STAT7111 & STAT8111 Generalized Linear Models Assignment 1

Umut Demirhan - Student ID: 46739106

25 August, 2023

Question 1

Part A: Examining graphically and numerically the correlation between variables

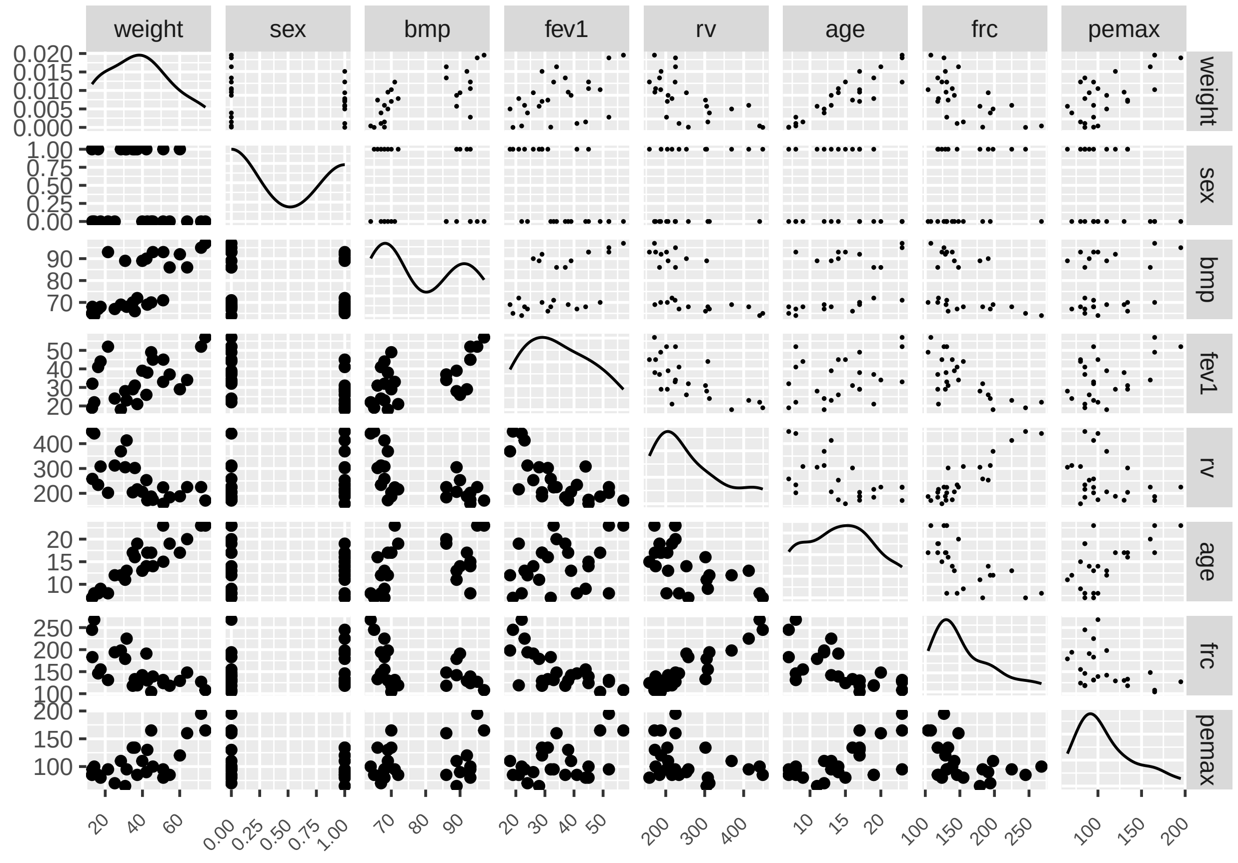
To see how variables linearly relate to each other, I'll calculate and present a correlation matrix:

```r
# Import the data
data(cystfibr)
# Select only the variables of interest
subdata <- cystfibr[,c("weight", "sex", "bmp", "fev1", "rv", "age", "frc", "pemax")]
# Calculate correlation matrix
cor_matrix <- cor(subdata)
# Display the correlation matrix
cor_matrix
```

```
##            weight        sex        bmp       fev1         rv        age
## weight  1.0000000 -0.1904400  0.6725463  0.4488393 -0.6215056  0.9058675
## sex    -0.1904400  1.0000000 -0.1375611 -0.5282571  0.2713516 -0.1671220
## bmp     0.6725463 -0.1375611  1.0000000  0.5455204 -0.5823729  0.3777643
## fev1    0.4488393 -0.5282571  0.5455204  1.0000000 -0.6658557  0.2944880
## rv     -0.6215056  0.2713516 -0.5823729 -0.6658557  1.0000000 -0.5519445
## age     0.9058675 -0.1671220  0.3777643  0.2944880 -0.5519445  1.0000000
## frc    -0.6172561  0.1836055 -0.4343888 -0.6651149  0.9106029 -0.6393569
## pemax   0.6352220 -0.2885692  0.2295148  0.4533757 -0.3155501  0.6134741
##               frc      pemax
## weight -0.6172561  0.6352220
## sex     0.1836055 -0.2885692
## bmp    -0.4343888  0.2295148
## fev1   -0.6651149  0.4533757
## rv      0.9106029 -0.3155501
## age    -0.6393569  0.6134741
## frc     1.0000000 -0.4172078
## pemax  -0.4172078  1.0000000
```

