Goals for the Week

Linear Regression: Review

Residual Plots

Conditions for Linear Models

Interpretation and Extrapolation

Inference for Linear Models

STAT 140: Week 12

Scheduling 1

- Week 12 (11/15-11/19): More on Linear Regression
- Week 13 (11/22 - 11/26): Multiple Linear Regression
  - Video Lecture rather than in-person–there is NO NEED to be in class on Monday, November 21.
  - Progress Check 12/13 due Monday, 11/29
  - Homework 12, Journal 12, due Wednesday, 12/01
- Week 14 (11/29 - 12/03)
  - Monday: Practice for Week 13 material
  - Wednesday: Week 13 Catch Up
  - Friday: Project Work Day (work completed will count towards HW14)
- Week 15 (12/06): TBD, but either final review or project work day.
  - Homework 13, Journal 13 due Monday, 12/06
  - Homework 14, Journal 14 due Monday, 12/06

Scheduling 2

Outstanding Assignments:

- PC: 12/13 (combined)
- Journal: 12, 13, 14
- HW 12, 13, 14*
- Final Project
- Final Exam

My hope is that after Week 12 (11/15-11/19), you will be able to:

- Build a simple linear regression model.
- Interpret the coefficients of a simple linear regression model.
- Check the conditions for a simple linear regression model.
- Make inferences regarding coefficients of a simple linear regression model.
- Identify outliers, leverage points, and influential points.
- Generate and interpret R output for a simple linear regression model.
- Create residual plots using `ggplot2`.

Linear Regression: Review

- **Simple linear regression** is the statistical method of fitting a line to data, where the relationship between **two** variables, x and y, can be modeled by a straight line (also known as the least squares line).

$$y = \beta_0 + \beta_1 x$$

- When we use x to predict y, we usually call x the explanatory or **predictor** variable and y the response.
- The values β_0 and β_1 represent the model's parameters.
- The parameters are estimated using data, and we write their point estimates as b_0 and b_1.

Rail Trail: Data 1

- The Pioneer Valley Planning Commission (PVPC) collected data north of Chestnut Street in Florence, MA for ninety days from April 5, 2005 to November 15, 2005.
- Data collectors set up a laser sensor, with breaks in the laser beam recording when a rail-trail user passed the data collection station. By counting the breaks, you can estimate how many people are using the trail during each day when data was collected.

We will be using this example dataset for the rest of the semester–make sure you are familiar with it!

Rail Trail: Data 2

`RailTrail` is a data frame with 90 observations on the following variables:

- `hightemp`: daily high temperature (in degrees F)
- `lowtemp`: daily low temperature (in degrees F)
- `avgtemp`: average of daily low and daily high temperature (in degrees F)
- `spring`: indicator of whether the season was Spring
- `summer`: indicator of whether the season was Summer
- `fall`: indicator of whether the season was Fall
- `cloudcover` measure of cloud cover (in oktas)
- `precip` measure of precipitation (in inches)
- `volume` estimated number of trail users that day (number of breaks recorded)
- `weekday` logical indicator of whether the day was a non-holiday weekday
- `dayType` one of "weekday" or "weekend"

Rail Trail: Linear Model 1

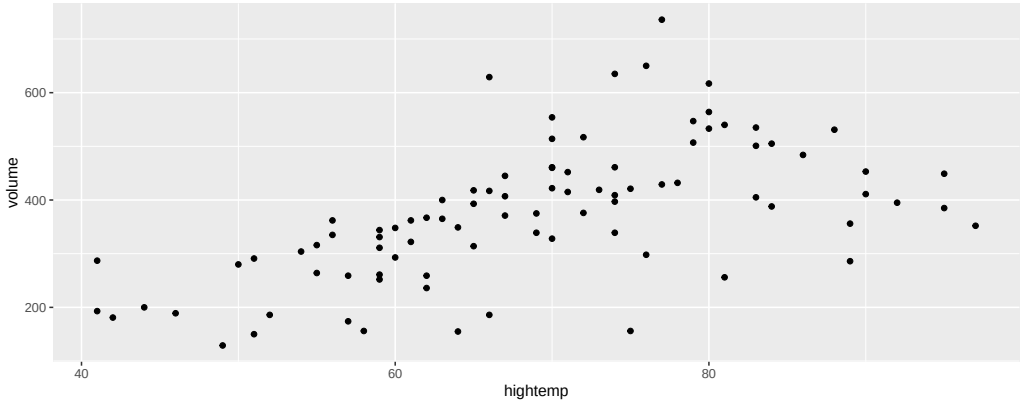

```
library(mosaicData)
colnames(RailTrail)
```

```
##  [1] "hightemp"   "lowtemp"    "avgtemp"    "spring"     "summer"
##  [6] "fall"       "cloudcover" "precip"     "volume"     "weekday"
## [11] "dayType"
```

- Let X, the predictor variable, be `hightemp`, the daily high temperature for a set of 90 randomly selected days.
- Let Y, the response variable, be `volume`, the number of trail users that day.
- Recall that even if none of the points falls exactly on the line, it is still useful to describe the relationship using a line.

