



STOR 320 Modeling IV

Lecture 19

Yao Li

Department of Statistics and Operations Research

UNC Chapel Hill

Introduction

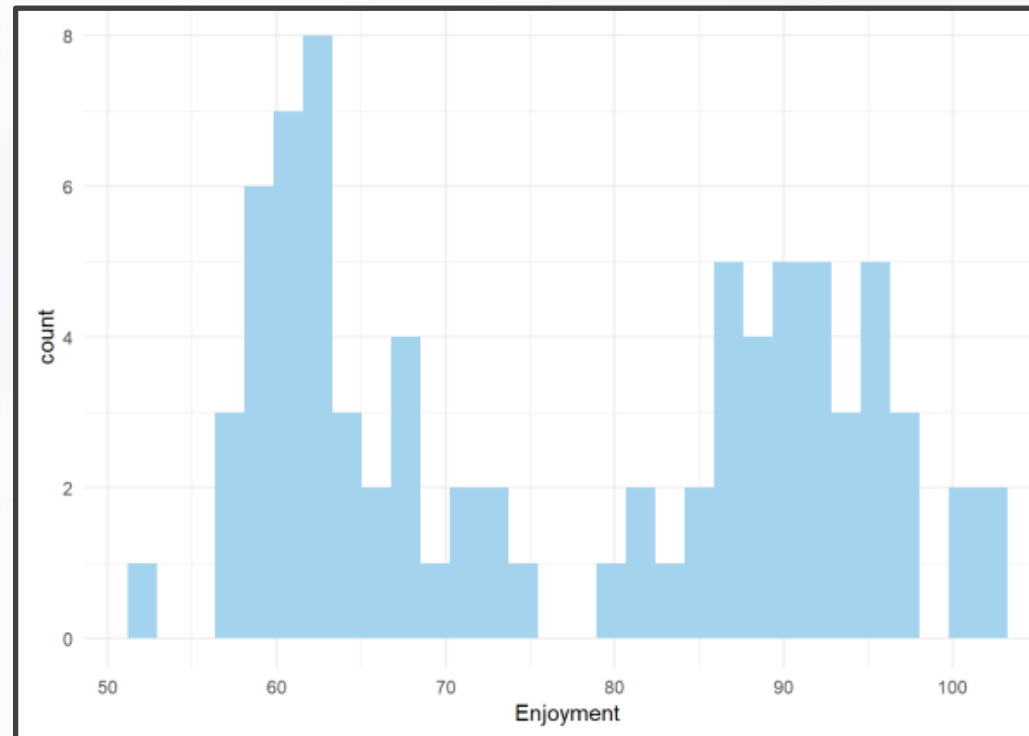
- Read Chapter 18 (R4DS)
- Previously: Numeric Variables
- New Focus
 - Categorical Predictor Variables
 - Interaction Effects
- Understand Using Multiple Datasets and Visualizations

Example 1: Data

- Data Overview
 - Enjoyment (E)
 - Food (F)
 - Condiment (C)
 - 80 Observations

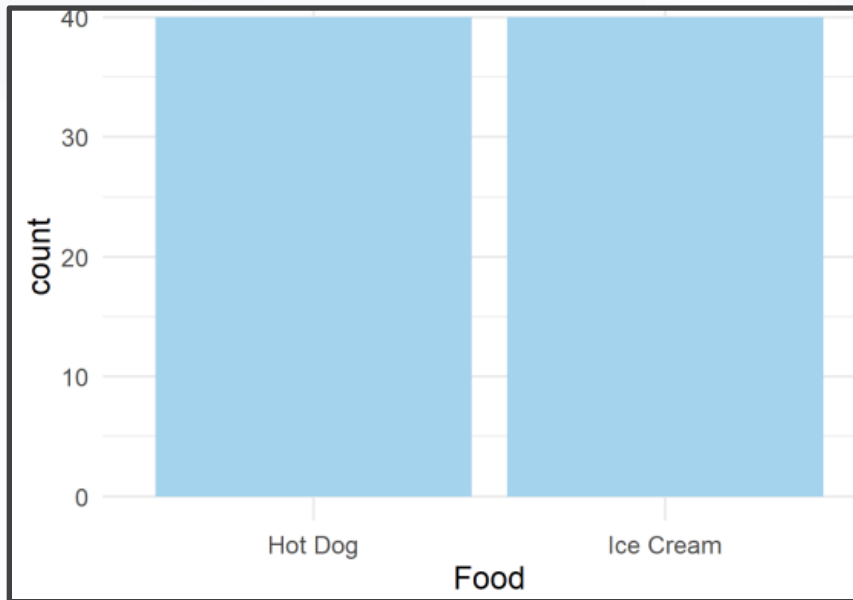
Enjoyment <dbl>	Food <chr>	Condiment <chr>
81.92696	Hot Dog	Mustard
84.93977	Hot Dog	Mustard
90.28648	Hot Dog	Mustard
89.56180	Hot Dog	Mustard
97.67683	Hot Dog	Mustard

