EMSE 4765: DATA ANALYSIS

For Engineers and Scientists

Session 10: Simple Linear Regression, Model Testing and Parameter Inference

Version: 3/22/2021

Lecture Notes by: J. René van Dorp[1]

www.seas.gwu.edu/~dorpjr

[1] Department of Engineering Management and Systems Egineering, School of Engineering and Applied Science, The George Washington University, 800 22nd Street, N.W., Suite 2800, Washington D.C. 20052. E-mail: dorpjr@gwu.edu.

- **Regression analysis** is probably the most widely used form of **linear dependence analysis.**

- It is used to **explore the relationships** between a set of **explanatory variables** $X_1, \ldots, X_p$ and **a single linearly dependent variable** Y.

In general regression analysis is used to answer questions of the following type:

1. ***Description:*** How can we describe the relationship between the dependent variable and the explanatory variables?

2. ***Inference:*** How strong is the relationship captured by the model? Is the relationship described by the model statistically significant? Which explanatory variables are the most important?

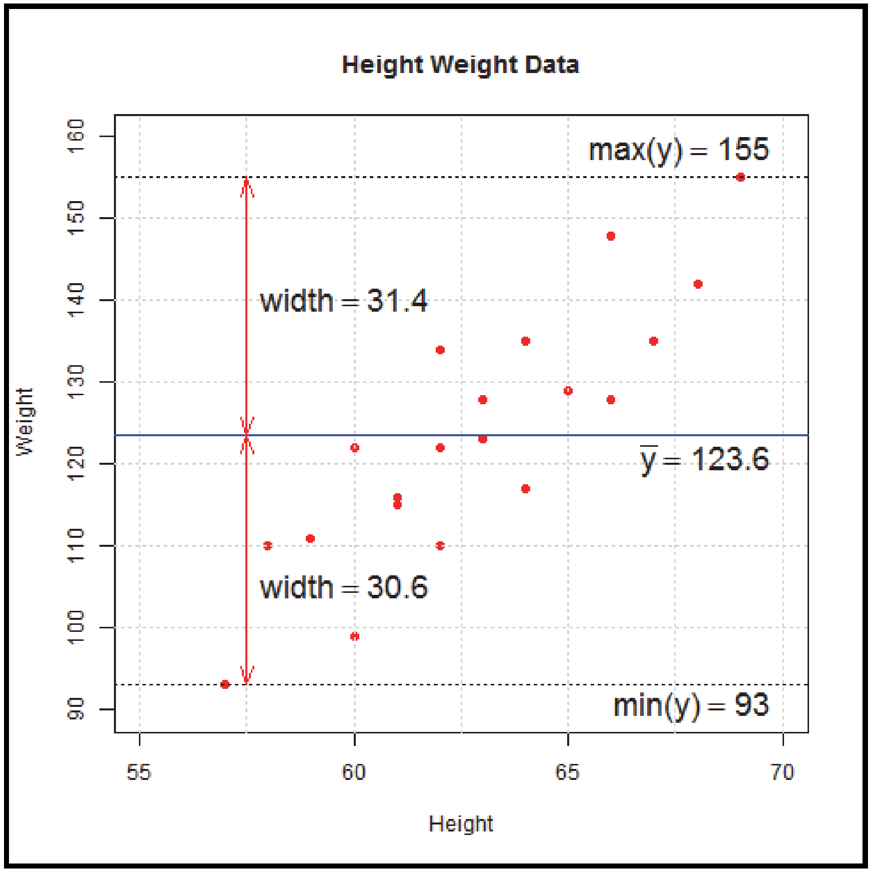
3. ***Prediction:*** **Given a new set of values for the explanatory variables what is the predicted value for the dependent variable** and **what is the uncertainty in the prediction of the dependent variable when using these values**?

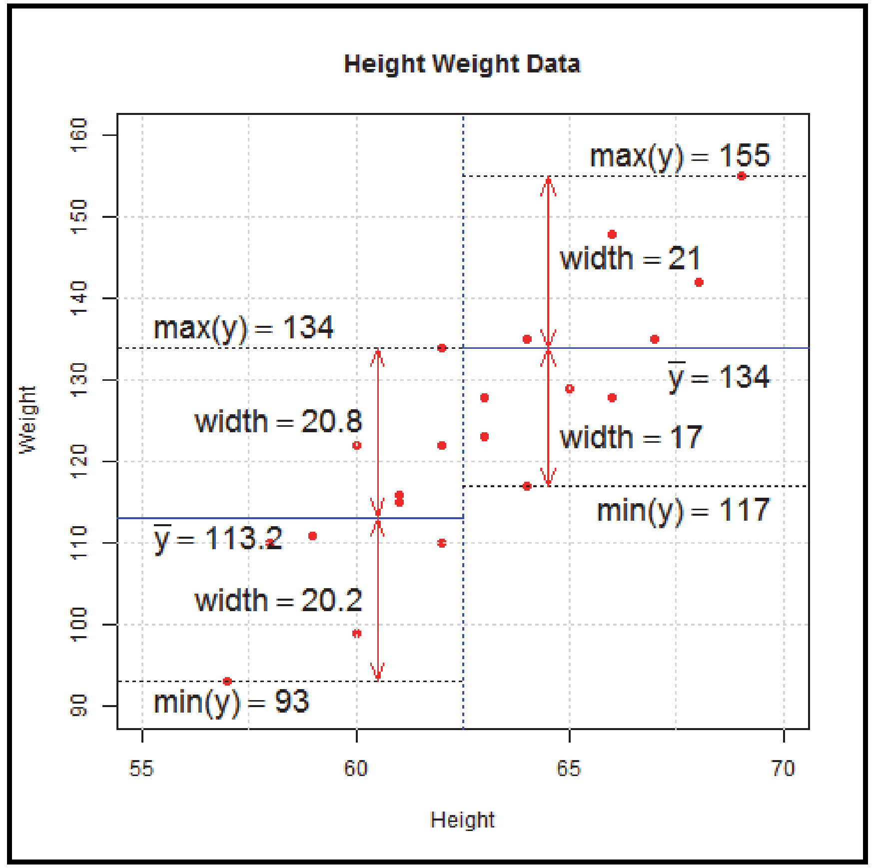
- In regression analysis, one accepts that **the relationship between a single dependent variable Y and a set of explanatory variables (the X's) is imperfect** due to other factors not captured by **the explanatory variables**.

