Data Analytics
CS301
Multiple Linear Regression
Understanding the Summary

Week 10-11: 28th October
Fall 2021
Oliver BONHAM-CARTER

Up To Now in Regression

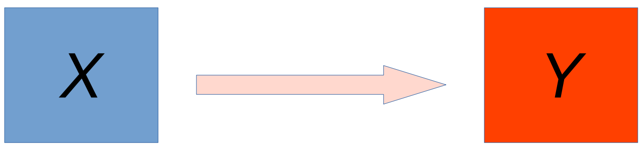
- We have discussed how one entity influences another.
- What about having two entities (independent) which may have some kind of influence on a dependent variable.
- Especially if a dependent variable has a high correlation with more multiple independent variable.

How is Grade Point Average Related to Time Studying?

- "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 healthy food AND … ??
  - *How do I begin to study this?*

So, Multiple Linear Regression Is What … ?

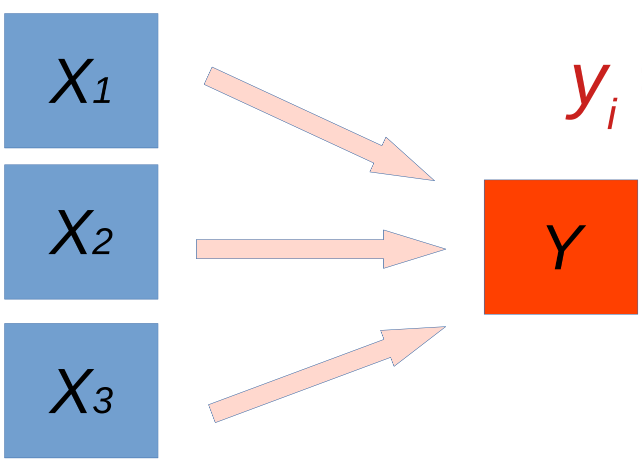
Simple regression considers a single explanatory (independent variable) and a response (dependent) variable



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

Multiple regression simultaneously considers the influence of multiple explanatory (independent variables) on a response (dependent) variable

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



Types of Questions to Address

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

Hypotheses

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

(Null Hyp) $H_0 : \beta_1 = \beta_2 \ldots \beta_{k-1} = \beta_k = 0$

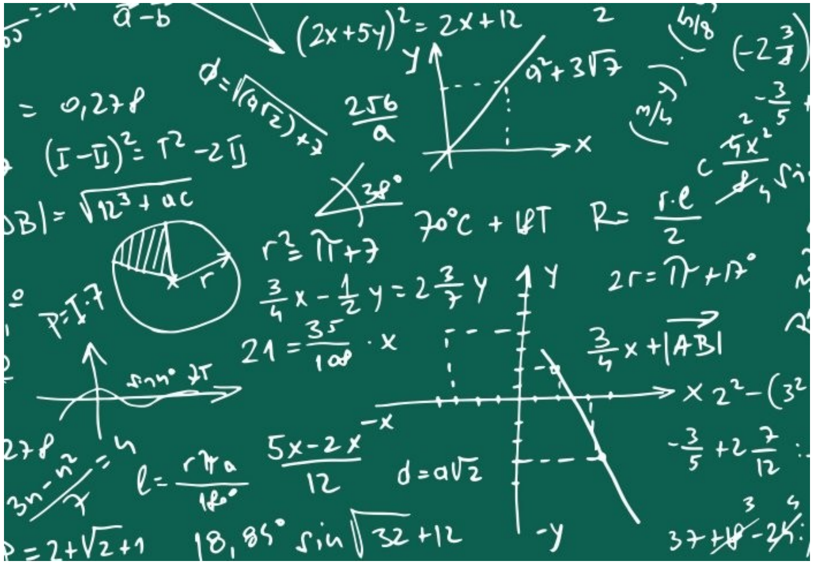
(Alt Hyp) H_A : Not all β's = 0

- The main null hypothesis of a multiple regression is that there is no relationship between the indep variables and the dep variables
- The fit of the observed dep values to those predicted by the multiple regression equation is no better than what you would expect by chance.

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)
# install.packages("psych")
library(psych)


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

Create the Multiple-Variable Regression Model

```
# model creation

mod <- lm(data = dataLungCap, 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 are 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.

Slope of Height:

"How is my *Height* variable related to *Age*?"

```
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   0.009926  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*)

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

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

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

Adjusted R-squared is a modified version of R-squared value. Value has been adjusted for the number of predictors in the model. The adjusted R-squared increases when the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected.

```
**
**
***
---
Signif. codes:  0 ‘***’ 0.001 ‘*           *’ 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
```

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 -24.6
Age           0.126368   0.017851    7.
Height        0.278432   0.009926   28
---
Signif. codes:  0 ‘***’ 0.001 ‘**’  .01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.056 o  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

Used also for finding values in F-table for hypothesis testing

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)
  - **Ha**: $\beta_{age} \mathrel{!=} \beta_{height} \mathrel{!=} 0$ (slopes not all zeros)

Nothing is happening between the k-number of 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
-3.4080 -0.7097 -0.0078  0.7167  3.1

Coefficients:
              Estimate Std. Error t
(Intercept) -11.747065   0.476899 -24.
Age           0.126368   0.017851   7.079         2 ***
Height        0.278432   0.009926  28.051         5 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.      .' 0.1 ' ' 1

Residual standard error: 1.056 on 722 degrees of    reedom
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 the Ho (i.e., all the model coefficients are zero (slope = 0).

Conclusion: 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”)

Assumptions: Linear Relationships

- A multiple regression model relating a *y*-dependent variable to *p* multiple independent variables