Next, I'll produce a scatterplot matrix for the *cystfibr* dataset. This will help visualize the relationships between the different variables:

```r
subdata %>%
  GGally::ggpairs(upper = list(continuous = wrap("points", size = 0.5, shape = 16))) +
  theme(axis.text.x = element_text(angle = 45, vjust = 0.5, hjust=1, size=7))
```



Interpretation:

- *Weight* and *age*: Strong positive correlation (0.9058675) suggests that as *weight* increases, *age* is likely to increase as well, and vice versa.
- *Weight* and *bmp*: Moderate positive correlation (0.6725463) suggests that an increase in weight is associated with an increase in *bmp*.
- *Weight* and *pemax*: Moderate positive correlation (0.6352220) indicates that as *weight* increases, *pemax* is also likely to increase.
- *rv* and *frc*: Strong positive correlation (0.9106029) suggests that *rv* and *frc* are likely to increase or decrease together.
- *fev1* and *bmp*: Moderate positive correlation (0.5455204) indicates that an increase in *fev1* is associated with an increase in *bmp*.
- *fev1* and *pemax*: Moderate positive correlation (0.4533757) indicates that an increase in *fev1* is likely associated with an increase in *pemax*.
- *sex* and *fev1*: Moderate negative correlation (-0.5282571) suggests that an increase in one variable is likely associated with a decrease in the other.
- *rv* and *fev1*: Moderate to strong negative correlation (-0.6658557) suggests that as *rv* increases, *fev1* is likely to decrease, and vice versa.
- *age* and *frc*: Moderate to strong negative correlation (-0.6393569) indicates that as *age* increases, *frc* is likely to decrease.
- *weight* and *rv*: Moderate to strong negative correlation (-0.6215056) suggests that as *weight* increases, *rv* is likely to decrease, and vice versa.
- *weight* and *frc*: Moderate to strong negative correlation (-0.6172561) suggests that as *weight* increases, frc is likely to decrease.

Potential multicollinearity issues should be monitored due to strong correlations among some variables.

Part B: Regression Analysis of the Variables

In this section, I analyze Model 1, where the dependent variable *'pemax'* is predicted using the independent variable *'weight'*. Mathematically, this relationship is represented as:

$$\text{pemax} = \beta_0 + \beta_1 \times \text{weight} + \varepsilon$$

Here, β_0 is the intercept, β_1 is the coefficient for the variable *'weight'*, and ε is the error term.

```r
# Fit a linear model with weight as the dependent variable
model1 <- lm(pemax ~ weight, data = cystfibr)

# Display the model summary
summary(model1)
```

```
##
## Call:
## lm(formula = pemax ~ weight, data = cystfibr)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -44.30 -22.69   2.23  15.91  48.41
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  63.5456    12.7016   5.003 4.63e-05 ***
## weight        1.1867     0.3009   3.944 0.000646 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.38 on 23 degrees of freedom
## Multiple R-squared:  0.4035, Adjusted R-squared:  0.3776
## F-statistic: 15.56 on 1 and 23 DF,  p-value: 0.0006457
```

Interpretation:

- *Intercept (β_0):* The estimated intercept is 63.5456. This means that when the weight is zero, the predicted value of pemax would be 63.5456. However,this intercept in the context of the problem should be approached cautiously as a weight of zero most probably is not meaningful in the given context.
- *Coefficient for Weight ((β_1):* The estimated coefficient is 1.1867, suggesting that for each additional unit increase in weight, we can expect *'pemax'* to increase by approximately 1.187 units.
- *Statistical Significance:* Both the intercept and the *'weight'* variable have p-values that are less than 0.05, indicating they are statistically significant predictors.
- *Residual Standard Error:* The value is 26.38, providing an estimate of the standard deviation of the residual errors. A smaller residual standard error would indicate a better fit of the model to the data.
- *R-squared Value:* The R-squared value is 0.4035, indicating that approximately 40.35% of the variability in *'pemax'* can be explained by *'weight'*.
- *Adjusted R-squared Value:* The value is 0.3776, which is a more accurate measure of the goodness-of-fit when we are comparing multiple models (not applicable here as we only have one predictor).
- *F-statistic and its p-value:* The F-statistic is 15.56 with a p-value of approximately 0.000646, which is significant at the 0.05 level.

```r
# Diagnostic plots
par(mfrow = c(3, 2))
plot(model1, which = 1:6)
```

