IE8990: Adv. Data Analytics for Complex Systems

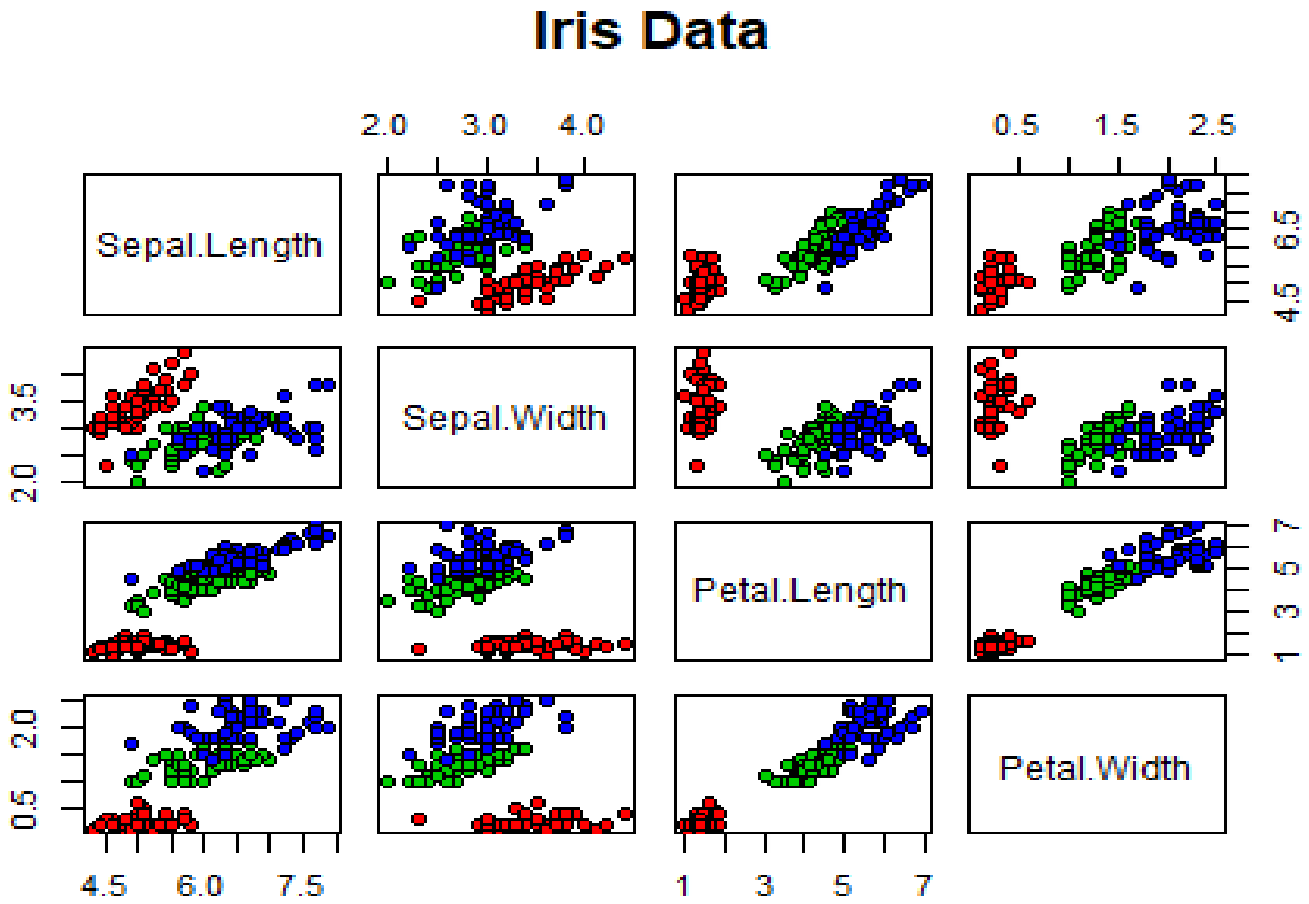
- Lab 2
  - R output and interpretation for iris example

Homework 1: Question 1 – Iris data cont.

