

# **Multiple Regression From the Familiar to the Multi-Dimensional**

Tom Short

West Chester University of Pennsylvania

[tshort@wcupa.edu](mailto:tshort@wcupa.edu)

# Two Examples

## Simple Linear Regression

“The Trench Data”

How long should it take to dig a trench?

## Multiple Linear Regression

*Journal of Statistics Education* Body Fat Data

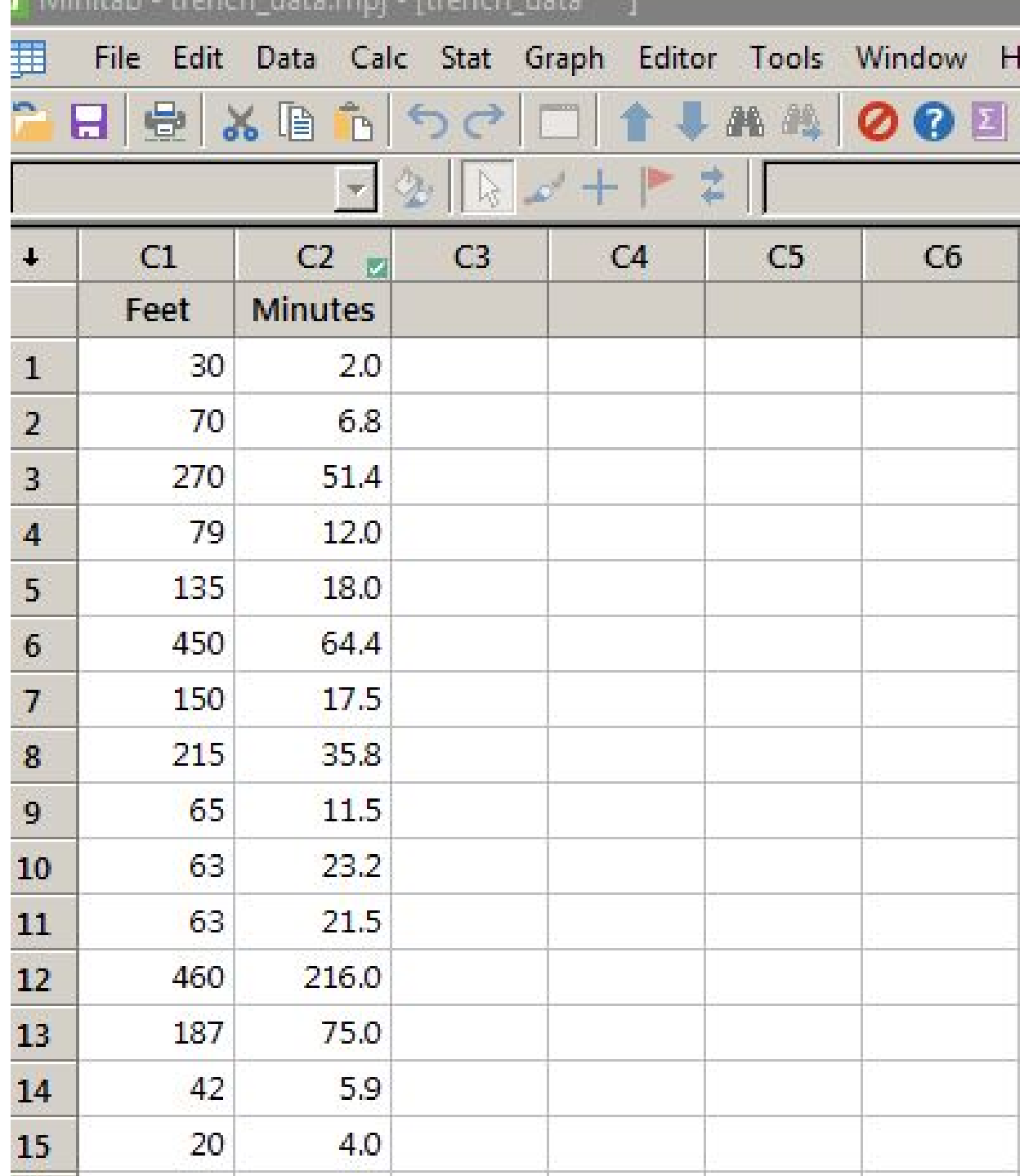
Predicting Body Fat Percentage from “dry” measurements

<http://ww2.amstat.org/publications/jse/v4n1/datasets.johnson.html>

## Minitab

<http://www.minitab.com/en-us/>

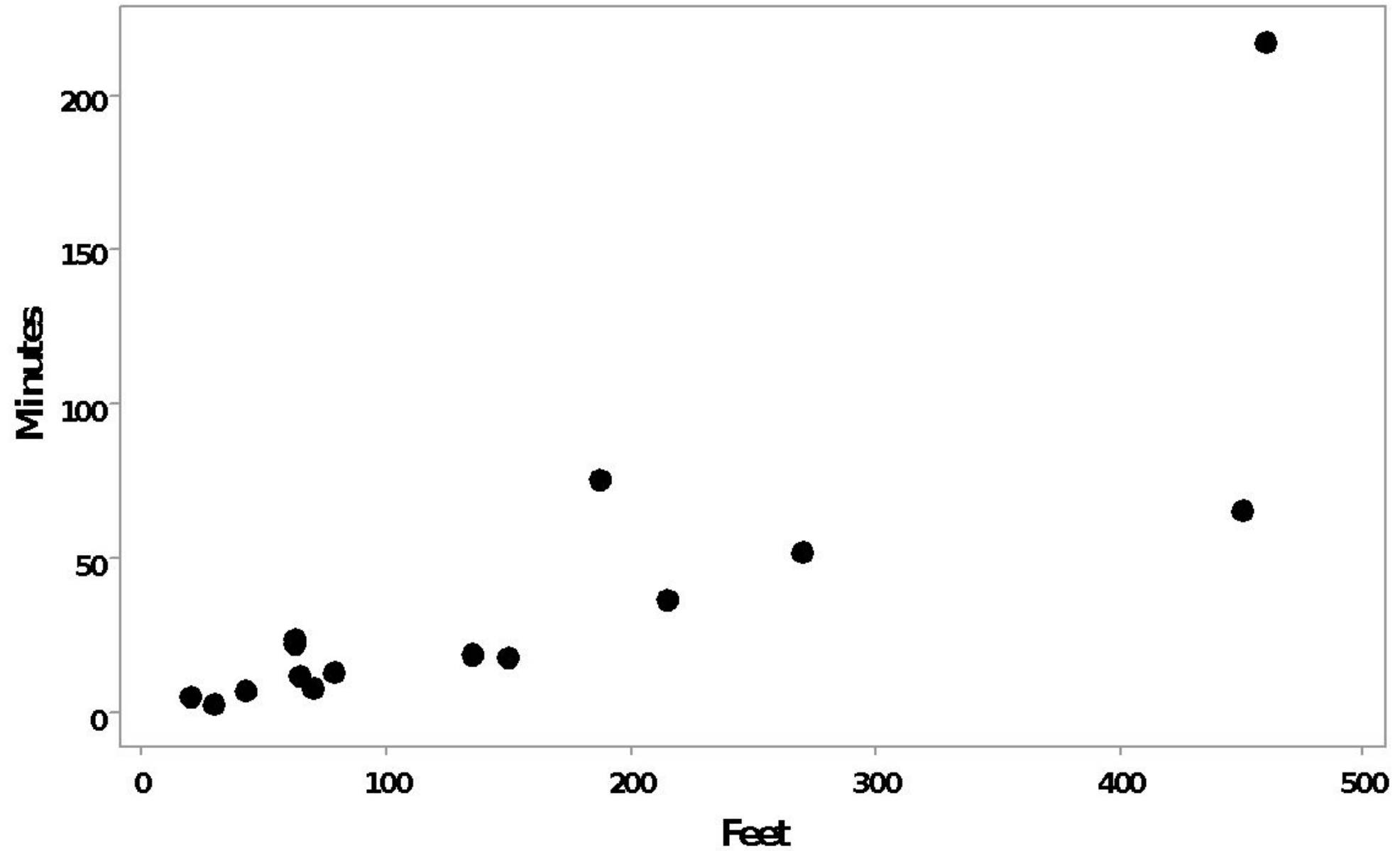
# The Trench Data



The image shows a Minitab spreadsheet window titled "Minitab - trench\_data.mtp" and "[trench\_data]". The menu bar includes File, Edit, Data, Calc, Stat, Graph, Editor, Tools, Window, and Help. The toolbar contains various icons for file operations, editing, and data analysis. The spreadsheet itself has a grid with columns labeled C1 through C6 and rows numbered 1 through 15. Column C1 is labeled "Feet" and column C2 is labeled "Minutes". The data in the spreadsheet is as follows:

	C1	C2	C3	C4	C5	C6
	Feet	Minutes				
1	30	2.0				
2	70	6.8				
3	270	51.4				
4	79	12.0				
5	135	18.0				
6	450	64.4				
7	150	17.5				
8	215	35.8				
9	65	11.5				
10	63	23.2				
11	63	21.5				
12	460	216.0				
13	187	75.0				
14	42	5.9				
15	20	4.0				

### Scatterplot of Minutes vs Feet



# Describe the Association

## Correlation: Feet, Minutes

```
Pearson correlation of Feet and Minutes = 0.814  
P-Value = 0.000
```

Assuming that a line is the correct model, then we have a strong, positive, and statistically significant correlation.