Rail Trail: Linear Model 2

Comment on the relationship between `hightemp` and `volume`, mentioning the form, direction, strength, and unusual values.



Rail Trail: Linear Model 3

- This graph is more or less linear (past 80 degrees, it starts to curve downwards again).
- The direction is mostly positive (again, past 80 degrees, the relationship starts to become negative).
- The relationship is fairly strong.
- There is maybe one unusual value (around 66 degrees with 600 riders).

Rail Trail: Linear Model 4

- The parameters of the linear regression model, β_0 and β_1, are estimated using observed data.
- These estimates are known as b_0 and b_1.
- In practice, this estimation is done using a computer.
  - In R, we use the `lm()` command and a formula linking the two variables.

```
RailTrail_model <- lm(volume ~ hightemp, data = RailTrail)
```

Rail Trail: Linear Model 5

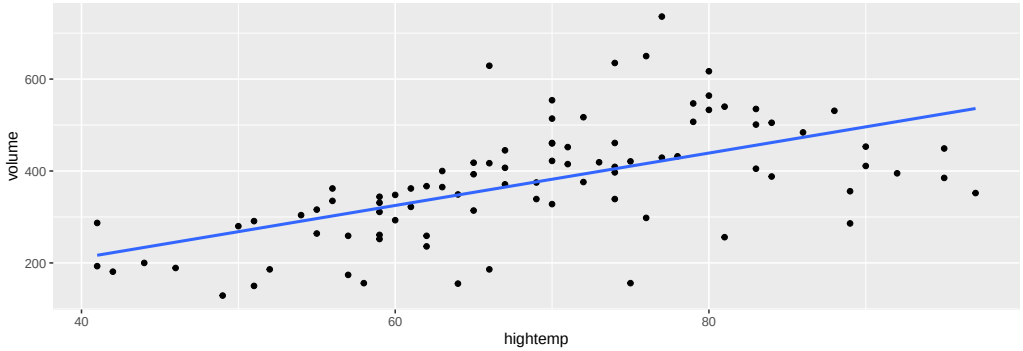

```
RailTrail_model
```

```
##
## Call:
## lm(formula = volume ~ hightemp, data = RailTrail)
##
## Coefficients:
## (Intercept)     hightemp
##     -17.079        5.702
```

It may be helpful to visualize these numbers by adding a line to the scatterplot between the two variables.

Rail Trail: Adding the Line to a Plot


```
ggplot(data = RailTrail, aes(x = hightemp, y = volume)) +
  geom_point() +
  geom_smooth(method = "lm", se = F)
```



Rail Trail: Linear Model 6

- We can describe the linear relationship between the number of trail users each day and the day's high temperature using the equation:

$$\widehat{volume} = -17.10 + 5.70 \times hightemp$$

Prediction and Residuals

Remember that we can use this linear relationship to predict the value of the response variable at a certain value of the explanatory variable.

- For example, we might want to predict the average number of people on the Rail Trail on a 60 degree day.

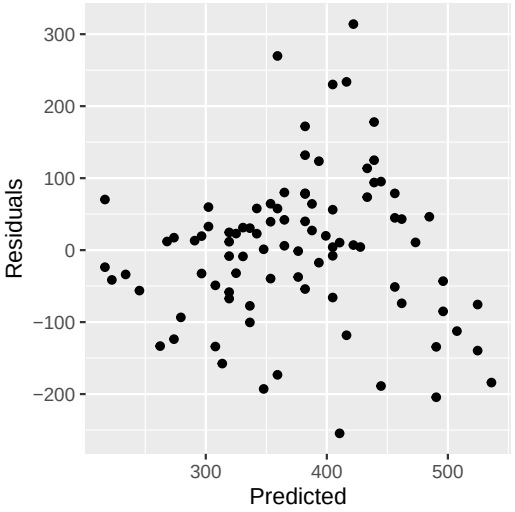
$$\widehat{volume} = -17.10 + 5.70 \times 60 = 324.9$$

Remember also that **residuals** are the leftover variation in the data after accounting for model fit.

- Each observation will have a residual. For example, if we expected 324.9 people to use the Rail Trail on a 60 degree day and only 293 people actually used it, the residual is $\hat{e} = 293 - 324.9 = -31.9$.

