due to the World war that took place around this time.

e. Why do we make these plots? Comment on your findings from these plots (1 sentence is fine). (2 points)

These plots are made to understand what type of relationship exists between the year and the total basement area. Knowing the kind of relationship that exists between the variables would inform the type of model we would build. It also helps us know if there would be need for a transformation before building the model.
From the plot it is observed that there is some linear relationship between total basement area and the year it was built for each building type. Therefore we can conclude that a linear model would be appropriate for the relationship Total_Bsmt_SF, Bldg_Type and Year_Built.

f. Now choose an appropriate plot to investigate the relationship between Bldg_Type and Year_Built in Ames2. Why should we consider this? What do you notice? (3 points)

```
#bar plot to investigate between Bldg_Type and Year_Built in Ames2
Ames2 %>%
  ggplot(mapping = (aes(x = Year_Built, fill = Bldg_Type)))+
  geom_bar(position = "dodge2") +
  labs(x = "Year built", y = "Frequency", fill = "Building type") +
  facet_wrap(~Bldg_Type) +
  ggtitle(label = "Frequency of different building types over the years") +
  scale_x_continuous(breaks = seq(1880, 2010, 10)) +
  scale_y_continuous(expand = expansion(mult = c(0, 0.1)))
```



It is necessary to consider this plot so that we can further visualise how the frequency of construction of different building types have changed over the years. Some buildings types were no longer featured over time.
For instance, after the year 1970 thereabout, it looks like the construction of TwoFmCon buildings stopped. In addition, from the year 2005, TwnhsE houses became more popular than the other building types. These observations would help determine how to treat the predictors and the model to be used. For example, do we need to drop certain categories because they are fewer in number when compared with the others. Also, filtering from one year to another will lead to the exclusion of some categories.

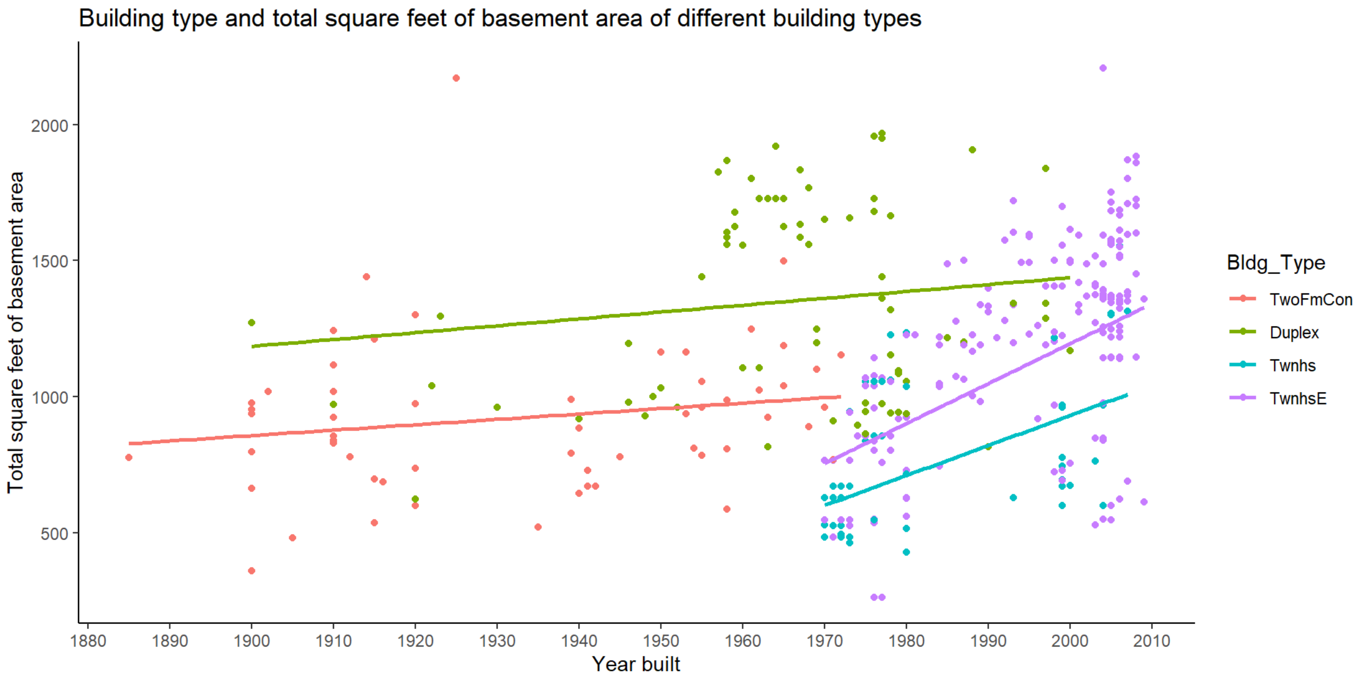
g. Use the lm command to build a linear model, linmod1, of Total_Bsmt_SF as a function of the predictors Bldg_Type and Year_Built for the “Ames2” dataset. (2 points)

```
#constructing a linear model (linmod1) of Total_Bsmt_SF as a function of Bldg_Type and Year_Built
linmod1 <- lm(Total_Bsmt_SF~Year_Built + Bldg_Type, data = Ames2)
summary(linmod1)
```

```


Call:
lm(formula = Total_Bsmt_SF ~ Year_Built + Bldg_Type, data = Ames2)

Residuals:
    Min      1Q  Median      3Q     Max 
-738.53 -223.35    7.68  238.36 1306.23 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -1.293e+04  1.833e+03  -7.054 6.30e-12 ***
Year_Built       7.166e+00  9.478e-01   7.560 2.15e-13 ***
Bldg_TypeDuplex  1.870e+02  6.504e+01   2.875  0.00422 ** 
Bldg_TypeTwnhs  -5.314e+02  7.252e+01  -7.327 1.04e-12 ***
Bldg_TypeTwnhsE -2.349e+02  7.678e+01  -3.059  0.00235 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 327.1 on 467 degrees of freedom
Multiple R-squared:  0.3339,	Adjusted R-squared:  0.3282 
F-statistic: 58.54 on 4 and 467 DF,  p-value: < 2.2e-16
```

h. State and evaluate the assumptions of the model. (6 points)

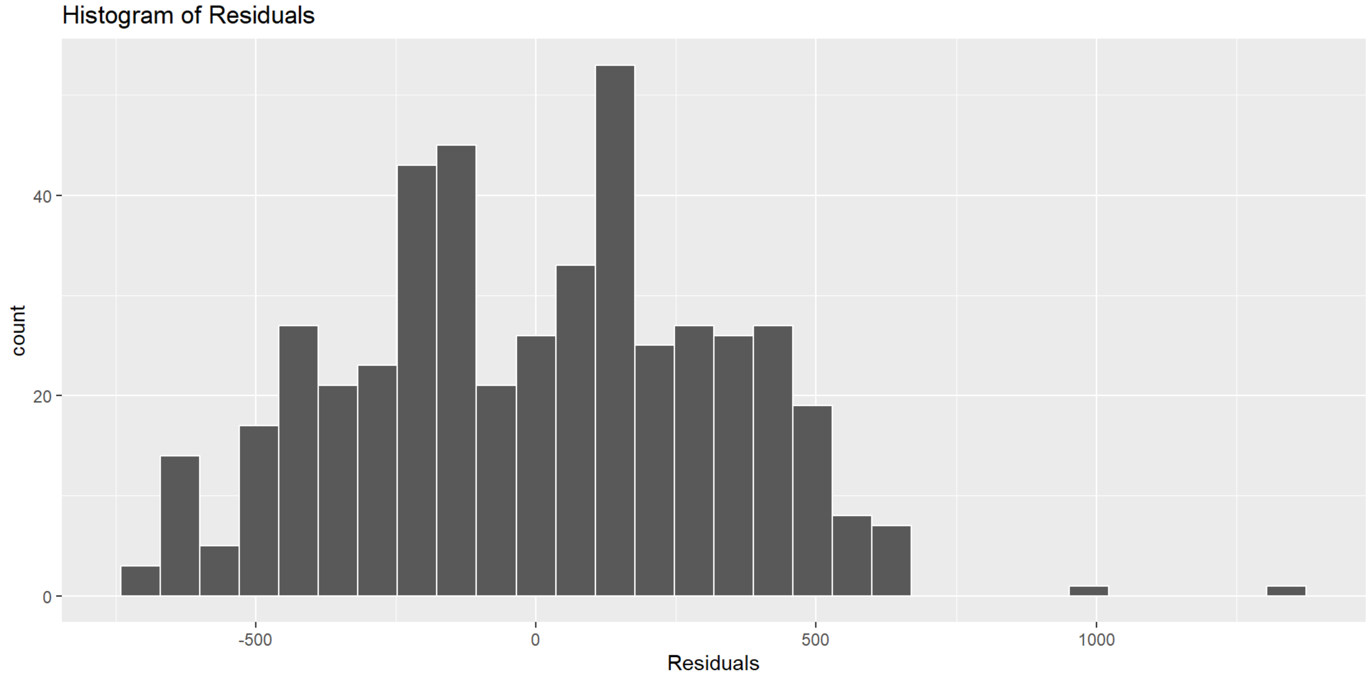
```
#evaluating the assumption of linearity
Ames2 %>%
  ggplot(mapping = (aes(x = Year_Built, y = Total_Bsmt_SF, colour = Bldg_Type)))+
  geom_point() +
  geom_smooth(method = lm, se = FALSE) +
  scale_x_continuous(breaks = seq(1800, 2010, 10)) +
  labs(x = "Year built", y = "Total square feet of basement area") +
  ggtitle(label = "Building type and total square feet of basement area of different building type
s") +
  theme_classic()
```

