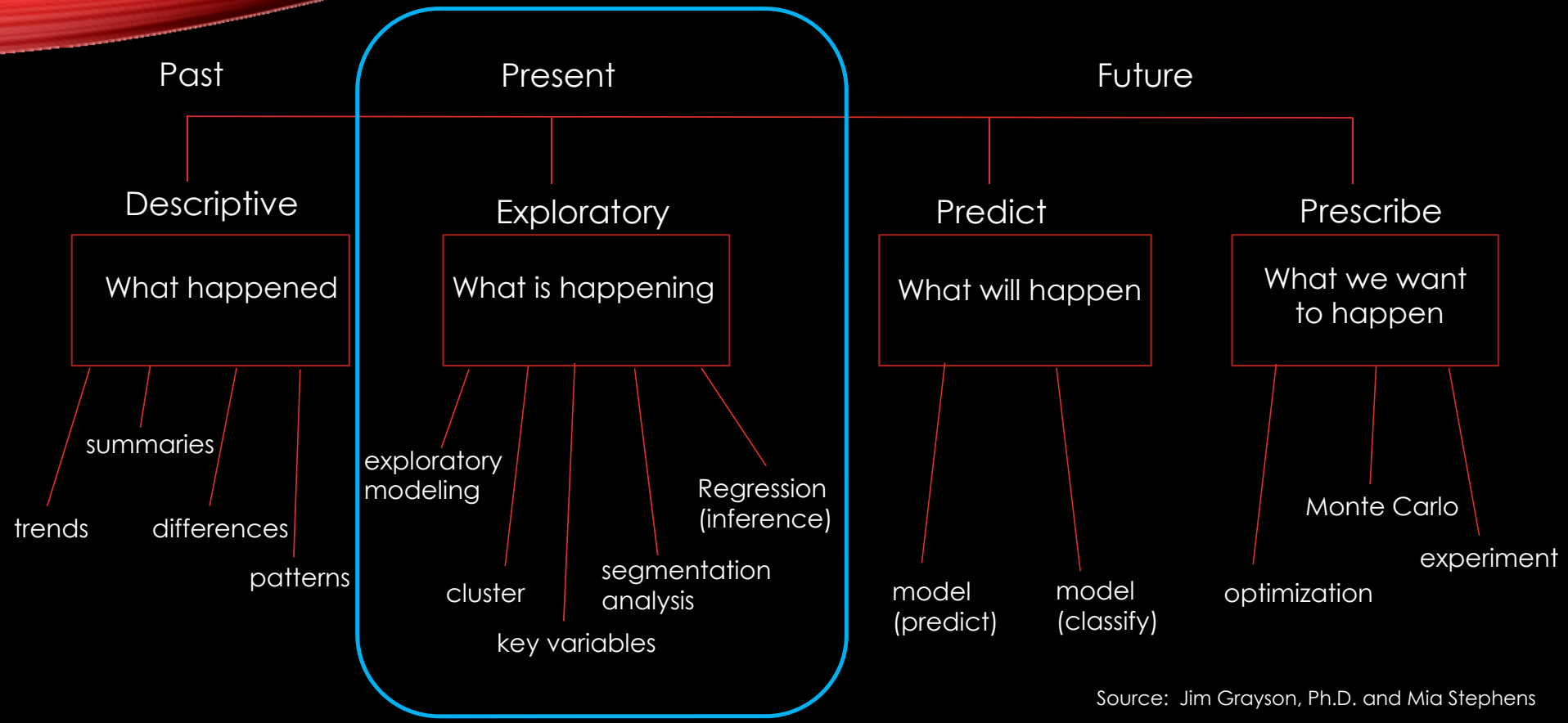
SIMPLE LINEAR REGRESSION



Source: Jim Grayson, Ph.D. and Mia Stephens

Inference

"We are often interested in understanding the way that *Y* is affected as *X*1, . . . , *Xp* change. In this situation we wish to estimate *f*, but our goal is not necessarily to make predictions for *Y* . We instead want to understand the relationship between *X* and *Y* , or more specifically, to understand how *Y* changes as a function of *X*1, . . . ,*Xp*. Now $\hat{f}$ cannot be treated as a black box, because we need to know its exact form."

Excerpts from pages 19-20, An Introduction to Statistical Learning, James, et al (Springer)

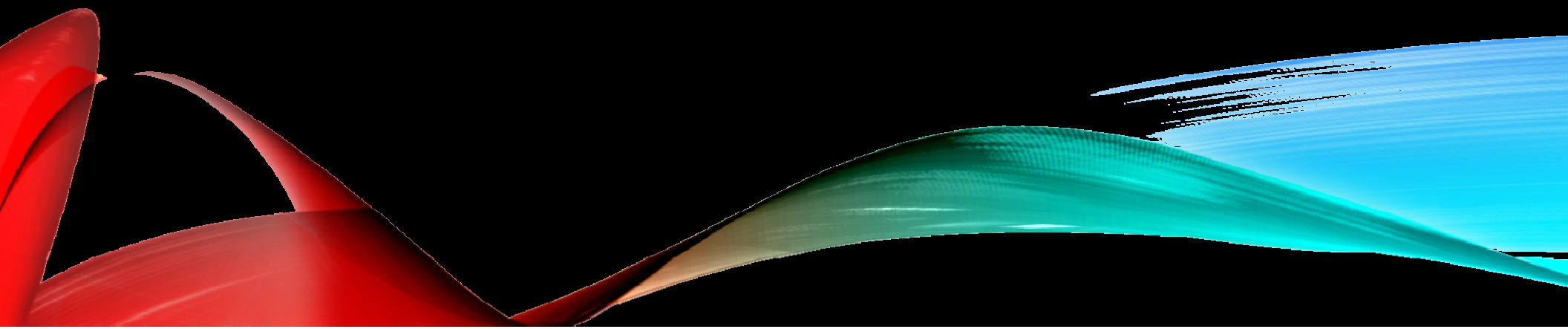
Prediction

"… consider a company that is interested in conducting a direct-marketing campaign. The goal is to identify individuals who will respond positively to a mailing, based on observations of demographic variables measured on each individual. In this case, the demographic variables serve as predictors, and response to the marketing campaign (either positive or negative) serves as the outcome. The company is not interested in obtaining a deep understanding of the relationships between each individual predictor and the response; instead, the company simply wants an accurate model to predict the response using the predictors. This is an example of modeling for prediction."

Excerpts from pages 19-20, An Introduction to Statistical Learning, James, et al (Springer)

- What is it?
- What can it do? (use cases)
- How does it work?
- JMP Mechanics
- Interpret results (statistically)
- Interpret results (operationally)
- How to implement the results
- How to understand the managerial implications

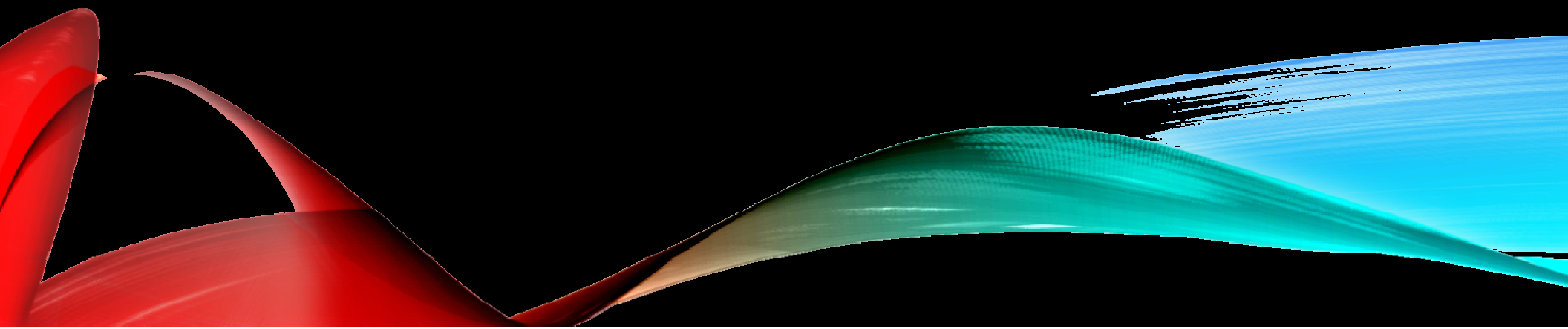
WHAT IS IT?



Simple linear regression is an approach for predicting a quantitative response variable (Y) with a single predictor variable (X).