Rail Trail: Residual Plots 1

- The equations for calculating slope and intercept are based on minimizing the sum of the squared residuals.
- Additionally, residuals are helpful in evaluating if a linear model is appropriate for the data.
- We often display them in a residual plot, which displays the predicted values on the x-axis and the residuals on the y-axis (note: this is slightly different than what your textbook says, but the principle is the same).

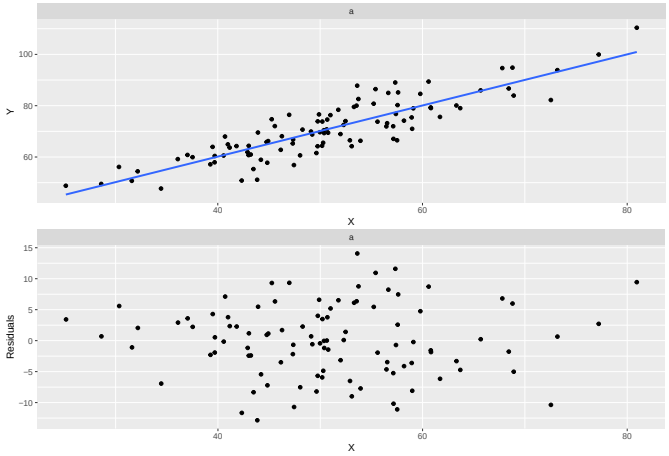


Rail Trail: Residual Plots 2

- To evaluate how well a *linear* model fits a data set, we can identify characteristics or patterns still apparent in data after fitting a model.
  - If there is a pattern, a linear model is *not* appropriate for the data.
  - We want to see a random scatter or cloud of residuals with no pattern.

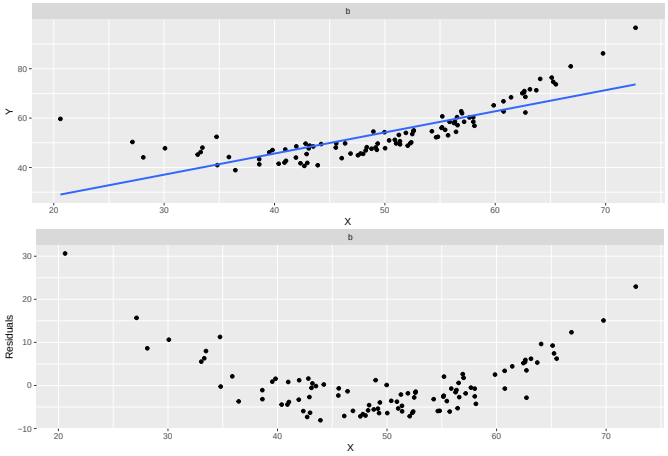
Residual Plots: Other Examples 1

- The original scatterplot is shown on the top, and the residual plot is on the bottom.
- This strong, linear relationship is a good example of a linear model that is appropriate and fits the data well.



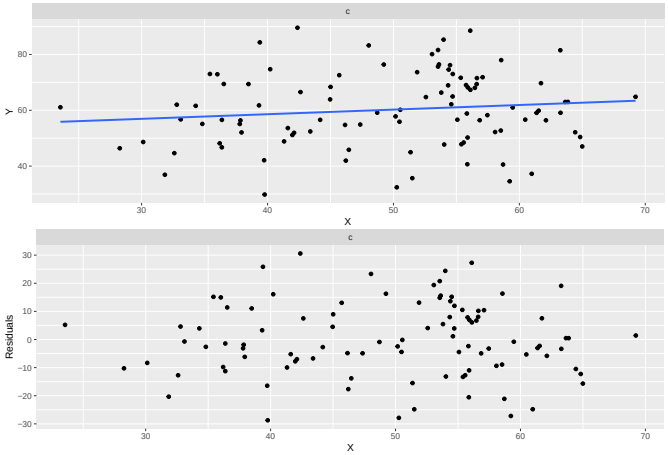
Residual Plots: Other Examples 2

- The original scatterplot is shown on the top, and the residual plot is on the bottom.
- This strong, quadratic relationship is a good example of a linear model that fits the data well, but is not appropriate (look at the patterns in the residual plot!).



Residual Plots: Other Examples 3

- The original scatterplot is shown on the top, and the residual plot is on the bottom.
- This weak linear relationship is a good example of a linear model that is appropriate, but does not fit the data very well (the points are far from the line).



Conditions for Linear Models 1

- When fitting a linear model, we should check our data to see if the following conditions hold.
  - If **ALL** of the conditions hold, a linear model is appropriate.
  - If **ANY** of the conditions do not hold, the linear model is not appropriate!!
- Before we do anything else, we need to check the conditions for the model.

