Data Analytics
CS301
Multiple Linear Regression
Understanding the Summary

Fall 2018

Oliver Bonham-Carter

Up To Now in Regression

- We have discussed how one entity influences another.
- What about having two entities (independent) which may have some kind of influence on a dependent variable.
- Especially if a dependent variable has a high correlation with more multiple independent variable.

Main Idea

- GPA could be dependent on studying
- Student performance may be based on more than just one entity.
- GPA could be dependent on studying AND getting enough rest
- OR maybe even more variables are involved?
- GPA could be dependent on studying AND rest AND eating good food AND …

So, Multiple Linear Regression Is What … ?

- Multiple linear regression is the most common form of linear regression analysis.
- A predictive analysis, the multiple linear regression is used to explain the relationship between one continuous dependent variable and two or more independent variables.
- The independent variables can be continuous or categorical (dummy coded as appropriate).

Types of Questions Answered

- Do age and IQ scores effectively predict GPA?
- Do weight, height, and age explain the variance in cholesterol levels?
- Are video game sales explained by their exciting graphics and inexpensive costs?
- Is road safety a combination of relaxed and defensive driving?
- Are there more independent variables to be used to answer to these above dependents?

Equation of Multiple Independent Variables

- The model is now a multi-independent variable equation.

Dependent Variable

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \epsilon_i$$

Independent Variables

Multiple Variables
Equation and Assumptions

- A population model for a multiple regression model that relates a *y*-variable to *p* − 1 predictor variables is written as the following.

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_{p-1} x_{i,p-1} + \varepsilon_i$$

- We assume that the ε_i have a normal distribution with mean 0 and constant variance σ^2. These are the same assumptions that we used in simple regression with one *x*-variable.

- The subscript *i* refers to the ith individual or unit in the population. In the notation for the *x*-variables, the subscript following *i* simply denotes which *x*-variable it is.

Same Hypothesis as Before...

As an example, to determine whether variable X_1 is a useful predictor variable in a model, we use the following hypothsis:

$$H_0 : \beta_1 = 0$$

$$H_A : \beta_1 \, != 0$$

(think slope values)

If the null hypothesis above were the case, then a change in the value of X_1 would not change Y, so Y and X_1 are not related.

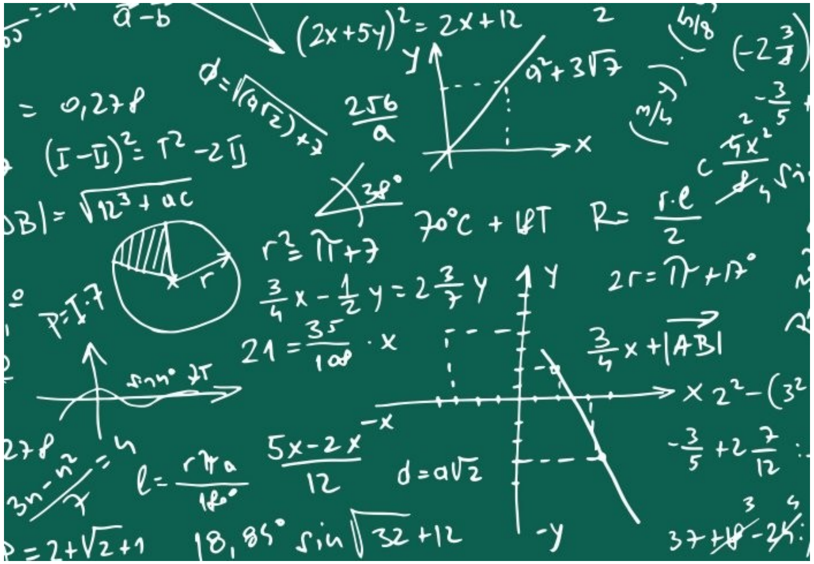
We would still be left with variables X_2 and X_3 being present in the model and so we could not reject the null hypothesis above. Instead we should say that we do not need variable X_1 in the model given that variables X_2 and X_3 will remain.

In general, the interpretation of a slope in multiple regression can be tricky. Correlations among the predictors can change the slope values dramatically from what they would be in separate simple regressions.

