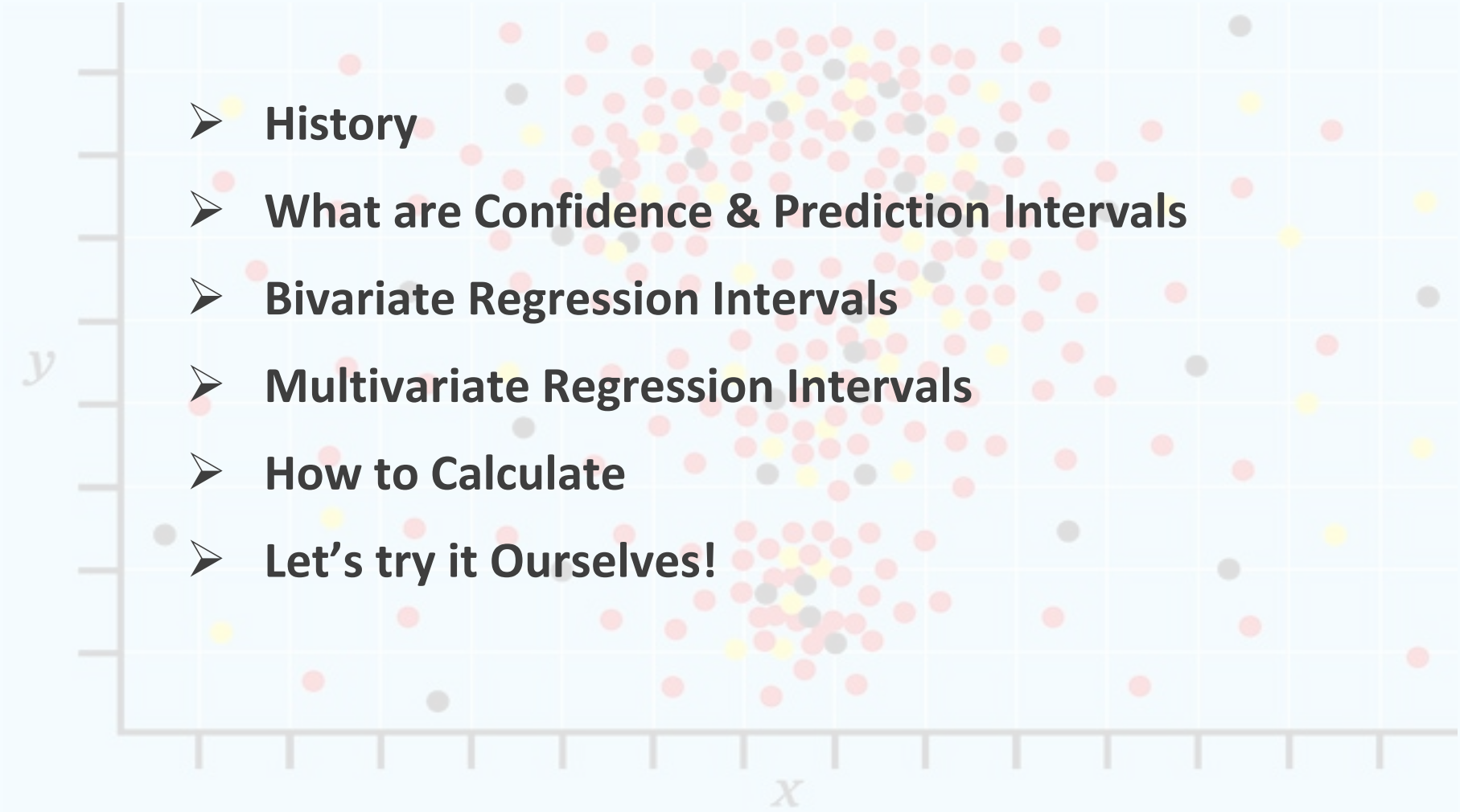
Presented at the 2016 ICEAA Professional Development & Training Workshop - www.iceaaonline.com/atlanta2016

Multiple Regression

Prediction & Confidence Intervals Demystified

Stacy Dean
The Boeing Company

Agenda

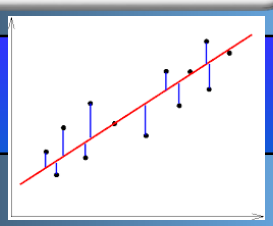


- History
- What are Confidence & Prediction Intervals
- Bivariate Regression Intervals
- Multivariate Regression Intervals
- How to Calculate
- Let's try it Ourselves!

