Live Session 7

1. Welcome/Intro (including polls)
2. Quiz 2 Review – see me with questions about your quiz
3. Correlation and Regression
4. Assignments for next 2 weeks
5. Wrap up and Feedback

Analyze

Description:

Analyze, describe, and present the data to discover the root cause(s), identify/prioritize critical inputs (x's), determine the inputs impact on the output.

Key Concepts:

Inferential statistics, common distributions, developing a hypothesis, determining the likelihood some event happens based on a sample (calculating probabilities), Using the normal distribution as the "go to" distribution.

Project:

Write a null and alternative hypothesis statement.

Tools:

Hypothesis testing
Chi-square test for independence

Key Concepts:

Collecting sample data, how confidence intervals and sample size are related.

Project:

Utilize the sample size formula.

Tools:

Confidence intervals.

Key Concepts:

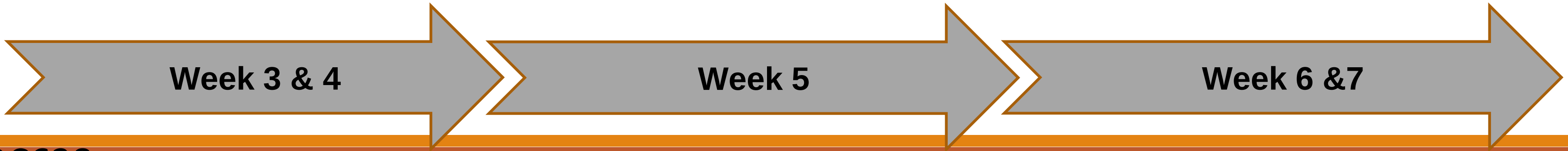
Determining input's (x) impact on the output (y).

Project:

Use regression to identify relationships between the output (y) and inputs (x's).

Tools:

Correlation
Simple linear regression
Multiple regression
Scatterplot
Trend/ line chart
Pareto chart
Fishbone (cause/effect) diagram



Week 3 & 4

Week 5

Week 6 &7

Correlation and Regression

Always remember P-G-A

- Practical – does the relationship make practical sense for my project?
- Graphical – what do I see when I graph the data?
- Analytical – apply the statistical calculations

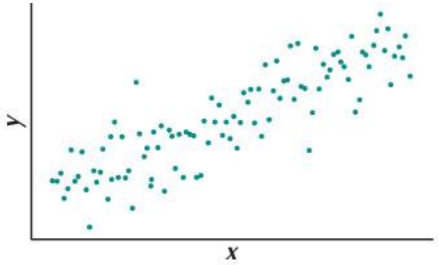
Correlation: is there a relationship between these variables?

Regression: what is the equation for this relationship?

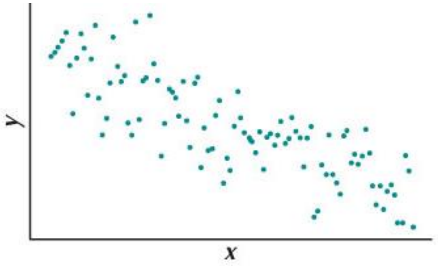
Correlation

Scatterplots

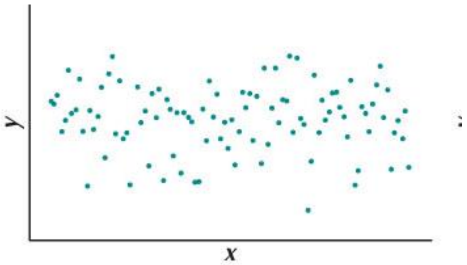
The relationship between two quantitative variables can take many different forms. Four of the most common are:



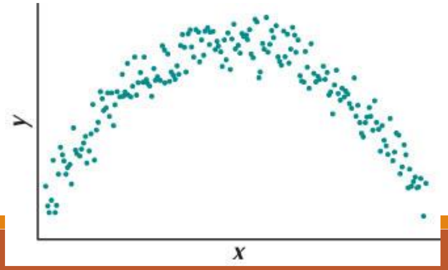
Positive linear relationship: As *x* increases, *y* also tends to increase.



