MBC 638

LIVE SESSION WEEK 7

Agenda

Topic	Time	Sunday Section	Wednesday Section
Introduction	5 min	6:30-6:35	9:00-9:05
Quiz 2 Recap	15 min	6:35-6:50	9:05-9:20
Highlights from Week 7 Video	65 min	6:50-7:55	9:20-10:25
Review of Upcoming Assignments and Open Question	5 min	7:55-8:00	10:25-10:30

Current Grade Status

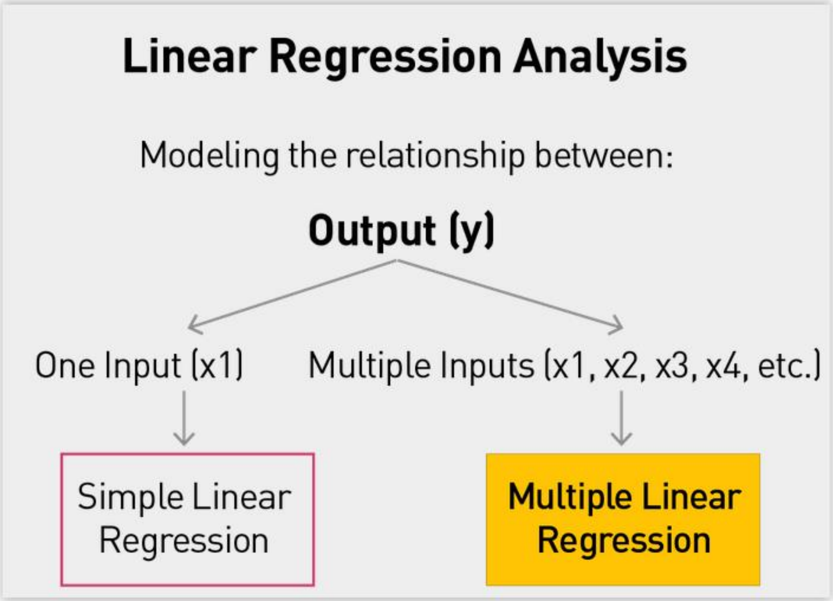
	January 2017	Points --->	10	10	10	5	3	2	10	5	3	2	15	5	20	100			
	Last Name, First Name		Participation	Prob Def Wkst	Quiz #1	Hwk #1	Hwk #2	Hwk #3	Quiz #2	Hwk #4	Hwk #5	Hwk #6	Paper	Storybrd	Final	Total	Points Still Availbable	Highest Potential Grade	Rounded
1	Doe	Jane	7	8.5	8	5	3	2	9	0	0	0				42.50	53	95.5	96.0

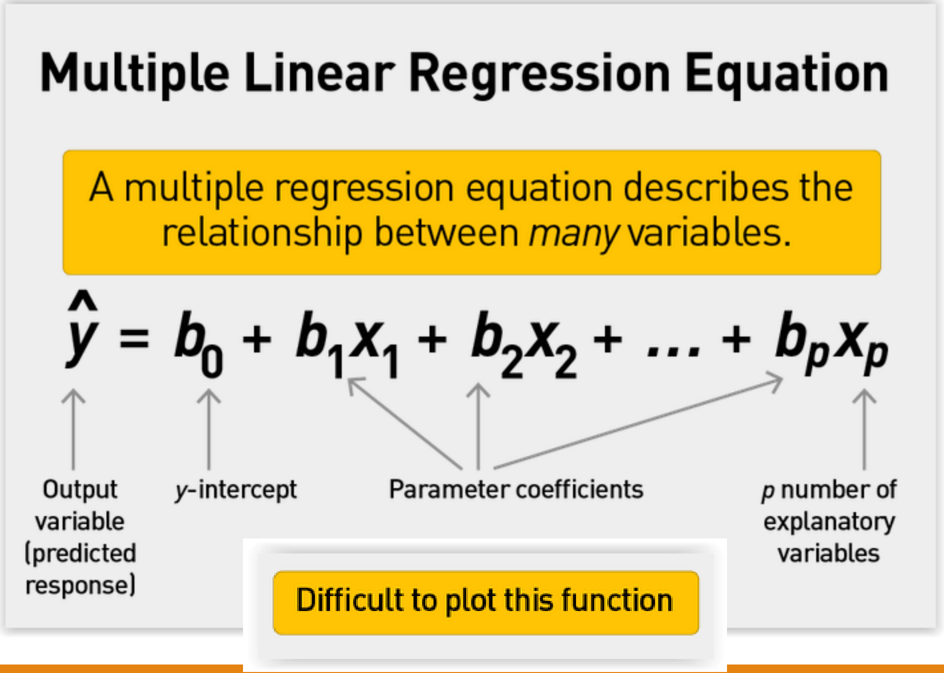
A	B	C	D	E	F	G	H	I	J	K	L	M	N
	January 2017	Participation	Live Session Attendance										
	Last Name, First Name		Wk1	Wk2	Wk3	Wk4	Wk5	Wk6	Wk7	Wk8	Wk9	Wk10	Totals
1	Doe	Jane	1	1	1	1	1	1	0	0	0	0	7