Analysis Question

- Two variables: Do *Age* and *Height* (both) influence the capacity of lungs (*LungCap*)?
- Asking actually, can we make a model that takes the following form?

$$\text{LungCap} = \text{Age}*b_1 + \text{Height}*b_2 + b_3$$



Lung Capacity Data

```
library(tidyverse)
library(psych)


#open lung capacity data
lc <-file.choose()
dataLungCap <- read.csv(lc)
View(dataLungCap)
```

Create the Multiple-Variable Regression Model

```
# model creation
mod <- lm(LungCap ~ Age + Height)
# get a report of the model
summary(mod)
```

Summary

```
Call:
lm(formula = LungCap ~ Age + Height)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.4080 -0.7097 -0.0078  0.7167  3.1679 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -11.747065   0.476899 -24.632  < 2e-16 ***
Age           0.126368   0.017851   7.079 3.45e-12 ***
Height        0.278432   0.009926  28.051  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.056 on 722 degrees of freedom
Multiple R-squared:  0.843,	Adjusted R-squared:  0.8425 
F-statistic:  1938 on 2 and 722 DF,  p-value: < 2.2e-16
```

Intercept Value:
"When the age and height of zero"

```
Call:
lm(formula = LungCap ~ Age + Height)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.4080 -0.7097 -0.0078  0.7167  3.1679 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -11.747065   0.476899 -24.632  < 2e-16 ***
Age           0.126368   0.017851   7.079 3.45e-12 ***
Height        0.278432   0.009926  28.051  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.056 on 722 degrees of freedom
Multiple R-squared:  0.843,     Adjusted R-squared:  0.8425 
F-statistic:  1938 on 2 and 722 DF,  p-value: < 2.2e-16
```

The estimated mean lung capacity of someone having an age and height of zero. Is this meaningful?

Slope of Age:

“How is my *Age* variable related to *Height*?”

```
Call:
lm(formula = LungCap ~ Age

Residuals:
    Min      1Q  Median
-3.4080 -0.7097 -0.0078  0.71

Coefficients:
              Estimate Std   ror t value Pr(>|t|)
(Intercept) -11.747065   0.476899 -24.632  < 2e-16 ***
Age           0.126368   0.017851   7.079 3.45e-12 ***
Height        0.278432   0.009926  28.051  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.056 on 722 degrees of freedom
Multiple R-squared:  0.843,     Adjusted R-squared:  0.8425
F-statistic:  1938 on 2 and 722 DF,  p-value: < 2.2e-16
```

The effect of *Age* on *Lung Capacity* adjusting or controlling for *Height*. We may associate an increase of 1 year in *Age* with an increase of 0.126 in *Lung Capacity* adjusting or controlling for *Height*

Test Statistic:

```
Call:
lm(formula = LungCap ~ Age + Height)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4080 -0.7097 -0.0078  0.7167  3.1679

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -11.747065   0.476899 -24.632  < 2e-16 ***
Age           0.126368   0.017851   7.079 3.45e-12 ***
Height        0.278432   0.009926  28.051  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.056 on 722 degrees of freedom
Multiple R-squared:  0.843,     Adjusted R-squared:  0.8425
F-statistic:  1938 on 2 and 722 DF,  p-value: < 2.2e-16
```

The test statistic that we use to perform the hypothesis test that the slope for *Age* = 0.

Test Statistic:

```
Call:
lm(formula = LungCap ~ Age + Height)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.4080 -0.7097 -0.0078  0.7167  3.1679 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -11.747065   0.476899 -24.632  < 2e-16 ***
Age           0.126368   0.017851   7.079 3.45e-12 ***
Height        0.278432   0.009926  28.051  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.056 on 722 degrees of freedom
Multiple R-squared:  0.843,	Adjusted R-squared:  0.8425 
F-statistic:  1938 on 2 and 722 DF,  p-value: < 2.2e-16
```

The estimated effect of *Height* on *Lung Capacity*, adjusted for *Age*.

Test Statistic:

```
Call:
lm(formula = LungCap ~ Age + Height)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.4080 -0.7097 -0.0078  0.7167  3.1679 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -11.747065   0.476899 -24.632  < 2e-16 ***
Age           0.126368   0.017851   7.079 3.45e-12 ***
Height        0.278432   0.009926  28.051  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.056 on 722 degrees of freedom
Multiple R-squared:  0.843,	Adjusted R-squared:  0.8425 
F-statistic:  1938 on 2 and 722 DF,  p-value: < 2.2e-16
```

The test statistic that we use to perform the hypothesis test that the slope for *Height* = 0.