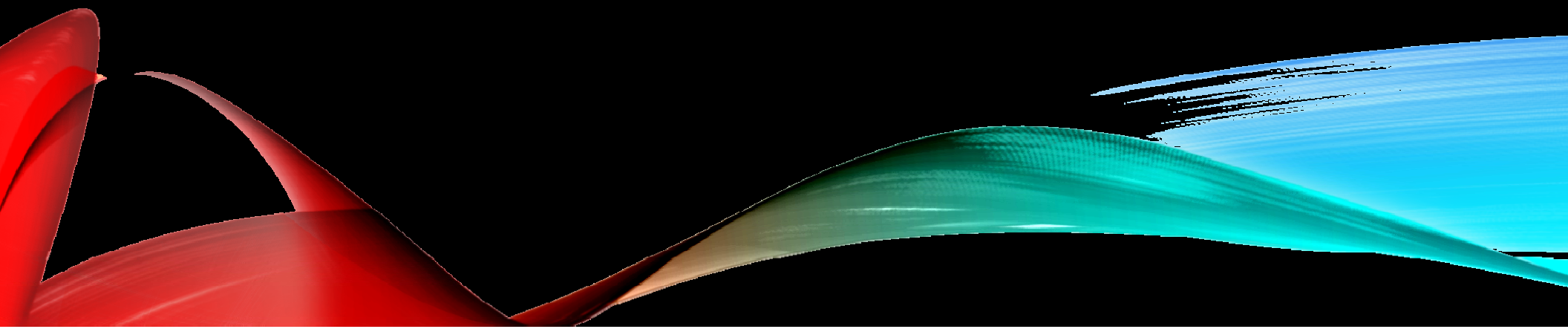
Y	X	Objective	Summary of Fit Measures	Statistical Significance Measure	Operational Significance
Continuous	Continuous or Categorical	Explanatory	RSquare Adj, Root Mean Square Error	Prob > F (p-value)	Mean and Individual Confidence Limits & Variable CI

WHAT CAN IT DO? (USE CASES)



- Is there a relationship between advertising budget and sales?
- Is there a relationship between sales and price?
- Which media contribute to sales?
- Which media contribute the most to sales?
- Understanding housing prices

HOW DOES IT WORK?



Least Square Regression Line

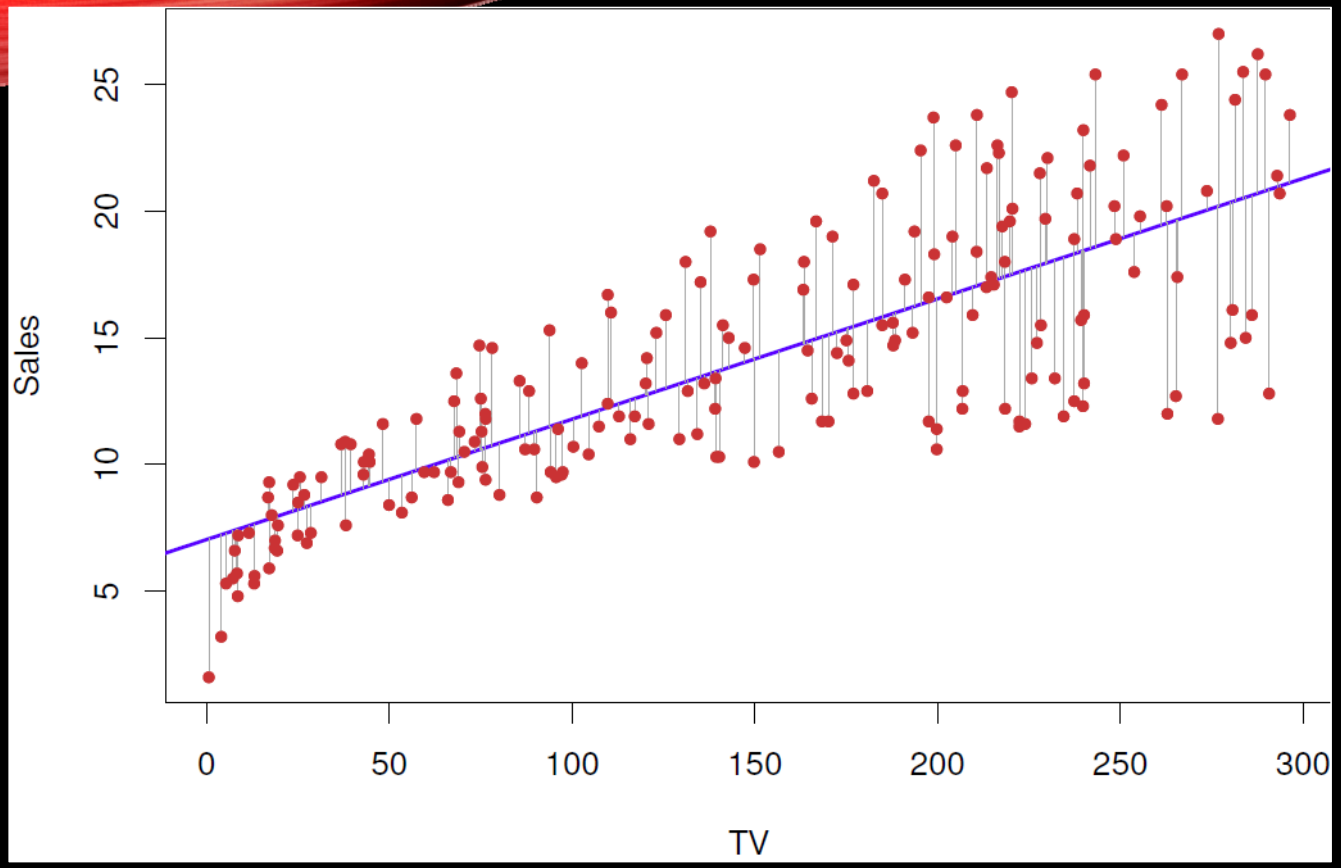


Figure 3.1, **An Introduction to Statistical Learning** by James, et al, (Springer 2013)

Least Squares Regression:

Fit line to minimize the sum of squared residuals
$e_1^2 + e_2^2 + \ldots + e_n^2$
where $e = y - \hat{y}$

Our model is:
$y = \beta_0 + \beta_1 + \varepsilon$

We estimate the β's by finding a regression line
$\hat{y} = b_0 + b_1 x$

Figure 10.1 Straight-Line Least Squares Regression

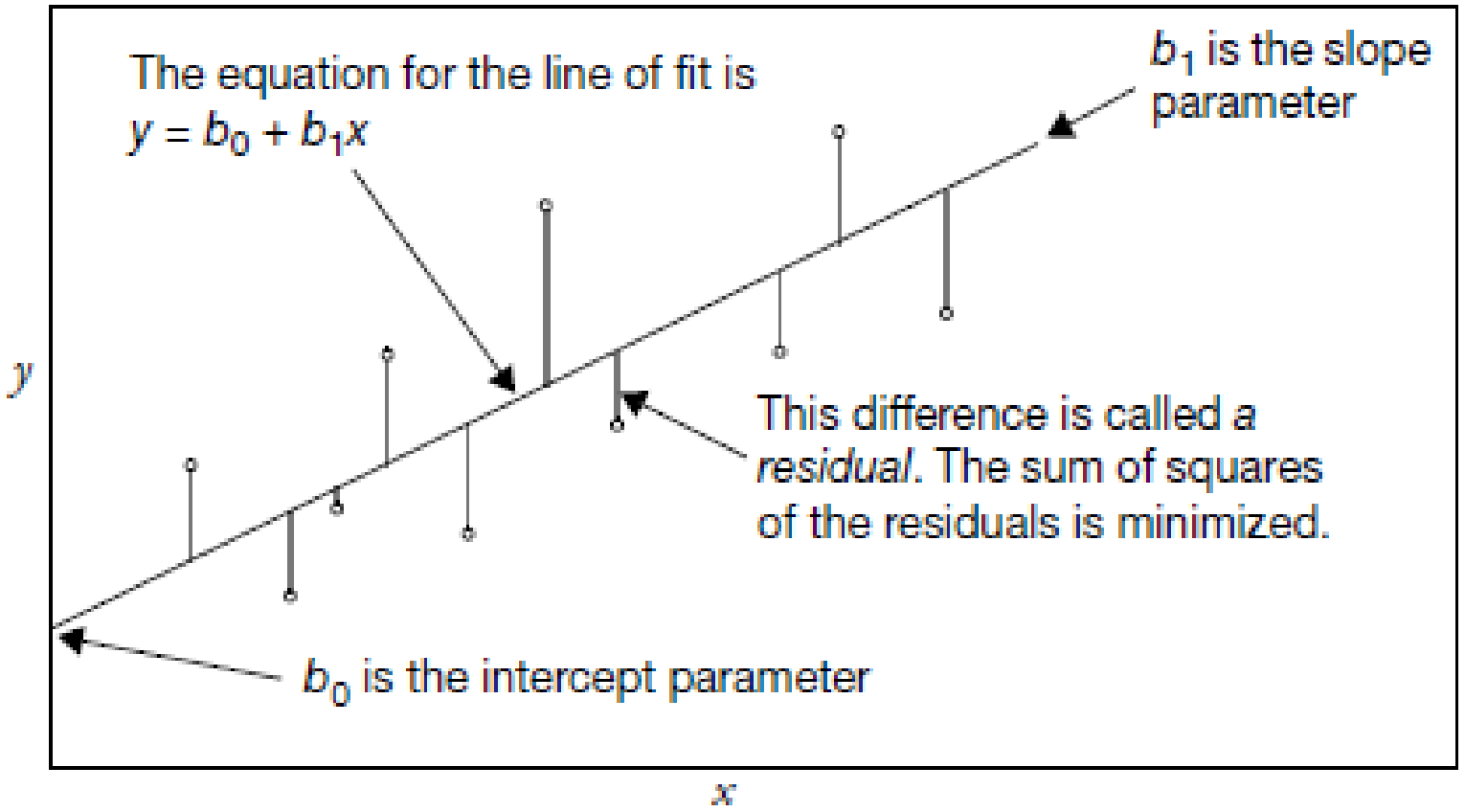
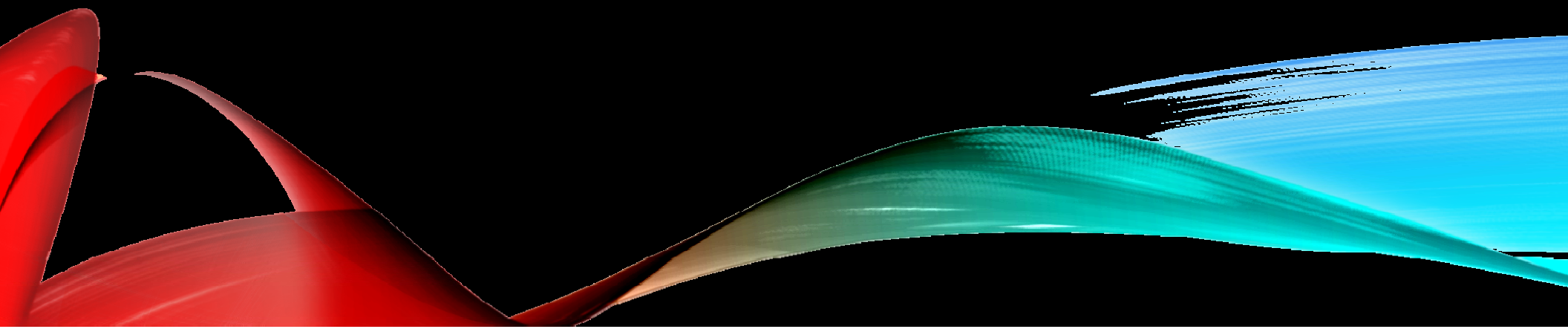