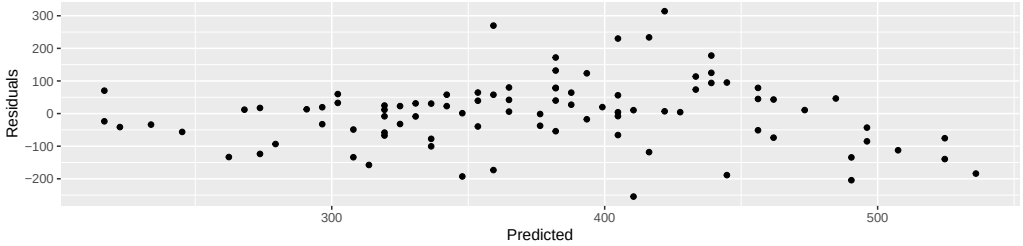
Conditions for Linear Models 2

- **Linearity**: The data should show a linear trend. If there is a nonlinear trend, an advanced regression method should be used.
  - Typically, we check this condition using the residual plot. If there is a pattern, the linear model is not appropriate.
- **Nearly Normal Residuals**: The residuals must be nearly normal.
  - Generally, we check this condition with a histogram. When this condition is found to be unreasonable, it is usually because of unusual values.
- **Constant Variability**: The variability of points around the least squares line remains roughly constant.
  - Generally, we check this condition with a scatterplot. The most common pattern observed when this condition fails is increasing variability, or larger variability of y when x is large.
- **Independent Observations**: Observations should be independent from each other. Be careful about applying regression to **time series** data, which are sequential observations.

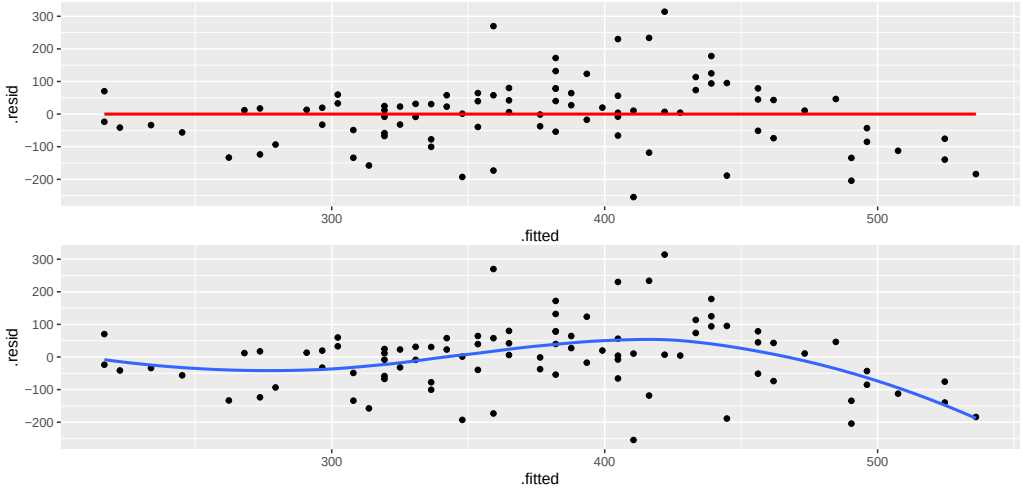
Rail Trail: Linearity? 1

Are our data linear?

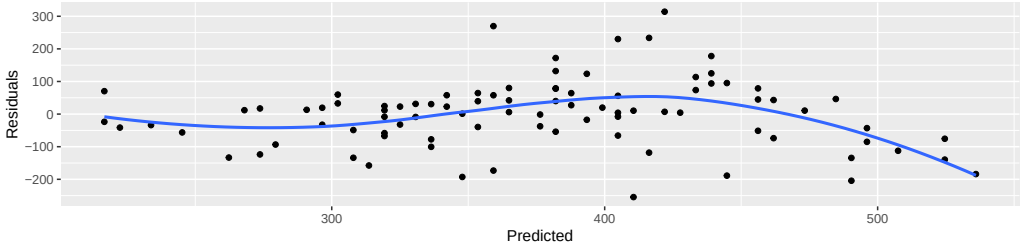
```
ggplot(data = RailTrail_model, aes(x = .fitted, y = .resid)) +
  geom_point() +
  xlab("Predicted") +
  ylab("Residuals")
```