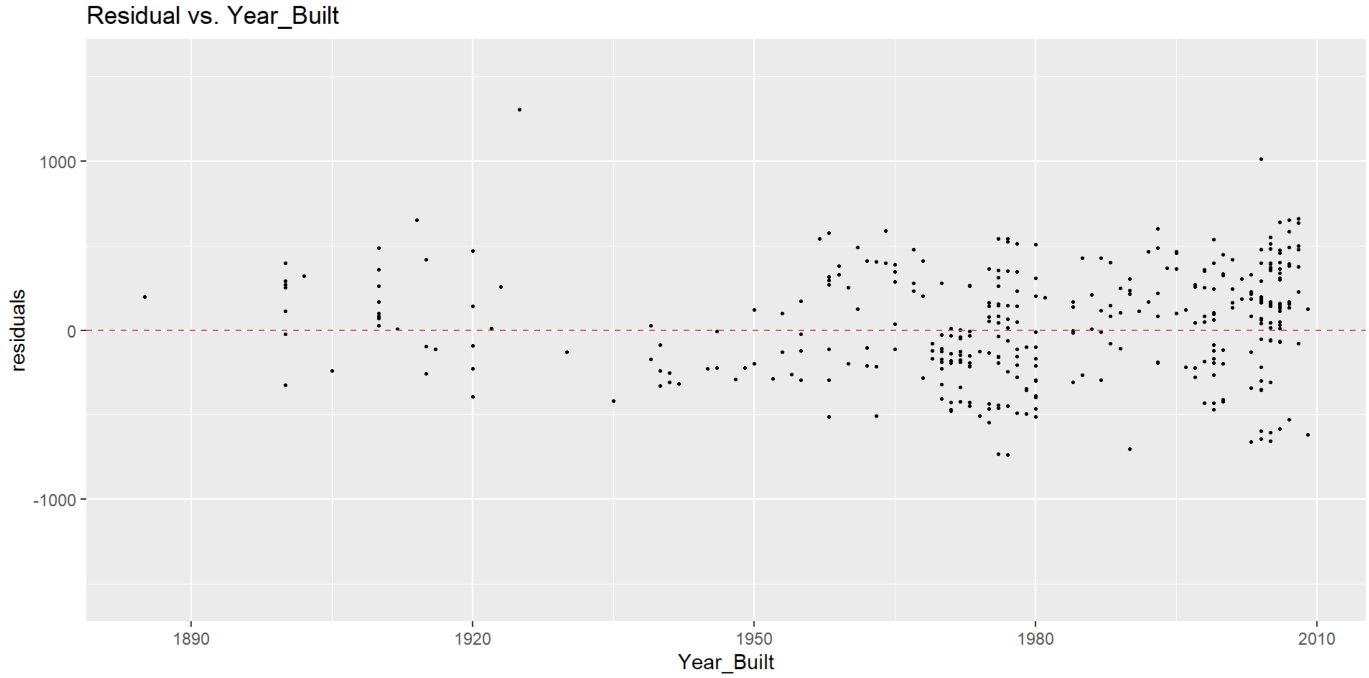


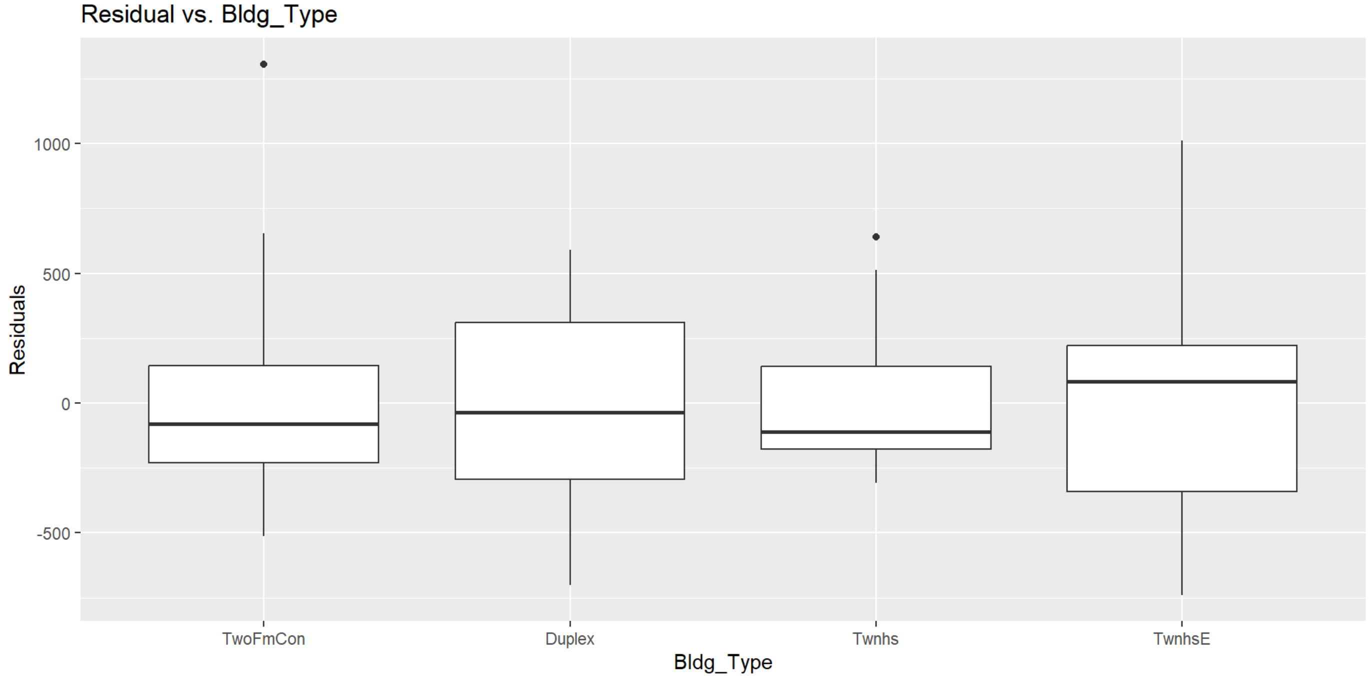
Assumption of linearity

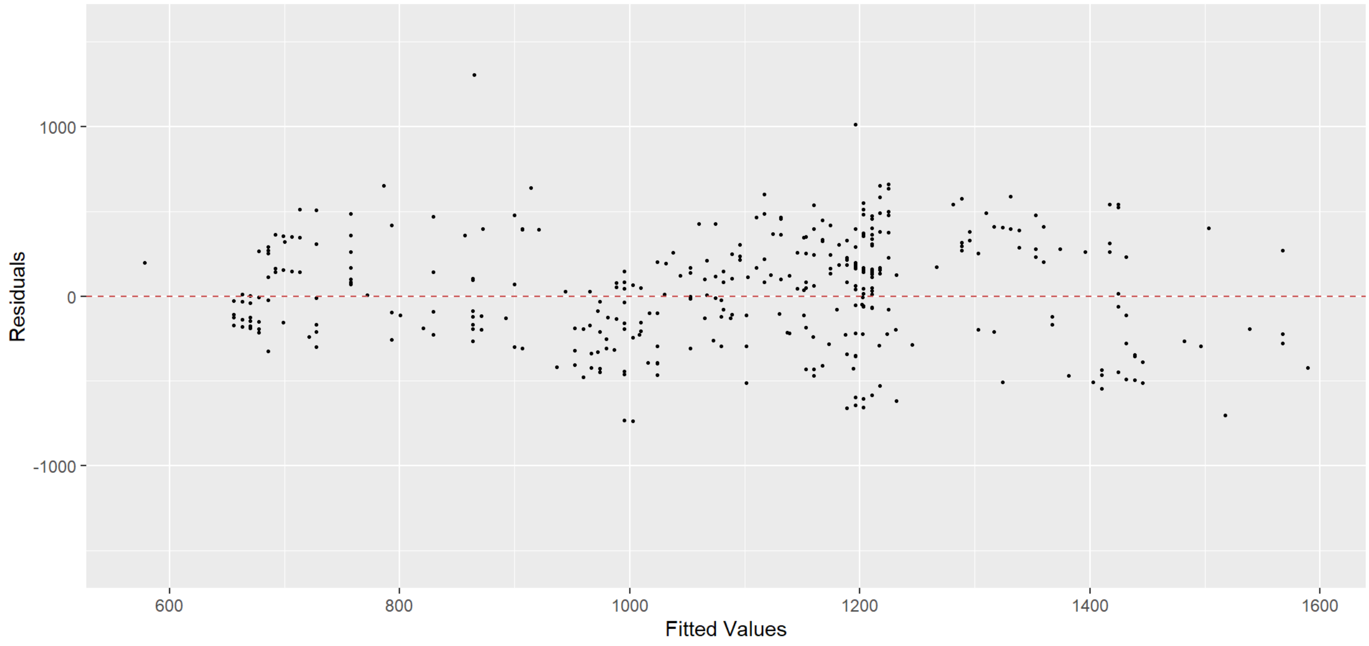
It looks like there is a roughly linear relationship between the total area of the basement of each building type and the year built. This is because as the year increases, the total basement area also increases.

```
linmod1 %>%
  gg_diagnose(max.per.page = 1)
```

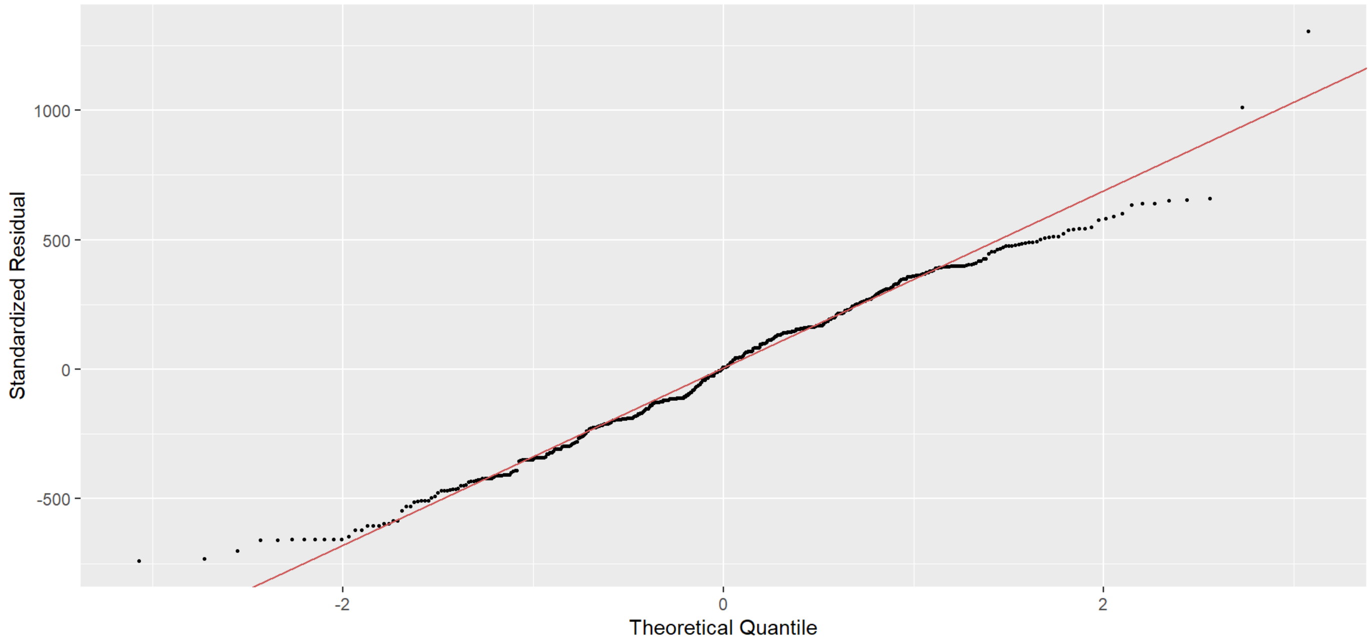

Histogram of Residuals
count
40
20
0
-500
0
500
1000
Residuals


Residual vs. Year_Built
residuals
1000
0
-1000
1890
1920
1950
1980
2010
Year_Built