Figure 10.1, **JMP Start Statistics**, 4e by Sall, et al, (SAS 2012)

JMP MECHANICS



Simple linear regression is used to model the relationship between two continuous variables.

Simple Linear Regression Using Fit Y by X

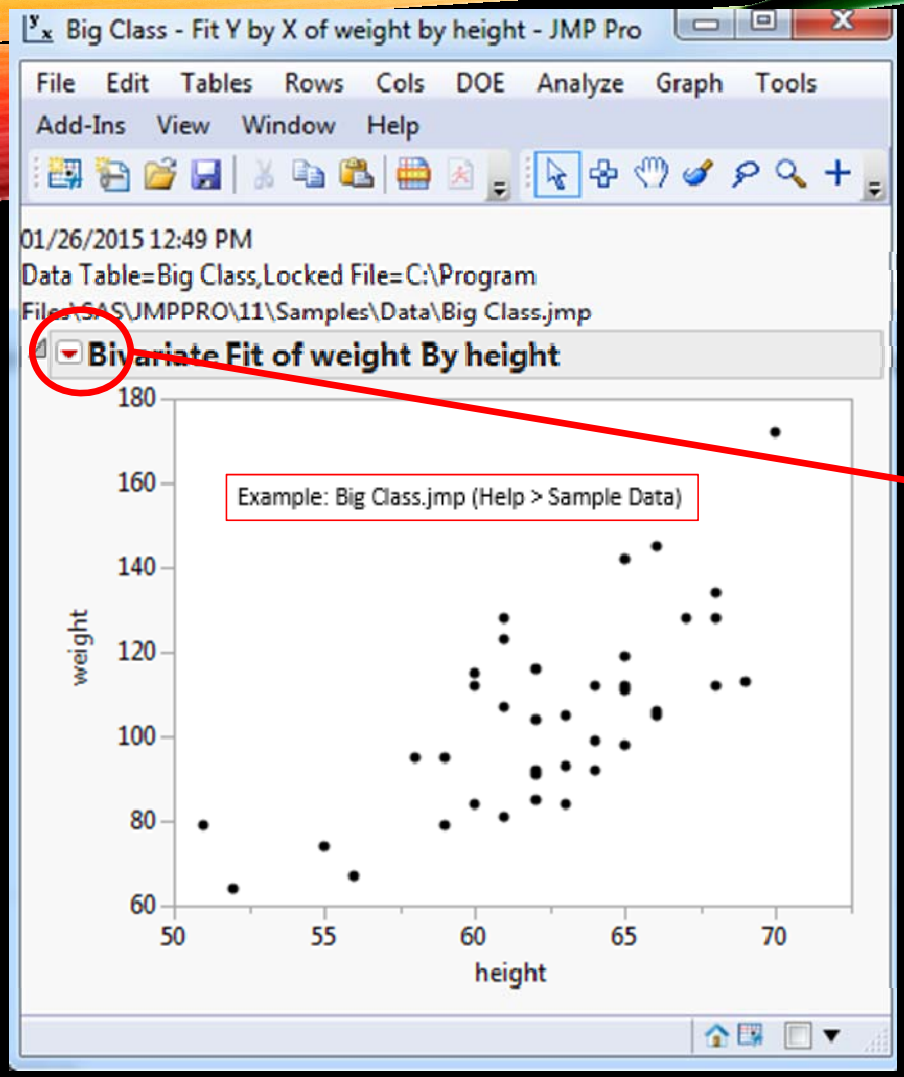
1. From an open JMP® data table, select **Analyze > Fit Y by X.**
2. Click on a continuous variable from **Select Columns**, and click **Y, Response** (continuous variables have blue triangles).
3. Select a second continuous variable, and click **X, Factor**.
4. Click **OK** to generate a scatterplot.
5. To fit a regression line, click on the **red triangle** and select **Fit Line**.

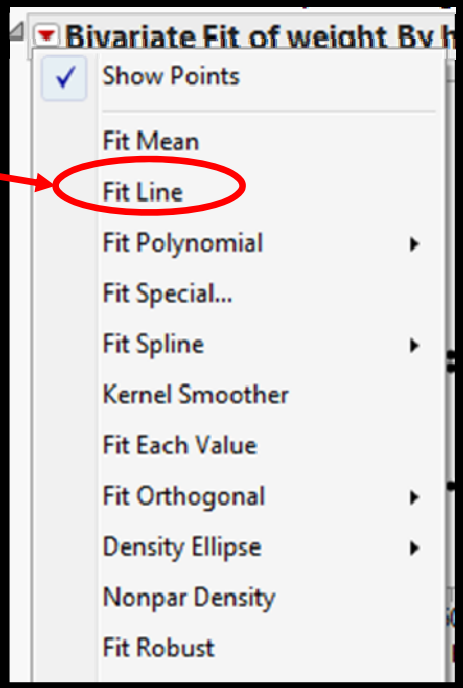
By default, JMP will provide the following results:

- The regression equation (under Linear Fit).
- The Summary of Fit.
- Lack of Fit (if the data table includes replicates of X values).
- The ANOVA table.
- The parameter estimates.