Negative linear relationship: As *x* increases, *y* tends to decrease.



No apparent relationship: As *x* increases, *y* tends to remain unchanged.



Nonlinear relationship: The *x* and *y* variable are related, but not in a way that can be approximated using a straight line.

Correlation Coefficient

Scatterplots provide a visual description of the relationship between two quantitative variables. The *correlation coefficient* is a numerical measure for quantifying the linear relationship between <u>two</u> quantitative variables.

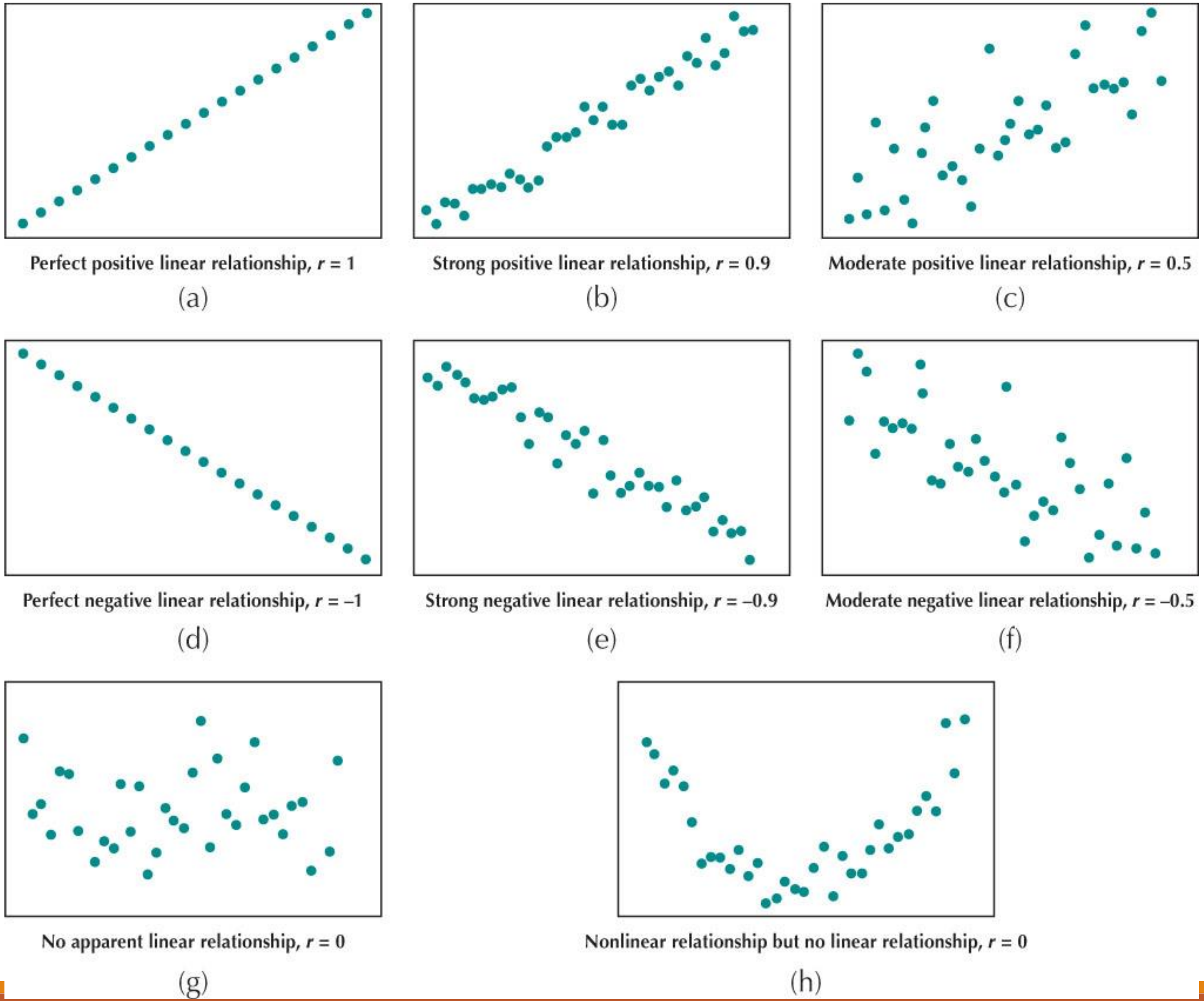
The **correlation coefficient *r*** measures the strength and direction of the linear relationship between two variables. The correlation coefficient *r* is

