STAT011 Statistical Methods I

Lecture 23 Simple Linear Regression II

Lu Chen
Swarthmore College
4/23/2019

Review

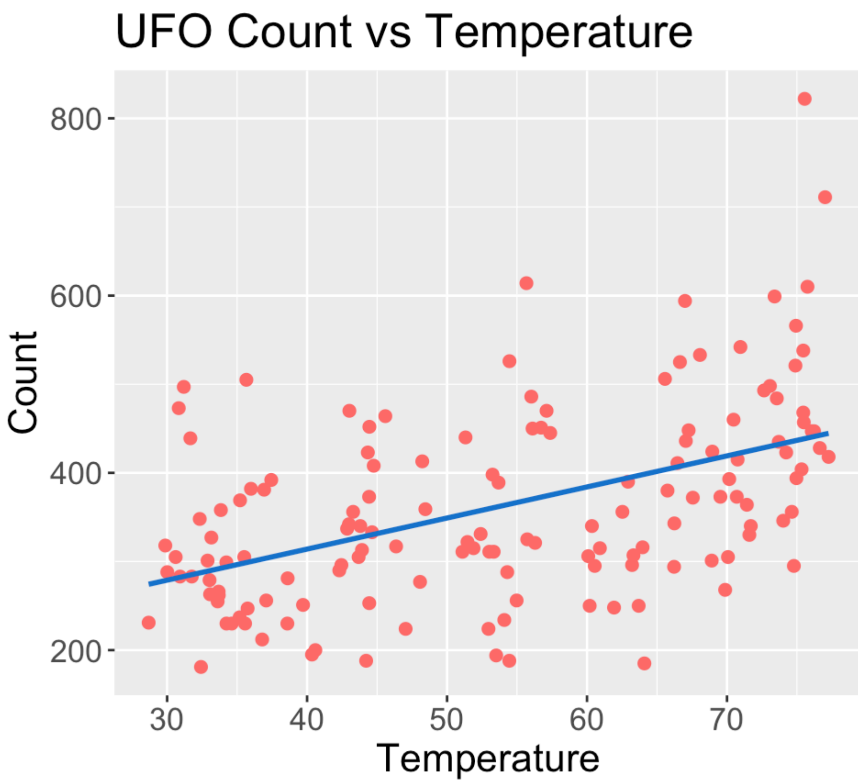
- Least-squares regression review
  - Scatterplot and correlation
  - Least-squares regression
  - Assessing the regression line: residual plot and r^2
- Simple linear regression
  - Idea
  - Model $y = \mu_y + \epsilon = \beta_0 + \beta_1 x + \epsilon$ where $\epsilon \sim N(0, \sigma)$
- Inference for the regression line
  - Confidence intervals $b_0 \pm t^* \mathrm{SE}_{b_0}$ and $b_1 \pm t^* \mathrm{SE}_{b_1}$
  - Significance test $t = \frac{b_1 - 0}{\mathrm{SE}_{b_1}} \overset{approx.}{\sim} t(n-2)$

Outline

Simple linear regression

- Model assumptions of SLR
  - Check assumptions
- Prediction
  - Mean response
  - Individual response
- Inference for predictions
  - Confidence interval for mean response
  - Prediction interval for individual response

Simple linear regression



Denote *Temperature* as x and *Count* as y.

$$\begin{array}{ccccc} y & = & \mu_y & + & \epsilon \\ \text{Data} & = & \text{Fit} & + & \text{Residual} \end{array}$$

$\mu_y = \beta_0 + \beta_1 x$ and $\epsilon \sim N(0, \sigma)$

Model: $y = \beta_0 + \beta_1 x + \epsilon$, where $\epsilon \sim N(0, \sigma)$

Parameters: β_0, β_1, σ

Estimated regression line:

$\hat{y} = b_0 + b_1 x = 173.8 + 3.5x$

```
cor(UFO$Count, UFO$Temperature)
```

```
## [1] 0.4824087
```

Simple linear regression

```r
summary(m <- lm(Count ~ Temperature, data=UFO))
```

```
## Call:
## lm(formula = Count ~ Temperature, data = UFO)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -213.47   -64.13   -14.56    64.82   383.39
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 173.7714    29.9539   5.801 4.11e-08 ***
## Temperature   3.5055     0.5341   6.563 9.20e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 97.64 on 142 degrees of freedom
## Multiple R-squared:  0.2327, Adjusted R-squared:  0.2273
## F-statistic: 43.07 on 1 and 142 DF,  p-value: 9.204e-10
```