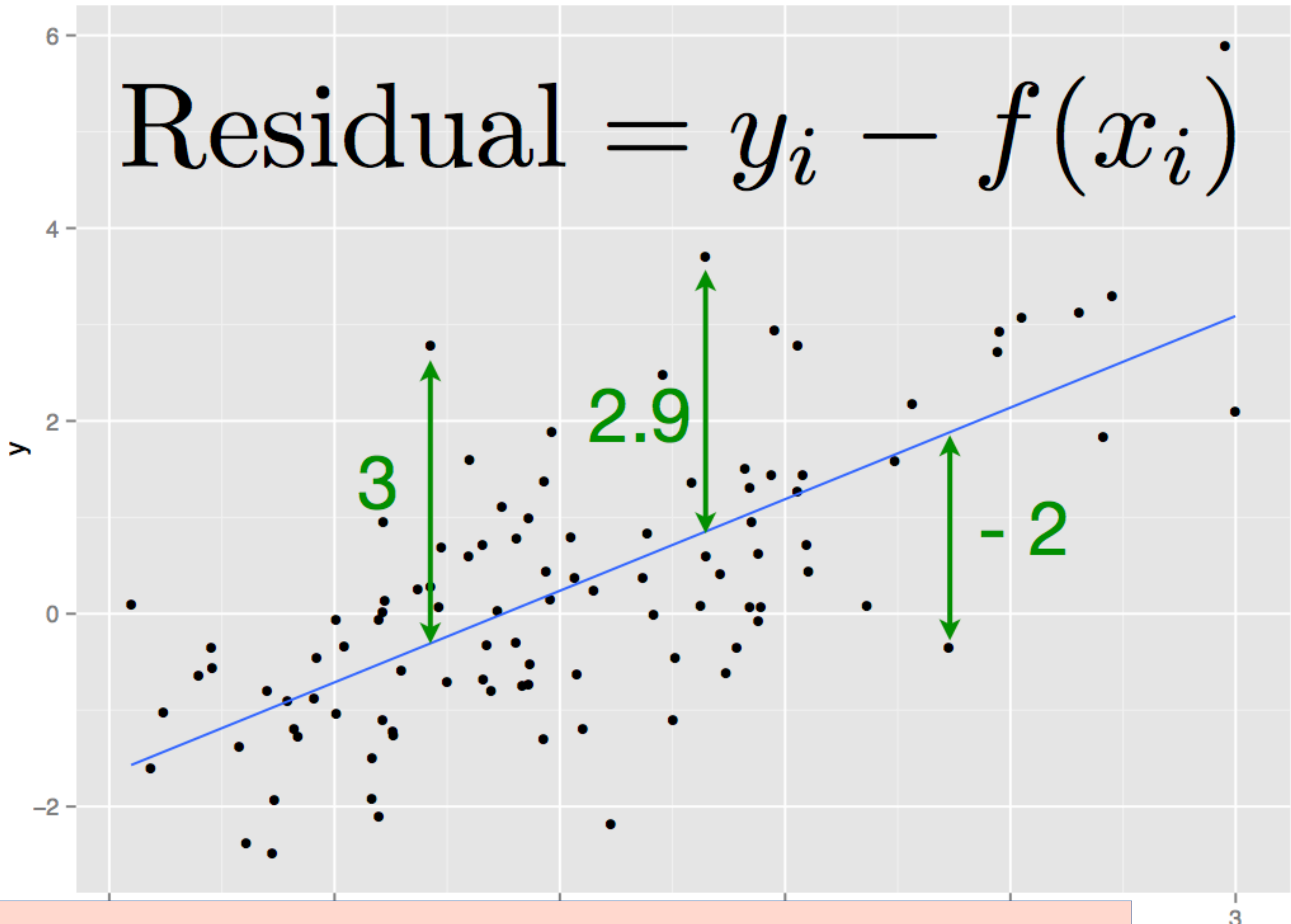
$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_{p-1} x_{i,p-1} + \varepsilon_i$$

- There must be a linear relationship between the independent variable and the independent variables.
  - Use a scatterplots to explore a linear or curvilinear relationship.

See sandbox/assumptions.R for quick demo of assumptions

Assumptions: Residuals (From Last Time)

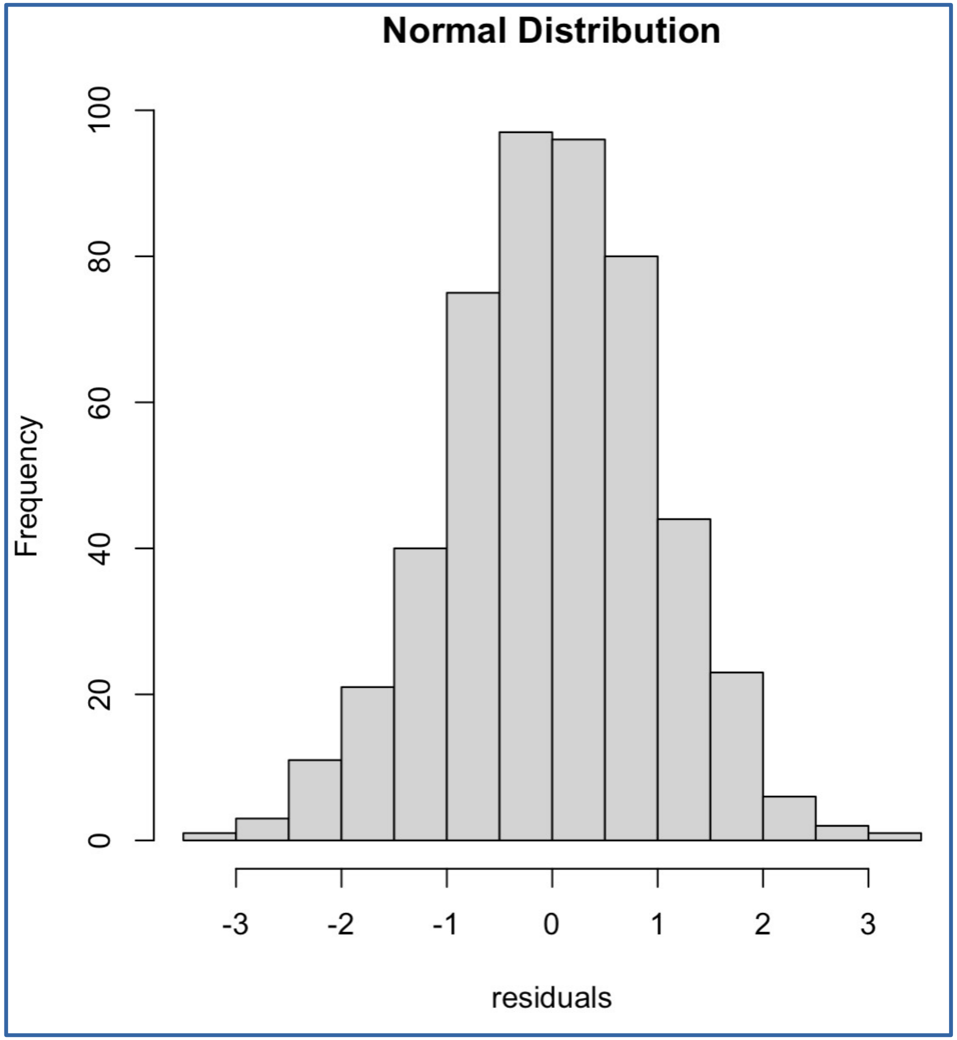
- A *residual* of an observed value is the difference between the observed value and the estimated value of the quantity of interest



See sandbox/assumptions.R for quick demo of assumptions

Assumptions: Residual Normal Distribution

- Multivariate Normality– Multiple regression assumes that the residuals are normally distributed.



See sandbox/assumptions.R for quick demo of assumptions

Assumptions: No Multicollinearity

- Multicollinearity occurs when independent variables in a regression model are correlated.
- This correlation is a problem because independent variables should be independent (of each other!).
- If the degree of correlation between variables is high enough, it can cause problems when you fit the model and interpret the results.

- Multiple regression assumes that the independent variables are not highly correlated with each other. This assumption is tested using Variance Inflation Factor (VIF) values.