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{(n - 1)s_x s_y}$$

where s_x is the sample standard deviation of the *x* data values, and s_y is the sample standard deviation of the *y* data values.

<u>Guideline</u>:
r value of ~ ± 0.7 desired, indicates meaningful relationship

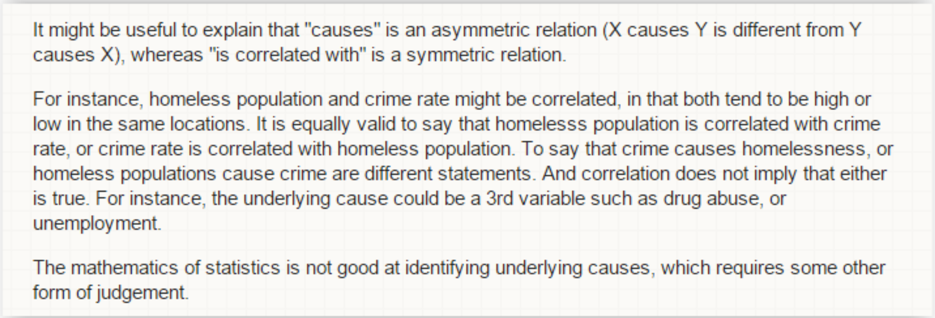
Properties of *r*



Correlation Coefficient (*r*)

Guideline:
r value of ~ ± 0.7 desired, indicates meaningful relationship

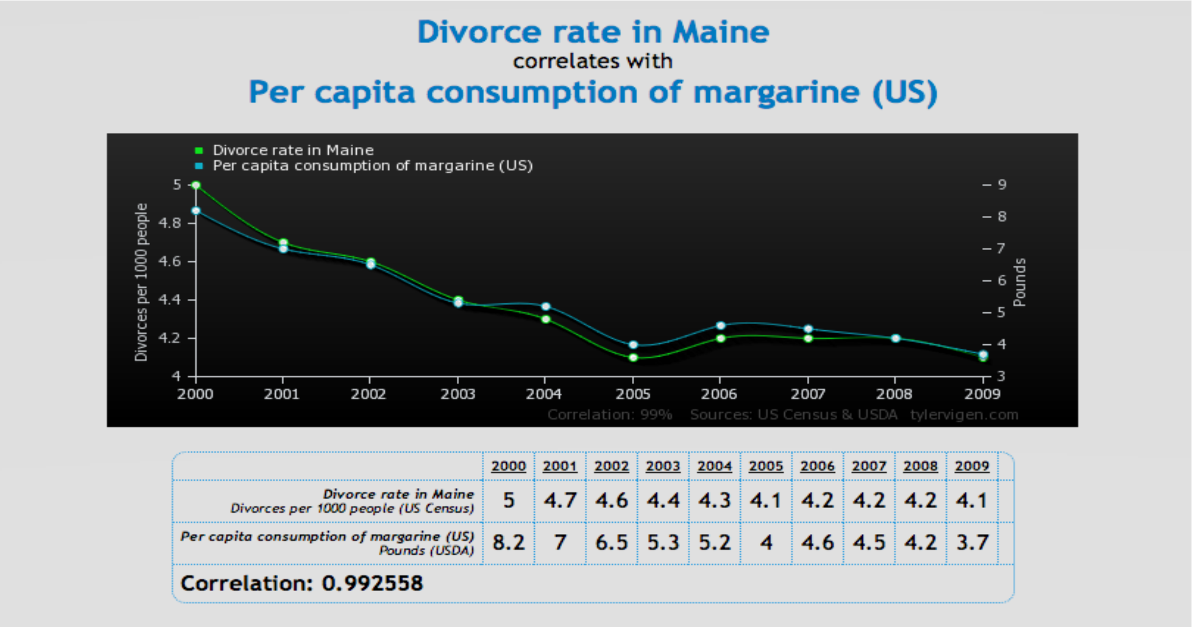
Highlights: Video Segment 6.3: Causation Video