# Fit the Least Squares Regression Line

## The Model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \text{ for } i = 1, \dots, n$$

## The Fitted Equation

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \text{ for } i = 1, \dots, n$$

# Simple Linear Regression

## Regression Analysis: Minutes versus Feet

### Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	27231.1	27231.1	25.56	0.000
Feet	1	27231.1	27231.1	25.56	0.000
Error	13	13847.3	1065.2		
Lack-of-Fit	12	13845.8	1153.8	798.49	0.028
Pure Error	1	1.4	1.4		
Total	14	41078.3			

### Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
32.6370	66.29%	63.70%	18.76%

### Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-9.9	12.6	-0.78	0.448	
Feet	0.3102	0.0614	5.06	0.000	1.00

### Regression Equation

$$\text{Minutes} = -9.9 + 0.3102 \text{ Feet}$$

# The Fitted Equation

## Regression Analysis: Minutes versus Feet

### Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	27231.1	27231.1	25.56	0.000
Feet	1	27231.1	27231.1	25.56	0.000
Error	13	13847.3	1065.2		
Lack-of-Fit	12	13845.8	1153.8	798.49	0.028
Pure Error	1	1.4	1.4		
Total	14	41078.3			

### Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
32.6370	66.29%	63.70%	18.76%

### Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-9.9	12.6	-0.78	0.448	
Feet	0.3102	0.0614	5.06	0.000	1.00

### Regression Equation

Minutes = -9.9 + 0.3102 Feet



# Statistical Significance

## Regression Analysis: Minutes versus Feet

### Analysis of Variance

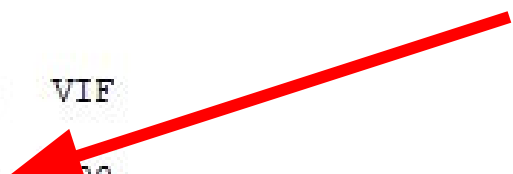
Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	27231.1	27231.1	25.56	0.000
Feet	1	27231.1	27231.1	25.56	0.000
Error	13	13847.3	1065.2		
Lack-of-Fit	12	13845.8	1153.8	798.49	0.028
Pure Error	1	1.4	1.4		
Total	14	41078.3			

### Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
32.6370	66.29%	63.70%	18.76%

### Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-9.9	12.6	-0.78	0.448	
Feet	0.3102	0.0614	5.06	0.000	1.00



### Regression Equation

Minutes = -9.9 + 0.3102 Feet

# Summary Statistics

## Regression Analysis: Minutes versus Feet

### Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	27231.1	27231.1	25.56	0.000
Feet	1	27231.1	27231.1	25.56	0.000
Error	13	13847.3	1065.2		
Lack-of-Fit	12	13845.8	1153.8	798.49	0.028
Pure Error	1	1.4	1.4		
Total	14	41078.3			

### Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
32.6370	66.29%	63.70%	18.76%



### Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-9.9	12.6	-0.78	0.448	
Feet	0.3102	0.0614	5.06	0.000	1.00

### Regression Equation

Minutes = -9.9 + 0.3102 Feet

# Summary Statistics

$s$  = The estimated standard deviation of the residuals  
(Smaller is better.)

$R^2$  = The proportion of the variability in  $y$  that is accounted for by  
the independent variable(s).  
(Larger is better.)

# Conditions for Inference

- + The model is correct
- + The errors follow a normal distribution
- + The errors have constant variability around the model
- + The errors are independent of each other (not autocorrelated)

# Are the conditions met?

+ Is the model correct? **Scatterplot**

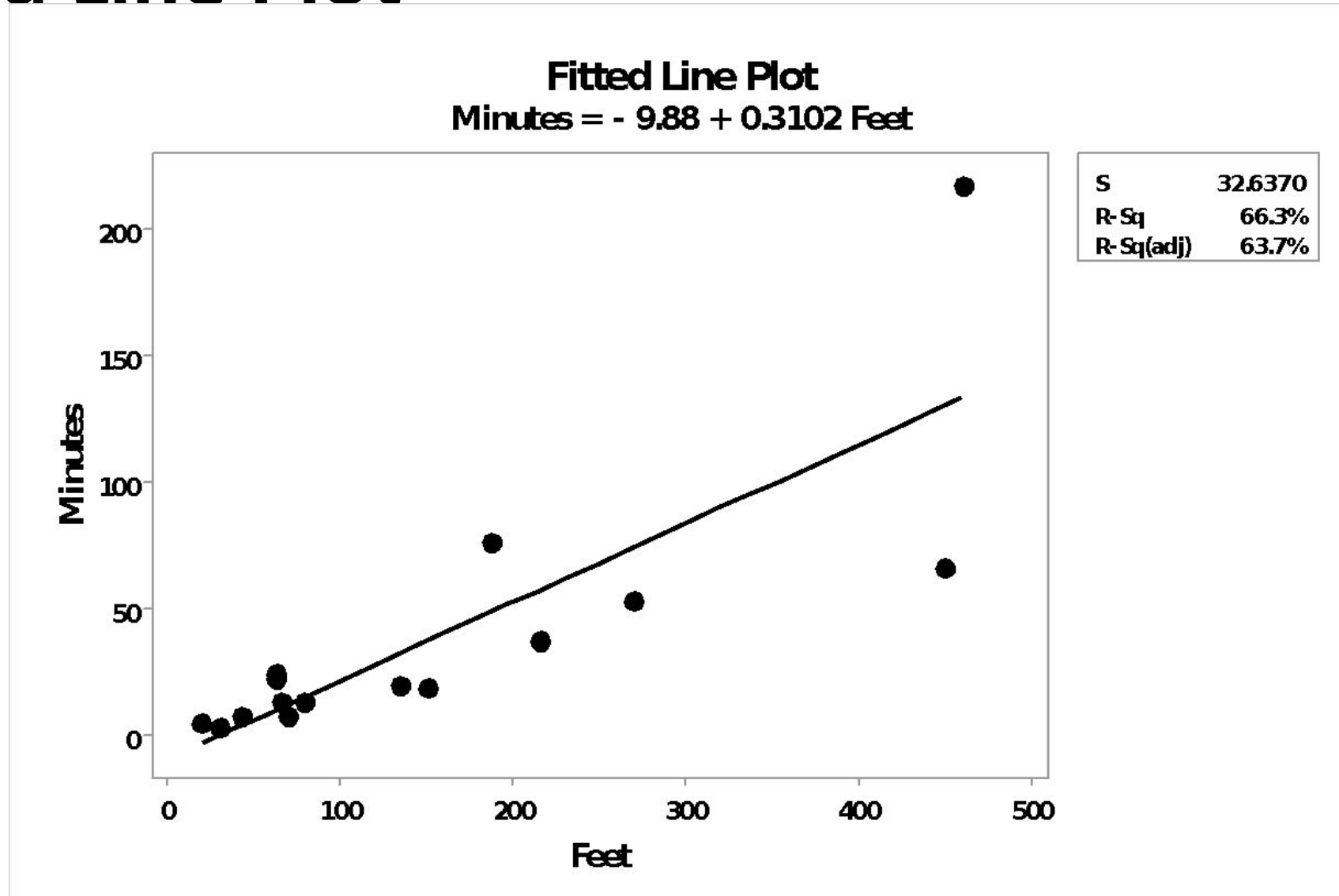
Residual = Observed – Predicted =  $e_i = y_i - \hat{y}_i = y - yhat$

+ Do the residuals follow a normal distribution? **Histogram**

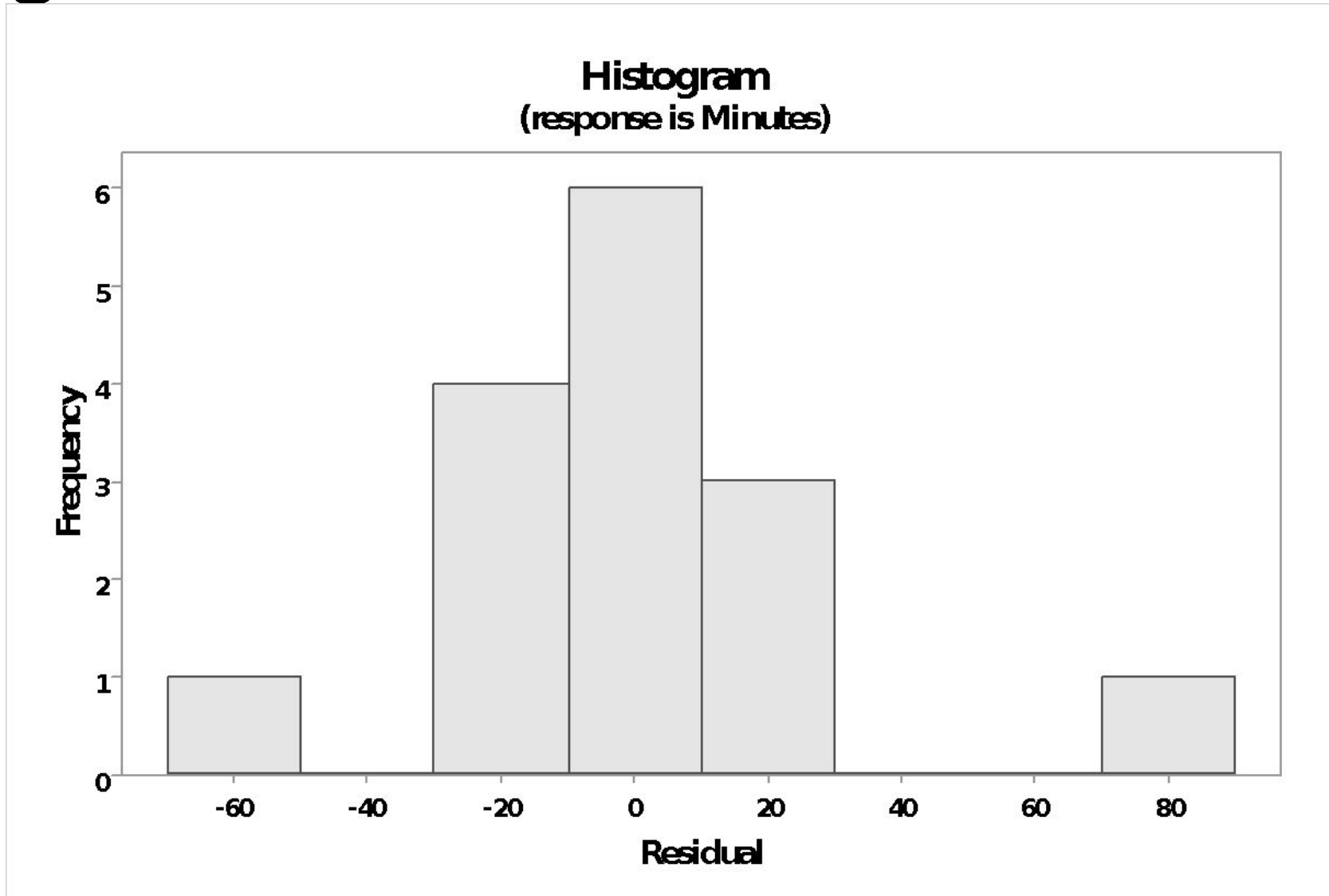
+ Do the residuals have constant variability around the model?  
**Scatterplot**. Or, the **residual plot**.