- Enjoyment Visualized

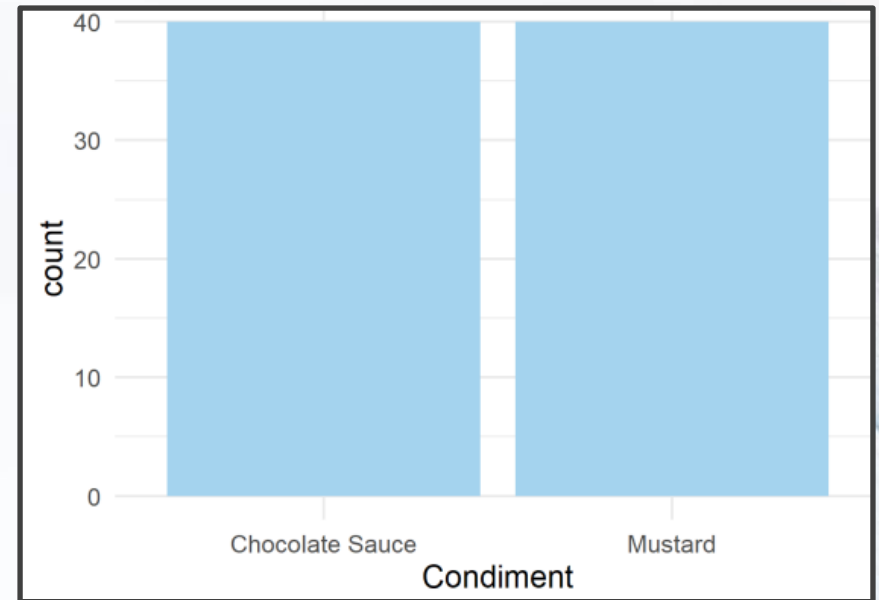


Example 1: Data

- Food Visualized



- Condiment Visualized



Example 1: Question

- Question of Interest

Can We Predict a Person's Culinary Enjoyment if...

We Serve Them a Particular Item:

- *Hot Dog*
- *Ice Cream*

With a Particular Condiment

- *Mustard*
- *Chocolate Sauce*



Example 1: Model 1

- Regressing E on F

```
EvsF.Model=lm(Enjoyment~Food,data=CONDIMENT)
tidy(EvsF.Model)
```

```
## # A tibble: 2 x 5
##   term                estimate std.error statistic  p.value
##   <chr>              <dbl>    <dbl>    <dbl>    <dbl>
## 1 (Intercept)       77.5        2.39    32.4    5.82e-47
## 2 FoodIce Cream   -0.283        3.39    -0.0835 9.34e- 1
```

- $\hat{E} = 77.5 - 0.283F$
- Questions:
 - What Does 77.5 Represent?
 - What About -0.283?

Example 1: Model 1

- What is R Doing?

```
CONDIMENT$Food[1:6]
```

```
## [1] "Hot Dog" "Hot Dog" "Hot Dog" "Hot Dog"  
" "Hot Dog" "Hot Dog"
```

```
head(model_matrix(CONDIMENT, Enjoyment~Food))
```