Agenda

Topic	Time	Sunday Section	Wednesday Section
Introduction	5 min	6:30-6:35	9:00-9:05
Quiz 2 Recap	15 min	6:35-6:50	9:05-9:20
Highlights from Week 7 Video	65 min	6:50-7:55	9:20-10:25
Review of Upcoming Assignments and Open Question	5 min	7:55-8:00	10:25-10:30

Highlights: Video Segment 7.3: Multiple Linear Regression





Highlights: Video Segment 7.4: Multiple Regression Using Excel

- P value helps you identify which input variable(s) are important/useful in describing Y.
- We want Ps lower than our alpha, assume .05, so if p is low Ho must go – our implied Ho is, X doesn't describe Y and Ha is, X does describe Y – so if we have a low p value we reject Ho and say X does describe Y.
- Then re-run your regression without the Xs that don't help describe Y to develop a better model.

	A	B	C	D	E	F	G	H	I	J	K	L
1				output (y)	inputs (x)							
2	First name	Last name	Team	Runs Scored	Hits	Doubles	Triples	Home Runs	RBIs	Walks	Bat Ave	Yankees?
3	Ichiro	Suzuki	SEA	111	238	22	7	6	68	49	0.351	0
4	Delmon	Young	TBD	65	186	38	0	13	93	26	0.288	0
5	Alexis	Rios	TOR	114	191	43	7	24	85	55	0.297	0
6	Derek	Jeter	NYY	102	206	39	4	12	73	56	0.322	1
7	Michael	Young	TEX	80	201	37	1	9	94	47	0.315	0
8	Orlando	Cabrera	LAA	101	192	35	1	8	86	44	0.301	0
9	Nick	Markakis	BAL	97	191	43	3	23	112	61	0.3	0
10	Grady	Sizemore	CLE	118	174	34	5	24	78	101	0.277	0
11	Brian	Roberts	BAL	103	180	42	5	12	57	89	0.29	0
12	Robinson	Cano	NYY	93	189	41	7	19	97	39	0.306	1
13	Curtis	Granderson	DET	122	185	38	23	23	74	52	0.302	0
14	Aaron	Hill	TOR	87	177	47	2	17	78	41	0.291	0
15	Bobby	Abreu	NYY	123	171	40	5	16	101	84	0.283	1
16	David	DeJesus	KCR	101	157	29	9	7	58	64	0.26	0
17	Torii	Hunter	MIN	94	172	45	1	28	107	40	0.287	0
18	Adrian	Beltre	SEA	87	164	41	2	26	99	38	0.276	0
19	Magglio	Ordonez	DET	117	216	54	0	28	139	76	0.363	0
20	Jose	Guillen	SEA	84	172	28	2	23	99	41	0.29	0
21	Justin	Morneau	MIN	84	160	31	3	31	111	64	0.271	0

	Coefficients	Standard Error	t Stat	P-value
Intercept	-6.677284091	8.358708258	-0.79884	0.4260
Hits	0.437991874	0.048707908	8.992213	0.00000000
Doubles	0.001881127	0.142779213	0.013175	0.9895
Triples	1.236783363	0.291004023	4.250056	0.0000
Home Runs	0.757792868	0.174045929	4.353982	0.00002863
RBIs	-0.203601646	0.075335531	-2.7026	0.0079
Walks	0.283549256	0.043817935	6.471078	0.0000
Bat Ave	12.65150623	38.82469554	0.325862	0.7451
Yankees?	9.194227286	3.330203822	2.76086	0.0067
				alpha = 0.05

$$\hat{y} = -6.678 + 0.438x_{hits} + 0.002x_{dbles} + 1.237x_{triples} + 0.758x_{homeruns}$$
$$-0.204x_{RBIs} + 0.284x_{walks} + 12.652x_{batave} + 9.194x_{yanks}$$

