1

(a)

Inflexible is best because simple models have enough data to learn the relationship with a large sample size. Also, inflexible methods can overfit with few predictors.

(b)

Flexible is best because it can capture complex relationships with higher dimensional data. Inflexible models are likely to underfit the data given the small amount of data.

(c)

Flexible is best because it can capture non-linear relationships. Inflexible is too simple and will most likely underfit our data if it's non-linear.

(d)

Inflexible is best because it generalizes noisy data with high variance. Flexible methods will likely overfit the data.

2

(a)

i. False, provided that the GPA is not high, college graduates can generally earn more than high school graduates on average, as their base salary is \$35k higher.
ii. False, if the graduate's GPA is 4.0, our model predicts high school graduates will earn more than college graduates as the negative coefficient of the interaction effect between GPA and level outweighs the base salary boost of college graduates.
iii. True, when the GPA is high enough, the negative coefficient of the interaction effect between GPA and level outweighs the base salary boost for college graduates, resulting in high school graduates earning more on average.
iv. False, when the GPA is high enough, the negative coefficient of the interaction effect between GPA and level outweighs the base salary boost for college graduates, resulting in college graduates earning less on average.

(b)

The model is as follows:

$$\hat{Y} = 50 + 20X_1 + 0.07X_2 + 35X_3 + 0.01X_1X_2 - 10X_1X_3$$

Given our inputs:

$$\hat{Y} = 50 + 20 * 4.0 + 0.07 * 110 + 35 + 0.01 * 4.0 * 110 - 10 * 4.0$$

$$\hat{Y} = 137.1$$

(c)

False, although the coefficient for the interaction effect appears small, we cannot claim an interaction effect without conducting a statistical significance test.

3

The RSS formula is as follows:

$$RSS = \sum (Y_i - \hat{Y}_i)^2$$

We are assuming that the data follows a polynomial relationship for all parts.

TODO: idk if (a) is correct ??

a. Generally, we would expect the training RSS of our cubic regression to be lower than our linear regression. Due to the randomness of ϵ, the cubic regression will likely overfit and capture random noise as it is more flexible and has more polynomial terms. When n is very large or the data is less noisy (ϵ has very low variance), overfitting is less of an issue, so we would expect the training RSS of our cubic regression to be similar to if not slightly lower than our linear regression. With less overfitting, we expect β_2 and β_3 to be close to 0 since the true relationship is linear, resulting in similar predictions between our cubic regression and linear regression.
b. We would expect the testing RSS of our cubic regression to be higher than our linear regression. The cubic regression likely overfitted the training data, resulting in lower performance on testing data.
c. If the true relationship is close to linear, then apply the answer from part (a). Otherwise, we would expect the training RSS of our cubic regression to be lower than our linear regression, because cubic regression is more flexible and has more polynomial terms.
d. If the true relationship is close to linear, then apply the answer from part (b). However, if the true relationship is far from linear, then then we would expect test RSS of our cubic regression to be lower than our linear regression, as our linear regression likely underfitted the model.

4

Let:

- $S_{YY} = \sum Y_i^2$ (Total Sum of Squares)
- $S_{\text{res}} = \sum (Y_i - \hat{Y}_i)^2$ (Residual Sum of Squares)
- $S_{\text{reg}} = S_{YY} - S_{\text{res}}$ (Regression Sum of Squares)

Given:

$$R^2 = \frac{\text{Explained Variance}}{\text{Total Variance}} = \frac{S_{YY} - S_{\text{res}}}{S_{YY}}$$

Then:

$$R^2 = 1 - \frac{S_{\text{res}}}{S_{YY}}.$$

The simple linear regression equation is:

$$\hat{Y} = \beta_1 X$$

where β_1 is estimated as:

$$\hat{\beta}_1 = \frac{\sum X_i Y_i}{\sum X_i^2} = \frac{S_{XY}}{S_{XX}}.$$

Then the fitted values are:

$$\hat{Y}_i = \hat{\beta}_1 X_i = \frac{S_{XY}}{S_{XX}} X_i.$$

The total sum of squares is:

$$S_{YY} = \sum Y_i^2.$$

The regression sum of squares is:

$$S_{\text{reg}} = \sum (\hat{Y}_i)^2 = \sum \left(\frac{S_{XY}}{S_{XX}} X_i \right)^2.$$