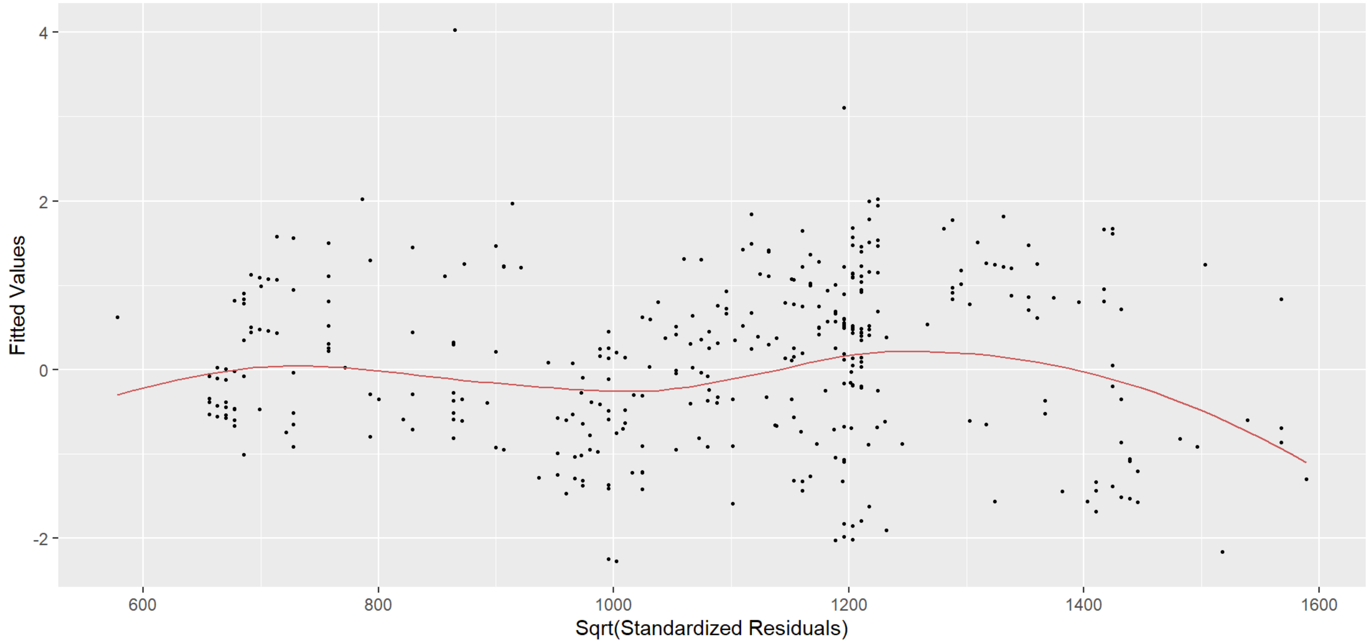

Residual vs. Bldg_Type
Residuals
1000
500
0
-500
TwoFmCon
Duplex
Twnhs
TwnhsE
Bldg_Type



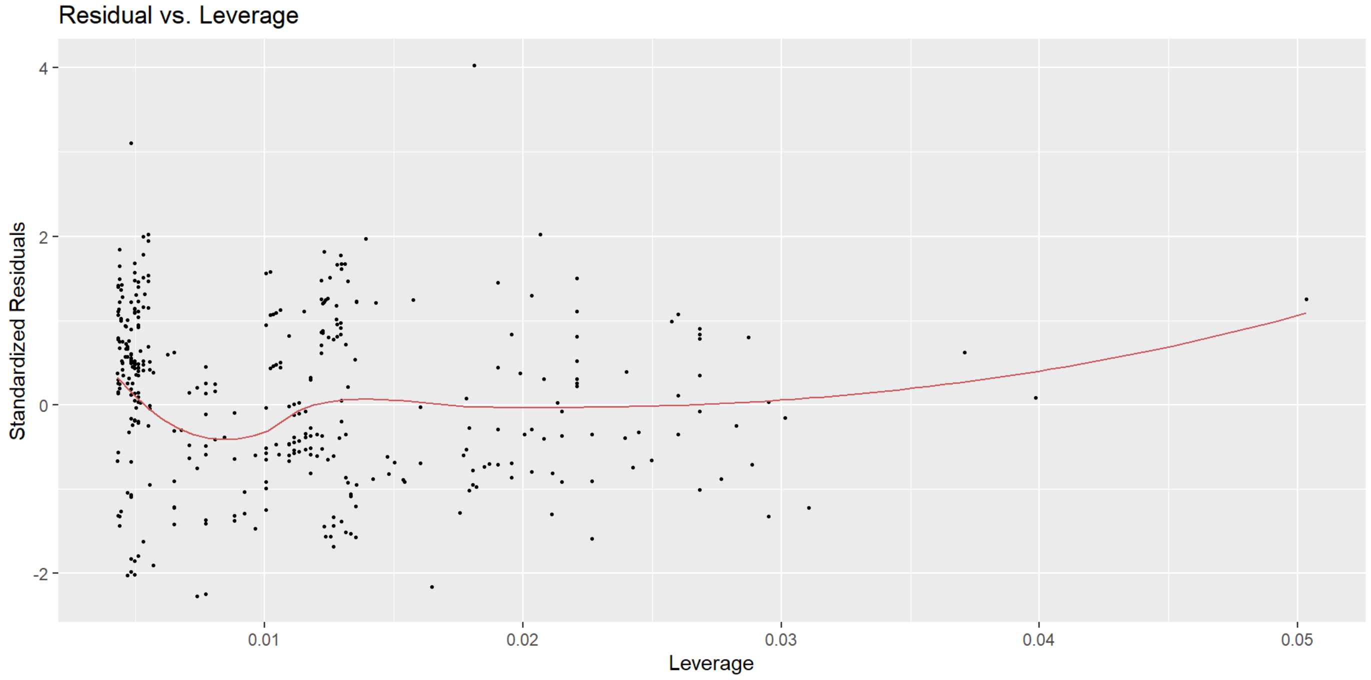
Residual vs. Fitted Value

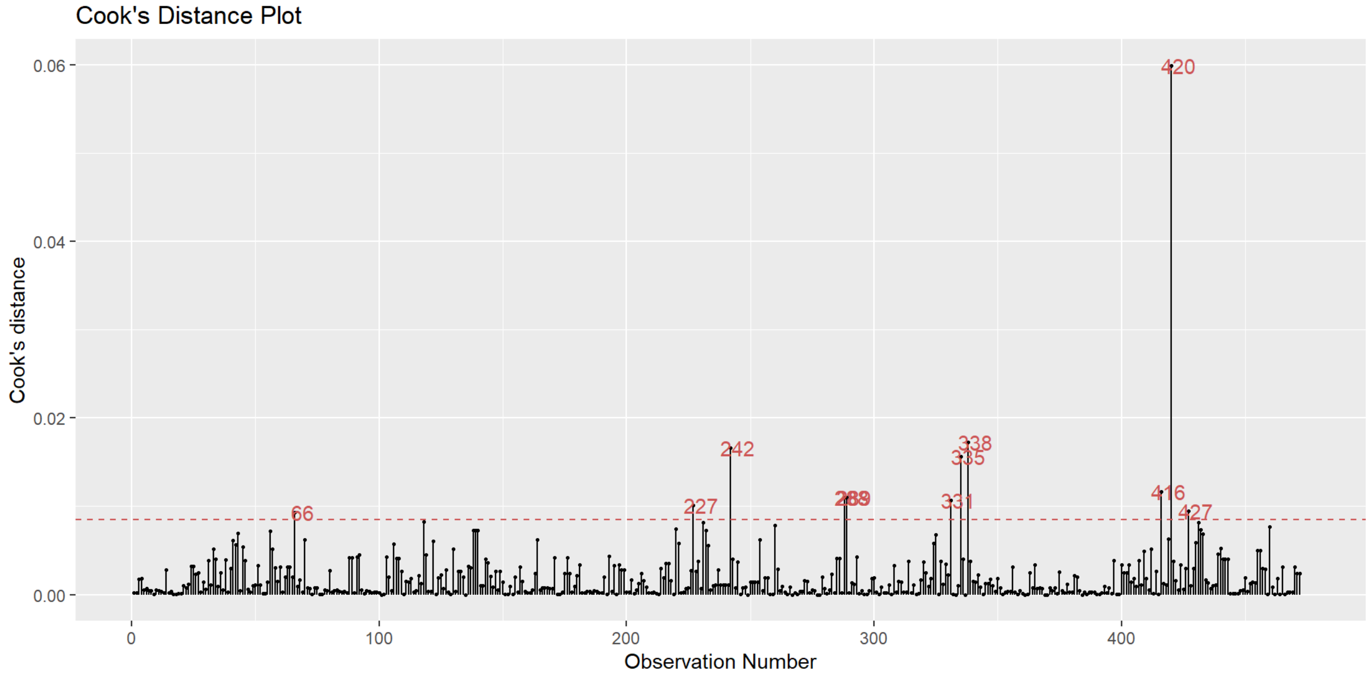


Normal-QQ Plot



Scale-Location Plot





Assumption of Normality

From the histogram and qq plot for residuals, the histogram looks normally distributed with some outliers to the right. From the qq plot, the residuals appear to be normally distributed except for some deviation at the tail ends of the plot. This is probably due to the outliers in the data set.