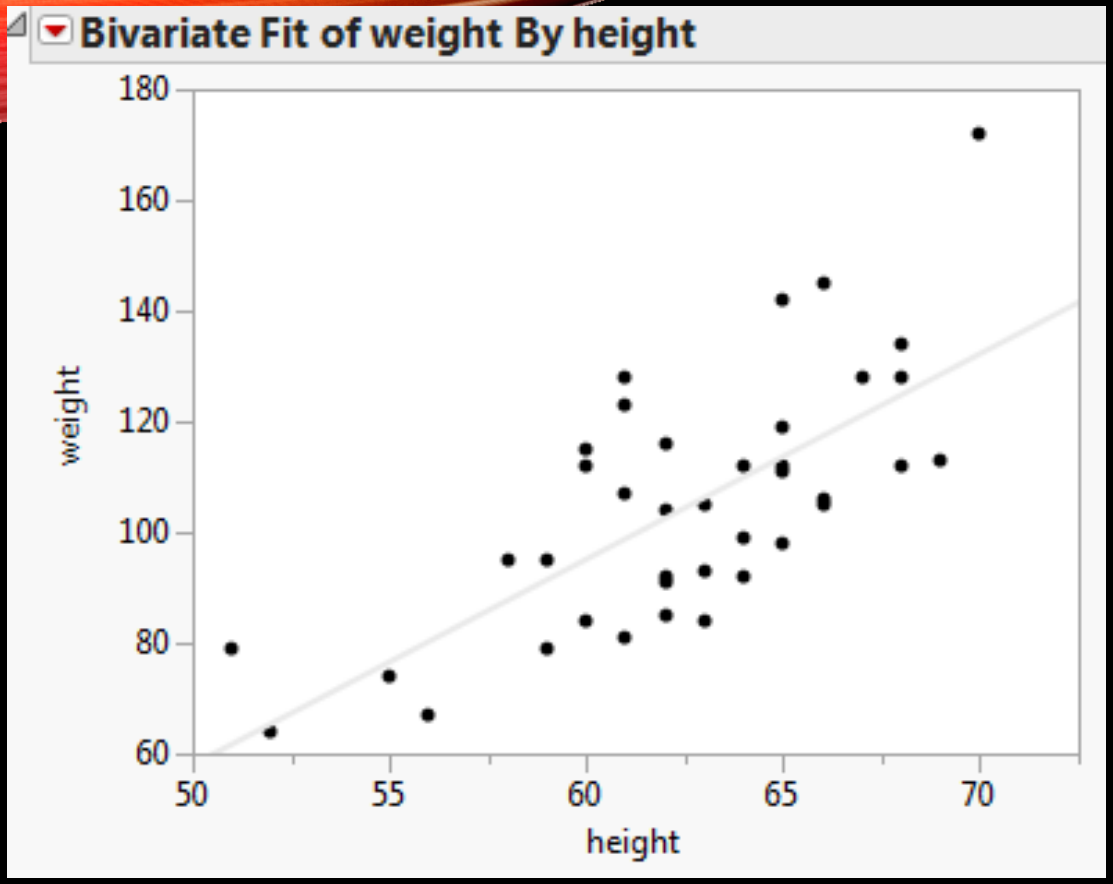
Make a scatterplot to display the relationship between two quantitative variables

From red triangle select Fit Line:





When the line is fit:



Notes: Simple linear regression can also be performed from **Analyze > Fit Model**. For more details on regression analysis, see the book ***Basic Analysis*** (under **Help > Books**) or search for "regression" in the JMP Help.

Linear Fit

Linear Fit

weight = -127.1452 + 3.7113549*height

Summary of Fit

RSquare	0.502917
RSquare Adj	0.489836
Root Mean Square Error	15.85786
Mean of Response	105
Observations (or Sum Wgts)	40

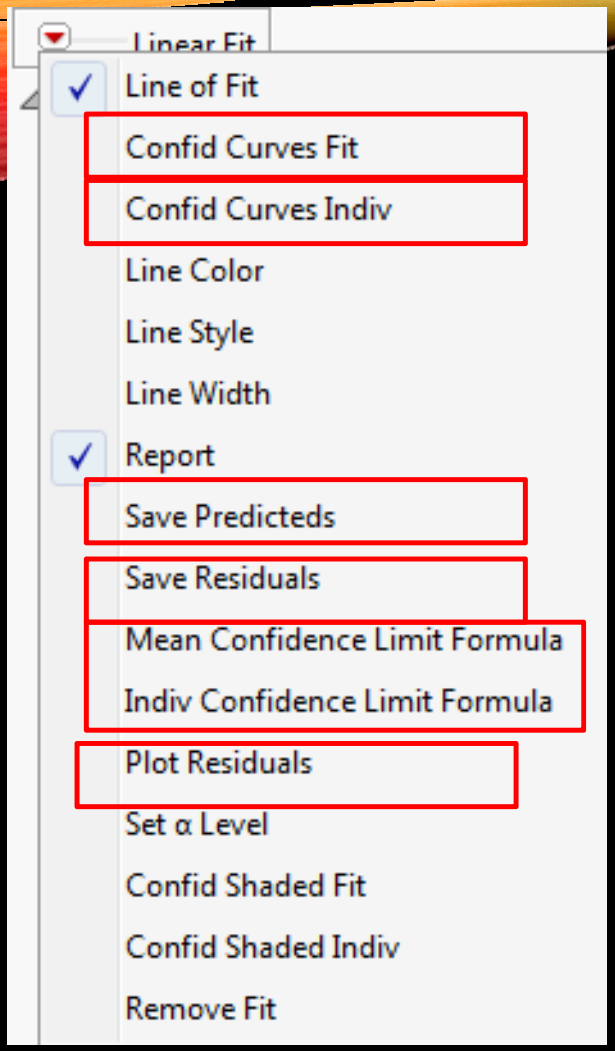
Lack Of Fit

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	9668.079	9668.08	38.4460
Error	38	9555.921	251.47	Prob > F
C. Total	39	19224.000		<.0001*

Parameter Estimates

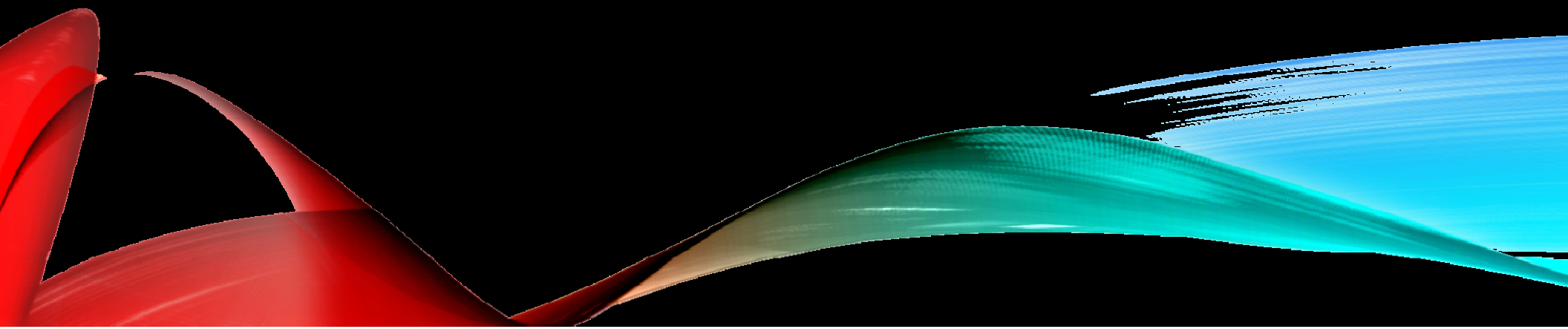
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-127.1452	37.52372	-3.39	0.0016*
height	3.7113549	0.598559	6.20	<.0001*



From red triangle select additional options:

Examine the residuals to assess the quality of the model

MODEL STATISTICAL VALIDITY



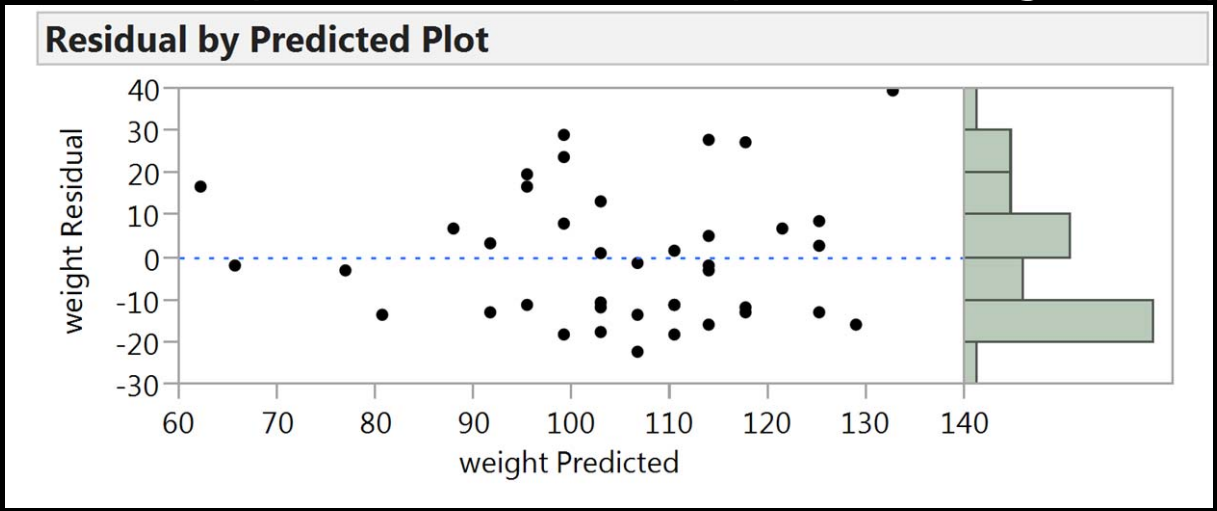
ASSUMPTIONS FOR INFERENCE ABOUT REGRESSION COEFFICIENTS

L – I – E – N

ASSUMPTIONS FOR INFERENCE ABOUT REGRESSION COEFFICIENTS

L – I – E – N

- Linearity Assumption
  - Scatterplot of y vs. x
  - Scatterplot of residuals plotted against predicted values