- $b_0 = 173.77, \mathrm{SE}_{b_0} = 29.95$
- $b_1 = 3.51, \mathrm{SE}_{b_1} = 0.53$
- $t = \frac{b_1 - 0}{\mathrm{SE}_{b_1}} = \frac{3.51}{0.53} = 6.56,$
- $P = 9.20 \times 10^{-10}$
- $s = 97.64, \mathrm{df} = 142 = 144 - 2$
- $r^2 = 0.23$

Model assumptions

To use the least-squares line as a basis for inference about a population, each of the following conditions should be approximately met:

1. The sample is an **SRS** from the population.
2. There is a **linear** relationship between x and y.
3. The **standard deviation** of the responses y about the population regression line is the **same** for all x.
4. The model residuals are **Normally** distributed.

For assumption 3, since $y = \mu_y + \epsilon$ and $\epsilon \sim N(0, \sigma)$,

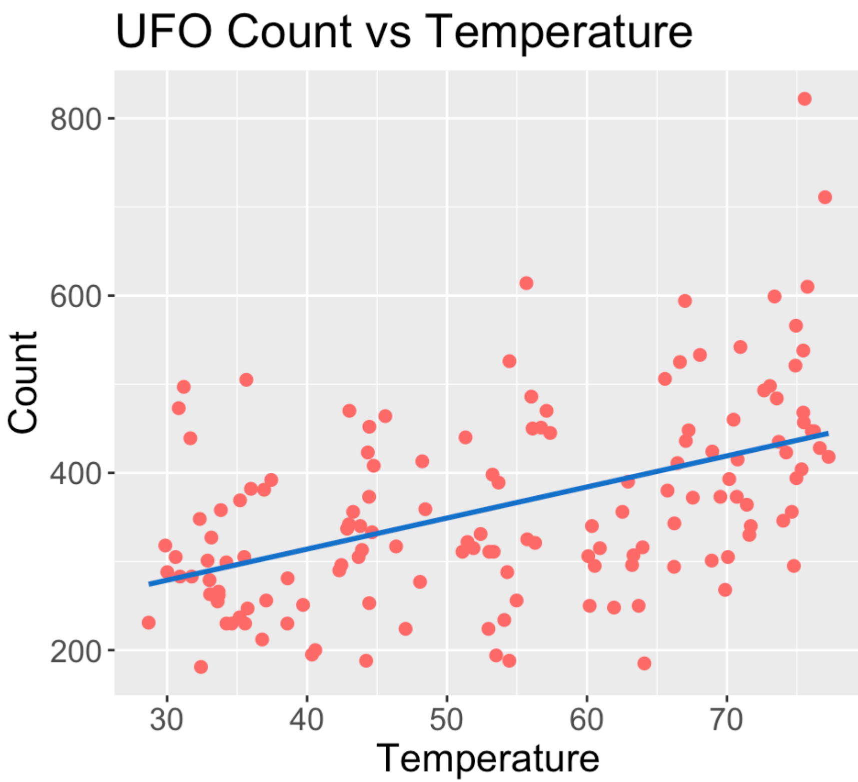
$$y \sim N(\mu_y, \sigma)$$

- Mean $\mu_y = \beta_0 + \beta_1 x$ is different for different x values.
- SD σ is the same for all x.

Check assumptions

1. The sample is an **SRS** from the population.
   - Check data collecting process.
2. There is a **linear** relationship between x and y.
   - Check scatterplot (linear) and residual plot (no pattern).
3. The **standard deviation** of the responses y about the population regression line is the **same** for all x.
   - Check residual plot: the spread of the residuals across the range of x should be roughly uniform.
4. The model residuals are **Normally** distributed.
   - Check Normal Q-Q plot: points should lie closely to the $y = x$ line.