See sandbox/assumptions.R for quick demo of assumptions

Test for Correlation Between *All* Variables

```
library(psych)
pairs.panels(dataLungCap)
pairs.panels(dataLungCap, lm = TRUE)
```

Allows for a quick study of correlation across the variables of your study

pairs.panels {psych} R Documentation

SPLOM, histograms and correlations for a data matrix

Description

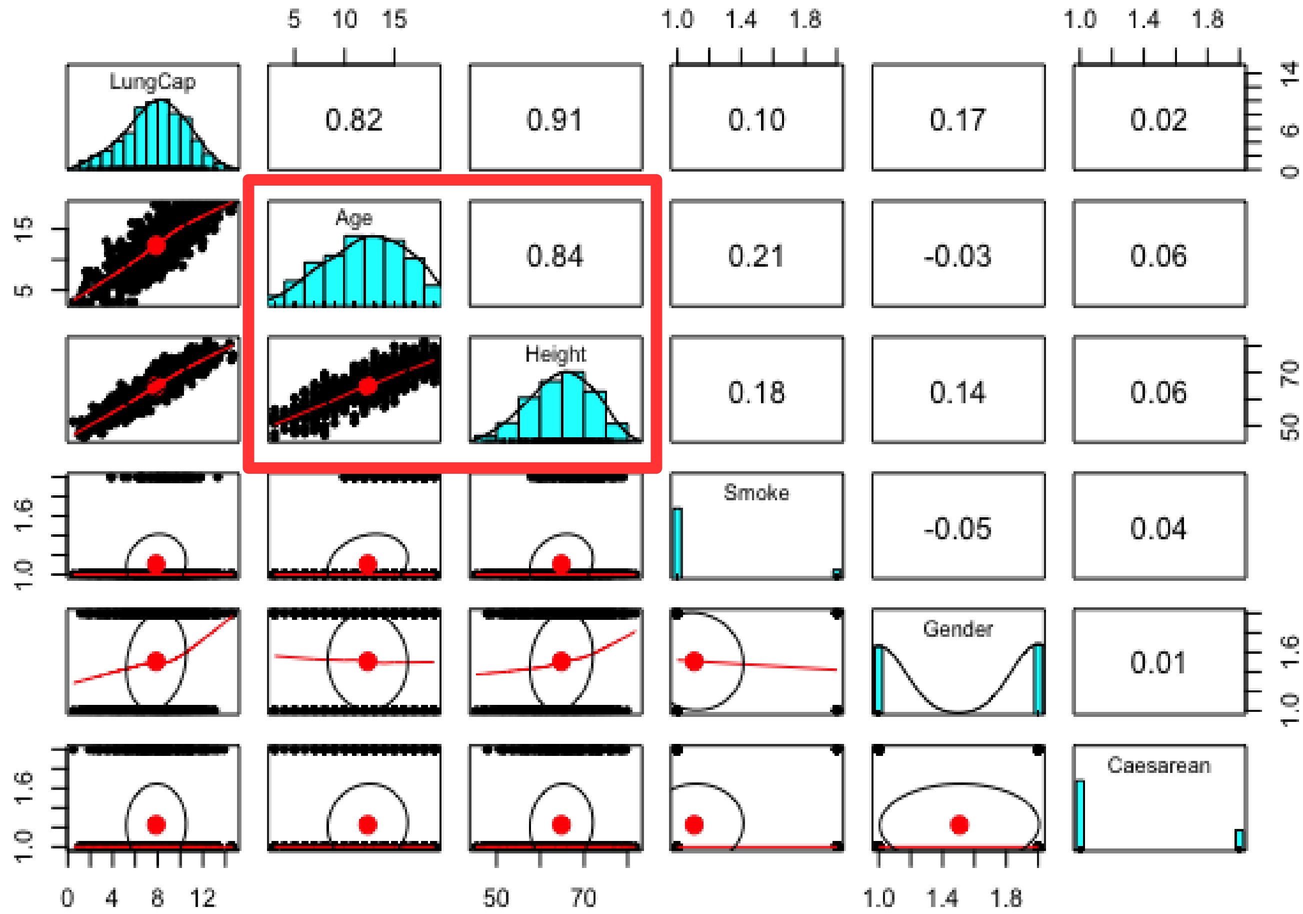
Adapted from the help page for pairs, pairs.panels shows a scatter plot of matrices (SPLOM), with bivariate scatter plots below the diagonal, histograms on the diagonal, and the Pearson correlation above the diagonal. Useful for descriptive statistics of small data sets. If lm=TRUE, linear regression fits are shown for both y by x and x by y. Correlation ellipses are also shown. Points may be given different colors depending upon some grouping variable. Robust fitting is done using lowess or loess regression. Confidence intervals of either the lm or loess are drawn if requested.

Correlation Between *Age* and *Height*

- Pearson correlation between Age and Height = 0.84

```
cor(dataLungCap$Age, dataLungCap$Height)
```

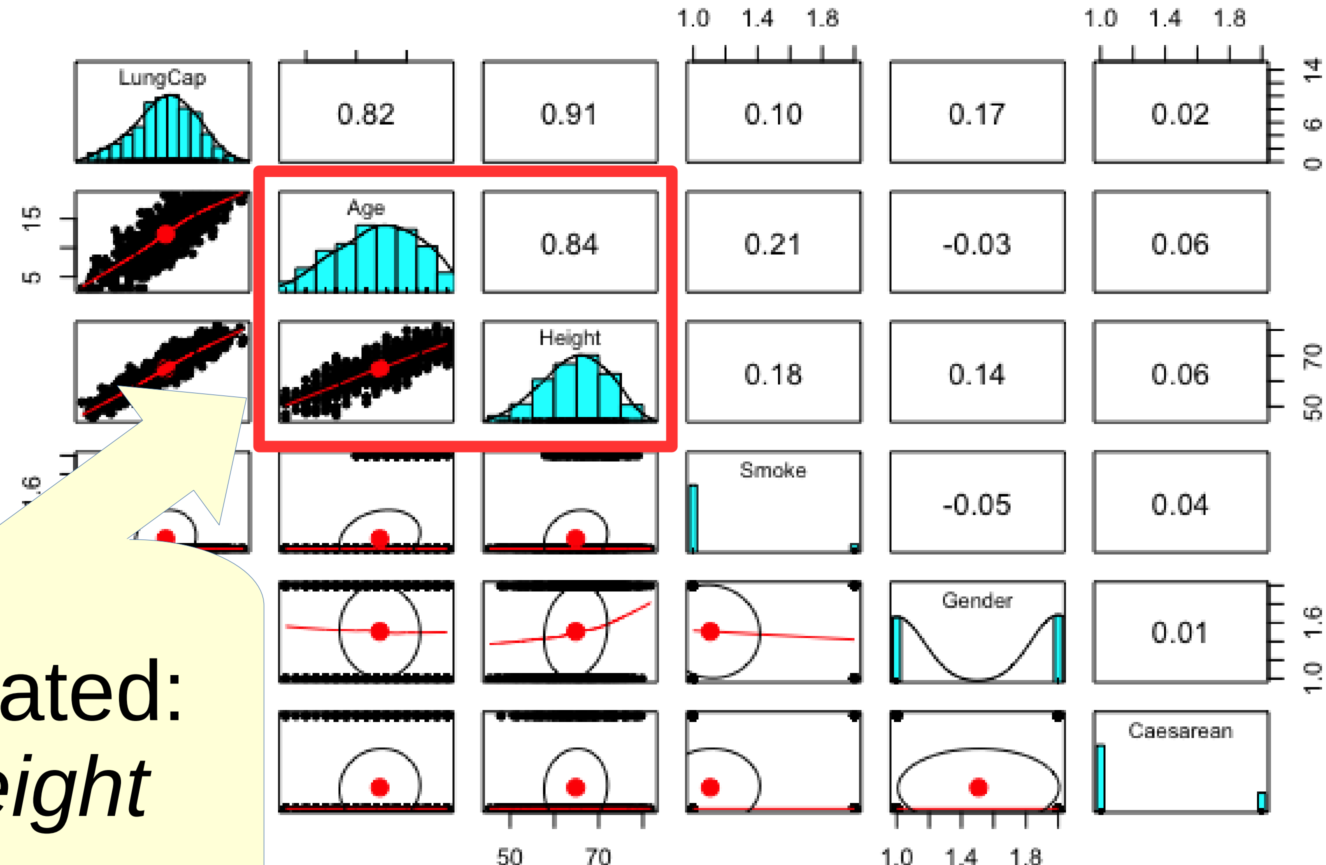
- The high correlation between *Age* and *Height* suggests that these two effects are related.