Chapter 15, Business Statistics 3e by Sharpe, De Veaux and Velleman (Pearson)

ASSUMPTIONS FOR INFERENCE ABOUT REGRESSION COEFFICIENTS

L – I – E – N

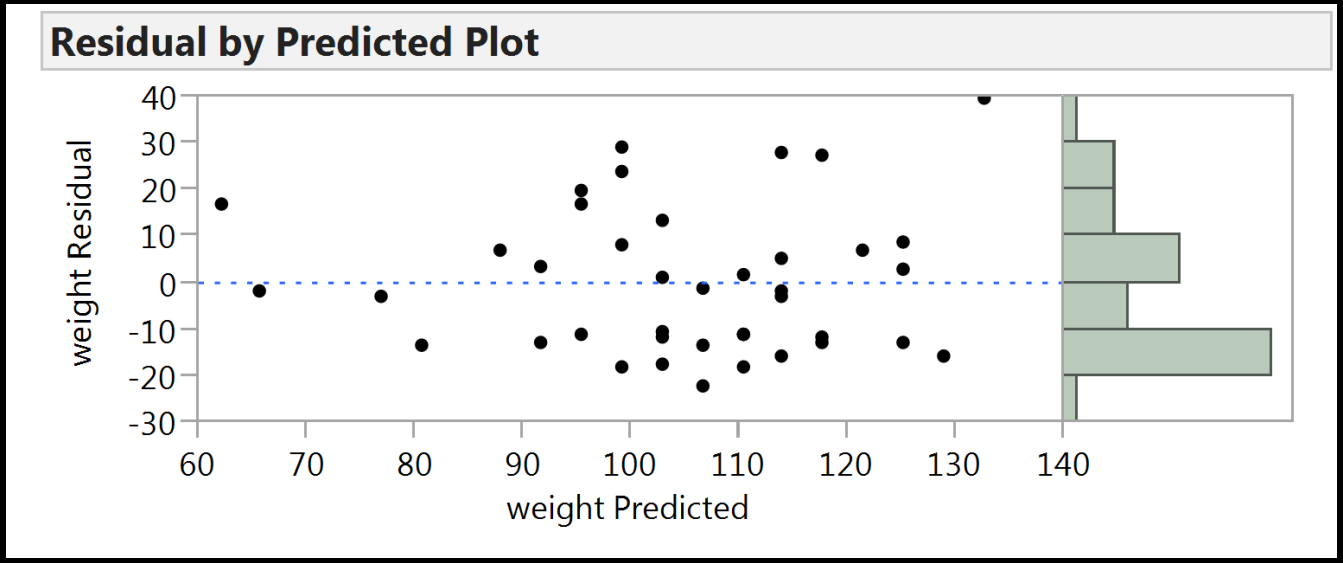
- Independence Assumption (Linearity Condition)
  - Plausible if collected with appropriate randomization

Chapter 15, Business Statistics 3e by Sharpe, De Veaux and Velleman (Pearson)

ASSUMPTIONS FOR INFERENCE ABOUT REGRESSION COEFFICIENTS

L – I – E – N

- Equal Variance Assumptions (Equal Spread Condition)
  - Scatterplot of the residuals vs predicted values

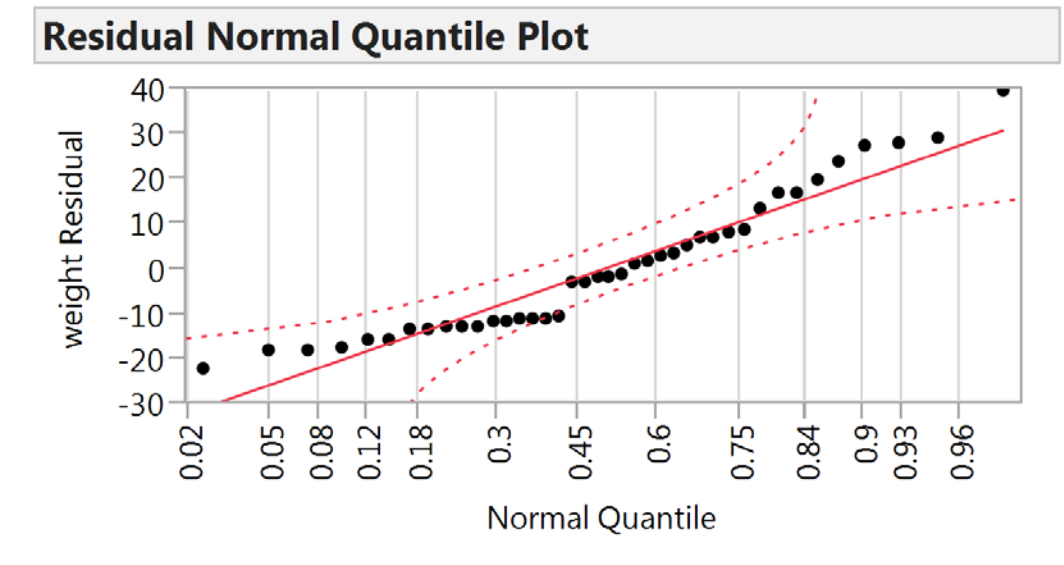


Chapter 15, Business Statistics 3e by Sharpe, De Veaux and Velleman (Pearson)

ASSUMPTIONS FOR INFERENCE ABOUT REGRESSION COEFFICIENTS

L – I – E – N

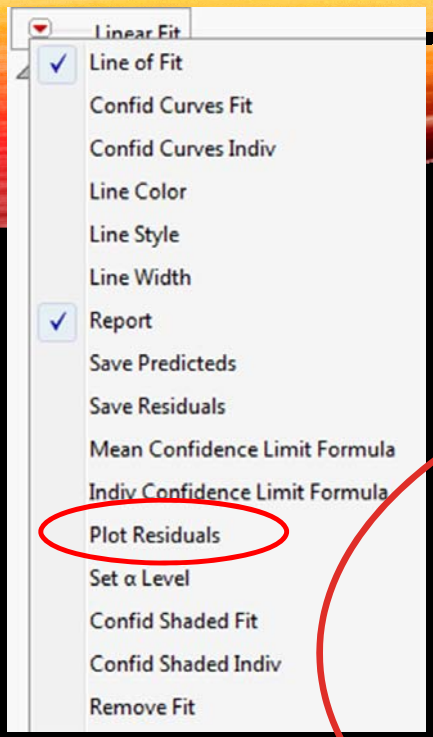
- Normal Probability Assumptions (Nearly Normal Condition)
  - Histogram or Normal Probability Plot of residuals