+ Are the residuals independent of each other? **Hard to check.**

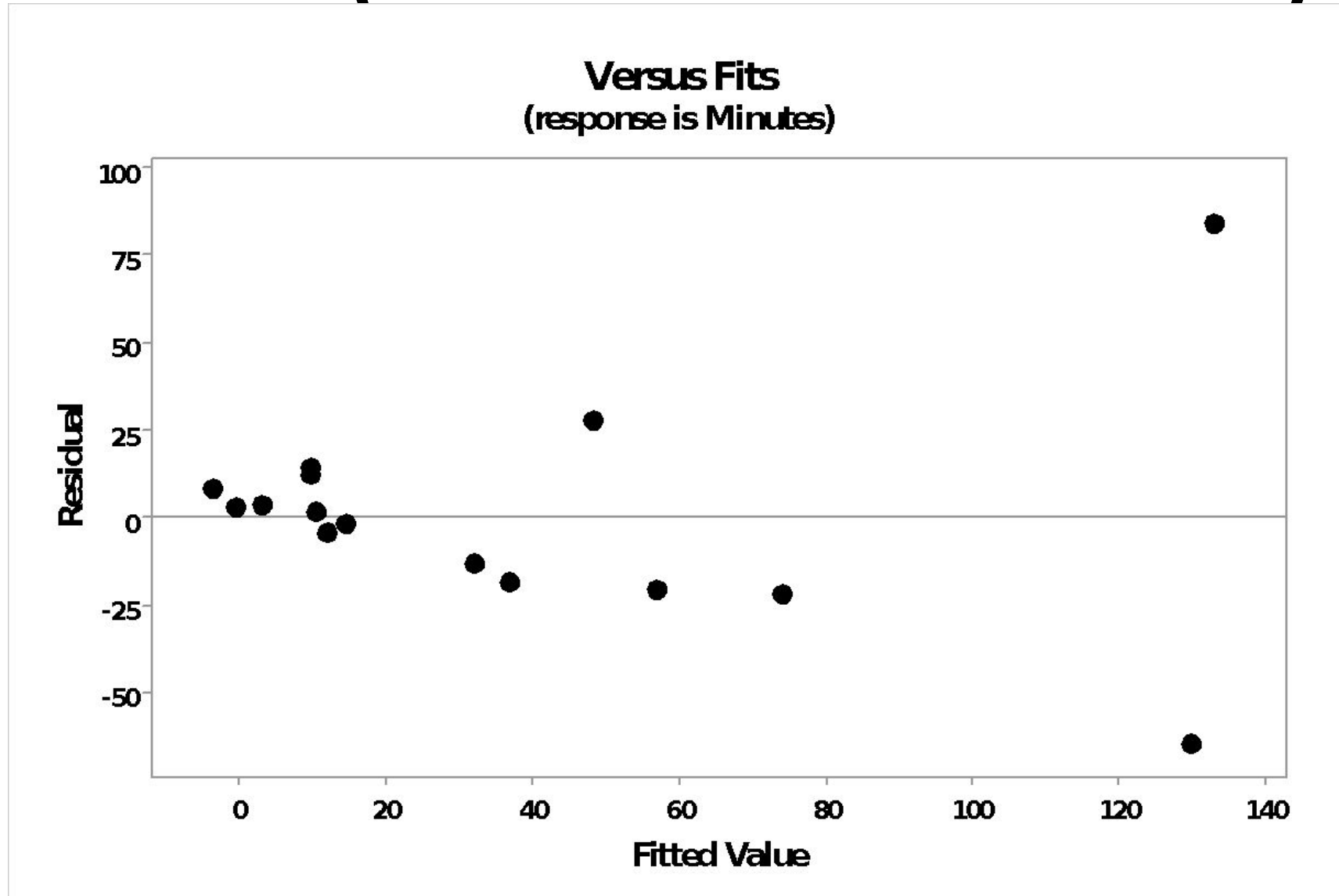
# Fitted Line Plot



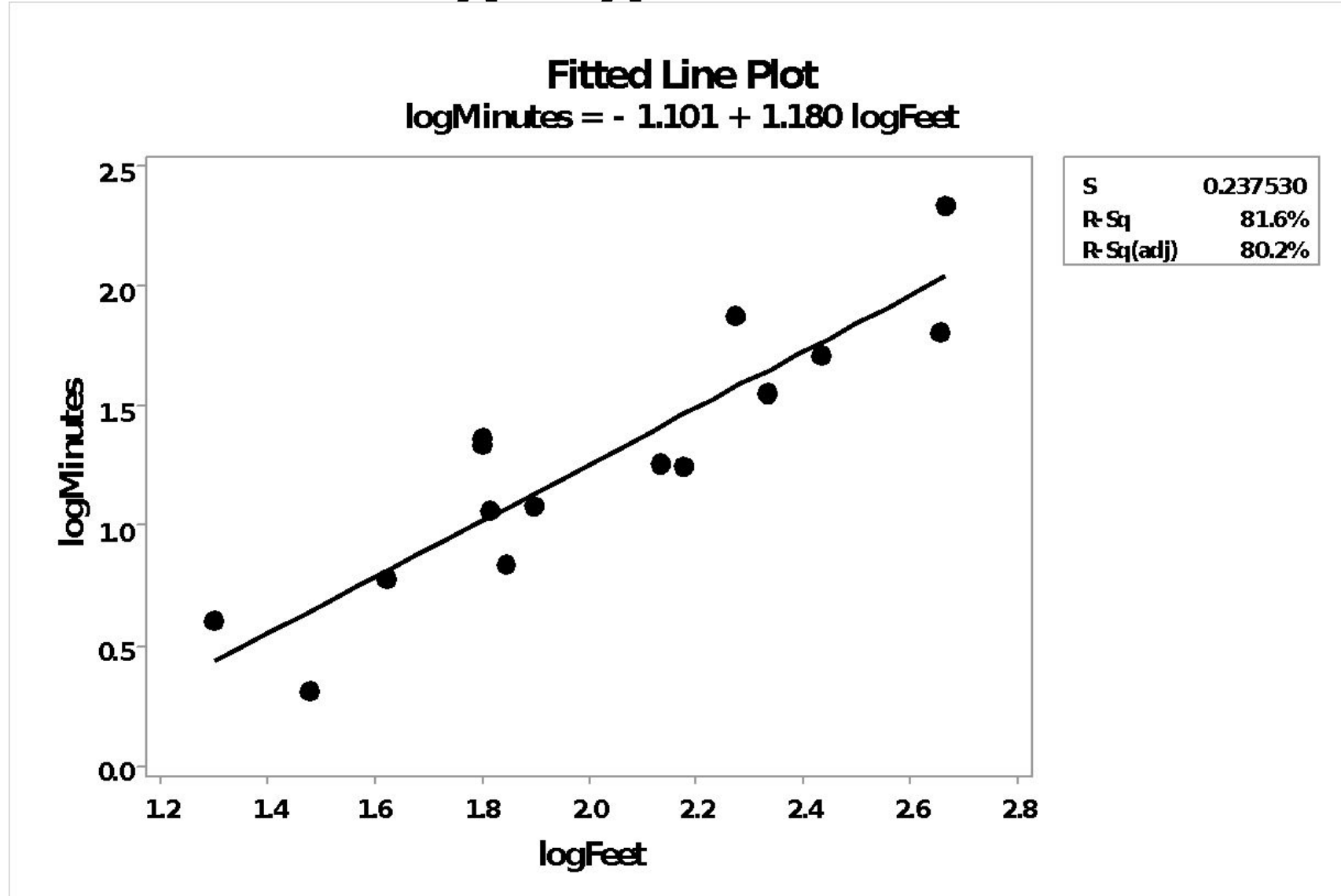
# Histogram of the Residuals



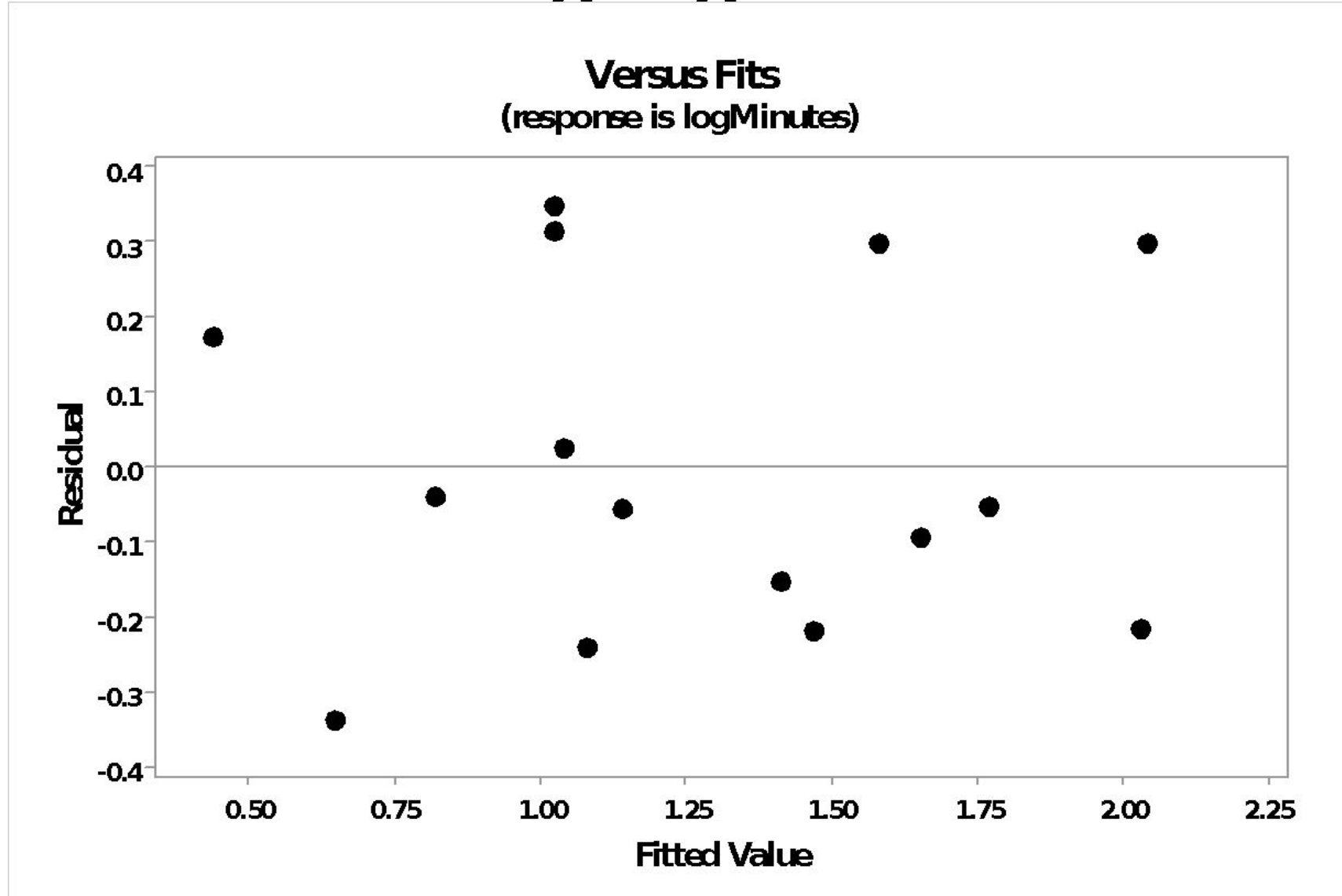
# Residual Plot (Residuals vs Predictions)



# Fitted Line for log-log Transformed Data



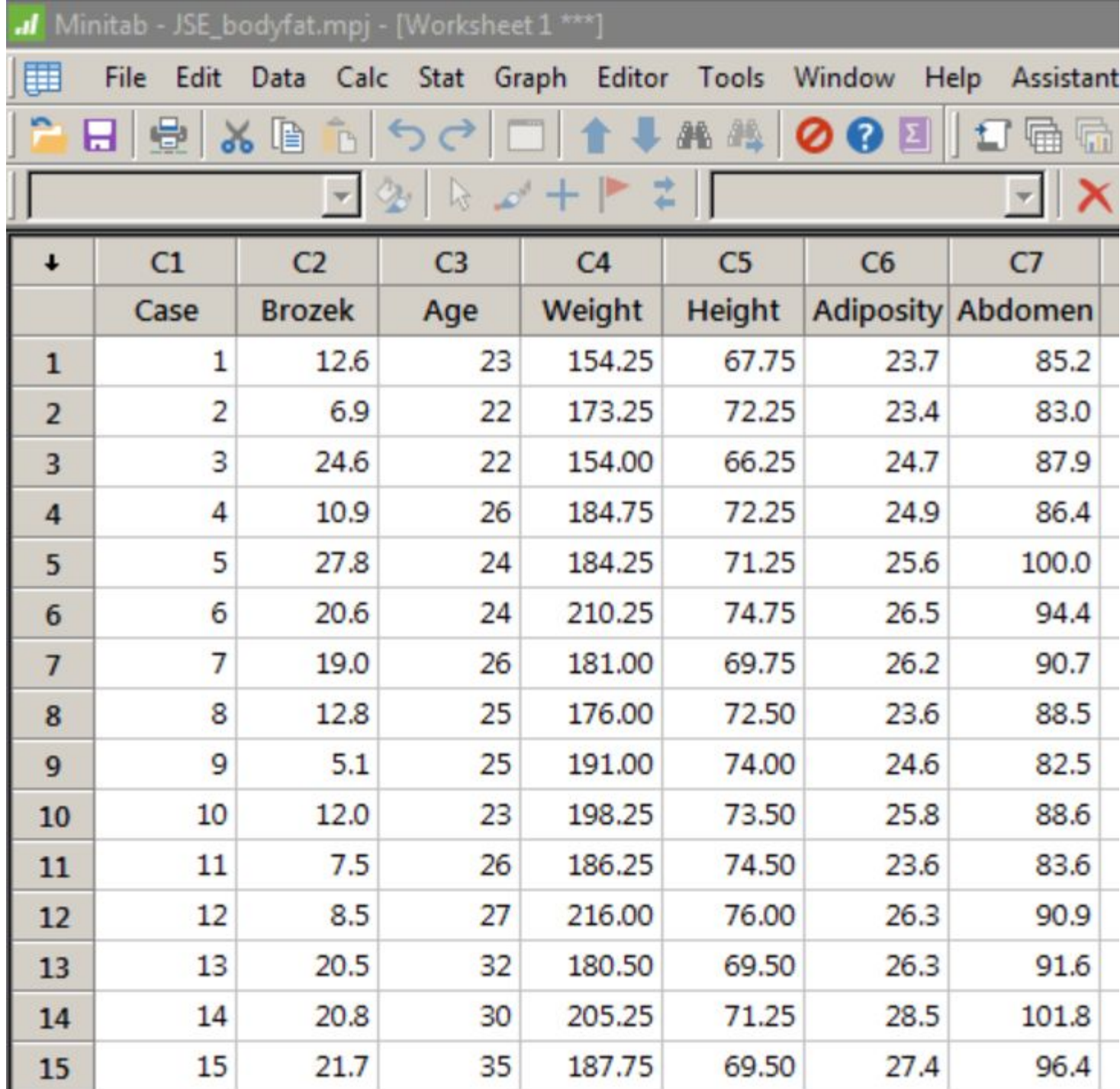
# Residual Plot for log-log Transformed Data



# *JSE* Body Fat Data

Minitab - JSE\_bodyfat.mpj - [Worksheet 1 \*\*\*]

File Edit Data Calc Stat Graph Editor Tools Window Help Assistant



↓	C1	C2	C3	C4	C5	C6	C7
	Case	Brozek	Age	Weight	Height	Adiposity	Abdomen
1	1	12.6	23	154.25	67.75	23.7	85.2
2	2	6.9	22	173.25	72.25	23.4	83.0
3	3	24.6	22	154.00	66.25	24.7	87.9
4	4	10.9	26	184.75	72.25	24.9	86.4
5	5	27.8	24	184.25	71.25	25.6	100.0
6	6	20.6	24	210.25	74.75	26.5	94.4
7	7	19.0	26	181.00	69.75	26.2	90.7
8	8	12.8	25	176.00	72.50	23.6	88.5
9	9	5.1	25	191.00	74.00	24.6	82.5
10	10	12.0	23	198.25	73.50	25.8	88.6
11	11	7.5	26	186.25	74.50	23.6	83.6
12	12	8.5	27	216.00	76.00	26.3	90.9
13	13	20.5	32	180.50	69.50	26.3	91.6
14	14	20.8	30	205.25	71.25	28.5	101.8
15	15	21.7	35	187.75	69.50	27.4	96.4

# Variables

## Columns

3 - 5 Case Number

10 - 13 Percent body fat using Brozek's equation,  
 $457/\text{Density} - 414.2$

18 - 21 Percent body fat using Siri's equation,  
 $495/\text{Density} - 450$

24 - 29 Density ( $\text{gm}/\text{cm}^3$ )

36 - 37 Age (yrs)

40 - 45 Weight (lbs)