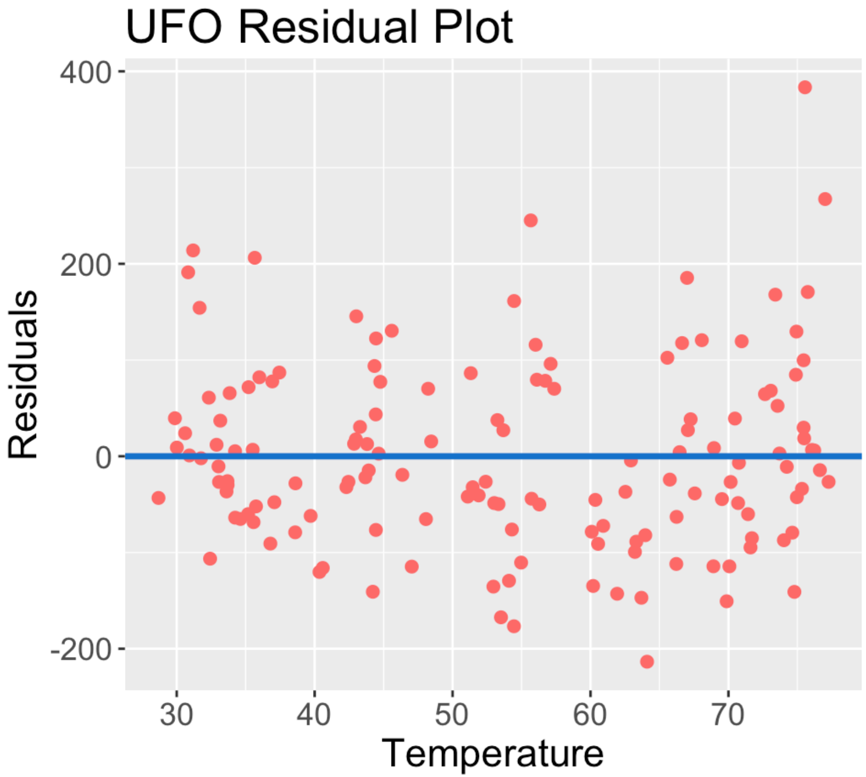
Check assumptions



Assumption 2: There is a linear relationship between x and y.

- Using scatterplot. The overall trend seems roughly linear but may be a little curved. There are one or two unusual points with very large *Count* values.

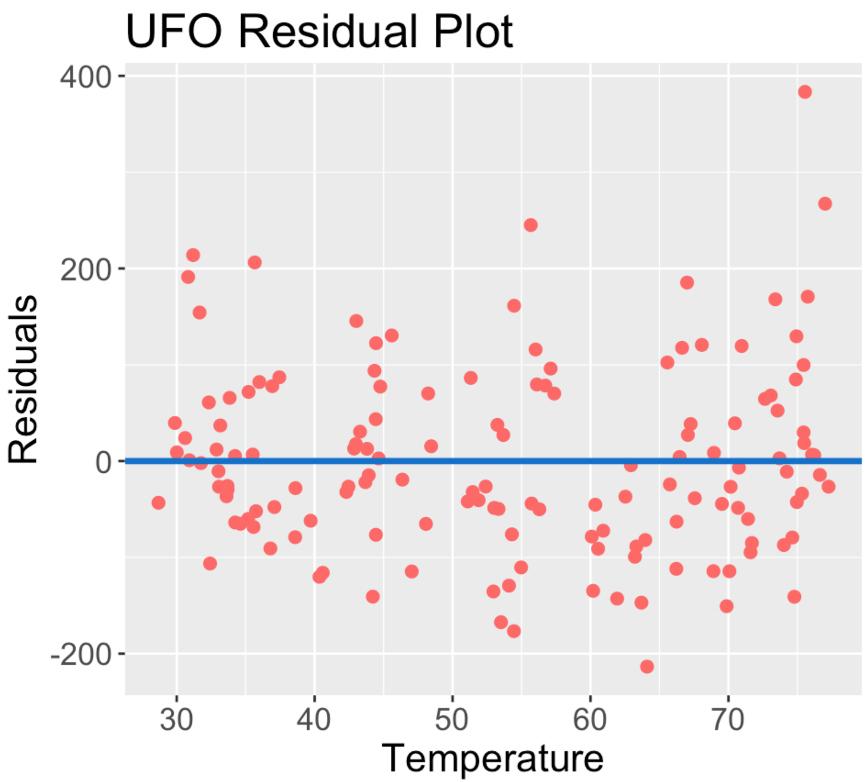
Check assumptions



Assumption 2: There is a linear relationship between x and y.

- Using residual plot. If the relationship is linear, the residual plot should show *no pattern* (points are evenly distributed above and below the $y = 0$ line). Here, it does not have any clear pattern but the overall trend seems to be a little curved.