R-squared Value:
“How do the independents explain the dependent?”

```
Call:
lm(formula = LungCap ~ Age + Height)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4080 -0.7097 -0.0078  0.7167  3.1679

Coefficients:
              Estimate Std. Error t valu    (>|t|)
(Intercept) -11.747065   0.476899 -24         < 2e-16 ***
Age           0.126368   0.017851     079 3.45e-12 ***
Height        0.278432   0.00992   28.051   < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.00  ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.056 on 722 degrees of freedom
Multiple R-squared:  0.843,     Adjusted R-squared:  0.8425
F-statistic:  1938 on 2 and 722 DF,  p-value: < 2.2e-16
```

Approximately 84% of the variation in *Lung Capacity* can be explained by our model (*Age* and *Height*)

F-Statistic of Test:
"What value do I look up in a table to check on significance?"

```
Call:
lm(formula = LungCap ~ Age + Height)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4080 -0.7097 -0.0078  0.7167  3.1679

Coefficients:
              Estimate Std. Error t value 
(Intercept) -11.747065   0.476899 -24.632
Age           0.126368   0.017851   7.07
Height        0.278432   0.009926   28.
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.056 on 722 degrees of freedom
Multiple R-squared:  0.843,     Adjusted R-squared:  0.8425
F-statistic:  1938 on 2 and 722 DF,  p-value: < 2.2e-16
```

Null Hyp. Test:
The test of the null hypothesis that all model coefficients are zero.

Degrees of Freedom:
There are 725 rows in the data and three groups.
722 = 725 – 3

Our Test of The Null Hypothesis

- Ho: $\beta_1 = \beta_2 = \ldots = \beta_k$
- In our case,
  - Ho: $\beta_{age} = \beta_{height} = 0$ (slopes are zero)

Nothing is happening between the variables

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \epsilon_i$$

The p-Value:
"Is this model statistically meaningful?"

```
Call:
lm(formula = LungCap ~ Age + Height)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4080 -0.7097 -0.0078  0.7167  3.1679

Coefficients:
              Estimate Std. Error t valu
(Intercept) -11.747065   0.476899 -24.632
Age           0.126368   0.017851   7.079 3.4        ***
Height        0.278432   0.009926  28.051  < 2       ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05   ' 0.1 ' ' 1

Residual standard error: 1.056 on 722 degrees of  eedom
Multiple R-squared:  0.843,     Adjusted R-squared.  0.8425
F-statistic:  1938 on 2 and 722 DF,  p-value: < 2.2e-16
```

The p-value is very close to zero and so we reject Ho (i.e., all the model coefficients are zero (slope = 0).

There is something non-random happening in this model.

Residual Errors:

"What is the estimation of the difference between observed and predicted values?"

```
Call:
lm(formula = LungCap ~ Age + Height)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4080 -0.7097 -0.0078  0.7167  3.1679

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -11.747065   0.476899 -24.632  < 2e-16 ***
Age           0.126368   0.017851   7.079 3.45e-12 ***
Height        0.278432   0.009926  28.051  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.056 on 722 degrees of freedom
Multiple R-squared:  0.843,     Adjusted R-squared:  0.8425
F-statistic:  1938 on 2 and 722 DF,  p-value: < 2.2e-16
```

