Linear Regression Analysis

CH 6

Bivariate Data

- Most statistical studies involve more than one variable.
- Suppose each individual in the sample provides two variable values.
- Objective is to discover a relationship (or lack thereof) between the variables.
  - Do some variables tend to vary together?
  - Do some variables explain variability in another?

New types of Variables

- A **response variable** measures or records an outcome of a study. (Also: y, dependent variable, predicted variable)

- An **explanatory variable** explains changes in the response variable. (Also: x, independent variable, predictor variable)

- Ex. From a survey, we could ask the questions:
  - Is there is a difference in gender and cell phone provider (categorical vs. categorical)
  - Height and favorite color (numerical vs. categorical)
  - Age and distance of commute (numerical vs. numerical)

Association

- Two variables measured on the same individuals are associated if: knowing the value of one of the variables tells you something about the values of the other variable that you would not know without this information.

- Overall tendencies, not absolute rules.

Examining Relationships

Considering the relationship between two quantitative variables.

- Start with a graph
- Look for an overall pattern and deviations from the pattern
- Use numerical descriptions of the data and overall pattern (if appropriate)
- Consider a mathematical model (regression)

Scatterplots

A **scatterplot** is a graph displaying the relationship between two quantitative variables measured on the same set of individuals.

If appropriate:

- response variable on y-axis
- explanatory variable on x-axis

- Each individual in the dataset appears as a point in the plot.

Problem

- In 1981, the average length of a game in Major League Baseball was 2 hours 33 minutes. Through 1,054 games in 2014, that had jumped to a whopping 3 hours and 2 minutes. (Quote from Forbes Magazine).
- Major League Baseball officials decided to intervene.

Example: MLB Length of Games

- Thus, MLB Commissioner Bud Selig created a committee to study the length of Major League Baseball games.

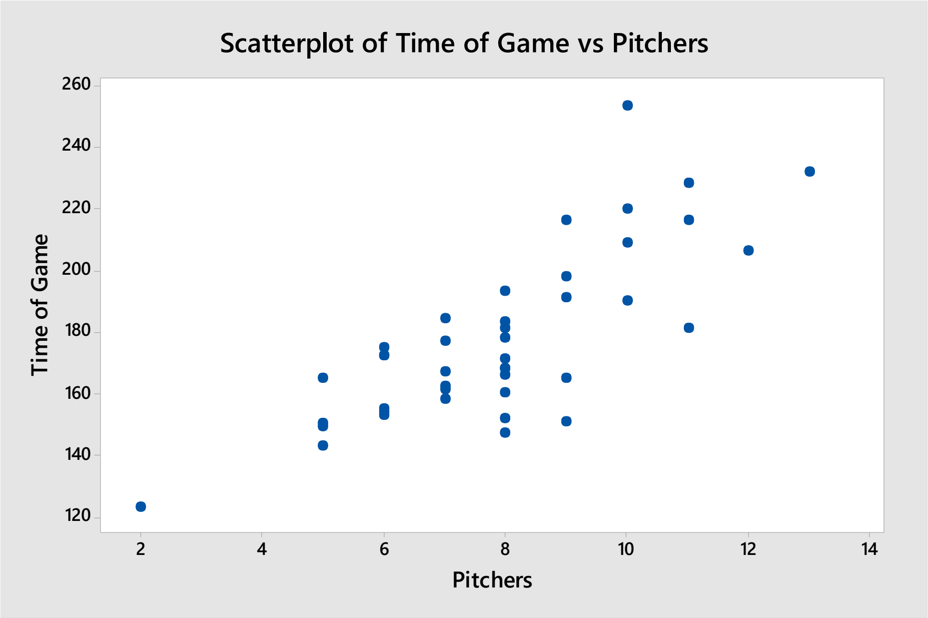
- The goals of the committee are to decrease the time of game and improve the overall pace of play in the 2015 regular season and beyond.

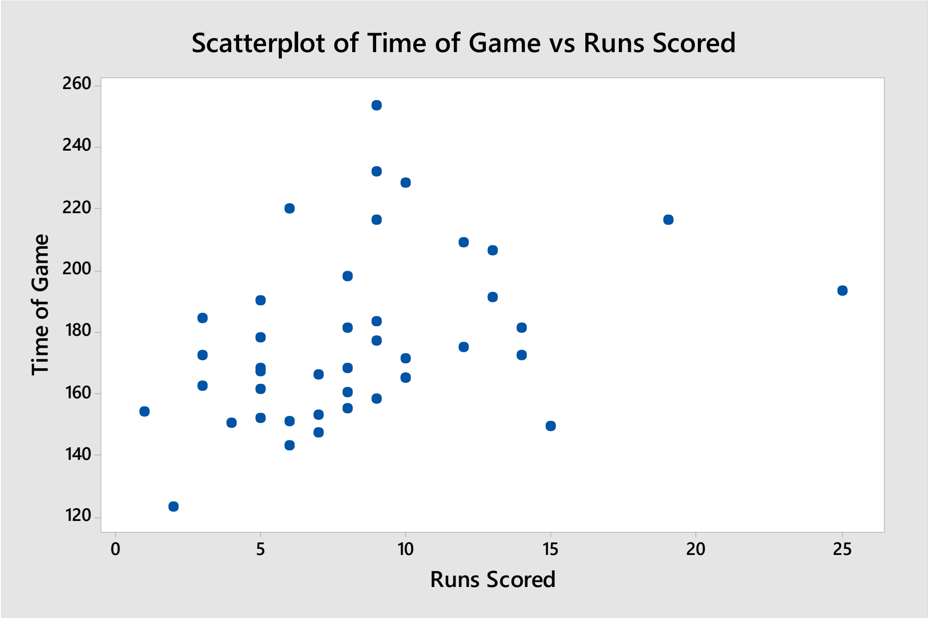
Data

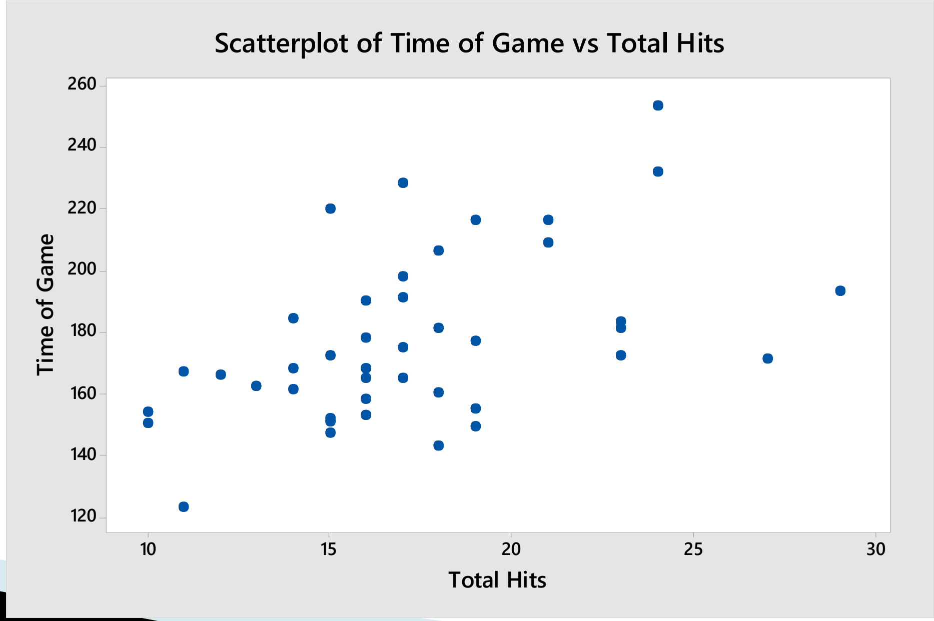
- The following data set in StatCrunch presents information on baseball games from April 24 to April 26, 2015.
- Variables include the length of the game in minutes, along with the number of runs, hits and pitchers used in the game.
- Do any of these explanatory variables have an association with the length of the game?

Simple Scatterplots in Minitab

- Graph → Scatter plot -> Simple
- Choose the explanatory variable as the x column
- Choose the response variable as the y column.
- General Title: “x-variable vs. y-variable”
- Click Ok


Scatterplot of Time of Game vs Pitchers
Time of Game
260
240
220
200
180
160
140
120
2
4
6
8
10
12
14
Pitchers


Scatterplot of Time of Game vs Runs Scored
Time of Game
260
240
220
200
180
160
140
120
0
5
10
15
20
25
Runs Scored


Scatterplot of Time of Game vs Total Hits
Time of Game
260
240
220
200
180
160
140
120
10
15
20
25
30
Total Hits

Interpreting Scatterplots

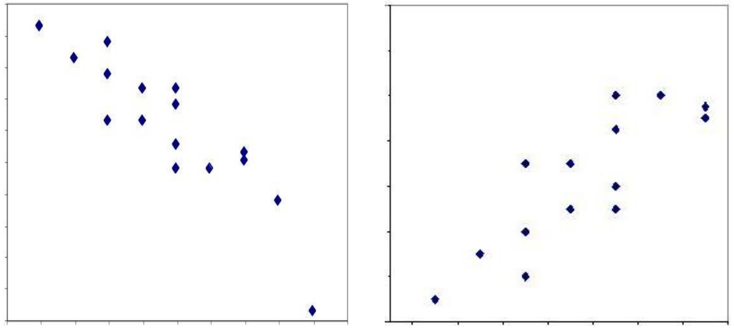
- We describe the relationship between the two variables by examining the shape (or form), trend (or direction), and strength of the association.

- We look for the overall pattern…
  - Shape: linear, curved, clusters, no pattern
  - Trend: positive, negative, no direction
  - Strength: how closely the points fit the "shape"

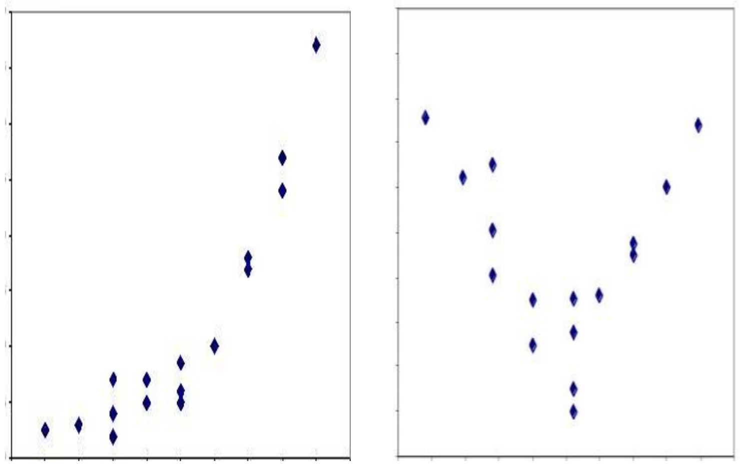
- Also, we will look for deviations from the pattern later (outliers).