Rail Trail: Linearity? 2



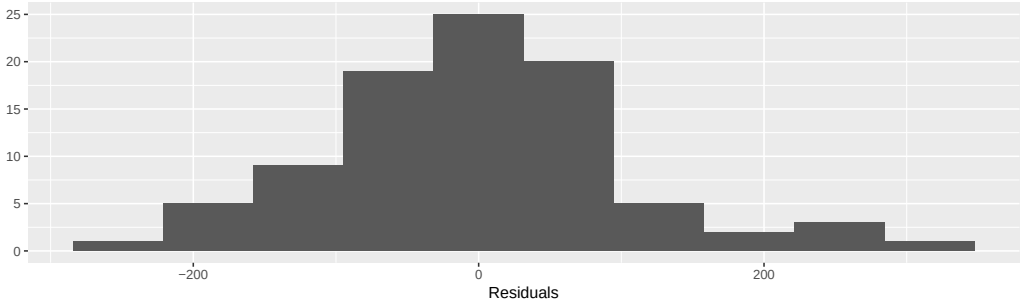
Rail Trail: Linearity? 3



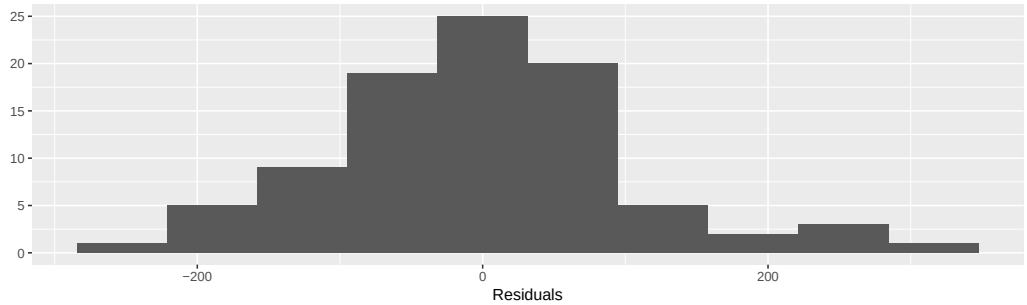
- A linear model might not be the best for this data—there are lots of negative values for very small or very large values of high temperature, and lots of positive values for values of high temperature in the middle.
- We will proceed, with caution!

Rail Trail: Nearly Normal Residuals? 1

Are the residuals normal?



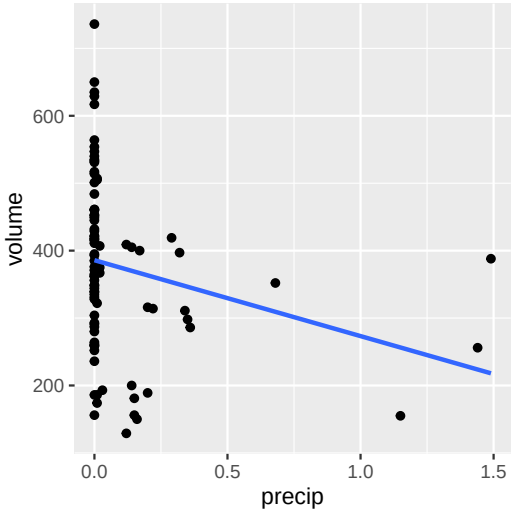
Rail Trail: Nearly Normal Residuals? 2



- The histogram is nearly normal and centered at approximately 0. Residuals range from -250 to 300.
- There do not appear to be any outliers.

Rail Trail: Unusual Values

- Recall that the most common violation of nearly normal residuals is the presence of unusual values.
- Let's look at a different example to understand the different types of unusual values, known as outliers, high leverage points, and influential points–this time, we will predict the volume of Rail Trail users (`volume`) from the total amount of precipitation (`precip`).

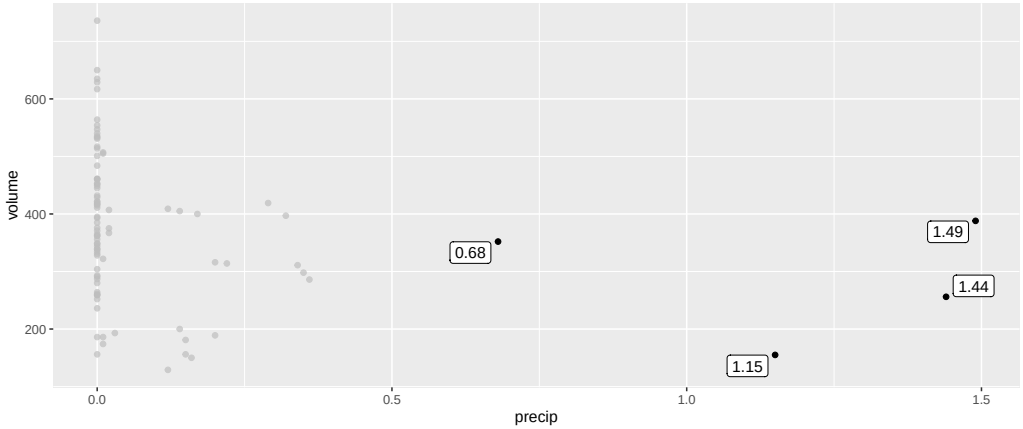


Outliers and High Leverage Points

- **Outliers** in regression are points that fall far away from the cloud of points.
  - These points are especially important because they can have a strong influence on the least squares line!
- Points with **high leverage** are points that fall horizontally away from the center of the cloud.
  - High leverage points tend to "pull harder" on the line.
  - High leverage points can strongly influence the slope of the least squares line.
- Some outliers are high leverage points, but all high leverage points are outliers.