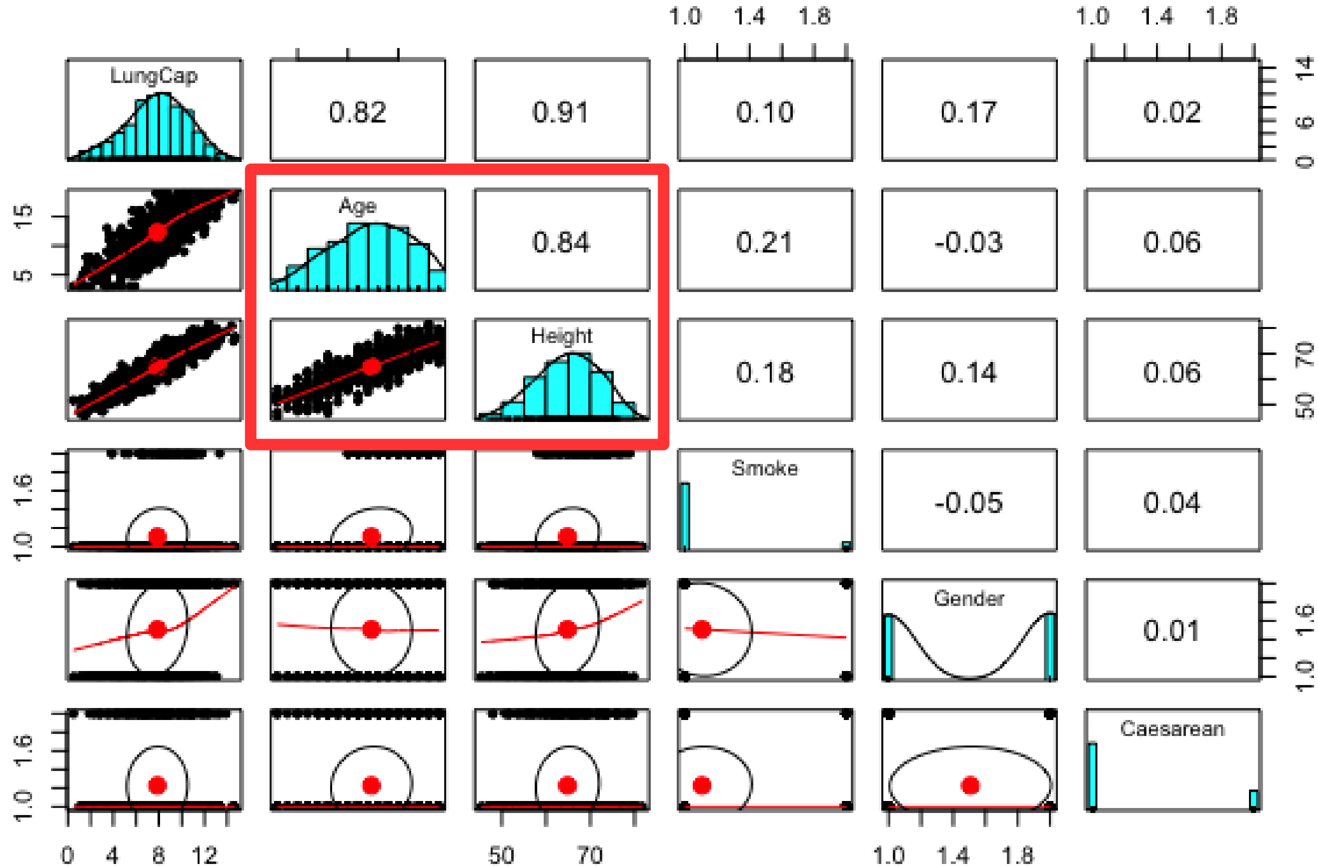
This error gives an idea about how far the observed *Lung Capacity* (dependent) values are from the predicted or fitted *Lung Capacity* (the "y-hats")

Correlation between *Age* and *Height*

- Pearson correlation between Age and Height = 0.84

```
> cor(Age, Height, method = "pearson")
[1] 0.8357368
```