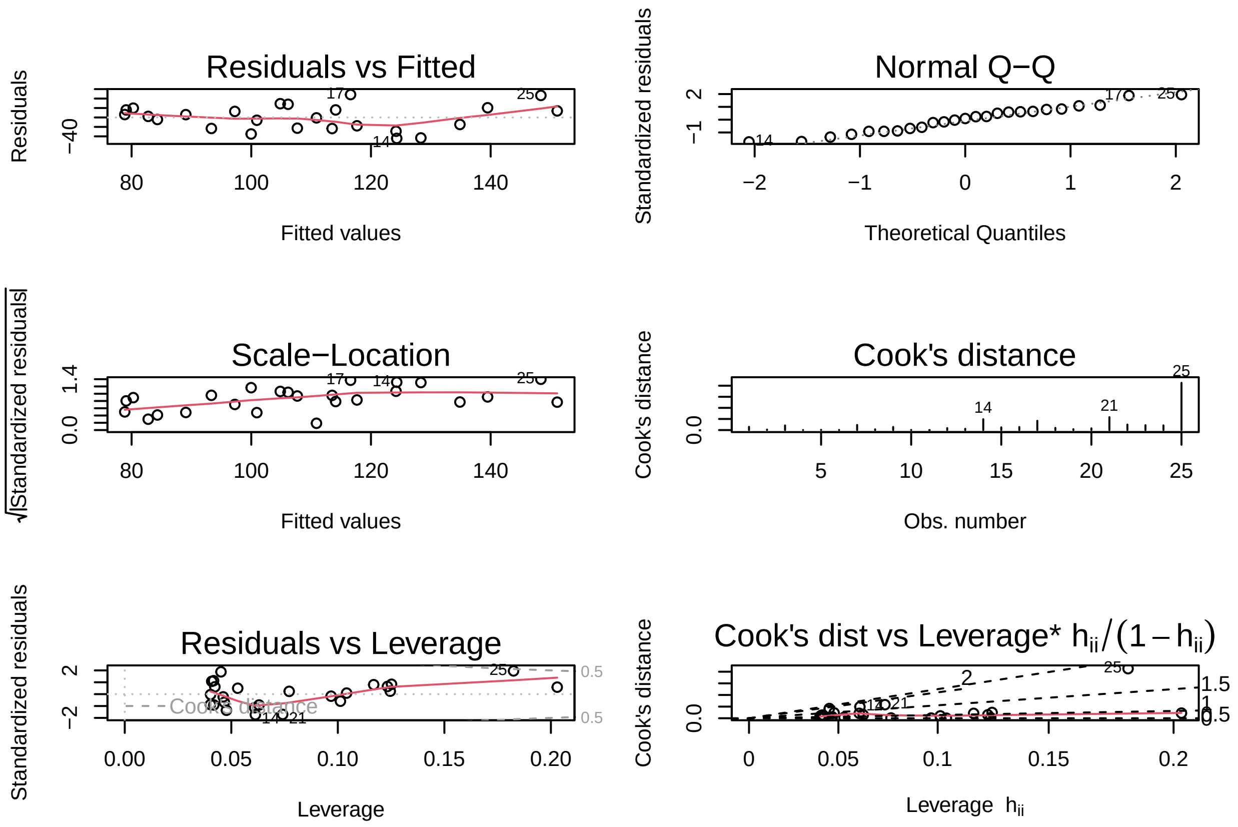


Figure 1: Regression Diagnostics Plots

Diagnostic Checking:

- **Residuals vs Fitted**: The red line in the plot is not perfectly straight and fluctuates around zero; However, they look spreading around zero randomly so I can assume linearity.
- **Normal Q-Q Plot**: The points largely follow a straight line, which suggests that the residuals are approximately normally distributed.
- **Scale-Location**: The spread of residuals does not appear to be constant across all levels of the fitted values, indicating potential heteroscedasticity.
- **Cook's Distance**: There are no points that have excessively large Cook's distance values, indicating that there are no highly influential points.
- **Residuals vs Leverage**: There is no alarming pattern, and the leverage values are within a reasonable range, which means that there are no highly influential outliers in the data.
- **Cook's Distance vs Leverage**: No points have both high leverage and high Cook's distance,except observation 25.Removing that observation may improve the model.

Part C: Proposing and analysing two more models to introduce *"sex"* variable.

Model 2: The second model considers both *'weight'* and *"sex"* as predictors for *'pemax'*.

$$\text{pemax} = \beta_0 + \beta_1 \times \text{weight} + \beta_2 \times \text{sex} + \varepsilon$$

Model 3 (with interaction term): In this model, we consider the interaction between *'weight'* and *'sex'*.

$$\text{pemax} = \beta_0 + \beta_1 \times \text{weight} + \beta_2 \times \text{weight} \times \text{sex} + \varepsilon$$

```
# Model 2
model2 <- lm(pemax ~ weight + sex, data = cystfibr)
summary(model2)
```

```
##
## Call:
## lm(formula = pemax ~ weight + sex, data = cystfibr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.388 -16.850   0.073  13.168  43.748
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  70.9719    14.4644   4.907 6.61e-05 ***
## weight        1.1248     0.3056   3.681  0.00131 **
## sex         -11.4776    10.7963  -1.063  0.29926
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.31 on 22 degrees of freedom
## Multiple R-squared:  0.4327, Adjusted R-squared:  0.3811
## F-statistic: 8.388 on 2 and 22 DF,  p-value: 0.00196
```

```
# Model 3
model3 <- lm(pemax ~ weight + weight * sex, data = cystfibr)
# weight * sex will include main effects and interaction
summary(model3)
```

```
##
## Call:
## lm(formula = pemax ~ weight + weight * sex, data = cystfibr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.464 -14.565  -2.096  14.247  42.973
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  61.3603    15.9335   3.851 0.000927 ***
## weight        1.3572     0.3471   3.910 0.000805 ***
## sex          22.0905    27.2923   0.809 0.427358
## weight:sex   -0.9240     0.6922  -1.335 0.196187
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.85 on 21 degrees of freedom
## Multiple R-squared:  0.477,  Adjusted R-squared:  0.4023
## F-statistic: 6.385 on 3 and 21 DF,  p-value: 0.003025
```

```
# To compare models based on AIC
AIC(model1,model2,model3)
```

```
##        df      AIC
## model1  3 238.4922
## model2  4 239.2398
## model3  5 239.2035
```

