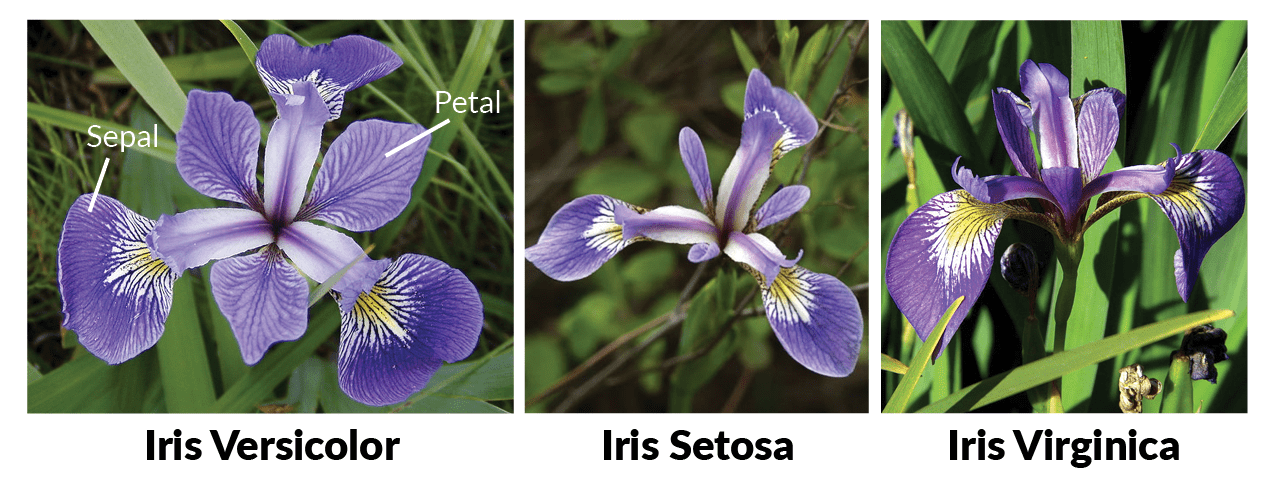
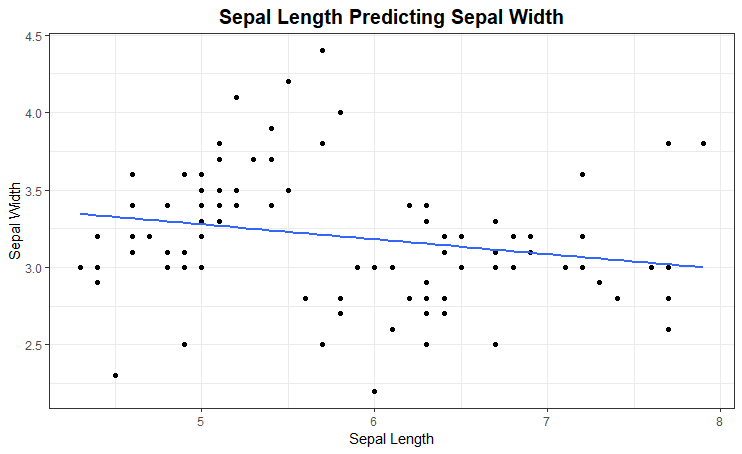
Motivation:

## Description of Iris Flower data set:

* 3 different flower species (Iris Setosa, Iris Versicolor, Iris Virginica)
* 50 measurements were collected for each flower
* Edgar Anderson collected the data to quantify the morphologic variation of Iris flowers of three related species
* All flowers were pasture and picked on the same day and measured at the same time by the same person with the same apparatus



**Simple Linear Regression**



What is the explanatory variable?

What is the response variable?

Discuss the relationship between two variables. Think about a hypothesis test for the slope parameter.

