Goals for the Week
Categorical Predictors
Multiple Regression
Conditions and Assumptions
Evaluating Models

STAT 140: Week 13

Statistical Toolbox

- One-proportion z-test, confidence interval
- Two-proportion z-test, confidence interval
- One-sample t-test, confidence interval
- Two-sample t-test, confidence interval
- Simple Linear Regression

My hope is that after Week 13 (11/22-11/26), you will be able to:

- Build regression models based on categorical predictors.
- Build multiple linear regression models using categorical and numerical predictors.
- Check assumptions for multiple linear regression.
- Use R^2 and adjusted R^2 to evaluate the strength of a regression model.

Categorical Predictors with Two Levels 1

- We may be interested in predicting a response based on a categorical predictor.
- To incorporate a categorical variable into a linear model, we use an **indicator variable**.
  - Indicator variables are categorical variables that have been coded as 0 or 1.
- Using an indicator variable in place of a category name allows for these variables to be directly used in a regression!
- Example: We will continue using the `RailTrail` dataset from the Pioneer Valley Planning Commission (PVPC) to illustrate concepts. `RailTrail` includes columns for `hightemp`, `lowtemp`, `avgtemp`, `spring`, `summer`, `fall`, `cloudcover`, `precip`, `volume`, `weekday`, and `dayType`.
- We will predict the number of Rail Trail users using the `weekday` variable.
  - `FALSE` if not weekday (weekend), `TRUE` if weekday
  - Remember that `FALSE` is equivalent to 0 and `TRUE` is equivalent to 1!

Categorical Predictors with Two Levels 2


```
head(RailTrail$weekday)
```

```
## [1]  TRUE  TRUE  TRUE FALSE  TRUE  TRUE
```

```
weekday_model <- lm(volume ~ weekday, data = RailTrail)
weekday_model
```

```
##
## Call:
## lm(formula = volume ~ weekday, data = RailTrail)
##
## Coefficients:
## (Intercept)  weekdayTRUE
##      430.71       -80.29
```

Categorical Predictors with Two Levels 3

- The estimated intercept if the value of the response variable for the first category (i.e., the category corresponding to an indicator variable of 0 and known as the **baseline** or **reference level**).
  - We expect 430.71 Rail Trail users during a weekend.
- The estimated slope is the average change in the reponse variable between the two categories.
  - On average, there were 80.29 fewer Rail Trail users during weekdays.
  - Overall, we expect 350.42 (or 430.71-80.29) Rail Trail users on weekdays.

Indicator Variables: Two Levels

- Now that we have learned how to incorporate categorical variables with two levels into a linear regression model, can we incorporate categorical variables with more than two levels?
- Consider

$$\widehat{volume} = b_0 + b_1 \times Season$$

- In the Rail Trail data set, we have observations for spring, summer, and fall... what do we do?
- Can we code the levels as 0, 1, 2, ... ?

Indicator Variables: Multiple Levels 1

- This enforces too much structure on your data! You would be assuming the difference between spring and summer is the same as the difference between summer and fall.

```
head(RailTrail$season)
```

```
## [1] 1 1 0 1 0 0
```

```
season_model1 <- lm(data = RailTrail, volume ~ season)
summary(season_model1)$coefficients
```

```
##             Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 384.3984   16.88781 22.7618884 1.221493e-38
## season      -16.5277   18.75138 -0.8814129 3.804955e-01
```

Indicator Variables: Multiple Levels 2


```
summary(season_model1)$coefficients
```

```
##             Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 384.3984   16.88781 22.7618884 1.221493e-38
## season      -16.5277   18.75138 -0.8814129 3.804955e-01
```

- Here, we are saying that every time the season changes (every time you add 1), the expected number of riders will decrease by about -16.5277.
- However, this increase is not significant (p-value = 0.380), and this interpretation doesn't really make sense!

Indicator Variables: Multiple Levels 3

- When fitting a regression model with a categorical variable that has k levels where $k > 2$, we should incorporate coefficients for $k - 1$ of those levels.
- The last level that does not receive a coefficient is known as the **reference level**.
- Coefficients for the other levels are all considered relative to the reference level.
- Example:

$$\widehat{volume} = b_0 + b_1 \times spring + b_2 \times summer$$

Rail Trail Example: Three Levels 1


```
head(RailTrail$season)
```

```
## [1] summer summer spring summer spring spring
## Levels: fall spring summer
```

```
season_model2 <- lm(data = RailTrail, volume ~ season)
summary(season_model2)$coefficients
```

```
##                Estimate Std. Error  t value     Pr(>|t|)
## (Intercept)  295.08333   35.47431 8.318226 1.113315e-12
## seasonspring  76.57704   39.28554 1.949243 5.448811e-02
## seasonsummer 126.79667   43.15636 2.938076 4.227037e-03
```

- What is the reference level for season?
- What can we say about the expected number of riders in spring? In summer?