Analysis:

- Model 1: This model only includes *'weight'* as a predictor for *'pemax'*. The R^2 value is 0.4035, which means that approximately 40.35% of the variability in *'pemax'* is explained by *'weight'* The AIC value is 238.4922.
- Model 2: This model includes *'weight'* and *'sex'* as predictors for *'pemax'*. The R^2 value is 0.4327, so approximately 43.27% of the variability in *'pemax'* is explained by these two predictors. The AIC value is 239.2398.
- Model 3: This model includes *'weight'*, *'sex'*, and the interaction term between *weight* and *sex* as predictors for *pemax*. The R^2 value is 0.477, meaning approximately 47.7% of the variability in *pemax* is explained by these predictors. The AIC value is 239.2035.

The R^2 value increases from *Model 1* to *Model 3*, indicating that each subsequent model accounts for more variability in the data. However, when looking at the AIC values, I can see that Model 1 has the smallest AIC, suggesting a better fit to the data when considering the number of predictors. The AIC for Model 2 and Model 3 are very close, yet the third model has an additional term (interaction term), making it more complex without a significant decrease in AIC. Therefore,I would choose Model 1. Although Model 3 has a slightly higher R^2, Model 1 provides a relatively good fit with a lower AIC and fewer predictors. It's a simpler model that still captures a significant amount of the variance in the data.

Part D: Constructing a Statistical Model for the Response Variable *'pemax'*

Model Fitting and Selection Process:

- Since the independent variable *'pemax'* is right-skewed, I initially applied log transformation.
- I categorized *'bmp'* into two groups (High for values over 80 and Low for *values* 80 and below) due it is distribution.

```
# Categorize bmp into 'High' and 'Low'
cystfibr$bmp_cat <- ifelse(cystfibr$bmp > 85, 'High', 'Low')
```

- Following, based on the distribution of the variables, a logarithmic transformation was initially applied to *pemax*, *rv*, and *frc*. But then considering the model's overall significance and The R^2, I reverse these transformations.
- I then form simple regression model for each variables to see if they are significant and had the following results:

1. *weight* p-value: 0.00131 (significant) Multiple R-squared: 0.3679 Conclusion: Should be included in the multivariate model.
2. *fev1* p-value: 0.0374 (significant) Multiple R-squared: 0.1751 Conclusion: Should be included in the multivariate model.
3. *rv_log* p-value: 0.1103 (meets the <0.2 criteria) Multiple R-squared: 0.1071 Conclusion: Should be considered for inclusion in the multivariate model.
4. *frc_log* p-value: 0.0244 (significant) Multiple R-squared: 0.2015 Conclusion: Should be included in the multivariate model.
5. *age* p-value: 0.00185 (significant) Multiple R-squared: 0.3498 Conclusion: Should be included in the multivariate model.
6. *sex* p-value: 0.185 (meets the <0.2 criteria) Multiple R-squared: 0.07507 Conclusion: Should be considered for inclusion in the multivariate model.
7. *bmp* (categorized) p-value: 0.5528 (not significant and >0.2) Multiple R-squared: 0.01553 Conclusion: It will be included anyway as it is one of the main explanatory variable of interest.

- I then created a multivariate model as below:

```
# Fit the multivariate linear regression model
multivar_model <- lm(pemax ~ weight + fev1 + rv + frc + age + as.factor(sex) + as.factor(bmp_cat), data

# Display the summary statistics of the model
summary(multivar_model)
```

Creating a multivariate model:

```
##
## Call:
## lm(formula = pemax ~ weight + fev1 + rv + frc + age + as.factor(sex) +
##     as.factor(bmp_cat), data = cystfibr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.244  -9.906   1.472  15.022  39.199
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             1.1554    85.8629   0.013   0.9894
## weight                  3.0353     1.4445   2.101   0.0508 .
## fev1                    0.8233     0.9483   0.868   0.3974
## rv                      0.2023     0.1780   1.137   0.2715
## frc                    -0.2604     0.4272  -0.610   0.5502
## age                    -4.6940     4.7226  -0.994   0.3342
## as.factor(sex)1        -7.6344    13.9249  -0.548   0.5906
## as.factor(bmp_cat)Low  41.0297    20.6127   1.991   0.0629 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.05 on 17 degrees of freedom
## Multiple R-squared:  0.6337, Adjusted R-squared:  0.4829
## F-statistic: 4.201 on 7 and 17 DF,  p-value: 0.007358
```

Overall Model Fit

- Multiple R-squared: 0.5914 - This suggests that approximately 59.14% of the variance in the dependent variable pemax_log is explained by the independent variables in the model.
- Adjusted R-squared: 0.4232 - This is the adjusted R-squared which takes into account the number of predictors in the model. It's useful for comparing models with different numbers of predictors.
- F-statistic: 3.515 on 7 and 17 degrees of freedom, p-value: 0.01614 - This indicates that the model as a whole is statistically significant at a 5% significance level (p < 0.05).

Individual Coefficients Significant at 5% level: *Weight*: The coefficient estimate is 0.027554 , and it's significant (p-value = 0.0471). This means that for each unit increase in *weight*, *pemax* is expected to increase by 0.030945, holding all other variables constant. *bmp_cat*: The coefficient estimate is 0.389244 , and it's significant (p-value = 0.0491). Since *bmp* is categorized, this coefficient represents the difference in the mean value of *pemax* between the low and high categories of *bmp*, holding all other variables constant.