Rail Trail: High Leverage

These points are all high leverage!



Influential Points

- **Influential Points** are high leverage points that actually invoke influence on the slope of the line.
- If we had fitted the line without a potential influential point AND the point is unusually far away from the line, it can be called influential.
  - Influential points could change a strong relationship to a weak relationship.
  - Influential points could change a weak relationship to a strong relationship.

Rail Trail: Influential Points 1

- Let's consider removing the point (1.15, 155).
- I can identify the location of this point using `which()`.

```
which(RailTrail$precip > 0.50)
```

```
## [1] 17 26 27 65
```

```
RailTrail[26,]
```

```
##    hightemp lowtemp avgtemp spring summer fall cloudcover precip v
## 26       81      65      73      0      1    0          10   1.44
##    dayType labelprecip  leverage
## 26 weekday        1.44 0.3058375
```

Rail Trail: Influential Points 2

Here is the model with all of the points.

```
(model1 <- lm(data = RailTrail, volume ~ precip))
```

```
##
## Call:
## lm(formula = volume ~ precip, data = RailTrail)
##
## Coefficients:
## (Intercept)       precip
##       385.8       -112.6
```

```
cor(RailTrail$precip, RailTrail$volume)
```

```
## [1] -0.232384
```

Rail Trail: Influential Points 3

- Let's remove row 27 to see how the new line behaves.

```
RailTrail_rm <- RailTrail[-27,]
(model2 <- lm(data = RailTrail_rm, volume ~ precip))
```

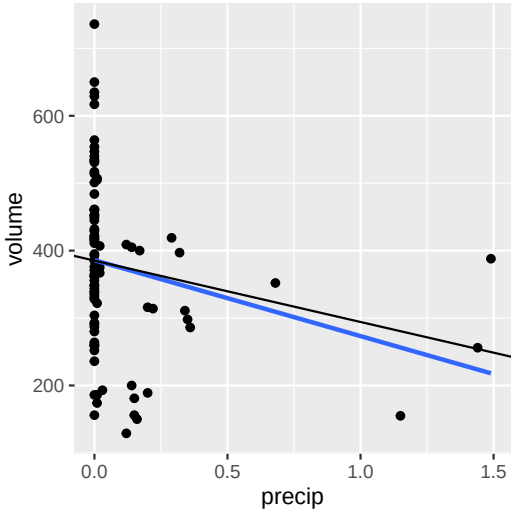
```
##
## Call:
## lm(formula = volume ~ precip, data = RailTrail_rm)
##
## Coefficients:
## (Intercept)       precip
##      385.22       -91.04
```

```
cor(RailTrail_rm$precip, RailTrail_rm$volume)
```

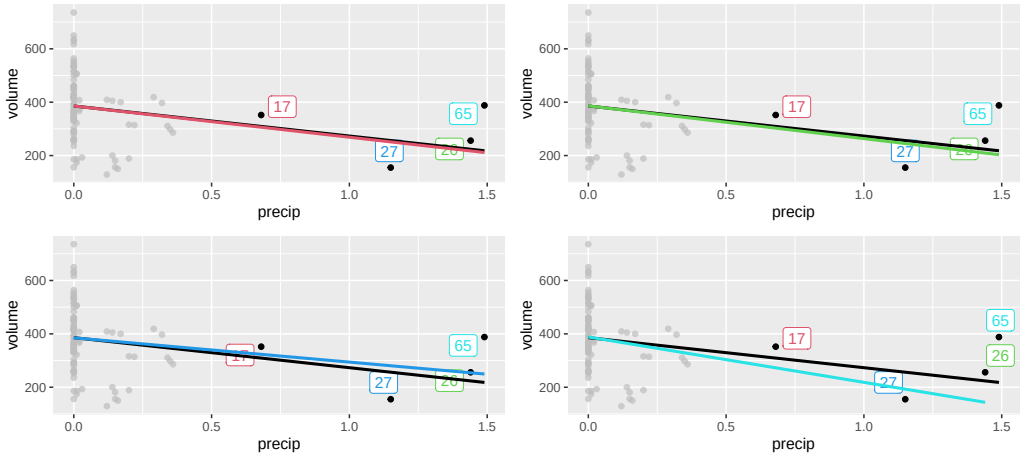
```
## [1] -0.1727494
```

Rail Trail: Influential Points 4

- Once this point is removed, the correlation goes from -0.23 to -0.17, and the model coefficients also change quite a bit.
- We can also look at the effect this point has visually. Since the new line (black) has a flatter slope, I would say the point is pulling the original line (blue) down.
- We would call this an influential point.