Rail Trail Example: Three Levels 1


```
summary(season_model2)$coefficients
```

```
##                Estimate Std. Error  t value     Pr(>|t|)
## (Intercept)  295.08333   35.47431 8.318226 1.113315e-12
## seasonspring  76.57704   39.28554 1.949243 5.448811e-02
## seasonsummer 126.79667   43.15636 2.938076 4.227037e-03
```

- The reference level is `fall`–notice how there is no row describing the coefficient on `fall` in the summary! The intercept, 295.1, describes how many users we expect in the fall.
- The expected number of users in the spring is 76.6 more than in the fall (371.7). This difference is nearly significant at the 5% level (p-value = 0.054).
- The expected number of users in the supper is 126.8 more than in the fall (421.9). This difference is significant at the 5% level (p-value = 0.004).

Multiple Regression 1

- So far, we have looked at simple linear regression, which models a response variable based on a single explanatory variable.
- Multiple linear regression, or multiple regression, models a response variable based on **more than one** explanatory variable.

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

when there are k predictors.

- This method is motivated by scenarios where many variables may be simultaneously connected to an output.
- **We always estimate the β_i parameters using statistical software!**

Multiple Regression 2

- Explanatory variables in multiple linear regression can be categorical or numerical! You can also mix the different types of variables.
- Most of the things we know about simple linear regression can be easily extended to multiple regression.
  - You can still predict the volume of riders on a given day.
  - You can still calculate a residual for a given observation.
  - You can still interpret the parameters.
  - You can still build confidence intervals and run hypothesis tests.

Rail Trail Example: Multiple Regression

$$\widehat{volume} = \beta_0 + \beta_{HT}\, hightemp + \beta_W\, Week\ Day$$

```
model1 <- lm(data = RailTrail, volume ~ hightemp + weekday)
summary(model1)$coefficients
```

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  42.806573 64.3444307  0.6652724 5.076361e-01
## hightemp      5.347817  0.8462988  6.3190648 1.086317e-08
## weekdayTRUE -51.553496 23.6744270 -2.1776027 3.214226e-02
```

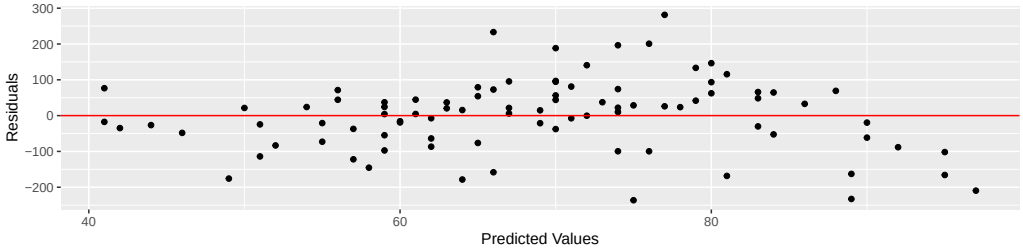
- Since their p-values are small, both the high temperature and whether the day is a weekday are considered to significantly effect the number of Rail Trail users.

Conditions for the Least Squares Line

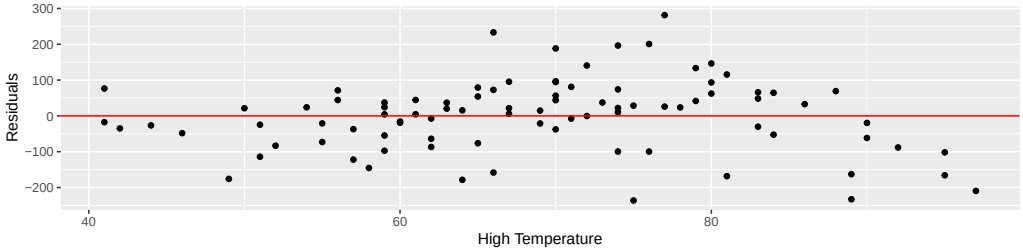
- In general, we refer to the plots used to check assumptions and conditions for linear regression as **diagnostic plots**.
- Checking the diagnostic plots for multiple linear regression is not exactly the same as for simple linear regression, but is very similar.
  - **Linearity**: Check the residuals plot to ensure that there are no patterns.
  - **Nearly Normal Residuals**: Check a histogram of residuals (most likely violation: outliers, leverage points, or influential points).
  - **Constant Variability**: Check the initial scatterplot (most likely violation: "wedge-shaped" pattern).
  - **Independent Observations**: Observations should be independent from each other (most likely violation: sequentially collected, or time-series, data. Check a plot of the residuals plotted in order of data collection).

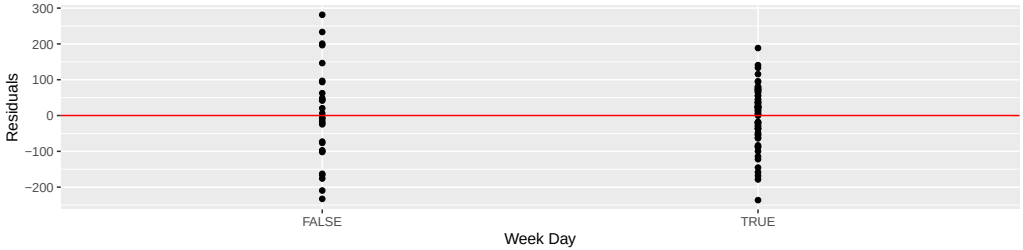
Linearity 1

- In simple linear regression, we check the residuals plotted against the predicted values to ensure that there are no patterns.
- What happens when you have more than one x value?
  - Check the residuals plotted against the predicted/fitted values
  - Check a residual plot for each predictor.
- **These plots can also be used to check constant variability!**



Linearity 2


Residuals
300
200
100
0
−100
−200
40
60
80
High Temperature


Residuals
300
200
100
0
−100
−200
FALSE
TRUE
Week Day

Diagnostic Plots

- Having reviewed the diagnostic plots, there are two options:
  - If we are not concerned about any observed issues, use this as the final model. If there are any abnormalities here, we should still mention them!
  - Try and improve the model fit by using more complicated methods–some of which are covered in STAT-242.
- Don't report results when conditions are grossly violated! If the conditions are very clearly violated, a new model using different techniques and techniques beyond those we have learned in this class might be better.

Evaluating Models

- Both of the models with one variable have significant slopes, but both of the slopes are significant in the model with two variables.
- How can we tell if the model with `hightemp` and `weekday` is better than the model with `hightemp` alone? `weekday` alone?
- We can talk about the R^2 associated with the models!

R^2 1

- We evaluated the strength of the linear relationship between two variables earlier using the correlation, R.
- It is more common to explain the strength of a linear fit using R^2, or **R squared**.
- The R^2 of a linear model describes the amount of variation in the response variable that is explained by the least squares line.
  - When there is only one predictor in the model, saying the variation can be explained by the predictor is the same as saying the variation can be explained by the least squares line.
- Example: 33.9% (0.583^2) of variation in the estimated number of Rail Trail users can be explained by the day's high temperature (or, 33.9% of variation in the estimated number of Rail Trail users can be explained by least squares line).

R^2 2

- When there are multiple predictors in a model, saying the variation can be explained by the least squares line is the same as saying the variation can be explained by all of the predictors together.
- A good model will explain a lot of the variation, and thus have a high R^2.