Not Significant at 5% level: *fev1*, *rv*, *frc*, *age*, *sex*: These variables have p-values greater than 0.05, indicating that they are not statistically significant in the presence of the other variables.

- I then simplify the multivariate model as below:

```
simplified_model <- lm(pemax ~ weight + as.factor(bmp_cat), data = cystfibr)
summary(simplified_model)
```

Simplifing the model and revise the log trasformation of *'pemax'*

```
##
## Call:
## lm(formula = pemax ~ weight + as.factor(bmp_cat), data = cystfibr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.938 -13.600   3.652  18.365  46.351
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)            34.329     19.672   1.745 0.094927 .
## weight                  1.599      0.360   4.441 0.000206 ***
## as.factor(bmp_cat)Low  23.906     12.720   1.879 0.073494 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.04 on 22 degrees of freedom
## Multiple R-squared:  0.486,  Adjusted R-squared:  0.4393
## F-statistic:  10.4 on 2 and 22 DF,  p-value: 0.0006612
```

Interpretation:

- Adjusted R-squared: The adjusted R-squared of 0.4144 is lower than the adjusted R-squared of 0.4232 from the previous full model. This indicates that the simplified model explains slightly less variability in the dependent variable *(pemax)* compared to the full model.

- Significance of Predictors: weight remains statistically significant at a very high level (p = 0.000302), while *as.factor(bmp_cat)* is only marginally significant (p = 0.060778), suggesting that the *bmp_cat* variable's impact on *pemax* is close to significant, but not below the usual threshold of 0.05.
- Overall Model Significance: The F-statistic is 9.492 with a p-value of 0.001066, indicating that the model is statistically much more significant.

```r
# Diagnostic plots for model5
par(mfrow = c(3, 2))
plot(simplified_model, which = 1:6)
```

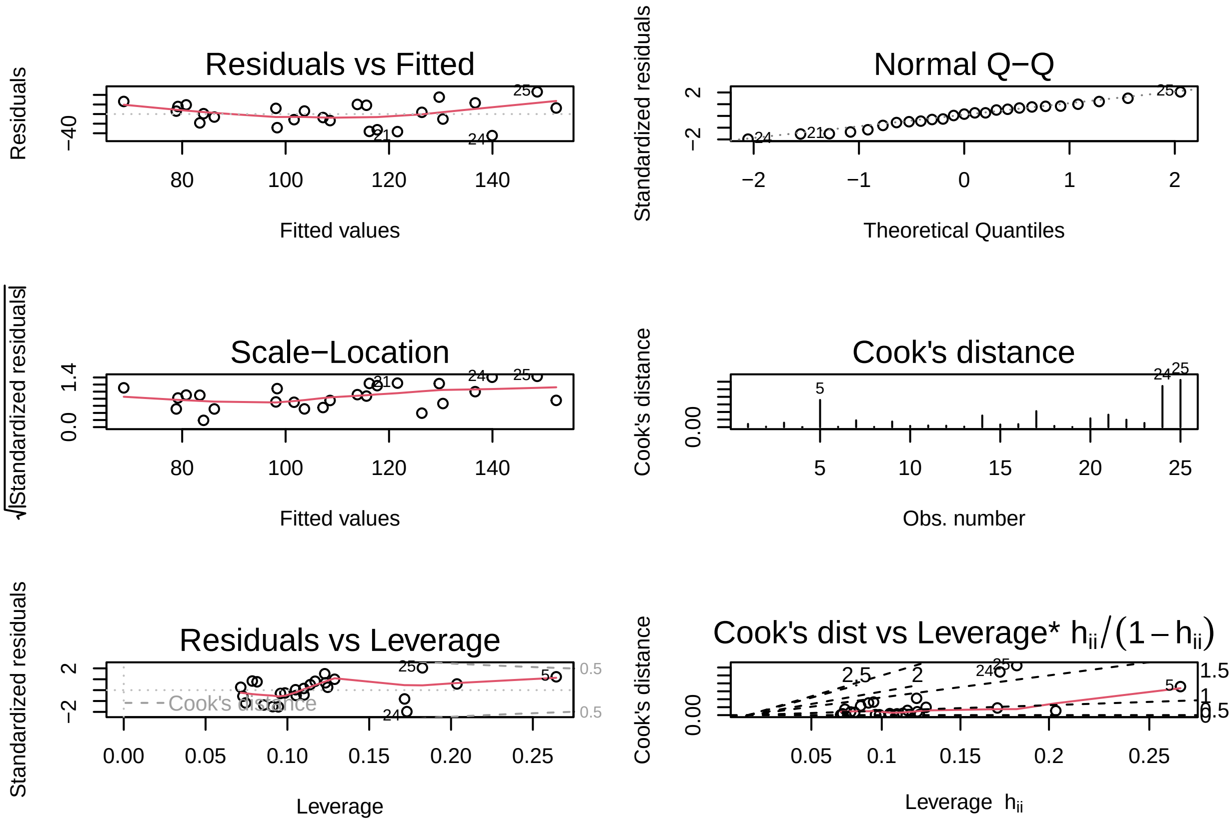


Figure 2: Regression Diagnostics Plots

Model Diagnostics interpretation:

- Conducted diagnostic plots to check assumptions of linearity, independence, homoscedasticity, and normality of residuals. The model was found to be reasonably well-fitted to the assumptions after removal of log-transformations for certain variables.

