STAT011 Statistical Methods I

Lecture 25 Multiple Linear Regression

Lu Chen
Swarthmore College
4/30/2019

Conducting simple linear regression analysis

1. State the statistical model for simple linear regression
   $y = \mu_y + \epsilon = \beta_0 + \beta_1 x + \epsilon$, where $\epsilon \sim N(0, \sigma)$
2. Do exploratory data analysis: scatterplot and correlation
3. Obtain the least-squares regression line and add the line to the scatterplot
4. Check assumptions 1. SRS 2. Linearity 3. Constant SD 4. Normaility
   If assumptions are violated, try transformation
5. Assess the fitting of the model: r^2
6. Make inferences:
   Confidence intervals for both intercept and slope $b_0 \pm t^* \mathrm{SE}_{b_0}$ and $b_1 \pm t^* \mathrm{SE}_{b_1}$
   Significance test for the slope $t = \frac{b_1 - 0}{\mathrm{SE}_{b_1}} \overset{approx.}{\sim} t(n-2)$
7. Predictions:
   Mean response and its confidence interval $\hat{\mu}_y \pm t^* \mathrm{SE}_{\hat{\mu}_y}$
   Individual response and its prediction interval $\hat{y} \pm t^* \mathrm{SE}_{\hat{y}}$

Outline

Multiple linear regression

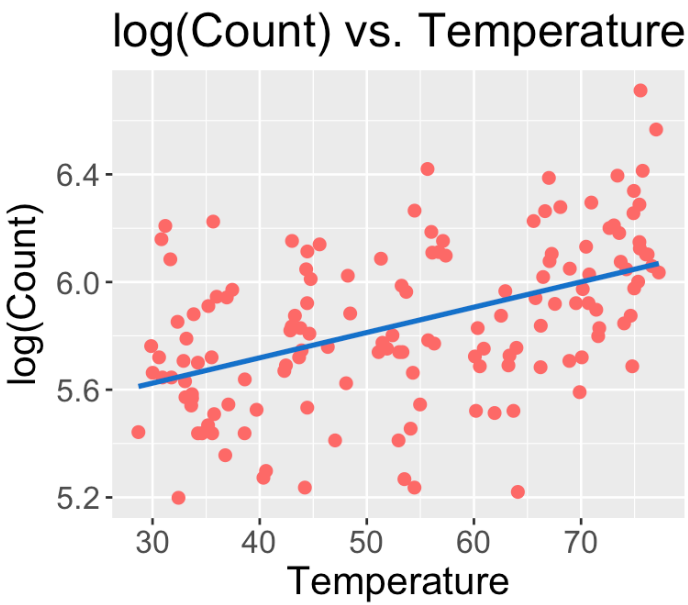
- Statistical model
- Model interpretation
- Inference for the slopes
- Model assessment

Data

```r
head(UFO, 15)
```

```
##    Year Month Temperature Precipitation Count
## 1  2000     1       34.24          2.03   230
## 2  2000     2       40.59          2.01   200
## 3  2000     3       47.04          2.38   224
## 4  2000     4       53.51          2.27   194
## 5  2000     5       64.10          2.62   185
## 6  2000     6       69.87          3.49   268
## 7  2000     7       74.63          2.30   356
## 8  2000     8       74.78          1.92   295
## 9  2000     9       66.22          2.17   294
## 10 2000    10       55.73          2.21   325
## 11 2000    11       38.59          2.68   230
## 12 2000    12       28.68          1.65   231
## 13 2001     1       31.76          1.81   283
## 14 2001     2       34.62          2.14   230
## 15 2001     3       42.30          2.50   290
```