1. Fit Model 0 using Sepal.Length as response and Sepal.Width as the only one predictor
   a. Find all the parameters, write down their interpretations
   b. Evaluate Model 0 using adj R^2 as the criteria
   c. Is Model 0 good for prediction? If no, how do you want to improve the model?
2. Based on your answer in 2c, fit a new model (Model 1) using Sepal.Length as response, Sepal.Width and Species as predictors
   a. Find all the parameters, write down their interpretations
   b. Is Model 1 a good model for prediction?
3. Can you come up with another model that have better adj R^2 than Model 1?

Data visualization

- Response variable
  - Sepal.Length
- Model 0
  - Sepal.Width
- Model 1
  - Sepal.Width + Species



Model 0

````
28
29 ### 3. Fit a linear regression models
30 #### 3.1 Fit Model 0
31 ```{r}
32 lmfit.sepal<-lm(formula=Sepal.Length ~ Sepal.Width, data=iris)
33 summary(lmfit.sepal )
34 ```
````

```
Call:
lm(formula = Sepal.Length ~ Sepal.Width, data = iris)

Residuals:
    Min      1Q  Median      3Q     Max
-1.5561 -0.6333 -0.1120  0.5579  2.2226

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.5262     0.4789   13.63   <2e-16 ***
Sepal.Width   -0.2234     0.1551   -1.44    0.152
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8251 on 148 degrees of freedom
Multiple R-squared:  0.01382,   Adjusted R-squared:  0.007159
F-statistic: 2.074 on 1 and 148 DF,  p-value: 0.1519
```

Model 1

```
#### 3.2 Fit Model 1
```{r}
lmfit.sepal2<-lm(formula=Sepal.Length ~ Sepal.Width + Species, data=iris)
summary(lmfit.sepal2)
```
```

```
Call:
lm(formula = Sepal.Length ~ Sepal.Width + Species, data = iris)

Residuals:
     Min       1Q   Median       3Q      Max
-1.30711 -0.25713 -0.05325  0.19542  1.41253

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         2.2514     0.3698   6.089 9.57e-09 ***
Sepal.Width         0.8036     0.1063   7.557 4.19e-12 ***
Speciesversicolor   1.4587     0.1121  13.012  < 2e-16 ***
Speciesvirginica    1.9468     0.1000  19.465  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.438 on 146 degrees of freedom
Multiple R-squared:  0.7259,    Adjusted R-squared:  0.7203
F-statistic: 128.9 on 3 and 146 DF,  p-value: < 2.2e-16
```

- What do those coefficients mean?
  - Sepal.width
  - Intercept
  - Speciesversicolor
  - Speciesvirginica

Dummy Variables in Regression

- A dummy variable (aka, an indicator variable) is a numeric variable that represents categorical data, such as gender, race, political affiliation, etc.
- Usually 0-1 variables
- Typically, 1 represents the presence of a qualitative attribute, and 0 represents the absence.

https://stattrek.com/multiple-regression/dummy-variables.aspx

How Many Dummy Variables?

- To represent a categorical variable that can assume *k* different values, a researcher would need to define *k-1* dummy variables
- For example, suppose we are interested in political affiliation, a categorical variable that might assume three values - Republican, Democrat, or Independent. We could represent political affiliation with two dummy variables:
  - $X_1 = 1$, if Republican; $X_1 = 0$, otherwise.
  - $X_2 = 1$, if Democrat; $X_2 = 0$, otherwise.

Avoid the Dummy Variable Trap!

- When defining dummy variables, a common mistake is to define too many variables.
- A k^{th} dummy variable is redundant; it carries no new information, and it creates a severe multicollinearity problem for the analysis.
- Using *k* dummy variables when only *k - 1* dummy variables are required is known as the dummy variable trap.

How to Interpret Dummy Variables?