- **Simple Regression: one explanatory variable** and **one dependent variable**

Height-Weight Sample of 20 Individuals:
An (imperfect) relationship is present between a person's height and weight. Many factors influence weight (besides height) such as: lifestyle, genetics, etc.

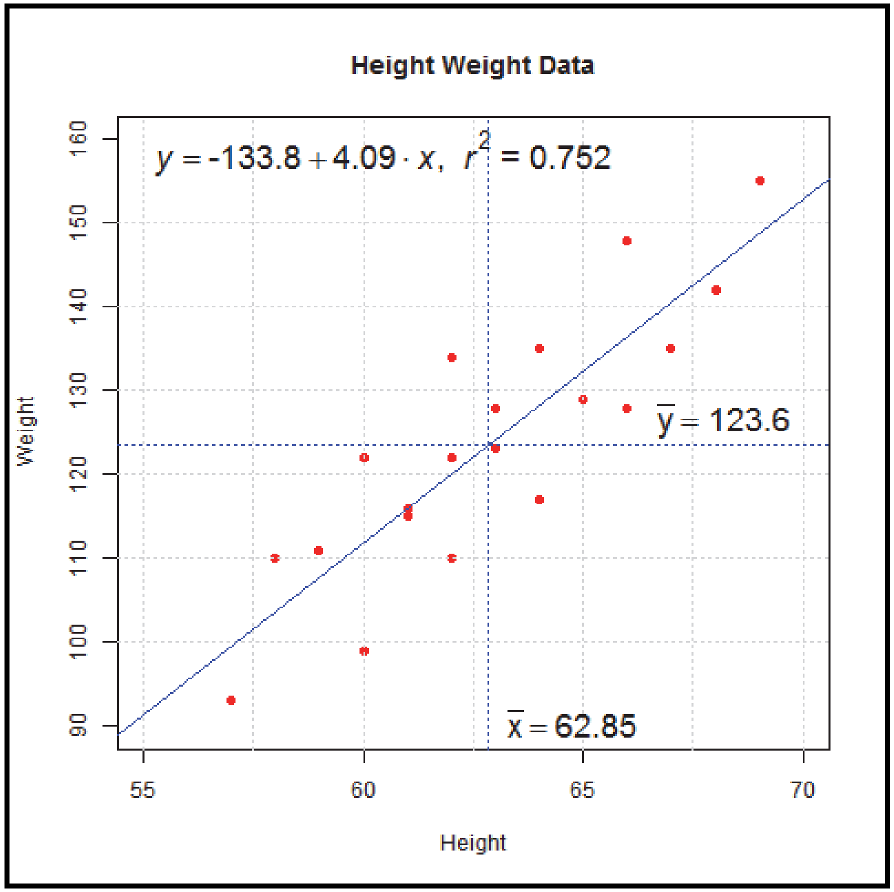
Condition	Sample Size	Best Guess	Weight Range	Half Width
No information	20	$\overline{y} = 123.6$	[93,155]	≈ 31
Below median height	10	$\overline{y} = 113.2$	[93,134]	≈ 20.5
Above median height	10	$\overline{y} = 134.0$	[117,155]	≈ 19





- We could go on, dividing height into smaller intervals and improve our guesses.
- With too many intervals, too few observations remain per interval $\Rightarrow$ results are too specific and not generalizable to a larger population ($\Rightarrow$ useless).

- An approach is needed that uses data more efficiently, but makes assumptions.



$$E[Y|x] = b_0 + b_1 x = \begin{pmatrix} b_0 & b_1 \end{pmatrix} \begin{pmatrix} 1 \\ x \end{pmatrix}, \; b_0 : \text{Intercept}, \; b_1 : \text{slope}$$

- Intercept b_0 and slope b_1 are chosen such that the mean values $E[Y|x]$ are as close as possible to the actual observed y values in the data. (More later.)