Shape Examples

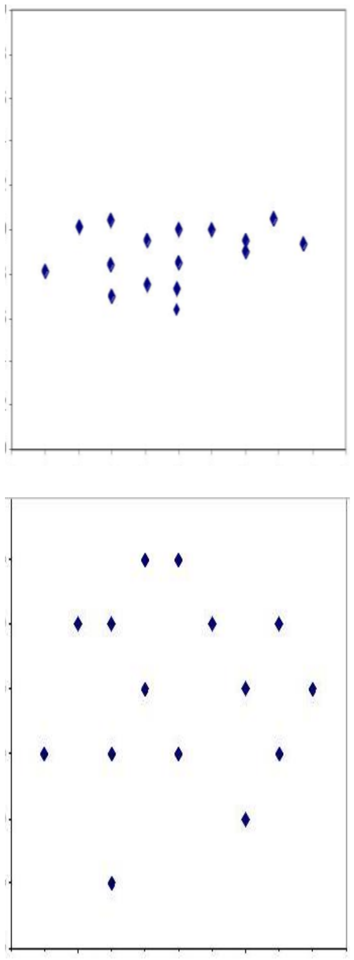
Linear



Nonlinear



No relationship



Notes on Shape

- ALWAYS look at Shape first

- For our purposes we will ONLY be able to work with Linear shapes.

Trend

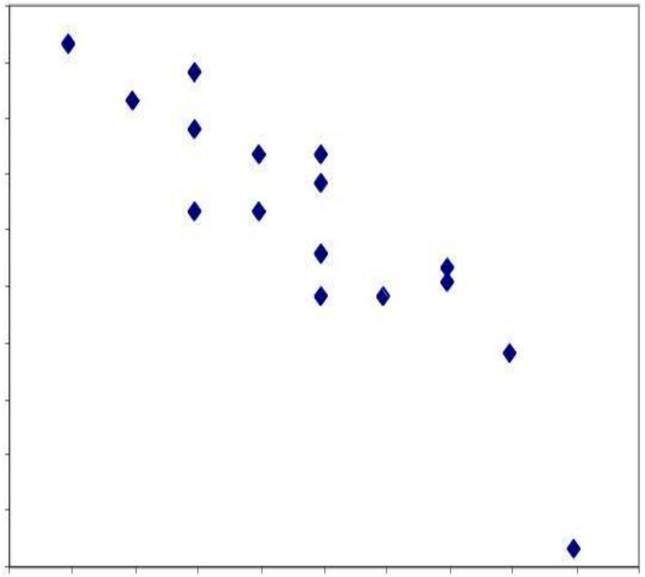
- The general tendency of the scatterplot as you read from left to right

- Typical trends:

1. Increasing (uphill), called a ***positive*** association
2. Decreasing (downhill), called a ***negative*** association
3. No trend, if there is neither an uphill nor downhill tendency

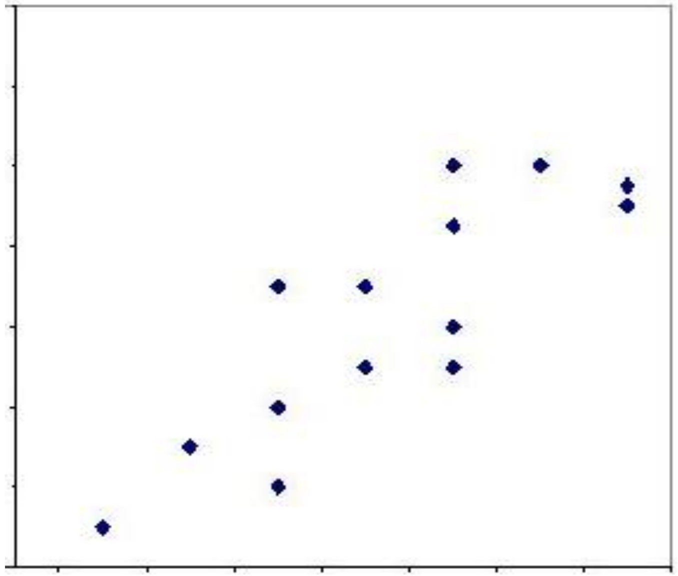
Trend Examples

Negative



high x ↔ low y
low x ↔ high y

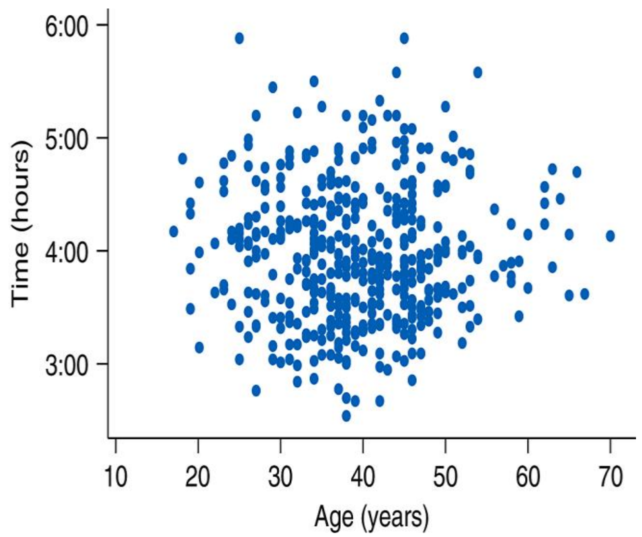
Positive



high x ↔ high y
low x ↔ low y

Marathon times

The following plot shows the relationship between martathon runners' age vs. time:



This scatterplot shows no trend because the points seem to follow no predictable pattern. This means that for every age group we can find relatively fast and relative slow runners. Marathon running speed does not seem to be related to age of runner.

Strength of an Association

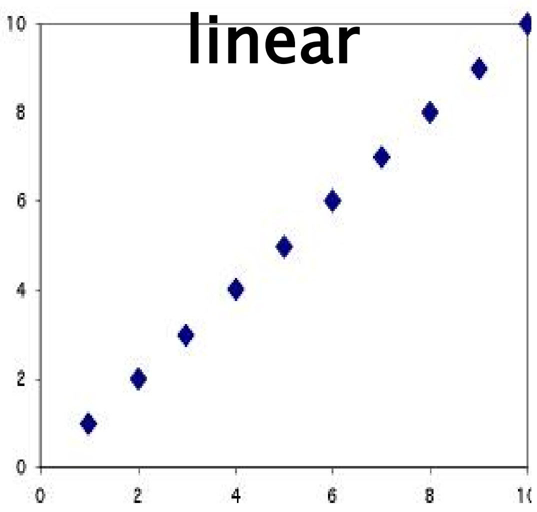
- Scatterplots with large amounts of scatter or vertical variation indicate a *weak* association.

- Scatterplots with small amounts of scatter or little vertical variation indicate a *strong* association.