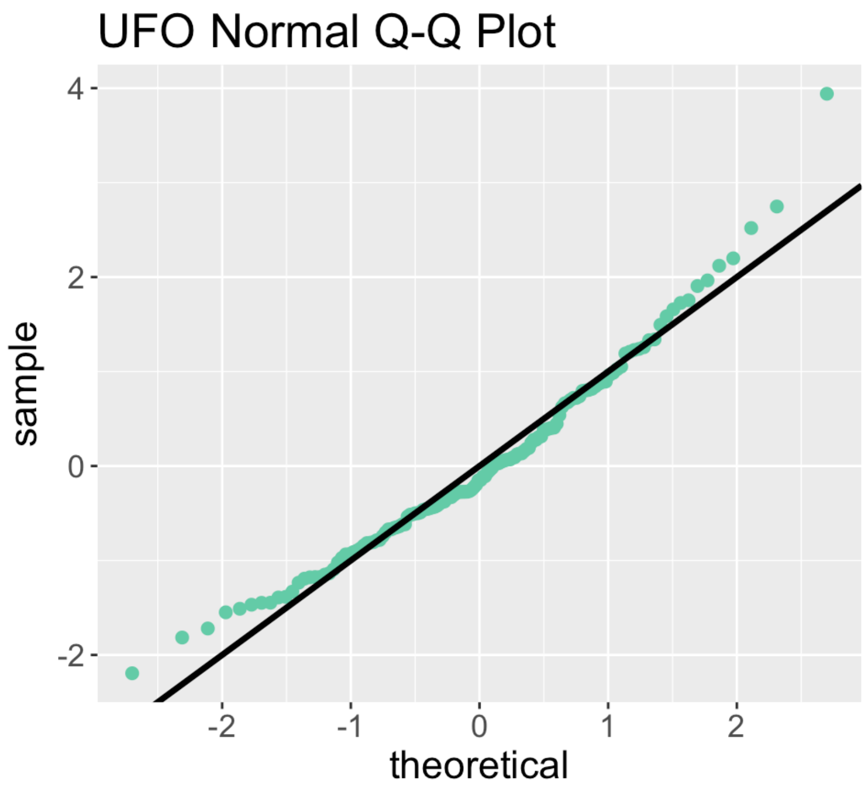
Check assumptions



Assumption 3: The SD of the responses y about the population regression line is the same for all x.

- Using residual plot. The spread of residuals is generally the same for all *Temperature* values except when *Temperature* is higher than 75 F, the spead seems to be much larger.

Check assumptions



Assumption 4: The model residuals are Normally distributed.

- Using Normal Q-Q plot. Most points lie quite closely to the $y = x$ line. But we see a little curved pattern in the points. The Normality assumption is mostly satisfied but could be slightly violated.

Conclusion:

- We don't see any clear violation of the assumptions. But there is probably one or more outliers and only a litte concern about linearity, constant SD and Normality assumptions.

Check assumptions R codes

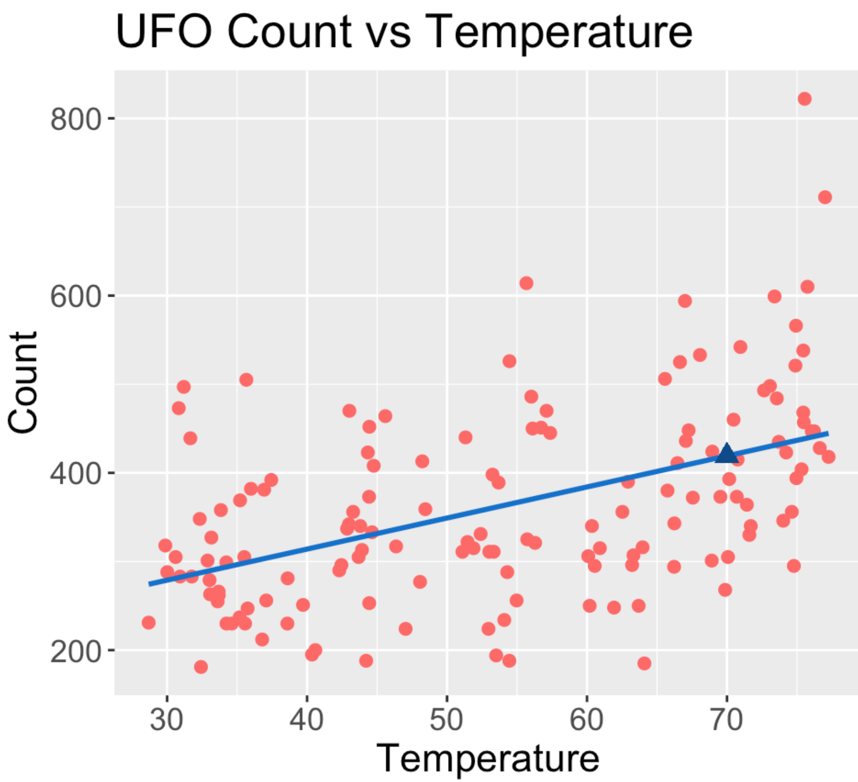
```
library(ggplot2) # use ggplot2 package

# Scatterplot with regression line
ggplot(data=UFO, aes(x=Temperature, y=Count))+
  geom_point(size=2)+
  geom_smooth(method="lm", se=F)+ # add regression line
  ggtitle("UFO Count vs Temperature")

# Residual plot
UFOcheck <- data.frame(Residuals = m$residuals, Temperature = UFO$Temperature)
ggplot(data=UFOcheck, aes(x=Temperature, y=Residuals))+
  geom_point(size=2)+
  geom_hline(yintercept=0, size=1.2)+ # add y=0 line
  ggtitle("UFO Residual Plot")

# Q-Q plot
ggplot(data=UFOcheck, aes(sample = scale(Residuals)))+
  stat_qq(size=2)+
  geom_abline(intercept=0, slope=1, size=1.2)+ # add y=x line
  ggtitle("UFO Normal Q-Q Plot")
```