Highlights: Video Segment 7.5: Correlation, *F* Test, and Model Building

Correlation Coefficients

- **Multiple correlation coefficient (R)**
  - Measures relationship between observed output y and predicted output y
- **Coefficient of determination (R^2)**
  - Proportion of the variation in response variable that is explained by model
  - Always increases when another $x$is added to model
    - E.g., if model to predict output y has two x inputs, adding a third x will increase R^2

Multiple R is a little different for multiple regression in that it is the relationship between the observed and predicted Y. In simple linear regression this was the correlation between the X and Y.

Adjusted R square is a better measure to look at when trying to determine how good your model is…how much of the variability in Y is explained by your equation

Correlation Coefficients (cont.)

- **Adjusted R^2**
  - Measure that helps account for too many unnecessary x inputs
  - Often x inputs are added in order to increase R^2
    - Higher R^2 makes model seem better, but having more inputs complicates forecast

Adjusted Coefficient of Determination

We measure the goodness of a regression equation using the coefficient of determination r^2 = SSR/SST. In multiple regression, we use the same formula for the coefficient of determination (though the letter r is promoted to a capital R).

Multiple Coefficient of Determination R^2

The multiple coefficient of determination is given by:

$$R^2 = \text{SSR/SST} \quad 0 \leq R^2 \leq 1$$

where SSR is the sum of squares regression and SST is the total sum of squares. **The multiple coefficient of determination represents the proportion of the variability in the response *y* that is explained by the multiple regression equation**.

Adjusted Coefficient of Determination

Unfortunately, when a new *x* variable is added to the multiple regression equation, the value of R^2 *always increases,* even when the variable is not useful for predicting *y*. So, we need a way to adjust the value of R^2 as a penalty for having too many unhelpful *x* variables in the equation.

Adjusted Coefficient of Determination R^2_{adj}

The adjusted coefficient of determination is given by:

$$R^2_{\text{adj}} = 1 - (1 - R^2)\left(\frac{n-1}{n-k-1}\right)$$

where *n* is the number of observations, *k* is the number of *x* variables, and R^2 is the multiple coefficient of determination.

Highlights: Video Segment 7.5: Correlation, *F* Test, and Model Building

	A	B
1	**8 Input Variables**	
2		
3	SUMMARY OUTPUT	
4	*Regression Statistics*	
5	Multiple R	0.9326355
6	R^2	0.8698089
7	Adjusted R^2	0.8609824
8	Standard Error	8.6228711
9	Observations	127

Multiple R = measure of the relationship of the observed output of Y and the predicted of Y

R Squared = proportion of the variation explained by the model, always increases when you add more Xs

Adjusted R Squared = accounts for excess Xs, a better measure of the variability explained by the model

Degrees of freedom

Sum of Squares
Amt of variation that can't be accounted for

Mean Square
Variance around the fitted line

F statistic

Probability value

Highlights: Video Segment 7.5: Correlation, *F* Test, and Model Building

Overall Model Significance

	A	B	C	D	E	F
1	**ANOVA**					
2		df	SS	MS	F	Significance F
3	Regression	8	58617.65646	7327.207	98.54502	1.444E−48
4	Residual	118	8773.760859	74.35391		
5	Total	126	67391.41732			

- *F*: intermediate step on the way to probability
- Significance *F*: *p*-value
  - "If *p* is low, H_0 must go."

ANOVA= analysis of variance

F provides a p value, our measure of goodness

Ho: coefficient of variables =0
Ha: At least something in my model is a good x input and should have a value, at least one variable helps forecast Y.

Must have a low F to have a valid model.