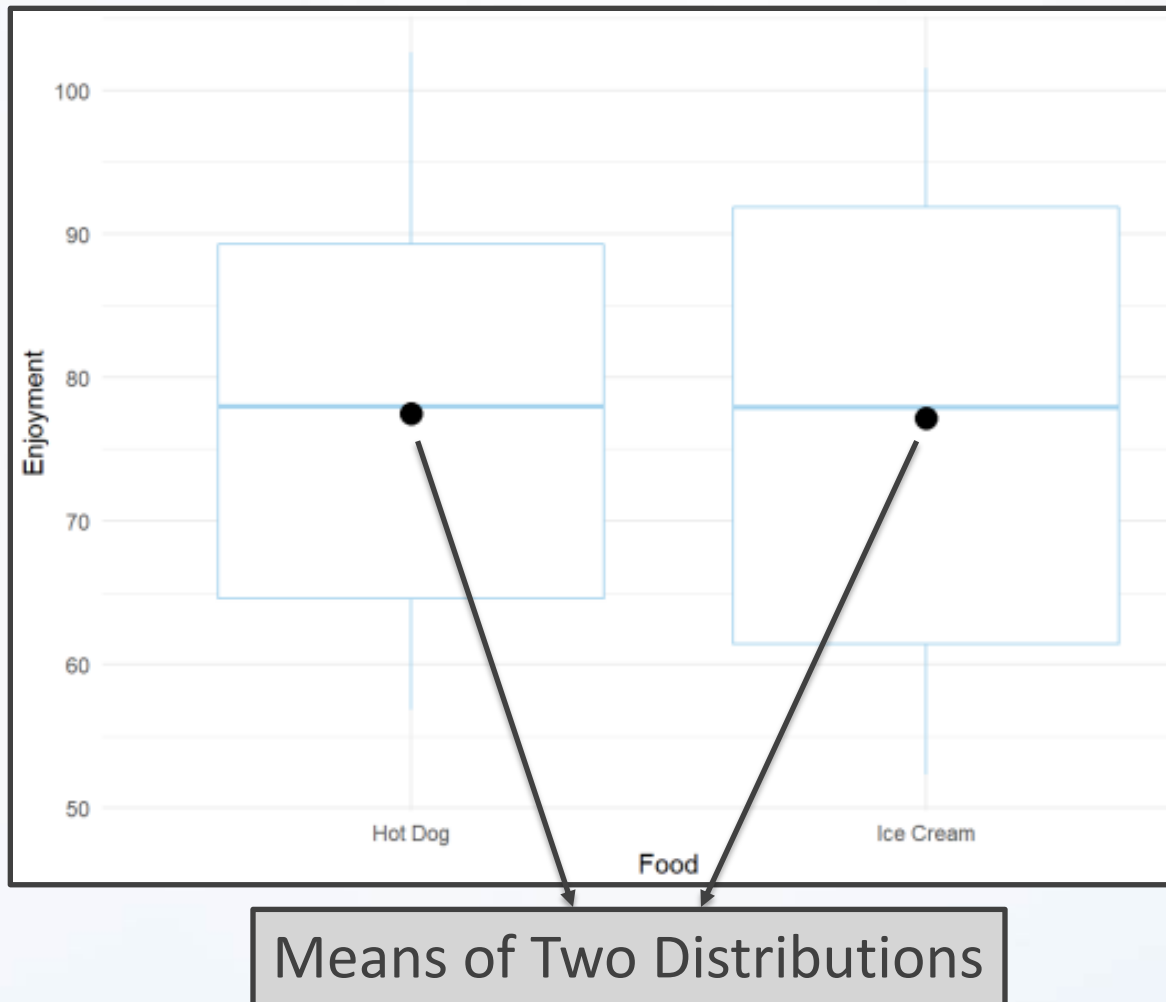
```
## # A tibble: 6 x 2  
##   `(Intercept)` `FoodIce Cream`  
##           <dbl>           <dbl>  
## 1             1             0  
## 2             1             0  
## 3             1             0  
## 4             1             0  
## 5             1             0  
## 6             1             0
```

Example 1: Interpretation

- Regressing E on F
 - $\hat{E} = 77.5 - 0.283F$
 - $F = \begin{cases} 0 & \text{if Hot Dog} \\ 1 & \text{if Ice Cream} \end{cases}$
 - If You Eat a Hot Dog,
 $\hat{E} = 77.5 - 0.283(0) = 77.5$
 - If You Eat Ice Cream,
 $\hat{E} = 77.5 - 0.283(1) = 77.217$
 - P-value = 0.934 for the Parameter Estimated by 0.283
(Not Statistically Significant)