Prediction



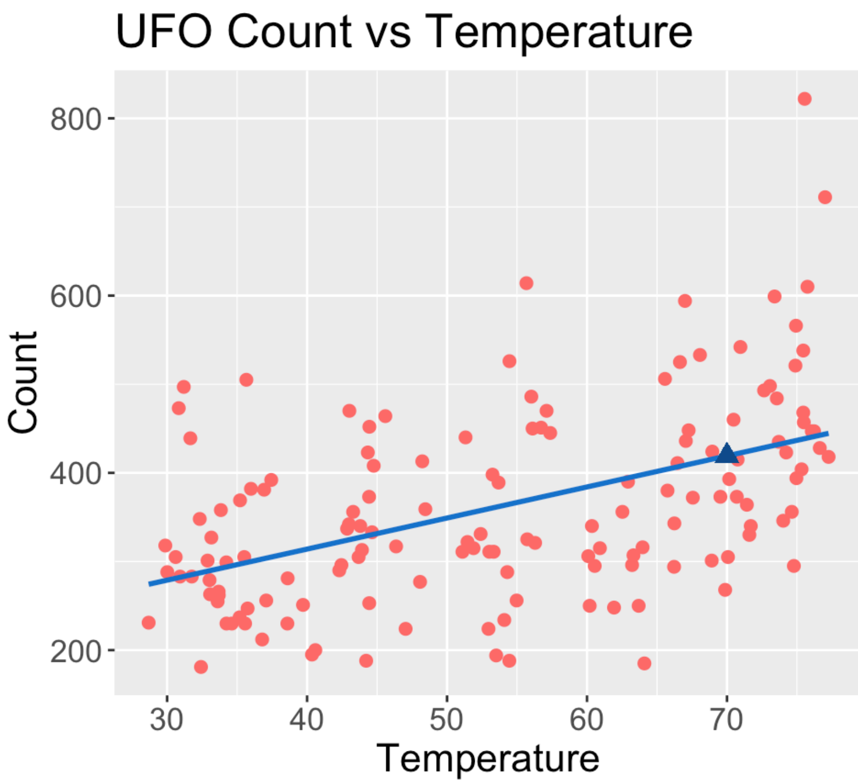
$$y = \beta_0 + \beta_1 x + \epsilon, \text{ where } \epsilon \sim N(0, \sigma)$$

For $b_0 = 173.8$ and $b_1 = 3.5$, we have $173.8 + 3.5 \times 70 = 418.8$ Which interpretation of the value 418.8 is correct?

1. The average UFO *Count* when mean *Temperature* of a month is 70 F is predicted to be 418.8.
2. The UFO *Count* when mean *Temperature* of a month is 70 F is predicted to be 418.8.

- **Both are correct.**

Prediction



Since the SLR model has $y = \beta_0 + \beta_1 x + \epsilon$ and $\mu_y = \beta_0 + \beta_1 x$, when β_0 and β_1 are estimated by b_0 and b_1, there are **two types of predictions**:

1. Mean response $\hat{\mu}_y = b_0 + b_1 x$
   - $\hat{\mu}_y = 418.8$. The average UFO *Count* when mean *Temperature* of a month is 70 F is predicted to be 418.8.
2. Individual response $\hat{y} = b_0 + b_1 x$
   - $\hat{y} = 418.8$. The UFO *Count* when mean *Temperature* of a month is 70 F is predicted to be 418.8.

Prediction

Predicted **mean response** $\hat{\mu}_y = b_0 + b_1 x$

Predicted **individual response** $\hat{y} = b_0 + b_1 x$

- What is the difference between the predictions for mean response and individual response?
- The interpretation: the former predicts the mean of the response y, while the latter predicts an individual response y.
- The variability: the former has smaller variability than the latter (we are more certain about a predicted average than a predicted invidual value).
- We use **confidence interval** and **prediction interval** to make inference about the two types of predictions.