UFO and temperature



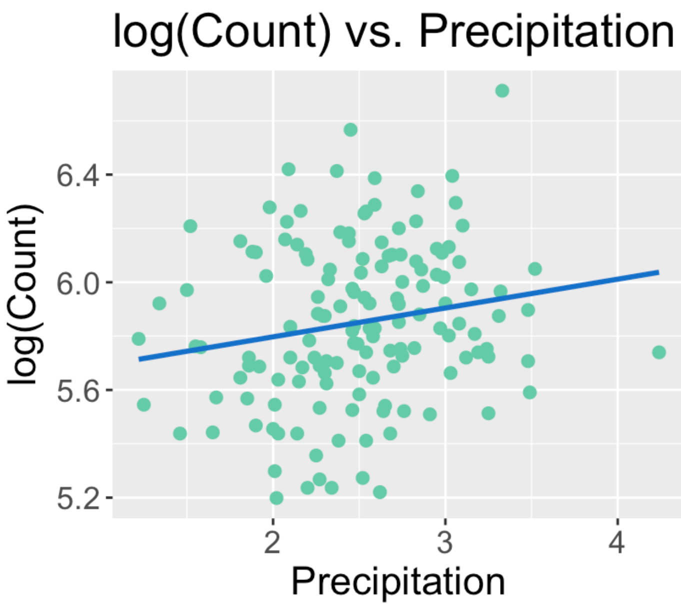
- $\widehat{\log(Count)} = 5.34 + 0.01 \times Temperature$
- As *Temperature* increases one degree, *log(Count)* increases 0.01 unit.
- t test for the slope of *Temperature* has $t = 6.55$ and $P = 1.01 \times 10^{-9} << 0.05$. *Temperature* and *log(Count)* have a highly significant positive relationship.

```r
m1 <- lm(log(Count) ~ Temperature, data=UFO)
summary(m1)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.342374   0.080562  66.314  < 2e-16 ***
## Temperature 0.009403   0.001437   6.546 1.01e-09 ***
```

UFO and raining

UFO and raining



- $\widehat{\log(Count)} = 5.58 + 0.11 \times Precipitation$
- As *Precipitation* increases 1 unit, *log(Count)* increases 0.11 unit.
- t test for the slope of *Precipitation* has $t = 2.21$ and $P = 0.029 < 0.05$. *Precipitation* and *log(Count)* have a significant positive relationship.

```r
m2 <- lm(log(Count) ~ Precipitation, data=UFO)
summary(m2)
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     5.58322    0.12315   45.34   <2e-16 ***
## Precipitation  0.10709    0.04847    2.21   0.0287 *
```

UFO and raining

```r
m1 <- lm(log(Count) ~ Temperature, data=UFO)
summary(m1)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.342374   0.080562  66.314  < 2e-16 ***
## Temperature 0.009403   0.001437   6.546 1.01e-09 ***
##
## Residual standard error: 0.2626 on 142 degrees of freedom
## Multiple R-squared:  0.2318, Adjusted R-squared:  0.2264
## F-statistic: 42.84 on 1 and 142 DF,  p-value: 1.005e-09
```

- $r^2 = 0.2318$.

```r
m2 <- lm(log(Count) ~ Precipitation, data=UFO)
summary(m2)
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.58322    0.12315   45.34   <2e-16 ***
## Precipitation  0.10709    0.04847    2.21   0.0287 *
##
## Residual standard error: 0.2946 on 142 degrees of freedom
## Multiple R-squared:  0.03324,    Adjusted R-squared:  0.02643
## F-statistic: 4.882 on 1 and 142 DF,  p-value: 0.02874
```

- $r^2 = 0.0332$.

Multiple Linear Regression Model

Denote $\log(Count)$ as y, *Precipitation* as x_1 and *Temperature* as x_2.

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

- $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
- $\epsilon \sim N(0, \sigma)$

Statistical Model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$, where $\epsilon \sim N(0, \sigma)$

Parameters: $\beta_0, \beta_1, \beta_2, \sigma$

Estimated regression line: $\hat{y} = b_0 + b_1 x_1 + b_2 x_2$

UFO, precipitation and temperature

```r
summary(m3 <- lm(log(Count) ~ Precipitation + Temperature, data=UFO))
```

```
## Call:
## lm(formula = log(Count) ~ Precipitation + Temperature, data = UFO)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.72722 -0.17891 -0.00509  0.19283  0.68442
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.42728    0.11260  48.199  < 2e-16 ***
## Precipitation -0.05451    0.05053  -1.079    0.283
## Temperature    0.01035    0.00168   6.157 7.29e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2625 on 141 degrees of freedom
## Multiple R-squared:  0.2381, Adjusted R-squared:  0.2273
## F-statistic: 22.03 on 2 and 141 DF,  p-value: 4.731e-09
```

- $b_0 = 5.43, b_1 = -0.05, b_2 = 0.01, \sigma = 0.26$
- Note: model estimates and fitting stays the **same** when *Temperature* goes before *Precipitation* in the model.

