

Stat632 HW 3

Brandon Keck

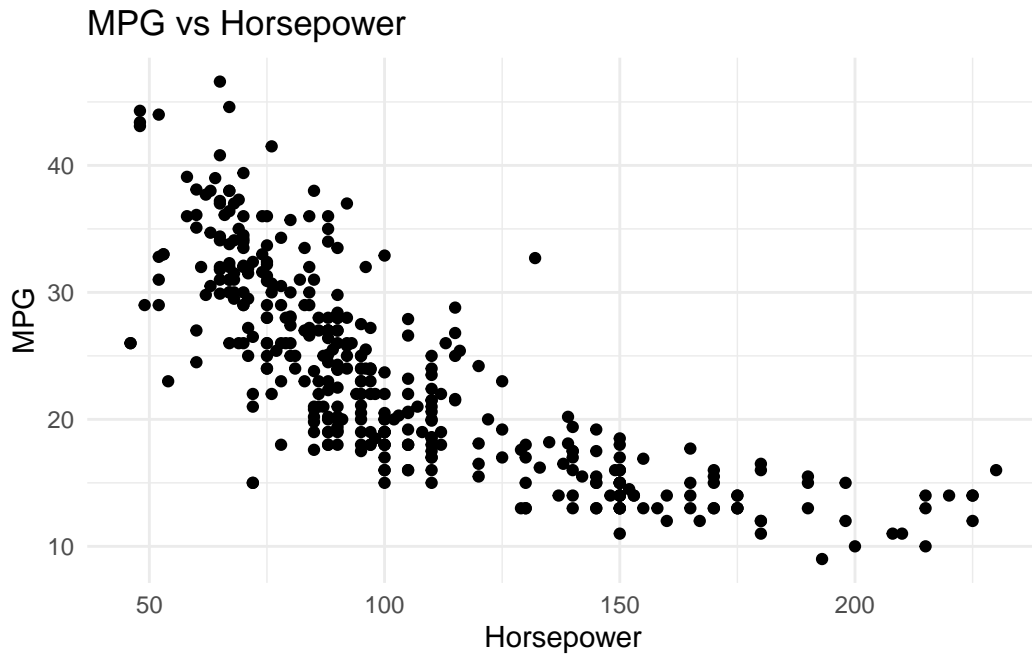
```
# Load the necessary libraries
library(ISLR) # ISLR package to access the Auto dataset
library(ggplot2)
library(dplyr)
```

```
data(Auto)
# head(Auto)
```

Exercise 1.

(a)

```
ggplot(Auto, aes(horsepower, mpg)) +
  geom_point() +
  labs(title = "MPG vs Horsepower",
       x = "Horsepower",
       y = "MPG") +
  theme_minimal()
```



(b)

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

Where:

Y = mpg

X = horsepower

```
lm_auto <- lm(mpg ~ horsepower + I(horsepower^2), data = Auto)
summary(lm_auto)
```

Call:

```
lm(formula = mpg ~ horsepower + I(horsepower^2), data = Auto)
```

Residuals:

Min	1Q	Median	3Q	Max
-14.7135	-2.5943	-0.0859	2.2868	15.8961

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	56.9000997	1.8004268	31.60	<2e-16 ***
horsepower	-0.4661896	0.0311246	-14.98	<2e-16 ***
I(horsepower^2)	0.0012305	0.0001221	10.08	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.374 on 389 degrees of freedom
Multiple R-squared: 0.6876, Adjusted R-squared: 0.686
F-statistic: 428 on 2 and 389 DF, p-value: < 2.2e-16

$$\hat{mpg} = 56.90 - 0.4662(horsepower) + 0.00123(horsepower^2)$$

Each additional 1-unit increase in horsepower decreases mpg by 0.466 mpg on average. All predictors have p-values <2e-16 meaning they are highly significant.

(c)

$$\hat{mpg} = 56.90 - 0.4662(horsepower) + 0.00123(horsepower^2)$$

```
56.9000997 - 0.4661896 * (150) + 0.0012305 * (150^2)
```

```
[1] 14.65791
```

Here I am making a prediction by hand. It looks like if a vehicle was to have a horsepower of 150 we should see mpg be about 14.66. Now let's double check with the prediction function in R to see how close we are.

```
pred <- predict(lm_auto, data.frame(horsepower = 150),
               interval = "prediction")
pred
```

```
      fit      lwr      upr
1 14.65872  6.027273 23.29016
```