- $\widehat{b}_0 = -133.8$: The weight of a person with 0 height, **far outside the observed data range!**

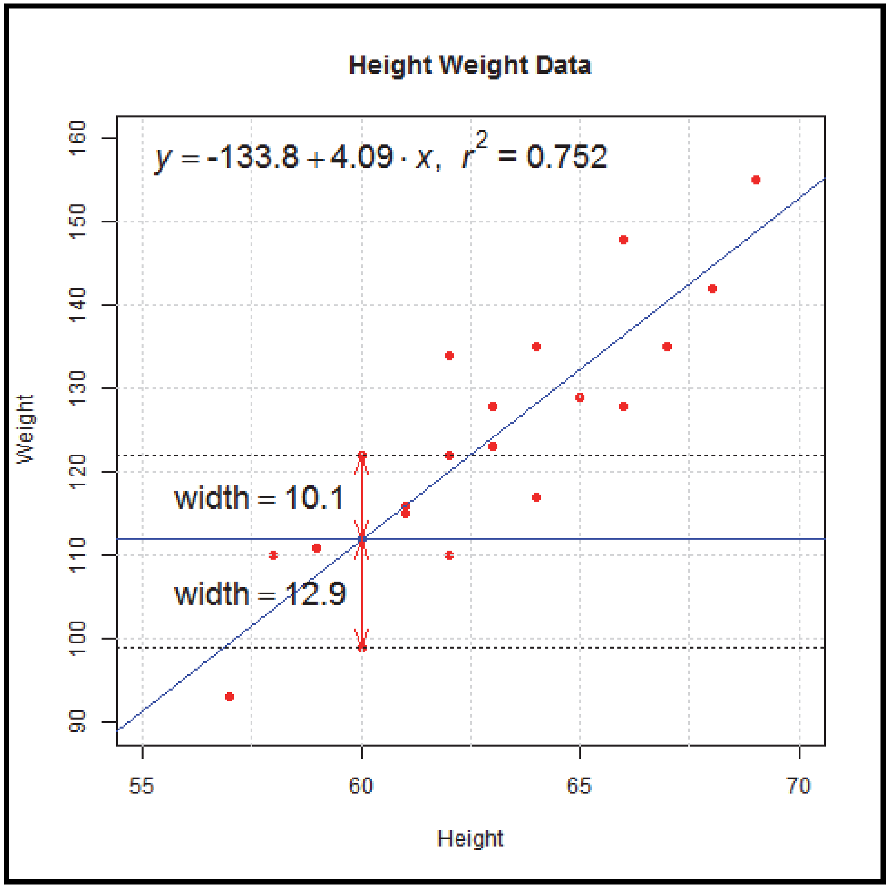
- $\widehat{b}_1 = 4.09\,\text{pounds/inch}$: Weight increases **on average** with 4.09 pounds per inch increase in height.

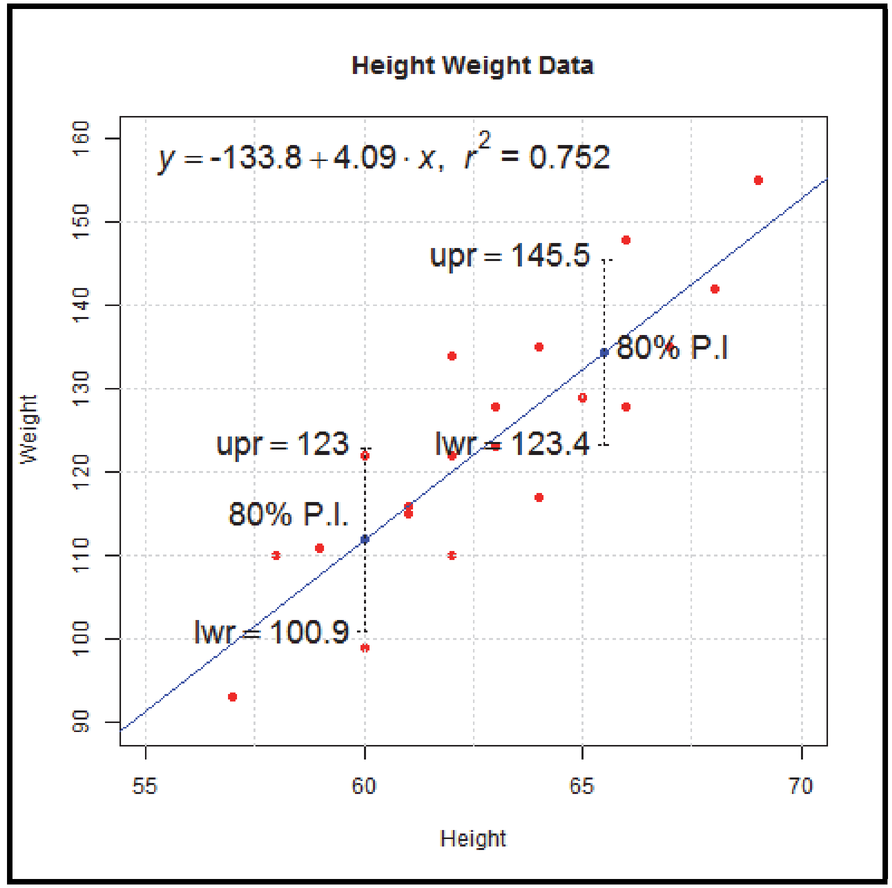
- Note that, regression line contains the point $(\overline{x}, \overline{y}) = (62.85, 123.6)$.

Best guess for the weight of a person who is 60 inches tall:

$$\widehat{y} = -133.8 + 4.09 \times 60 \approx 111.6\,\text{pounds}$$

- Two individuals in the data of height 60 inches: one weighs 99 pounds and the other weighs 122 pounds. Half-width ≈ 11.5 pounds.