It might be useful to explain that "causes" is an asymmetric relation (X causes Y is different from Y causes X), whereas "is correlated with" is a symmetric relation.

For instance, homeless population and crime rate might be correlated, in that both tend to be high or low in the same locations. It is equally valid to say that homelesss population is correlated with crime rate, or crime rate is correlated with homeless population. To say that crime causes homelessness, or homeless populations cause crime are different statements. And correlation does not imply that either is true. For instance, the underlying cause could be a 3rd variable such as drug abuse, or unemployment.

The mathematics of statistics is not good at identifying underlying causes, which requires some other form of judgement.



	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Divorce rate in Maine *Divorces per 1000 people (US Census)*	5	4.7	4.6	4.4	4.3	4.1	4.2	4.2	4.2	4.1
Per capita consumption of margarine (US) *Pounds (USDA)*	8.2	7	6.5	5.3	5.2	4	4.6	4.5	4.2	3.7

Correlation: 0.992558

http://www.businessinsider.com/spurious-correlations-by-tyler-vigen-2014-5

Causality is also a relationship between two things, but it is not mathematical, it is physical (or philosophical). Something causes something else if there is a chain of events between the first thing and the second thing, each of which causes the next thing in the chain to happen. Causality involves time; the first thing happens, and then later the second thing happens as a result. We say the first thing is the cause, and the second thing is the effect. Note that unlike correlation, the relationship is unsymmetrical.

Correlation is the mathematical relationship between two things which are measured. It is given as a value between -1 and 1. A correlation of 0 means the two things are unrelated; given the first value, there is no way to predict the second. A correlation of 1 means the two things are completely related, the first thing always predicts the second. As an example, let's say you measure the heights and weights of a group of people. These have a high correlation, somewhere around .8; height is a good predictor of weight, and vice-versa. Now say you took the same group and measured eye color. There is a low correlation between eye color and height, pretty close to 0. They are basically independent, knowing one doesn't tell you anything about the other.

http://www.w-uh.com/posts/030302a_correlation_vs_ca.html

Slide courtesy of Lynn Gill

Simple Linear Regression

The Regression Line

Equation of the Regression Line

The **equation of the regression line** that approximates the relationship between *x* and *y* is

$$\hat{y} = b_1 x + b_0$$

where the *regression coefficients* are the **slope**, *b_1,* and the ***y* intercept**, *b_0*.

The equations of these coefficients are

$$b_1 = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} \qquad b_0 = \bar{y} - (b_1 \cdot \bar{x})$$

Note: The “hat” over the *y* (pronounced “*y*-hat”) indicates this is an estimate of *y* and not necessarily an actual value of *y*.

Coefficient of Determination r^2

SSR (sum of squares residuals) represents the amount of variability in the response variable that is accounted for by the regression equation.

SSE (sum of squares error) represents the amount of variability in the *y* that is left unexplained after accounting for the relationship between *x* and *y*.