Assumption of homoscedasticity

From the scatter plots of residuals against year_built and residuals against fitted values. They both look like they are randomly distributed across the plot with no indication of trend. However, it does look like there are more data points between the year 1975 and 2010.
Considering the plot of the residuals versus the categorical variable (building type): There is some similarity among the interquartile ranges of the boxplots. All these suggest that the assumption of homoscedasticity is satisfied.

i. Use the lm command to build a second linear model, linmod2, for Total_Bsmt_SF as a function of Bldg_Type, Year_Built and Lot_Area. (2 points)

```
#constructing a linear model (linmod2) of Total_Bsmt_SF as a function of Year_Built, Lot_Area and
 Bldg_Type
linmod2 <- lm(Total_Bsmt_SF~Year_Built + Lot_Area + Bldg_Type, data = Ames2)
summary(linmod2)
```

```
Call:
lm(formula = Total_Bsmt_SF ~ Year_Built + Lot_Area + Bldg_Type,
    data = Ames2)

Residuals:
    Min      1Q  Median      3Q     Max
-810.32 -212.07   -5.72  233.88 1232.65

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -1.176e+04  1.828e+03  -6.435 3.08e-10 ***
Year_Built       6.509e+00  9.476e-01   6.868 2.09e-11 ***
Lot_Area         7.793e-03  1.960e-03   3.977 8.10e-05 ***
Bldg_TypeDuplex  2.378e+02  6.529e+01   3.642 0.000301 ***
Bldg_TypeTwnhs  -4.120e+02  7.745e+01  -5.319 1.62e-07 ***
Bldg_TypeTwnhsE -1.265e+02  8.035e+01  -1.575 0.115942
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 322.1 on 466 degrees of freedom
Multiple R-squared:  0.3558,    Adjusted R-squared:  0.3489
F-statistic: 51.48 on 5 and 466 DF,  p-value: < 2.2e-16
```

j. Use Analysis of variance (ANOVA) and Adjusted R-squared to compare these two models, and decide which is a better model. (6 points)

```
#anova for linmod1 and linmod2
anova(linmod1,linmod2) #anova comparing linmod1 and linmod2

#retrieving the adjusted R squared values for linmod1 and linmod2
summary(linmod1)
summary(linmod2)
```

```
Analysis of Variance Table

Model 1: Total_Bsmt_SF ~ Year_Built + Bldg_Type
Model 2: Total_Bsmt_SF ~ Year_Built + Lot_Area + Bldg_Type
  Res.Df      RSS Df Sum of Sq      F    Pr(>F)
1    467 49980160
2    466 48339705  1   1640455 15.814 8.099e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Call:
lm(formula = Total_Bsmt_SF ~ Year_Built + Bldg_Type, data = Ames2)

Residuals:
    Min      1Q  Median      3Q     Max
-738.53 -223.35    7.68  238.36 1306.23

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -1.293e+04  1.833e+03  -7.054 6.30e-12 ***
Year_Built       7.166e+00  9.478e-01   7.560 2.15e-13 ***
Bldg_TypeDuplex  1.870e+02  6.504e+01   2.875  0.00422 **
Bldg_TypeTwnhs  -5.314e+02  7.252e+01  -7.327 1.04e-12 ***
Bldg_TypeTwnhsE -2.349e+02  7.678e+01  -3.059  0.00235 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 327.1 on 467 degrees of freedom
Multiple R-squared:  0.3339,    Adjusted R-squared:  0.3282
F-statistic: 58.54 on 4 and 467 DF,  p-value: < 2.2e-16


Call:
lm(formula = Total_Bsmt_SF ~ Year_Built + Lot_Area + Bldg_Type,
    data = Ames2)

Residuals:
    Min      1Q  Median      3Q     Max
-810.32 -212.07   -5.72  233.88 1232.65

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -1.176e+04  1.828e+03  -6.435 3.08e-10 ***
Year_Built       6.509e+00  9.476e-01   6.868 2.09e-11 ***
Lot_Area         7.793e-03  1.960e-03   3.977 8.10e-05 ***
Bldg_TypeDuplex  2.378e+02  6.529e+01   3.642 0.000301 ***
Bldg_TypeTwnhs  -4.120e+02  7.745e+01  -5.319 1.62e-07 ***
Bldg_TypeTwnhsE -1.265e+02  8.035e+01  -1.575 0.115942
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 322.1 on 466 degrees of freedom
Multiple R-squared:  0.3558,    Adjusted R-squared:  0.3489
F-statistic: 51.48 on 5 and 466 DF,  p-value: < 2.2e-16
```

From the anova results, the p-value 0.000081 which is less than 0.05. This implies that linmod2 which includes Lot_Area as a predictor has a significant effect on Total_Bsmt_SF. This indicates that linmod2 is a better model than linmod1 which includes the predictors year_built and Bldg_type alone.

The first model (linmod1) has an adjusted R squared value of 0.3282 which means that it explains 32.8% of the variance in Total_Bsmt_SF while the second model (linmod2) has an adjusted R squared value of 0.3489 which means that it explains 34.9% of the variance in Total_Bsmt_SF. This also suggests that linmod2 is a better model than linmod1.

k. Construct a confidence interval and a prediction interval for the basement area of a Twnhs built in 1980, with a lot Area of 7300. Explain what these two intervals mean. (6 points)

```
predict(linmod2, newdata = data.frame(Year_Built = 1980, Bldg_Type = "Twnhs", Lot_Area = 7300), in
terval = "confidence") # calculating the confidence interval
predict(linmod2, newdata = data.frame(Year_Built = 1980, Bldg_Type = "Twnhs", Lot_Area = 7300), in
terval = "prediction") # calculating the prediction interval
```

```
       fit      lwr      upr
1 768.7589 702.1423 835.3755
       fit      lwr      upr
1 768.7589 132.3605 1405.157
```

We are 95% confident that the a Twnhs house built in 1980 has a mean Total_Bsmt_Ft that falls between 702.14 and 837.38. For the prediction interval, we are 95% confident that the Total_Bsmt_Ft of any Twnhs house built in 1980 will be between 132.36 and 1405.16.
It is observed that the prediction interval is wider than the confidence interval. This is because there is a greater uncertainty when predicting individual values than the mean value (Minitab Blog, 2013)

l. Now build a linear mixed model, linmod3, for Total_Bsmt_SF as a function of Year_Built,MS_Zoning and Bldg_Type. Use Neighborhood as random effect. What is the critical number to pull out from this, and what does it tell us? (4 points)

```
linmod3 <- lmer(Total_Bsmt_SF~Year_Built + MS_Zoning + Bldg_Type + (1|Neighborhood), data = Ames2)
summary(linmod3)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula:
Total_Bsmt_SF ~ Year_Built + MS_Zoning + Bldg_Type + (1 | Neighborhood)
   Data: Ames2

REML criterion at convergence: 6566.8

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.1502 -0.5804 -0.0394  0.6330  4.4359

Random effects:
 Groups       Name        Variance Std.Dev.
 Neighborhood (Intercept) 35128    187.4
 Residual                 68517    261.8
Number of obs: 472, groups:  Neighborhood, 27

Fixed effects:
                                     Estimate Std. Error t value
(Intercept)                         -4890.652   2262.148  -2.162
Year_Built                              2.876      1.151   2.499
MS_ZoningResidential_High_Density     148.504    211.630   0.702
MS_ZoningResidential_Low_Density      288.369    198.824   1.450
MS_ZoningResidential_Medium_Density   109.234    197.814   0.552
Bldg_TypeDuplex                       264.530     59.046   4.480
Bldg_TypeTwnhs                        -63.140     87.020  -0.726
Bldg_TypeTwnhsE                       105.171     83.438   1.260

Correlation of Fixed Effects:
            (Intr) Yr_Blt MS_ZR_H MS_ZR_L MS_ZR_M Bld_TD Bld_TT
Year_Built  -0.996
MS_ZnnR_H_D -0.158  0.081
MS_ZnnR_L_D -0.169  0.085  0.912
MS_ZnnR_M_D -0.142  0.061  0.901   0.963
Bldg_TypDpl  0.378 -0.395  0.014  -0.017  -0.001
Bldg_TypTwn  0.546 -0.570  0.004   0.059  -0.006   0.617
Bldg_TypTwE  0.602 -0.626 -0.005   0.052  -0.010   0.662  0.911
```

The critical number to pull out of this model is the standard deviation of the residuals and neighbourhood. The standard deviation of the effect of neighbourhood is 187.4 and the residual standard deviation is 261.8. This means that neighbourhood is not an important consideration in this model.

m. Construct 95% confidence intervals around each parameter estimate for linmod3. What does this tell us about the significant of the random effect? (3 points)

```
#constructing the confidence intervals around each parameter estimate
confint(linmod3)
```

```
                                        2.5 %     97.5 %
.sig01                             114.916972  253.19244
.sigma                             244.221572  278.77404
(Intercept)                      -9699.022207 -595.99553
Year_Built                           0.691417    5.33084
MS_ZoningResidential_High_Density -254.190137  549.38121
MS_ZoningResidential_Low_Density   -91.829020  665.77648
MS_ZoningResidential_Medium_Density -266.136920  487.72182
Bldg_TypeDuplex                    145.073466  377.00462
Bldg_TypeTwnhs                    -249.148897  102.22240
Bldg_TypeTwnhsE                    -68.266514  263.18536
```

From the confidence intervals above, variables such as MS_ZoningResidential_High_Density, MS_ZoningResidential_Low_Density, MS_ZoningResidential_Medium_Density, Bldg_TypeTwnhs and Bldg_TypeTwnhsE range from negative to positive and contain zero in them. Therefore, this suggests that neighbourhood as a random effect is not significant.

n. Write out the full mathematical expression for the model in linmod2 and for the model in linmod3. Round to the nearest integer in all coefficients with modulus (absolute value) > 10 and to three decimal places for coefficients with modulus < 10. (4 points)

Mathematicl expression for linmod2

\$\$

$$\begin{aligned}
\mathrm{E(Total\_Bsmt\_SF)} =&(-11760 + 6.509 \times \mathrm{YEARBUILT} + 0.008 \times \mathrm{LOTAREA} \\
&+ 238 \times \mathrm{isDUPLEX} - 412 \times \mathrm{isTWNHS} \\
&- 127 \times \mathrm{isTWNHSE}, 322). \\
\mathrm{mean\ total\_Bsmt\_SF} \sim N&(-11760 + 6.509 \times \mathrm{YEARBUILT} + 0.008 \times \mathrm{LOTAREA} \\
&+ 238 \times \mathrm{isDUPLEX} - 412 \times \mathrm{isTWNHS} \\
&- 127 \times \mathrm{isTWNHSE}, 322)
\end{aligned}$$

\$\$

Mathematical expression for linmod3

\$\$

$$\begin{aligned}
\mathrm{E(Total\_Bsmt\_SF)} =&(-4891 + 2.876 \times \mathrm{YEARBUILT} + 149 \times \mathrm{isHIGHDENSITY} \\
&+ 288 \times \mathrm{isLOWDENSITY} + 109 \times \mathrm{isMEDIUMDENSITY} \\
&+ 265 \times \mathrm{isDUPLEX} \\
&- 63 \times \mathrm{isTWNHS} \\
&+ 105 \times \mathrm{isTWNHSE} + U). \\
\mathrm{U} \sim N(0, 262)& \\
\mathrm{mean\ Total\_Bsmt\_SF} {\sim} N(\mathrm{Total\_Bsmt\_SF}), 187)&
\end{aligned}$$