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

(Remember that we are studying variable slope)

```
# Pearson correlation test
cor(dataLungCap$Age, dataLungCap$Height)
# output: 0.8357368
# Examine the 95 percent confidence level
confint(mod, conf.level = 0.95)
```

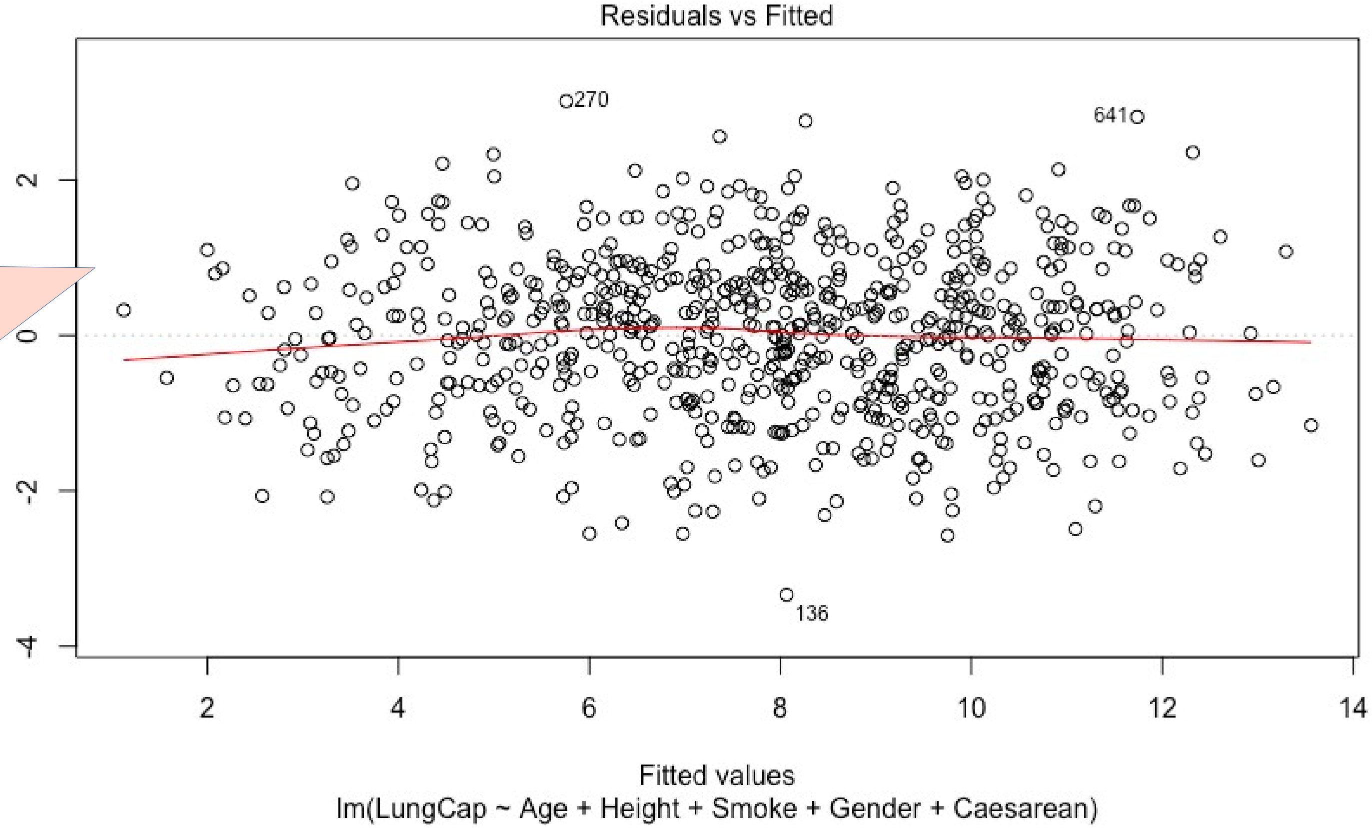
The **estimated slope** for Age is 0.126 and we are 95 percent sure that the **true slope** of *Age* is between 0.09 and 0.16.

```
> 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 a Bigger Model!!

