



STOR 320 Modeling VI

Lecture 29

Yao Li

Department of Statistics and Operations Research

UNC Chapel Hill

Introduction

- Read Chapter 23 (R4DS)
- Previously: Interactions (two categorical)
- New Focus
 - Interaction (categorical and continuous)
 - Interaction (two continuous)

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)
```

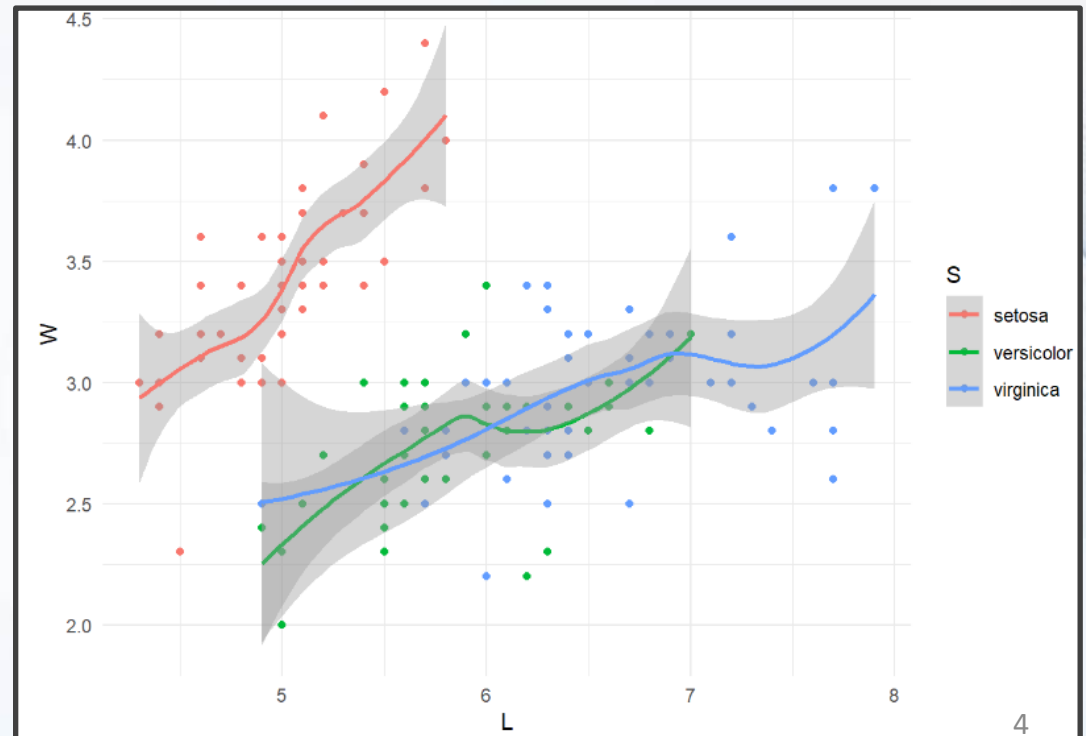
```
##      L      W      S  
## 1 5.1 3.5 setosa  
## 2 4.9 3.0 setosa  
## 3 4.7 3.2 setosa  
## 4 4.6 3.1 setosa  
## 5 5.0 3.6 setosa  
## 6 5.4 3.9 setosa
```

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

```
modell1=lm(W~L, IRIS)
tidy(modell1)
```

```
## # 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()