**R Output for predicting Sepal Width from Sepal Length**

Call:

lm(formula = Sepal.Width ~ Sepal.Length, data = iris\_1)

Residuals:

Min 1Q Median 3Q Max

-1.02561 -0.25890 -0.02297 0.22483 1.18968

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.75795 0.25599 14.680 <2e-16 \*\*\*

Sepal.Length -0.09608 0.04359 -2.204 0.0299 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.41 on 98 degrees of freedom

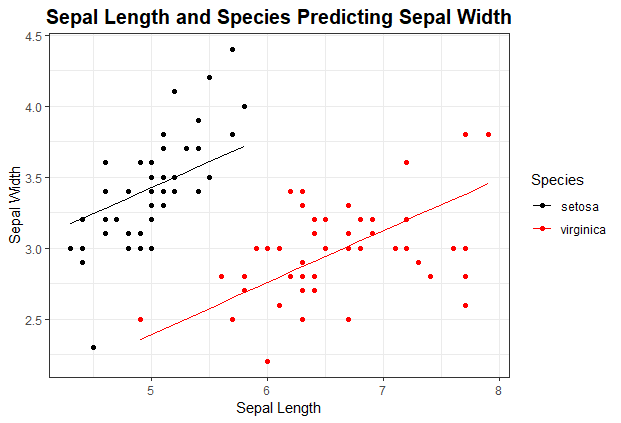
Multiple R-squared: 0.04723, Adjusted R-squared: 0.03751

F-statistic: 4.858 on 1 and 98 DF, p-value: 0.02986

Write and interpret the regression model.

What are the model metrics?

**Dummy Variables Models**



What are the explanatory variables?

What is the response variable?

Do you think both species have the same slope?

**R Output for predicting Sepal Width from Sepal Length and Species**

Call:

lm(formula = Sepal.Width ~ Sepal.Length + Species, data = iris\_1)

Residuals:

Min 1Q Median 3Q Max

-0.94327 -0.17184 0.01209 0.19702 0.71863

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.60040 0.29735 5.382 5.11e-07 \*\*\*

Sepal.Length 0.36508 0.05879 6.209 1.32e-08 \*\*\*

Speciesvirginica -1.03156 0.11060 -9.327 3.81e-15 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2992 on 97 degrees of freedom

Multiple R-squared: 0.4977, Adjusted R-squared: 0.4873

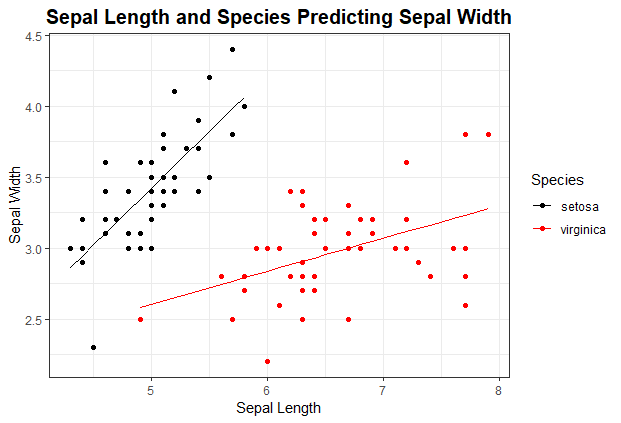
F-statistic: 48.06 on 2 and 97 DF, p-value: 3.14e-15

State and interpret the regression model.

State the model metrics.

Compare the simple linear regression model metrics to this model metrics. What do you notice?

**Interaction Models**



What are the explanatory variables?

What is the response variable?

**R Output for predicting Sepal Width from Sepal Length and Species with an interaction**

Call:

lm(formula = Sepal.Width ~ Sepal.Length \* Species, data = iris\_1)

Residuals:

Min 1Q Median 3Q Max

-0.72394 -0.18795 -0.00152 0.16416 0.56814

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.5694 0.5566 -1.023 0.30882

Sepal.Length 0.7985 0.1109 7.200 1.35e-10 \*\*\*

Speciesvirginica 2.0157 0.6894 2.924 0.00431 \*\*

Sepal.Length:Speciesvirginica -0.5666 0.1268 -4.468 2.16e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2737 on 96 degrees of freedom

Multiple R-squared: 0.5842, Adjusted R-squared: 0.5712

F-statistic: 44.96 on 3 and 96 DF, p-value: < 2.2e-16

State and interpret the regression model.

State the model metrics.

Compare the simple linear regression and Dummy Variable model metrics to this model metrics. What do you notice?

**Quadratic Models**

**Model Utility**

**Collinearity**

**Model Metrics**

**Stepwise Regression**