- **Residuals vs Fitted**: The red line is going up and down around zero which is good but the residuals form a funnel shape, I may have a slight heteroscedasticity here, which could violate my assumption of linear regression.

- **Normal Q-Q Plot**: The points reasonably follow a straight line which suggests that the residuals are approximately normally distributed.

- **Scale-Location**: The variance of the residuals is not constant, which may be a violation of the assumption of homoskedasticity.It is not easy to determine that with this graph; therefore, further test may be conducted.
- **Cook's Distance**: The Cook's distance values for all points are within an acceptable range, suggesting that none of the points are highly influential.
- **Residuals vs Leverage**: In this case, there are no outliers, so I can be confident that the regression model is not being affected by any single point.Although, I would be careful of observation 24,25 and 5.
- **Cook's Distance vs Leverage**: Observation 24,25 and 5 are the points both far away from the rest of the data and has a significant impact on the fit of the regression model.Removing this observation may improve the model by making it more accurate and robust. However, we have only 25 observation, thus every observation is valuable.

Final Model After removing non-significant variables and changing transformations, the model that provided the better fit and more meaningful results included *weight* and *bmp_cat* as predictors.There probably a much better model than this; However, in the scope of this assigment, this model will provide sufficient resolution. The final selected model is:

$$\text{pemax} = 34.329 + 1.599 \times \text{weight} + 23.906 \times \text{bmp\_cat (if low)}$$

Interpreting the model parameters. **Intercept (34.329)**

The intercept of 34.329 represents the estimated value of the *pemax* response variable when all other predictor variables are zero. In this context, it means that when the weight is zero and the *bmp_cat* is not 'Low,' the predicted *pemax* value would be 34.329. This is more of a theoretical baseline rather than a practical interpretation, as having zero weight isn't realistic.

Weight (1.599)

The coefficient for *weight* is 1.599, which suggests that for each unit increase in weight, the *pemax* value is expected to increase by approximately 1.599 units, keeping all other variables constant. This suggests a positive association between *weight* and *pemax.*

bmp_cat (23.906)

The coefficient for *bmp_cat* being 'Low' is 23.906. This means that, all else being equal, having a 'Low' *bmp_cat* is associated with an increase in *pemax* by 23.906 units compared to not being in the 'Low' category. Given that *bmp_cat* is one of your main variables of interest, this could be a significant finding, although it's worth noting that the p-value is slightly above the common alpha level of 0.05 (p = 0.073494), making it marginally significant.

Question 2

The general form for an exponential family distribution is

$$f(x;\theta) = h(x)\exp\left(\eta(\theta)T(x) - A(\theta)\right)$$

where $h(x)$ is called the base measure. $\eta(\theta)$ is the natural parameter. $T(x)$ is the sufficient statistic.$A(\theta)$ is the log partition function.

a. Showing that the Inverse Gaussian distribution is a member of the exponential family.

The probability density function $f(x;\mu,\gamma)$ for the Inverse Gaussian distribution is:

$$f(x;\mu,\gamma) = \sqrt{\frac{\gamma}{2\pi x^3}}\exp\left(-\frac{\gamma(x-\mu)^2}{2\mu^2 x}\right)$$

I need to rewrite this PDF in the form of the exponential family $h(x)\exp(\eta T(x) - A(\eta))$.

First,let's focus on simplifying the term inside the exponential function $-\frac{\gamma(x-\mu)^2}{2\mu^2 x}$:

Insert this into $-\frac{\gamma(x-\mu)^2}{2\mu^2 x}$:

$$-\frac{\gamma(x^2 - 2\mu x + \mu^2)}{2\mu^2 x}$$

Distribute the terms in the numerator:

$$-\frac{\gamma x^2}{2\mu^2 x} + \frac{\gamma\mu x}{\mu^2 x} - \frac{\gamma\mu^2}{2\mu^2 x}$$

Simplify the terms:

$$-\frac{\gamma x}{2\mu^2} + \frac{\gamma}{\mu} - \frac{\gamma}{2x}$$

Now I insert this expanded term back into the original PDF:

$$f(x;\mu,\gamma) = \sqrt{\frac{\gamma}{2\pi x^3}}\exp\left(-\frac{\gamma x}{2\mu^2} + \frac{\gamma}{\mu} - \frac{\gamma}{2x}\right)$$

I separate the terms inside the exponential function into two groups—those dependent on x and those independent of x:

$$f(x;\mu,\gamma) = \sqrt{\frac{\gamma}{2\pi}}x^{-\frac{3}{2}}\exp\left(-\frac{\gamma}{2x}\right)\exp\left(-\frac{\gamma x}{2\mu^2} + \frac{\gamma}{\mu}\right)$$

Here, $h(x) = \sqrt{\frac{\gamma}{2\pi}}x^{-\frac{3}{2}}\exp\left(-\frac{\gamma}{2x}\right)$, $\eta = -\frac{\gamma}{2\mu^2}$, $T(x) = x$, and $A(\eta) = -\frac{\gamma}{\mu} = 2\mu\eta$.

Thus, the Inverse Gaussian distribution is a member of the exponential family.

b. Giving the natural parameter and the scale parameter.