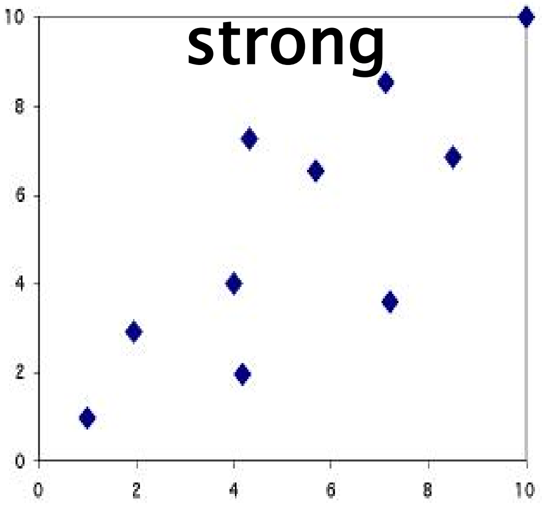
Strength Examples

A stronger relationship has points falling more closely to a clear from

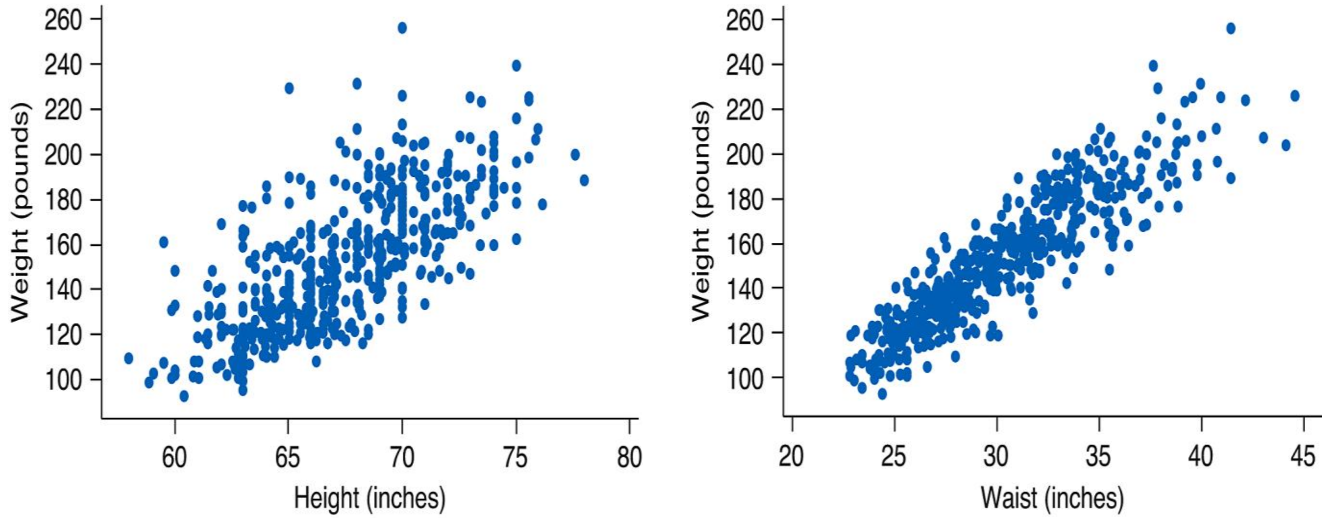
Perfect linear



Less strong



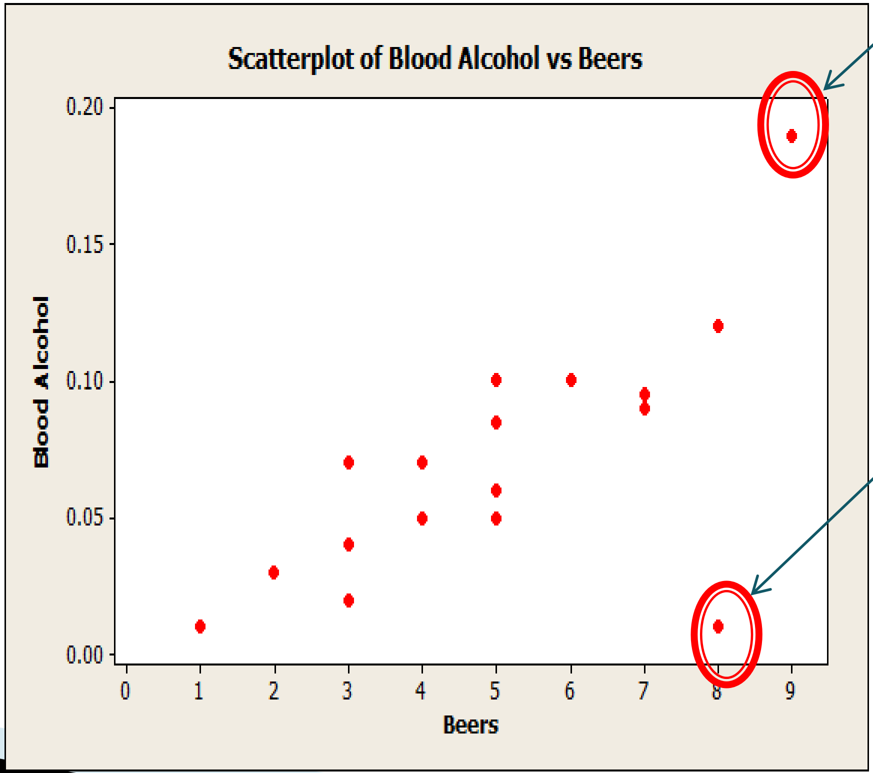
Example: Strength of Association



Is there a stronger association between height and weight or between waist size and weight?

Outliers

- An outlier in two-variable analysis is a point that falls outside the overall pattern of the relationship.



Outlier in x and y? Not a relationship outlier.

Outlier from relationship

- More on this later

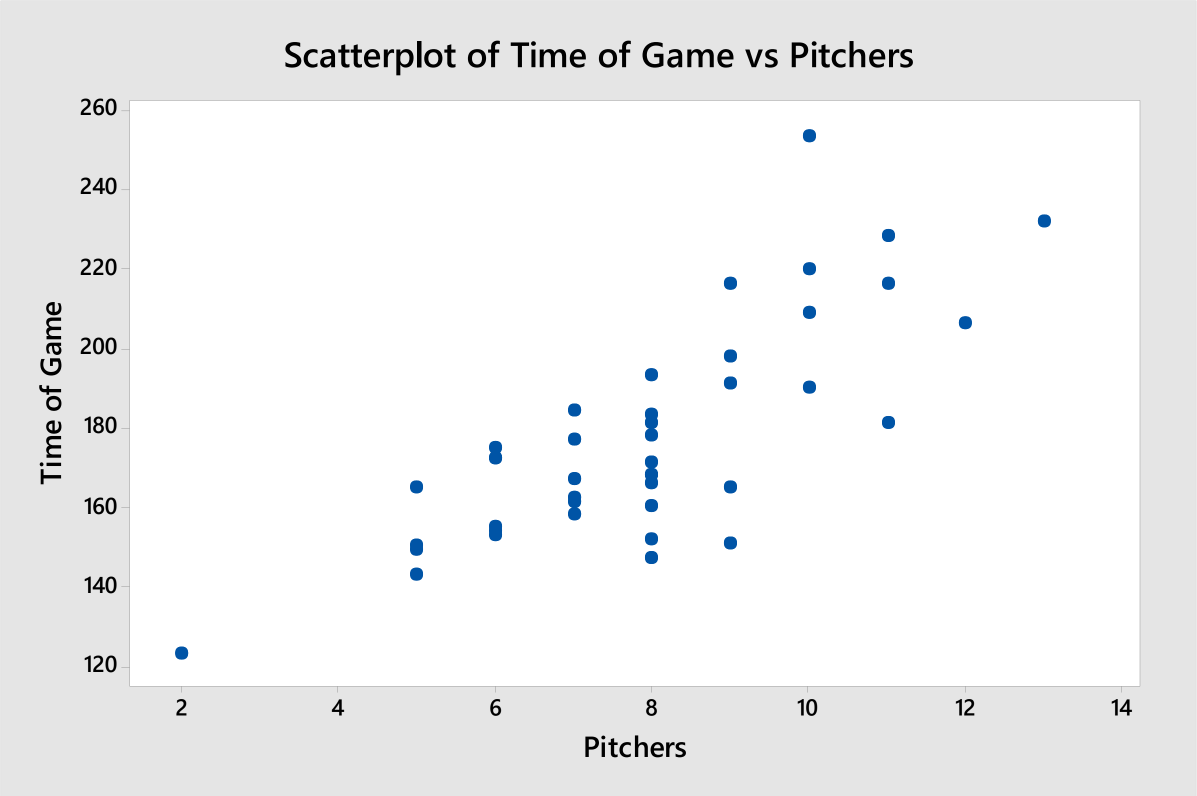
Writing Descriptions of Associations

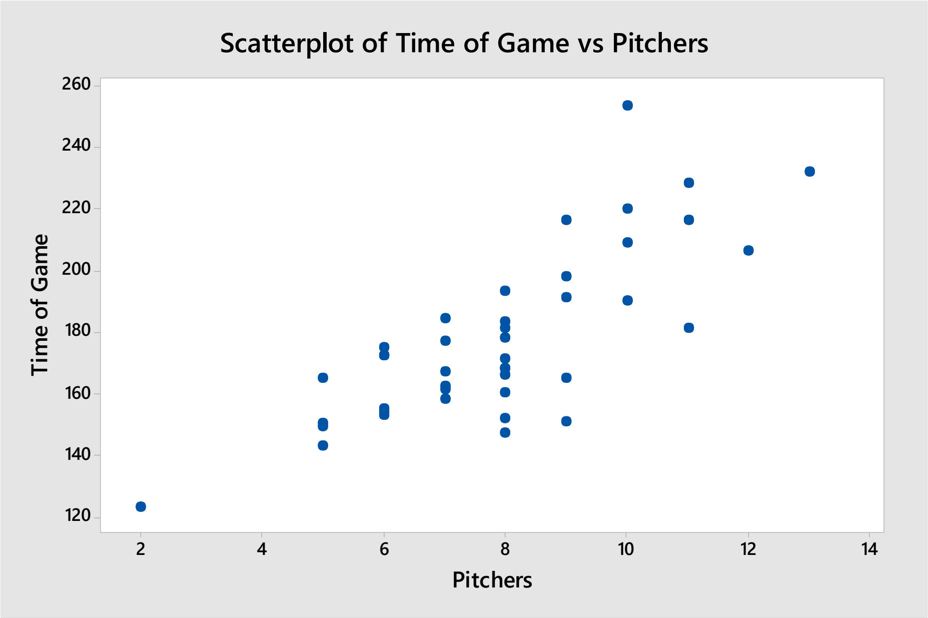
- When writing a description of an association between two numerical variables, always include:

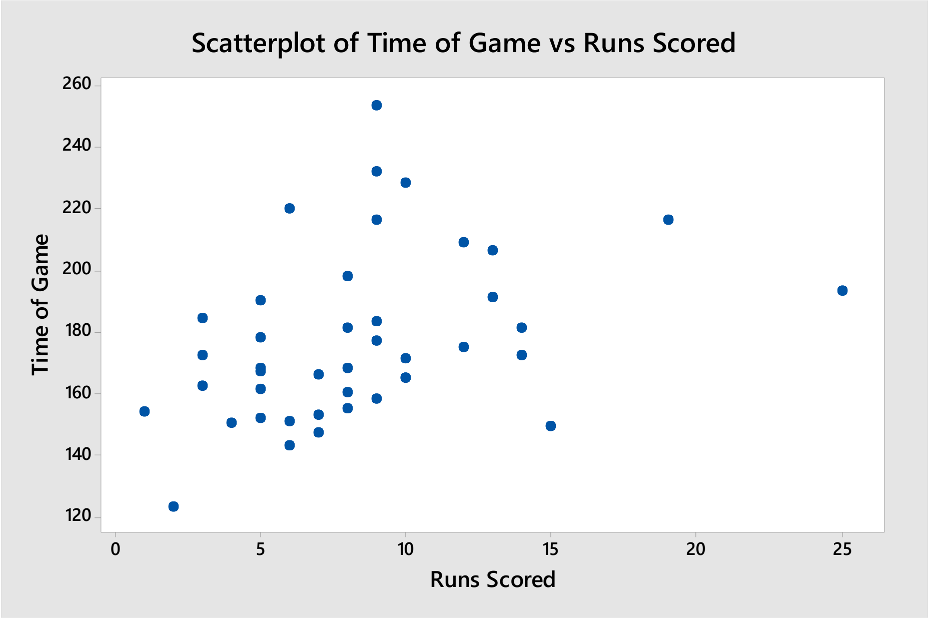
1. Trend
2. Shape
3. Strength

- In addition, mention any observations that don't fit the general trend (if any).

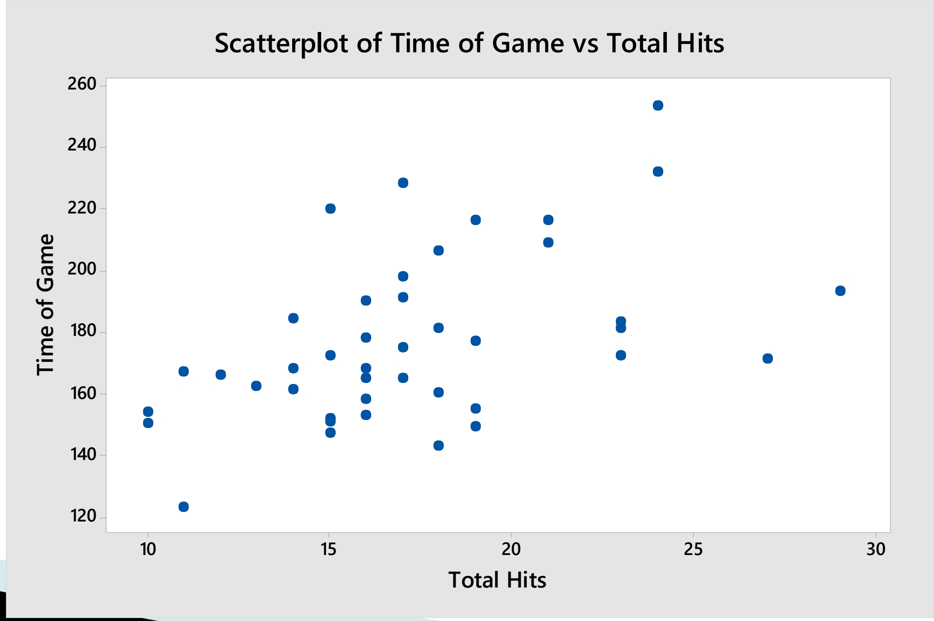
Scatterplot interpretation







Best Predictor?