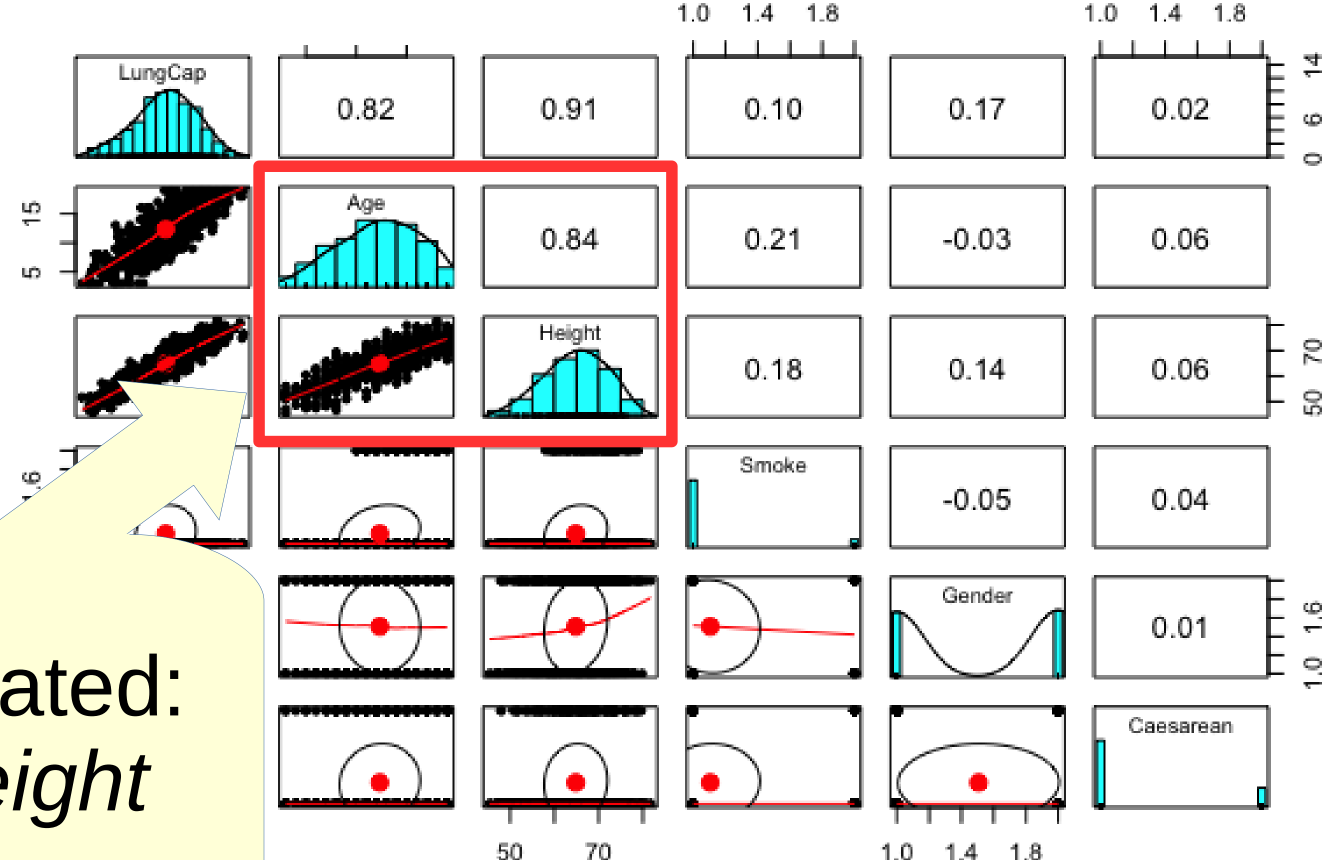
- The high correlation between *Age* and *Height* suggests that these two effects are related.



Correlation between *Age* and *Height*

- Pearson correlation between Age and Height = 0.84

```
> cor(Age, Height, method = "pearson")
[1] 0.8357368
```



Highly correlated: *age* and *height*

Correlation and Confidence

```
# Pearson correlation test
cor(Age, Height, method = "pearson")
# output: 0.8357368
# Examine the 95 per cent confidence level
confint(mod, conf.level = 0.95)
```

The **estimated slope** for Age is 0.126 and we are 95 per cent sure that the **true slope** of Age is between 0.09 and 0.16.

```
> confint(mod, conf.level = 0.95)
                  2.5 %      97.5 %
(Intercept) -12.68333877 -10.8107918
Age           0.09132215   0.1614142
Height        0.25894454   0.2979192
```

Create Bigger Model!!

- Use this data set to make a bigger model.
- Fit a linear model using ALL x variables.

Create Bigger Model!!

- `mod2 <- lm(LungCap ~ Age + Height + Smoke + Gender + Caesarean)`
- `summary(mod2)`
- `plot(mod2) # check the four visuals!`

Residuals Vs. Fitted: The relationship between *Age*, *Height* and *Lung Capacity* is approx linear.