Since we know that SST represents the sum of SSR and SSE, it makes sense to consider the *ratio* of SSR and SST, called the **coefficient of determination** $\boldsymbol{r^2}$**.**

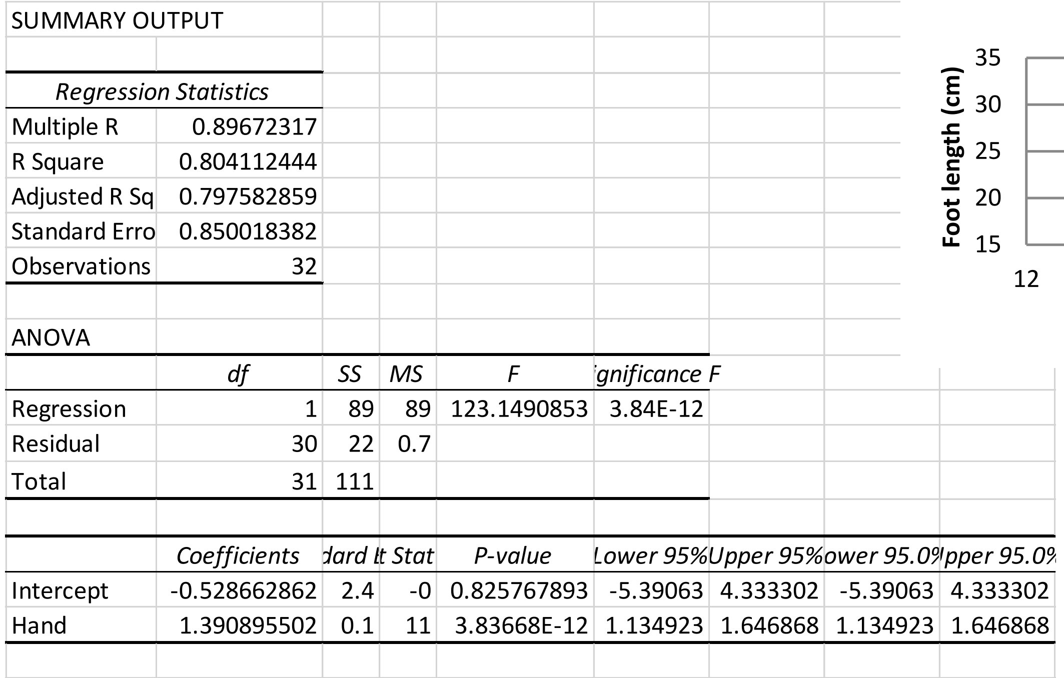
Coefficient of Determination r^2

The **coefficient of determination** $\boldsymbol{r^2}$ = *SSR/SST* measures the goodness of fit of the regression equation to the data. We interpret r^2 as the proportion of the variability in *y* that is accounted for by the linear relationship between *y* and *x*. The values that r^2 can take are $0 \le r^2 \le 1$.

Coefficient of Determination ($\boldsymbol{r^2}$**)**

Guideline:
r^2 value of greater than 0.77 desired, indicates good fit

Simple linear regression: Hand to foot



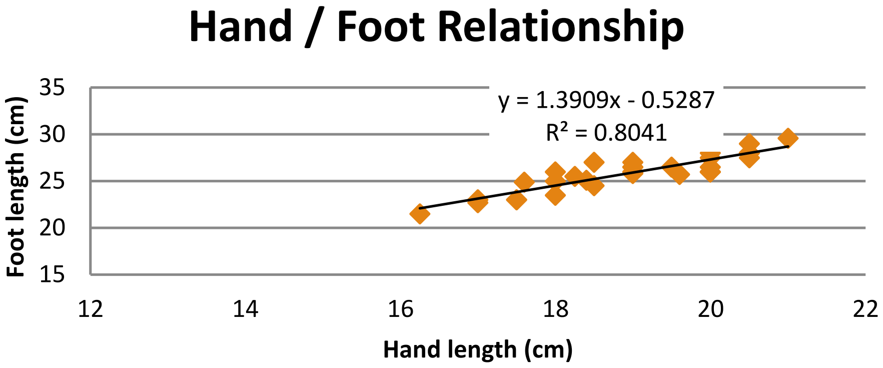
SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.89672317
R Square	0.804112444
Adjusted R Sq	0.797582859
Standard Erro	0.850018382
Observations	32

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regression	1	89	89	123.1490853	3.84E-12
Residual	30	22	0.7		
Total	31	111			

	Coefficients	*dard E*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	-0.528662862	2.4	-0	0.825767893	-5.39063	4.333302	-5.39063	4.333302
Hand	1.390895502	0.1	11	3.83668E-12	1.134923	1.646868	1.134923	1.646868



Questions to answer:

1. Is this a good fit? How do you know?
2. What is the equation for y=f(x)?
3. What % change in y is related to change in x?
4. If my hand length is 15.24 cm, what would you estimate my foot length to be?

