Simple Linear Regression Model Inference

Module 3

DATA 5600
Introduction to Regression and Machine Learning for Analytics
Marc Dotson

Module Overview

- Confidence Interval for the Slope
- Hypothesis Test for the Slope
- Confidence Interval for the Mean
- Prediction Interval for Individual Observations
- Model Evaluation Metrics

Overview

- We can use a linear regression model for inference only *after* checking assumptions
- We are interested in inference for $\hat{\beta}_1$
- Recall:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \text{where} \quad \epsilon_i \sim N(0, \sigma^2)$$

$$\hat{\beta}_1 = \frac{\mathrm{SS}_{XY}}{\mathrm{SS}_{XX}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$s^2 = MSE = \frac{\text{SSE}}{\text{degrees of freedom for error}}$$

$$= \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n} e^2}{n-2}$$

Overview

- Mathematically, we can show the estimated slope $\hat{\beta}_1$ from our least-squares regression fit is approximately normally distributed (under common circumstances – assuming that the sample size is sufficiently large)
- The standard error of $\hat{\beta}_1$ is given by

$$\text{s.e.}\left(\hat{\beta}_1\right) = \frac{s}{\sqrt{SS_{XX}}} = \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

- Since we need to estimate the unknown model variance σ^2 with the MSE s^2, we use a: t distribution for inference, with $n - 2$ degrees of freedom.

Car Gas Mileage

- In a real analysis, we would use the log-transformed response model, but for the sake of this module, we will use the original non-transformed data for illustration.
- Recall for the Car Gas Mileage data set:

$$\hat{\beta}_1 = -0.0098$$

$$s^2 = 22.31$$

- The standard error of $\hat{\beta}_1$ is

$$\text{s.e.}(\hat{\beta}_1) = \frac{\sqrt{22.31}}{\sqrt{(3436 - 2535)^2 + (3433 - 2535)^2 + \cdots + (3449 - 2535)^2}}$$
$$= 0.0005749$$

Confidence Interval for the Slope

Confidence Interval for the Slope

- A $(1-\alpha)100\%$ confidence interval for β_1 is given by

$$\hat{\beta}_1 \pm t_{\alpha/2}\text{s.e.}(\hat{\beta}_1),$$

where $t_{\alpha/2}$ represents the upper $\alpha/2$ critical value from the t distribution with $n-2$ degrees of freedom.

Car Gas Mileage

n = 289

df = 287

Alpha/2 = 0.025

- A 95% confidence interval for the slope is given by

$$\hat{\beta}_1 \pm t_{0.025}\text{s.e.}(\hat{\beta}_1)$$
$$= -0.0098 \pm (1.9683)(0.0005749)$$
$$= (-0.011, -0.009)$$

- Interpretation (including "significance"):

Hypothesis Test for the Slope

Hypothesis Test for the Slope

- To test $H_0: \beta_1 = 0$, versus $H_a: \beta_1 \neq 0$, the t statistic is given by

$$t = \frac{\hat{\beta}_1 - 0}{\text{s.e.}(\hat{\beta}_1)}$$

where t = number of standard errors your value is from the null value.

- The *p*-value is computed as $p = 2P(t > |t_{obs}|)$, where t_{obs} is the observed value of the test statistic, and the tail probability is based on a t distribution with $n - 2$ degrees of freedom.

- If we use a significance level of α, then we reject H_0 if $p < \alpha$. In other words, this indicates evidence at the α-level that there is a significant linear relationship between x and y.

p-Value

- Recall a p-value is the probability of seeing a result as or more "extreme" than what you observed, if there really is no difference (i.e., if the null hypothesis is true).
- In a hypothesis test for the slope, the p-value is the probability of having a slope as "steep" or steeper as the one our data produced, if the "truth" is that there is no linear association between x and y.
- Remember:
  - if p-value < α, we "reject" the null hypothesis in favor of the alternative hypothesis
  - if p-value > α, we "fail to reject" the null hypothesis – we do not "accept" the null hypothesis, we simply have insufficient evidence to accept the alternative hypothesis

Car Gas Mileage

- For the car gas mileage data, the test statistic for the slope is given by

$$t = \frac{\hat{\beta}_1 - 0}{\text{s.e.}(\hat{\beta}_1)} = \frac{-0.0098 - 0}{0.0005749} = -17.11 (computer)$$

- The *p*-value is given by $p = 2P(t > |-17.11|)$, based on a t distribution with 287 degrees of freedom. Using computer software, it turns out that $p = 1.04x10^{-45}$.