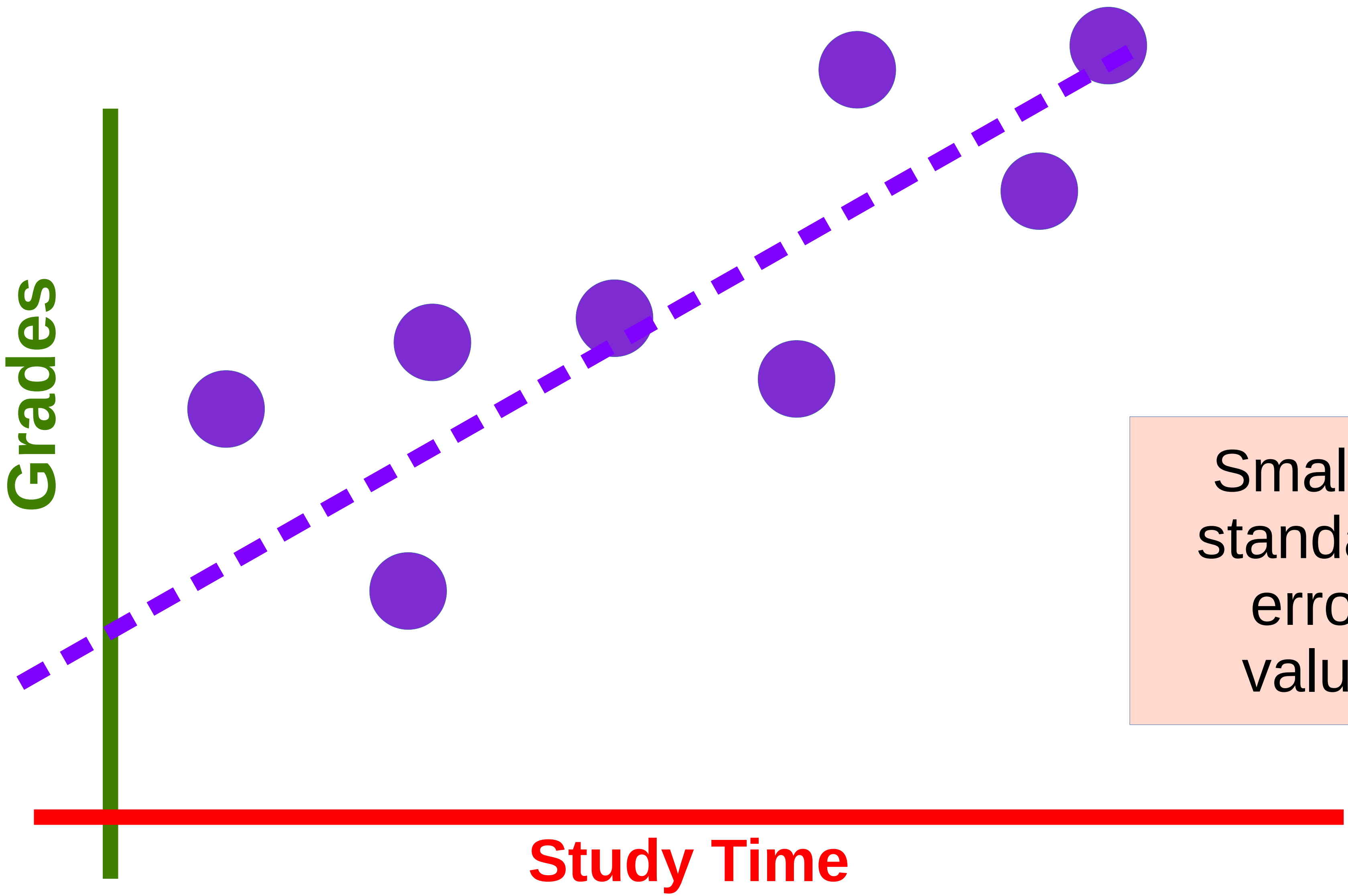
- `mod2 <- lm(data = dataLungCap, LungCap ~ Age + Height + Smoke + Gender + Caesarean)`
- `summary(mod2)`
- `plot(mod2) # check the four plots!`

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



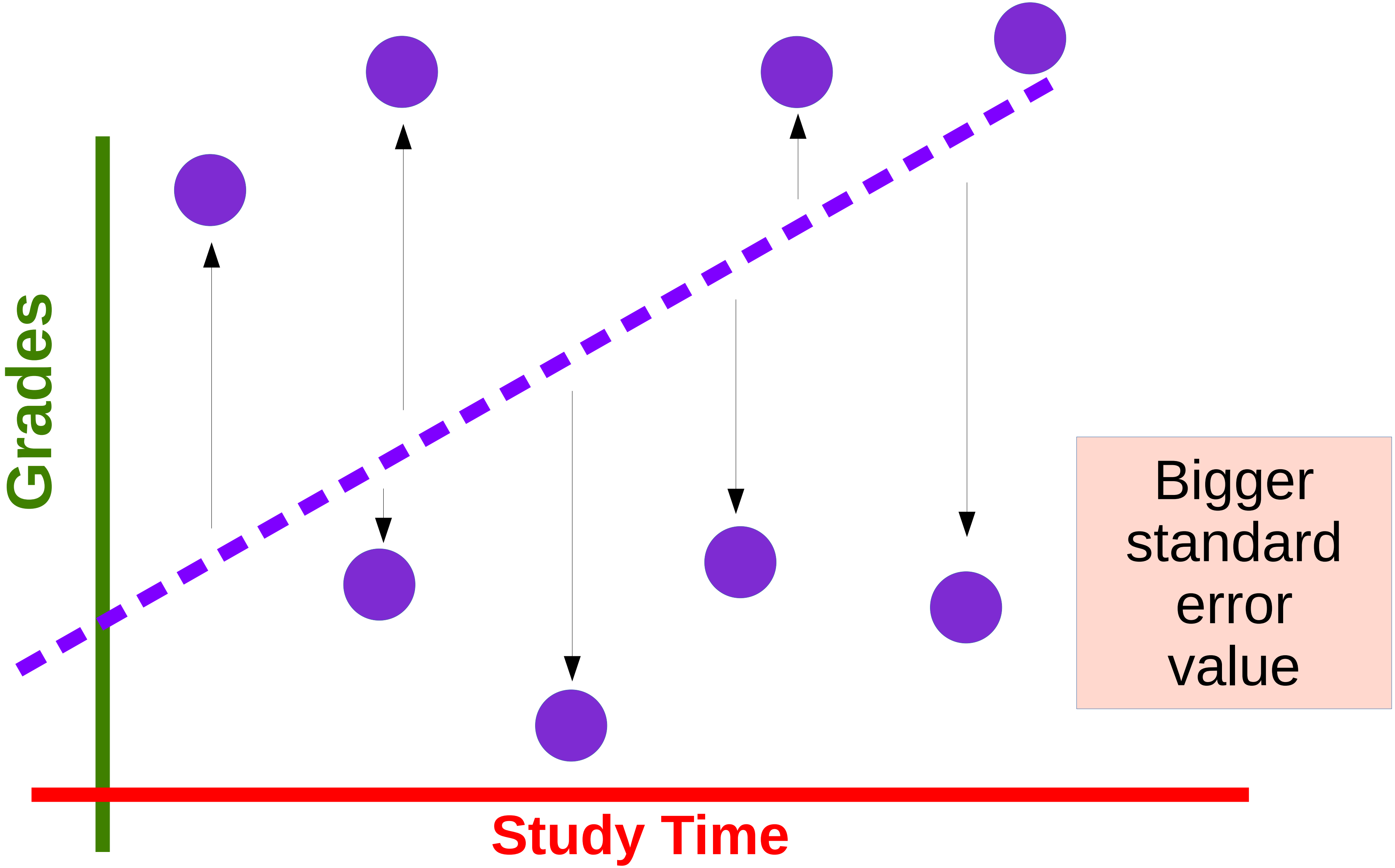
Lemme Say Something More About Residuals!!

Assumptions: Residuals
”Good” Residual Standard Error

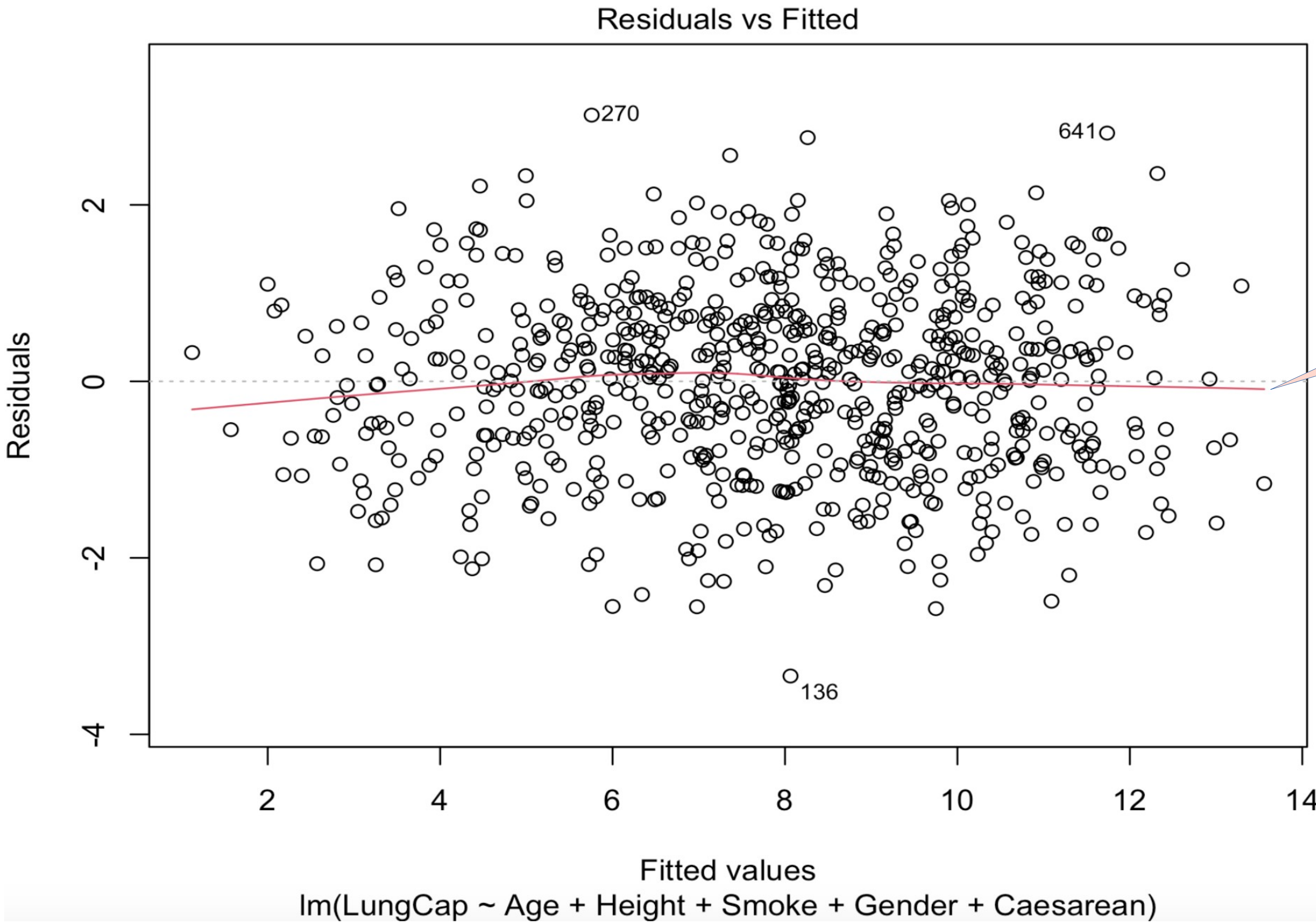


Smaller standard error value

Assumptions: Residuals