49 - 53 Height (inches)

58 - 61 Adiposity index =  $\text{Weight}/\text{Height}^2$  ( $\text{kg}/\text{m}^2$ )

65 - 69 Fat Free Weight  
=  $(1 - \text{fraction of body fat}) * \text{Weight}$ ,  
using Brozek's formula (lbs)

74 - 77 Neck circumference (cm)

81 - 85 Chest circumference (cm)

89 - 93 Abdomen circumference (cm) "at the umbilicus  
and level with the iliac crest"

97 - 101 Hip circumference (cm)

106 - 109 Thigh circumference (cm)

114 - 117 Knee circumference (cm)

122 - 125 Ankle circumference (cm)

130 - 133 Extended biceps circumference (cm)

138 - 141 Forearm circumference (cm)

146 - 149 Wrist circumference (cm) "distal to the  
styloid processes"

# Describe Association through Correlations

## Correlation: Brozek, Age, Weight, Height, Adiposity, Abdomen

	Brozek	Age	Weight	Height	Adiposity
Age	0.289 0.000				
Weight	0.613 0.000	-0.013 0.840			
Height	-0.023 0.721	-0.245 0.000	0.489 0.000		
Adiposity	0.728 0.000	0.119 0.060	0.887 0.000	0.041 0.514	
Abdomen	0.814 0.000	0.230 0.000	0.888 0.000	0.192 0.002	0.924 0.000

Cell Contents: Pearson correlation  
P-Value

# Describe Association through Correlations

## Correlation: Brozek, Age, Weight, Height, Adiposity, Abdomen

	Brozek	Age	Weight	Height	Adiposity
Age	0.289 0.000				
Weight	0.613 0.000	-0.013 0.840			
Height	-0.023 0.721	-0.245 0.000	0.489 0.000		
Adiposity	0.728 0.000	0.119 0.060	0.887 0.000	0.041 0.514	
Abdomen	0.814 0.000	0.230 0.000	0.888 0.000	0.192 0.002	0.924 0.000