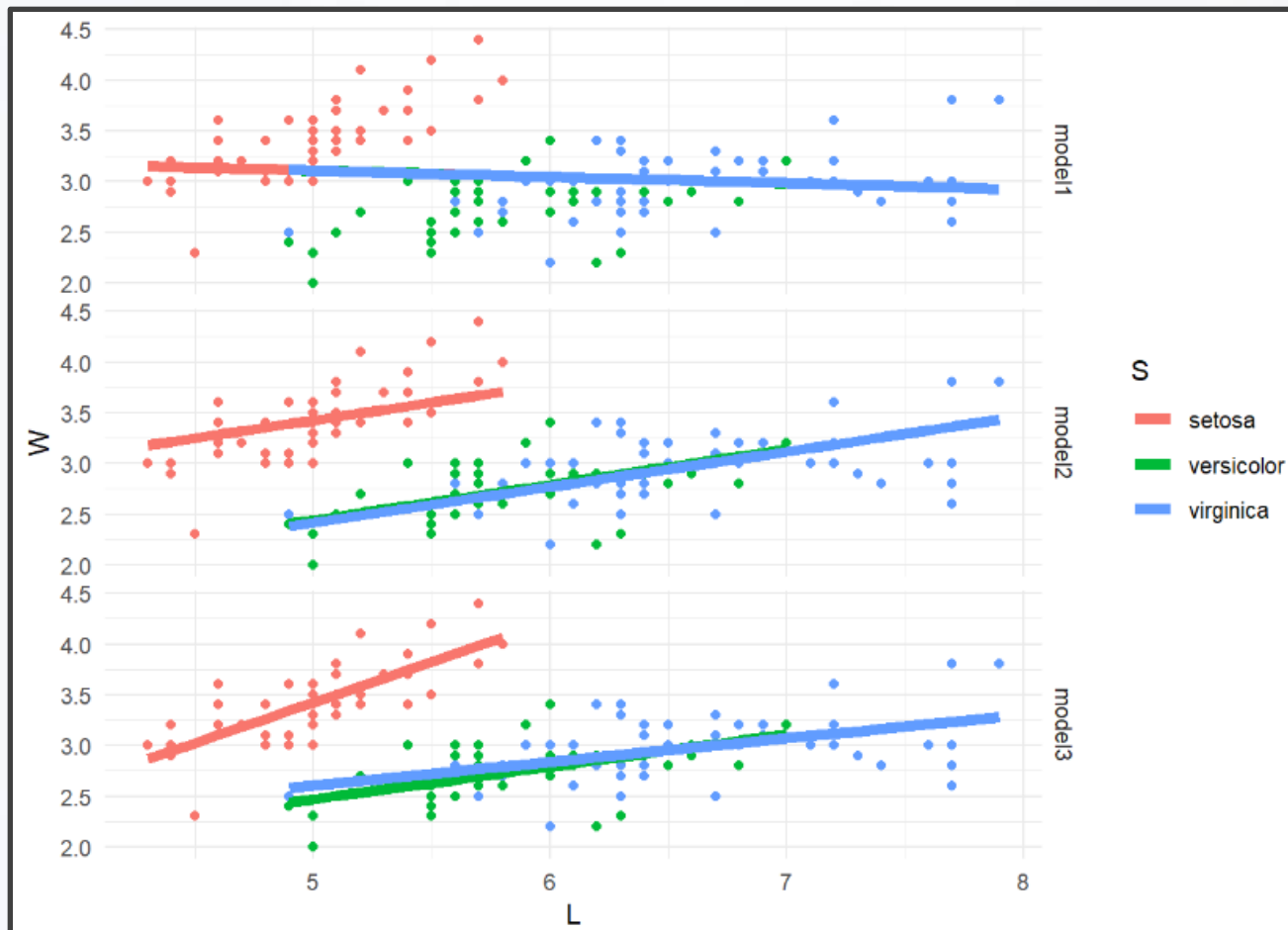
## # Observations: 450
## # Variables: 5
## # $ model <chr> "model1", "model1", "model1", "model1", "model1", "model1..."
## # $ L      <dbl> 5.1, 4.9, 4.7, 4.6, 5.0, 5.4, 4.6, 5.0, 4.4, 4.9, 5.4, 4.4...
## # $ W      <dbl> 3.5, 3.0, 3.2, 3.1, 3.6, 3.9, 3.4, 3.4, 2.9, 3.1, 3.7, 3.0...
## # $ S      <fct> 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

Example 3: Data

- Data Overview
 - Advertising Data
 - Sales
 - TV
 - Radio
 - 200 Observations

```
```{r, message=F}  
Ad = read_csv("Advertising.txt")[,c(2,3,5)]
head(Ad)
```
```