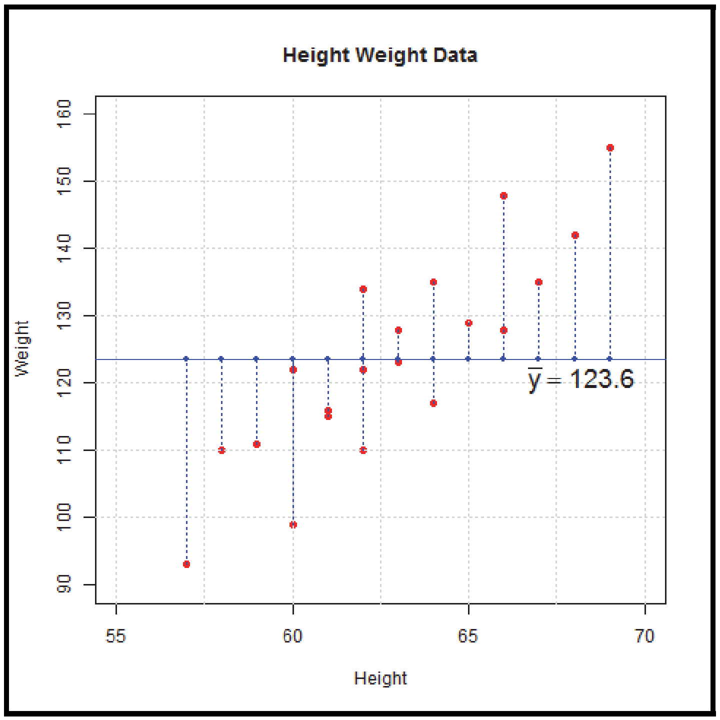




Plot on the right formalizes the uncertainty in weight at 60 inches and 65.5 inches in height. **Predicition Intervals (P.I.) have a probability interpretation.**

Step 1: How do we choose the parameters intercept b_0 and slope b_1?

- Uncertainty about Y is **the greatest in the absense of any information about x**. **One measure of uncertainty is the variance**, **which is proportional to** $\sum_{i=1}^{n}(y_i - \overline{y})^2 \approx 4606.8$, called "the sum of squares".



Suppose we set **the slope $b_1 = 0$** and **the intercept $b_0 = \overline{y}$** of the dependent variable observarions. That is:

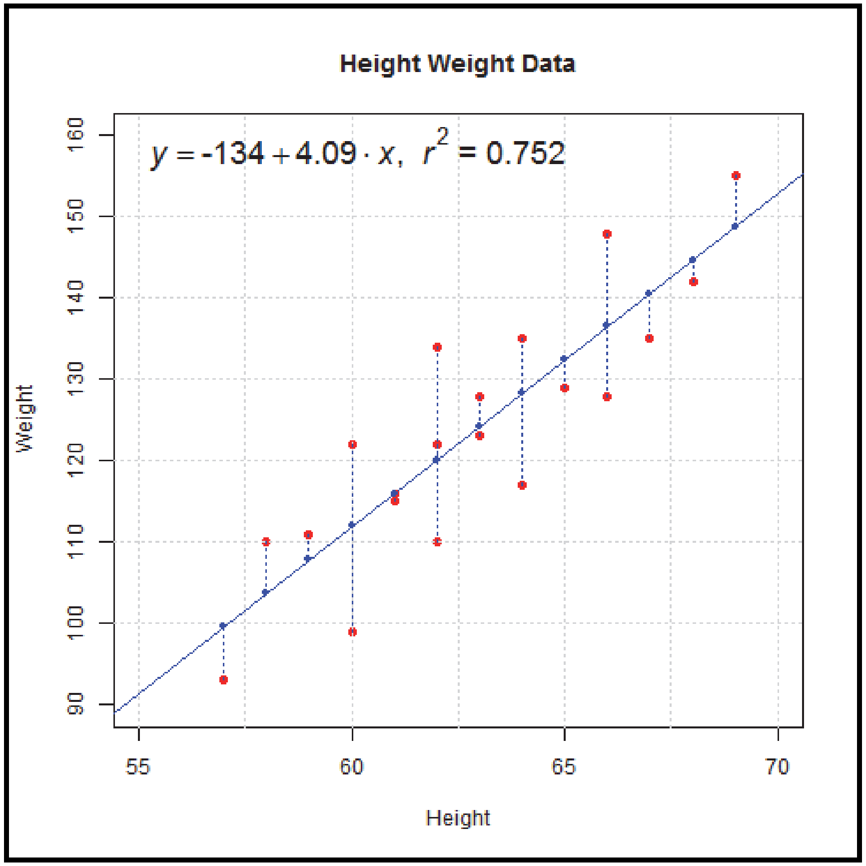
$$E[Y|x] = \overline{y}$$

The accuracy of that model can be summarized by:

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 \approx 4606.8$$

If there is any relationship between x and Y in this model? **No!**
Can we improve accuracy (i.e. reduce our uncertainty about Y)? **Yes!**

Suppose we set: $E[Y|x] = -134 + 4.09 \times x$



Errors were previously measured from: $\overline{y}$

Errors are now measured from the fitted value: $\widehat{y}_i = -134 + 4.09 \times x_i$

$$\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 \approx 1143.3$$

Using **height information $\boldsymbol{x_i}$** we reduced the uncertainty from 4606.8 to 1143.3

Choose slope b_0 and intercept b_1 that **minimizes the remaining uncertainty.**

- To summarize **goodness-of-fit of the regression line**, we compare the **uncertainty in Y without x** to the **uncertainty in Y with x** as measured by their sum of squares:

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 - \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$$

- **The relative amount of uncertainty in the sum of squares** explained **by the regression line** then equals:

$$R^2 = \frac{\sum_{i=1}^{n}(y_i - \overline{y})^2 - \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = 1 - \frac{1143.3}{4606.8} \approx 75.18\%$$

- Using notation R^2 to denote this measure is not a coincidence: **the R^2 estimate** is **equivalent to the squared correlation (ρ)** between the fitted values $\widehat{y}$ and the actual values y.