Cell Contents: Pearson correlation  
P-Value

# Fit the Multiple Regression Model

## The Model

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \varepsilon_i, \text{ for } i = 1, \cdots, n$$

## The Fitted Equation

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_k x_{ki}, \text{ for } i = 1, \cdots, n$$

# Multiple Linear Regression Part 1

## Regression Analysis: Brozek versus Age, Weight, Height, Adiposity, Abdomen

Method

Rows unused 1

### Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	5	10733.4	2146.68	125.76	0.000
Age	1	0.8	0.79	0.05	0.830
Weight	1	96.9	96.89	5.68	0.018
Height	1	14.2	14.21	0.83	0.362
Adiposity	1	21.0	21.02	1.23	0.268
Abdomen	1	2137.4	2137.37	125.21	0.000
Error	245	4182.1	17.07		
Total	250	14915.5			

### Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
4.13157	71.96%	71.39%	69.68%

# Multiple Linear Regression Part 2

## Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-73.3	35.6	-2.06	0.041	
Age	-0.0053	0.0245	-0.22	0.830	1.40
Weight	-0.2319	0.0973	-2.38	0.018	119.96
Height	0.465	0.509	0.91	0.362	25.95
Adiposity	0.796	0.718	1.11	0.268	100.19
Abdomen	0.8751	0.0782	11.19	0.000	10.41

## Regression Equation

Brozek = -73.3 - 0.0053 Age - 0.2319 Weight + 0.465 Height + 0.796 Adiposity + 0.8751 Abdomen

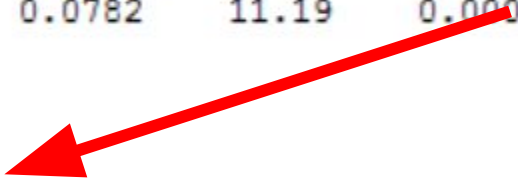
# The Fitted Equation

## Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-73.3	35.6	-2.06	0.041	
Age	-0.0053	0.0245	-0.22	0.830	1.40
Weight	-0.2319	0.0973	-2.38	0.018	119.96
Height	0.465	0.509	0.91	0.362	25.95
Adiposity	0.796	0.718	1.11	0.268	100.19
Abdomen	0.8751	0.0782	11.19	0.000	10.41

## Regression Equation

Brozek = -73.3 - 0.0053 Age - 0.2319 Weight + 0.465 Height + 0.796 Adiposity + 0.8751 Abdomen



## Regression Analysis: Brozek versus Age, Weight, Height, Adiposity, Abdomen

Method

Rows unused 1

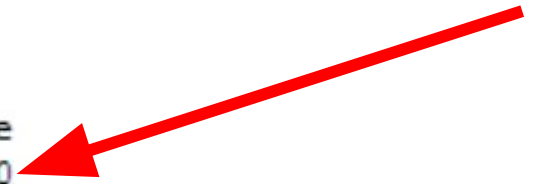
### Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	5	10733.4	2146.68	125.76	0.000
Age	1	0.8	0.79	0.05	0.830
Weight	1	96.9	96.89	5.68	0.018
Height	1	14.2	14.21	0.83	0.362
Adiposity	1	21.0	21.02	1.23	0.268
Abdomen	1	2137.4	2137.37	125.21	0.000
Error	245	4182.1	17.07		
Total	250	14915.5			

### Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
4.13157	71.96%	71.39%	69.68%

**Statistical  
Significance**



# The Overall F-Test

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

$H_a$  : At least one of the coefficients is different from 0

$H_0$  : The “reduced model” is adequate.

$H_a$  : A more complex model is necessary.

$H_0$  : There is nothing worthwhile in the model.

$H_a$  : There is something worthwhile in the model.

## Regression Analysis: Brozek versus Age, Weight, Height, Adiposity, Abdomen

Method

Rows unused 1

### Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	5	10733.4	2146.68	125.76	0.000
Age	1	0.8	0.79	0.05	0.830
Weight	1	96.9	96.89	5.68	0.018
Height	1	14.2	14.21	0.83	0.362
Adiposity	1	21.0	21.02	1.23	0.268
Abdomen	1	2137.4	2137.37	125.21	0.000
Error	245	4182.1	17.07		
Total	250	14915.5			

### Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
4.13157	71.96%	71.39%	69.68%

**Summary  
Statistics**



# More Statistical Significance

## Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-73.3	35.6	-2.06	0.041	
Age	-0.0053	0.0245	-0.22	0.830	1.15
Weight	-0.2319	0.0973	-2.38	0.018	1.99
Height	0.465	0.509	0.91	0.362	1.59
Adiposity	0.796	0.718	1.11	0.268	1.19
Abdomen	0.8751	0.0782	11.19	0.000	10.41

## Regression Equation

Brozek = -73.3 - 0.0053 Age - 0.2319 Weight + 0.465 Height + 0.796 Adiposity + 0.8751 Abdomen

# Conditions for Inference

- + The model is correct
- + The errors follow a normal distribution
- + The errors have constant variability around the model
- + The errors are independent of each other (not autocorrelated)

# Are the conditions met?

+ Is the model correct? **Multivariate data, so hard to check.**

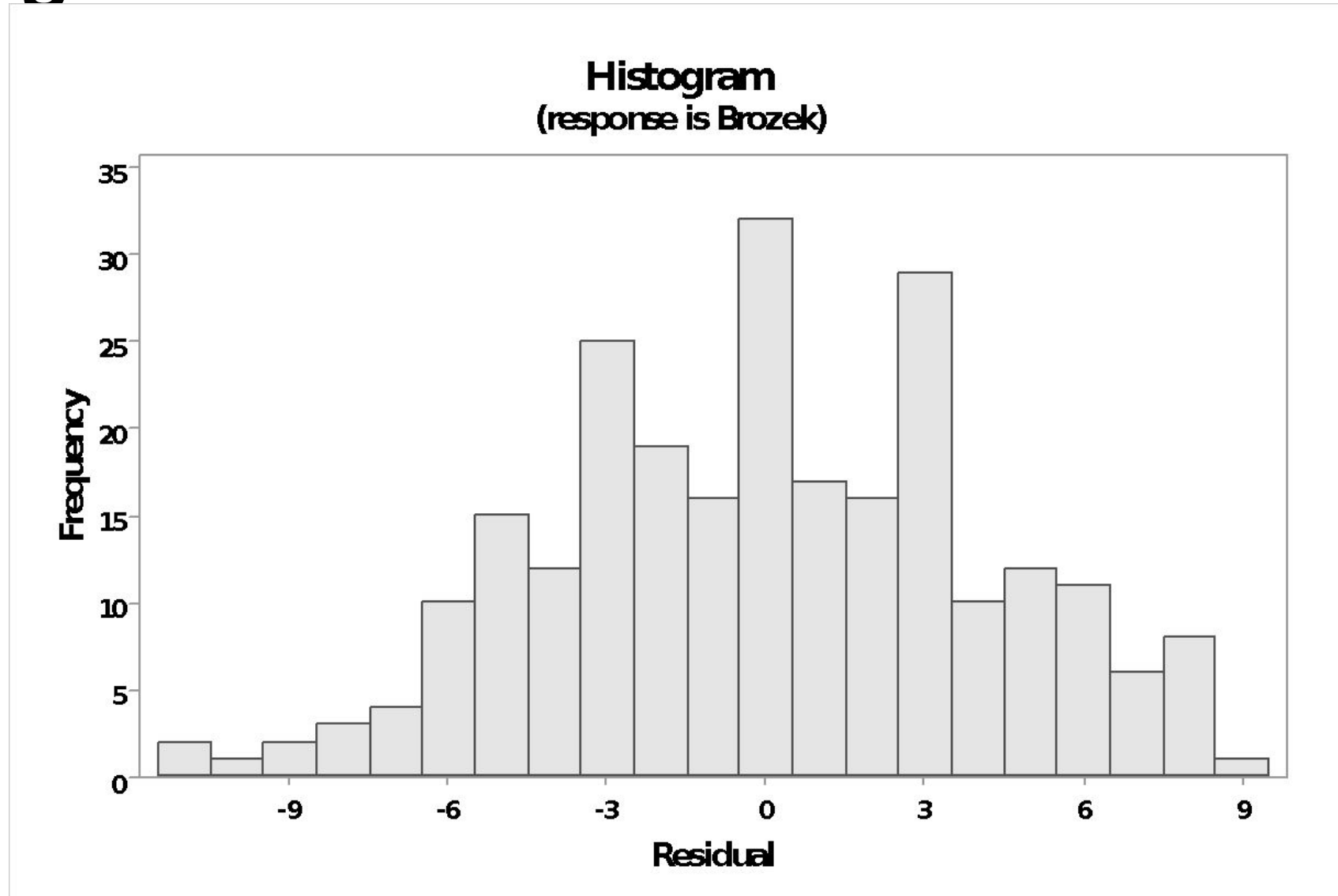
Residual = Observed – Predicted =  $e_i = y_i - \hat{y}_i = y - yhat$