Example 1: Interpretation

- Understanding This Visually



Example 1: Model 2

- Regressing E on C

```
Evsc.Model=lm(Enjoyment~Condiment,data=CONDIMENT)  
tidy(Evsc.Model)
```

```
## # A tibble: 2 x 5  
##   term                estimate std.error statistic  p.value  
##   <chr>              <dbl>    <dbl>    <dbl>    <dbl>  
## 1 (Intercept)        79.2      2.38     33.3  6.67e-48  
## 2 CondimentMustard  -3.73     3.36     -1.11  2.71e- 1
```

Significant: P-value < 0.05

- $\hat{E} = 79.2 - 3.73C$

Not Significant: P-value > 0.05

- $$C = \begin{cases} 0 & \text{if Chocolate Sauce} \\ 1 & \text{if Mustard} \end{cases}$$

Example 1: Model 3

- Regressing E on C + F

```
Evscf.Model=lm(Enjoyment~Food+Condiment,data=CONDIMENT)
tidy(Evscf.Model)
```



```
## # A tibble: 3 x 5
##   term                estimate std.error statistic  p.value
##   <chr>              <dbl>    <dbl>    <dbl>    <dbl>
## 1 (Intercept)        79.3        2.93     27.1    4.07e-41
## 2 FoodIce Cream     -0.283        3.38     -0.0836 9.34e- 1
## 3 CondimentMustard  -3.73        3.38     -1.10    2.74e- 1
```

- $\hat{E} = 79.3 - 0.283F - 3.73C$
- $F = \begin{cases} 0 & \text{if Hot Dog} \\ 1 & \text{if Ice Cream} \end{cases}$
- $C = \begin{cases} 0 & \text{if Chocolate Sauce} \\ 1 & \text{if Mustard} \end{cases}$
- What does 79.3 Represent?

Example 1: Model 3

- Obtaining Predicted Values

```
GRID=CONDIMENT %>%  
  data_grid(  
    Food=unique(Food),  
    Condiment=unique(Condiment)  
  )  
print(GRID)
```

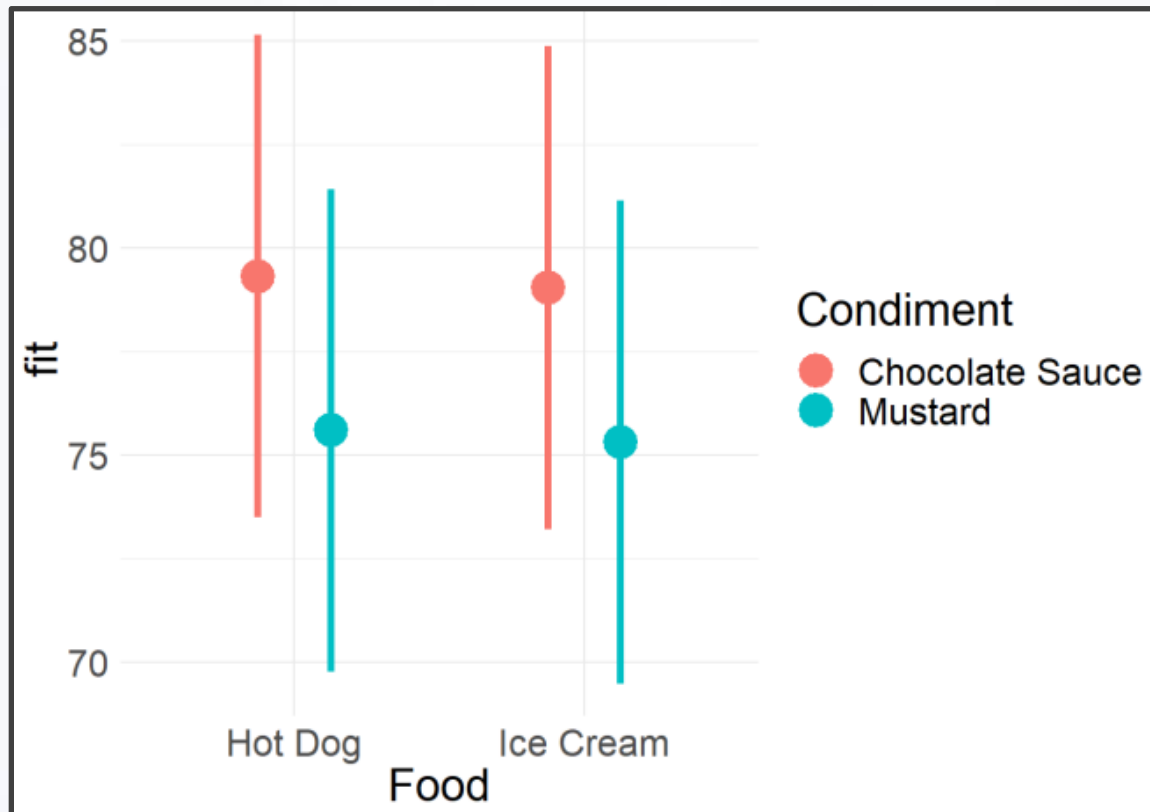
```
## # A tibble: 4 x 2  
##   Food      Condiment  
##   <chr>    <chr>  
## 1 Hot Dog  Chocolate Sauce  
## 2 Hot Dog  Mustard  
## 3 Ice Cream Chocolate Sauce  
## 4 Ice Cream Mustard
```

```
GRID2=cbind(GRID,predict(EvsCF.Model,  
                          newdata=GRID,  
                          interval="confidence"))  
print(GRID2)
```