Sometimes hard to say visually

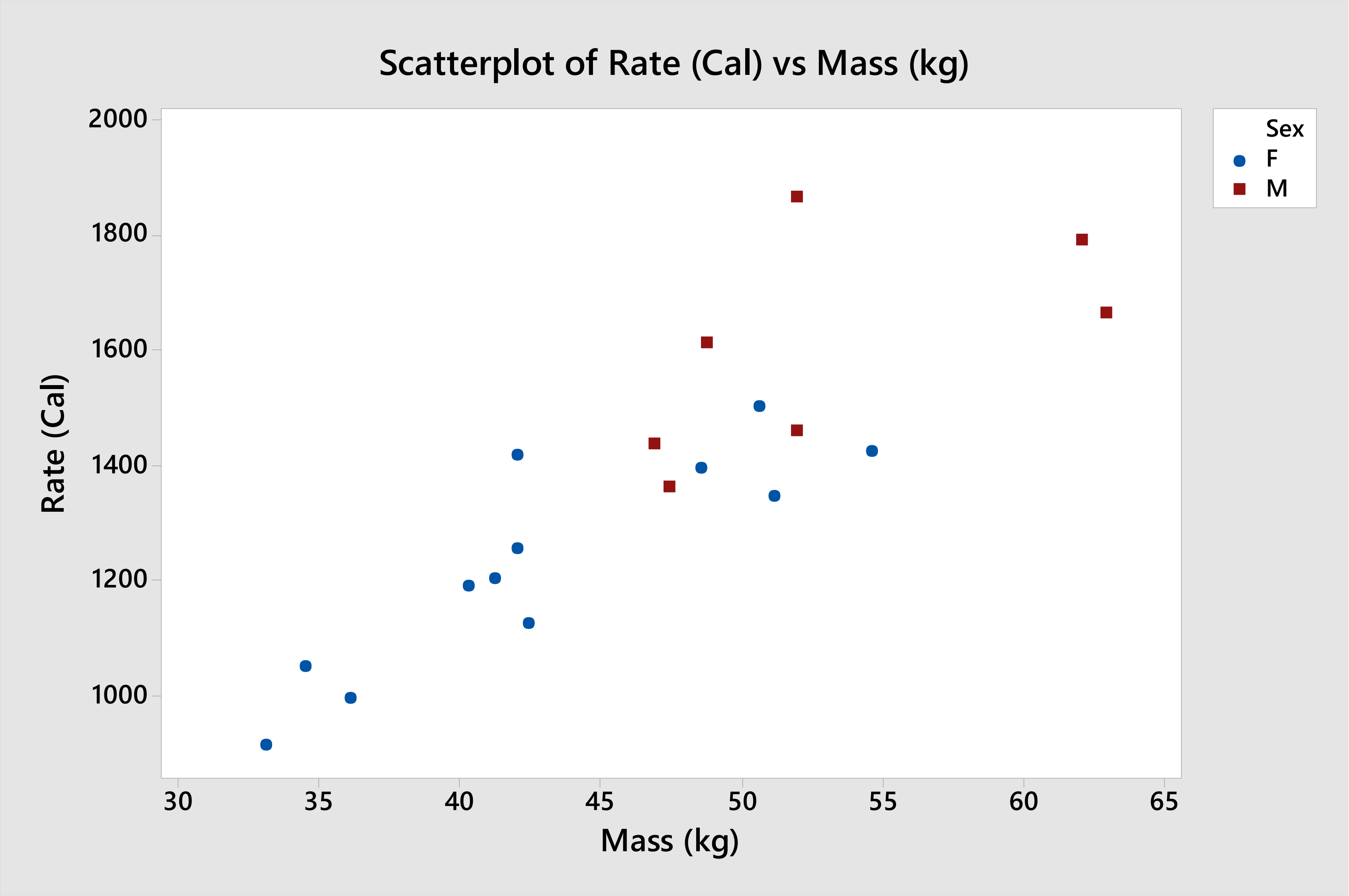
Categorical variables in Scatterplots

- Consider the data called ‘Calories’
- We have data for 19 subjects on their Gender, Lean Body Mass, and Metabolic rate.
- We want to see if there is a difference in associations between genders.

Scatterplots with Groups

- Graph → Scatter plot -> With Groups
- Choose the explanatory variable as the x column (Here: Mass)
- Choose the response variable as the y column. (Here: Rate)
- Choose your categorical variable (Here: Sex)
- Click Ok

Scatterplots with Group



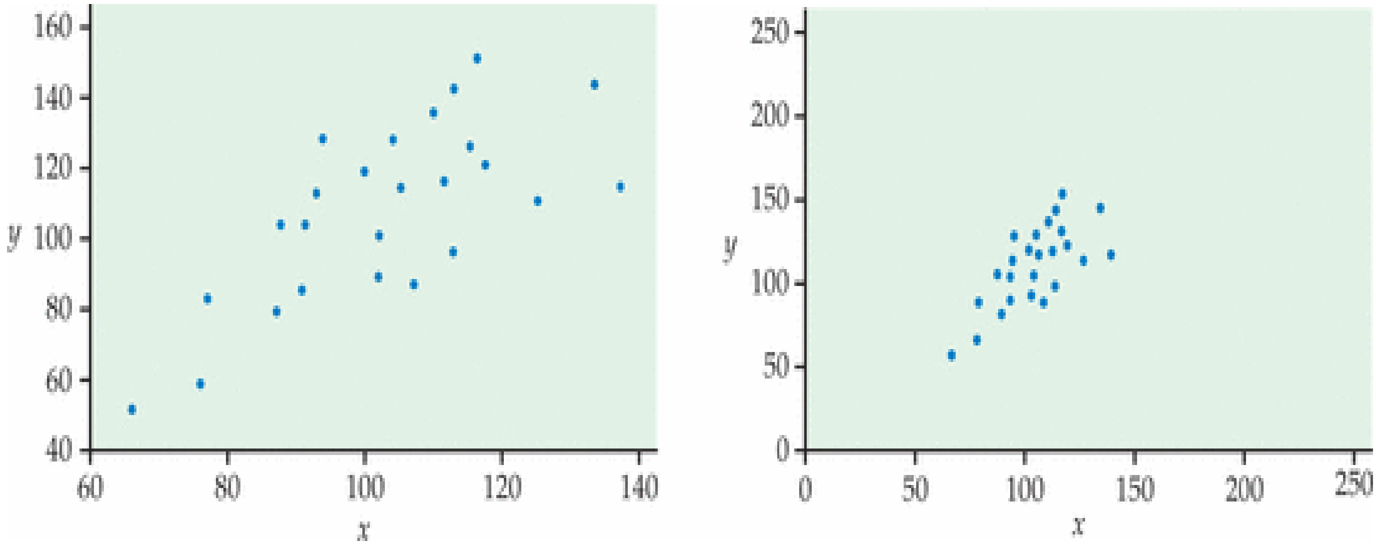
Be Careful Describing Associations

- Always use a phrase like "tends to" when describing an association because the trend you are describing has variability – the association you are describing may not be true for all individuals.

- Always point out any data points that appear to be unusual or not part of the general pattern.

Correlation

- Our eyes are not always good judges of how strong a relationship is.



- These graphs depict exactly the same data, however the right graph used a larger scale on the x and y axes.

Correlation

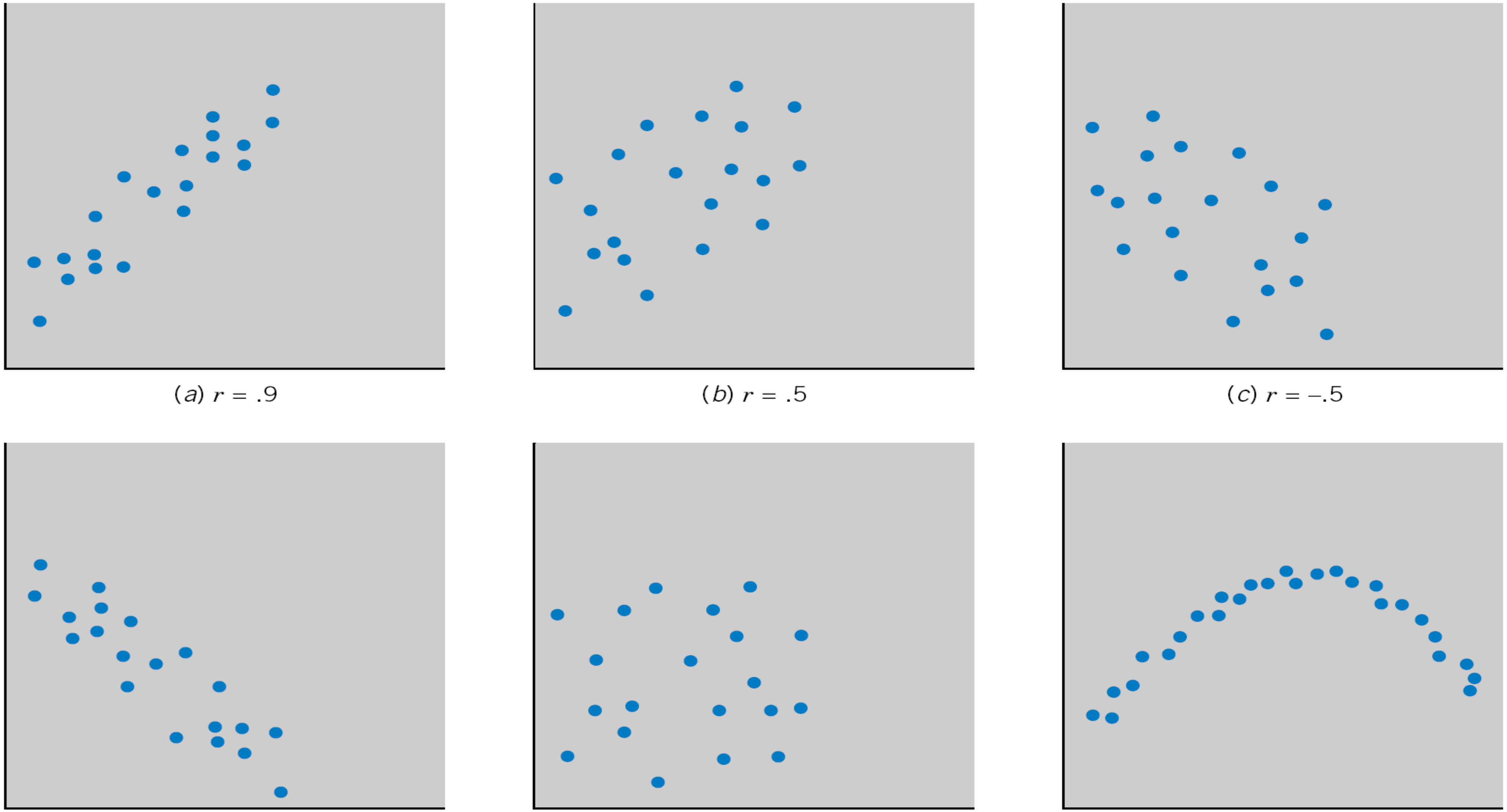
- The correlation measures the direction and strength of a linear relationship between two numerical variables.
- The symbol for the sample correlation is r.
- Also known as the Pearson's correlation coefficient.