Confidence interval for a mean response

A **level** C **confidence interval for the mean response** $\mu_y = \beta_0 + \beta_1 x$ when x takes value x^* is

$$\hat{\mu}_y \pm t^* \mathrm{SE}_{\hat{\mu}_y}$$

where

$$\mathrm{SE}_{\hat{\mu}_y} = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

and t^* is the value for the $t(n-2)$ density curve with area C between $-t^*$ and t^*.

- x^* is a specific x value that one is interested in. For example, $x^* = 70$ for the UFO-Temperature example.

Prediction interval for an invidual response

A **level** C **prediction interval for an individual response** on the response variable $y = \beta_0 + \beta_1 x + \epsilon$ when x takes value x^* is

$$\hat{y} \pm t^* \mathrm{SE}_{\hat{y}}$$

where

$$\mathrm{SE}_{\hat{y}} = s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

and t^* is the value for the $t(n-2)$ density curve with area C between $-t^*$ and t^*.

- Note: *prediction interval* is essentially a confidence interval with a different name refering spefically to the interval for an individual response.

Confidence interval and prediction interval

- Confidence interval for $\mu_y = \beta_0 + \beta_1 x$

$$\hat{\mu}_y \pm t^* \mathrm{SE}_{\hat{\mu}_y}, \text{ where } \mathrm{SE}_{\hat{\mu}_y} = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

- Prediction interval for $y = \beta_0 + \beta_1 x + \epsilon$

$$\hat{y} \pm t^* \mathrm{SE}_{\hat{y}}, \text{ where } \mathrm{SE}_{\hat{y}} = s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

- $\hat{\mu}_y = \hat{y}$. The predictions for a mean response and an invidual response have the same value.
- $\mathrm{SE}^2_{\hat{y}} = \mathrm{SE}^2_{\hat{\mu}_y} + s^2$. The prediction for a **mean response** has smaller variability and thus narrower confidence interval than the prediction for an **individual response**.

Confidence interval and prediction interval

```r
predict(m, list(Temperature=70))
```

```
##        1
## 419.1532
```

```r
# confidence interval
predict(m, list(Temperature=70), interval="confidence")
```

```
##        fit     lwr      upr
## 1 419.1532 395.803 442.5034
```

- $\hat{\mu}_y = 419.15$
- The 95% confidence interval for μ_y is [395.80, 442.50].
- We are 95% confident that the true average UFO count at *Temperature* 70 F is within 395.80 and 442.50.

Confidence interval and prediction interval

```r
# prediction interval
predict(m, list(Temperature=70), interval="prediction")
```

```
##        fit      lwr      upr
## 1 419.1532 224.7331 613.5733
```

```r
predict(m, list(Temperature=70), interval="prediction", level=0.99)
```

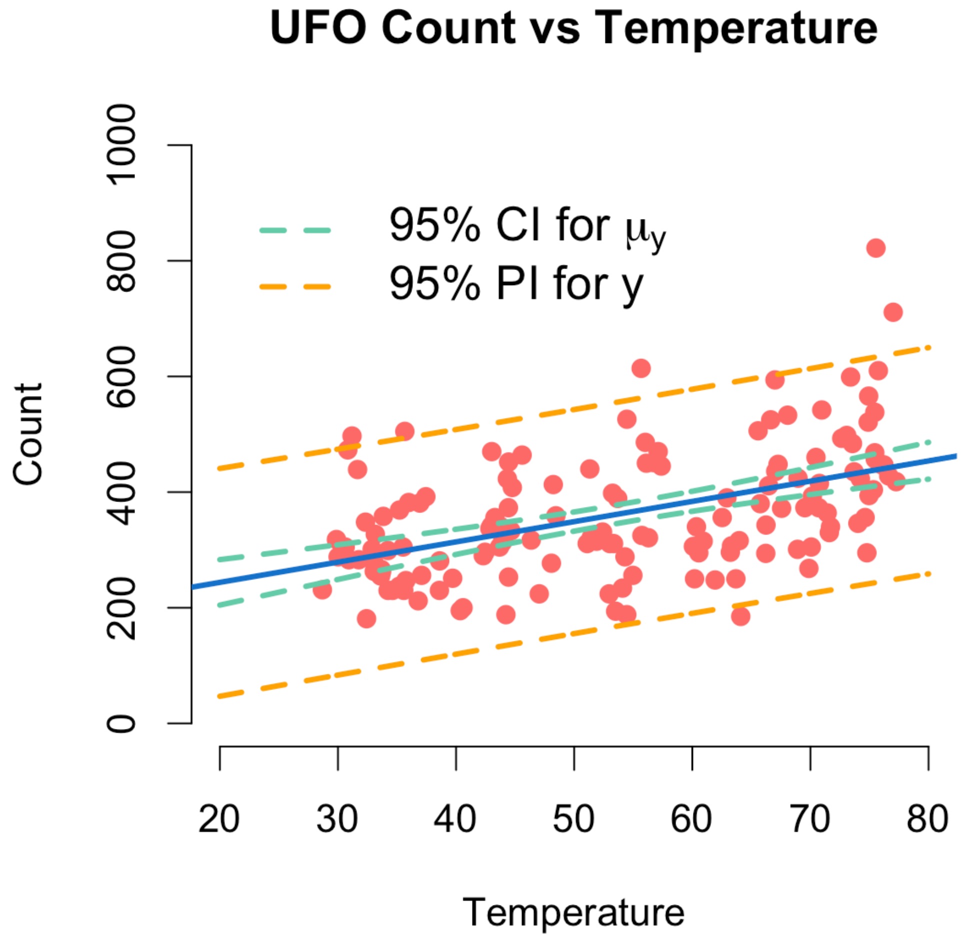
```
##        fit      lwr      upr
## 1 419.1532 162.3707 675.9357
```