UFO, precipitation and temperature

```r
summary(m3)
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.42728    0.11260  48.199  < 2e-16 ***
## Precipitation -0.05451    0.05053  -1.079    0.283
## Temperature    0.01035    0.00168   6.157 7.29e-09 ***
```

Estimated regression line:

$$\widehat{\log(Count)} = 5.43 - 0.05 \times Precipitation + 0.01 \times Temperature$$

- As *Precipitation* increases one unit, $\log(Count)$ descreases 0.05 unit, given that *Temperature* is held constant.
  - When temperature stays the same: more rain $\Rightarrow$ less UFO reports.
- As *Temperature* increases one unit, $\log(Count)$ increases 0.01 unit, given that *Precipitation* is held constant.
  - When precipitation stays the same: higher temperature $\Rightarrow$ more UFO reports.

UFO, precipitation and temperature

```r
summary(m3)
```

```
##                Estimate Std. Error t value Pr(>|t|)    
## (Intercept)     5.42728    0.11260  48.199  < 2e-16 ***
## Precipitation -0.05451    0.05053  -1.079    0.283    
## Temperature     0.01035    0.00168   6.157 7.29e-09 ***
```

- **Test the significance of *Precipitation*** $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$
  - $t = -1.08, P = 0.28 > 0.05$
  - Given that *Temperature* is held constant, the relationship between *Precipitation* and UFO count is negative but not significant.
- **Test the significance of *Temperature*** $H_0 : \beta_2 = 0$ vs. $H_a : \beta_2 \neq 0$
  - $t = 6.16, P = 7.3 \times 10^{-9} < 0.05$
  - Given that *Precipitation* is held constant, the relationship between UFO count and *Temperature* is positive and highly significant.

UFO, precipitation and temperature

- Model 1: $\log(\widehat{Count}) = 5.34 + 0.01 \times Temperature$
  - $t = 6.55$ and $P = 1.01 \times 10^{-9} << 0.05$
  - $r^2 = 0.2318$
- Model 2: $\log(\widehat{Count}) = 5.58 + 0.11 \times Precipitation$
  - $t = 2.21$ and $P = 0.029 < 0.05$
  - $r^2 = 0.0332$
- Model 3: $\log(\widehat{Count}) = 5.43 - 0.05 \times Precipitation + 0.01 \times Temperature$
  - $t = -1.08, P = 0.28 > 0.05$ for *Precipitation*
  - $t = 6.16, P = 7.3 \times 10^{-9} << 0.05$ for *Temperature*
  - $r^2 = 0.2381$
- How do the relationships change when both *Precipitation* and *Temperature* are in the model?

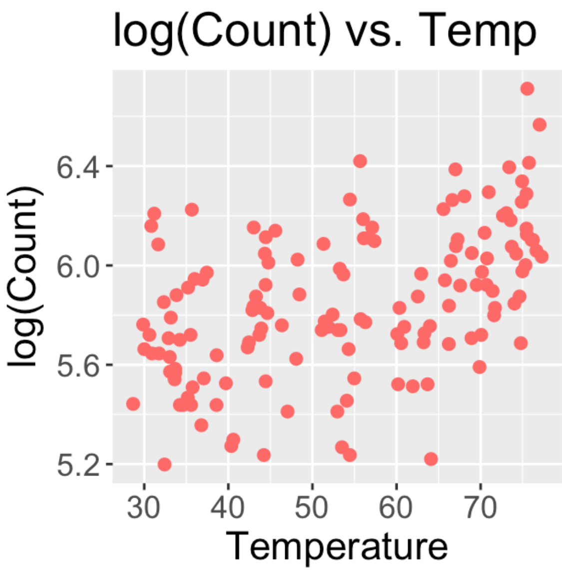
UFO, precipitation and temperature

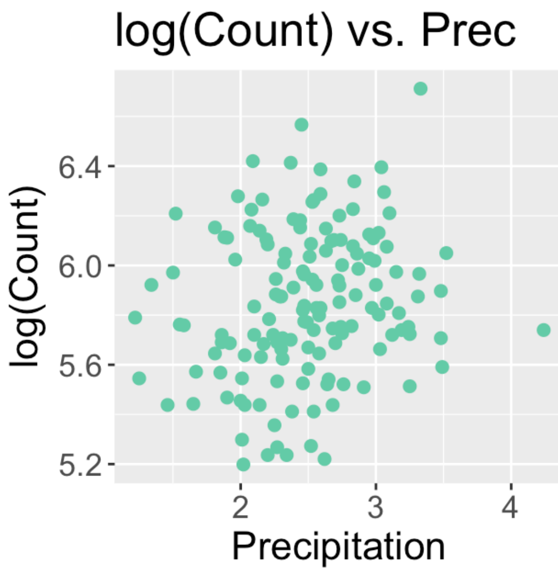
Consider the t tests and r^2 values in Model 1 & 3, the relationship between *Temperature* and *log(Count)* stays similar with or without *Precipitation*.