+ Do the residuals follow a normal distribution? **Histogram**

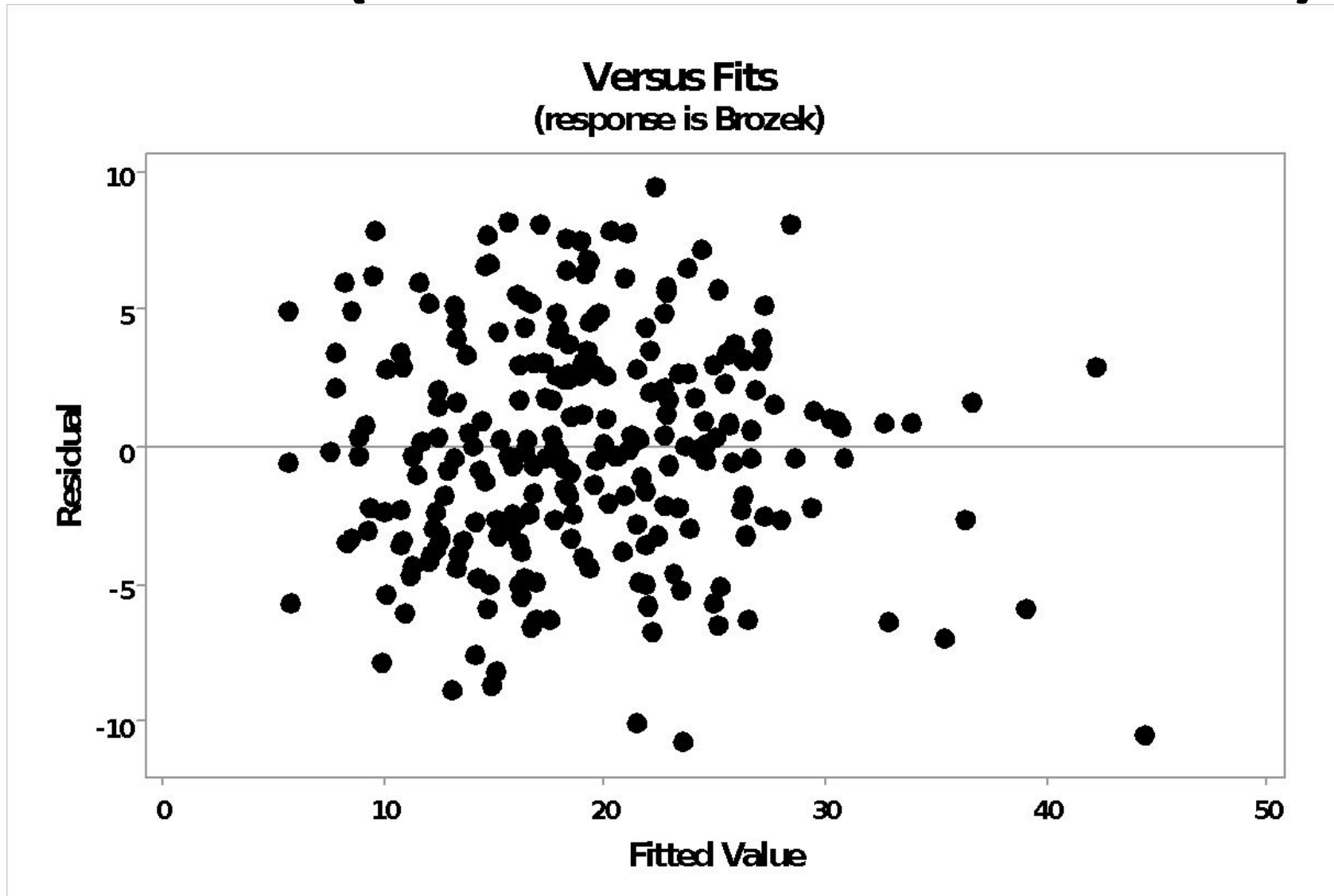
+ Do the residuals have constant variability around the model?  
**Residual plot**, also helps to confirm that the model is correct.

+ Are the residuals independent of each other? **Hard to check.**

# Histogram of the Residuals



# Residual Plot (Residuals vs Predictions)



# Describe Association through Correlations

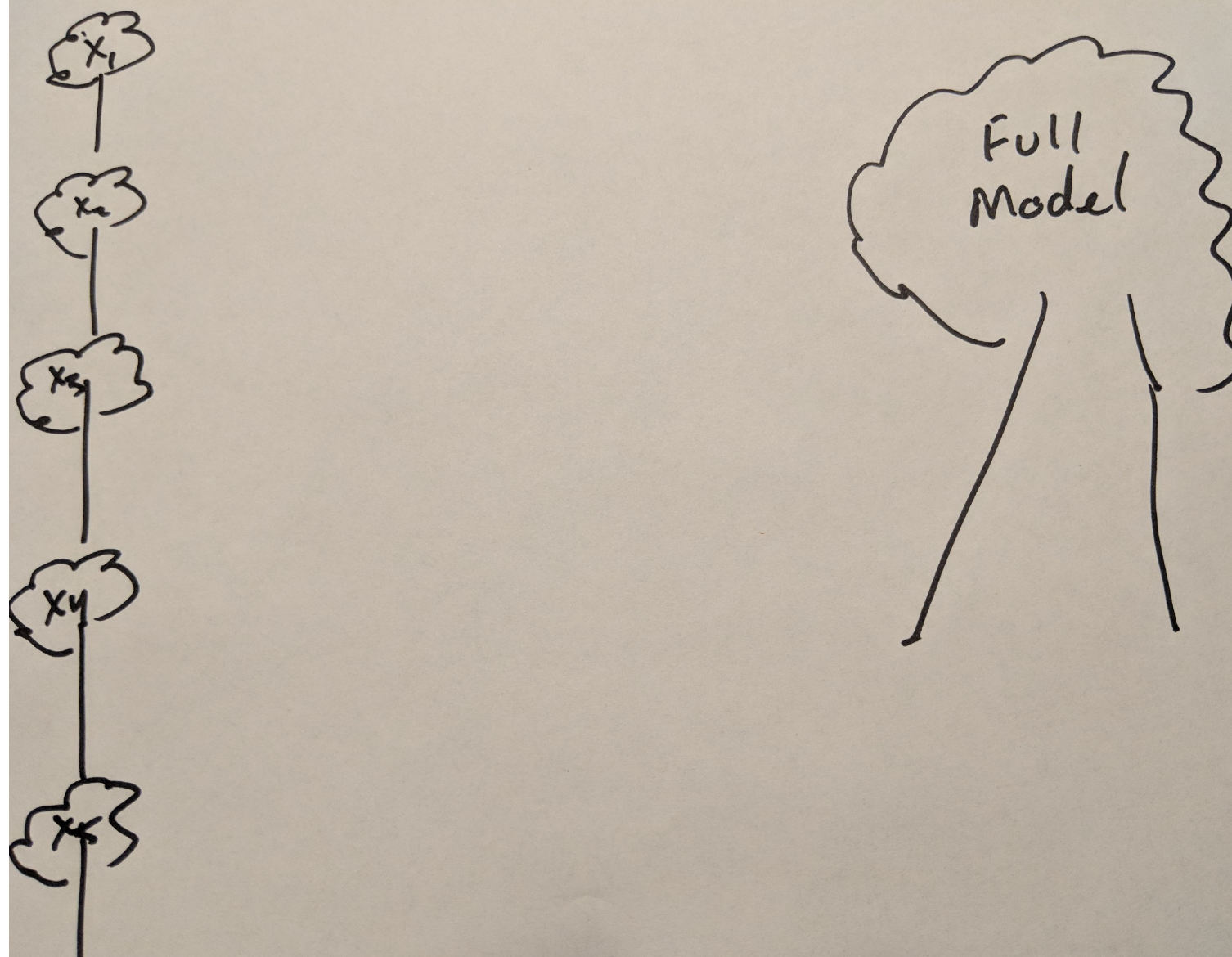
Correlation: Brozek, Age, Weight, Height, Adiposity, Abdomen

Recall

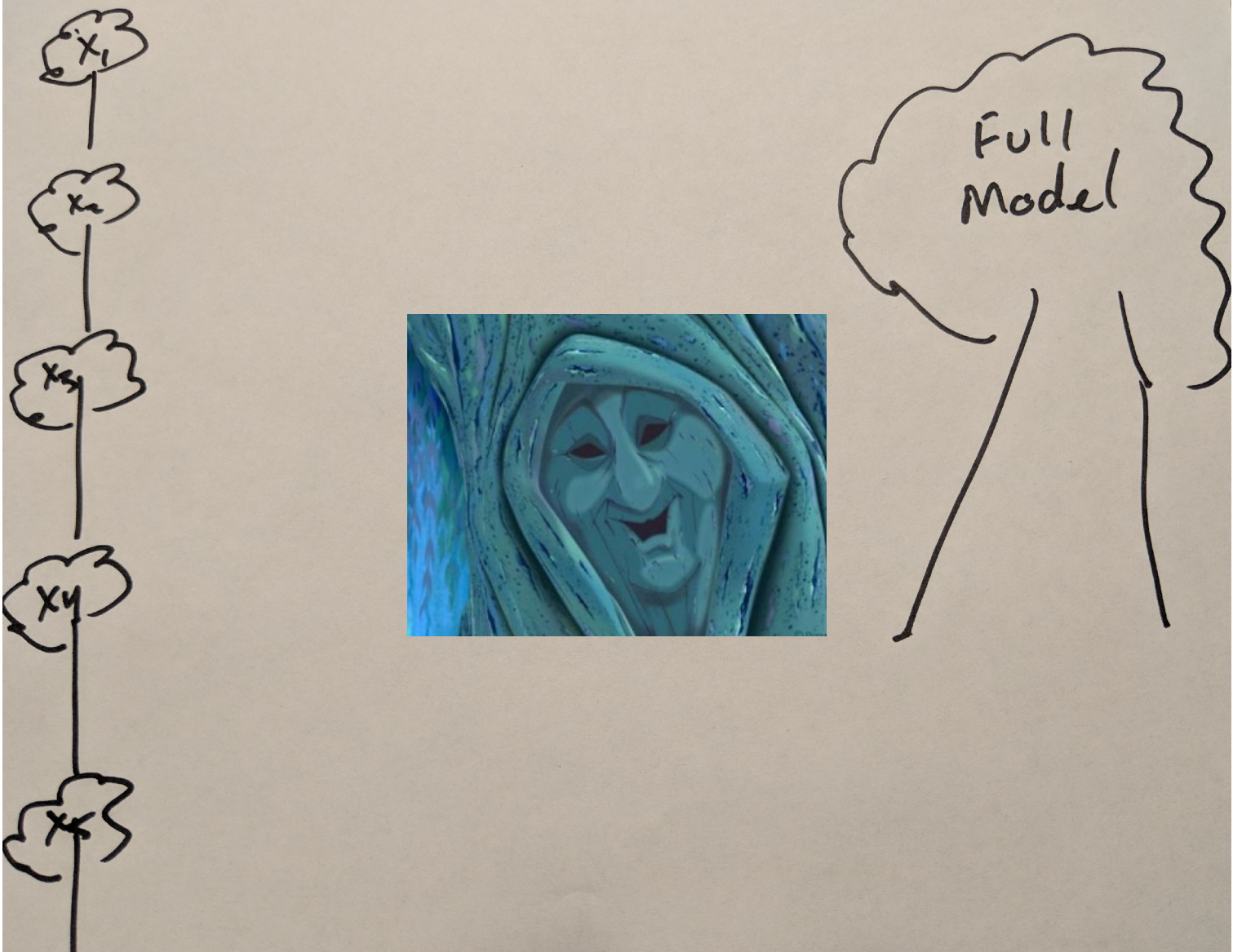
	Brozek	Age	Weight	Height	Adiposity
Age	0.289 0.000				
Weight	0.613 0.000	-0.013 0.840			
Height	-0.023 0.721	-0.245 0.000	0.489 0.000		
Adiposity	0.728 0.000	0.119 0.060	0.887 0.000	0.041 0.514	
Abdomen	0.814 0.000	0.230 0.000	0.888 0.000	0.192 0.002	0.924 0.000