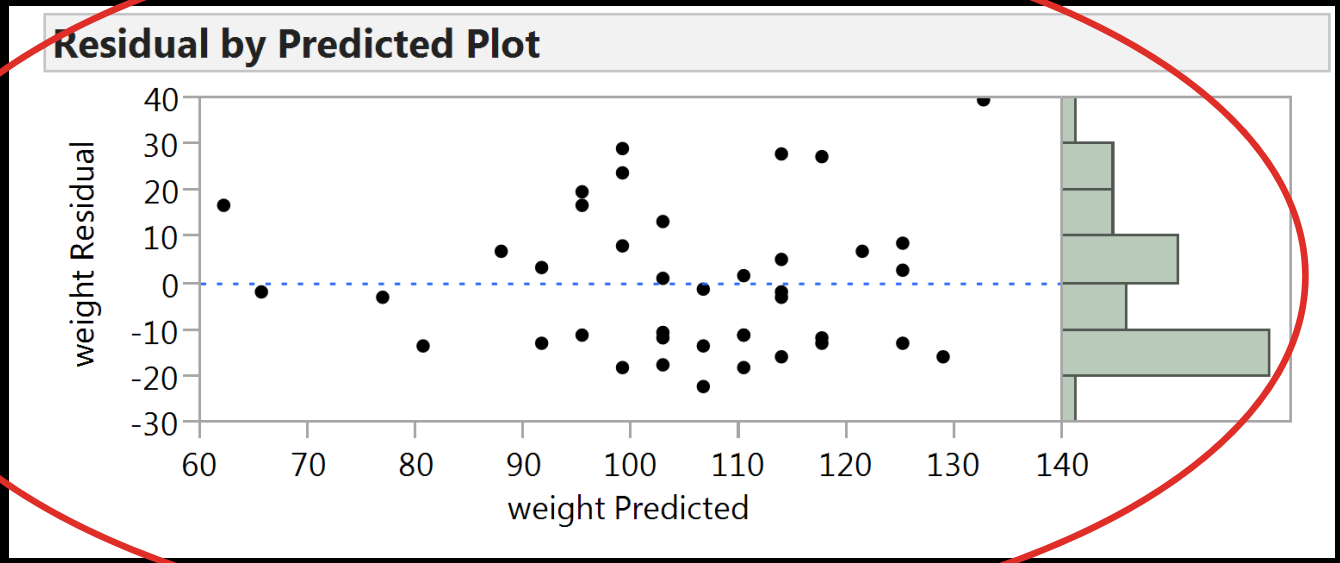


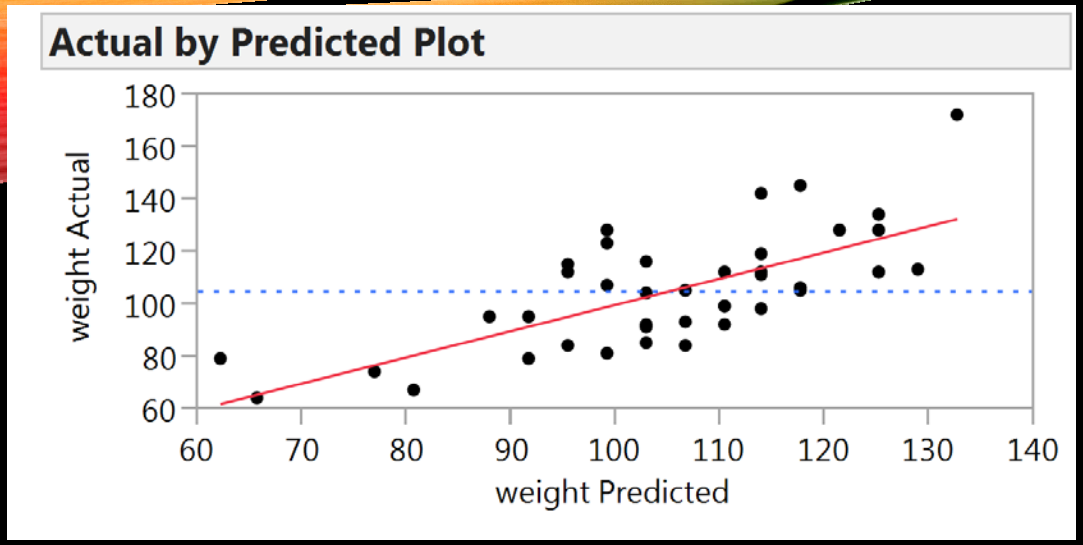
Chapter 15, Business Statistics 3e by Sharpe, De Veaux and Velleman (Pearson)

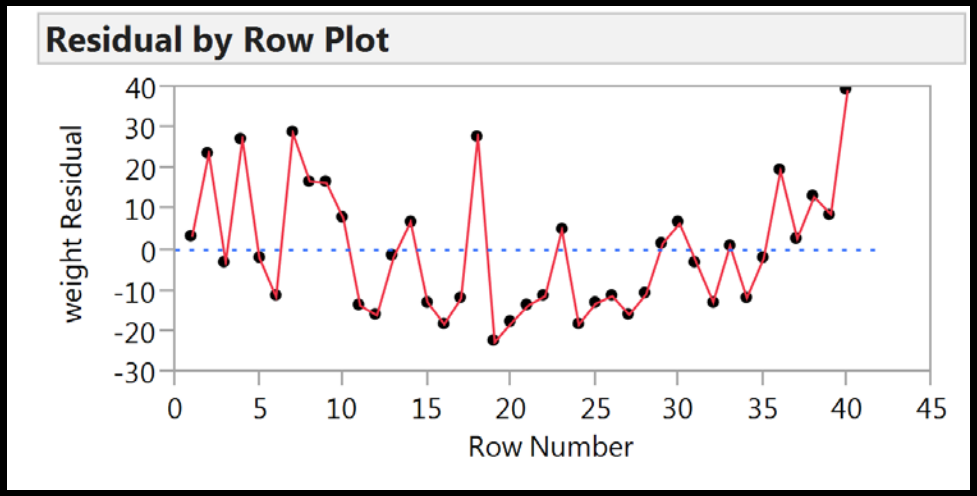
Red Triangle Linear Fit > Plot Residuals

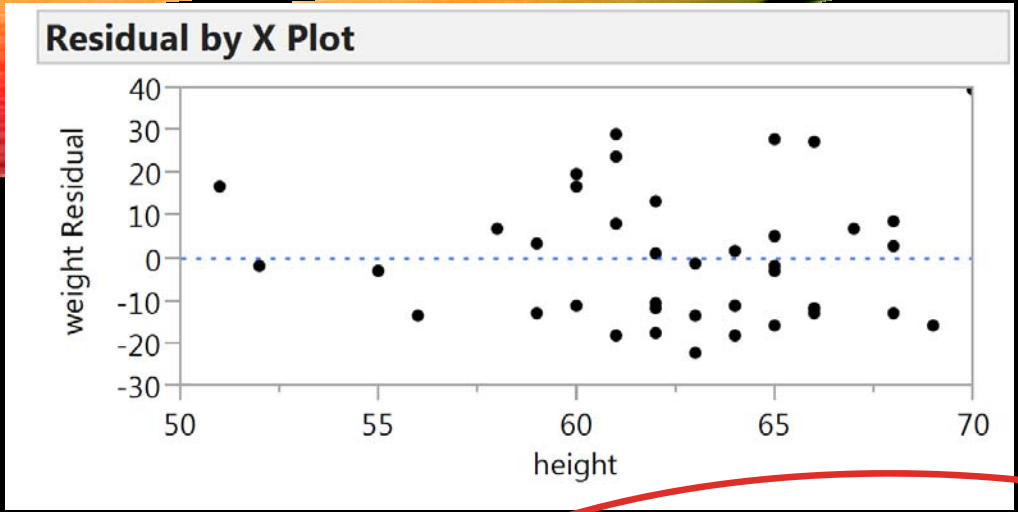
Diagnostic Plots

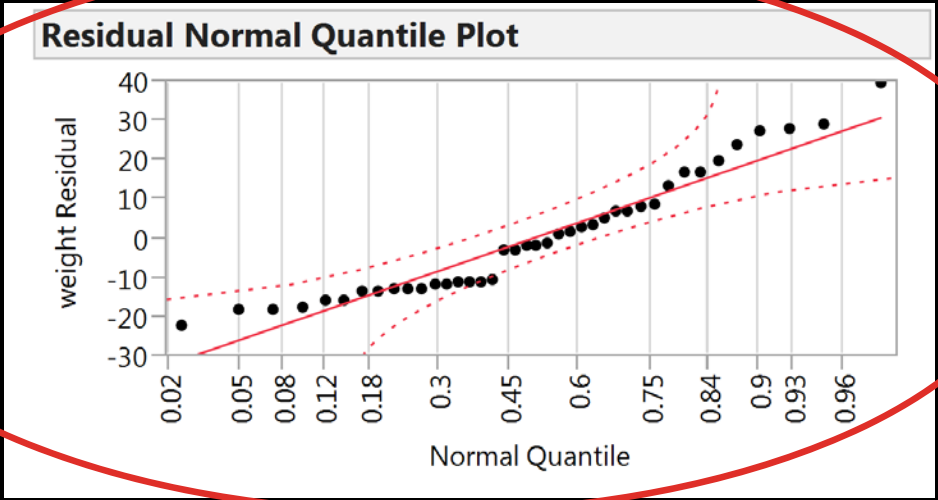




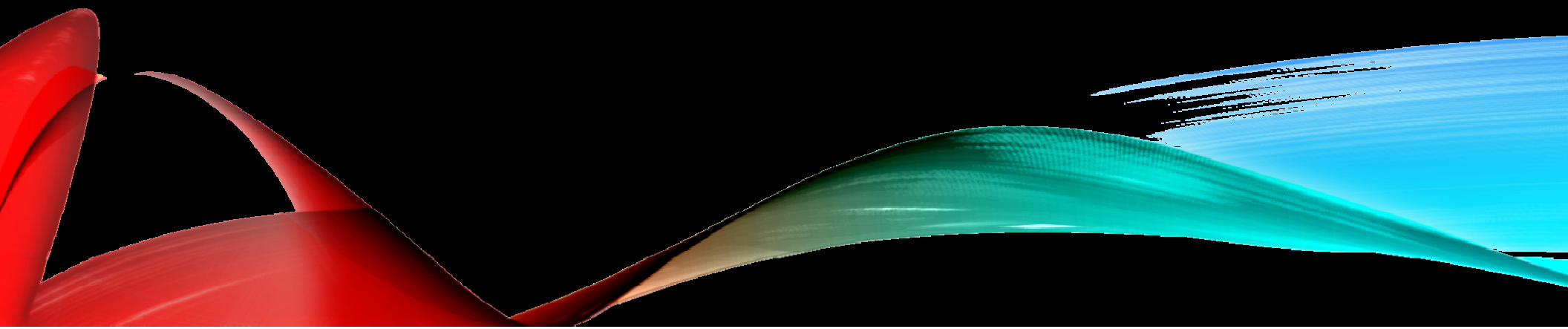

Actual by Predicted Plot
weight Actual
180
160
140
120
100
80
60
60
70
80
90
100
110
120
130
140
weight Predicted


Residual by Row Plot
weight Residual
40
30
20
10
0
-10
-20
-30
0
5
10
15
20
25
30
35
40
45
Row Number