Natural Parameter: $\eta = -\frac{\gamma}{2\mu^2}$

Scale Parameter: γ

c.Hence, deriving the mean and variance of Inverse Gaussian distribution

The mean μ and variance σ^2 of the Inverse Gaussian distribution are:

$$\mu = \mu$$

$$\sigma^2 = \frac{\mu^3}{\gamma}$$

Yet here, in the exponential family, the mean $\mathbb{E}[T(X)]$ and variance $\mathrm{Var}[T(X)]$ can be expressed in terms of the log partition function $A(\eta)$ and the natural parameter η as:

$$\text{Mean: } \frac{d}{d\eta}A(\eta)$$

$$\text{Variance: } \frac{d^2}{d\eta^2}A(\eta)$$

For our Inverse Gaussian distribution, $\eta = -\frac{\gamma}{2\mu^2}$ and $A(\eta) = -\frac{\gamma}{\mu} = 2\mu^2\eta$.

Mean First, let's find the mean $\mathbb{E}[T(X)]$:

$$\frac{d}{d\eta}A(\eta) = \frac{d}{d\eta}(2\mu^2\eta) = 2\mu^2$$

Since $T(x) = x$, I get $\mu = \mathbb{E}[T(X)] = 2\mu^2$, which appears to be inconsistent with the standard interpretation of the Inverse Gaussian mean.

Variance For the variance, I have:

$$\frac{d^2}{d\eta^2}A(\eta) = \frac{d^2}{d\eta^2}(2\mu^2\eta) = 0$$

This result also doesn't provide a meaningful interpretation for the variance of the Inverse Gaussian distribution.

Given these inconsistencies, it appears that using the exponential family form for deriving the mean and variance of the Inverse Gaussian distribution may not be directly applicable or may require additional assumptions. Typically, these parameters are known to be μ for the mean and $\frac{\mu^3}{\gamma}$ for the variance, usually derived through direct integration methods.

Question 3

a. Equivalence of Maximum Likelihood and Least Squares Estimates under Normally Distributed Noise

The linear model is: $Y_i = x_i^\top \beta + \epsilon_i$ for $i = 1, \ldots, n$, where $\epsilon_i \sim N(0, \sigma^2)$ i.i.d.

The likelihood function $L(Y|X, \beta, \sigma^2)$ is given as:

$$L(Y|X, \beta, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - x_i^\top \beta)^2}{2\sigma^2}\right)$$

Taking the natural logarithm, I get the log-likelihood $\log L(Y|X, \beta, \sigma^2)$:

$$\log L(Y|X, \beta, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i^\top \beta)^2$$

Differentiating the log-likelihood with respect to β and setting it to zero, I find that the maximum likelihood estimator for β is also the least squares estimator. The term I need to minimize is:

$$-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i^\top \beta)^2$$

Minimizing this term with respect to β is equivalent to minimizing the least squares error, which is defined by:

$$LS(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^\top \beta)^2$$

b. Maximizing Log-Likelihood Under Laplace Distributed Noise and Its Equivalence to Minimizing Absolute Error Loss

The noise ϵ_i follows a Laplace distribution $L(0, \sigma^2)$. The PDF of a Laplace-distributed random variable z with mean μ and variance $2\sigma^2$ is:

$$f(z) = \frac{1}{2\sigma}\exp\left(-\frac{|z - \mu|}{\sigma}\right)$$

The likelihood of a single observation y_i given x_i and β is:

$$f(y_i|x_i, \beta, \sigma^2) = \frac{1}{2\sigma}\exp\left(-\frac{|y_i - x_i^\top \beta|}{\sigma}\right)$$

The likelihood of the entire sample is:

$$L(Y|X, \beta, \sigma^2) = \prod_{i=1}^{n}\frac{1}{2\sigma}\exp\left(-\frac{|y_i - x_i^\top \beta|}{\sigma}\right)$$

Taking the logarithm, I get the log-likelihood $\log L(Y|X, \beta, \sigma^2)$:

$$\log L(Y|X,\beta,\sigma^2) = -n\log(2\sigma) - \frac{1}{\sigma}\sum_{i=1}^{n}|y_i - x_i^\top\beta|$$

To maximize this log-likelihood, I minimize:

$$\frac{1}{\sigma}\sum_{i=1}^{n}|y_i - x_i^\top\beta|$$

This is equivalent to minimizing the absolute error loss $AL(\beta)$ defined as:

$$AL(\beta) = \frac{1}{n}\sum_{i=1}^{n}|y_i - x_i^\top\beta|$$

c. Comparative Plot of PDFs for Normal and Laplace Distributions with Matching Variances

The variance for a standard normal distribution, $\mathcal{N}(0,1)$ is $\sigma^2 = 1$.

The variance for a Laplace distribution, $\mathcal{L}(0,b)$ is $\sigma^2 = 2b^2$.