Rail Trail: Influential Points 5

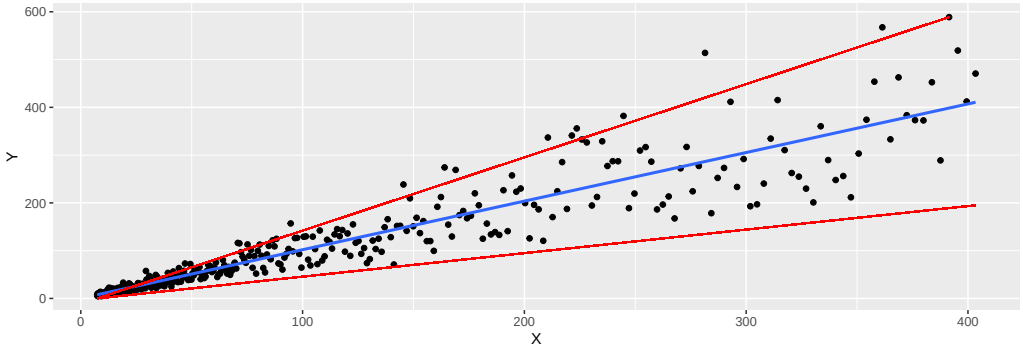


Removing Outliers

- "It is tempting to remove outliers. Don't do this without a **very good reason**."
  - Outliers are often the most interesting cases!
  - Example: large market swings
- What is an example of a very good reason?
  - Women's Health Dataset: Weights recorded in pounds for women over 18
  - Why is there a value of 19 pounds? Should it be 119? 190? If I don't know how to fix it and I know for sure it is a typo, I can get rid of it.
- Outliers are supposed to be **rare**. If outliers comprise a large portion of your data, they are not outliers!
- Analyze the data with and without the outlier–if your analysis does not change, report the analysis with the outlier. If it does change, report both.

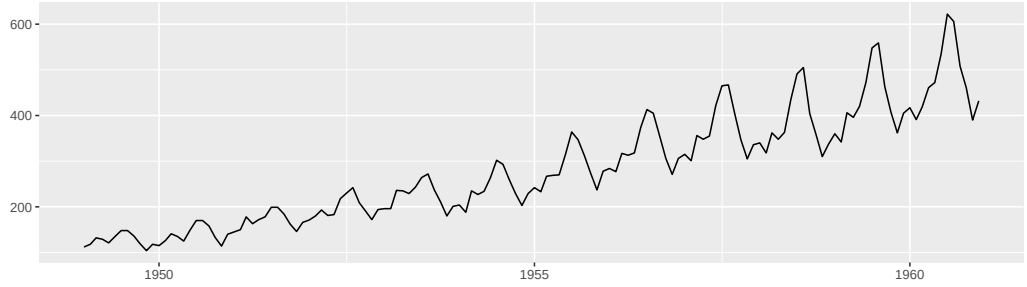
Example: Constant Variability

- The most likely violation of constant variability is a wedge-shaped pattern!



Example: Independent Observations

- The most likely violation of independence is data collected sequentially!
- Example: Airline Passenger Data from 1949-1961–we can see that there is an overall increase as well as a seasonal trend (people tend to fly a lot in the summer and near the holidays, but not as much in between).



Interpreting Parameter Estimates

Once we have confirmed that a linear model is appropriate, we can interpret the estimates of the parameters.

- The slope of a linear model can be interpreted as the expected increase or decrease in the Y variable that would occur if the explanatory variable X increased by one unit.
- The intercept of a linear model can be interpreted as the expected value of the Y variable if the explanatory variable X were 0.

Rail Trail: Interpretation 1

- If the high temperature increases by one degree, we can expect 5.702 additional people to use the Rail Trail.
- When the high temperature is 0 degrees, we can expect -17.079 people to use the Rail Trail.
  - Does this intercept make sense?
  - Why doesn't the intercept make sense?

Rail Trail: Extrapolation 1

- Our linear model is not valid all the way to $X = 0$.
- Applying a model estimate to values outside the realm of the original data is called **extrapolation**.
- Example:

```
summary(RailTrail$hightemp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   41.00   59.25   69.50   68.83   77.75   97.00
```

The minimum value of `hightemp` is 41 degrees! 0 degrees is well outside this range and therefore, we should not consider the intercept as valid.

```
?RailTrail
```

Rail Trail: Extrapolation 2

- "The collectors left the motion detector out during the winter, but because the counter drops data when the temperature falls below 14 degrees Fahrenheit, there is no data for the cold winter months."
- Because we do not have any data collected for days when the high temperature was less than 41, so we cannot extrapolate to anything below that range!
- Linear regression is simply a modeling framework, and the truth is almost always more complex than our simple line.