\$\$

3. Logistic Regression

a. Do the following:

i. Create a new dataset called “Ames3” that contains all data in “Ames” dataset plus a new variable “excellent_heating” that indicates if the heating quality and condition “Heating_QC” is excellent or not. (2 points)

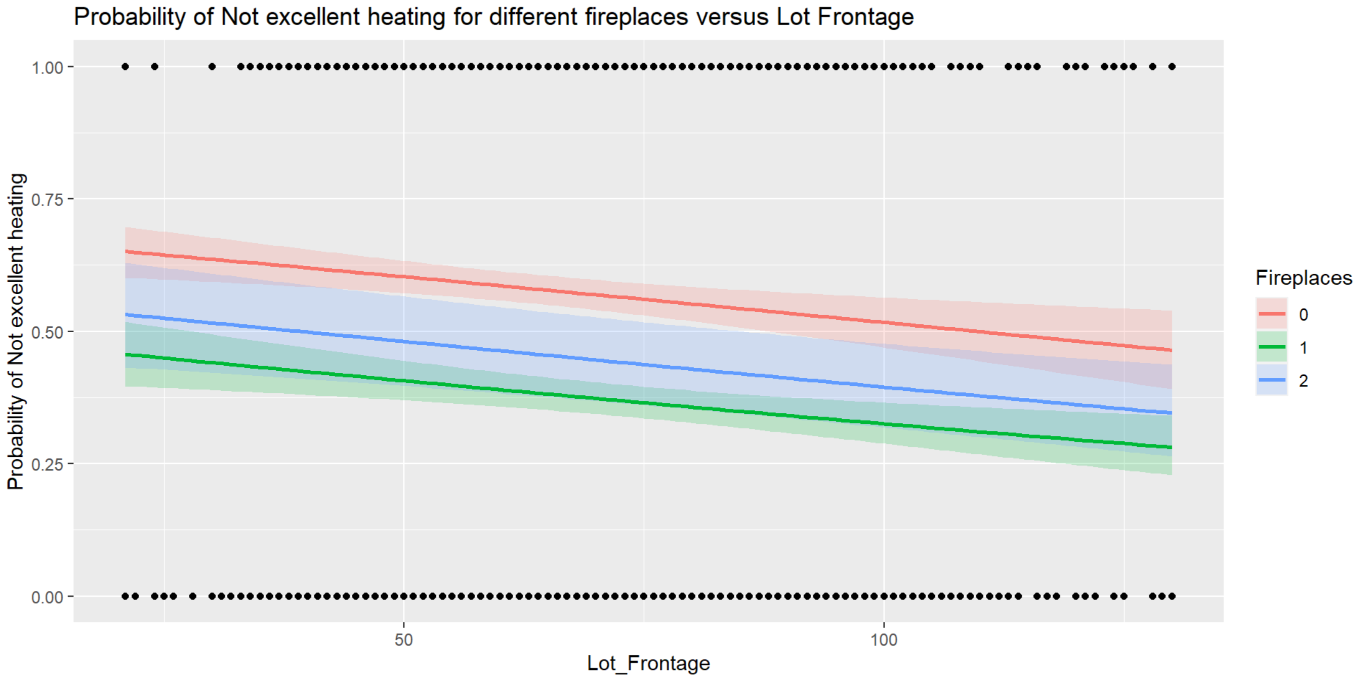
ii. In “Ames3” dataset, remove all cases “3” and “4” corresponding to the Fireplaces variable. Remove all cases where Lot_Frontage is greater than 130 or smaller than 20. Drop the unused levels from the dataset. (2 points)
iii. Save “Fireplaces” as factor in “Ames3” dataset (1 point)
iv. Construct a logistic regression model glmod for excellent_heating as a function of Lot_Frontage and Fireplaces for the dataset “Ames3”. (2 points)

```
Ames3 <- Ames %>%
  mutate(excellent_heating =
           as.factor(dplyr::recode
                     (Heating_QC, Fair = "Not excellent",
                       Good = "Not excellent",
                       Poor = "Not excellent",
                       Typical = "Not excellent",
                       Excellent = "Excellent"))) %>% # recoding the  categorical variables in Hea
ting_QC into 2 categories excellent and not excellent, naming it excellent_heating and making it a
factor
  filter(!Fireplaces %in% c("3", "4"),
         Lot_Frontage > 20 & Lot_Frontage < 131) %>% #removing cases with lot frontage less than 2
0 and greater than 130
  mutate(Fireplaces = as.factor(Fireplaces)) %>% #saving fireplaces as a factor
  droplevels() #dropping unused fireplaces levels

glmod <- glm(as.factor(excellent_heating)~Lot_Frontage + Fireplaces, family = "binomial", data = A
mes3) #building the regression model
summary(glmod)
```

```
Call:
glm(formula = as.factor(excellent_heating) ~ Lot_Frontage + Fireplaces,
    family = "binomial", data = Ames3)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4503  -1.0442  -0.8695   1.0761   1.5932

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.769387   0.147977   5.199    2e-07 ***
Lot_Frontage -0.007018   0.002137  -3.285  0.00102 **
Fireplaces1  -0.796183   0.088528  -8.994  < 2e-16 ***
Fireplaces2  -0.494887   0.176308  -2.807  0.00500 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3324.2  on 2399  degrees of freedom
Residual deviance: 3213.3  on 2396  degrees of freedom
AIC: 3221.3

Number of Fisher Scoring iterations: 4
```

b. Construct confidence bands for the variable excellent_heating as a function of Lot_Frontage for each number of Fireplaces (hint: create a new data frame for each number of Fireplaces). Colour these with different transparent colours for each number of Fireplaces and plot them together on the same axes. Put the actual data on the plot, coloured to match the bands, and jittered in position to make it possible to see all points. Ensure you have an informative main plot title, axes labels and a legend. (7 points)

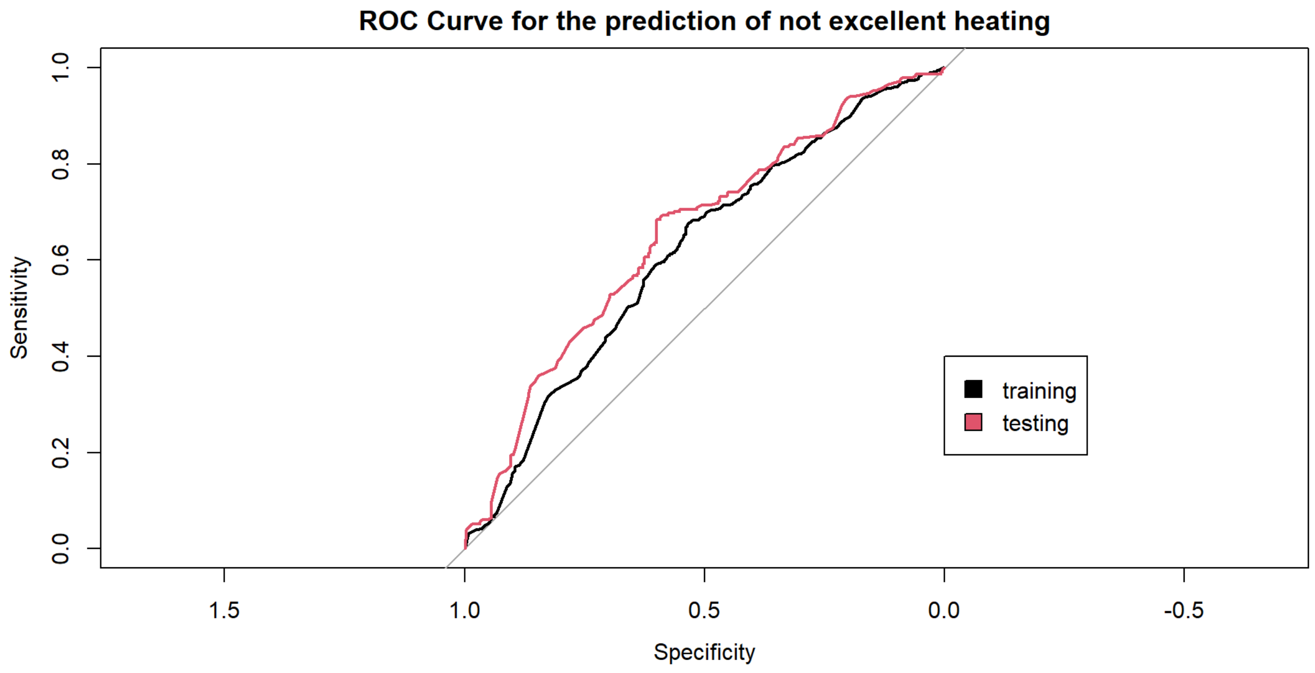
```
#constructing confidence bands for the variable excellent_heating as a function of Lot_Frontage for each number of Fireplaces
ilink <- family(glmod)$linkinv
newlotfrontage <- with(Ames3, data.frame(Lot_Frontage = seq(min(Ames3$Lot_Frontage),max(Ames3$Lot_Frontage), length = 100),Fireplaces = factor(rep(0:2, each = 100))))

newlotfrontage <- cbind(newlotfrontage,
                        predict(glmod, newlotfrontage,
                                type = "link", se.fit = TRUE)[1:2])
newlotfrontage <- transform(newlotfrontage, Fitted = ilink(fit),
                            Upper = ilink(fit+(1.96 *se.fit)),
                            Lower = ilink(fit-(1.96*se.fit)))

ggplot(Ames3, mapping = aes(x = Lot_Frontage, y = as.numeric(excellent_heating)- 1)) +
  geom_ribbon(data = newlotfrontage,
              aes(ymin = Lower, ymax = Upper,
                  x = Lot_Frontage, fill = Fireplaces),
              alpha = 0.2, inherit.aes = FALSE) +
  geom_line(data = newlotfrontage,
            aes(y= Fitted, x = Lot_Frontage,
                colour = Fireplaces), size = 1) +
  geom_point() +
  labs(x = "Lot_Frontage", y = "Probability of Not excellent heating") +
  ggtitle("Probability of Not excellent heating for different fireplaces versus Lot Frontage")
```



From the plot above, it is observed that the probability of not excellent heating decreases for all the types of fireplaces as the Lot Frontage increases.

c. Split the data using set.seed(120) and rebuild the model on 80% of the data. Cross validate on the remaining 20%. Plot the ROCs for both data and comment on your findings. (6 points)

```
set.seed(120)
training_sample1 <- c(Ames3$excellent_heating) %>%
  createDataPartition(p = 0.8, list = FALSE) #splitting the data into training and testing data
train_data1 <- Ames3[training_sample1, ]
test_data1 <- Ames3[-training_sample1, ]
train_model1 <- glm(excellent_heating~Lot_Frontage + Fireplaces,
                    data = train_data1,
                    family = "binomial")#creating the model on the training data

predtrain <- predict(train_model1,
                     type = "response") #predict the values on the training data
predtest <- predict(train_model1,
                    newdata = test_data1,
                    type = "response") #predict the values on the testing data
roctrain <- roc(response = train_data1$excellent_heating,
                predictor = predtrain,
                plot= TRUE,
                main = "ROC Curve for the prediction of not excellent heating", auc = TRUE) #roc c
urve for the training data

roc(response = test_data1$excellent_heating,
    predictor = predtest,
    plot = TRUE,
    auc = TRUE,
    add = TRUE,
    col = 2) #roc curve for the testing data
legend(0, 0.4, legend = c("training", "testing"), fill = 1:2)
```



```
Call:
roc.default(response = test_data1$excellent_heating, predictor = predtest,     auc = TRUE, plot =
TRUE, add = TRUE, col = 2)

Data: predtest in 248 controls (test_data1$excellent_heating Excellent) < 231 cases (test_data1$ex
cellent_heating Not excellent).
Area under the curve: 0.6517
```

From the plot above, the training and testing curve are similar. There is only a small gap between the two curves which is nothing to worry about. Therefore, there is no overfitting in this model.