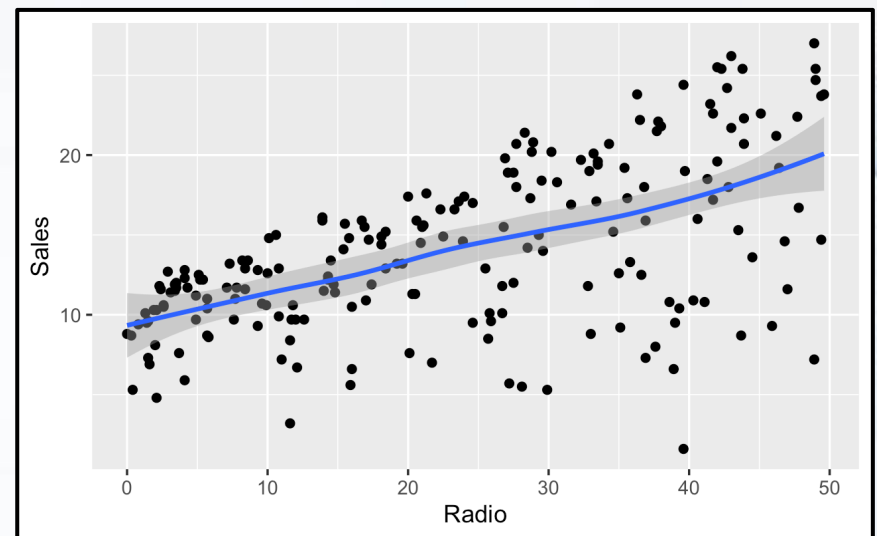
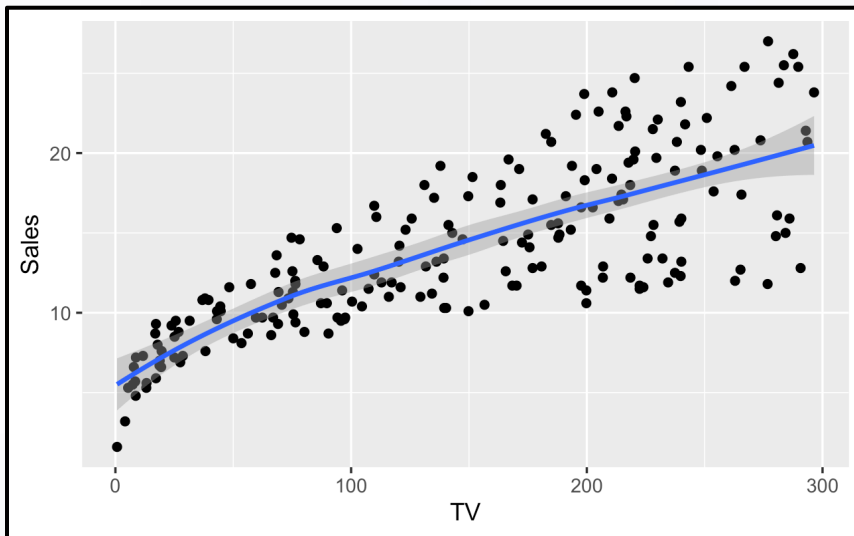
| TV
<dbl> | Radio
<dbl> | Sales
<dbl> |
|--------------------|-----------------------|-----------------------|
| 230.1 | 37.8 | 22.1 |
| 44.5 | 39.3 | 10.4 |
| 17.2 | 45.9 | 9.3 |
| 151.5 | 41.3 | 18.5 |
| 180.8 | 10.8 | 12.9 |
| 8.7 | 48.9 | 7.2 |