From the regression model

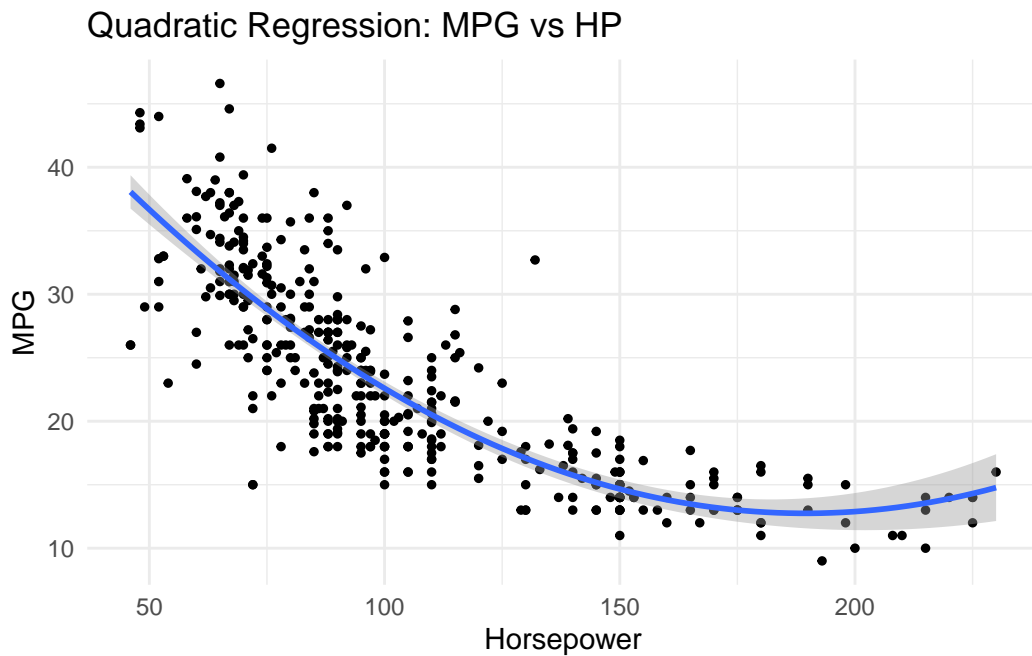
$$\hat{mpg} = 56.90 - 0.4662(horsepower) + 0.00123(horsepower^2)$$

we predict that a vehicle with 150 horsepower will have an estimated mpg 14.66. Additionally, using a 95% prediction interval, we estimate that the actual mpg could fall between 6.03 and 23.29, accounting for variability in individual observations.

(d)

Next we plot the estimated regression curve on the scatterplot.

```
ggplot(data = Auto, aes(horsepower, mpg)) +  
  geom_point(size = 1) +  
  stat_smooth(method = 'lm', formula = y ~ poly(x, 2), se = TRUE) +  
  labs(title = "Quadratic Regression: MPG vs HP",  
        x = "Horsepower",  
        y = "MPG") +  
  theme_minimal()
```



(e)

Plot of the residuals vs fitted values, and a QQ plot of the Standardized Residuals

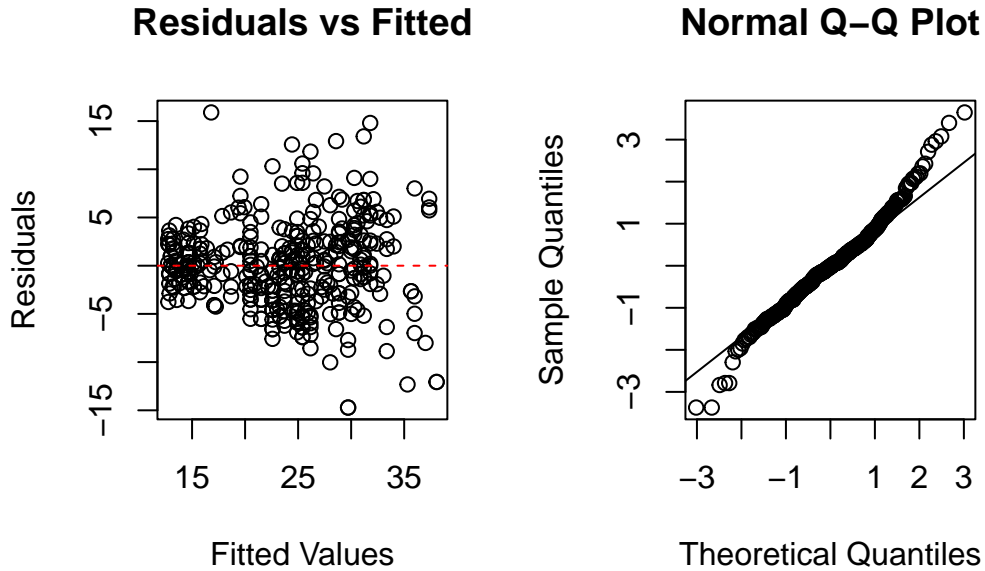
```
# residuals vs fitted values  
par(mfrow = c(1, 2))  
plot(predict(lm_auto), resid(lm_auto),  
      main = "Residuals vs Fitted",  
      xlab = "Fitted Values",
```

```

        ylab = "Residuals")
abline(h = 0, col = "red", lty = 2)

# Q-Q plot
qqnorm(rstandard(lm_auto))
qqline(rstandard(lm_auto))

```



The Residuals vs Fitted plot suggests that variance might not be perfectly equal across all fitted values. Observations 334, 323, and 155 stand out as potentially influential points. However, there is no extreme funnel shape, meaning heteroscedasticity is not a major issue.