Since:

$$S_{\text{reg}} = \frac{S_{XY}^2}{S_{XX}^2} \sum X_i^2 = \frac{S_{XY}^2}{S_{XX}},$$

we substitute into R^2:

$$R^2 = \frac{S_{\text{reg}}}{S_{YY}} = \frac{S_{XY}^2}{S_{XX} S_{YY}}.$$

By definition, the Pearson correlation coefficient r is:

$$r = \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}}.$$

Squaring both sides:

$$r^2 = \frac{S_{XY}^2}{S_{XX} S_{YY}}.$$

This is exactly the expression for R^2, so:

$$R^2 = r^2.$$

TODO: ?? Verify the proof is correct and figure out the tricks used.

5

TODO: ??

```r
# Load required packages
library(ISLR)
library(ggplot2)

# Load the Auto dataset
data("Auto")
```

(a)

```
# (a) Fit simple linear regression: mpg ~ horsepower
lm_fit <- lm(mpg ~ horsepower, data = Auto)
lm_summary <- summary(lm_fit)
lm_summary
```

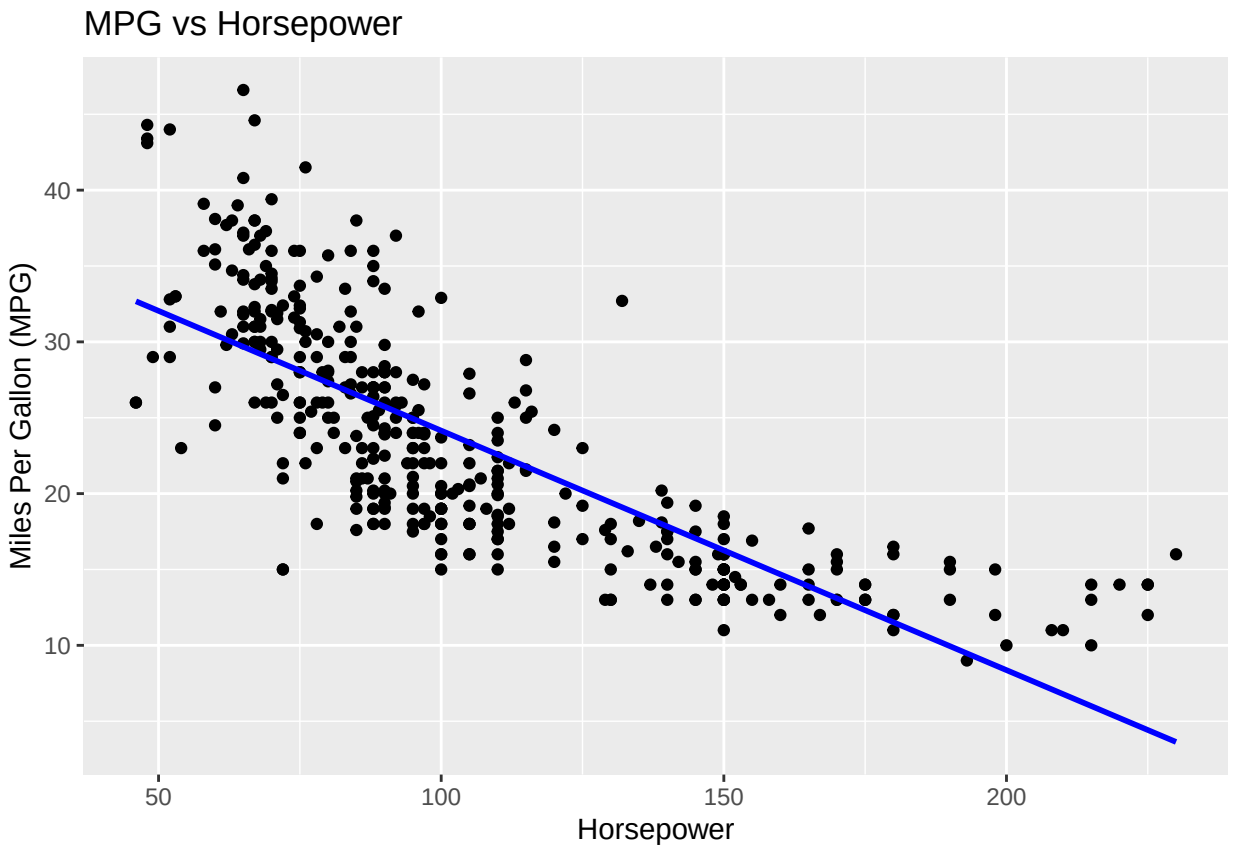
```
##
## Call:
## lm(formula = mpg ~ horsepower, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.5710  -3.2592  -0.3435   2.7630  16.9240
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 39.935861   0.717499   55.66   <2e-16 ***
## 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
```