Correlation Formula

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

- s_x and y_x are the sample standard deviations for the x and y columns of data, respectively.
- Note: you are actually finding the z-scores for each ordered pair (x, y) and multiplying.

Examples



(a) $r = .9$ (b) $r = .5$ (c) $r = -.5$ (d) $r = -.9$ (e) $r = 0$ (f) $r = 0$

Correlation Properties

- Correlation always satisfies $-1 \le r \le 1$.
  - +1 means perfect positive correlation
  - 0 means no correlation or no linear relationship (could have curved relationship).
  - −1 means perfect negative correlation.
- Response and explanatory variables are interchangeable
- Unitless and not resistant to outliers.

Correlation: Baseball Game Times

- Stat → Basic Statistics → Correlation
- Double click both variables you are using into the box to the right
- Use Default Method
- Unlick box for 'Display p-value'
- Click Ok

Correlation: Time of Game, Pitchers

Correlations

Pearson correlation	0.779

Correlation: Time of Game, Runs Scored

Correlations

Pearson correlation	0.373

Correlation: Time of Game, Total Hits

Correlations

Pearson correlation	0.489

Coefficient of Determination: r^2

- **r^2, the coefficient of determination,** is the square of the correlation coefficient.

- r^2 represents **the proportion of the variability in y** (vertical scatter from the regression line) **that can be explained by changes in x (or explained by the linear relationship).**

- For our example using pitchers as the explanatory Variable:

$$r = 0.779, \text{ thus } r^2 = 0.779^2 = 0.606841$$

Coefficient of Determination: r^2

- Usually converted to a percentage, thus always between 0% and 100%
- Measures how much variation in the response variable is explained by the explanatory variable
- The larger r^2, the smaller the amount of variation or scatter about the regression line.

Modeling Linear Trends

- A regression line is a straight line that describes how a response variable y changes as an explanatory variable x changes.

- We often use a regression line to predict the value of the response variable y for a given value of x.

- The distinction between the explanatory and response variables is necessary.

Equation of the Regression Line

- The least-squares regression equation is

$$y = b_0 + b_1 x$$

  - y is the predicted response for any value x
  - b_0 is the y-intercept
  - b_1 is the slope

Slope and y-intercept calculation

The **slope of the line,** b_1 has the formula…

$$b_1 = r\frac{s_y}{s_x}$$

r is the correlation.
s_y is the standard deviation of the response variable *y.*
s_x is the the standard deviation of the explanatory variable *x.*

Once we know b_1, the slope, calculate b_0 , **the *y*–intercept:**

$$b_0 = \bar{y} - b_1\bar{x}$$

where $\bar{x}$ and $\bar{y}$ are the sample means of the *x* and *y* variables

The computation of the coefficients should be left up to Minitab.

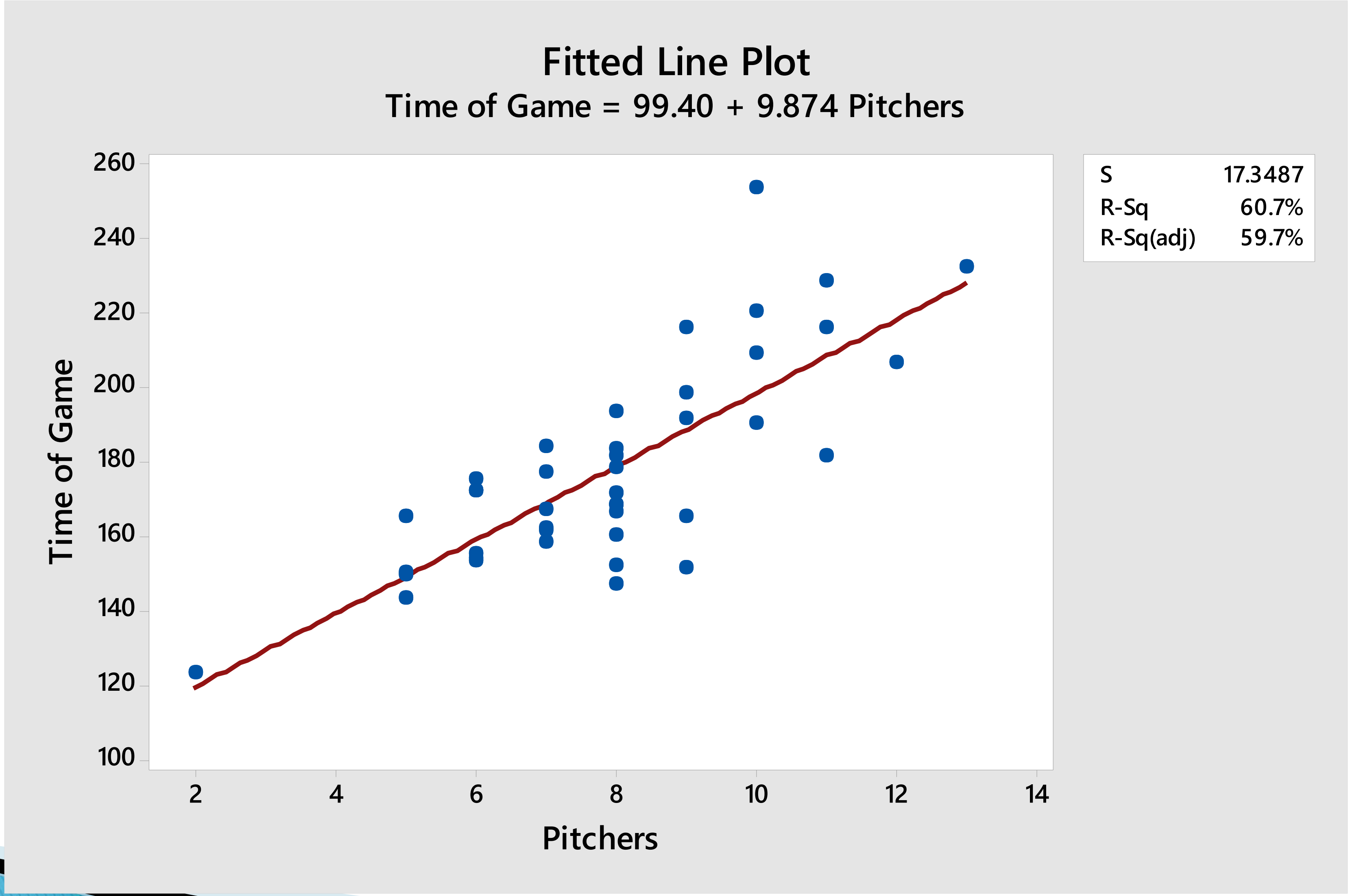
Example: Baseball Game Times

Stat → Regression -> Fitted Line plot

Correctly choose the response, y and explanatory, x variables.

Linear is already selected by default.

Minitab Output



Example: Baseball Game Times More Info

Stat → Regression -> Regression -> Fit Regression Model

Choose the response, y and predictors, x variables.

Click the results button to choose your desired output

This can give us more detailed numerical information on the relationship

We will explore this output in detail in the future

Example: Baseball Game Times More Info

Regression Analysis: Time of Game versus Pitchers

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	18564	18564.0	61.68	0.000
Pitchers	1	18564	18564.0	61.68	0.000
Error	40	12039	301.0		
Lack-of-Fit	8	2838	354.8	1.23	0.312
Pure Error	32	9201	287.5		
Total	41	30603			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
17.3487	60.66%	59.68%	57.13%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	99.4	10.3	9.68	0.000	
Pitchers	9.87	1.26	7.85	0.000	1.00

Regression Equation

Time of Game = 99.4 + 9.87 Pitchers

Fits and Diagnostics for Unusual Observations

Obs	Time of Game	Fit	Resid	Std Resid	
13	123.00	119.15	3.85	0.25	X
15	151.00	188.26	-37.26	-2.18	R
23	232.00	227.76	4.24	0.27	X
37	253.00	198.14	54.86	3.24	R

R Large residual
X Unusual X

Interpretation: Slope

- The slope $b_1 = 9.874$. This says for every unit increase (pitcher used) the average game time increases by 9.874 minutes.

- In general, for every unit change in x, y changes on average by the slope b_1.

- This phrasing does not always make sense. Always consider the context.

Interpretation: y-intercept

- The y-intercept $b_0 = 99.397$. Theoretically, the average game length when 0 pitchers are used would be 99.397 minutes.

- Only meaningful when the straight line pattern intersects zero.

- Note: you cannot conclude anything from the size of these coefficients.