Multiple Regression

Multiple Regression

Thus far, we have examined the relationship between the response variable *y* and a single predictor variable *x*. In our data-filled world, however, we often encounter situations where we can use more than one *x* variable to predict the *y* variable.

Multiple regression describes the linear relationship between one response variable *y* and more than one predictor variable x_1, x_2, …. The **multiple regression equation** is an extension of the regression equation

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$$

where *k* represents the number of *x* variables in the equation and b_0, b_1, … represent the **multiple regression coefficients.**

The interpretation of the regression coefficients is similar to the interpretation of the slope in simple linear regression, except that we add that the other *x* variables are held constant.

F Test for Multiple Regression

The multiple regression model is an extension of the model from Section 13.1, and approximates the relationship between *y* and the collection of *x* variables.

Multiple Regression Model

The **population multiple regression equation** is defined as:

$$y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

where β_1, β_2, …, β_k are the parameters of the population regression equation, *k* is the number of *x* variables, and ε is the error term that follows a normal distribution with mean 0 and constant variance.

The population parameters are unknown, so we must perform inference to learn about them. We begin by asking: *Is our multiple regression useful?* To answer this, we perform the ***F* test for the overall significance of the multiple regression.**

F Test for Multiple Regression

The hypotheses for the *F* test are:

H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$
H_a: At least one of the β*'s* ≠ 0.

The *F* test is not valid if there is strong evidence that the regression assumptions have been violated.

F Test for Multiple Regression

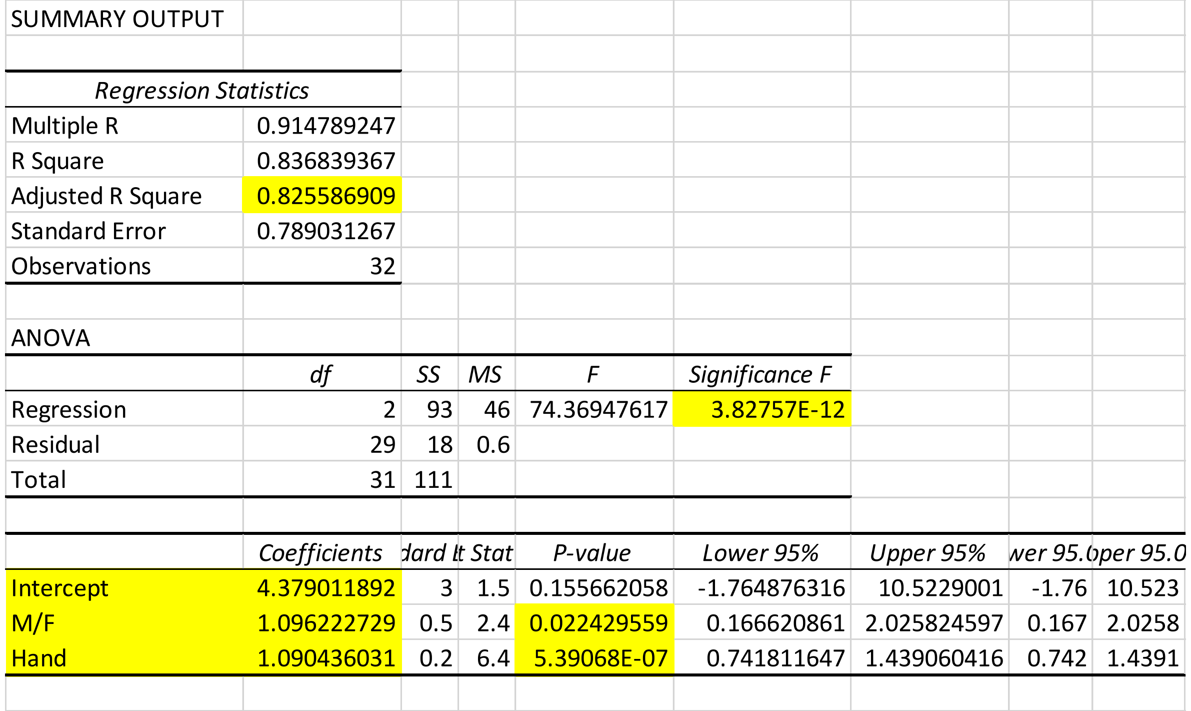
If the conditions for the regression model are met