Hypothesis Test

- H_0: $\beta_1 = \beta_2 = \ldots = \beta_p = 0$
- H_a: at least *one* $\beta_j \neq 0$

$$\hat{y} = -6.678 + 0.438x_{hits} + 0.002x_{dbles} + 1.237x_{triples} + 0.758x_{\text{homeruns}} - 0.204x_{RBIs} + 0.284x_{walks} + 12.652x_{batave} + 9.194x_{yanks}$$

- Significant *p*-value does not mean all *x*'s have significant influence on *y*
  - Does mean one or more *x*'s has significant influence
  - Refine model to eliminate less useful variables
- Failing to reject H_0: no evidence that any coefficient $\beta_j \neq 0$
- To assess model, look at ANOVA
  - Low *F* statistic indicates good model

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 $\neq 0$.

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.

Highlights: Video Segment 7.5: Correlation, *F* Test, and Model Building

Model Building

- Nonlinear regression models exist
  - Remember: **practical, graphical, statistical**
  - Plot data whenever possible
- Curved relationship: x^2 variable(s); possibly quadratic function(s)
- Reciprocals: $1/x$
- Interaction terms: $x_1 \times x_2$
- Too many variables → pare down
  - Best models: fewer x inputs but same predictive value

Highlights: Video Segment 7.6:Just Correlation

	Runs Scored	Hits	Doubles	Triples	Home Runs	RBIs	Walks	Bat Ave	Yankees?
Runs Scored	1								
Hits	0.843722818	1							
Doubles	0.672563335	0.788082	1						
Triples	0.32311721	0.277987	0.123538	1					
Home Runs	0.521110368	0.346261	0.424893	-0.17987	1				
RBIs	0.665750495	0.672397	0.682796	-0.06029	0.808734	1			
Walks	0.630325512	0.384619	0.352119	-0.00888	0.604142	0.544677	1		
Bat Ave	0.568341926	0.727978	0.540431	0.166983	0.124469	0.417129	0.185354	1	
Yankees?	0.34809467	0.263619	0.189084	0.098172	0.172827	0.272852	0.203198	0.199359	1

Look through the table to identify high correlation variables.

Multiple R is the observed to predicted Y relationship.

Adjusted R Square accounts for the fact that you have possibly too many Xs, discounts the Xs that are not relevant to the model- the high p value Xs. Overall, how well do my Xs describe my output.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.932635474
R Square	0.869808928
Adjusted R Square	0.860982415
Standard Error	8.622871076
Observations	127

	Coefficients	andard Err	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%
Intercept	-6.677284091	8.358708	-0.79884	0.425987	-23.2298	9.875234	-23.2298	9.875234
Hits	0.437991874	0.048708	8.992213	4.81E-15	0.341537	0.534447	0.341537	0.534447
Doubles	0.001881127	0.142779	0.013175	0.98951	-0.28086	0.284623	-0.28086	0.284623
Triples	1.236783363	0.291004	4.250056	4.29E-05	0.660516	1.813051	0.660516	1.813051
Home Runs	0.757792868	0.174046	4.353982	2.86E-05	0.413135	1.102451	0.413135	1.102451
RBIs	-0.203601646	0.075336	-2.7026	0.007895	-0.35279	-0.05442	-0.35279	-0.05442
Walks	0.283549256	0.043818	6.471078	2.31E-09	0.196778	0.370321	0.196778	0.370321
Bat Ave	12.65150623	38.8247	0.325862	0.745106	-64.232	89.53497	-64.232	89.53497
Yankees?	9.194227286	3.330204	2.76086	0.006687	2.599517	15.78894	2.599517	15.78894

This can be used to look through to identify high p-value Xs to potentially eliminate.

Highlights: Video Segment 7.7: Categorical Input Variables

Categorical Input Variables (x's)

- Discrete, categorical x variable may be influencing output y
  - E.g., x_{yankees}: 0 or 1
  - Put categorical x variables in regression by assigning each variable 0 or 1

Example: Months of the Year

$K = N - 1 = 12 - 1 = 11$ input variables

Video Slide Presentation

Y	J	F	M	A	M	J	J	A	S	O	N
10	1	0	0	0	0	0	0	0	0	0	0
22	0	1	0	0	0	0	0	0	0	0	0

Highlights: Video Segment 7.7: Categorical Input Variables

Example: Power Tools

What drives the price of a particular power tool?