Predicting y for a value of x

- Predict the time when 8 total pitchers were used.
  - 99.397149 + 9.8740778 (8) = 178.39

- Predict the time when 20 total pitchers were used.
  - 99.397149 + 9.8740778 (20) = 296.88

Predicting in Minitab

- After you have ran the regression model you can use minitab to predict y values for given x values as well
- Stat -> Regression -> Regression – > Predict

Prediction for Time of Game

Regression Equation

Time of Game = 99.4 + 9.87 Pitchers

Settings

Variable	Setting
Pitchers	8

Prediction

Fit	SE Fit	95% CI	95% PI
178.390	2.68114	(172.971, 183.809)	(142.910, 213.869)

Settings

Variable	Setting
Pitchers	20

Prediction

Fit	SE Fit	95% CI	95% PI	
296.879	15.4703	(265.612, 328.145)	(249.900, 343.858)	XX

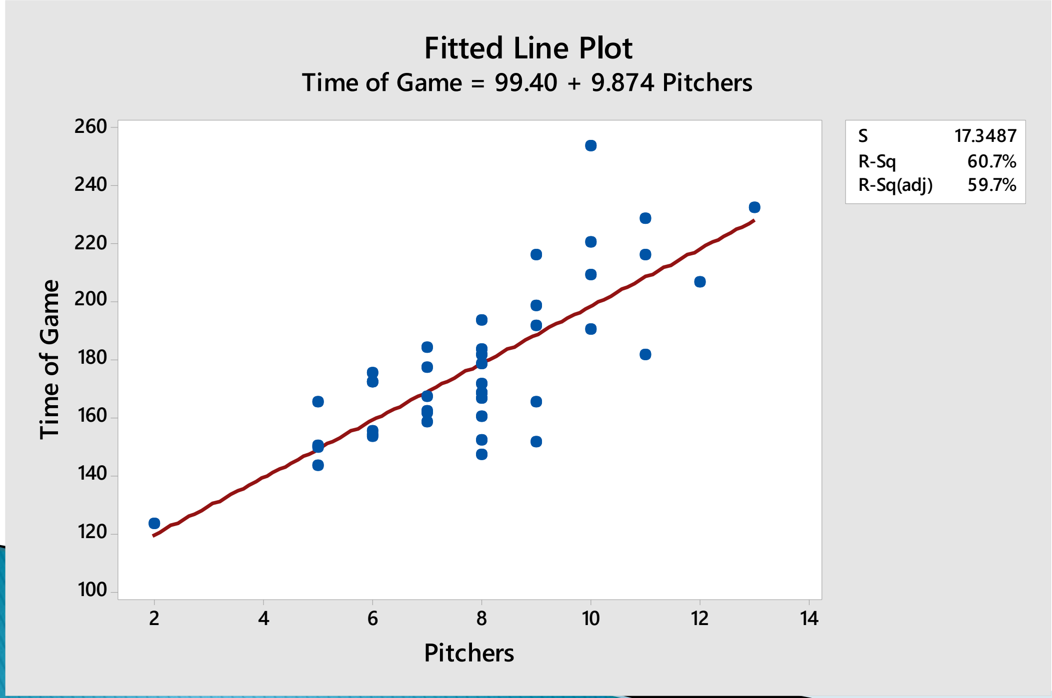
XX denotes an extremely unusual point relative to predictor levels used to fit the model.

Cautionary Notes Regarding Regression

- Do not use linear models to describe non-linear associations.
- Don't extrapolate!
- Beware of influential points that can have a big effect on *r*.
- Correlation is not causation!

Extrapolation

- We are not sure that the linear trend will continue beyond the range of the data, so these predictions may not be accurate
- Often the y-intercept is extrapolation



Prediction for Time of Game

Regression Equation

Time of Game = 99.4 + 9.87 Pitchers

Settings

Variable	Setting
Pitchers	8

Prediction

Fit	SE Fit	95% CI	95% PI
178.390	2.68114	(172.971, 183.809)	(142.910, 213.869)

Settings

Variable	Setting
Pitchers	20

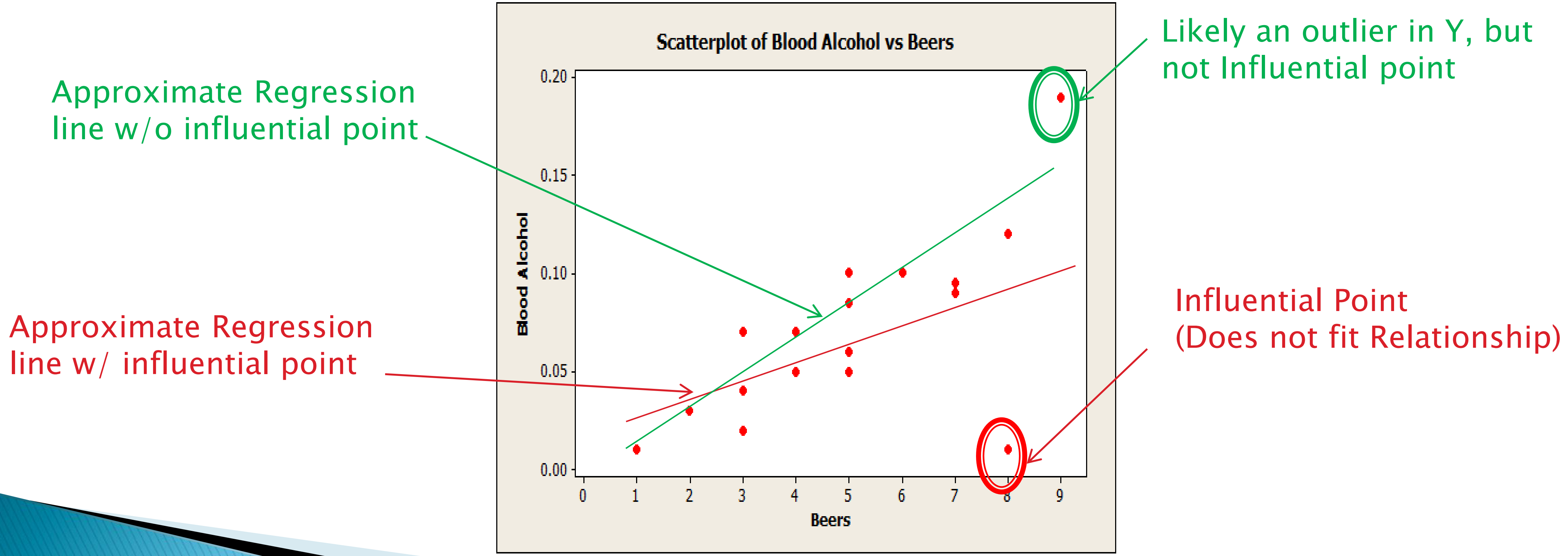
Prediction

Fit	SE Fit	95% CI	95% PI	
296.879	15.4703	(265.612, 328.145)	(249.900, 343.858)	XX

XX denotes an extremely unusual point relative to predictor levels used to fit the model.

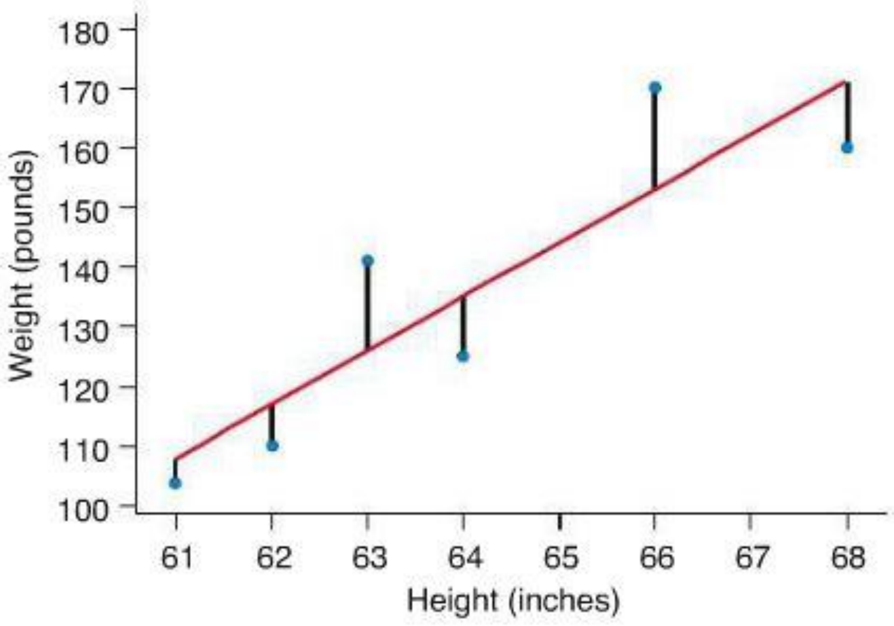
Influential Observations

- An <u>Influential Observation</u> is an observation whose deletion would drastically change the regression line.



Residuals

- Remember we use the line to predict *y* from x.
- Error = observed y – predicted y
- Also called the <u>residual</u>
- The least-squares regression line of y on x is the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible.



Association versus Causation

- An association between x and y, even if it is very strong, is not by itself good evidence that changes in x actually cause changes in y.

- Our outcomes could be influenced by a confounding (lurking) variable

- An experiment that controls confounding variables is best for establishing causation.

Facts about Regression

- If we reverse the roles of the explanatory and response variables, we will get a different regression line

- The slope, b is related to the correlation coefficient, *r.*

- The least-squares line passes through the means of the x and y variables.

Inference for Regression

- Our scatterplot, regression equation, and parameters were constructed from a sample and are used to estimate the actual model. We can use inference ideas with this sample data to generalize about the population

- Inference for regression
  - Thinking about the regression parameters
  - Checking the conditions for inference
  - Testing the hypothesis of no linear relationship
    - Testing for lack of correlation
  - Confidence intervals for the regression slope β
  - Inference about prediction

The regression parameters

- We are using *Least Squares estimation* methods to give us the line:

$$y = b_0 + b_1 x$$

- Remember, this sample is one of many. Therefore these parameter estimates have their own sampling distributions

- At the population level, the model becomes:

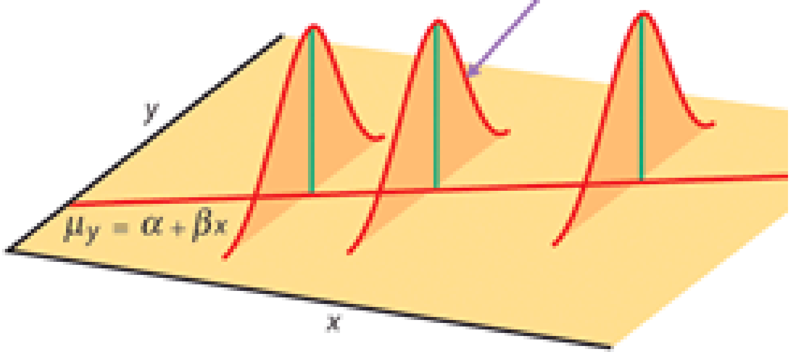
$$y_i = (\beta_0 + \beta_1 x_i) + (\varepsilon_i)$$

w/residuals e_i independent and Normally distributed $N(0,\sigma)$.

The regression parameters

- r is an unbiased estimate for the population correlation, ρ
- $\hat{y}$ is an unbiased estimate for mean response, μ_y
- b_0 is an unbiased estimate for the Y-intercept, β_0
- b_1 is an unbiased estimate for slope, β_1

For any fixed x, the responses y follow a Normal distribution with standard deviation σ.



➔Regression assumes equal variance of Y (σ is the same for all values of x).

