Homework 5

## Description

The Auto dataset from the ISLR2 package contains information on various automobile models from the 1970s and 1980s, providing a useful context for exploring relationships between vehicle characteristics and fuel efficiency. It includes **392 observations** on **nine variables**, such as mpg (miles per gallon), horsepower, weight, acceleration, displacement, cylinders, and year. These variables are a mix of quantitative and categorical data, with the name column identifying each car model. The dataset is particularly valuable for regression analysis due to its real-world relevance and the presence of nonlinear relationships, multicollinearity, and opportunities for transformation—making it ideal for studying how predictor variables influence fuel efficiency.

library(here) ## File Path Management  
library(ISLR2) ## Data Extraction  
library(dplyr) ## Data Transformation  
library(tidyr) ## Data Transformation  
library(ggplot2) ## Data Visualization  
library(broom) ## Data Analysis  
source(here("R","assessment\_regression.R"))

**Question:** *Fit a linear regression model to predict mpg using horsepower. How well does the model fit the data, and what does the residual plot suggest about the relationship between the two variables?*

## Summary Statistics

auto\_df <- Auto %>% drop\_na()  
  
# Summary statistics: mean, sd, min, max  
summary\_stats <- auto\_df %>%  
 select(mpg, horsepower) %>%  
 summarise(across(everything(), list(  
 mean = mean,  
 sd = sd,  
 min = min,  
 max = max  
 ), .names = "{.col}\_{.fn}")) %>%  
 pivot\_longer(  
 everything(),  
 names\_to = c("variable", "statistic"),  
 names\_sep = "\_",  
 values\_to = "value"  
 )   
  
summary\_stats

# A tibble: 8 × 3  
 variable statistic value  
 <chr> <chr> <dbl>  
1 mpg mean 23.4   
2 mpg sd 7.81  
3 mpg min 9   
4 mpg max 46.6   
5 horsepower mean 104.   
6 horsepower sd 38.5   
7 horsepower min 46   
8 horsepower max 230

The summary statistics show that `mpg` ranges from 9.0 to 46.6 with a mean of 23.4, indicating a wide spread in fuel efficiency among cars. `Horsepower` has a mean of 104 with a standard deviation of 38.5, ranging from 46 to 230, suggesting substantial variation in engine strength across the dataset.

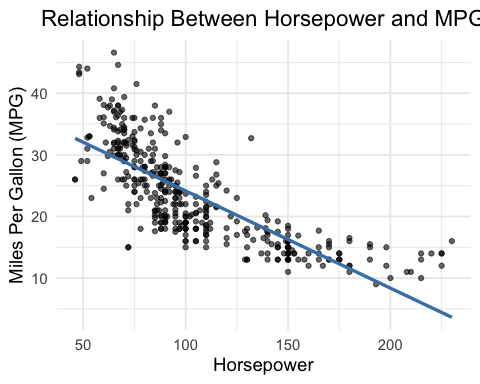
# Compute correlation  
cor\_val <- auto\_df %>%  
 summarise(correlation = cor(mpg, horsepower))  
  
cor\_val

correlation  
1 -0.7784268

The correlation between `mpg` and `horsepower` is approximately -0.78, indicating a strong negative linear relationship: as horsepower increases, fuel efficiency tends to decrease.

## Visualization

ggplot(auto\_df, aes(x = horsepower, y = mpg)) +  
 geom\_point(alpha = 0.6) +  
 geom\_smooth(method = "lm", se = FALSE, color = "steelblue", linewidth = 1.2) +  
 labs(  
 title = "Relationship Between Horsepower and MPG",  
 x = "Horsepower",  
 y = "Miles Per Gallon (MPG)"  
 ) +  
 theme\_minimal(base\_size = 14) +  
 theme(plot.title = element\_text(hjust = 0.5))



While the scatterplot initially suggests that transforming the response variable `mpg` might help, the pattern more strongly supports transforming the predictor `horsepower` to better linearize the relationship. This approach helps achieve more constant variance and a better-fitting linear model.

## Analysis

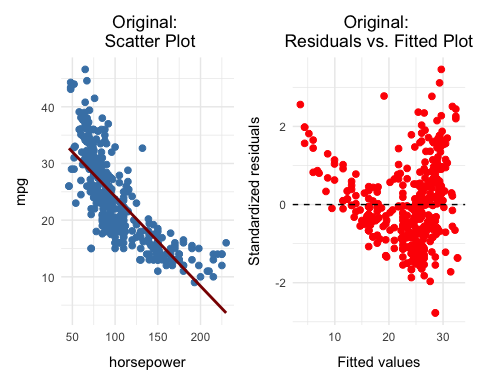
### Regular Model

model\_0 <- lm(mpg ~ horsepower, data = auto\_df)  
summary(model\_0)

Call:  
lm(formula = mpg ~ horsepower, data = auto\_df)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-13.5710 -3.2592 -0.3435 2.7630 16.9240   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 39.935861 0.717499 55.66 <2e-16 \*\*\*  
horsepower -0.157845 0.006446 -24.49 <2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 4.906 on 390 degrees of freedom  
Multiple R-squared: 0.6059, Adjusted R-squared: 0.6049   
F-statistic: 599.7 on 1 and 390 DF, p-value: < 2.2e-16

The intercept of approximately 39.94 suggests that a car with 0 horsepower is expected to achieve 39.94 mpg—though not realistic, this sets the baseline for interpretation. The slope of -0.158 indicates that, on average, each additional unit of horsepower is associated with a decrease of 0.158 mpg in fuel efficiency.

plot\_scatter\_resid(model\_0,auto\_df ,"Original")



The scatterplot reveals a nonlinear pattern between `horsepower` and `mpg`, with diminishing drops in mpg at higher horsepower. The residual plot shows a curved pattern, indicating nonlinearity and non-constant variance—violating linear model assumptions.

