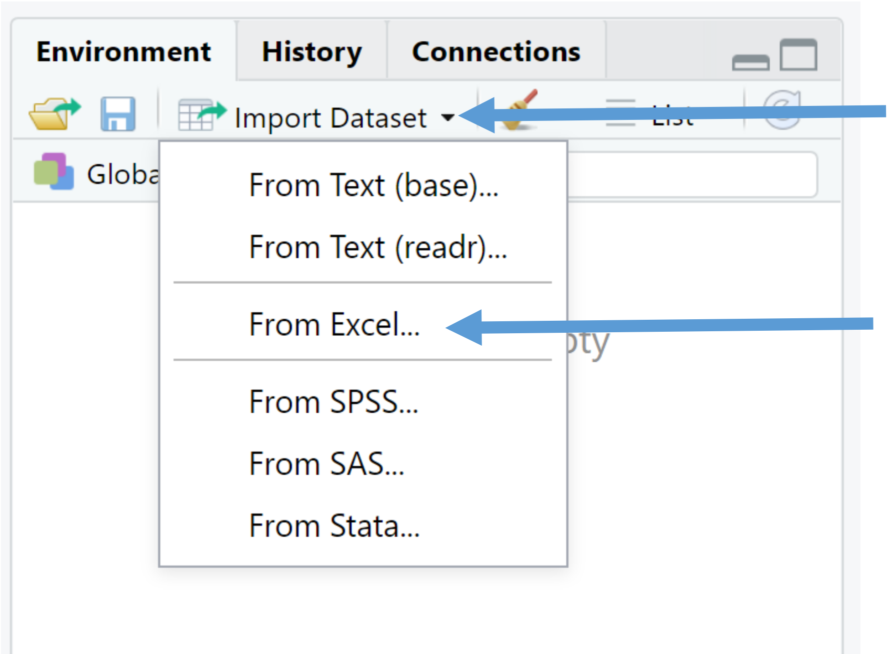
Simple Linear Regression Using RStudio to Model Weight and Height