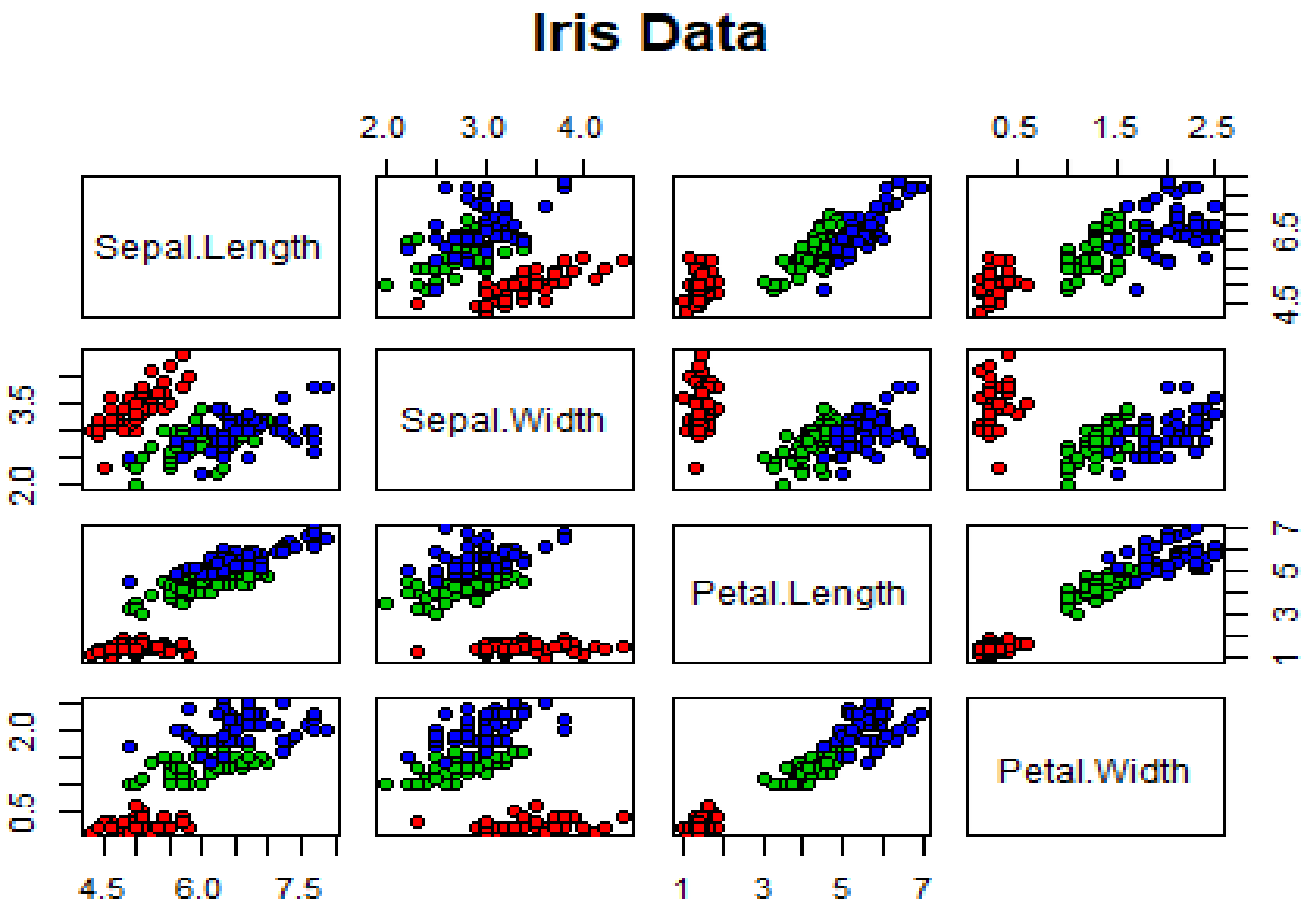
- The value of the categorical variable that is *not* represented explicitly by a dummy variable is called the reference group. In this example, the reference group consists of Independent voters.
- Back to Model 1: What do those coefficients mean?
  - Sepal.width
  - Intercept
  - Speciesversicolor
  - Speciesvirginica

Model 2 from the Class

Model #	Coefficients Estimated	R^2	Adj R^2
2a	(Intercept) Petal.Length	0.76	0.7583
2b	(Intercept) Sepal.Width Petal.Length Speciesversicolor Speciesvirginica	0.8633	0.8595
2c	(Intercept) Sepal.Width Speciesveriscolor Speciesvirginica Petal.Width Petal.Length	0.8673	0.8627
2d	(Intercept) I(Petal.Length * Petal.Width) I(Sepal.Width^2) I(Petal.Length * Species_versicolor) I(Petal.Length^2)	0.8735	0.8701
2e	Speciessetosa Speciesversicolor Speciesvirginica Sepal.Width:Speciessetosa Sepal.Width:Speciesversicolor Sepal.Width:Speciesvirginica	0.9947	0.9944
2f	Sepal.Width (0.43222) Petal.Length (0.77563) Speciessetosa (2.39039) Speciesversicolor (1.43458) Speciesvirginica (0.99629)	0.9973	0.9972

Is there any other terms that can *potentially* help characterizing the variance of y?