”Bad” Residual Standard Error



Residuals Vs. Fitted

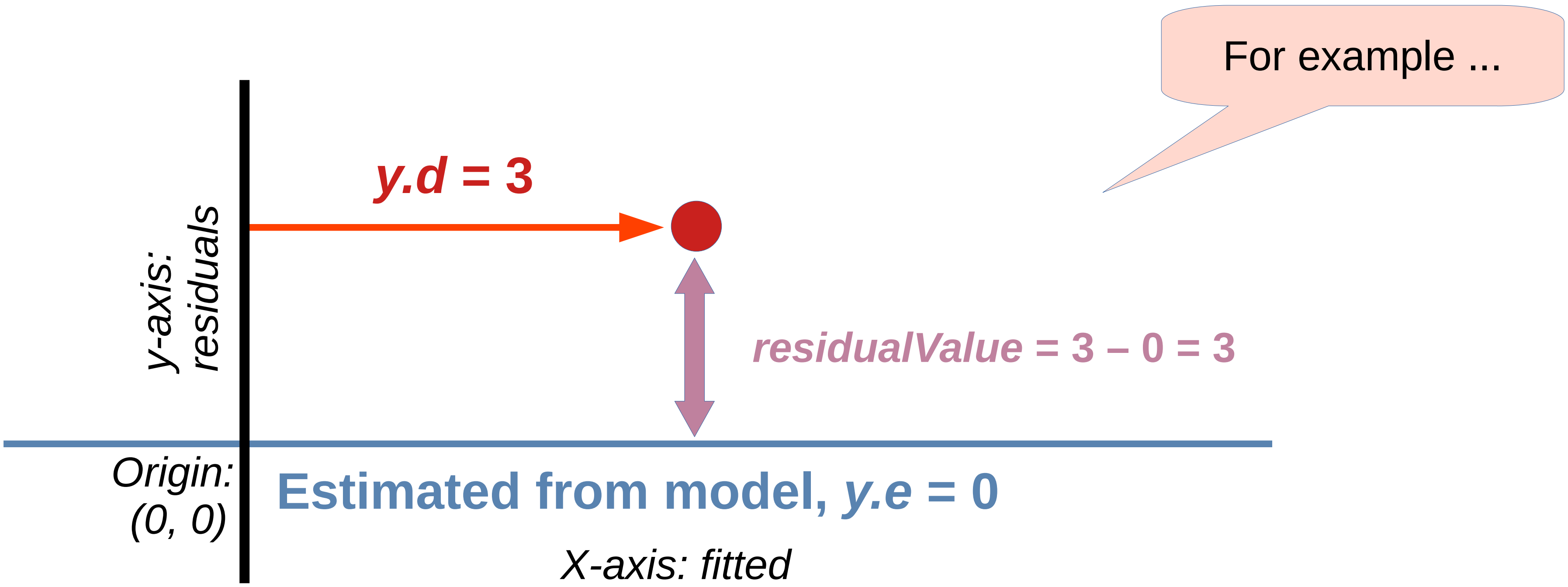


Expected

- Points are vertical distances between *y.d* (i.e., a datapoint) and *y.e* (i.e., an estimated point)

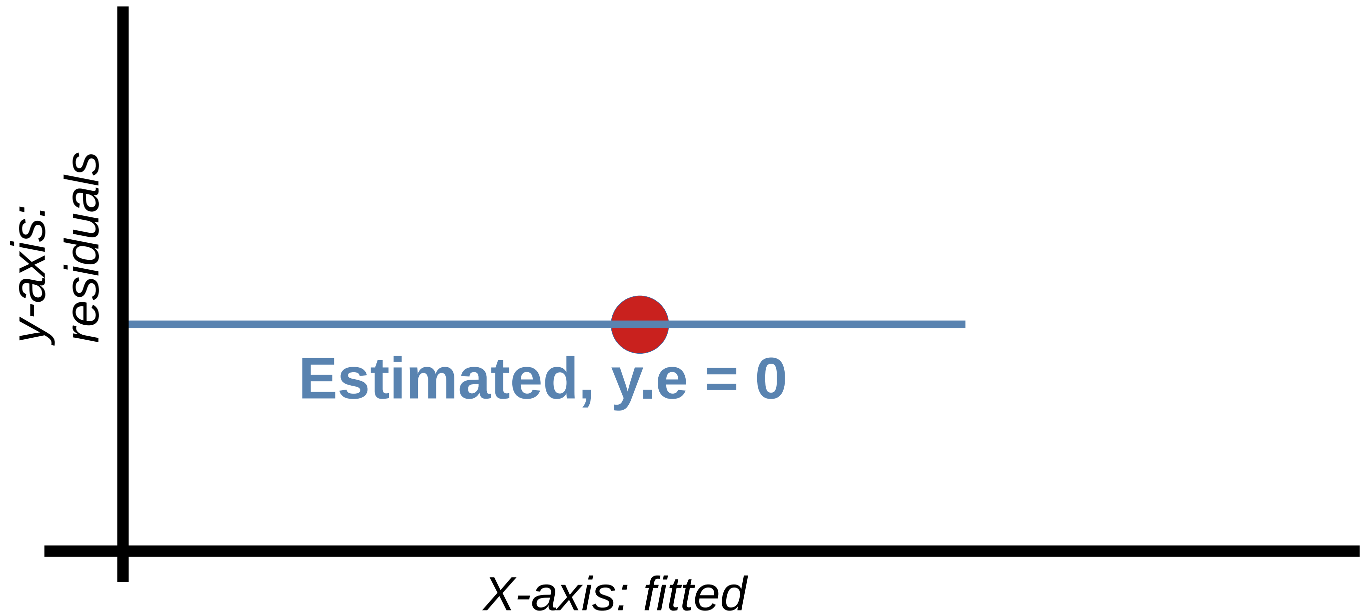
Residuals Vs. Fitted

- **Residual value**: a vertical distance between any one data point *y.d* (i.e.,datapoint) and the estimated value *y.e* (i.e.,estimated)
- *residualValue* = *y.d* – *y.e*

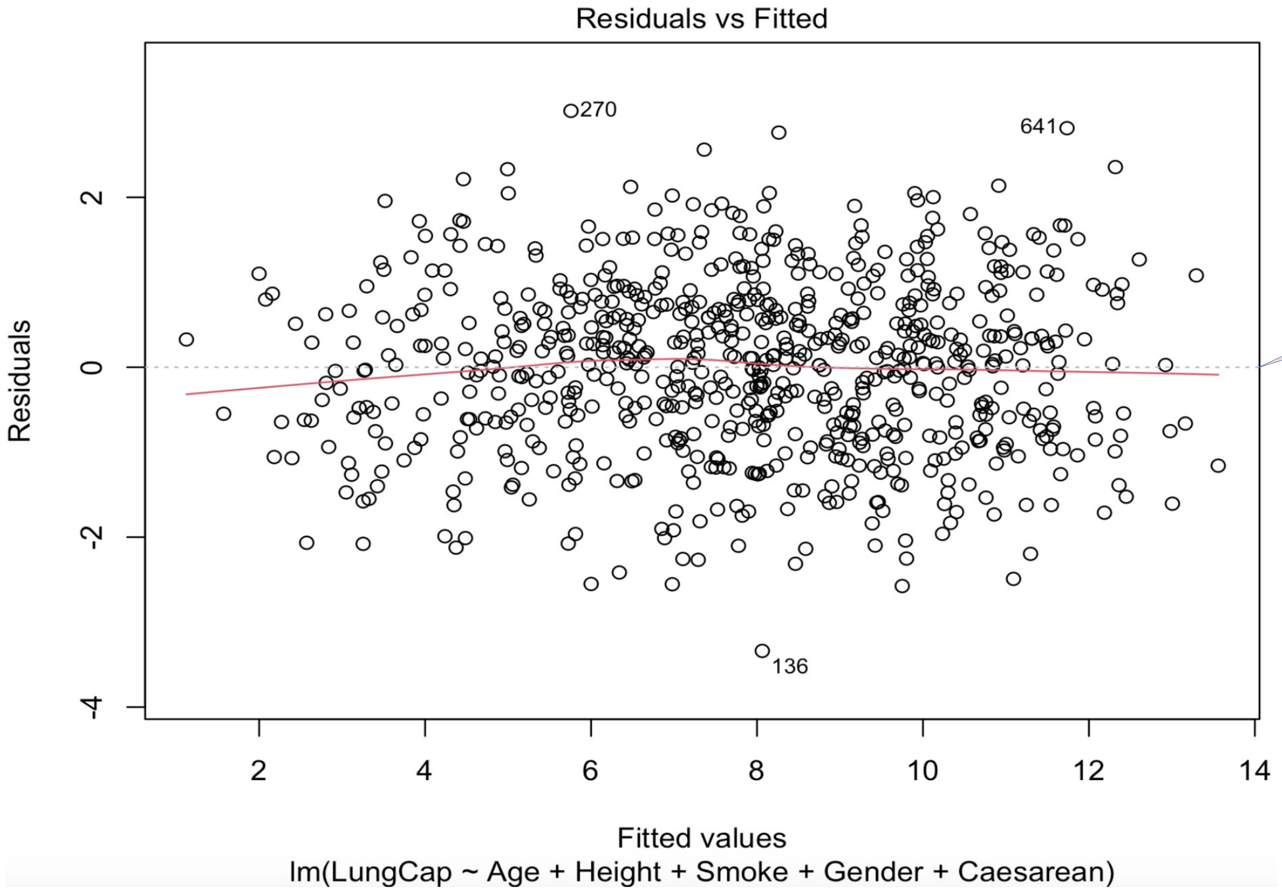