- Interpretation (including "significance"):

Confidence Intervals and Hypothesis Testing

- Two branches of statistical inference: confidence intervals and hypothesis testing, will *always* result in the same substantive conclusions.
- If a confidence interval does not contain the null value, then the hypothesis test will result in a "significant" p-value, and vice-versa.
- Often, a confidence interval is more useful for assessing the magnitude of an effect than a single p-value.
- Note: We could follow the exact same steps and get a confidence interval and do a hypothesis test for the intercept $\hat{\beta}_0$, but this is usually less interesting and unnecessary to the research question.


p-values
confidence intervals
Corporate needs you to find the differences between this picture and this picture.
They're the same picture.
imgflip.com

Confidence and Prediction Intervals

Recall…

- In Module 1, we used the linear regression model to predict the MPG for a car weighing 3000 lbs.

$$\begin{aligned}\hat{y}_{3000} &= \hat{\beta}_0 + \hat{\beta}_1(3000)\\ &= 51.59 - 0.0098(3000)\\ &= 22.19 \approx 22.09 \text{ (Python result)}\end{aligned}$$

- If we took another sample of cars and fit a regression line, would this prediction be the same?

- How do we incorporate sampling variability (uncertainty) into our prediction of Y?

Two Types of Intervals

- There are two types of intervals that we can use to incorporate a measure of uncertainty into our predictions.

1. Confidence intervals for the mean of Y
   - Interval for the <u>average</u> car's MPG
   - Predict the average MPG of cars weighing 3000 lbs.
   - Target is a fixed parameter ($E(Y|X = x_i^*)$)

2. Prediction intervals for individual observations
   - Interval for a new <u>single</u> car's MPG
   - Predict the MPG of a car weighing 3000 lbs.
   - Target is a random variable (y_i^*)

Confidence Intervals for the Mean of Y

Confidence Intervals for the Mean of Y

- We want to create an interval, or band, around our regression line that indicates the variability of our estimate of the line (average value of Y for a given x_i^*)

- The interval will be centered at the line $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i^*$
- The standard error for $\hat{y}_i$ is given by

$$SE_{CI}\left(\hat{\beta}_0 + \hat{\beta}_1 x_i^*\right) = s\sqrt{\frac{1}{n} + \frac{\left(x_i^* - \bar{x}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Variability depends on where you are predicting

- So, a $(1-\alpha)$-level confidence interval is

$$\left(\hat{\beta}_0 + \hat{\beta}_1 x_i^*\right) \pm t_{\alpha/2,n-2} SE_{CI}\left(\hat{\beta}_0 + \hat{\beta}_1 x_i^*\right)$$

Recall s is the standard deviation of the residuals:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}}$$

Confidence Intervals for the Mean of Y

- CI for the average MPG of cars weighing $\mathrm{x}_i^* = 3000$ lbs.

$$\hat{y}_{3000} = 22.09$$

$$SE_{CI}(\hat{y}_{3000}) = \sqrt{22.31}\sqrt{\frac{1}{289} + \frac{(3000-2535)^2}{(3436-2535)^2 + \cdots + (2720-2535)^2}}$$
$$= 0.3856$$