The Normal Q-Q plot shows that most residuals follow the theoretical normal distribution, indicating approximate normality. However, there is some deviation in the tails, suggesting potential outliers or slight skewness.

Overall, while there are minor concerns with variance and influential points, the model does not show any severe violations of assumptions.

Exercise 2:

```
data("Carseats")
# head(Carseats)
```

(a)

```
lm_seats <- lm(Sales ~ Price + Urban + US, data = Carseats)
summary(lm_seats)
```

Call:

```
lm(formula = Sales ~ Price + Urban + US, data = Carseats)
```

Residuals:

Min	1Q	Median	3Q	Max
-6.9206	-1.6220	-0.0564	1.5786	7.0581

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	13.043469	0.651012	20.036	< 2e-16 ***
Price	-0.054459	0.005242	-10.389	< 2e-16 ***
UrbanYes	-0.021916	0.271650	-0.081	0.936
USYes	1.200573	0.259042	4.635	4.86e-06 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.472 on 396 degrees of freedom

Multiple R-squared: 0.2393, Adjusted R-squared: 0.2335

F-statistic: 41.52 on 3 and 396 DF, p-value: < 2.2e-16

$$\hat{Sales} = 13.04 - 0.544(Price) - 0.0219(Urban) + 1.2006(US)$$

For each \$1 increase in Price decreases Sales by 0.0544 cents.

The coefficient for Urban is not significant. It has a high p-value of 0.936 which is greater than our significance value of $\alpha = 0.05$. Meaning there is no evidence that being in an urban area affects Sales.

(b)

$$\hat{Sales} = 13.04 - 0.544(Price) - 0.0219(Urban) + 1.2006(US)$$

1. Intercept ($\beta_0 = 13.04$)

- When Price = 0, Urban = No, and Us = No, the predicted Sales = 13.03
- Not practically meaningful because a price of 0 is unrealistic.

2. Price ($\beta_1 = 0.0544$)

- For each \$1 increase in Price, Sales decreases by 0.0544 cents on average.

3. Urban ($\beta_2 = -0.0219$)

- If a store is in an Urban area, Sales are lower by 0.0219 cents compared to a non-urban area.

4. US ($\beta_3 = 1.2006$)

- If a store is in the US, Sales are higher by 1.20 units compared to a store outside the US.

(c)

$$\hat{Sales} = 13.04 - 0.0544(Price) - 0.0219(Urban) + 1.2006(US)$$

(d)

$$H_0 : \beta_j = 0$$

From the summary of the multiple linear regression model we can reject the H_0 for Price and US. The output shows that these predictors have p-values very close to 0, which means that the β coefficients are significantly different from 0. From the model so far only Price and US are useful predictors of Sales. Urban has a p-value of 0.936 and we cannot reject H_0 meaning it does not significantly affect Sales and can be removed from the model.

(e)

```
lm_seats2 <- lm(Sales ~ Price + US, data = Carseats)
summary(lm_seats2)
```

Call:

```
lm(formula = Sales ~ Price + US, data = Carseats)
```

Residuals:

Min	1Q	Median	3Q	Max
-6.9269	-1.6286	-0.0574	1.5766	7.0515

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	13.03079	0.63098	20.652	< 2e-16 ***
Price	-0.05448	0.00523	-10.416	< 2e-16 ***
USYes	1.19964	0.25846	4.641	4.71e-06 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.469 on 397 degrees of freedom

Multiple R-squared: 0.2393, Adjusted R-squared: 0.2354

F-statistic: 62.43 on 2 and 397 DF, p-value: < 2.2e-16

After removing the predictor variable Urban from the model we see a slightly improved Adjusted R-squared from 0.2335 to 0.2354 confirming that the Urban predictor was unnecessary. The residual standard error also slightly decreased, suggesting a better fit model.

(f)

Comparing the models from part (a) and part (e) we find that the reduced model in part (e) provides a slightly better fit. The Adjusted R-squared increased slightly, indicating that the reduced model is more efficient without sacrificing explanatory power. Additionally, the residual standard error decreased, suggesting a modestly improved fit. However, the most compelling evidence for the superiority of the reduced model is the higher F-statistic, which indicated a stronger overall model.

(g)

```
confint(lm_seats2)
```

	2.5 %	97.5 %
(Intercept)	11.79032020	14.27126531
Price	-0.06475984	-0.04419543
USYes	0.69151957	1.70776632

We calculated a 95% confidence interval for the coefficients in the reduced model. The results confirm that both Price and US are statistically significant predictors of Sales as their confidence intervals do not include 0.

Price: The interval (-0.0648, -0.0442) suggests that for every \$1 increase in Price, Sales decreases by approximately 0.044 to 0.065 units.

US: The interval (0.6915, 1.7078) indicates that being in the US increases Sales by approximately 0.69 to 1.71 units on average.