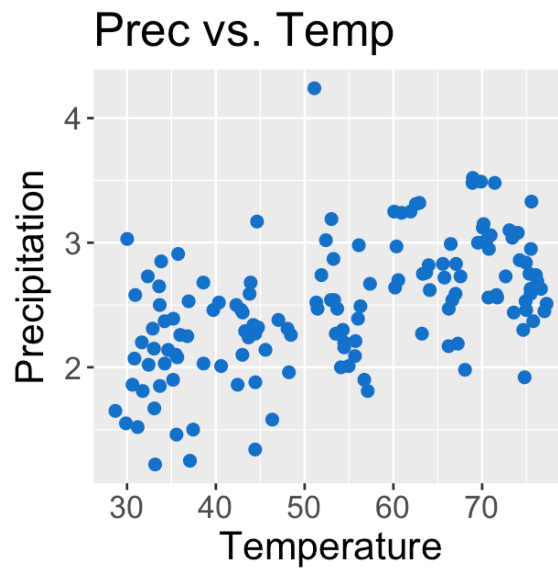
Consider the t tests and r^2 values in Model 2 & 3, the relationship between *Precipitation* and *log(Count)* changes dramatically when adding *Temperature* to the model, which

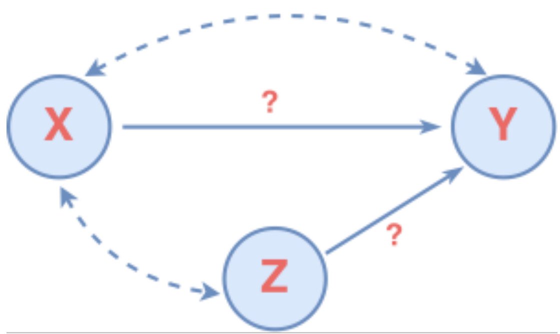
- reverses the direction of the relationship between *Precipitation* and *log(Count)* from positive to negative - Simpson's paradox.
- explains 21% more variability in *log(Count)*.
- makes *Precipitation* no longer a significant predictor for *log(Count)*.