Residual by X Plot
weight Residual
40
30
20
10
0
-10
-20
-30
50
55
60
65
70
height


Residual Normal Quantile Plot
weight Residual
40
30
20
10
0
-10
-20
-30
0.02
0.05
0.08
0.12
0.18
0.3
0.45
0.6
0.75
0.84
0.9
0.93
0.96
Normal Quantile

INTERPRET RESULTS (STATISTICALLY)



Linear Fit

Linear Fit

weight = -127.1452 + 3.7113549*height

Summary of Fit

RSquare	0.502917
RSquare Adj	0.489836
Root Mean Square Error	15.85786
Mean of Response	105
Observations (or Sum Wgts)	40

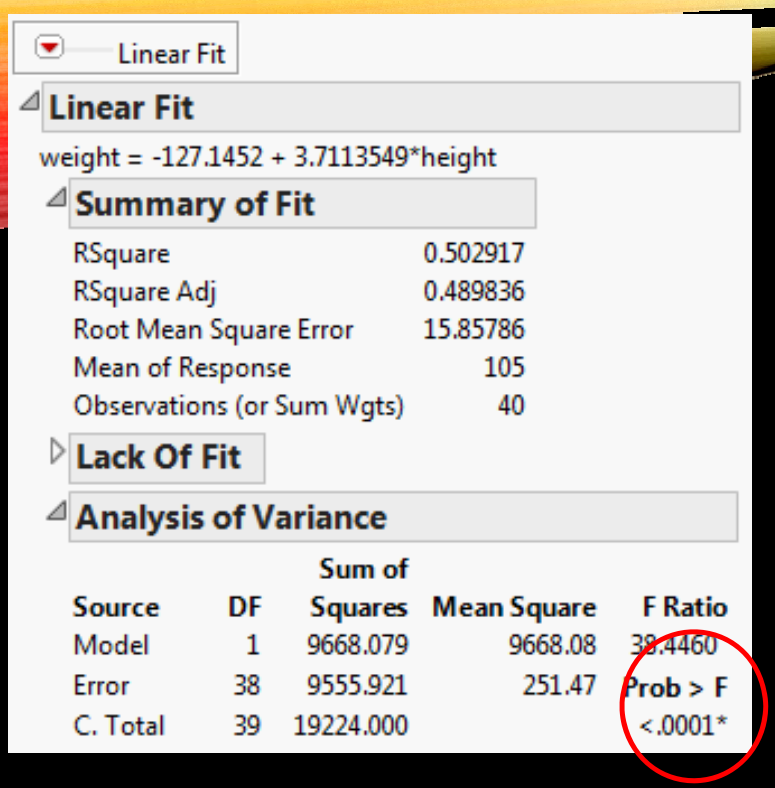
Lack Of Fit

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	9668.079	9668.08	38.4460
Error	38	9555.921	251.47	Prob > F
C. Total	39	19224.000		<.0001*

The Rsquare / Rsquare Adj gives us a measure of the variability in the data explained with our model. Larger values are preferred.

The Root Mean Square Error tells us the standard deviation of the error (ε) – a measure of variability of the model. We might multiply this by 3 to give us some insight into the variability range of roughly ± 48.


Linear Fit

Linear Fit

weight = -127.1452 + 3.7113549*height

Summary of Fit

RSquare	0.502917
RSquare Adj	0.489836
Root Mean Square Error	15.85786
Mean of Response	105
Observations (or Sum Wgts)	40

Lack Of Fit

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	9668.079	9668.08	38.4460
Error	38	9555.921	251.47	Prob > F
C. Total	39	19224.000		<.0001*

Generally the alternate hypothesis for a Regression model is that at least one of the parameters is significant (not equal to 0). In simple linear regression there is only one predictor parameter, so the null for this case is that the β_1 (height) = 0, which would imply that there is no slope and that a horizontal line at the mean weight would describe the association between height and weight.

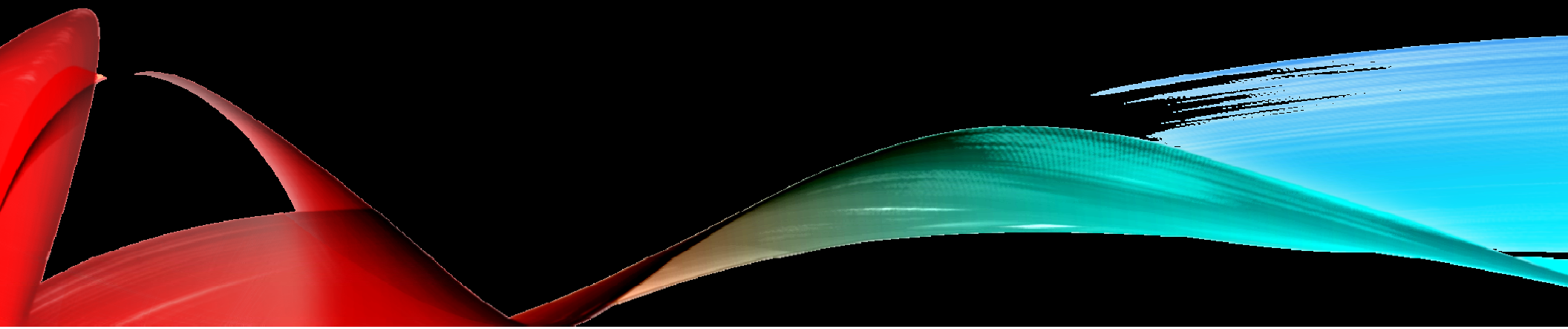
With a p-value (Prob > F) of less than 0.0001 we can reject the Ho and accept Ha that the model parameter (height) has a slope (is not flat) - we are accepting Ha that $\beta_1 \neq 0$.

Parameter Estimates

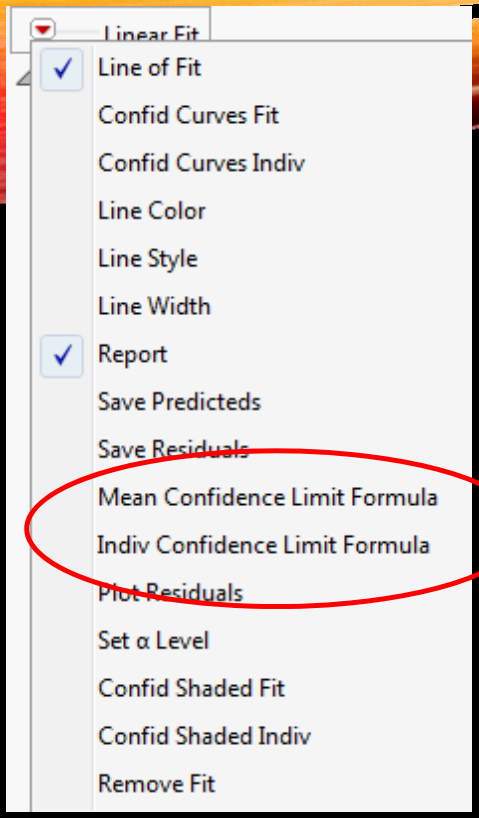
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-127.1452	37.52372	-3.39	0.0016*
height	3.7113549	0.598559	6.20	<.0001*

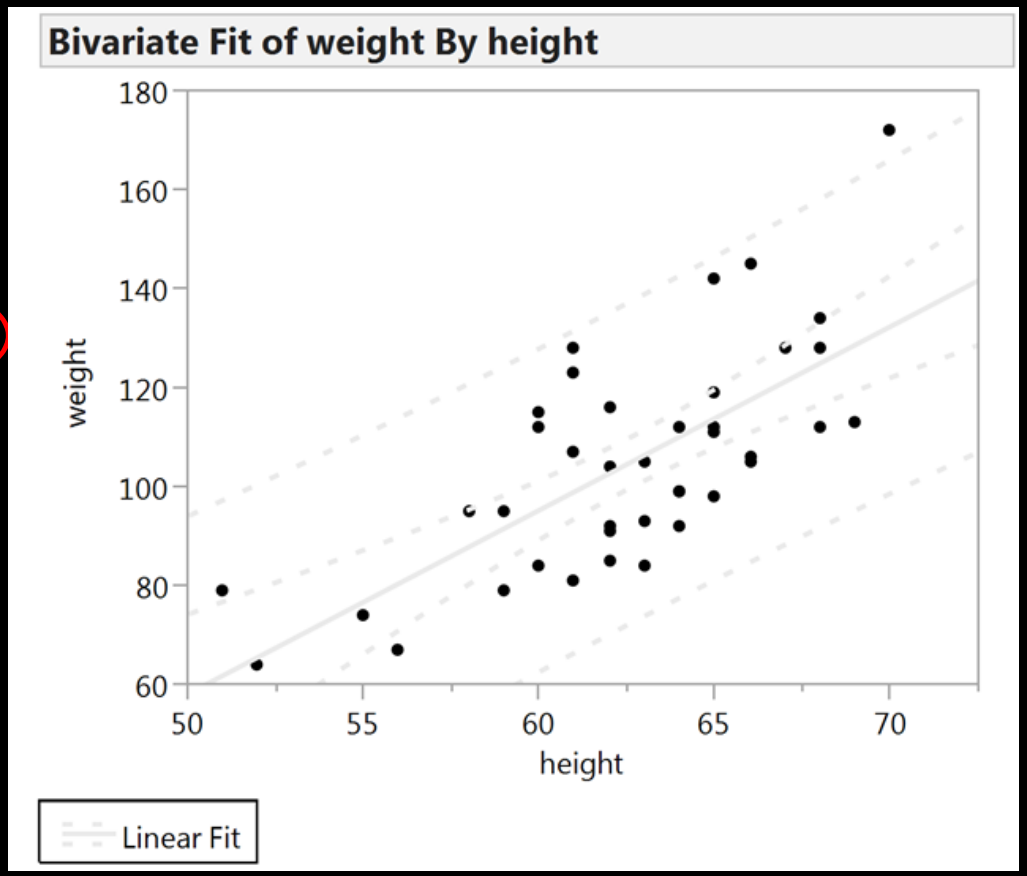
This is somewhat redundant for one predictor, but having rejected the null and accepted the alternate we then look at the p-value of each parameter to determine if it is significant. For height in our example this confirms that $b_1 \neq 0$ because the p-value (Prob > | t |) is <0.0001