Brief History of Confidence Intervals

A long time ago

1805 – Adrien-Marie Legendre
Least squares method for fitting data
"Nouvelles Méthodes pour la Détermination des Orbites des Comètes"

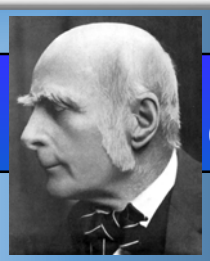


+ 3yrs

1808 – Carl Friedrich Guass and Laplace
The normal distribution

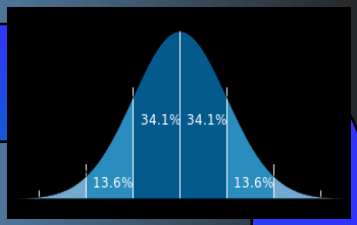
+ 71yrs

1877 – Francis Galton
Regression to the mean
1888 – Correlation



+ 6yrs

1894 – Karl Pearson
Standard deviations & Central limits theorem



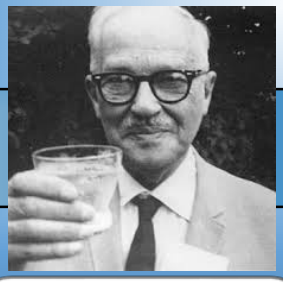
+ 14yrs

1908 – William Gosset
The t-test

+ 27yrs

1935 – R.A Fisher
Concept of statistical significance
"Design of Experiments"

+ 2yrs



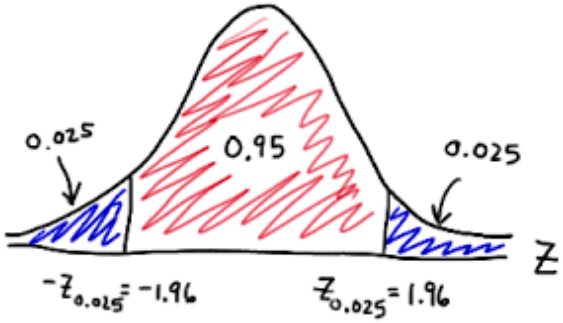
1937 – Jerzy Neyman
Confidence intervals in statistical testing
"Outline of a Theory of Statistical Estimation based on the Classical Theory of Probability"

2016

So, yeah… it's only 79 years old!

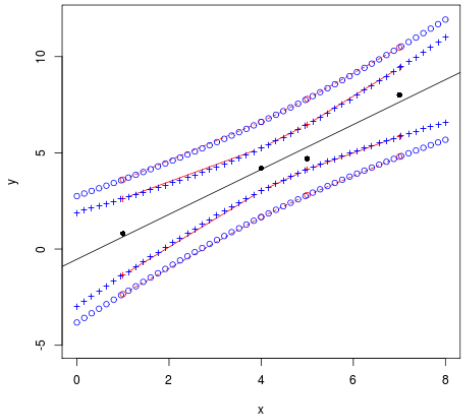
What are Confidence / Prediction Intervals?

Interval measuring the region that the **true** value may actually fall around the *estimated* value (given a *specified level of certainty*)



What's it good for?

- Measures the range around an estimated value
- Sanity check for statistical significance
- Foundation for hypothesis testing



Mighty helpful in measuring risk in CER estimating risk

How do you calculate them for a CER estimate?

$$CI = \hat{y} \pm t_{\left(\frac{\alpha}{2}, n-2\right)} * SEE \sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{\sum X^2 - n\bar{X}^2}}$$

$$PI = \hat{y} \pm t_{\left(\frac{\alpha}{2}, n-2\right)} * SEE \sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{\sum X^2 - n\bar{X}^2}}$$

What the heck is the formula doing???