STAT 441/541 Statistical Methods II

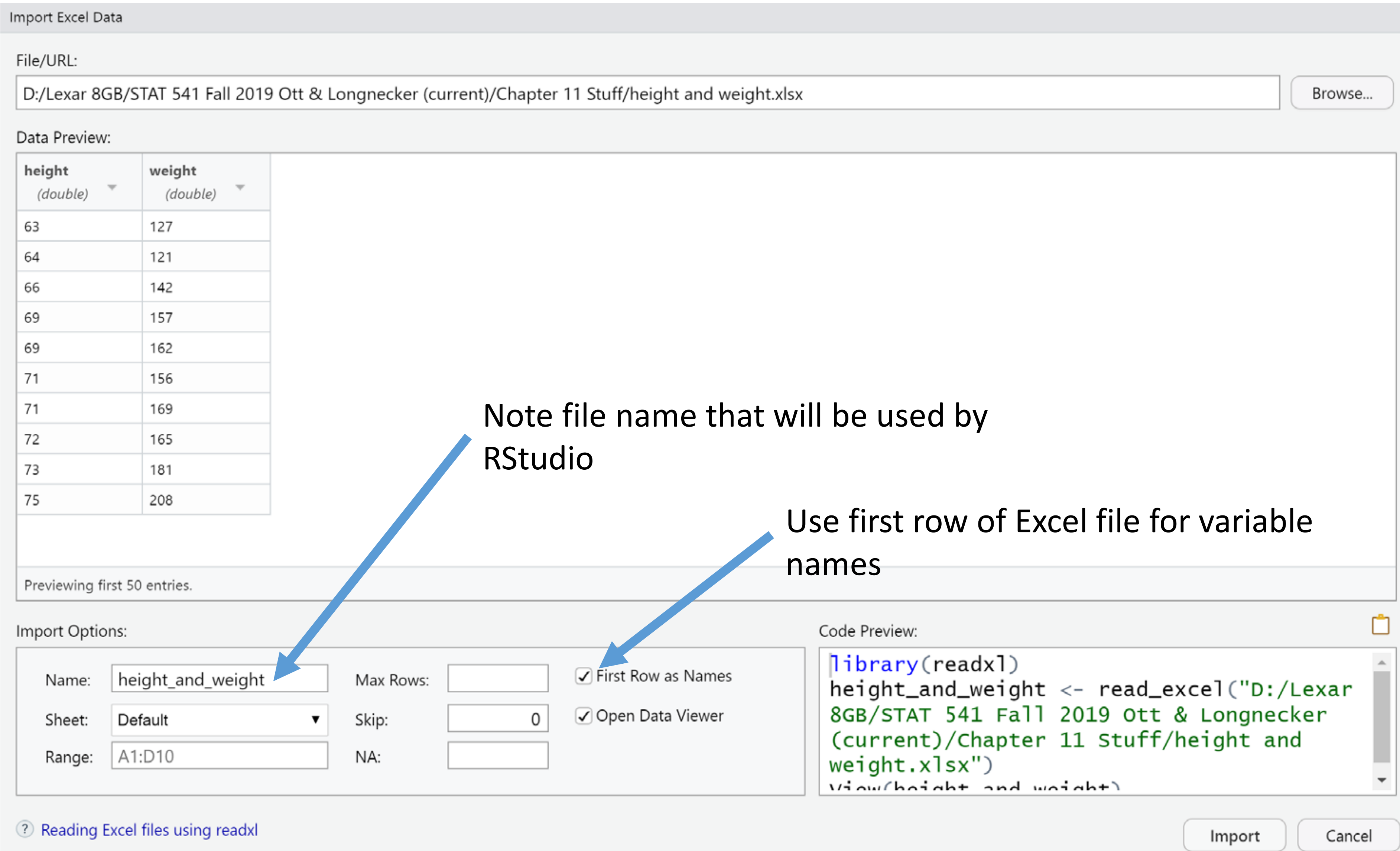
R Code and Excel File

- Simple Linear Regression height weight example.R
- height and weight.xlsx

RStudio Import Dataset



Window for "Import Excel Data"



height (double)	weight (double)
63	127
64	121
66	142
69	157
69	162
71	156
71	169
72	165
73	181
75	208

Note file name that will be used by RStudio

Use first row of Excel file for variable names

R Code for Height and Weight Example

Put the cursor on the first row

Then click Run


```
#
# Chapter 11 Simple Linear Regression for height and weight data
#

# In RStudio, use File, Import Dataset, From Excel...
#  to get Excel data file

# Note the name for the imported Excel file
str(height_and_weight)
```

```
> # Note the name for the imported Excel file
> str(height_and_weight)
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':      10 obs. of  2 variables:
 $ height: num  63 64 66 69 69 71 71 72 73 75
 $ weight: num  127 121 142 157 162 156 169 165 181 208
> 
```

R Code for Height and Weight Example

Click Run two times to process the next two active lines of code

```
# To have most of our R code reuseable for future
#  analyses, we will use a data object called dataobj
dataobj <- as.data.frame(height_and_weight)
str(dataobj)
```

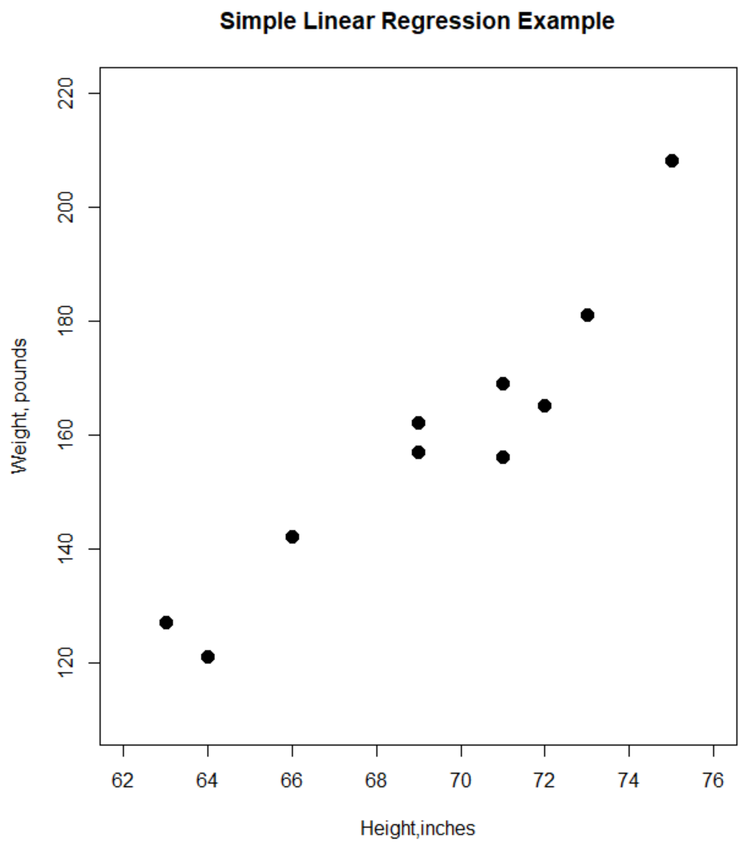
```
> # To have most of our R code reuseable for future
> #  analyses, we will use a data object called dataobj
> dataobj <- as.data.frame(height_and_weight)
> str(dataobj)
'data.frame':   10 obs. of  2 variables:
 $ height: num  63 64 66 69 69 71 71 72 73 75
 $ weight: num  127 121 142 157 162 156 169 165 181 208
```