4. Multinomial Regression

a. For the dataset “Ames”, create a model multregmod to predict BsmtFin_Type_1 from Total_Bsmt_SF and Year_Remod_Add. (3 points)

```
#constructing multigremod model to predict BsmtFin_Type_1 as a function of Total_Bsmt_SF and Year_
Remod_Add
multigremod <- multinom(BsmtFin_Type_1~Total_Bsmt_SF + Year_Remod_Add, data = Ames)
multigremod
```

```
# weights:  28 (18 variable)
initial  value 5701.516737
iter  10 value 4611.614897
iter  20 value 4159.256251
iter  30 value 4153.561922
iter  40 value 4150.324235
iter  50 value 4146.549269
iter  60 value 4144.509436
iter  70 value 4144.474970
final  value 4144.474825
converged
Call:
multinom(formula = BsmtFin_Type_1 ~ Total_Bsmt_SF + Year_Remod_Add,
    data = Ames)

Coefficients:
            (Intercept) Total_Bsmt_SF Year_Remod_Add
BLQ           34.465254  6.282504e-05   -0.017706708
GLQ         -105.324418  1.030040e-03    0.052676145
LwQ           39.566891  1.243787e-05   -0.020550529
No_Basement    4.876103 -1.729079e-01    0.004007989
Rec           56.710979  1.596801e-06   -0.028929851
Unf          -29.377212 -6.987213e-04    0.015514051

Residual Deviance: 8288.95
AIC: 8324.95
```

b. Write out the formulas for this model in terms of P(No_Basement), P(Unf) P(Rec),P(BLQ), P(GLQ), P(LwQ), You may round coefficients to 3 dp. (4 points)

$$\mathrm{logit}(\mathrm{P}(\mathrm{No\_Basement})) = 4.876 - 0.173 \times \mathrm{Total\_Bsmt\_SF} + 0.004 \times \mathrm{year\_Remod\_Add}$$

$$\mathrm{P}(\mathrm{No\_Basement}) \sim B(\mathrm{inverselogit}(4.876 - 0.173 \times \mathrm{Total\_Bsmt\_SF} + 0.004 \times \mathrm{year\_Remod\_Add}), 1)$$

$$\mathrm{logit}(\mathrm{P}(\mathrm{UNF})) = -29.377 - 0.0006 \times \mathrm{Total\_Bsmt\_SF} + 0.016 \times \mathrm{year\_Remod\_Add}$$

$$\mathrm{P}(\mathrm{UNF}) \sim B(\mathrm{inverselogit}(-29.377 - 0.0006 \times \mathrm{Total\_Bsmt\_SF} + 0.016 \times \mathrm{year\_Remod\_Add}), 1)$$

$$\mathrm{logit}(\mathrm{P}(\mathrm{REC})) = 56.711 + 0.000002 \times \mathrm{Total\_Bsmt\_SF} - 0.029 \times \mathrm{year\_Remod\_Add}$$

$$\mathrm{P}(\mathrm{REC}) \sim B(\mathrm{inverselogit}(56.711 + 0.000002 \times \mathrm{Total\_Bsmt\_SF} - 0.029 \times \mathrm{year\_Remod\_Add}), 1)$$

$$\mathrm{logit}(\mathrm{P}(\mathrm{BLQ})) = 34.465 + 0.00006 \times \mathrm{Total\_Bsmt\_SF} - 0.002 \times \mathrm{year\_Remod\_Add}$$

$$\mathrm{P}(\mathrm{BLQ}) \sim B(\mathrm{inverselogit}(34.465 + 0.00006 \times \mathrm{Total\_Bsmt\_SF} - 0.002 \times \mathrm{year\_Remod\_Add}), 1)$$

$$\mathrm{logit}(\mathrm{P}(\mathrm{GLQ})) = -105.324 + 0.001 \times \mathrm{Total\_Bsmt\_SF} + 0.053 \times \mathrm{year\_Remod\_Add}$$

$$\mathrm{P}(\mathrm{GLQ}) \sim B(\mathrm{inverselogit}(-105.324 + 0.001 \times \mathrm{Total\_Bsmt\_SF} + 0.053 \times \mathrm{year\_Remod\_Add}), 1)$$

$$\mathrm{logit}(\mathrm{P}(\mathrm{LWQ})) = 39.567 + 0.00001 \times \mathrm{Total\_Bsmt\_SF} - 0.021 \times \mathrm{year\_Remod\_Add}$$

$$\mathrm{P}(\mathrm{LWQ}) \sim B(\mathrm{inverselogit}(39.567 + 0.00001 \times \mathrm{Total\_Bsmt\_SF} - 0.021 \times \mathrm{year\_Remod\_Add}), 1)$$

$$\mathrm{P}(\mathrm{ALQ}) = 1 - P(\mathrm{No\_Basement}) - P(\mathrm{UNF}) - \mathrm{P}(\mathrm{REC}) - \mathrm{P}(\mathrm{BLQ}) - \mathrm{P}(\mathrm{GLQ}) - \mathrm{P}(\mathrm{LWQ})$$

c. Evaluate the performance of this model using a confusion matrix and by calculating the sum of sensitivities for the model. Comment on your findings. (4 points)

```
#evaluating the perfomance of the model using a confusion matrix
multitable <- table(Ames$BsmtFin_Type_1, predict(multigremod, type = "class"))
names(dimnames(multitable)) <- list("Actual", "Predicted")
multitable
```

```
             Predicted
Actual        ALQ BLQ GLQ LwQ No_Basement Rec Unf
  ALQ           1   0 117   0           0  18 293
  BLQ           0   0  50   0           0  30 189
  GLQ           1   0 579   0           0   2 277
  LwQ           1   0  38   0           0  30  85
  No_Basement   0   0   0   0          80   0   0
  Rec           3   0  31   0           0  46 208
  Unf           6   0 291   0           0  76 478
```

```
#calculating the percentage sensitivity for each Basement type
Sensitivity_AlQ <-  1/(1+117+18+293+0+0+0)*100
Sensitivity_BLQ <- 0/(50+30+189+0+0+0)*100
Sensitivity_GLQ <- 579/(1++579+0+2+277)*100
Sensitivity_LwQ <- 0/(1+0+38+0+0+30+85)*100
Sensitivity_No_Basement <- 80/(0+0+0+0+80+0+0)*100
Sensitivity_Rec <- 46/(3+0+31+0+0+46+208) *100
Sensitivity_Unf <- 478/(6+0+291+0+0+76+478) *100

Sensitivity_AlQ
Sensitivity_BLQ
Sensitivity_GLQ
Sensitivity_LwQ
Sensitivity_No_Basement
Sensitivity_Rec
Sensitivity_Unf
```

```
[1] 0.2331002
[1] 0
[1] 67.40396
[1] 0
[1] 100
[1] 15.97222
[1] 56.16921
```

```
SSens <-multitable[1,1]/sum(Ames$BsmtFin_Type_1 == "ALQ") +
  multitable[2,2]/sum(Ames$BsmtFin_Type_1 == "BLQ") +
  multitable[3,3]/sum(Ames$BsmtFin_Type_1 == "GLQ") +
  multitable[4,4]/sum(Ames$BsmtFin_Type_1 == "LwQ") +
  multitable[5,5]/sum(Ames$BsmtFin_Type_1 == "No_Basement") +
  multitable[6,6]/sum(Ames$BsmtFin_Type_1 == "Rec") +
  multitable[7,7]/sum(Ames$BsmtFin_Type_1 == "Unf") #calculating the sum of sensitivities for mult
igremod as SSens

SSens #retrieving the sum of sensitivities

CCR <- (multitable[1,1] + multitable[2,2] +  multitable[3,3] + multitable[4,4] + multitable[5,5] +
multitable[6,6] + multitable[7,7]) /length(Ames$BsmtFin_Type_1) # calculating the correct classifi
cation rate

CCR
```

```
[1] 2.397785
[1] 0.4040956
```

From the percentage of sensitivities calculated for each category, it is observed that no BLQ or LwQ houses were correctly predicted (0%). It has also not predicted well for ALQ (15.97%) and REC(15%). It has a 67.4% prediction for GLQ , and a 56.2% prediction for UNF which is okay. It predicted all the NO_Basements correctly.
From the correct classification rate calculated, only 40.4% of the data was identified correctly which means that more than half was classified wrongly and this is not good enough. The sum of sensitivities for this model is 2.40 which is less than the sum of sensitivities (5) when all categories are rightly predicted. Overall,this indicates that this model is not great.