##	Food	Condiment	fit	lwr	upr
## 1	Hot Dog	Chocolate Sauce	79.32368	73.49373	85.15363
## 2	Hot Dog	Mustard	75.59862	69.76867	81.42857
## 3	Ice Cream	Chocolate Sauce	79.04103	73.21108	84.87098
## 4	Ice Cream	Mustard	75.31598	69.48603	81.14593

Example 1: Model 3

- Prediction Visualization



Example 1: Model 4

- Interaction Effect

```
EvFC.Full.Model=lm(Enjoyment~Food+Condiment+Food*Condiment,data=CONDIMENT)
tidy(EvFC.Full.Model)
```

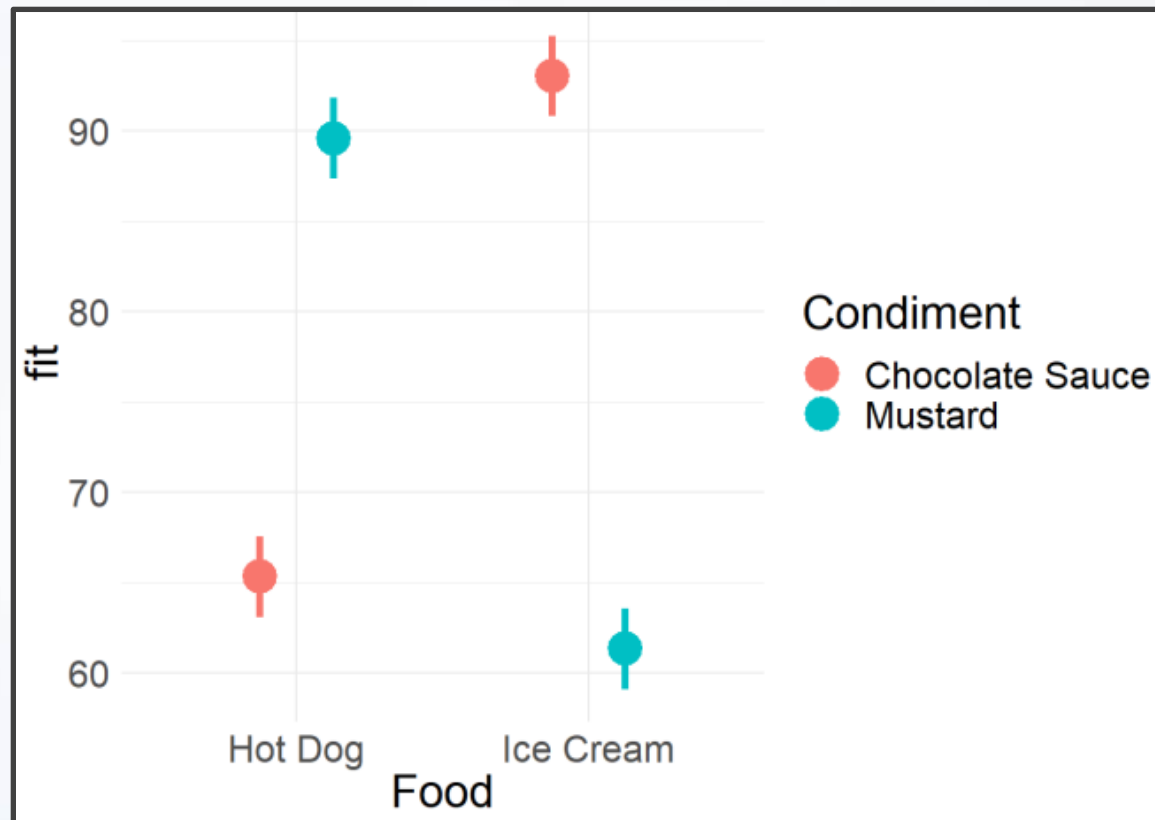
```
## # A tibble: 4 x 5
##   term                                estimate std.error statistic  p.value
##   <chr>                                <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)                        65.3       1.12      58.3 7.18e-65
## 2 FoodIce Cream                       27.7       1.58      17.5 2.11e-28
## 3 CondimentMustard                    24.3       1.58      15.3 5.58e-25
## 4 FoodIce Cream:CondimentMustard    -56.0       2.24     -25.0 1.95e-38
```