INTERPRET RESULTS (OPERATIONALLY)



The Mean Confidence Limit and Individual Confidence Limit provide a better understanding of prediction variability



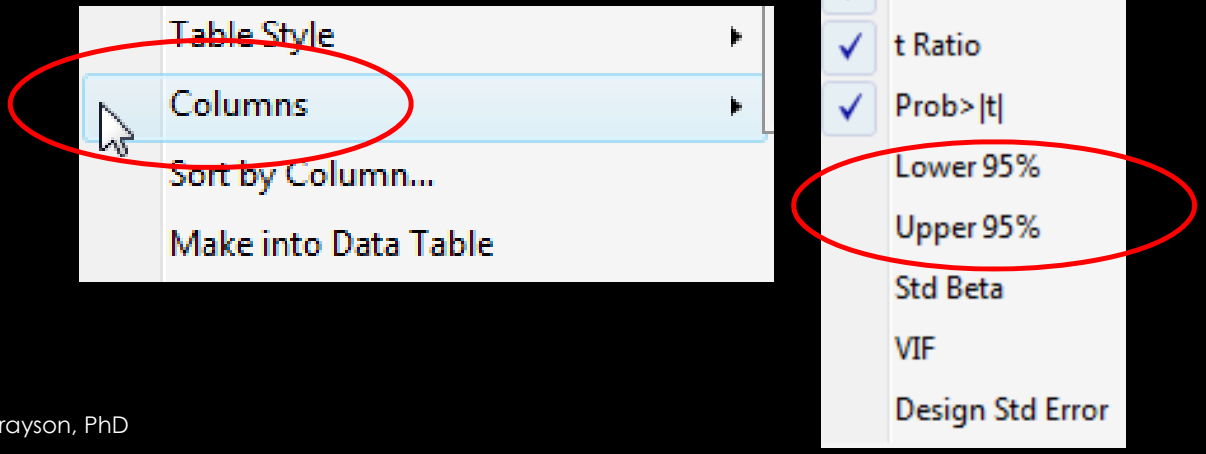


Parameter (Coefficients) Estimation and Variability

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-127.1452	37.52372	-3.39	0.0016*
height	3.7113549	0.598559	6.20	<.0001*

Select the "parameter" table and right click >
Select Columns >
Select Lower 95% and Upper 95%



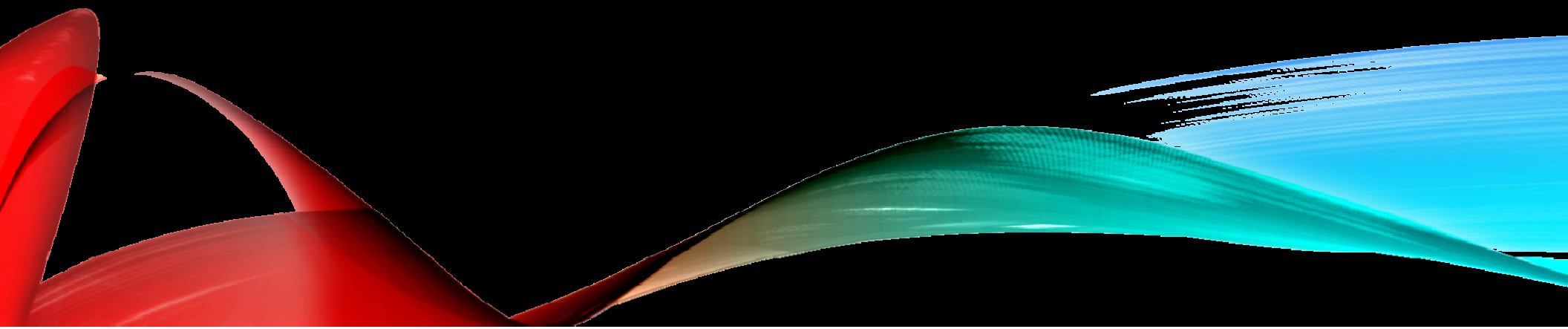
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%
Intercept	-127.1452	37.52372	-3.39	0.0016*	-203.108	-51.18245
height	3.7113549	0.598559	6.20	<.0001*	2.4996359	4.9230739

The Lower 95% and Upper 95% are confidence intervals around the coefficients.

Our coefficient of interest is height. The height coefficient was determined with sample data; therefore, there is uncertainty in the "true" height coefficient. We can interpret the confidence interval values to mean we are 95% confidence the true height coefficient is in the interval (2.5, 4.9).

HOW TO IMPLEMENT THE RESULTS



We apply the results by providing a value for X (height) and then using the simple linear regression formula to compute Y (predicted weight) and other measures such as confidence intervals

	name	age	sex	height	weight	Predicted weight	Lower 95% Mean weight	Upper 95% Mean weight	Lower 95% Indiv weight	Upper 95% Indiv weight
41			•	72	•	140.07230375	127.54697353	152.59763396	105.6127961	174.53181139

Enter 72 under height in a blank row

-127.14524861092

+ 3.71135489385956 * *height*

-127.14524861092

+ 3.71135489385956 * *height*

2.02439416391197

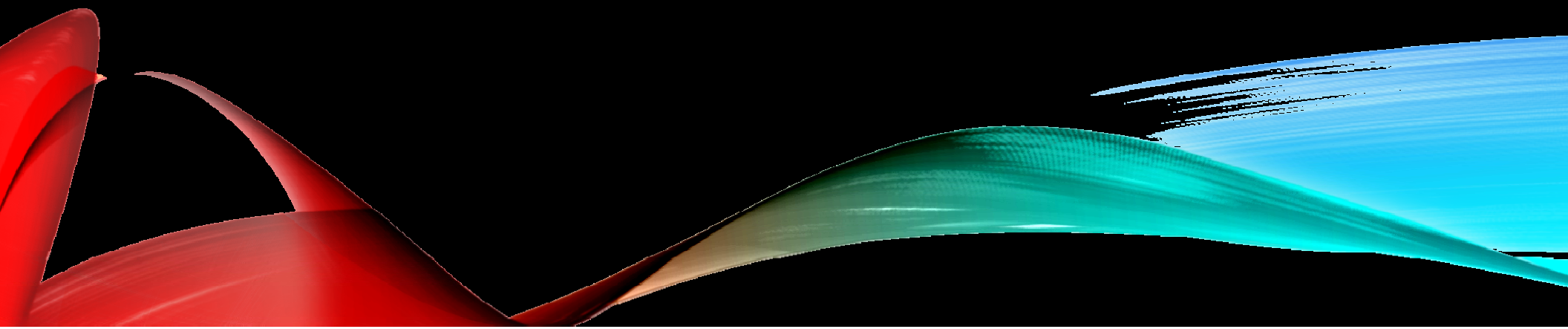
- * Vec Quadratic

[5.59915942441948 -0.0891152585838446,
-0.0891152585838446 0.00142470437384244]

[1] || *height*

* 251.471592144628

HOW TO UNDERSTAND THE MANAGERIAL IMPLICATIONS, OR THE OPERATIONAL VALIDITY



We developed the explanatory model to help us "explain" and understand the relationship between height and weight. What did we learn?

- There is a positive linear relationship between height and weight
- Our model explains about 50% of the variability in the data – maybe there are other factors that might help us explain our data
- The height coefficient (3.7) tells us for every additional inch in height, there is on average, an additional 3.7 pounds in weight

- For a person that is 72 inches tall the 95% mean confidence interval is (128, 153) – we think about whether this variability is acceptable for providing understanding about the height – weight association.