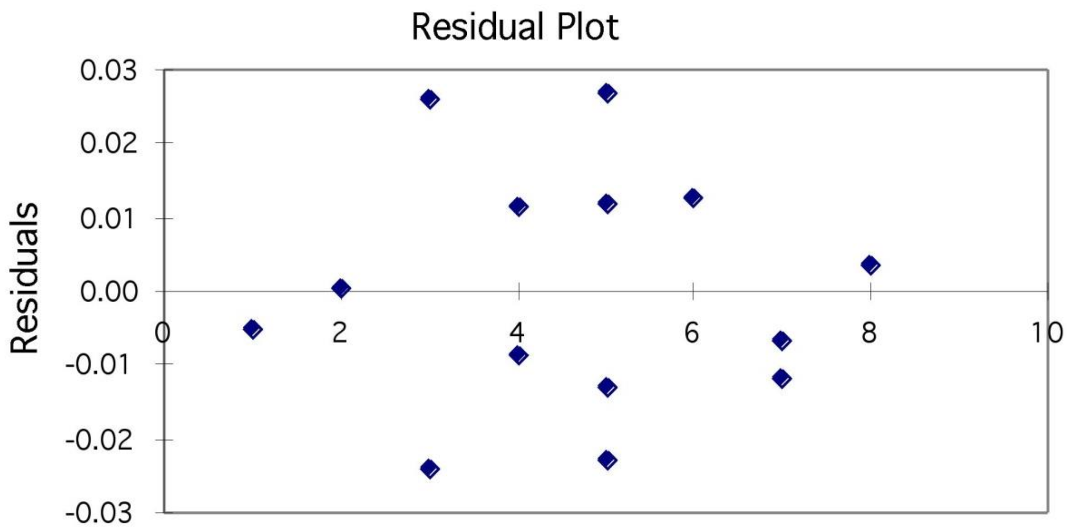
Conditions for Inference

- The observations are independent
  - Good sampling techniques
- The relationship is linear
  - Scatterplot
- The standard deviation of y, σ, is the same for all values of x
  - Residual plots
- The response y varies Normally around its mean
  - Normal plot/histogram

Residual Plots

- The residuals ($y - \hat{y}$) give useful information about the contribution of individual data points to the overall pattern of scatter.

- If residuals appear to be scattered randomly around 0 with uniform variation, it indicates that the data fit a linear model, have Normally distributed residuals for each value of *x*, and have constant standard deviation

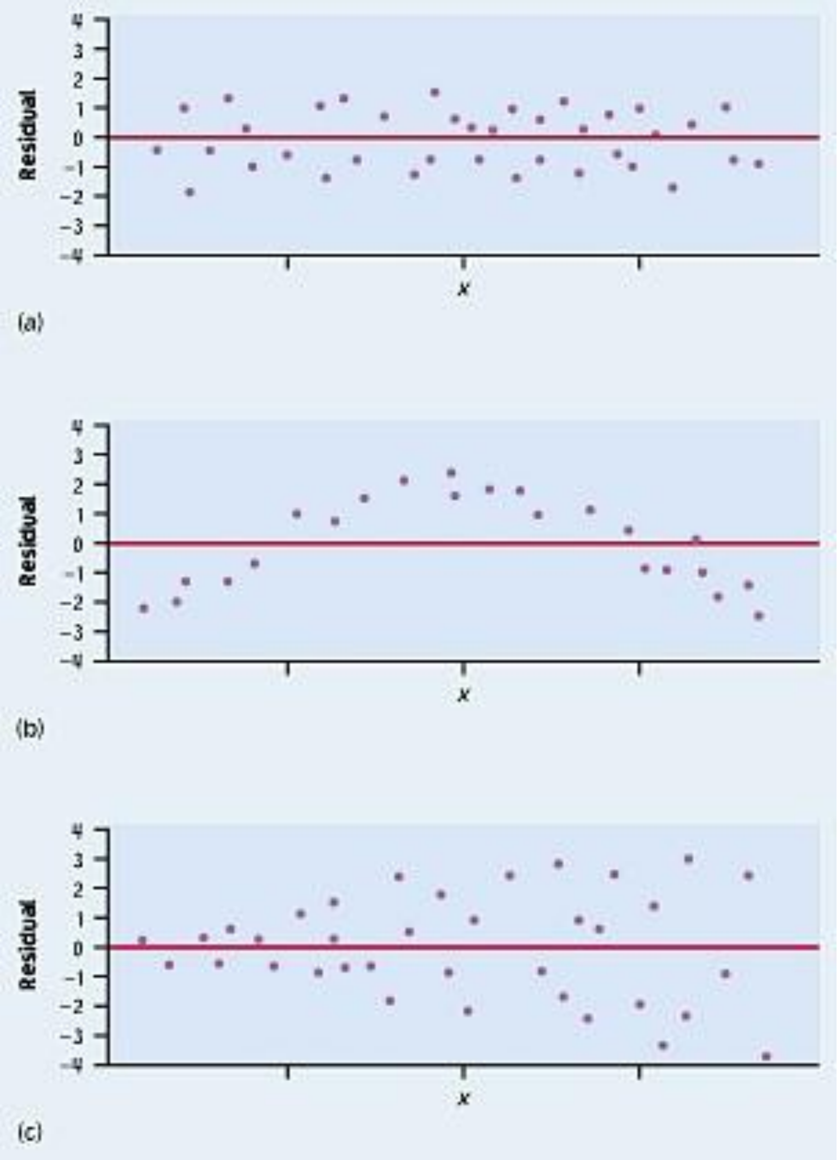


Residual Plots cont…

Residuals are randomly scattered
→ good!

Curved pattern
→ the relationship is **not linear.**

Change in variability across plot
→ ***σ* not equal** for all values of *x*.



More Regression output

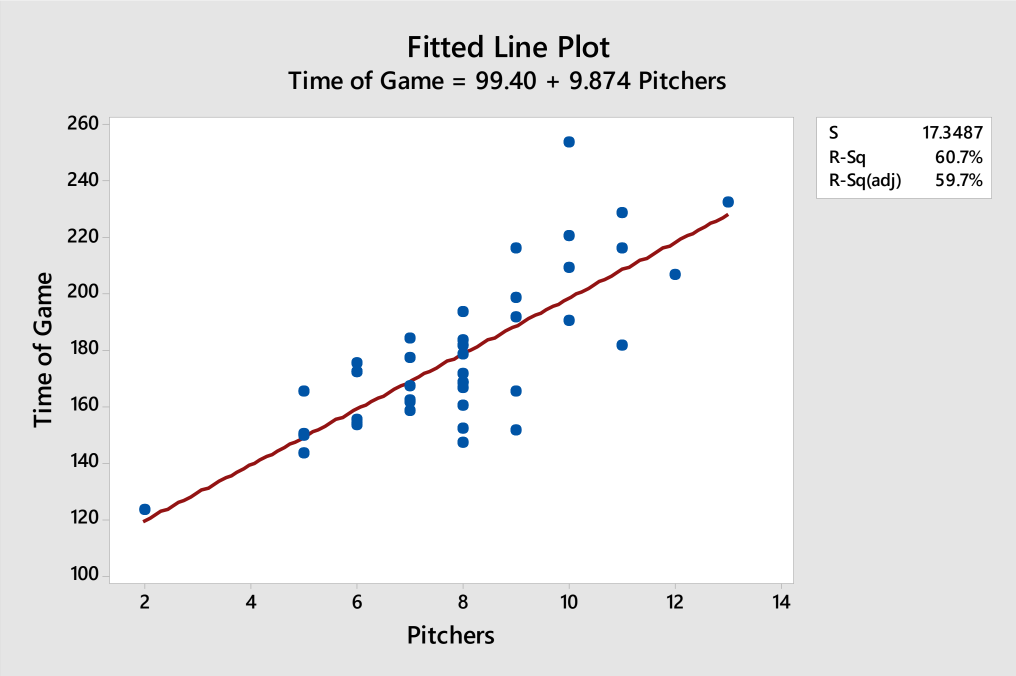
- To truly perform between these variables we need to check:
  - We can assume data are a random sample.
  - Check to see if relationship is linear. (Scatterplot)
  - Residuals look to be Normally distributed. (Normal plot/histogram)
  - No apparent patterns in the variance of residuals (Residual plots)

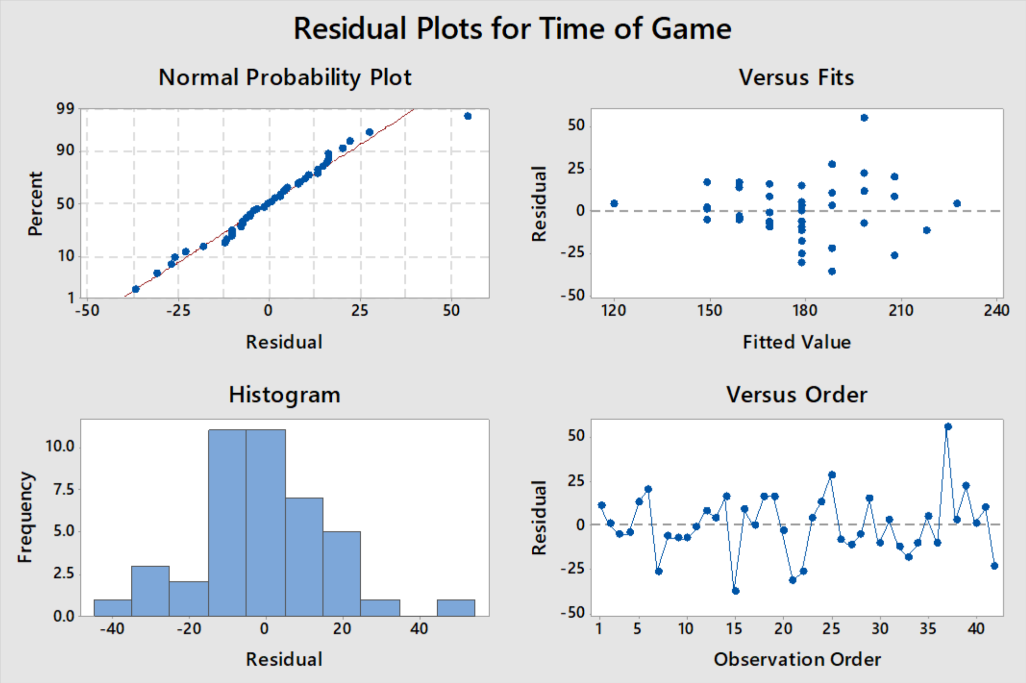
Regression in Minitab

- Stat → Regression -> Fitted Line plot (seen previously to do initial examination)

- Stat → Regression -> Regression -> Fit Regression Model
  - Choose the response, y and predictors, x variables.
  - Click the results button to choose your desired numerical output (default is fine)
  - Click the results button to choose your desired numerical output (four in one gives all the info we need)

Minitab Regression Plots





- The data are a random sample.
- The relationship is clearly linear.
- The residuals are roughly Normally distributed.
- The spread of the residuals around 0 is fairly homogenous along all values of *x*.