Step 1: State the hypotheses and the rejection rule.

Step 2: Find the *F* statistic and the *p*-value. (Located in the ANOVA table of computer output.)

Step 3: State the conclusion and the interpretation.

Multiple regression: Hand and gender to foot

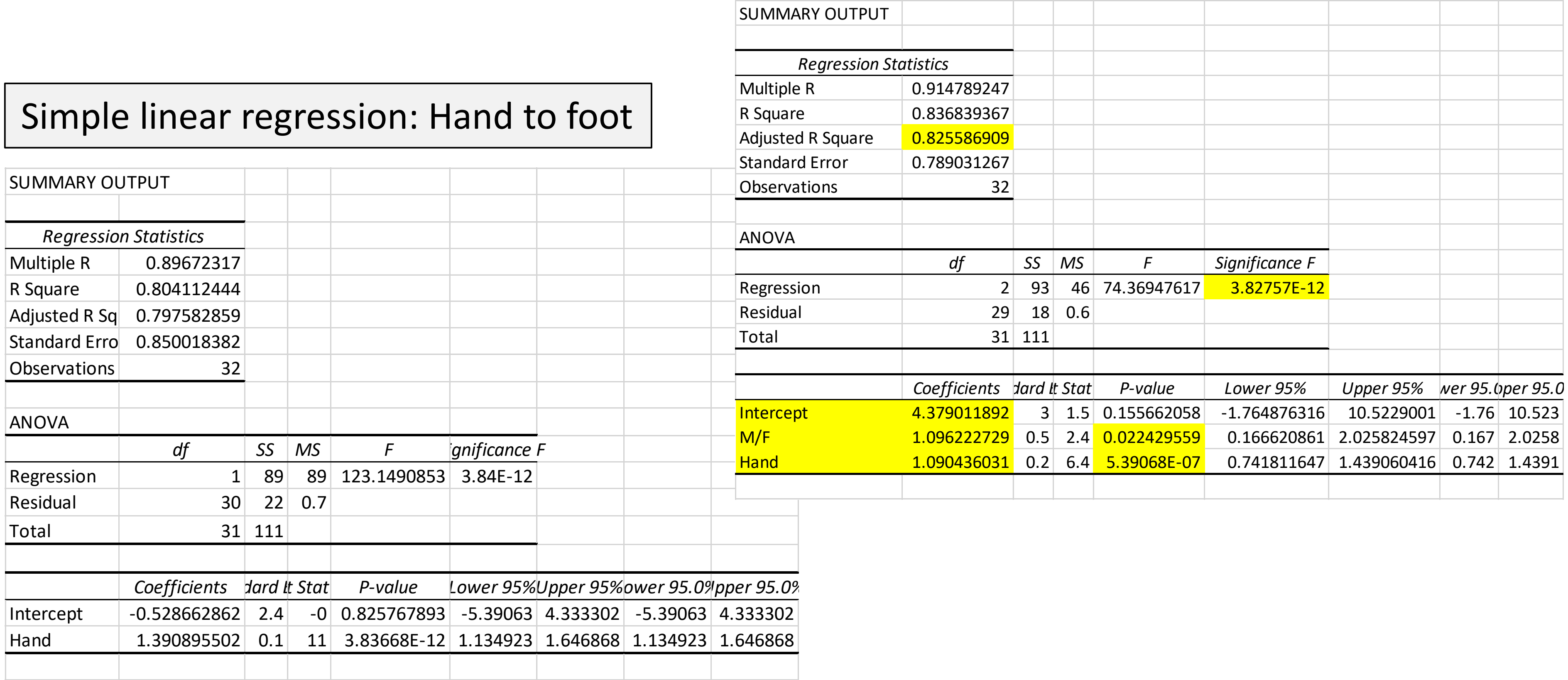


SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.914789247							
R Square	0.836839367							
Adjusted R Square	0.825586909							
Standard Error	0.789031267							
Observations	32							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	2	93	46	74.36947617	3.82757E-12			
Residual	29	18	0.6					
Total	31	111						
	Coefficients	*dard*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*wer 95.*	*per 95.0*
Intercept	4.379011892	3	1.5	0.155662058	-1.764876316	10.5229001	-1.76	10.523
M/F	1.096222729	0.5	2.4	0.022429559	0.166620861	2.025824597	0.167	2.0258
Hand	1.090436031	0.2	6.4	5.39068E-07	0.741811647	1.439060416	0.742	1.4391

Questions to answer (assume an alpha of 0.05):

1) What is the Ho and Ha for this multiple regression?
2) What is the p-value for this multiple regression? What is your conclusion?
3) What are the variables?
4) Which variables are significant?
5) How many samples were used to create this model?
6) What is the correlation for this model?
7) What is the equation for y=f(x)?

Recap of regression BLT



Simple linear regression: Hand to foot

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.89672317
R Square	0.804112444
Adjusted R Sq	0.797582859
Standard Erro	0.850018382
Observations	32

ANOVA

	df	SS	MS	F	ignificance F
Regression	1	89	89	123.1490853	3.84E-12
Residual	30	22	0.7		
Total	31	111			

	Coefficients	dard E	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%