$$\hat{E} = 65.32 + 27.73F + 24.29C - 56.03FC$$

- $F = \begin{cases} 0 & \text{if Hot Dog} \\ 1 & \text{if Ice Cream} \end{cases}$
- $C = \begin{cases} 0 & \text{if Chocolate Sauce} \\ 1 & \text{if Mustard} \end{cases}$
- $FC = \begin{cases} 0 & \text{otherwise} \\ 1 & \text{if Ice Cream and Mustard} \end{cases}$

Example 1: Model 4

- Understanding This Visually
 - What Is Different?



Example 1: Summary

- Summary
 - Categorical Predictors
 - Purpose:
 - Generalize t-test
 - Estimate Difference in Means Between Groups

Example 2: Data

- Data Overview
 - Popular Built-in Data
 - Sepal.Width (W)
 - Sepal.Length (L)
 - Species (S)
 - 150 Observations

```
IRIS=iris[,c(1,2,5)]  
names(IRIS)=c("L", "W", "S")  
head(IRIS)
```

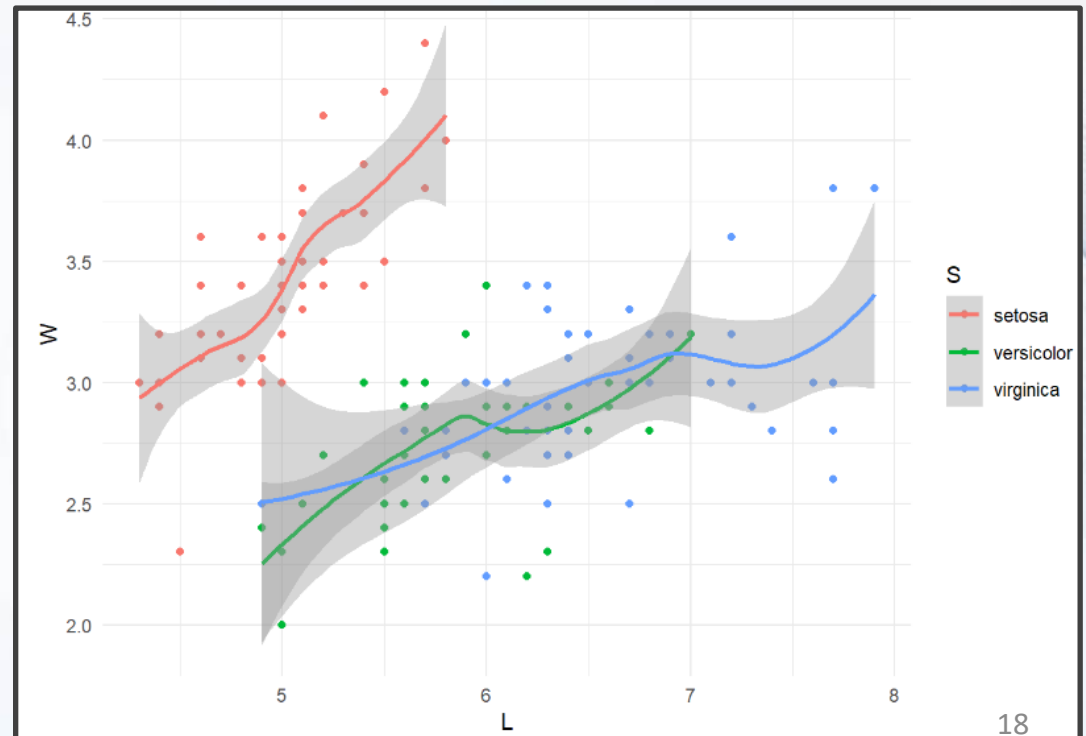
```
##      L      W      S  
## 1 5.1 3.5 setosa  
## 2 4.9 3.0 setosa  
## 3 4.7 3.2 setosa  
## 4 4.6 3.1 setosa  
## 5 5.0 3.6 setosa  
## 6 5.4 3.9 setosa
```