Parameters and Statistics

Let's refresh our memories–so far, we have learned about the following parameters and their equivalent statistics.

	Population	Sample
Mean	μ	$\bar{x}$
Standard Deviation	σ	s
Proportion	p	$\hat{p}$
Rate	λ	–
Correlation	–	R
Intercept	β_0	b_0
Slope	β_1	b_1

Inference for Linear Regression

- We can use a z-interval or test to draw conclusions about a proportion, p, and we can use a t-interval or test to draw conclusions about a mean, μ.
- Let's extend the simple linear regression tools we learned earlier so that we can draw conclusions about the population slopes and intercepts, β_0 and β_1.
- We've talked a lot about uncertainty and error in this class.
- Just like there is variability in a sample mean or proportion, there is variability in the point estimates of the slope and intercept.
- What is the variability of a statistic called?

Confidence Intervals 1

- The variability of a statistic is called the Standard Error!
- We can use the standard errors and estimates of the slope and intercept to draw conclusions about the real slopes and intercept using confidence intervals and hypothesis tests.
- Let's say we want a confidence interval for the slope of a linear regression model.
  - What do we need to build a confidence interval?

Confidence Intervals 2

- To build a confidence interval, we need a point estimate, standard error, and some z^* or t^*.
- We can use R to find a point estimate for the slope of a simple linear regression model–the point estimate is b_1.
- It turns out that we need a t^*_{df} to build the interval. The formula is:

$$point\ estimate \pm t^* \times SE_{slope}$$

- What are the degrees of freedom for t^*?

```
RailTrail_model$df.residual
```

```
## [1] 88
```

Confidence Intervals 3

- How can we find the standard error?

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

- Use the `summary()` function on the different models!
- Remember `lm()` is a very powerful function–use `attributes()` to explore all of the different things packed into a linear model.
- What does the first column represent? The second column?

Confidence Intervals 4


```
summary(RailTrail_model)$coefficients
```

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) -17.079281 59.3953040 -0.2875527 7.743652e-01
## hightemp      5.701878  0.8480074  6.7238541 1.705138e-09
```

- The first column represents estimates of the intercept (b_0) and slope (b_1).
- The second column represents the standard errors for b_0 and b_1, which we can use to build confidence intervals.

Rail Trail: Confidence Intervals


```
slope <- 5.701878
se_slope <- 0.8480074
tstar <- qt(0.975, df = RailTrail_model$df.residual)
slope - tstar*se_slope
```

```
## [1] 4.016641
```

```
slope + tstar*se_slope
```

```
## [1] 7.387115
```

- We are 95% confident that the slope for predicting the day's RailTrail volume from the high temperature is between 4.01 and 7.39.

Hypothesis Testing 1

- To formally test if there is a relationship between two variables, we check to see if the slope is equal to zero.
  - $H_0 : \beta_1 = 0$
  - $H_A : \beta_1 \neq 0$
- We would reject the null in favor of the alternative if the data provide strong evidence that the true slope parameter is different than zero.
- To assess the hypotheses, we use the standard error for the estimate to compute an appropriate test statistic and identify the p-value.

Hypothesis Testing 2

- The generic form for the test statistic is

$$t = \frac{\text{point estimate} - \text{null value}}{SE}$$

- In a hypothesis test for the slope, the test statistic is

$$t = \frac{b_1 - 0}{SE}$$

- This corresponds to the `t value` (third) column in the R output!

Rail Trail: Hypothesis Testing 1


```
summary(RailTrail_model)$coefficients
```

```
##               Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) -17.079281 59.3953040 -0.2875527 7.743652e-01
## hightemp      5.701878  0.8480074  6.7238541 1.705138e-09
```

- The associated p-values are also in the R output (fourth column, `Pr(>|t|)`).
- Conclude:
  - If the p-value is less than α, we reject the null hypothesis. There is significant evidence showing that the response variable is related to the explanatory variable.
  - If the p-value is greater than α, we fail to reject the null hypothesis. There is not sufficient evidence showing that the response variable is related to the explanatory variable.

Rail Trail: Hypothesis Testing 2

- Since the p-value for the slope on `hightemp` is less than 0.05 (or any signficance level), we reject the null hypothesis.
- There is significant evidence of a positive linear relationship between the day's high temperature and the volume of riders on the Rail Trail (which we could have guessed from the confidence interval).

Inference for Regression

- Warning:

"We usually rely on statistical software to identify point estimates, standard errors, test statistics, and p-values in practice. However, be aware that software will not generally check whether the method is appropriate, meaning we must verify conditions are still met."

Statistical Toolbox, part 2

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