Cell Contents: Pearson correlation  
P-Value

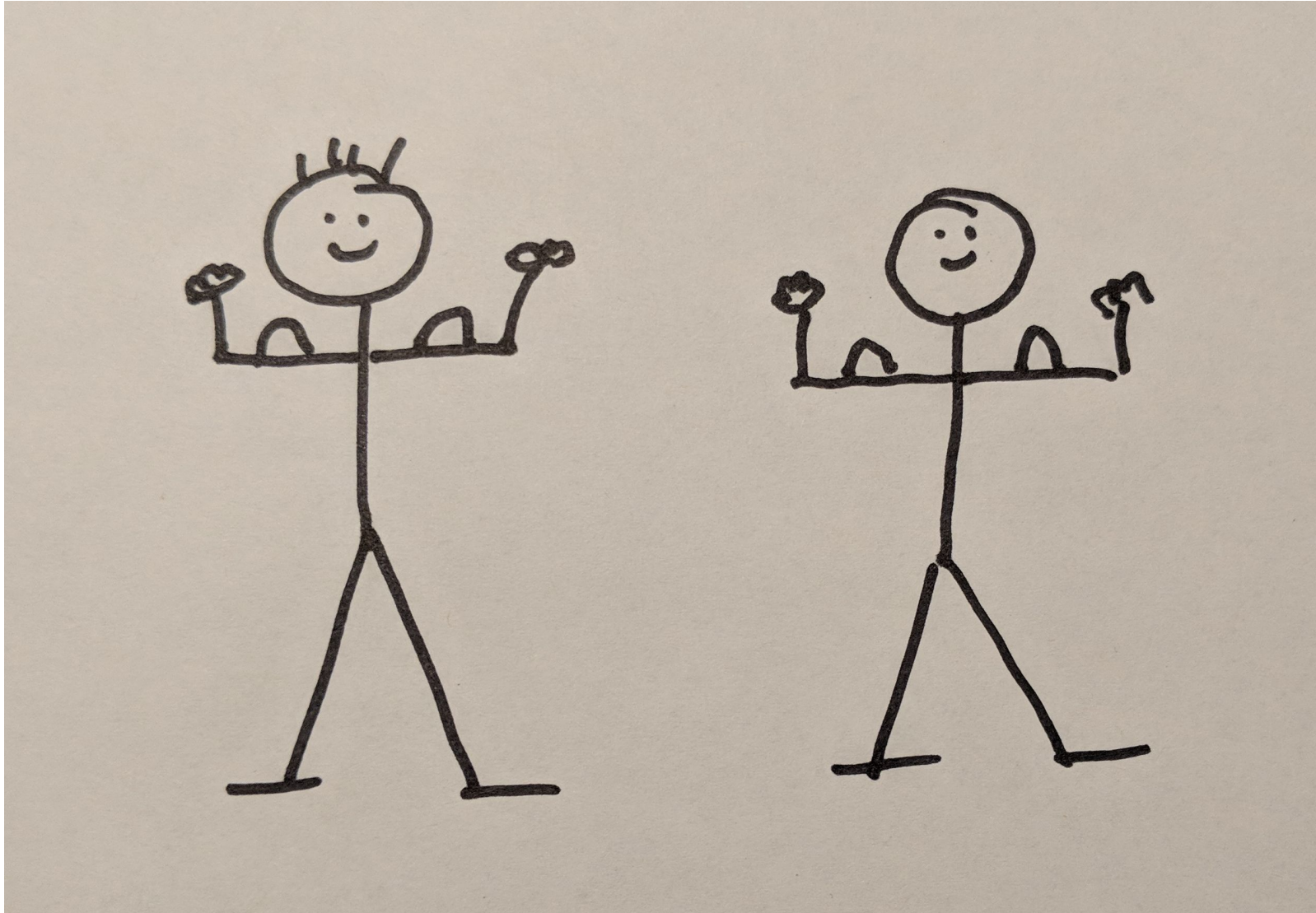
# The "Forest" of Regression Models



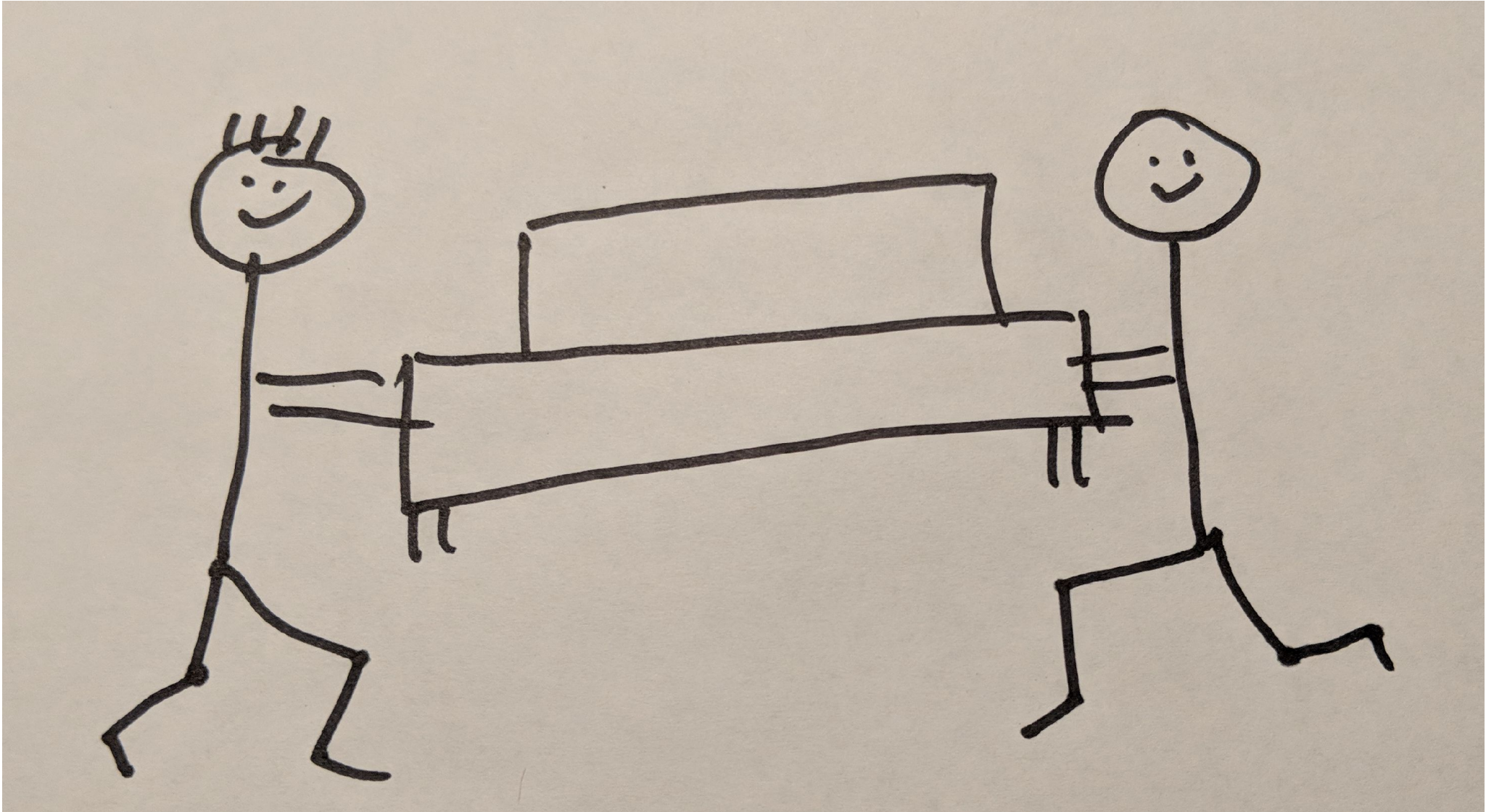
# The "Forest" of Regression Models



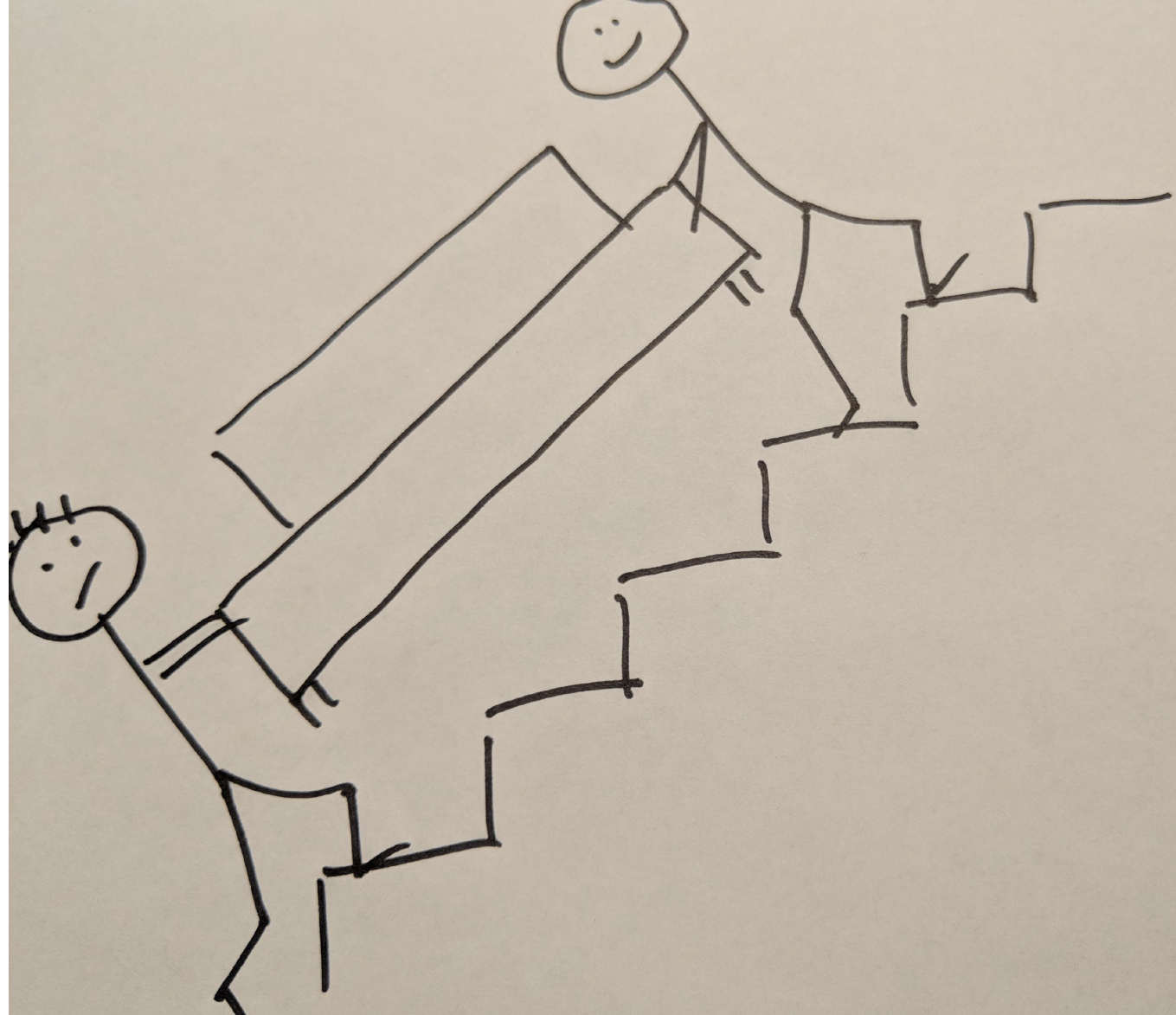
# Two Strong “Movers”



# Unloading the truck ...



# Navigating the stairs ...



# Regression Analysis: Brozek versus Abdomen, Adiposity

## Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	10042.3	5021.14	248.23	0.000
Adiposity	1	58.2	58.19	2.88	0.091
Abdomen	1	2050.8	2050.77	101.38	0.000
Error	249	5036.7	20.23		
Lack-of-Fit	248	5022.2	20.25	1.39	0.603
Pure Error	1	14.6	14.58		
Total	251	15079.0			

## Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
4.49754	66.60%	66.33%	65.04%

**Which variable  
is doing the  
“heavy lifting”?**

## Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-36.40	2.55	-14.25	0.000	
Adiposity	-0.345	0.203	-1.70	0.091	6.83
Abdomen	0.6927	0.0688	10.07	0.000	6.83

## Regression Equation

$$\text{Brozek} = -36.40 - 0.345 \text{ Adiposity} + 0.6927 \text{ Abdomen}$$

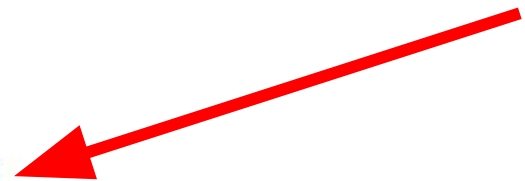
# Describe Association through Correlations

Correlation: Brozek, Age, Weight, Height, Adiposity, Abdomen