- Let's get back to the very beginning of our class. What can be used as input variables for regression?



Model 3a

```
#### 3.3 Model 3a

```{r}
lmfit.sepal3<-lm(Sepal.Length~ Petal.Length : Species+ Species, data=iris)
summary(lmfit.sepal3)
```
```

```
Call:
lm(formula = Sepal.Length ~ Petal.Length:Species + Species, data = iris)

Residuals:
     Min       1Q   Median       3Q      Max
-0.73479 -0.22785 -0.03132  0.24375  0.93608

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                  4.21317    0.40742  10.341  < 2e-16 ***
Speciesversicolor           -1.80565    0.59843  -3.017  0.00302 **
Speciesvirginica            -3.15351    0.63407  -4.973 1.85e-06 ***
Petal.Length:Speciessetosa   0.54229    0.27677   1.959  0.05200 .
Petal.Length:Speciesversicolor  0.82828    0.10228   8.098 2.16e-13 ***
Petal.Length:Speciesvirginica   0.99574    0.08709  11.433  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3365 on 144 degrees of freedom
Multiple R-squared:  0.8405,    Adjusted R-squared:  0.8349
F-statistic: 151.7 on 5 and 144 DF,  p-value: < 2.2e-16
```

Model 3b

```
#### 3.3 Model 3b

```{r}
lmfit.sepal3<-lm(Sepal.Length~ Petal.Length : Species+ Species - 1, data=iris)
summary(lmfit.sepal3)
```
```

```
Call:
lm(formula = Sepal.Length ~ Petal.Length:Species + Species -
    1, data = iris)

Residuals:
     Min       1Q   Median       3Q      Max
-0.73479 -0.22785 -0.03132  0.24375  0.93608

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
Speciessetosa                4.21317    0.40742  10.341  < 2e-16 ***
Speciesversicolor            2.40752    0.43832   5.493 1.74e-07 ***
Speciesvirginica             1.05966    0.48586   2.181   0.0308 *
Petal.Length:Speciessetosa   0.54229    0.27677   1.959   0.0520 .
Petal.Length:Speciesversicolor  0.82828    0.10228   8.098 2.16e-13 ***
Petal.Length:Speciesvirginica   0.99574    0.08709  11.433  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3365 on 144 degrees of freedom
Multiple R-squared:  0.9969,    Adjusted R-squared:  0.9967
F-statistic:  7667 on 6 and 144 DF,  p-value: < 2.2e-16
```

Model 3a vs Model 3b

```
#### 3.3 Model 3a

```{r}
lmfit.sepal3<-lm(Sepal.Length~ Petal.Length : Species+ Species, data=iris
summary(lmfit.sepal3)
```
```

```
Call:
lm(formula = Sepal.Length ~ Petal.Length:Species + Species, data = iris)

Residuals:
     Min       1Q   Median       3Q      Max
-0.73479 -0.22785 -0.03132  0.24375  0.93608

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                  4.21317    0.40742  10.341  < 2e-16 ***
Speciesversicolor           -1.80565    0.59843  -3.017  0.00302 **
Speciesvirginica            -3.15351    0.63407  -4.973 1.85e-06 ***
Petal.Length:Speciessetosa   0.54229    0.27677   1.959  0.05200 .
Petal.Length:Speciesversicolor 0.82828  0.10228   8.098 2.16e-13 ***
Petal.Length:Speciesvirginica  0.99574  0.08709  11.433  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3365 on 144 degrees of freedom
Multiple R-squared:  0.8405,    Adjusted R-squared:  0.8349
F-statistic: 151.7 on 5 and 144 DF,  p-value: < 2.2e-16
```

```
#### 3.3 Model 3b