5. Poisson/quasipoisson Regression

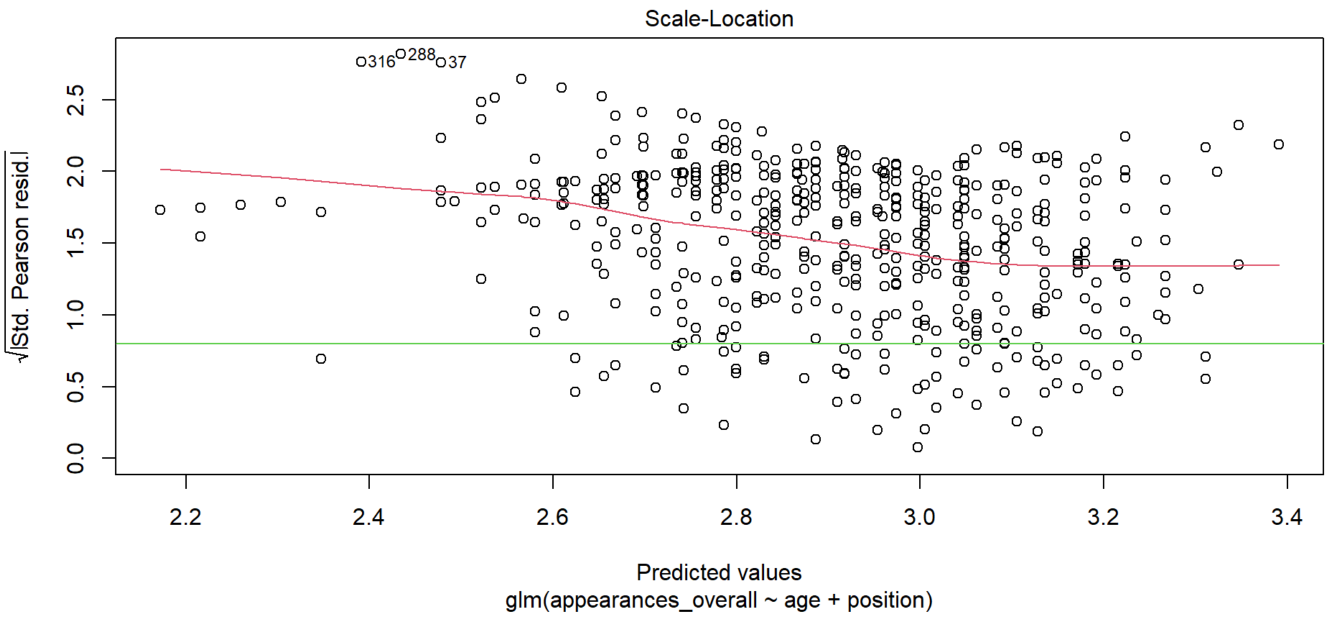
a. For the “footballer_data” dataset, create a model appearances_mod to predict the total number of overall appearances a player had based on position and age. (2 points)

```
#building appearances_mod model as a function of position and age
appearances_mod <- glm(appearances_overall~age + position, data = footballer_data2, family = "pois
son")
summary(appearances_mod)
```

```
Call:
glm(formula = appearances_overall ~ age + position, family = "poisson",
    data = footballer_data2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-7.5377  -3.5215   0.0351   2.1892   6.1853

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         1.575316   0.074884  21.037  < 2e-16 ***
age                 0.043704   0.002392  18.275  < 2e-16 ***
positionForward     0.110606   0.027448   4.030 5.59e-05 ***
positionGoalkeeper -0.364605   0.040780  -8.941  < 2e-16 ***
positionMidfielder  0.118259   0.023309   5.074 3.90e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 6539.7  on 564  degrees of freedom
Residual deviance: 6114.4  on 560  degrees of freedom
AIC: 8417.1

Number of Fisher Scoring iterations: 5
```

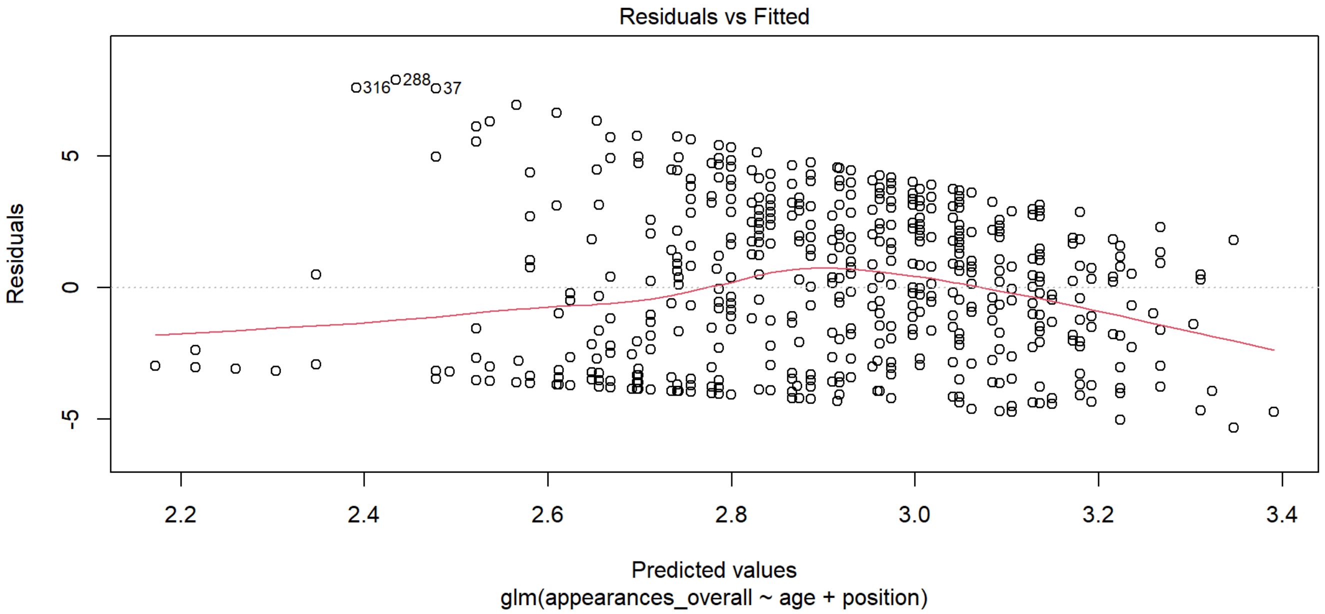
b. Check the assumption of the model using a diagnostic plot and comment on your findings. (3 points)

```
#evaluating the dispersion assumption
plot(appearances_mod, which = 3)
abline(h = 0.8, col = 3)
```



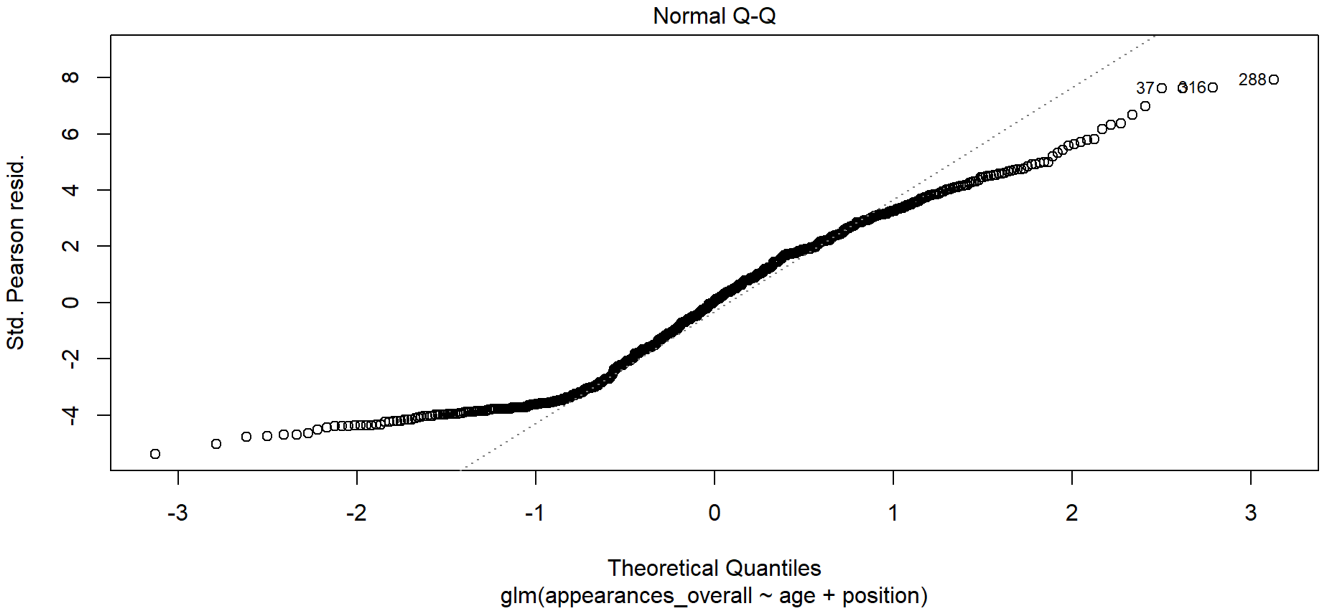
The red line is not flat and rises above the 0.8 (green) line. Also majority of the residuals lie above this line. This suggests that there is some overdispersion in the data that decreases as the prediction increases. This may be because we haven't accounted for all predictors in the model. A quasipoission estimation may be more suitable to build the model.

```
#evaluating for linearity
plot(appearances_mod, which = 1)
```



The red line is not flat when compared with the black line. It seems like there is some pattern with the residuals decreasing when the predicted values increase.

```
#evaluating normality
plot(appearances_mod, which = 2)
```

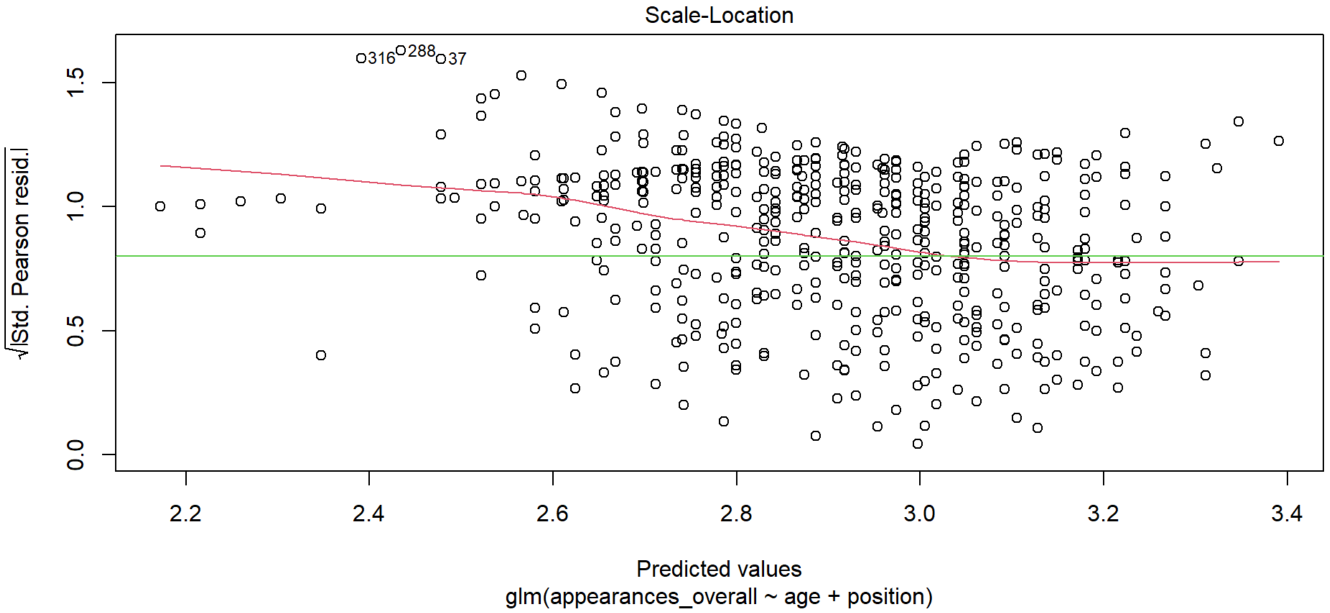


The plot above shows that there is a deviation from the black line especially at the tail. This suggests that there is a deviation from normality and the data is not normally distributed.

```
#building the model using quasipoisson as family
appearances_mod2 <- glm(appearances_overall~age + position, data = footballer_data2, family = "qua
sipoisson")
summary(appearances_mod2)
```