Example 3: Question

- Question of Interest

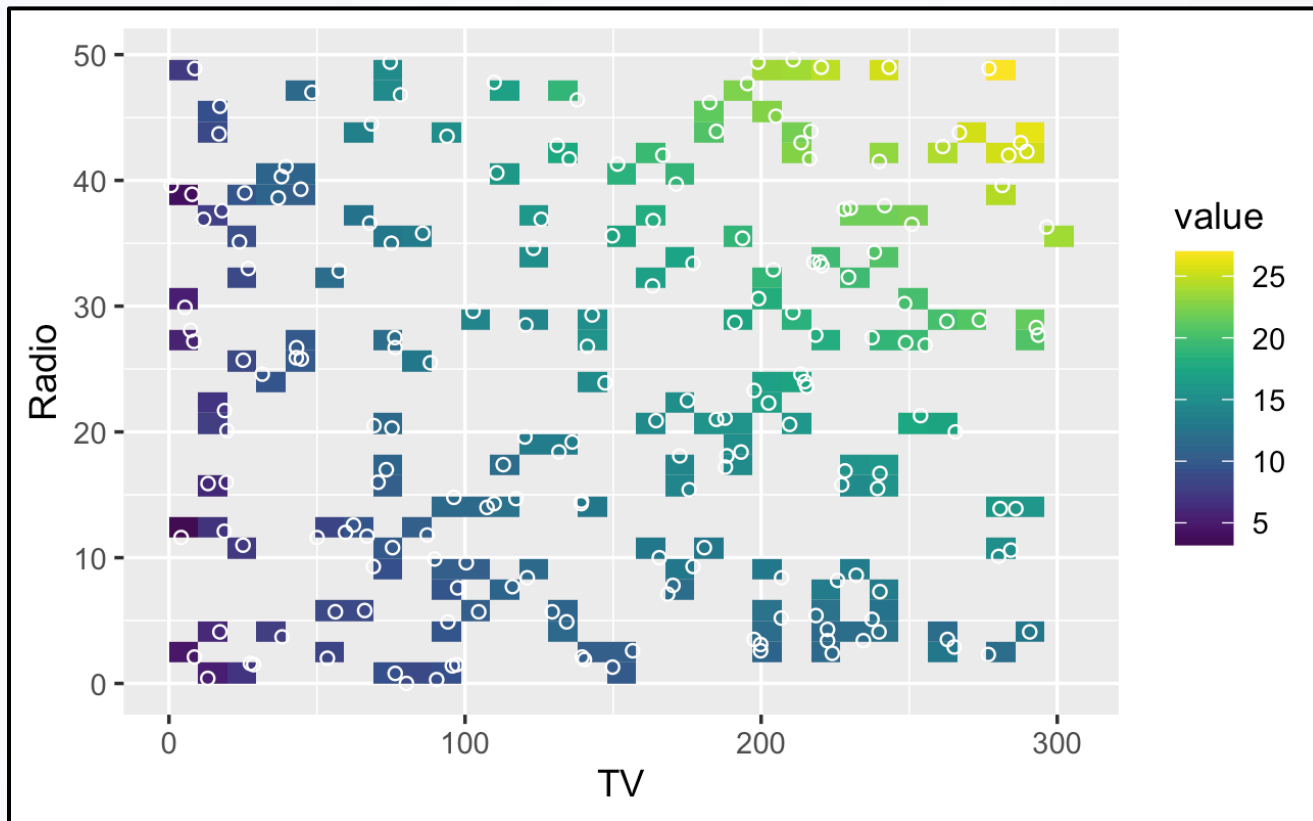
Can We Explain the Variation in Sales Using TV and Radio advertising budget?

- Visual of Relationship



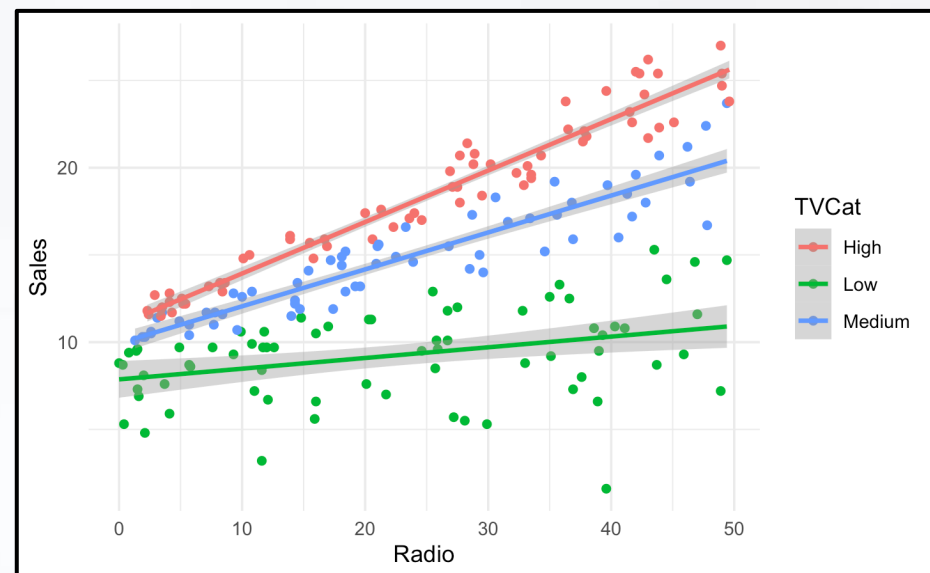
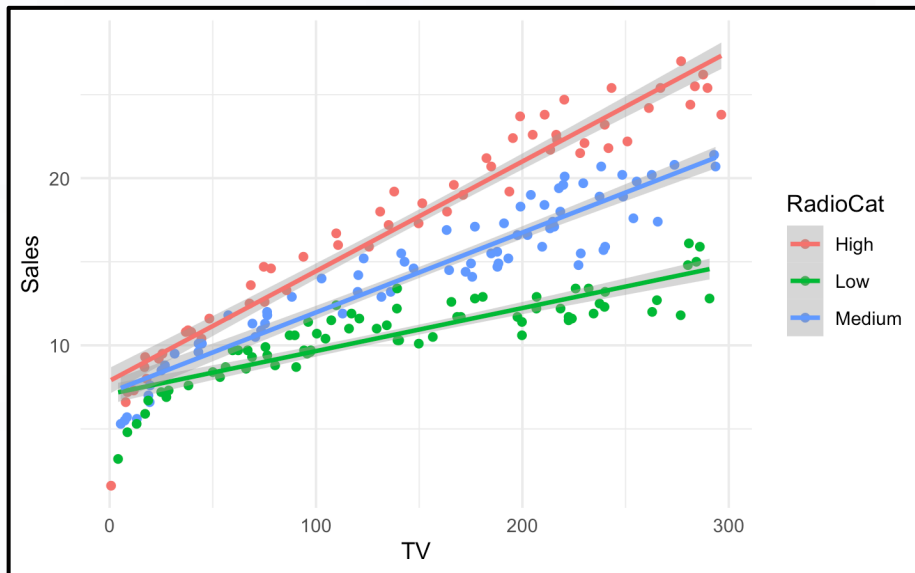
Example 3: Question