Residuals Vs. Fitted

- Any data point found (directly) on the estimated regression line has a residual of 0.
- The residual = 0 line corresponds to the estimated regression line.
- Suggestions for the *Appropriateness* of linear regression model



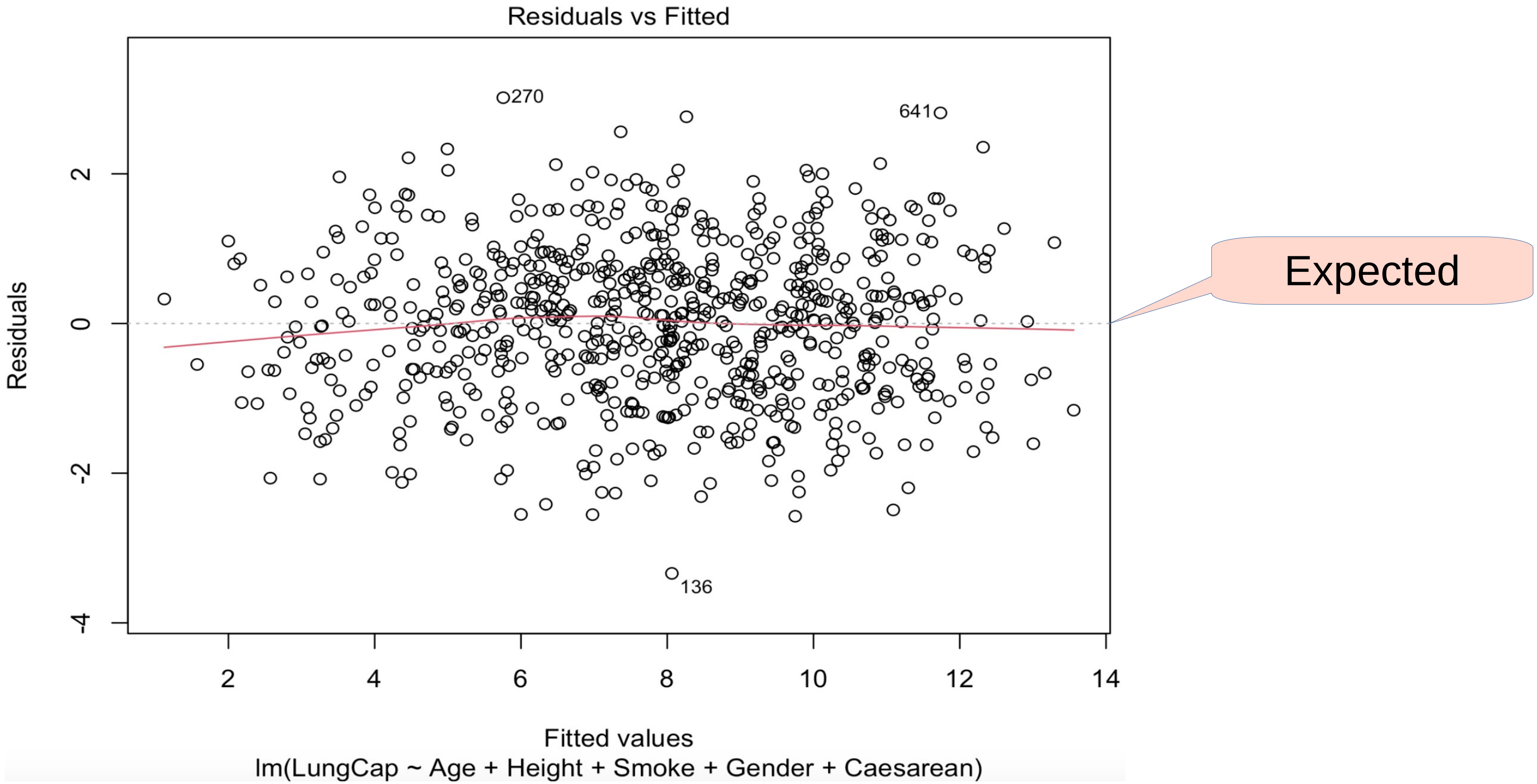
Suggestion: 1



Expected

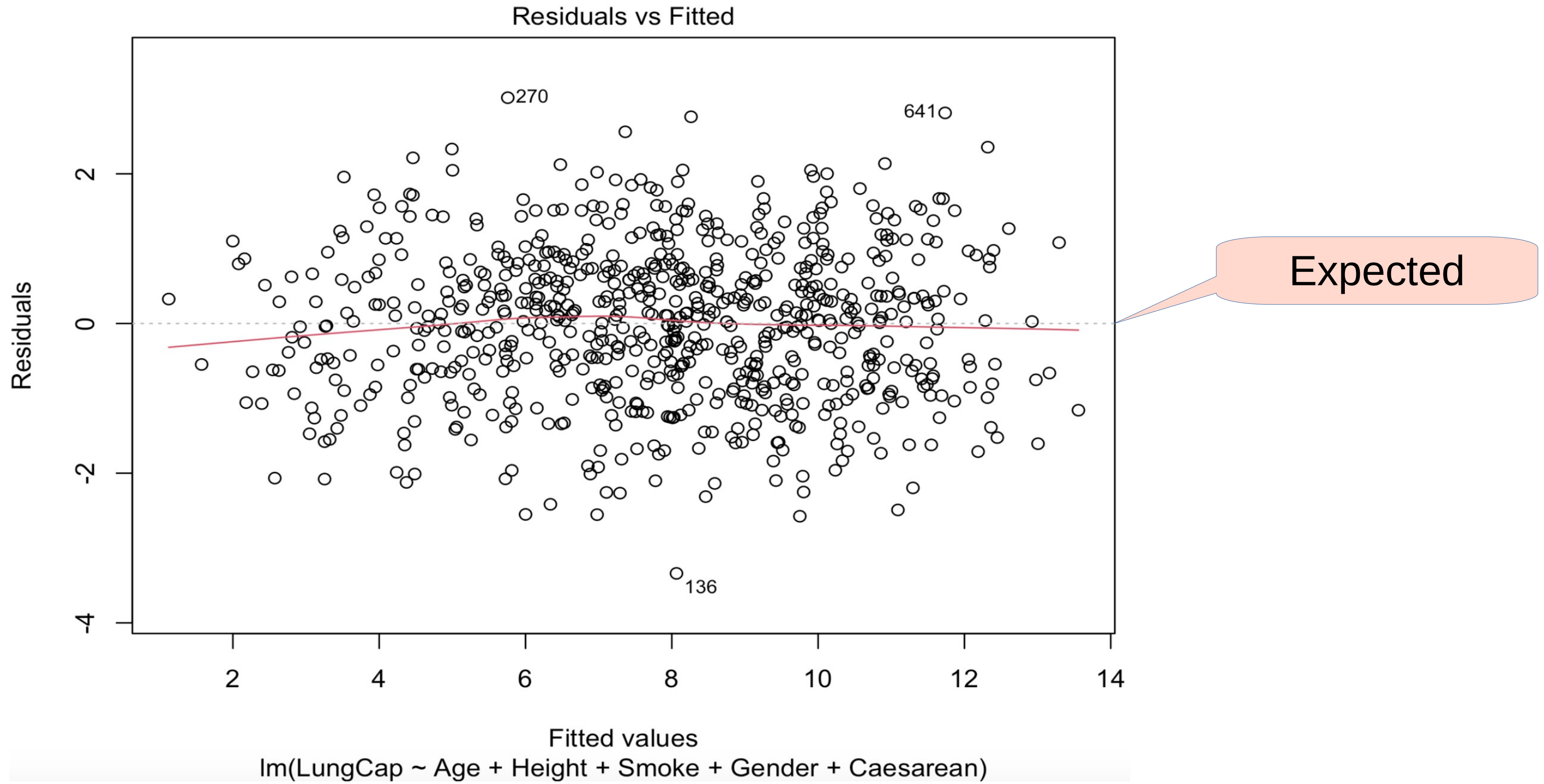
- The residuals are generally (randomly) situated around the 0 line.
- This suggests that the assumption that the relationship is linear is reasonable.

Suggestion: 2



- The residuals roughly form a "horizontal band" around the 0 line.
- This suggests that the variances of the error terms are equal.

Suggestion: 3



- Few residuals are "standing out" from the basic random pattern of residuals.
- This suggests that there are few outliers
- Is further study necessary to explain?

Go Create a Bigger Model!!

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

Or try a different data set and do more model fitting!