- Output (y) = price of the tool (continuous data)
- Inputs (x) =
  - Product brands (discrete data)
  - Types of accessories (discrete data)
  - Weight of the tool (continuous data)

Example: Power Tools: Data

	Brand 1	Brand 2	Brand 3	Accessories		
Price y	Sears x_1	Toshiba x_2	Dremel x_3	Basic x_4	Xtra x_5	Weight x_6
20	1	0	0	1	0	10
40	0	1	0	1	0	30
60	0	0	1	0	1	32
30	1	0	0	1	0	12
35	0	1	0	1	0	11

1 less than N is needed for regression

	Brand 1	Brand 2	Accessories	
Price y	Sears x_1	Toshiba x_2	Basic x_4	Weight x_6
20	1	0	1	10
40	0	1	1	30
60	0	0	0	32
30	1	0	1	12
35	0	1	1	11

Highlights: Video Segment 7.7: Categorical Input Variables

Example: Power Tools: Regression Equation

	Coefficients	*Std Error*	*t Stat*	*P-Value*	*Lower 95%*
Intercept	48.3628029	14.17717	3.411315	0.019016	11.91922486
x1	−21.377046	9.757484	−2.19084	0.080012	−46.45945872
x2	−12.478716	7.311039	−1.70683	0.148563	−31.27233932
x4	−4.2239686	8.881025	−0.47562	0.6544	−27.0533713
x6	0.2848723	0.391725	0.727225	0.499698	−0.722089489

$$\hat{y}=48.36-21.38x_1-12.48x_2-4.22x_4+0.28x_6$$

Estimating Price (y)

$$\hat{y}=48.36-21.38x_1-12.48x_2-4.22x_4+0.28x_6$$

Tool characteristics:

- Dremel (Brand 3): $x_1 = 0$ and $x_2 = 0$
- Extra accessories: $x_4 = 0$
- Weight 25 lb: $x_6 = 25$

$$\hat{y}=48.36-21.38(0)-12.48(0)-4.22(0)+0.28(25)=\$55.36$$

Highlights: Video Segment 7.8: Test your knowledge

SUMMARY OUTPUT

Regression Statistics			
Multiple R	0.914789247	when only using hand to predict foot	
R Square	0.836839367		0.79
Adjusted R Square	0.825586909		0.8
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 95.0*	*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

Regression Equation	y'=4.379+1.096X + 1.0904X2
my actual hand size in inches	7.625
my actual hand size in cms	19.3675
Use formula to predict foot size	y'=4.379+1.096X + 1.0904X2
predicted foot size	25.497322
actual foot in cms	25.4
residual	=Actual foot size - predicted foot size
	=25.4-25.497322
residual	-0.097322000

So the multiple regression when considering gender creates a smaller residual(error), it is a better predictor of my foot size

Using Simple Linear Regression - only Hand	
y' = (1.39X 19.3675)-.52866	
predicted foot size	26.392165000
actual foot in cms	25.4
residual	-0.992165000

Agenda

Topic	Time	Sunday Section	Wednesday Section
Introduction	5 min	6:30-6:35	9:00-9:05
Quiz 2 Recap	15 min	6:35-6:50	9:05-9:20
Highlights from Week 7 Video	65 min	6:50-7:55	9:20-10:25
Review of Upcoming Assignments and Open Question	5 min	7:55-8:00	10:25-10:30

Review of Upcoming Assignments: Wednesday

1. HMWK #4, Learning Curve on Chapter 4, in LaunchPad is due until Saturday, 3/4, midnight EST.

2. Understanding Variation Book –Live Class #8 is focused on the material from this book

3. Optional Learning Opportunity: 8.8 Relate Control Charts to Your Project

4. Projects…. Should begin Improve in the next week or so, such that you have time to collect “after” data

	March 2017						
	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
Week #7	26	27	28	1	2	3	4
				Live Class #7			***Homework #4 Due:*** 1. **CH 4** Learning Curve **Reminder:** Start reading Understanding Variation
Week #8	5	6	7	8	9	10	11
				Live Class #8			***Homework #5 Due:*** 1. Problems 1-10 pg 114-116 in Understanding Variation
Week #9	12	13	14	15	16	17	18
				Live Class #9			***Homework #6 Due:*** 1. Time Series Problem posted in Excel
Week #10	19	20	21	22	23	24	25
				Live Class #10			***Data collection and Analysis Paper DUE***
	26	27	28	29	30	31	1
		Project Storyboard DUE		***Final Exam DUE***			