i. Yes, because $p << 0.001$, indicating there is a significant relationship between horsepower and mpg.
ii. The R^2 value is around 0.6059 which means that 60.59% of variability in mpg is explained by horsepower.
iii. The relationship is negative since the coefficient for horsepower is $-0.1578 < 0$.
iv. The prediction for 98hp is 24.4671mpg.

(b)

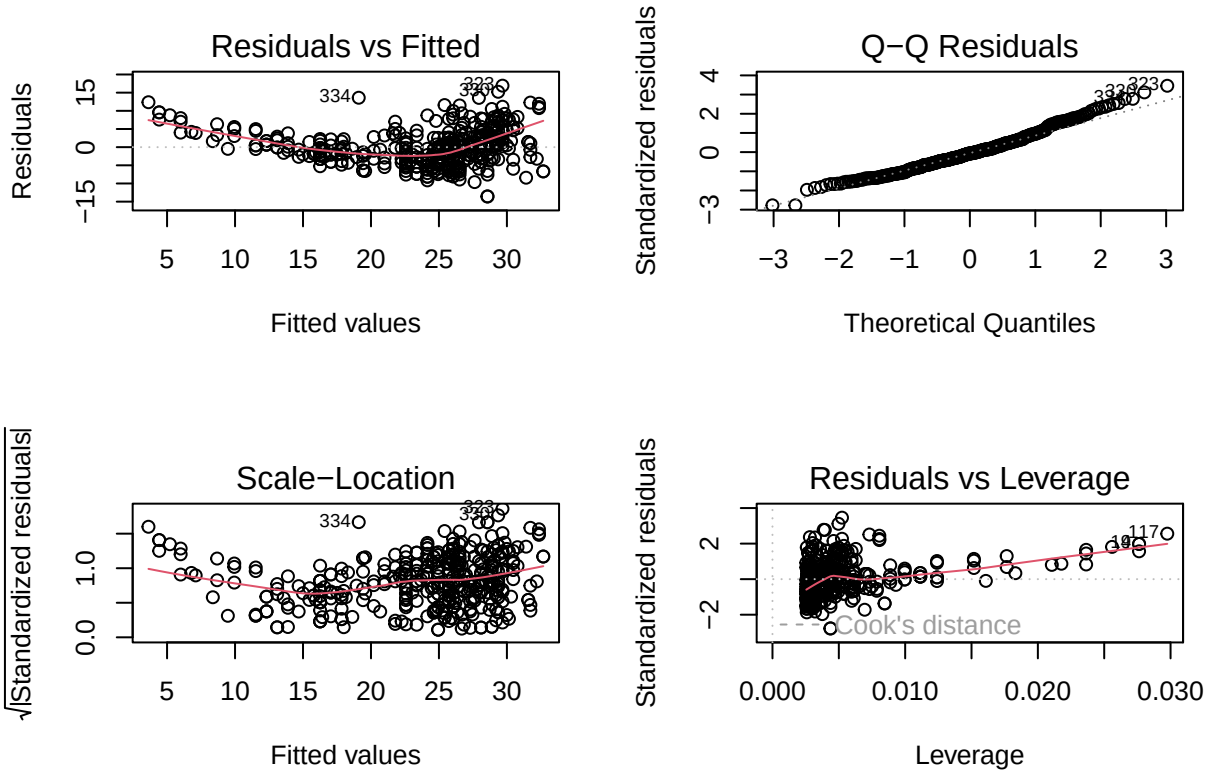
```
# (b) Scatterplot with regression line
ggplot(Auto, aes(x = horsepower, y = mpg)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "MPG vs Horsepower",
       x = "Horsepower",
       y = "Miles Per Gallon (MPG)")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



(c)

```
# (c) Diagnostic plots
par(mfrow = c(2, 2))  # Show all 4 plots in one window
plot(lm_fit)
```



i. The residuals vs fitted shows that the errors are not white noise. The constant variance assumption is violated as there seems to be more residual variance for higher fitted values.

ii. TODO: ??