- $\hat{y} = 419.15$
- The 95% prediction interval for y is $[224.73, 613.57]$.
- We are 95% confident that the true UFO count at *Temperature* 70 F is within 224.71 and 613.59.

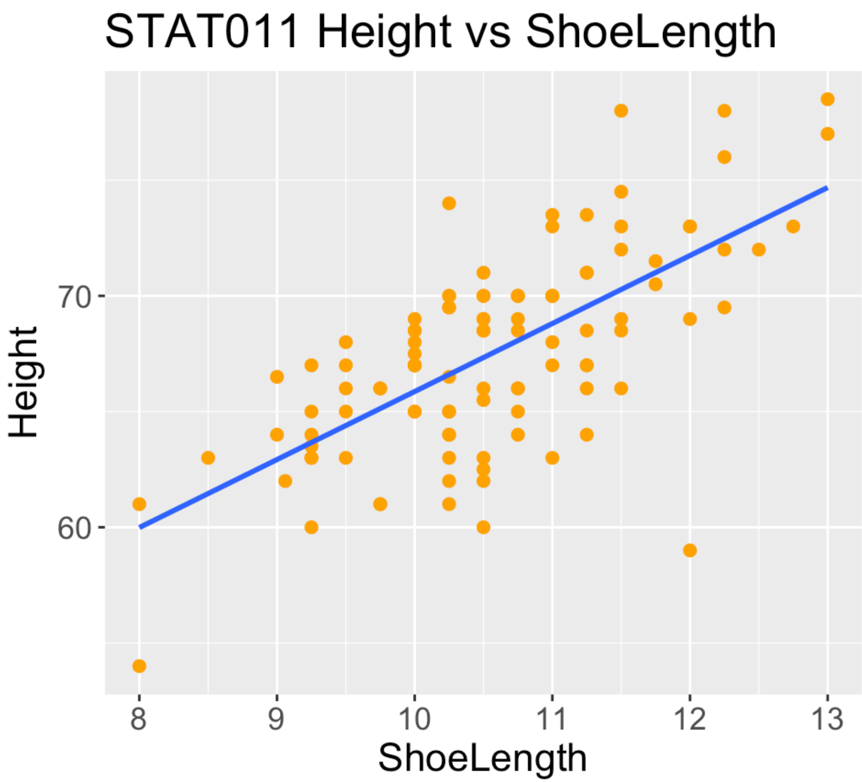
Confidence interval and prediction interval

For $x^* = 70$

- $\hat{\mu}_y = 419.15$ with 95% CI $[395.80, 442.50]$.
- $\hat{y} = 419.15$ with 95% PI $[224.73, 613.57]$.
- The prediction interval for individual response is much wider than the confidence interval for mean response at the same x value.
- This is true for x taking all possible values in the data set.