### Transformations

mod\_1\_auto\_df <- auto\_df %>%  
 mutate(  
 log\_hp = log(horsepower),  
 sqrt\_hp = sqrt(horsepower)  
 )

We selected `log(horsepower)` and `sqrt(horsepower)` as ad hoc transformations based on the curvature observed in the residual plots. Although a formal power transformation like Box-Tidwell could have been applied to the predictor, we opted not to pursue it, as it is beyond the course scope.

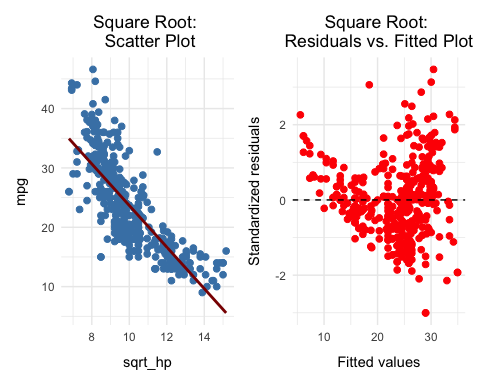
#### Square Root of hp

# Linear model with sqrt(horsepower)  
model\_2 <- lm(mpg ~ sqrt\_hp, data = mod\_1\_auto\_df)  
summary(model\_2)

Call:  
lm(formula = mpg ~ sqrt\_hp, data = mod\_1\_auto\_df)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-13.9768 -3.2239 -0.2252 2.6881 16.1411   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 58.705 1.349 43.52 <2e-16 \*\*\*  
sqrt\_hp -3.503 0.132 -26.54 <2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 4.665 on 390 degrees of freedom  
Multiple R-squared: 0.6437, Adjusted R-squared: 0.6428   
F-statistic: 704.6 on 1 and 390 DF, p-value: < 2.2e-16

With the square root transformation applied to `horsepower`, the intercept of 58.71 represents the expected `mpg` when `sqrt(horsepower)` is zero. The slope of -3.50 means that for each unit increase in the square root of horsepower, the expected fuel efficiency decreases by 3.50 mpg, on average.

plot\_scatter\_resid(model\_2,mod\_1\_auto\_df ,"Square Root")



After the square root transformation, the scatterplot shows a more linear trend and the residual plot reveals reduced curvature. This suggests a better fit compared to the original model, although some mild heteroscedasticity remains.

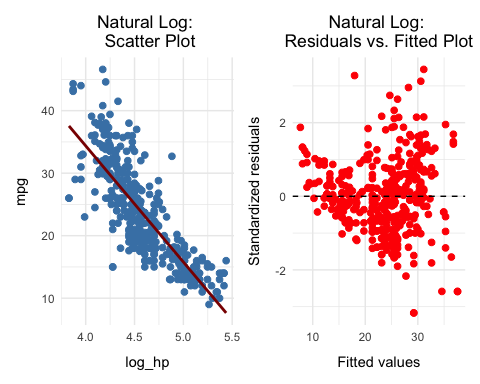
#### Natural Log of hp

# Linear model with log10(horsepower)  
model\_1 <- lm(mpg ~ log\_hp, data = mod\_1\_auto\_df)  
summary(model\_1)

Call:  
lm(formula = mpg ~ log\_hp, data = mod\_1\_auto\_df)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-14.2299 -2.7818 -0.2322 2.6661 15.4695   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 108.6997 3.0496 35.64 <2e-16 \*\*\*  
log\_hp -18.5822 0.6629 -28.03 <2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 4.501 on 390 degrees of freedom  
Multiple R-squared: 0.6683, Adjusted R-squared: 0.6675   
F-statistic: 785.9 on 1 and 390 DF, p-value: < 2.2e-16

Applying the natural log to `horsepower`, the intercept of 108.70 indicates the expected `mpg` when `log(horsepower)` is zero. The slope of -18.58 implies that, on average, a one-unit increase in `log(horsepower)` corresponds to a decrease of 18.58 mpg in fuel efficiency.

plot\_scatter\_resid(model\_1,mod\_1\_auto\_df ,"Natural Log")



After the square root transformation, the scatterplot shows a more linear trend and the residual plot reveals reduced curvature. This suggests a better fit compared to the original model, although some mild heteroscedasticity remains.

### Interpretation of Results

summary\_model\_0 <- summarize\_reg\_model(model\_0,'original')  
summary\_model\_1 <- summarize\_reg\_model(model\_1,'log\_hp')  
summary\_model\_2 <- summarize\_reg\_model(model\_2,'sqrt\_hp')  
  
  
bind\_rows(  
 summary\_model\_0,  
 summary\_model\_1,  
 summary\_model\_2  
)

type RSS RSE R2 Adj\_R2 AIC BIC  
1 original 9385.92 4.91 0.61 0.60 1246.88 1250.85  
2 log\_hp 7899.93 4.50 0.67 0.67 1179.31 1183.28  
3 sqrt\_hp 8486.62 4.66 0.64 0.64 1207.39 1211.37

Among the three models, the log-transformed model (`log\_hp`) has the lowest residual standard error (4.50), highest R² (0.67), and lowest AIC/BIC values, indicating it provides the best overall fit. The transformation successfully linearizes the relationship and reduces model error, making it the preferred choice.

### Box-Cox Transformation

A Box-Cox transformation should not be used here because it is only applicable to the **response variable** (y), and horsepower is the **predictor**. Applying Box-Cox to x violates its assumptions and intended use.