Example 2: Question

- Question of Interest

Can We Explain the Variation in Sepal Width Using Sepal Length and Species (setosa, versicolor, virginica)?

- Visual of Relationship



Example 2: Models

- Multiple Models

```
model1=lm(W~L, IRIS)
tidy(model1)
```

```
## # A tibble: 2 x 5
##   term      estimate std.error statistic  p.value
##   <chr>      <dbl>    <dbl>    <dbl>    <dbl>
## 1 (Intercept)  3.42      0.254     13.5 1.55e-27
## 2 L          -0.0619    0.0430     -1.44 1.52e- 1
```

```
model2=lm(W~L+S, IRIS)
tidy(model2)
```

```
## # A tibble: 4 x 5
##   term      estimate std.error statistic  p.value
##   <chr>      <dbl>    <dbl>    <dbl>    <dbl>
## 1 (Intercept)  1.68      0.235      7.12 4.46e-11
## 2 L           0.350    0.0463     7.56 4.19e-12
## 3 Sversicolor -0.983     0.0721    -13.6 7.62e-28
## 4 Svirginica  -1.01     0.0933    -10.8 2.41e-20
```

Setosa: $\hat{E} = 1.68 + 0.35L$

Versicolor: $\hat{E} = 1.68 + 0.35L - 0.983$

Virginica: $\hat{E} = 1.68 + 0.35L - 1.01$

Example 2: Models

- Full Model Estimated

```
model3=lm(W~L+S+L*S, IRIS)
tidy(model3)
```

```
## # A tibble: 6 x 5
##   term          estimate std.error statistic  p.value
##   <chr>          <dbl>    <dbl>    <dbl>    <dbl>
## 1 (Intercept)   -0.569    0.554    -1.03 3.06e- 1
## 2 L              0.799    0.110     7.23 2.55e-11
## 3 Sversicolor    1.44     0.713     2.02 4.51e- 2
## 4 Svirginica      2.02     0.686     2.94 3.85e- 3
## 5 L:Sversicolor  -0.479    0.134    -3.58 4.65e- 4
## 6 L:Svirginica  -0.567    0.126    -4.49 1.45e- 5
```