```{r}
lmfit.sepal3<-lm(Sepal.Length~ Petal.Length : Species+ Species - 1, data=iris
summary(lmfit.sepal3)
```
```

```
Call:
lm(formula = Sepal.Length ~ Petal.Length:Species + Species -
    1, data = iris)

Residuals:
     Min       1Q   Median       3Q      Max
-0.73479 -0.22785 -0.03132  0.24375  0.93608

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
Speciessetosa                4.21317    0.40742  10.341  < 2e-16 ***
Speciesversicolor            2.40752    0.43832   5.493 1.74e-07 ***
Speciesvirginica             1.05966    0.48586   2.181   0.0308 *
Petal.Length:Speciessetosa   0.54229    0.27677   1.959   0.0520 .
Petal.Length:Speciesversicolor 0.82828  0.10228   8.098 2.16e-13 ***
Petal.Length:Speciesvirginica  0.99574  0.08709  11.433  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3365 on 144 degrees of freedom
Multiple R-squared:  0.9969,    Adjusted R-squared:  0.9967
F-statistic:  7667 on 6 and 144 DF,  p-value: < 2.2e-16
```

MISSISSIPPI STATE UNIVERSITY

Model 3a

- $y = \beta_0^a + \beta_1^a x_1 + \beta_2^a x_2 + \beta_3^a x_3 + \beta_4^a x_4 + \beta_5^a x_5 + \varepsilon$
- Input variables:
  - $x_1 = 1$ if species= versicolor, 0 o.w.
  - $x_2 = 1$ if species= virginica, 0 o.w.
  - x_3 =petal.length*I(species= setosa)
  - x_4 =petal.length*x_1
  - x_5 = petal.length*x_2

```
#### 3.3 Model 3a

```{r}
lmfit.sepal3<-lm(Sepal.Length~ Petal.Length : Species+ Species, data=iris)
summary(lmfit.sepal3)
```
```

```
Call:
lm(formula = Sepal.Length ~ Petal.Length:Species + Species, data = iris)

Residuals:
     Min       1Q   Median       3Q      Max
-0.73479 -0.22785 -0.03132  0.24375  0.93608

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                      4.21317    0.40742  10.341  < 2e-16 ***
Speciesversicolor               -1.80565    0.59843  -3.017  0.00302 **
Speciesvirginica                -3.15351    0.63407  -4.973 1.85e-06 ***
Petal.Length:Speciessetosa       0.54229    0.27677   1.959  0.05200 .
Petal.Length:Speciesversicolor   0.82828    0.10228   8.098 2.16e-13 ***
Petal.Length:Speciesvirginica    0.99574    0.08709  11.433  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3365 on 144 degrees of freedom
Multiple R-squared:  0.8405,    Adjusted R-squared:  0.8349
F-statistic: 151.7 on 5 and 144 DF,  p-value: < 2.2e-16
```

Model 3a

- $y = \beta_0^a + \beta_1^a x_1 + \beta_2^a x_2 + \beta_2^a x_3 + \beta_4^a x_4 + \beta_5^a x_5 + \varepsilon$
- Coefficients:
  - β_0^a: intercept: reference category (Setosa)
  - β_1^a: intercept difference between versicolor and reference category
  - β_2^a: intercept difference between virginica and reference category
  - β_3^a: when species=setosa, increasing one unit of petal.length will result in β_3^aunits of increase in Sepal.Length
  - …

```
#### 3.3 Model 3a

```{r}
lmfit.sepal3<-lm(Sepal.Length~ Petal.Length : Species+ Species, data=iris)
summary(lmfit.sepal3)
```
```

```
Call:
lm(formula = Sepal.Length ~ Petal.Length:Species + Species, data = iris)

Residuals:
     Min       1Q   Median       3Q      Max
-0.73479 -0.22785 -0.03132  0.24375  0.93608

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                  4.21317    0.40742  10.341  < 2e-16 ***
Speciesversicolor           -1.80565    0.59843  -3.017  0.00302 **
Speciesvirginica            -3.15351    0.63407  -4.973 1.85e-06 ***
Petal.Length:Speciessetosa   0.54229    0.27677   1.959  0.05200 .
Petal.Length:Speciesversicolor  0.82828    0.10228   8.098 2.16e-13 ***
Petal.Length:Speciesvirginica   0.99574    0.08709  11.433  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3365 on 144 degrees of freedom
Multiple R-squared:  0.8405,    Adjusted R-squared:  0.8349
F-statistic: 151.7 on 5 and 144 DF,  p-value: < 2.2e-16
```