Height ~ ShoeLength



- Denote *Height* as Y and *ShoeLength* as X
- **Statistical Model:**

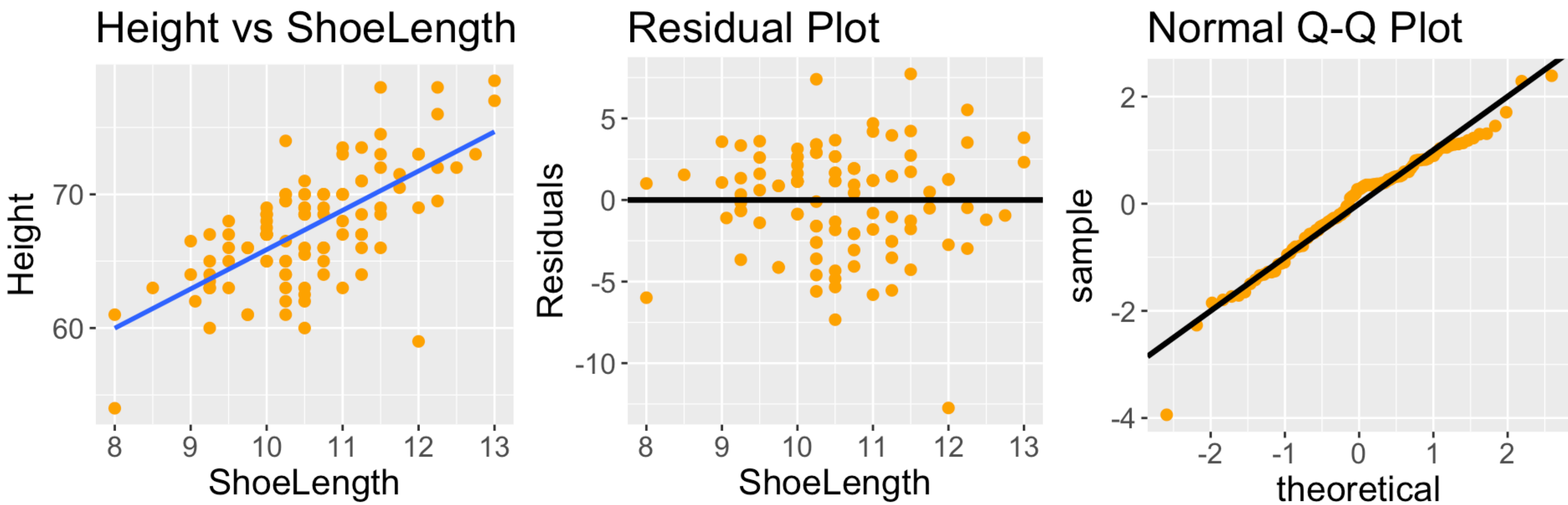
$$y = \beta_0 + \beta_1 x + \epsilon$$

where $\epsilon \sim N(0, \sigma)$

- **Estimated regression line:**

$$\hat{y} = 36.5 + 2.9x$$

Height ~ ShoeLength Check assumptions



- The scatterplot shows a linear trend; the residual plot has no clear pattern and the points have similar spread for all *ShoeLength* values (when *ShoeLength* is small or large, the spread is relatively smaller probably because there are fewer observations); all the points on the Normal Q-Q plot are close to the $y = x$ line. Therefore, except for two suspicious outliers, there is no clear violation of the linearity, constant SD and Normality assumptions.

Height ~ *ShoeLength* Prediction

Predict the mean and individual *Height* for $ShoeLength = 9$ and 11 and provide their corresponding 95% intervals.

```r
heightModel <- lm(Height ~ ShoeLength, data=Survey)
predict(heightModel, list(ShoeLength = c(9, 11)), interval="confidence")
```

```
##        fit      lwr      upr
## 1 62.92717 61.75498 64.09937
## 2 68.80523 68.11939 69.49108
```

```r
predict(heightModel, list(ShoeLength = c(9, 11)), interval="prediction")
```

```
##        fit      lwr      upr
## 1 62.92717 56.37485 69.47949
## 2 68.80523 62.32224 75.28823
```

Interpretation example: the predicted average *Height* for students with $ShoeLength = 11$ inches is 68.8 inches with 95% CI [68.1, 69.5]. We are 95% confident that the true average *Height* at $ShoeLength = 11$ inches is between 68.1 and 69.5 inches.

Summary

Simple linear regression

- Model assumptions
  - Check assumptions 1. SRS 2. Linearity 3. Constant SD 4. Normaility
- Prediction
  - Mean response $\hat{\mu}_y = b_0 + b_1 x$
  - Individual response $\hat{y} = b_0 + b_1 x$
- Inference for predictions
  - Confidence interval for mean response $\hat{\mu}_y \pm t^* \mathrm{SE}_{\hat{\mu}_y}$
  - Prediction interval for individual response $\hat{y} \pm t^* \mathrm{SE}_{\hat{y}}$