What Is The Math Doing??

$$SEE = \sqrt{\frac{\sum(Y-\hat{Y})^2}{n-1-k}}$$

$$Se_b = \sqrt{\frac{SEE^2}{\sum(x-\bar{x}^2)}}$$

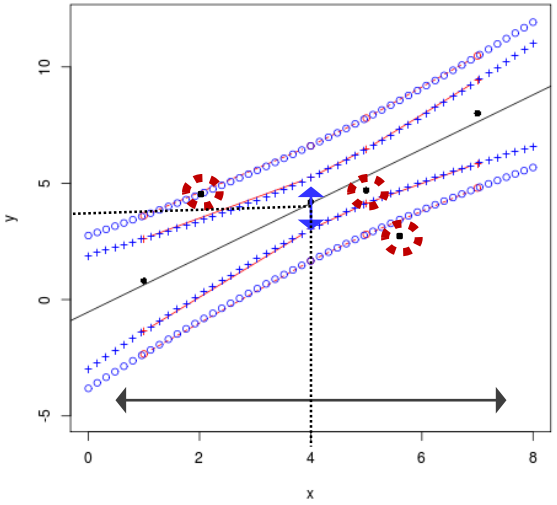
- **We are most confident in our estimate $(\hat{Y})$ at the mean $(\bar{X})$**
  - The farther away we are from $\bar{X}$ the less certain we are in the $\hat{Y}$

$$CI = \hat{y} \pm t_{\left(\frac{\alpha}{2},n-2\right)} * SEE\sqrt{\frac{1}{n}+\frac{(X-\bar{X})^2}{\sum X^2 - n\bar{X}^2}} \Rightarrow \sqrt{\frac{SEE^2}{n} + Se_b * \left((X-\bar{X})^2\right)}$$

Movement in Y (bounce) — *Movement in X (wiggle)*

$$PI = \hat{y} \pm t_{\left(\frac{\alpha}{2},n-2\right)} * SEE\sqrt{1+\frac{1}{n}+\frac{(X-\bar{X})^2}{\sum X^2 - n\bar{X}^2}} \Rightarrow \sqrt{SEE^2 + \frac{SEE^2}{n} + Se_b * \left((X-\bar{X})^2\right)}$$

Irreducible Noise — *Movement in Y (bounce)* — *Movement in X (wiggle)*



- **Confidence intervals measure the estimate variation (bounce) with respect to how far X is away from $\bar{X}$ (wiggle)**
  - Prediction intervals just account for the fact that our model didn't explain 100% of the data

So…. How do you deal with multiple Xs???

Multiple Regression Covariance
Moving from a Solo Act to a Choir

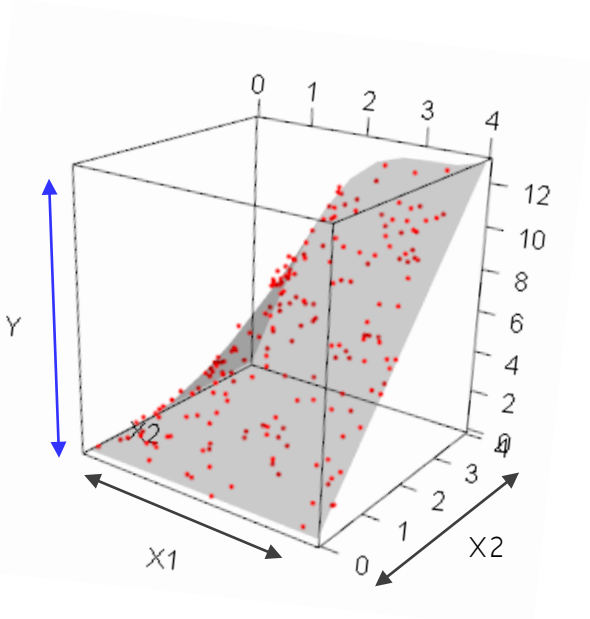
- **With multivariate regression the confidence and prediction interval must account for the simultaneous *wiggle* of multiple X variables**
  - To calculate the simultaneous *wiggle* we'll need to account for the *variance* between the variables (covariance) as well as with Y

$$CI = \hat{y} \pm t_{\left(\frac{\alpha}{2},DF\right)} * \sqrt{CovInt_i + 2 * \left(\sum x_i * CovInt_i\right) + \left(\left(\sum x_i * CovX_i\right) * x_i\right)}$$