- Visual of Relationship



Example 3: Question

- Visual of Relationship



Example 3: Model1

- Model 1

```
model1=lm(Sales~TV+Radio,Ad)
tidy(model1)
```

```
## # A tibble: 3 x 5
##   term          estimate std.error statistic  p.value
##   <chr>          <dbl>    <dbl>    <dbl>    <dbl>
## 1 (Intercept)    2.92      0.294      9.92 4.57e-19
## 2 TV             0.0458    0.00139    32.9 5.44e-82
## 3 Radio          0.188     0.00804    23.4 9.78e-59
```

Model1: $\hat{E} = 2.92 + 0.046TV + 0.188Radio$

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC   BIC
##   <dbl>         <dbl> <dbl>    <dbl>    <dbl> <dbl> <dbl> <dbl> <dbl>
## 1    0.897         0.896  1.68    860. 4.83e-98     2  -386.  780.  794.
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

Example 3: Model Selection

- $AIC = -2 \ln(\hat{L}) + 2p$
 - goodness of fit: $2 \ln(\hat{L})$
 - \hat{L} : the maximized value of the likelihood of the model
 - p : number of parameters in the model
- $BIC = -2 \ln(\hat{L}) + p \ln(n)$
 - n : number of observations in the data

Example 3: Model 2

```
model2=lm(Sales~TV*Radio,Ad)
tidy(model2)
```

```
## # A tibble: 4 x 5
##   term          estimate std.error statistic  p.value
##   <chr>          <dbl>    <dbl>    <dbl>    <dbl>
## 1 (Intercept)    6.75      0.248     27.2 1.54e-68
## 2 TV             0.0191    0.00150     12.7 2.36e-27
## 3 Radio          0.0289    0.00891      3.24 1.40e- 3
## 4 TV:Radio       0.00109   0.0000524    20.7 2.76e-51
```

Adjustment
In Slope

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic  p.value    df logLik  AIC  BIC
##   <dbl>      <dbl> <dbl>    <dbl>    <dbl> <dbl> <dbl> <dbl> <dbl>
## 1    0.968      0.967 0.944    1963. 6.68e-146     3  -270.  550.  567.
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