Minitab Regression output

Regression Analysis: Time of Game versus Pitchers

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	18564	18564.0	61.68	0.000
Pitchers	1	18564	18564.0	61.68	0.000
Error	40	12039	301.0		
Lack-of-Fit	8	2838	354.8	1.23	0.312
Pure Error	32	9201	287.5		
Total	41	30603			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
17.3487	60.66%	59.68%	57.13%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	99.4	10.3	9.68	0.000	
Pitchers	9.87	1.26	7.85	0.000	1.00

Regression Equation

Time of Game = 99.4 + 9.87 Pitchers

Fits and Diagnostics for Unusual Observations

Obs	Time of Game	Fit	Resid	Std Resid	
13	123.00	119.15	3.85	0.25	X
15	151.00	188.26	-37.26	-2.18	R
23	232.00	227.76	4.24	0.27	X
37	253.00	198.14	54.86	3.24	R

R Large residual
X Unusual X

Residual Plots for Time of Game

The regression Standard error

- The regression standard error, s, for n sample data points is computed from the residuals $(y_i - \hat{y}_i)$:

$$s = \sqrt{\frac{\sum residual^2}{n-2}} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}$$

- Notice DoF= n−2 here!

Calculating regression standard error

- We can have Minitab print out all residuals then calculate s

Fits and Diagnostics for All Observations

Obs	Time of Game	Fit	Resid	Std Resid	
1	209.00	198.14	10.86	0.64	
2	149.00	148.77	0.23	0.01	
3	153.00	158.64	-5.64	-0.33	
4	154.00	158.64	-4.64	-0.27	
5	172.00	158.64	13.36	0.79	
6	228.00	208.01	19.99	1.20	
7	152.00	178.39	-26.39	-1.54	
8	162.00	168.52	-6.52	-0.38	
9	171.00	178.39	-7.39	-0.43	
10	161.00	168.52	-7.52	-0.44	
11	167.00	168.52	-1.52	-0.09	
12	216.00	208.01	7.99	0.48	
13	123.00	119.15	3.85	0.25	X
14	184.00	168.52	15.48	0.91	
15	151.00	188.26	-37.26	-2.18	R
16	177.00	168.52	8.48	0.50	
17	178.00	178.39	-0.39	-0.02	
18	165.00	148.77	16.23	0.97	
19	175.00	158.64	16.36	0.96	
20	155.00	158.64	-3.64	-0.21	
21	147.00	178.39	-31.39	-1.83	
22	181.00	208.01	-27.01	-1.62	
23	232.00	227.76	4.24	0.27	X
24	172.00	158.64	13.36	0.79	
25	216.00	188.26	27.74	1.62	
26	190.00	198.14	-8.14	-0.48	
27	206.00	217.89	-11.89	-0.73	
28	143.00	148.77	-5.77	-0.34	
29	193.00	178.39	14.61	0.85	
30	168.00	178.39	-10.39	-0.61	
31	181.00	178.39	2.61	0.15	
32	166.00	178.39	-12.39	-0.72	
33	160.00	178.39	-18.39	-1.07	
34	158.00	168.52	-10.52	-0.61	
35	183.00	178.39	4.61	0.27	
36	168.00	178.39	-10.39	-0.61	
37	253.00	198.14	54.86	3.24	R
38	191.00	188.26	2.74	0.16	
39	220.00	198.14	21.86	1.29	
40	150.00	148.77	1.23	0.07	
41	198.00	188.26	9.74	0.57	
42	165.00	188.26	-23.26	-1.36	

R Large residual
X Unusual X

$$s = \sqrt{\frac{\sum residual^2}{n-2}} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}$$

In our example s=17.3487

Testing the significance of the slope

- To test for a significant relationship, we ask if the parameter for the slope β_1 is equal to zero, using a one-sample t test.

- We test the hypotheses H_0: $\beta_1 = 0$ vs. (typically two-sided) H_a.

- The standard error of the slope is

$$SE_{b_1} = \frac{s}{\sqrt{\sum(x - \bar{x})^2}}$$

- Thus the Test Statistic is:

$$t = \frac{b_1}{SE_{b_1}}$$ w/ **(*n* – 2) DoF.**

- From there we can find a p-value

Testing for Lack of correlation

- The regression slope b_1 and the correlation coefficient r are related and $b_1 = 0 \rightarrow r = 0$.

$$\text{slope } b_1 = r\frac{s_y}{s_x}$$

- Similarly, the population parameter for the slope β_1 is related to the population correlation coefficient ρ, and when $\beta_1 = 0 \rightarrow \rho = 0$.

- Thus, testing the hypothesis H_0: $\beta_1 = 0$ is the same as testing the hypothesis of no correlation between x and y in the population from which our data were drawn.

Confidence Interval for the slope

- We know the slope follows a t distribution w/ n–2 DoF and

$$SE_{b_1} = \frac{s}{\sqrt{\sum(x - \bar{x})^2}}$$

- Using the general form of a CI:

estimate ± *CV*SE*

$$\boldsymbol{b_1 \pm t^* SE_{b_1}}$$

HT example:

- In our example: n = 20, df = 18
- We can test:

H_0: $\beta_1 = 0$
H_a: $\beta_1 \neq 0$

$$SE_{b_1} = \frac{s}{\sqrt{\sum(x-\bar{x})^2}} = \frac{17.3487}{\sqrt{\sum(x-\bar{x})^2}} = 1.26$$

$t = b_1 / SE_{b_1}$ = 9.87/ 1.26 = 7.85 with df = n – 2 = 40

→ $P < 0.001$ (two-sided test), highly significant.

- Our CI for the slope would be, t^* = 2.021

$$b_1 \pm t * SE_{b_1} = 9.87 \pm 2.021 * 1.26 = (7.32, 12.42)$$

- Interpretation?

Predicting in Minitab

- After you have ran the regression model you can use Minitab to predict y values for given x values as well
- Stat -> Regression -> Regression – > Predict
- There are different formulas for a CI for μ_y and a Prediction Interval for $\hat{y}$. The are calculated using slightly different Standard errors.

Prediction for Time of Game

Regression Equation

Time of Game = 99.4 + 9.87 Pitchers

Settings

Variable	Setting
Pitchers	8

Prediction

Fit	SE Fit	95% CI	95% PI
178.390	2.68114	(172.971, 183.809)	(142.910, 213.869)

$$SE_{\hat{\mu}} = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x - \bar{x})^2}}$$

$$SE_{\hat{y}} = s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x - \bar{x})^2}}$$

Analysis of Variance (ANOVA) tables

- The Other part of regression output is the ANOVA table. The basic format is as follows:

Source	D.o.F	SS	MS	F
Model	DF_m	SSM	MSM	Test Statistic
Error	DF_E	SSE	MSE	–
Total	DF_T	SST	–	–

Analysis of Variance (ANOVA) tables

- To begin our calculations we need the following pieces:
  - Sum of x values (Σx)
  - Sum of y values(Σy)
  - Sum of x values squared(Σx^2)
  - Sum of y values(Σy^2)
  - Sum of the product of x and y (Σxy)

- We then find the "Sums of Squares". In General: $SS_{ab} = \Sigma a*b - (1/n)\Sigma a*\Sigma b$. So:
  - $SS_{xy} = \Sigma xy - (1/n)\Sigma x\Sigma y$
  - $SS_{xx} = \Sigma x^2 - (1/n)(\Sigma x)^2$
  - $SS_{yy} = \Sigma y^2 - (1/n)(\Sigma y)^2$

- We can use these sums of squares to first estimate the slope:

$$b_1 = \frac{SS_{xy}}{SS_{xx}}$$

- Once we have the slope, we can solve for the y intercept:

$$b_0 = \bar{y} - b_1\bar{x}$$

Analysis of Variance (ANOVA) tables

- From there we begin filling out our table. The easiest place to start is D.o.F:

Source	D.o.F	SS	MS	F
Model	DF_m	SSM	MSM	Test Statistic
Error	DF_E	SSE	MSE	–
Total	DF_T	SST	–	–

- D.o.F. for the Model Row:

$$DF_m = \# \text{ of estimated parameters } -1$$

- Next we need the total D.o.F.

$$DF_T = n-1$$

(*where n is our # of pairs in the regression context*)

- From there:

$$DF_E = DF_T - DF_m$$

Analysis of Variance (ANOVA) tables

- We keep filling out our table moving right:

Source	D.o.F	SS	MS	F
Model	DF_m	SSM	MSM	Test Statistic
Error	DF_E	SSE	MSE	–
Total	DF_T	SST	–	–

- Next step is SST:

$$SST = SS_{yy}$$

- From there:

$$SSM = b_1{}^2 * SS_{xx} = \frac{SS_{xy}{}^2}{SS_{xx}}$$

- Finally

$$SSE = SST - SSM$$

Analysis of Variance (ANOVA) tables

- Once we have our SS column calculated we then scale by the D.o.F. to find the MS

Source	D.o.F	SS	MS	F
Model	DF_m	SSM	MSM	Test Statistic
Error	DF_E	SSE	MSE	–
Total	DF_T	SST	–	–

- $MSM = SSM/DF_m$
- $MSE = SSE/DF_E$

Analysis of Variance (ANOVA) tables

- Finally we find our F Test Statistic by looking at the ratio of the MS terms

Source	D.o.F	SS	MS	F
Model	DF_m	SSM	MSM	Test Statistic
Error	DF_E	SSE	MSE	–
Total	DF_T	SST	–	–

- F Test Stat:

$$F = MSM/MSE$$

- We then go to the F table w/ both D.o.F. and find a p-val

- If we find a significant p-val here, then our model is a "good" (significant) one

Easy R^2 calculation

- A quick and easy way to get R^2 from our ANOVA table is:

$$R^2 = SSM/SST \text{ or } 1-(SSE/SST)$$

- We also may opt for an adjusted version of R^2 which accounts for the number of parameters in a model

$$R^2_{adj} = 1 - (MSE/MST)$$

where $MST = SST/DF_T$

ANOVA table Calculation Example

- Let's build an ANOVA table with the following dataset from the textbook. See p. 375 for description of data and calculations posted in Canvas

Reflux Ratio	Concentration
20	0.446
30	0.601
40	0.786
50	0.928
60	0.95

⇒

ANOVA Table					
Source	D.o.F	SS	MS	F	p-value
Model	1	0.1782225	0.178223	58.947058	0.0045906
Error	3	0.0090703	0.003023	–	–
total	4	0.1872928	–	–	–