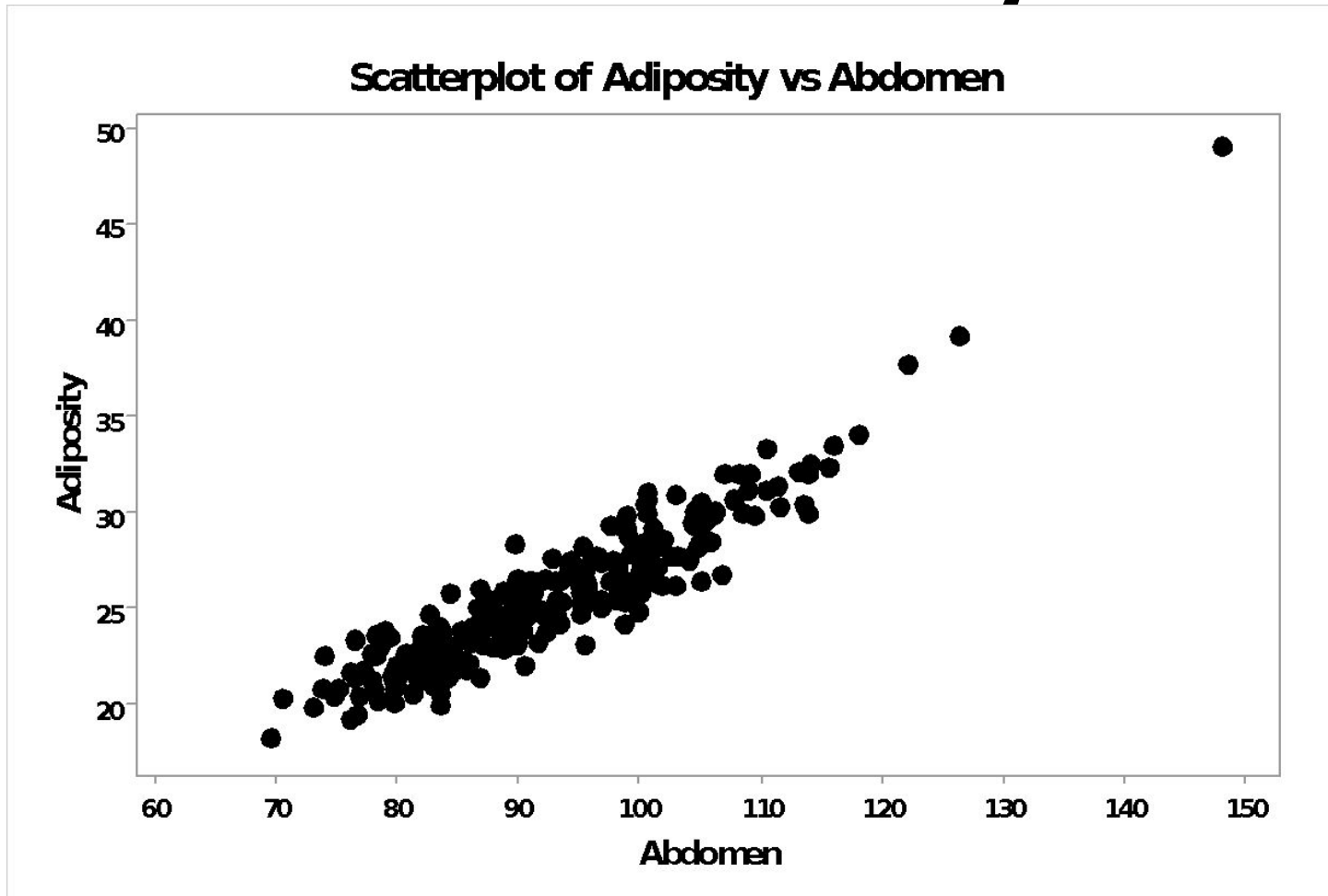
	Brozek	Age	Weight	Height	Adiposity	Abdomen
Age	0.289 0.000					
Weight	0.613 0.000	-0.013 0.840				
Height	-0.023 0.721	-0.245 0.000	0.489 0.000			
Adiposity	0.728 0.000	0.119 0.060	0.887 0.000	0.041 0.514		
Abdomen	0.814 0.000	0.230 0.000	0.888 0.000	0.192 0.002	0.924 0.000	

**Collinearity!**

Cell Contents: Pearson correlation  
P-Value



“Tell me what collinearity looks like!”



$r = 0.924$

“This is what collinearity looks like!”

# Navigating the Forest

- + Check the residual plots at least at the beginning and the “end.”
- + Maintain statistical significance.
- + Monitor the summary statistics, such as the  $R^2$ .
- + Are you in the correct forest?
  - Transformations?
  - Interactions?
  - Powers?
  - Principal components?
  - Unobserved (latent) variables?

# Navigating the Forest (continued)

- + “Surround” the model with simple models and the full model.
- + Celebrate parsimony.
- + Lean toward interpretability.
- + Consider stepwise (automated) variable selection methods.
- + Identify “candidate” models, examine the predictions they make.

# Thank you!

Tom Short

West Chester University of Pennsylvania

[tshort@wcupa.edu](mailto:tshort@wcupa.edu)