Our data is now in a data frame called dataobj ready for regression analysis

Scatterplot

Click Run two times to process the next two active lines of code

```
# Scatterplot to examine relationship
par(mfrow=c(1,1))
plot(dataobj$height,dataobj$weight,xlab="Height,inches", ylab="Weight, pounds",
     ylim=c(110,220),xlim=c(62,76),main="Simple Linear Regression Example",
     pch=19,cex=1.5)
```



It is always a good idea to examine a scatterplot of the data. The dependent variable goes on the vertical axis and the independent variable goes on the horizontal axis.

Always comment on the nature of the relationship between the dependent and independent variable.

R Code for Height and Weight Example

Use the lm function and store results in object called "model"

Then use summary function to get regression results

```
> # model weight as a function of height
> model <- lm(weight ~ height , data=dataobj)
> summary(model)

Call:
lm(formula = weight ~ height, data = dataobj)

Residuals:
     Min       1Q   Median       3Q      Max 
-13.2339  -4.0804  -0.0963   4.6445  14.2158 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -266.5344    51.0320  -5.223    8e-04 ***
height         6.1376     0.7353   8.347 3.21e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.641 on 8 degrees of freedom
Multiple R-squared:  0.897,     Adjusted R-squared:  0.8841 
F-statistic: 69.67 on 1 and 8 DF,  p-value: 3.214e-05
```

Coefficients Table

Estimates σ_ε

Regression Summary Information

Model, Prediction Equation, and Estimated Model

Use the Coefficients table to get parameter estimates for the intercept and slope of the independent variable

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -266.5344    51.0320  -5.223    8e-04 ***
height         6.1376     0.7353   8.347 3.21e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Model for weight and height: $weight = \beta_0 + \beta_1 height + \varepsilon$

Prediction equation for weight and height: $\widehat{weight} = \hat{\beta}_0 + \hat{\beta}_1 height$

Estimated model for weight and height: $\widehat{weight} = -266.5344 + 6.1376 * height$

Hypothesis Test for Slope Parameter

We also use information in the Coefficients Table to test hypotheses for the slope parameter, usually at $\alpha = 0.05$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -266.5344    51.0320  -5.223    8e-04 ***
height         6.1376     0.7353   8.347 3.21e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Test slope parameter for independent variable height

Using the five-step method to test the slope parameter β_1:

Hypotheses: H_0: $\beta_1 = 0$

H_a: $\beta_1 \neq 0$

Test Statistic: 8.347

P-value: 0.0000321

Decision about the null hypothesis: Reject the null hypothesis since p-value < alpha=0.05

Conclusion: Since $\beta_1 \neq 0$ there is a statistically significant relationship between weight and height

Confidence Interval on Slope Parameter β_1

Run the next line of code to get confidence intervals on model parameters β_0 and β_1

```
29  confint(model,level=0.95)
```

```
> confint(model,level=0.95)
                 2.5 %      97.5 %
(Intercept) -384.214357 -148.854434
height         4.441894    7.833269
```

The difference between the upper confidence and lower confidence tells us we have 97.5% - 2.5% = 95% confidence

Confidence limits for the slope parameter for independent variable height

The interpretation of confidence intervals consists of three parts: (1) The level of confidence, (2) The confidence limits, and (3) The parameter that we are estimating

The interpretation for the slope parameter for the weight and height example:
We are 95% confident that the interval from 4.44 to 7.83 captures the true slope parameter β_1. So plausible values of the slope parameter for height are 4.44 to 7.83.