To match the variance of the two distributions, I will set the variances equal to each other:

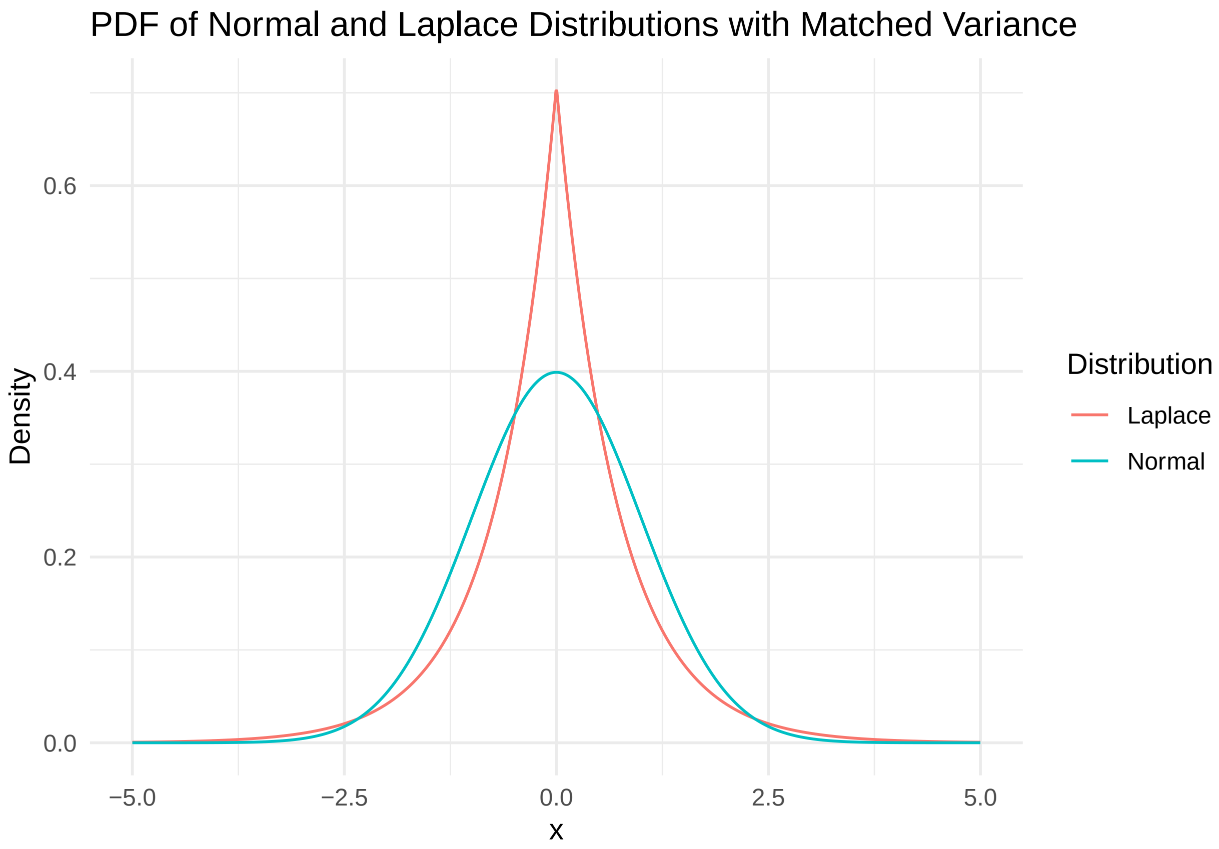
$$\sigma^2_{\text{Normal}} = \sigma^2_{\text{Laplace}}$$

$$1 = 2b^2$$

Solving for b, I find:

$$b = \sqrt{\frac{1}{2}}$$

```
# Defining the Normal and Laplace PDF functions
normal_pdf <- function(x) {
  return(dnorm(x, mean = 0, sd = 1))
}

laplace_pdf <- function(x) {
  b <- sqrt(1 / 2)
  return(0.5 / b * exp(-abs(x) / b))
}

# Generating data points
x_values <- seq(-5, 5, length.out = 1000)

# Calculating PDF values
normal_values <- normal_pdf(x_values)
laplace_values <- laplace_pdf(x_values)

# Creating dataframe for plotting
df <- data.frame(x = c(x_values, x_values),
                 y = c(normal_values, laplace_values),
                 Distribution = c(rep("Normal", length(x_values)), rep("Laplace", length(x_values))))

# Generating the plot
ggplot(df, aes(x = x, y = y, color = Distribution)) +
  geom_line() +
  theme_minimal() +
  ggtitle("PDF of Normal and Laplace Distributions with Matched Variance") +
  xlab("x") +
  ylab("Density")
```



d. Robustness of Linear Model with Laplace Error Noise to Outliers: Insights from PDF Comparison

The Laplace error model is generally more tolerant of outliers compared to the Normal error model. There are several key reasons that explain this:

- *Heavier Tails:* If you look at the plot we generated, you'll see that the Laplace distribution has "heavier" tails compared to the Normal distribution. What this means is that it accounts for extreme values more effectively, making it a better fit when outliers are present.
- *Error Sensitivity:* The Laplace model aims to minimize the absolute error, which is a lot less sensitive to outliers than minimizing the squared error, which is the case in the Normal model. The squared error can dramatically inflate the impact of outliers due to the squaring operation.
- *Focus on Median:* Unlike the Normal distribution, which is all about the mean, the Laplace distribution centers around the median. In statistics, the median is often a better measure of central tendency when outliers are involved, as it is less sensitive to extreme values.
- *Statistical Robustness:* In non-technical terms, the Laplace model is just built better for "real-world" data that may not always follow textbook distributions. It's less likely to produce misleading results when outliers are present.

By both looking at the plot and understanding these points, it becomes clear why the Laplace error model is considered more robust to outliers. This is a useful property, especially in business analytics where data is often messy and outliers are common.