- For each data point $\underline{x}_i^T = (\, x_{1i} \quad x_{2i} \quad \ldots \quad x_{pi}\,)$, the **expected value of the dependent variable $\bar{Y}$**, depends on the info contained in the explanatory variables and is given by:

$$E[Y|\underline{x}_i] = b_0 + b_1 x_{1i} + b_1 x_{2i} \; \ldots + b_p x_{pi}$$

- **To capture that the observations $\boldsymbol{y_i}$ of the dependent variable are not perfect**, **a realization $\boldsymbol{\epsilon_i}$ of an error term $\boldsymbol{\epsilon_i}$** is introduced:

$$y_i = E[Y|\underline{x}_i] + \epsilon_i, \; i = 1, \ldots, n$$

These ϵ_i, $i = 1, \ldots, n$ are called **residual observations or the residuals**.

- Combining these two equations yields **with $(p+1)$ parameters $\boldsymbol{b_i}$**:

$$y_i = b_0 + b_1 x_{1i} + b_1 x_{2i} \; \ldots + b_p x_{pi} + \epsilon_i, \; i = 1, \ldots, n$$

- In matrix form:

$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{\epsilon}$, where $\boldsymbol{y}, \boldsymbol{\epsilon}$ are n-vectors,
$\boldsymbol{X}$ is an $[n \times (p+1)]$-matrix and $\boldsymbol{b}$ is a $(p+1)$-vector,

- A draw-back of the R^2 measure is that **it always increases when an explanatory variable is added to the model.** Thus, by adding variables we can eventually obtain an R^2 of 100%, but lesser data per coefficient estimated.

- When building a model one would like to have **a model that is parsimonious while adequately describing the variation in the dependent variable.**

$$R^2_{adj} = 1 - \frac{s^2_\epsilon}{s^2_Y} = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2/[n-(p+1)]}{\sum_{i=1}^{n}(y_i - \overline{y})^2/(n-1)} =$$

$$= 1 - \frac{(n-1)}{(n-p-1)} \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = 1 - \frac{19}{18} \cdot \frac{1143.3}{4606.8} \approx 73.8\%$$

- When adding variables R^2_{adj} **eventually will have to go down. Pragmatic modeling approach**: add explanatory variables until the R^2_{adj} goes down.

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & & x_{1p} \\ 1 & x_{21} & x_{22} & & x_{2p} \\ & & & & \\ 1 & x_{n1} & x_{n2} & & x_{np} \end{pmatrix}$$, n-vector $\underline{1}$ is multiplied by the intercept b_0

1. The matrix $\boldsymbol{X}$ is of full rank: Their is no perfect redundancy in the matrix. **No column can be written as a linear combination of the others**.

2. **The explanatory data matrix $\boldsymbol{X}$ is fixed: it is not random**. When $\boldsymbol{X}$ is fixed, it cannot be correlated with the random error term ϵ.

3. The residual random error term ϵ has a mean of 0 and a variance σ^2, i.e.

$$E[\epsilon] = 0 \text{ and } V[\epsilon] = \sigma^2.$$

4. **The residual vector $\boldsymbol{\epsilon}^T = (\epsilon_1, \ldots, \epsilon_n)$** is **a realization of a random sample** of that random error term requiring independence and constant variance!

Note: No assumption has been made (yet) regarding the distributional form of ϵ.

Parameters that yield the highest R^2 (**i.e. the best fit**) are:

$$\widehat{\boldsymbol{b}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

1. The vector estimate $\widehat{\boldsymbol{b}}$ for the coefficient vector $\boldsymbol{b}$ is unbiased.

2. The covariance matrix of $\widehat{\boldsymbol{b}}$ equals $\Sigma(\boldsymbol{b}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

- The covariance matrix of $\widehat{\boldsymbol{b}}$ is used **to make statistical inferences about the values of the regression parameters/coefficients.**

- **The fitted values of the regression model** are given by: $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{b}}$

- The difference between the actual values $\boldsymbol{y}$ and the fitted values $\widehat{\boldsymbol{y}}$ are called **the residuals** and are denoted as follows:

$$\epsilon_i = y_i - \widehat{y}_i,\ i = 1, \ldots, n \text{ or in vector form } \boldsymbol{\epsilon} = \boldsymbol{y} - \widehat{\boldsymbol{y}} .$$

Is the relationship between Weight (Y) and Height (X) statistically significant?

- It can be shown that **the total sum of squares partitions as follows**:

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 + \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 \Leftrightarrow$$

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 - \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2$$

- We have:

$$R^2 = \frac{\sum_{i=1}^{n}(y_i - \overline{y})^2 - \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = \frac{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2},$$

- Thus

$$1 - R^2 = \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \Rightarrow \frac{R^2}{1 - R^2} = \frac{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}$$

- Finally, **if residuals ϵ_i form a normal random sample** it follows that:

$$F = \frac{(n-p-1)}{p} \times \frac{R^2}{1-R^2} = \frac{\sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2/p}{\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2/(n-p-1)} \sim F_{p,n-p-1}$$

Hence, **the larger the value of R^2**, **the larger the value of the F-statistic**.

Is there a relationship between weight (Y) and Height (X)?