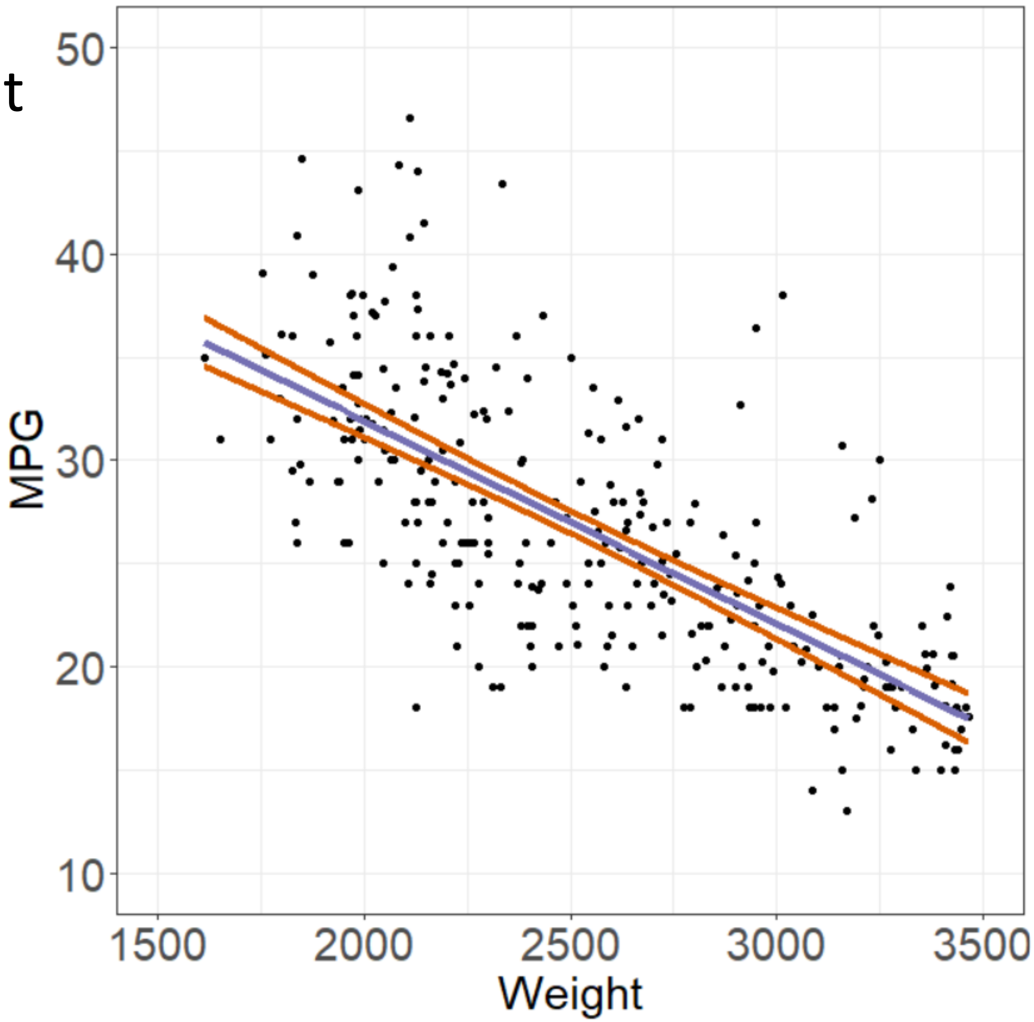
- So, a 95% confidence interval is

$$22.09 \pm t_{0.025,287} 0.3856 = (21.33, 22.85)$$

- Interpretation:

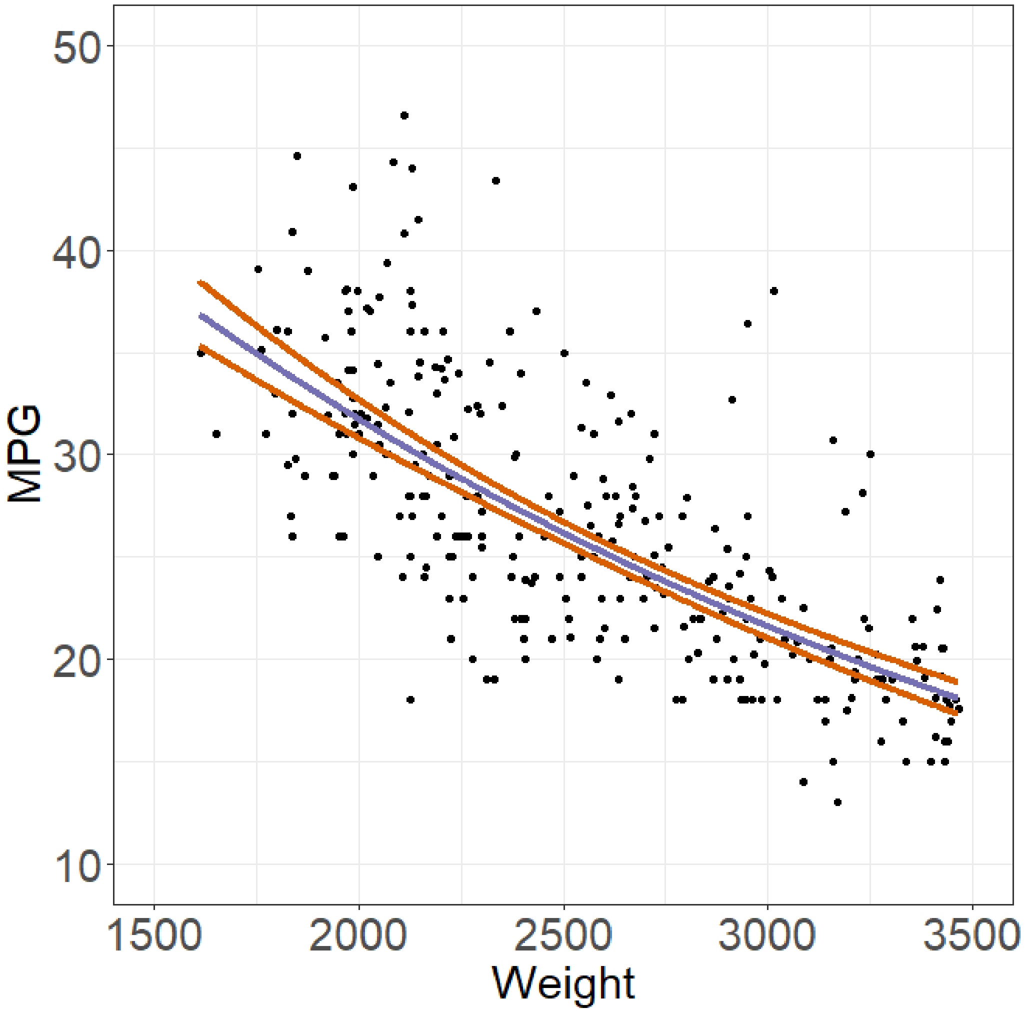
Confidence Intervals for the Mean of Y

- What do you observe about the confidence interval?



Confidence Intervals for the Mean of Y

- Using the appropriate log-transformed response model



Prediction Intervals for Individual Observations

Prediction Intervals for Individual Observations

- We just calculated a confidence interval for the mean of Y given x_i^* $\left(E(Y|X = x_i^*)\right)$

- What if instead of the mean we are now interested in predicting a future value of Y for a given x_i^*?

- Would you expect there to be more uncertainty/variability around an average or a specific predicted value?

Prediction Intervals for Individual Observations

- We want to create an interval, or band, around our regression line that indicates the variability of our estimate of the value of Y at a specific x_i^*

- The interval will be centered at the line $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i^*$
- The standard error for $\hat{y}_i$ is given by

$$SE_{PI}\left(\hat{\beta}_0 + \hat{\beta}_1 x_i^*\right) = s\sqrt{1 + \frac{1}{n} + \frac{\left(x_i^* - \bar{x}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Variability depends on where you are predicting

- So, a $(1-\alpha)$-level confidence interval is

$$\left(\hat{\beta}_0 + \hat{\beta}_1 x_i^*\right) \pm t_{\alpha/2,n-2} SE_{PI}\left(\hat{\beta}_0 + \hat{\beta}_1 x_i^*\right)$$

Recall s is the standard deviation of the residuals:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}}$$

Prediction Intervals for Individual Observations

- PI for the MPG of a car weighing $\mathrm{x}_i^* = 3000$ lbs.

$$\hat{y}_{3000} = 22.09$$

$$SE_{PI}(\hat{y}_{3000}) = \sqrt{22.31}\sqrt{1 + \frac{1}{289} + \frac{(3000-2535)^2}{(3436-2535)^2 + \cdots + (2720-2535)^2}}$$
$$= 4.7391$$