UFO, precipitation and temperature



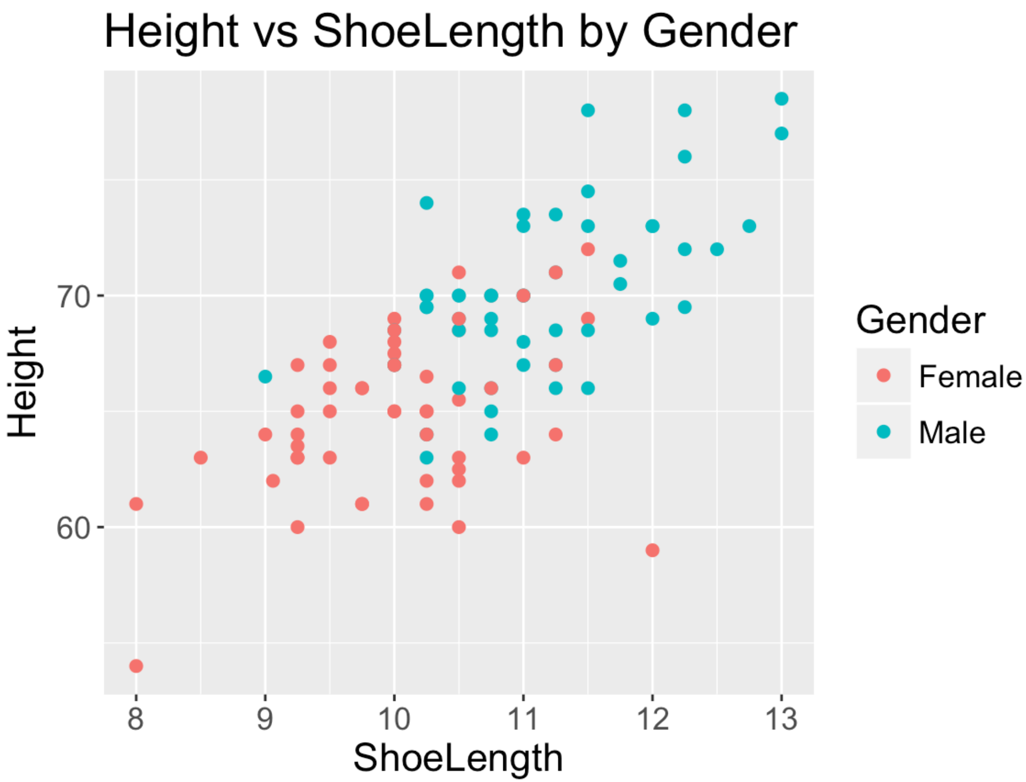






Temperature is a confounder in the relationship between *Precipitation* and *log(Count)*. It affects the relationship so heavily because it is associated with *Precipitation* ($r = 0.52$).

Height, shoe length and gender



Denote *Height* as y, *ShoeLength* as x_1 and *Gender* as x_2.

Model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$, where $\epsilon \sim N(0, \sigma)$

Parameters: $\beta_0, \beta_1, \beta_2, \sigma$

- x_1 is quantitative
- x_2 is categorical, $x_2 = 1$ for male and 0 for female.

Height, shoe length

```r
summary(m4 <- lm(Height ~ ShoeLength, data=Survey))
```

```
## Call:
## lm(formula = Height ~ ShoeLength, data = Survey)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.7443  -1.8357   0.8686   1.9295   7.7253
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  36.4759     3.3775  10.800  < 2e-16 ***
## ShoeLength    2.9390     0.3182   9.236 3.69e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.251 on 103 degrees of freedom
## Multiple R-squared:  0.453,	Adjusted R-squared:  0.4477
## F-statistic: 85.31 on 1 and 103 DF,  p-value: 3.688e-15
```

Height, shoe length and gender

```r
summary(m5 <- lm(Height ~ ShoeLength + Gender, data=Survey))
```

```
## Call:
## lm(formula = Height ~ ShoeLength + Gender, data = Survey)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.3595  -1.9134   0.2278   2.0169   7.0396
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  43.1009     3.6403  11.840  < 2e-16 ***
## ShoeLength    2.1882     0.3606   6.068 2.22e-08 ***
## GenderMale    2.6950     0.7192   3.747 0.000297 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.062 on 102 degrees of freedom
## Multiple R-squared:  0.5192, Adjusted R-squared:  0.5098
## F-statistic: 55.08 on 2 and 102 DF,  p-value: < 2.2e-16
```

Height, shoe length and gender

```r
summary(m5)
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    43.1009     3.6403  11.840  < 2e-16 ***
## ShoeLength      2.1882     0.3606   6.068 2.22e-08 ***
## GenderMale      2.6950     0.7192   3.747 0.000297 ***
```

$$\widehat{Height} = 43.1 + 2.2 \times ShoeLength + 2.7 \times Gender$$

- As *ShoeLength* increases 1 inch, *Height* increases 2.2 inches, given that *Gender* is held constant.
  - When *Gender* stays the same: longer shoe length $\Rightarrow$ larger height value.
- As *Gender* increases one unit (from *Female* to *Male*), *Height* increases 2.7 inches, given that *ShoeLength* is held constant.
  - When *ShoeLength* stays the same: males are 2.7 inches taller than females on average.

Height, shoe length and gender

```r
summary(m5)
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   43.1009     3.6403  11.840  < 2e-16 ***
## ShoeLength     2.1882     0.3606   6.068 2.22e-08 ***
## GenderMale     2.6950     0.7192   3.747 0.000297 ***
```