See EXCEL spreadsheet "height_weight_regression.xls"

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.867072479
R Square	0.751814685
Adjusted R Square	0.738026611
Standard Error	7.96987422
Observations	20

$$F = \frac{\sum_i (\hat{y}_i - \bar{y})^2 / p}{\sum_i (\hat{y}_i - y_i)^2 / (n-p-1)}$$

When model fits well F-value will be high

$$H_0 : b_1 = 0$$

Reject

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	3463.459889	3463.459889	54.5264505	7.52114E-07
Residual	18	1143.340111	63.51889508		
Total	19	4606.8			

Low P-value

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-133.7639797	34.89885153	-3.832904919	0.00121867	-207.0838028	-60.44415657
X	4.094892278	0.554547648	7.384202767	7.5211E-07	2.92983	5.259954556

same in case of simple linear regression

One sided *F*-hypothesis test provides the significance (p-value) of the overall model given the model R^2 value provided **the residual vector** $\boldsymbol{\epsilon}^T = (\epsilon_1, \ldots, \epsilon_n)$ is **a realization of a normal distributed random sample.**

Although the F-Statistic is statistically significant it is still possible that individual parameters are statistically insignificant (and thus are possibly of zero value).

$b=(X^TX)^{-1}X^Ty$

-133.764
4.095

See EXCEL spreadsheet "height_weight_regression.xls"

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.867072479
R Square	0.751814685
Adjusted R Square	0.738026611
Standard Error	7.96987422
Observations	20

$$t=\frac{\hat{b}_k-0}{\sqrt{\nu_{kk}}}\sim$$ T- distribution with (n-p-1) degrees of freedom

$$H_0: b_k = 0$$ Reject for all coefficients

ANOVA

	df	SS	MS	F	Significance F
Regression	1	3463.459889	3463.459889	54.5264505	7.52114E-07
Residual	18	1143.340111	63.51889508		
Total	19	4606.8			

Low P-values

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-133.7639797	34.89885153	-3.832904919	0.00121867	-207.0838028	-60.44415657
X	4.094892278	0.554547648	7.384202767	7.5211E-07	2.92983	5.259954556

$(Standard\_Error)^2*(X^TX)^{-1}$

1217.930	-19.328
-19.328	0.308

Root

Root

Minitab - Output

Regression Analysis: Weight versus Height

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	3463	3463.46	54.53	0.000
Error	18	1143	63.52		
Total	19	4607			

Model Summary

S	R-sq	R-sq(adj)
7.96987	75.18%	73.80%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	-133.8	34.9	-3.83	0.001
Height	4.095	0.555	7.38	0.000

Regression Equation

Weight = -133.8 + 4.095 Height

R - Output

Model Summary

R	0.867	RMSE	7.970
R-Squared	0.752	Coef. Var	6.448
Adj. R-Squared	0.738	MSE	63.519
Pred R-Squared	0.697	MAE	6.345

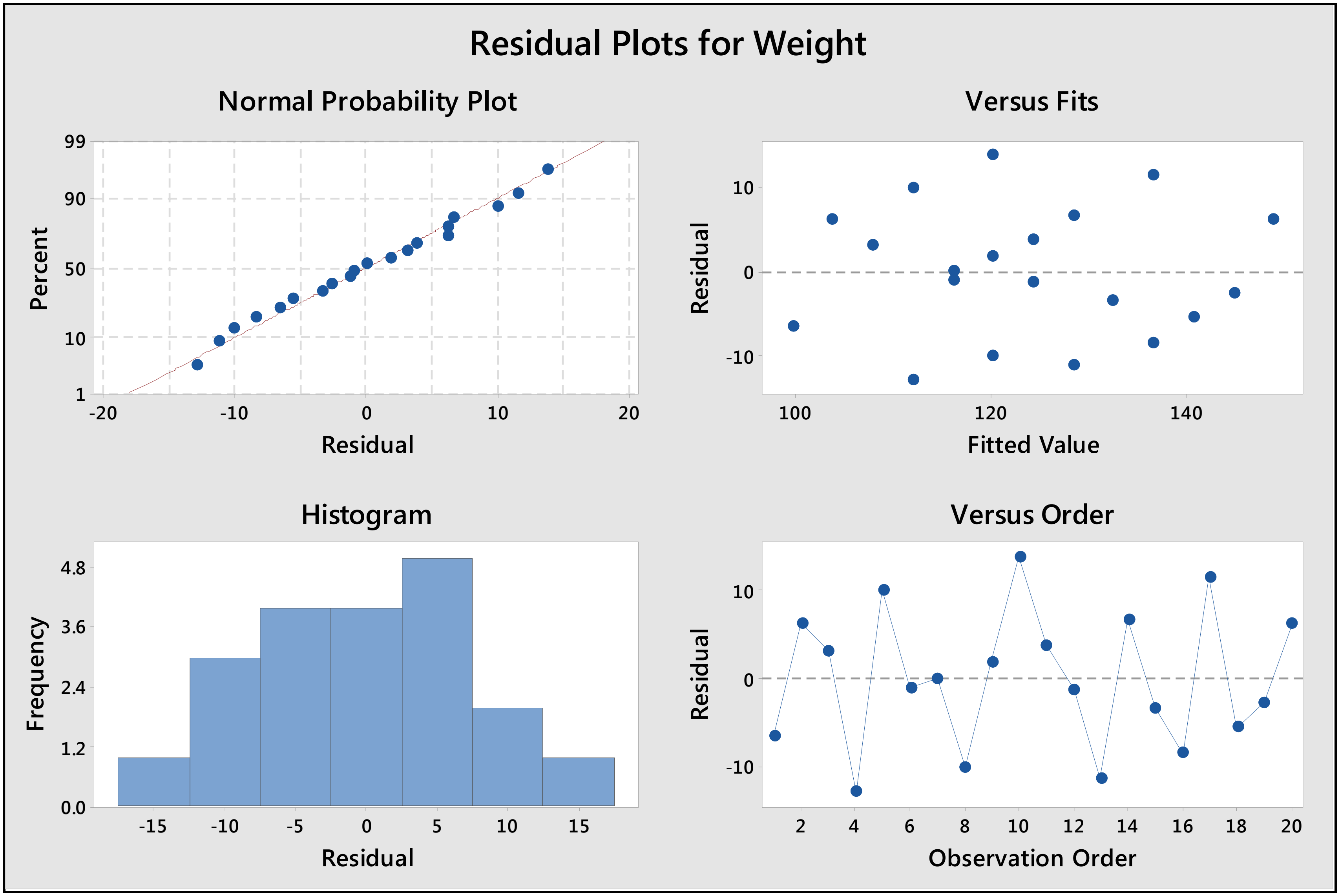
RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error