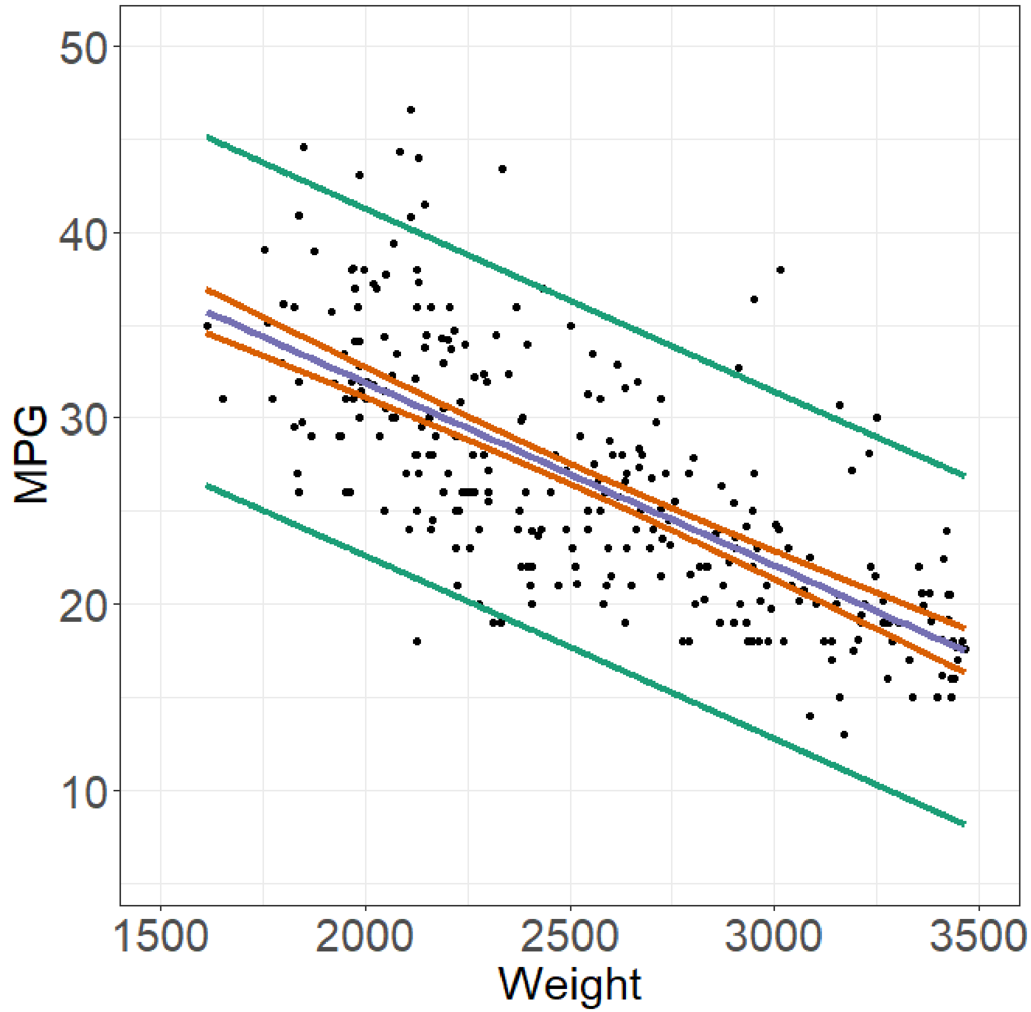
- So, a 95% confidence interval is

$$22.09 \pm t_{0.025,287} 4.7391 = (12.76, 31.41)$$

- Interpretation:

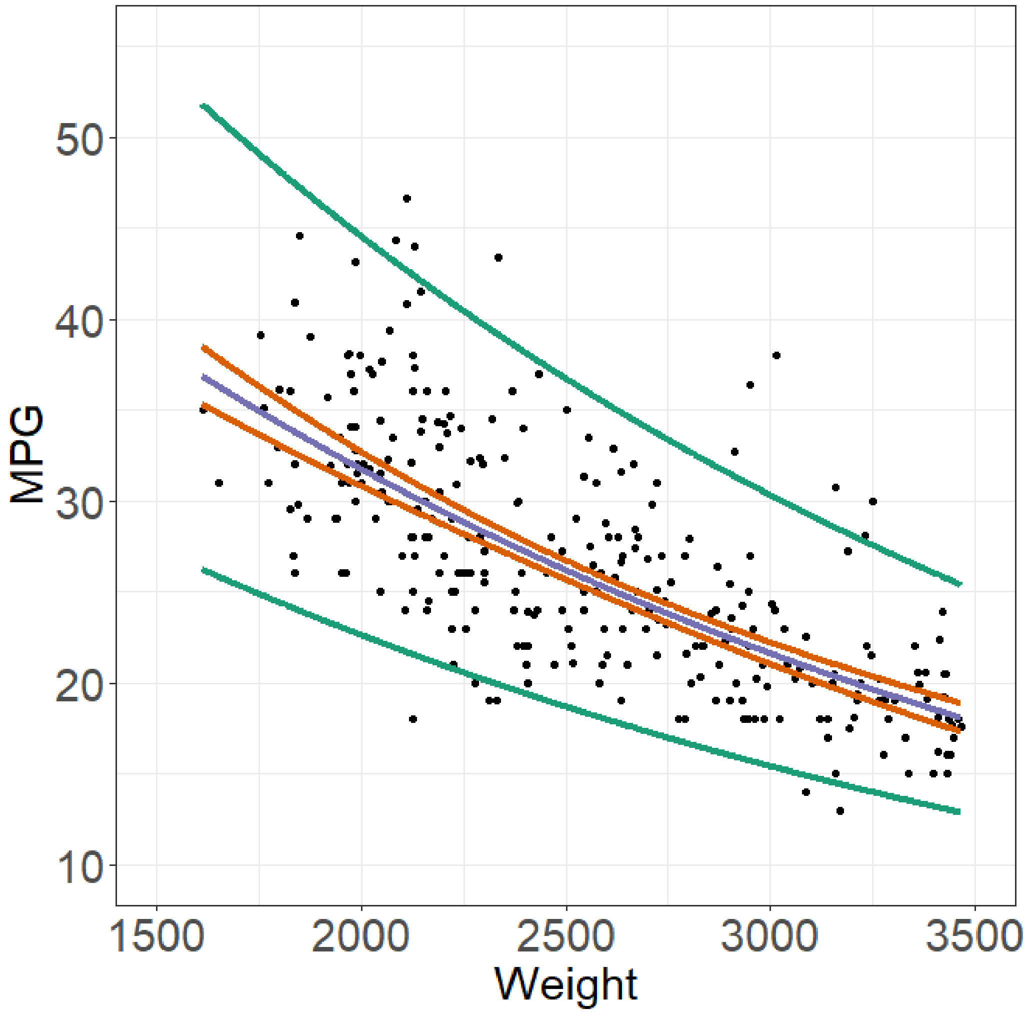
Prediction Intervals for Individual Observations