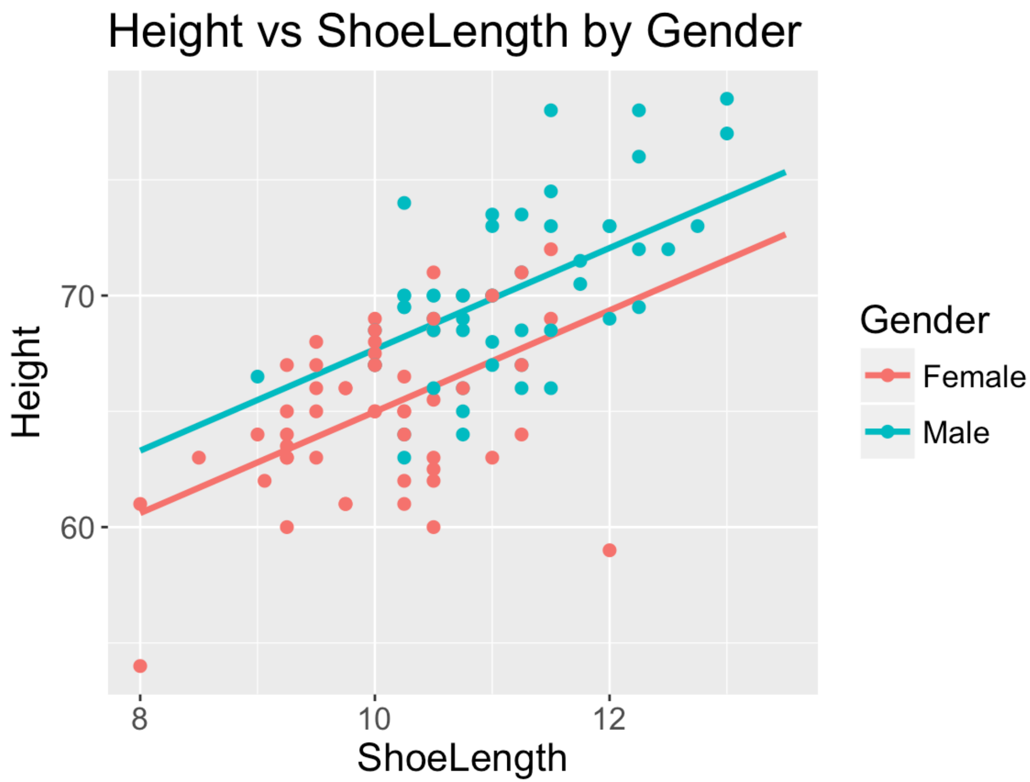
$$\widehat{Height} = 43.1 + 2.2 \times ShoeLength + 2.7 \times Gender$$

- For females, $Gender = 0$,

  $\widehat{Height} = 43.1 + 2.2 \times ShoeLength$
- For males, $Gender = 1$,

  $\widehat{Height} = 43.1 + 2.2 \times ShoeLength + 2.7 = 45.8 + 2.2 \times ShoeLength$