$$\text{Model2: } \hat{E} = 6.75 + 0.019TV + 0.029Radio + 0.001TV \times Radio$$

$$\hat{E} = 6.75 + (0.019 + 0.001Radio) \times TV + 0.029Radio$$

$$\hat{E} = 6.75 + 0.019TV + (0.029 + 0.001TV) \times Radio$$

Example 3: Predictions

- Gathering Predictions

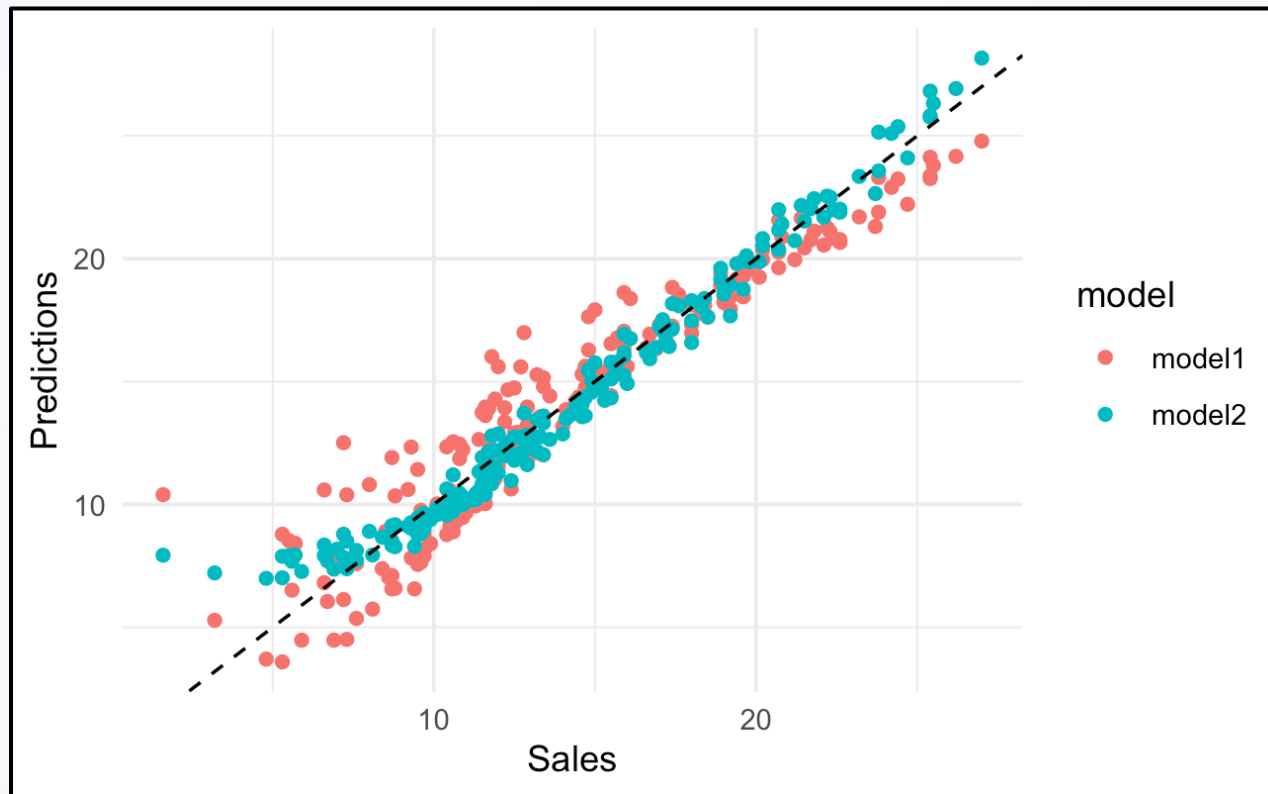
```
`` {r}
Ad %>%
  gather_predictions(model1,model2)%>%
  glimpse()
``
```

Rows: 400
Columns: 5
\$ model <chr> "model1", "model1", "model1", "model1", "model1", "...
\$ TV <dbl> 230.1, 44.5, 17.2, 151.5, 180.8, 8.7, 57.5, 120.2, ...
\$ Radio <dbl> 37.8, 39.3, 45.9, 41.3, 10.8, 48.9, 32.8, 19.6, 2.1...
\$ Sales <dbl> 22.1, 10.4, 9.3, 18.5, 12.9, 7.2, 11.8, 13.2, 4.8, ...
\$ pred <dbl> 20.555465, 12.345362, 12.337018, 17.617116, 13.2239...

200 Predictions for 2 Models

Example 3: Visualization

- Visualizing Prediction vs. True Value



Example 3: Summary

- Summary for Lectures on Categorical Predictor and Interactions
 - Numerical Response Variable
 - Categorical Predictor
 - Interaction between Two Categorical Predictors
 - Interaction between Two Categorical and Numerical Predictor
 - Interaction between Two Numerical Predictors