Model 3b

- $y = \beta_0^b x_0 + \beta_1^b x_1 + \beta_2^b x_2 + \beta_3^b x_3 + \beta_4^b x_4 + \beta_5^b x_5 + \varepsilon$
- Input variables:
  - $x_0 = 1$ if species= setosa, 0 o.w.
  - $x_1 = 1$ if species= versicolor, 0 o.w.
  - $x_2 = 1$ if species= virginica, 0 o.w.
  - x_3 =petal.length*x_0
  - x_4 =petal.length*x_1
  - x_5 = petal.length*x_2

```
#### 3.3 Model 3b

```{r}
lmfit.sepal3<-lm(Sepal.Length~ Petal.Length : Species+ Species - 1, data=iris)
summary(lmfit.sepal3)
```
```

```
Call:
lm(formula = Sepal.Length ~ Petal.Length:Species + Species -
    1, data = iris)

Residuals:
     Min       1Q   Median       3Q      Max
-0.73479 -0.22785 -0.03132  0.24375  0.93608

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
Speciessetosa                 4.21317    0.40742  10.341  < 2e-16 ***
Speciesversicolor             2.40752    0.43832   5.493 1.74e-07 ***
Speciesvirginica              1.05966    0.48586   2.181   0.0308 *
Petal.Length:Speciessetosa    0.54229    0.27677   1.959   0.0520 .
Petal.Length:Speciesversicolor 0.82828   0.10228   8.098 2.16e-13 ***
Petal.Length:Speciesvirginica 0.99574    0.08709  11.433  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3365 on 144 degrees of freedom
Multiple R-squared:  0.9969,    Adjusted R-squared:  0.9967
F-statistic:  7667 on 6 and 144 DF,  p-value: < 2.2e-16
```

Model 3b

- $y = \beta_0^b x_0 + \beta_1^b x_1 + \beta_2^b x_2 + \beta_3^b x_3 + \beta_4^b x_4 + \beta_5^b x_5 + \varepsilon$
- Coefficients:
  - β_0^b: intercept for species = setosa
  - β_1^b: intercept for species = versicolor
  - β_2^b: intercept when species =virginica
  - β_3^b: when species=setosa, increasing one unit of petal.length will result in β_3^bunits of increase in Sepal.Length
  - …

```
#### 3.3 Model 3b

```{r}
lmfit.sepal3<-lm(Sepal.Length~ Petal.Length : Species+ Species - 1, data=iris)
summary(lmfit.sepal3)
```
```

```
Call:
lm(formula = Sepal.Length ~ Petal.Length:Species + Species -
    1, data = iris)

Residuals:
     Min       1Q   Median       3Q      Max
-0.73479 -0.22785 -0.03132  0.24375  0.93608

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
Speciessetosa                4.21317    0.40742  10.341  < 2e-16 ***
Speciesversicolor            2.40752    0.43832   5.493 1.74e-07 ***
Speciesvirginica             1.05966    0.48586   2.181   0.0308 *
Petal.Length:Speciessetosa   0.54229    0.27677   1.959   0.0520 .
Petal.Length:Speciesversicolor  0.82828    0.10228   8.098 2.16e-13 ***
Petal.Length:Speciesvirginica   0.99574    0.08709  11.433  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3365 on 144 degrees of freedom
Multiple R-squared:  0.9969,    Adjusted R-squared:  0.9967
F-statistic:  7667 on 6 and 144 DF,  p-value: < 2.2e-16
```

Why their R^2 values are so different?

- Recall the definition of R^2

$$R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

The first equality *only* occurs because of the inclusion of the intercept in the model.

- When there is no intercept in the model, R uses the modified form

$$R_0^2 = \frac{\sum_i \hat{y}_i^2}{\sum_i y_i^2} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i y_i^2}$$

Homework 1: Question 1.4

- Use subset selection method (*step*) to explore if we can further improve the fit by incorporating more interaction terms? This will be your Model 3 in Q1 HW1