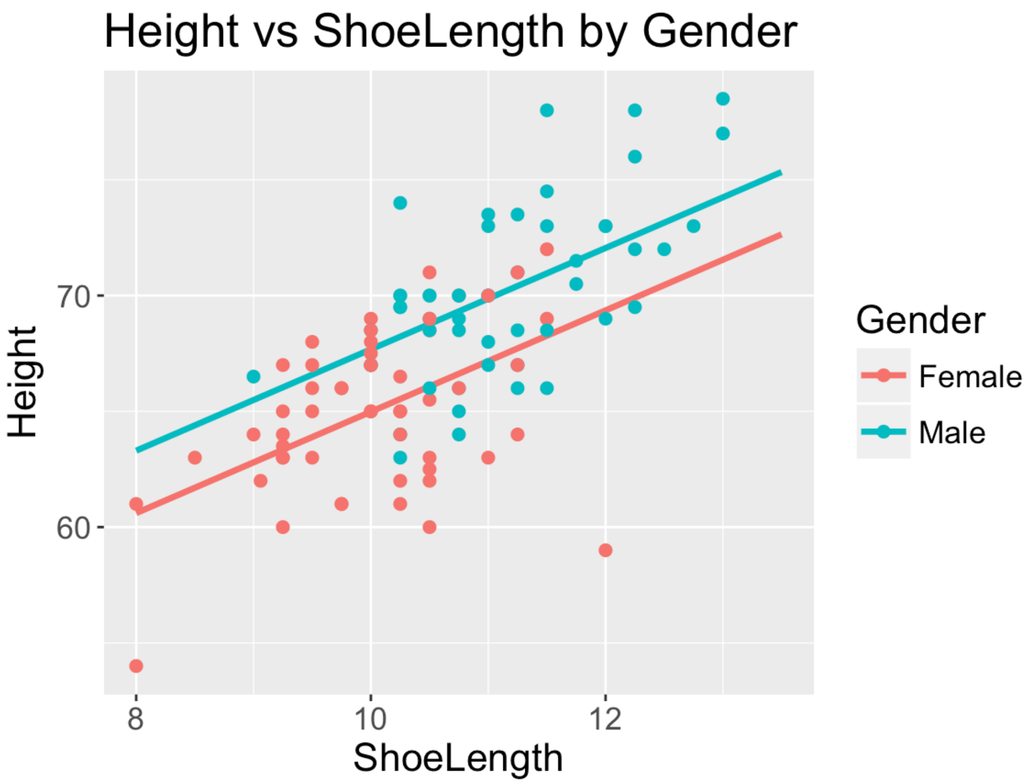
Height, shoe length and gender



$$\widehat{Height} = 43.1 + 2.2 \times ShoeLength + 2.7 \times Gender$$

- $b_1 = 2.2$ measures the **rate of change in *Height*** as *ShoeLength* changes **for both male and female students**.
- $b_2 = 2.7$ measures the **difference in *Height* between *Gender=1* (male) and *Gender=0* (female)** given the same *ShoeLength*.

Model assessment



$$\widehat{Height} = 43.1 + 2.2 \times ShoeLength + 2.7 \times Gender$$

- Model without *Gender*: *ShoeLength* and *Height* have a significant relationship; $r^2 = 0.453$.
- Model with *Gender*: Both *Gender* and *ShoeLength* have significant relationships with *Height*; $r^2 = 0.519$.
- Adding *Gender* to the model improves model fitting.

Model assessment

For multiple linear regression model with k explanatory variables,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

where $\epsilon \sim N(0, \sigma)$, how do we know the significance/effectiveness of the whole model?

- F test with degrees of freedom k and $n - k - 1$.

  $P \leq 0.05$, **the model** is significant at level 0.05.

  $P > 0.05$, **the model** is not significant at level 0.05 - all explanatory variables are not statistically associated with the response variable.
- r^2 : fraction of variability in the response variable explained by **all the explanatory variables** together.

Model assessment

```r
summary(m3)
```

```
## Call:
## lm(formula = log(Count) ~ Precipitation + Temperature, data = UFO)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.72722 -0.17891 -0.00509  0.19283  0.68442
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.42728    0.11260  48.199  < 2e-16 ***
## Precipitation -0.05451    0.05053  -1.079    0.283
## Temperature    0.01035    0.00168   6.157 7.29e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2625 on 141 degrees of freedom
## Multiple R-squared:  0.2381, Adjusted R-squared:  0.2273
## F-statistic: 22.03 on 2 and 141 DF,  p-value: 4.731e-09
```

- $F = 22.03$, $P = 4.7 \times 10^{-9}$
- $r^2 = 0.2381$

Model assessment

```r
summary(m5)
```

```
## Call:
## lm(formula = Height ~ ShoeLength + Gender, data = Survey)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.3595  -1.9134   0.2278   2.0169   7.0396
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  43.1009     3.6403  11.840  < 2e-16 ***
## ShoeLength    2.1882     0.3606   6.068 2.22e-08 ***
## GenderMale    2.6950     0.7192   3.747 0.000297 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.062 on 102 degrees of freedom
## Multiple R-squared:  0.5192, Adjusted R-squared:  0.5098
## F-statistic: 55.08 on 2 and 102 DF,  p-value: < 2.2e-16
```

- $F = 55.8$, $P < 2.2 \times 10^{-16}$
- $r^2 = 0.5192$

Summary

Multiple linear regression

- Statistical model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$ where $\epsilon \sim N(0, \sigma)$
- Model interpretation given that xxx is held constant
- Inference for the slopes
- Model assessment F test and r^2

Final Exam

Final Exam

- Tuesday 5/14 9am-12pm **SCI 101**
- Practice problems available on Thursday 5/2

Important dates

- Office hours
  - Tuesday 4/30 (today), 2:40 - 4:30pm
  - Thursday 5/9, 9:30 - 11:30am
- Muses sessions
  - Wednesday 5/1, 8 - 10pm
  - Monday 5/13, 7 - 9pm
- Stat Clinics
  - Friday 5/10 & Saturday 5/11, 3 - 6pm
  - Sunday 5/12, 6 - 9pm