Confidence Intervals on $E(y)$

Run the next lines of code to get confidence intervals on $E(y)$ when height=68 inches

```r
# confidence interval for mean weight given height=68 inches
newdatamu <- data.frame(height=68)
predict(model,newdatamu,interval="confidence")
```

```
> # confidence interval for mean weight given height=68 inches
> newdatamu <- data.frame(height=68)
> predict(model,newdatamu,interval="confidence")
       fit      lwr      upr
1 150.8211 144.1452 157.4971
```

Lower and Upper Confidence Limits when height=68 inches

Predicted value when height=68 inches

Interpretation of confidence interval on $E(y)$:
We are 95% confident that the interval from 144.15 to 157.50 pounds captures the mean weight when height=68 inches.

Prediction Intervals on y

Run the next lines of code to get prediction intervals on y when height=68 inches

```r
# prediction interval of weight for height=68 inches
newdatay <- data.frame(height=68)
predict(model,newdatay,interval="predict")
```

```
> # prediction interval of weight for height=68 inches
> newdatay <- data.frame(height=68)
> predict(model,newdatay,interval="predict")
       fit      lwr      upr
1 150.8211 129.8056 171.8367
```

Lower and Upper Prediction Limits when height=68 inches

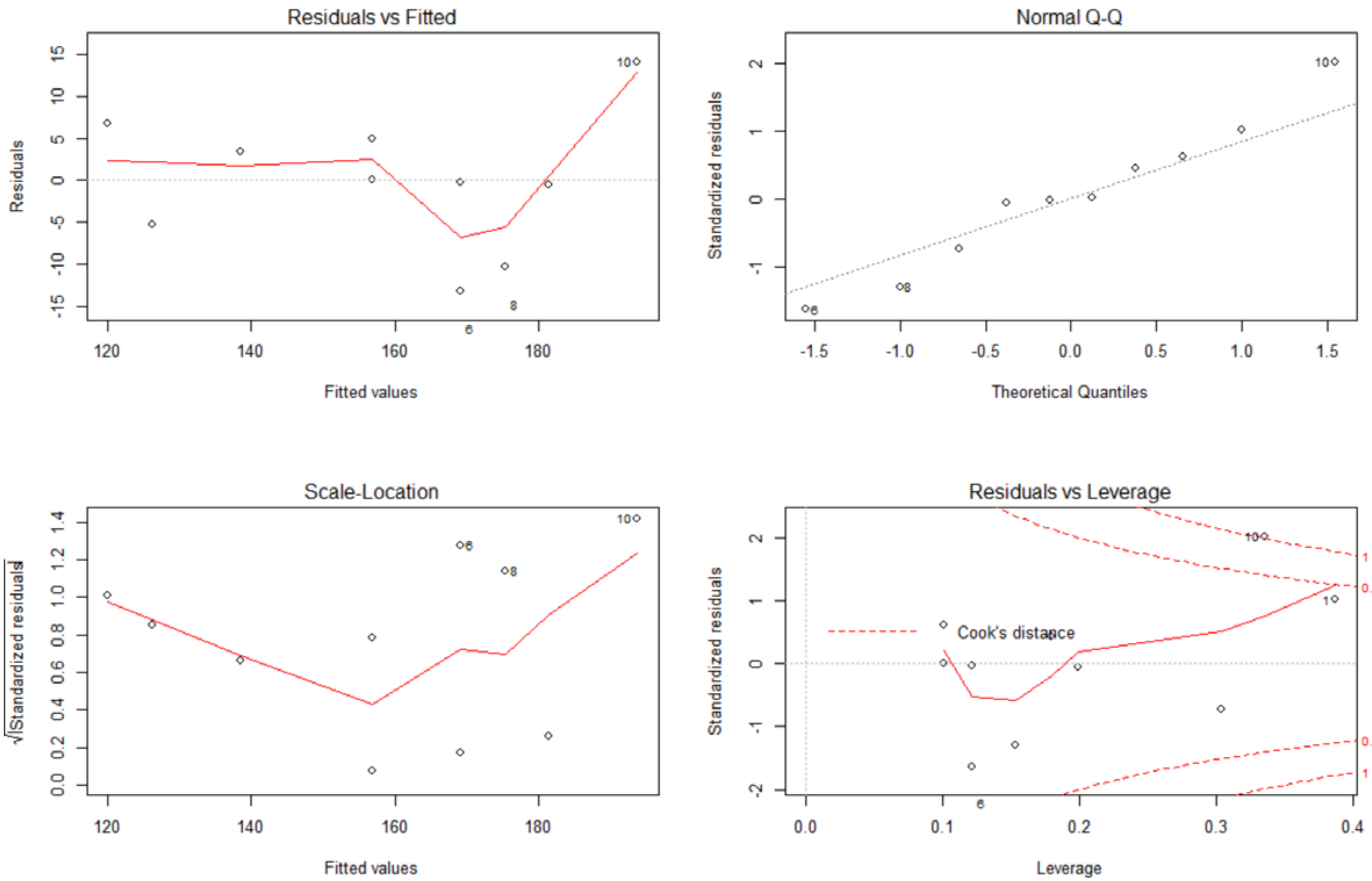
Predicted value when height=68 inches

Interpretation of prediction interval on y for weight and height example:
We are 95% confident that the interval from 129.81 to 171.84 pounds captures an individual weight when their height is 68 inches.

Diagnostic Plots for Height and Weight Example

```
> plot(model)
```

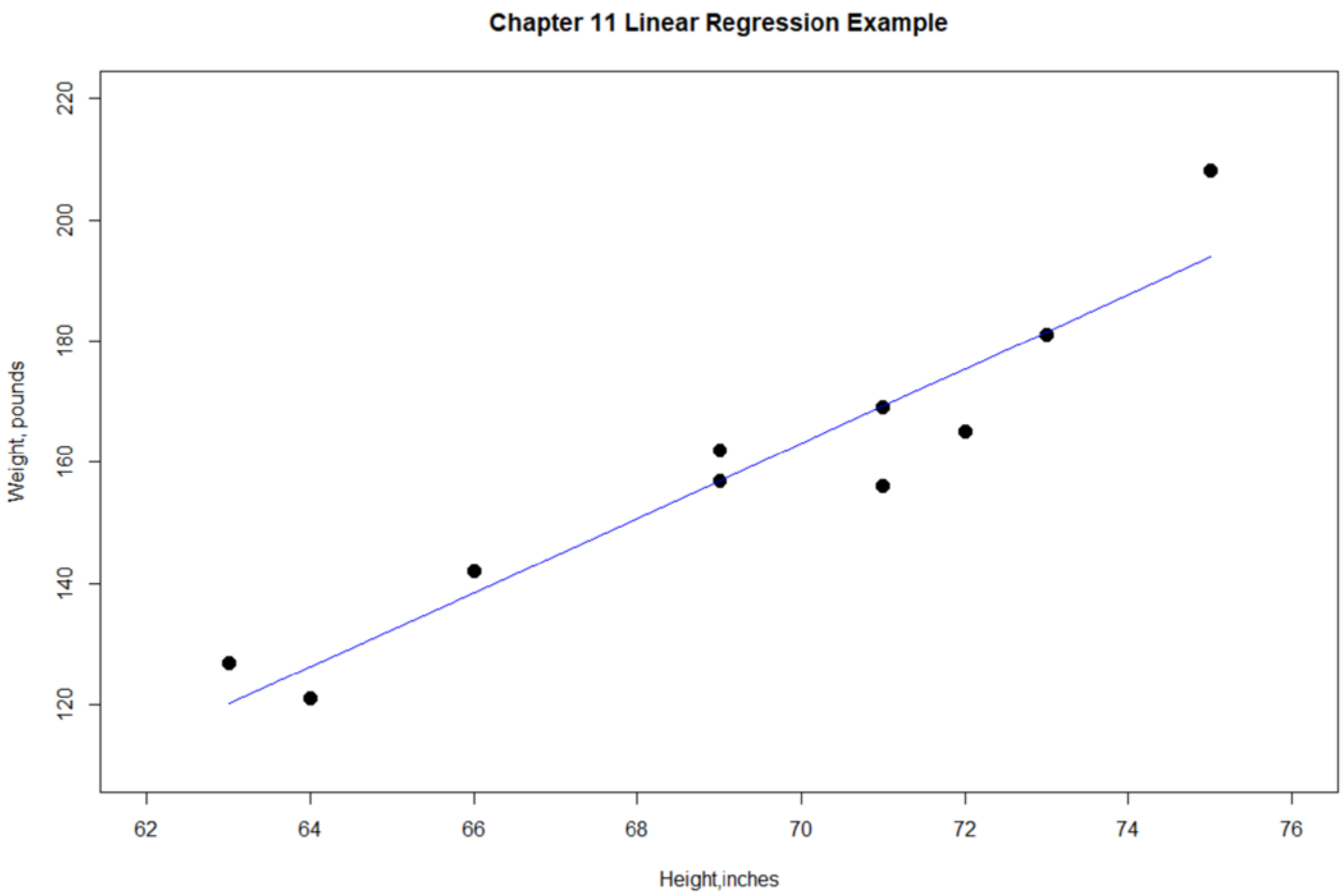
This line of code generates four diagnostic plots to check assumptions for regression analysis. We will cover this in Multiple Regression.