ANOVA

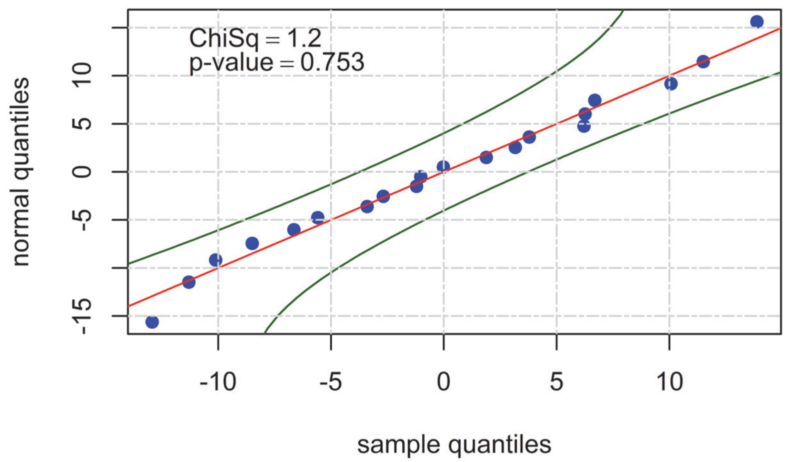
	Sum of Squares	DF	Mean Square	F	Sig.
Regression	3463.460	1	3463.460	54.526	0.0000
Residual	1143.340	18	63.519		
Total	4606.800	19			

Parameter Estimates

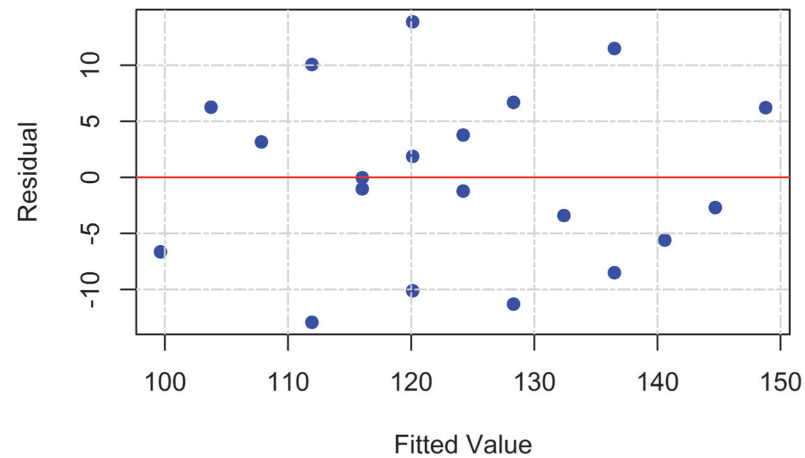
model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	-133.764	34.899		-3.833	0.001	-207.084	-60.444
Height	4.095	0.555	0.867	7.384	0.000	2.930	5.260


Residual Plots for Weight
Normal Probability Plot
Percent
Residual
Versus Fits
Residual
Fitted Value
Histogram
Frequency
Residual
Versus Order
Residual
Observation Order