- What do you observe about the prediction interval?
- How does it compare to the confidence interval?



Prediction Intervals for Individual Observations

- Using the appropriate log-transformed response model



Model Evaluation Metrics

Inference

- Statistical inference has to do with the "usefulness" of a model: is X useful at predicting Y?

- There are other methods of assessing this (in addition to hypothesis tests, confidence intervals, and prediction intervals)

Mean Squared Error (MSE)

For the cars data set: MSE = 22.31

- $MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \hat{\sigma}^2 = s^2$
- $0 \leq MSE < \infty$ The lower the MSE, the better the model
- Our best estimate of the true error variance
- Average squared distance between the observed outcome and the predicted outcome from the model (measures the amount of spread in the residuals)
- Pros: Easy to work with mathematically
- Cons: Not very interpretable since the units are squared, highly influenced by outliers, adding more variables in model always lowers MSE

Root Mean Squared Error (RMSE)

For the cars data set: RMSE = 4.72

- $RMSE = \sqrt{MSE} = \sqrt{\hat{\sigma}^2}$
- $0 \leq RMSE < \infty$

The lower the RMSE, the better the model

- Average error performed by the model in predicting the outcome (measures the amount of spread in the residuals)
- Pros: More interpretable than the MSE since it is on the scale of the data
- Cons: Slightly less mathematically friendly, highly susceptible to outliers, adding more variables in model always lowers RMSE

Mean Absolute Error (MAE)

For the cars data set: MAE = 3.64

- $MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n-2}$
- $0 \leq MAE < \infty$

The lower the MAE, the better the model

- Average absolute difference between the outcome and the model prediction of the outcome
- Pros: Less susceptible to outliers than RMSE
- Cons: Harder to work with mathematically (not differentiable) than RMSE, and adding more variables in model always lowers MAE

Multiple R-Squared (R^2) (a.k.a. the Coefficient of Determination)

- $R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = r^2$
- $0 \leq R^2 \leq 1$
- The proportion of total variation in Y explained by the predictor(s) (X) in the model
- The higher R^2, the better the model
- Pros:
- Cons:

Adjusted R-Squared (R^2)

- $R^2_{adj} = 1 - \left[\frac{(1-R^2)(n-1)}{n-k-1}\right]$ (k is the number of predictors in the model)
- $0 \le R^2_{adj} \le 1$
- The proportion of total variation in Y explained by the predictor(s) (X) in the model, adjusted for the number of variables in the model
- $R^2_{adj} \le R^2$
- Pros:
- Cons:

F-Statistic & p-value

- $$F = \frac{\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{2-1}}{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}}$$

- The test statistic for H_0: the predictor(s) in the model have no linear association with Y

- The further the F-statistic is from 1, the more evidence to reject H_0 (the associated p-value can be used to determine if there is *significant* evidence)