Movement in Y (bounce) — *Movement in X (wiggle)*

$$PI = \hat{y} \pm t_{\left(\frac{\alpha}{2},DF\right)} * \sqrt{MSE + CovInt_i + 2 * \left(\sum x_i * CovInt_i\right) + \left(\left(\sum x_i * CovX_i\right) * x_i\right)}$$

Irreducible Noise — *Movement in Y (bounce)* — *Movement in X (wiggle)*



The equation requires a few more steps, but nothing we can't handle…

Multivariate Confidence/Prediction Interval
3 Major Steps

1 Perform Regression

1. Calculate Regression Coefficients
2. Calculate Regression MSE (ANOVA)

2 Calculate Covariance Matrix

4. Dataset Matrix Multiplication
5. Inverse Matrix
6. Covariance Matrix

3 Calculate Prediction Interval

7. Regression MSE
8. Covariance of Intercept
9. Covariation of X's & Intercept
10. Covariation of X's

Step #1 Begin!

- **So we got our data….**

x1	x2	x3	y
1.69	148,521.20	(3,433.15)	$555.75
10.67	150,681.86	(4,262.38)	$503.14
596.37	153,337.06	(4,596.91)	$383.19
604.08	152,282.06	(4,560.57)	$374.61
500.24	153,147.59	(4,515.39)	$356.78
685.64	152,315.21	(5,182.86)	$350.75
618.71	152,489.73	(4,654.51)	$339.69
209.40	152,165.43	(5,170.26)	$401.26
113.83	153,895.14	(4,219.08)	$424.72
243.95	153,262.72	(4,350.00)	$424.29

Summary of Fit	
R Square	0.952
R Square Adj.	0.928
Root Mean Square Error	18.687
Mean of Response	411.418
Observations (or Sum Wgts)	10

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Model	3	41557.7	13852.6	39.6689	0.0002
Error	6	2095.23	349.2		
C. Total	9	43652.9			

Parameter Estimates				
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	3031.41	714.806	4.24	0.0054
x1	-0.1423	0.03017	-4.72	0.0033
x2	-0.0159	0.00488	-3.25	0.0174
x3	0.03463	0.01705	2.03	0.0886

- **We build our regression equation**
  - Calculate the Regression Coefficient Estimates
  - MSE ANOVA Statistic

Step 1 complete!

Step #2
Enter the Matrix!

1. **Start with the data points for your independent variables add one additional column consisting of 1 for each row in your dataset**
   - The 1's represents the intercept term
   - Make sure the row consisting of 1s is in the first column
2. **Now, calculate the covariance for the dependent variables using matrix multiplication**
   - You can use the SUMPRODUCT() function for this

Intercept	x1	x2	x3
1	1.69	148,521.20	(3,433.15)
1	10.67	150,681.86	(4,262.38)
1	596.37	153,337.06	(4,596.91)
1	604.08	152,282.06	(4,560.57)
1	500.24	153,147.59	(4,515.39)
1	685.64	152,315.21	(5,182.86)
1	618.71	152,489.73	(4,654.51)
1	209.40	152,165.43	(5,170.26)
1	113.83	153,895.14	(4,219.08)
1	243.95	153,262.72	(4,350.00)

Add this guy here

Data Matrix

10	3,585	1,522,098	(44,945)
3,585	1,940,148	547,455,583	(16,863,929)
1,522,098	547,455,583	231,700,368,370	(6,844,966,617)
(44,945)	(16,863,929)	(6,844,966,617)	204,254,755

Here's the Math

$=\sum(A_i*A_i)$	$=\sum(A_i*B_i)$	$=\sum(A_i*C_i)$	$=\sum(A_i*D_i)$
$=\sum(B_i*A_i)$	$=\sum(B_i*B_i)$	$=\sum(B_i*C_i)$	$=\sum(B_i*D_i)$
$=\sum(C_i*A_i)$	$=\sum(C_i*B_i)$	$=\sum(C_i*C_i)$	$=\sum(C_i*D_i)$
$=\sum(D_i*A_i)$	$=\sum(D_i*B_i)$	$=\sum(D_i*C_i)$	$=\sum(D_i*D_i)$

So far so good?