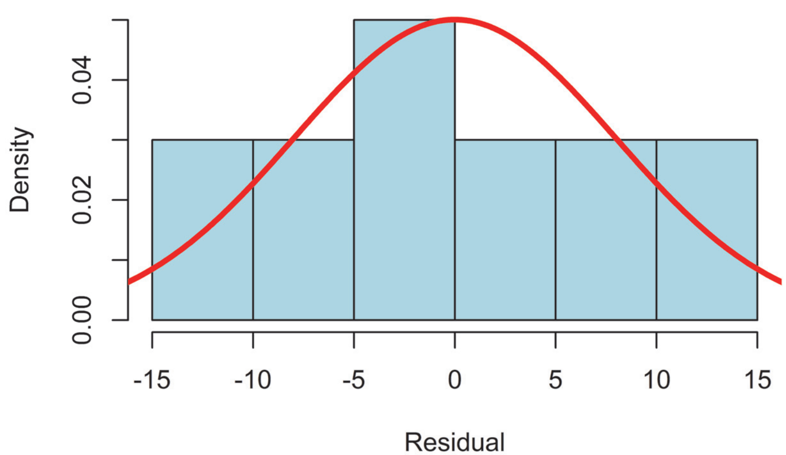
Normal Probability Plot of Residuals



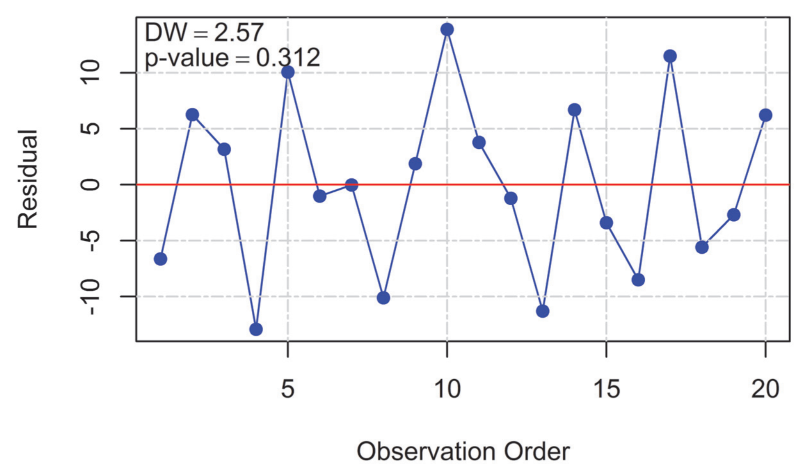
Residuals versus Fitted Values

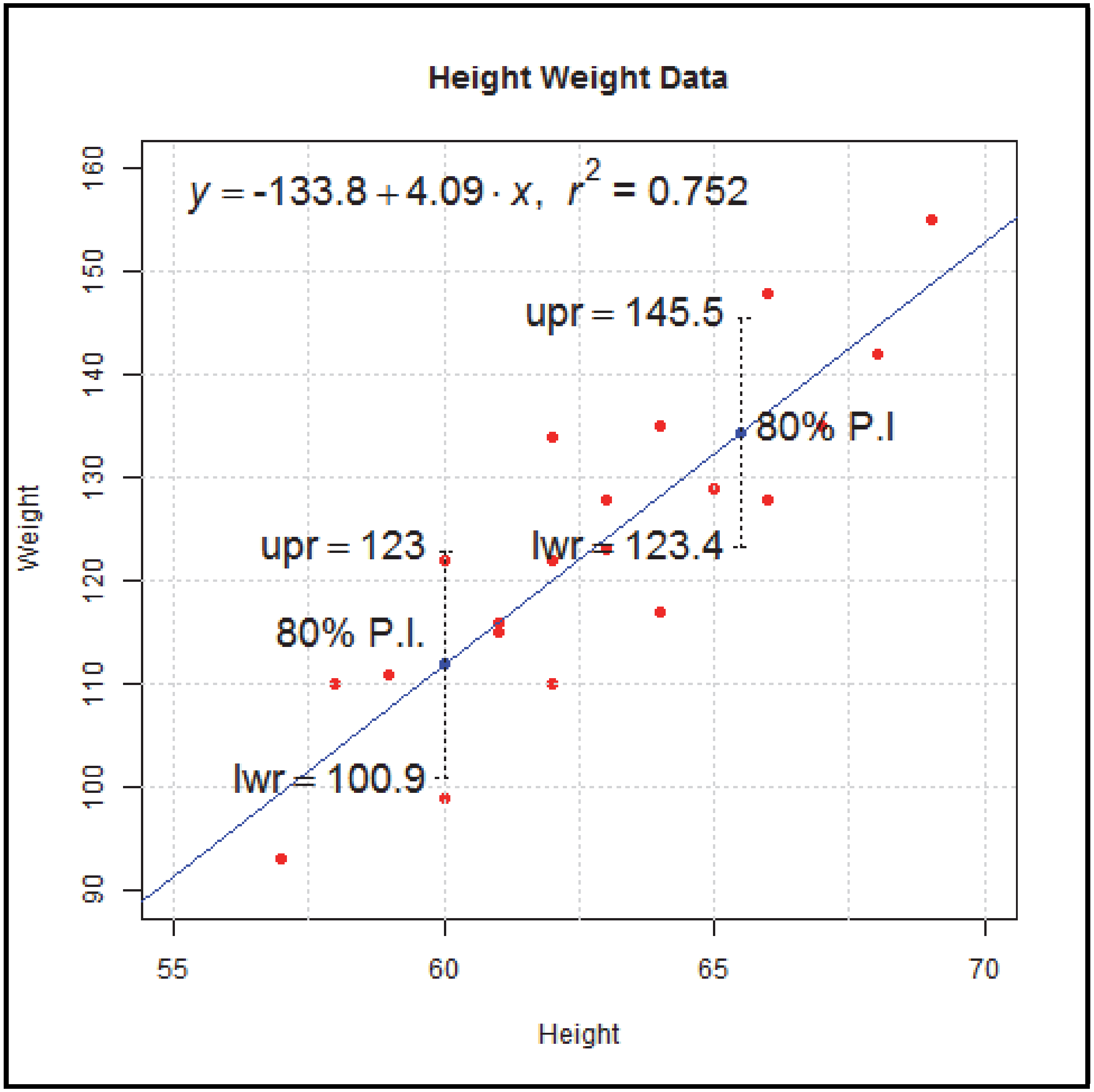


Historgram of Residuals



Residuals versus Order




Height Weight Data
$y = -133.8 + 4.09 \cdot x,\ r^2 = 0.752$
upr = 145.5
80% P.I
lwr = 123.4
upr = 123
80% P.I.
lwr = 100.9
Weight
Height
90
100
110
120
130
140
150
160
55
60
65
70