```
hightemp_model <- lm(data = RailTrail, volume ~ hightemp)
summary(hightemp_model)$r.squared
```

```
## [1] 0.33939
```

```
summary(weekday_model)$r.squared
```

```
## [1] 0.08600583
```

```
summary(model1)$r.squared
```

```
## [1] 0.3735356
```

Adjusted R^2 1

- R^2 by itself is acceptable when there is just a single variable, but is less helpful when there are many variables.
- Consider a scenario in which we start to add a lot of extra things to our model...
  - We could add reasonable things, like the cloud cover.
  - Eventually, we would run out of reasonable things, and start adding irrelevant terms, such as the score of the Red Sox game the night before.
- If you add a predictor to the model, R^2 will always increase, even if the predictor doesn't make sense.

Adjusted R^2 2

- We can use a new quantity, known as the adjusted R^2.

$$R^2 = 1 - \frac{\text{variability in residuals}}{\text{variability in the outcome}} = 1 - \frac{Var(e_i)}{Var(y_i)}$$

$$R^2_{adj} = 1 - \frac{s^2_{residuals}/(n-k-1)}{s^2_{outcome}/(n-1)} = 1 - \frac{s^2_{residuals}}{s^2_{outcome}} \times \frac{n-1}{n-k-1}$$

- The adjusted R^2 is adjusting for the number of predictors in your model.

Rail Trail Example: Adjusted R^2


```
RailTrail_model <- lm(data = RailTrail, volume ~ hightemp)
summary(RailTrail_model)$adj.r.squared
```

```
## [1] 0.3318831
```

```
model1 <- lm(data = RailTrail, volume ~ hightemp + weekday)
summary(model1)$adj.r.squared
```

```
## [1] 0.3591341
```

- The adjusted R^2 is always smaller than R^2.
- However, even after adjusting for the extra predictor, the bigger model seems to be better!

Model Selection

- The best model is not always the most complicated!
- Sometimes, including variables that are not evidently important can actually reduce the accuracy of predictions.
- Model selection strategies help us eliminate variables from the model that are found to be less important.
- It's common to refer to these models as **parsimonious**.

Warning

- While we remain cautious about making any causal interpretations using multiple regression on observational data, such models are a common first step in gaining insights or providing some evidence of a causal connection.