Intercept	-0.528662862	2.4	-0	0.825767893	-5.39063	4.333302	-5.39063	4.333302
Hand	1.390895502	0.1	11	3.83668E-12	1.134923	1.646868	1.134923	1.646868

Multiple regression: Hand and gender to foot

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.914789247
R Square	0.836839367
Adjusted R Square	0.825586909
Standard Error	0.789031267
Observations	32

ANOVA

	df	SS	MS	F	Significance F
Regression	2	93	46	74.36947617	3.82757E-12
Residual	29	18	0.6		
Total	31	111			

	Coefficients	dard E	t Stat	P-value	Lower 95%	Upper 95%	wer 95.0	pper 95.0
Intercept	4.379011892	3	1.5	0.155662058	-1.764876316	10.5229001	-1.76	10.523
M/F	1.096222729	0.5	2.4	0.022429559	0.166620861	2.025824597	0.167	2.0258
Hand	1.090436031	0.2	6.4	5.39068E-07	0.741811647	1.439060416	0.742	1.4391

Questions to answer:

1) How could you compare the regression models with/without gender included?
2) Which model is "better" and why?

Next two weeks

1. Project Next Steps – Measure/Analyze Phases

Measure/Analysis tools
Data Stratification Tree or Data Measurement Plan
Data collection should be complete/near complete
Use "soft tools" and statistical tools to gain insights into the problem
Begin identifying solutions to try

2. Coursework BLT's:

7.8 Test Your Knowledge: Categorical Input Variable

7.9* Relate Regression to Your Project

8.7 Test Your Knowledge: Measurement System

8.8* Relate Control Charts to Your Project

3. Assignments:

Homework #4: *(worth 5 points)*
3 days after live session 7

LaunchPad Assignments

- **LearningCurve** for Chapter **4**

Upcoming assignment:

Homework #5: *(worth 3 points)* 3 days after live session 8

Assignments and Deliverables folder on 2SU

- Complete Control Chart problems **#1-10 on pgs 114 -116** from the ***Understanding Variation*** Book.