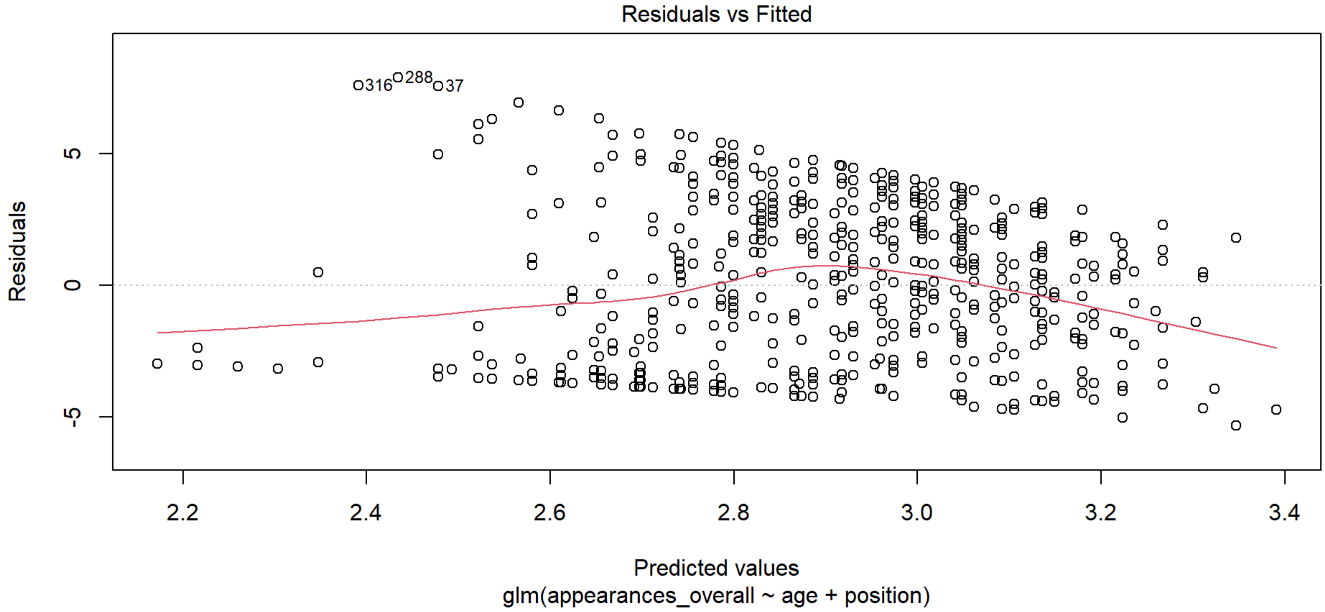
```
Call:
glm(formula = appearances_overall ~ age + position, family = "quasipoisson",
    data = footballer_data2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-7.5377  -3.5215   0.0351   2.1892   6.1853

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.575316   0.223980   7.033 5.90e-12 ***
age                 0.043704   0.007153   6.110 1.87e-09 ***
positionForward     0.110606   0.082097   1.347  0.17844
positionGoalkeeper -0.364605   0.121975  -2.989  0.00292 **
positionMidfielder  0.118259   0.069717   1.696  0.09039 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 8.946343)

    Null deviance: 6539.7  on 564  degrees of freedom
Residual deviance: 6114.4  on 560  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

```
#evaluating the dispersion assumption
plot(appearances_mod2, which = 3)
abline(h = 0.8, col = 3)
```



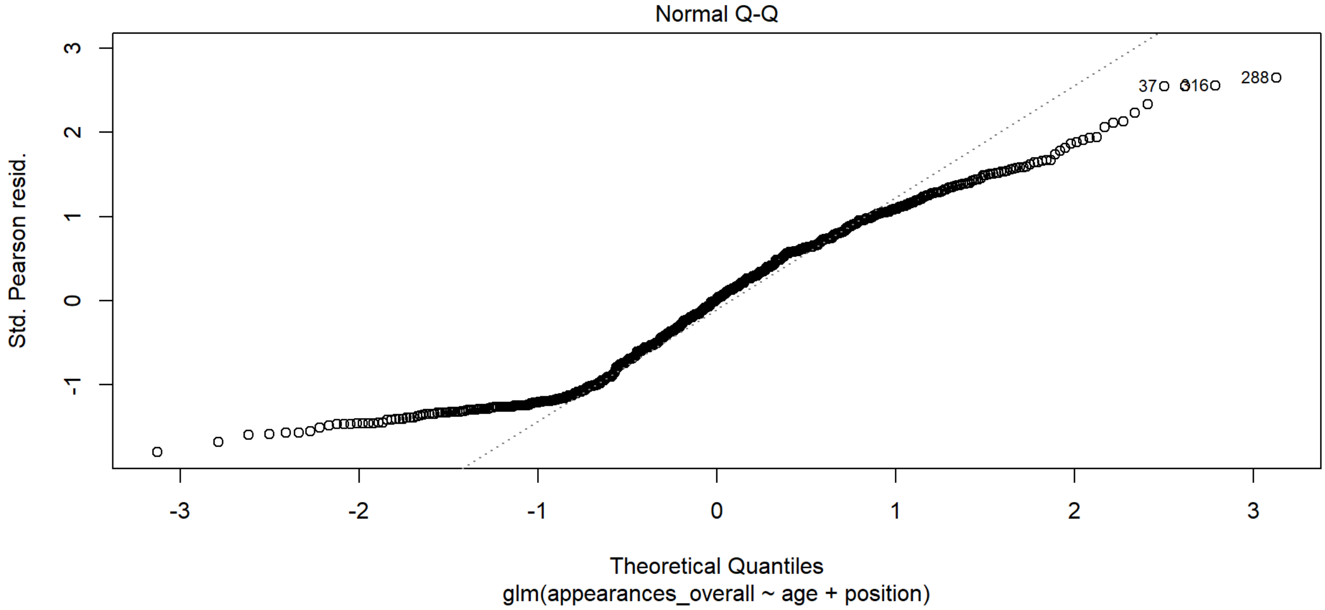
The red line is still not flat and it is above the green line especially on the left hand side of the plot. The data points look more evenly distributed across the zero line. It still looks like there is a little overdispersion but this looks way better than that of the poisson model.

```
#evaluating the linearity assumption
plot(appearances_mod2, which = 1)
```



The red line ia not flat when compared with the black line. It also looks like there is some sort of pattern. As the residuals decrease the predicted values increase.

```
#evaluating the normality assumption
plot(appearances_mod2, which = 2)
```



There is still some deviation from the line at the tail end this suggests that there is some deviation from normality

c. What do the coefficients of the model tell us about? which position has the most appearances? How many times more appearances do forwards get on average than goalkeepers? (3 points)

Mathematical expression for the appearances_mod2 model

$$

$$
\begin{aligned}
\log(\text{E(appearances)}) =& 1.575 + 0.044 \times \text{AGE} + 0.111 \times \text{isFORWARD} \\
& - 0.365 \times \text{isGOALKEEPER} + 0.118 \times \text{isMIDFIELDER}) \\
\text{mean appearances overall} \sim & Pois(exp(1.575 + 0.044 \times \text{AGE} + 0.111 \times \text{isFORWARD} \\
& - 0.365 \times \text{isGOALKEEPER} + 0.118 \times \text{isMIDFIELDER})
\end{aligned}
$$

$$

The coefficients tell us that the log mean number of appearances will increase by 0.044 when the age by 1 unit while keeping all other predictors remain constant. Midfielders have the most apperances.
In addition, after controlling for age of the player, goalkeepers' log mean number of appearances were 0.365 lower than those of defenders. For forwards, it is 0.111 higher than for defenders while for midfielders it is 0.118 higher than for defenders.
In other words, when age is held constant, the mean number of appearances of goalkeepers is $e^{-0.365}$ = 0.69 times higher (i.e 31% lower) than defenders (Roback and Legler, 2021 p.110).
In comparison to goalkeepers, forwards make $e^{0.476}$ = 1.61 appearances on average than goalkeepers which is approximately 2 appearances on average than goalkeepers.

REFERENCES

Roback, P. and Legler, J.M. (2021) “Poisson Regression,” in Beyond multiple linear regression: Applied generalized linear models and multilevel models in R. Boca Raton, FL: CRC Press, p. 110.

The Open University (no date) Interpreting data: Boxplots and tables, Interpreting data: boxplots and tables: View as single page. Available at: https://www.open.edu/openlearn/science-maths-technology/mathematics-statistics/interpreting-data-boxplots-and-tables/content-section-0/?printable=1 (https://www.open.edu/openlearn/science-maths-technology/mathematics-statistics/interpreting-data-boxplots-and-tables/content-section-0/?printable=1) (Accessed: December 1, 2022).

Minitab Blog Editor. (2013) When should I use confidence intervals, prediction intervals, and tolerance intervals, Minitab Blog. Available at: https://blog.minitab.com/en/adventures-in-statistics-2/when-should-i-use-confidence-intervals-prediction-intervals-and-tolerance-intervals (https://blog.minitab.com/en/adventures-in-statistics-2/when-should-i-use-confidence-intervals-prediction-intervals-and-tolerance-intervals) (Accessed: December 1, 2022).