Step #2 Cont.
On to the Covariance Matrix

4. **Now we need to calculate the inverse of our data matrix**
   - Why do with??? Well, we're trying to calculate the reciprocal of the matrix (basically 1/matrix), but you can't actually divide matrices
   - The way to get around this limitation is calculating the inverse matrix or Matrix^{1}

Data Matrix

10	3,585	1,522,098	(44,945)
3,585	1,940,148	547,455,583	(16,863,929)
1,522,098	547,455,583	231,700,368,370	(6,844,966,617)
(44,945)	(16,863,929)	(6,844,966,617)	204,254,755

$^{-1}$ =

Inverse Matrix

1463.174	0.016351	-0.009936	-0.009673
0.016351	0.00000261	-0.00000009	0.00000071
-0.009936	-0.00000009	0.00000007	0.00000009
-0.009673	0.00000071	0.00000009	0.00000083

To make the math easier we can just use the Minverse() function in Excel

   - Finally, we multiple the Inverse Matrix by the Mean Squared Error of the regression…

MSE
349.2

×

1463.174	0.016351	-0.009936	-0.009673
0.016351	0.00000261	-0.00000009	0.00000071
-0.009936	-0.00000009	0.00000007	0.00000009
-0.009673	0.00000071	0.00000009	0.00000083

=

$CovInt$	$CovX_1$	$CovX_2$	$CovX_3$
510947.36	5.70991	-3.46983	-3.37801
5.70991	0.00091	-0.00003	0.00025
-3.46983	-0.00003	0.00002	0.00003
-3.37801	0.00025	0.00003	0.00029

Viola! We have our Covariance matrix!

Step #3
Putting it All Together!

- **The last step is the prediction interval formula…**

$$PI = \hat{y} \pm t_{\left(\frac{\alpha}{2}, n-2\right)} * \sqrt{MSE + CovInt_i + 2 * \left(\sum x_i * CovInt\right) + \left(\sum (x_i * CovX_i)_i * x_i\right)}$$

  - Select a value for each X:

X
1.69
148,521.20
(3,433.15)

  - Use the SumProduct() to multiple the X values by the Covariance Matrix

	$CovInt$	$CovX_1$	$CovX_2$	$CovX_3$		X		$x_i * CovInt$	$x_i * CovX_1$	$x_i * CovX_2$	$x_i * CovX_3$
$CovInt$	510947.36	5.70991	-3.46983	-3.37801							
$CovX_1$	5.70991	0.00091	-0.00003	0.00025	✖	1.69	=	9.65	0.002	(0.0001)	0.0004
$CovX_2$	-3.46983	-0.00003	0.00002	0.00003		148,521.20		(515,343)	(4.797)	3.5295	4.4845
$CovX_3$	-3.37801	0.00025	0.00003	0.00029		(3,433.15)		11,597.20	(0.855)	(0.1037)	(0.9985)
						Sum		**(503,736)**	**(5.651)**	**3.4257**	**3.4864**

✖

$(x_i * CovX_1) * x_i$	1.69	148,521.20	(3,433.15)	Total
	(9.55)	508,795.40	(11,969.49)	**496,816.36**

  - Lastly, pick a significance level and plug the values into the formula

$$PI = \hat{y} \pm t_{\left(\frac{5\%}{2}, 6\right)} * \sqrt{349.2 + 510,947 + 2 * (503,736) + (496,816)} = (495.5 \quad -619.4)$$

Let's try it for real now!