Adjustment
In Mean

Adjustment
In Slope

$$\text{Setosa: } \hat{E} = 0.799L - 0.569$$

$$\text{Versicolor: } \hat{E} = (0.799 - 0.479)L + 1.44 - 0.569$$

$$\text{Virginica: } \hat{E} = (0.799 - 0.567)L + 2.02 - 0.569$$

Example 2: Predictions

- Gathering Predictions

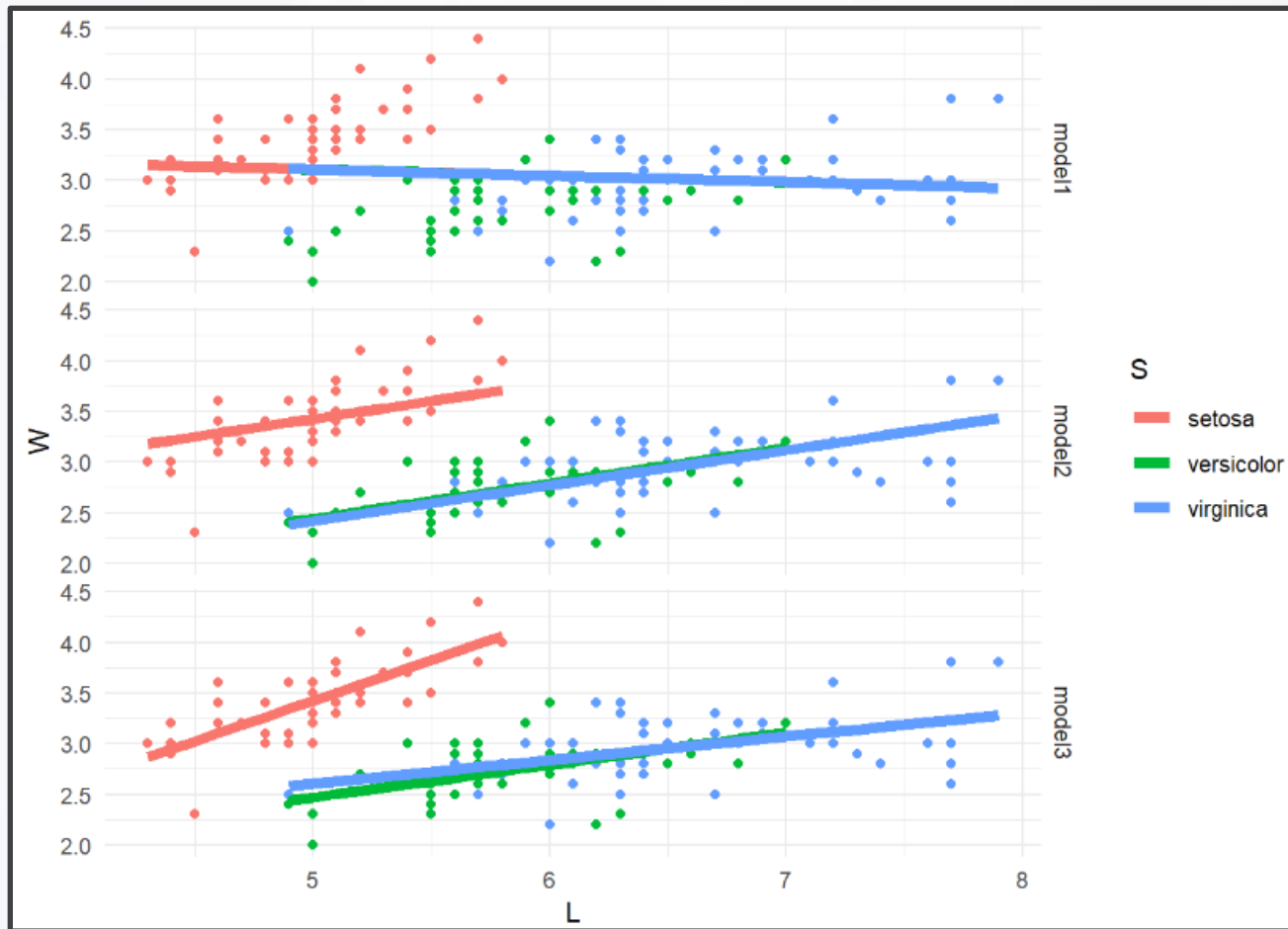
```
IRIS %>%  
  gather_predictions(model1,model2,model3)%>%  
  glimpse()  
  
## Observations: 450  
## Variables: 5  
## $ model <chr> "model1", "model1", "model1", "model1", "model1", "model1", ...  
## $ L      <dbl> 5.1, 4.9, 4.7, 4.6, 5.0, 5.4, 4.6, 5.0, 4.4, 4.9, 5.4, 4.4, ...  
## $ W      <dbl> 3.5, 3.0, 3.2, 3.1, 3.6, 3.9, 3.4, 3.4, 2.9, 3.1, 3.7, 3.4, ...  
## $ S      <fct> setosa, setosa, setosa, setosa, setosa, setosa, setosa, setosa, ...  
## $ pred   <dbl> 3.103334, 3.115711, 3.128088, 3.134277, 3.109523, 3.0847...
```

150 Predictions for 3 Models

- Variable Named “model”
- Allows Us To Quickly Create Graphics That Compare Models

Example 2: Visualization

- Visualizing Models



Example 2: Summary

- Summary
 - Numerical Response Variable
 - Categorical & Numerical Explanatory Variables