R Code for Height and Weight Example

These lines of code will put one graph in the plot window, create a scatterplot of the data, and then add the predicted regression line

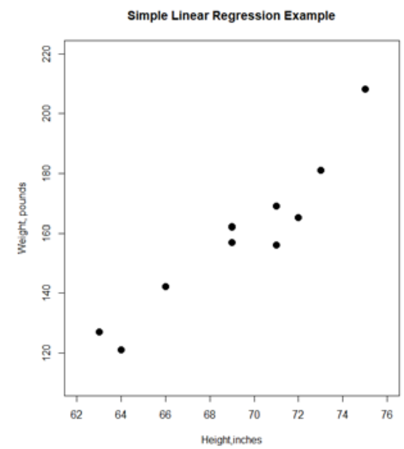
```
> par(mfrow=c(1,1))
> plot(dataobj$height,dataobj$weight,xlab="Height,inches", ylab="Weight, pounds",
+    ylim=c(110,220),xlim=c(62,76),main="Chapter 11 Linear Regression Example",
+    pch=19,cex=1.5)
> lines(sort(dataobj$height),fitted(model)[order(dataobj$height)], col="blue", type="l")
```



Correlation Analysis of Weight and Height

These lines of code generate a scatterplot, compute the sample correlation coefficient for weight and height, and perform a hypothesis test that the true correlation is zero

```r
# Investigate correlation between height and weight
par(mfrow=c(1,1))
plot(dataobj$height,dataobj$weight,xlab="Height,inches", ylab="Weight, pounds",
     ylim=c(110,220),xlim=c(62,76),main="Simple Linear Regression Example",
     pch=19,cex=1.5)
cor(dataobj$height, dataobj$weight, method="pearson")
# Hypothesis test for population correlation equal to zero
cor.test(dataobj$height, dataobj$weight, method="pearson",
         alternative="two.sided",conf.level=0.95)
```



First, examine the scatterplot to determine if there is a linear relationship between weight and height.

Second, interpret the correlation coefficient if there is a linear relationship.

```
> cor(dataobj$height, dataobj$weight, method="pearson")
[1] 0.9470984
```

Interpretation: There is a very strong positive correlation between weight and height with a sample correlation coefficient of 0.947

Hypothesis Test that Population Correlation Coefficient is Zero

R output for test of population correlation coefficient

```
> # Hypothesis test for population correlation equal to zero
> cor.test(dataobj$height, dataobj$weight, method="pearson",
+          alternative="two.sided",conf.level=0.95)

	Pearson's product-moment correlation

data:  dataobj$height and dataobj$weight
t = 8.3466, df = 8, p-value = 3.214e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7864411 0.9877259
sample estimates:
      cor
0.9470984
```

t test of population correlation coefficient

Use the five-step method to test the population correlation coefficient ρ:

Hypotheses: $H_0: \rho = 0$

$H_a: \rho \neq 0$

Test Statistic: 8.347

P-value: 0.0000321

Decision about the null hypothesis: Reject the null hypothesis since p-value < alpha=0.05

Conclusion: Since $\rho \neq 0$ there is a statistically significant positive correlation between weight and height.