Statistics Review

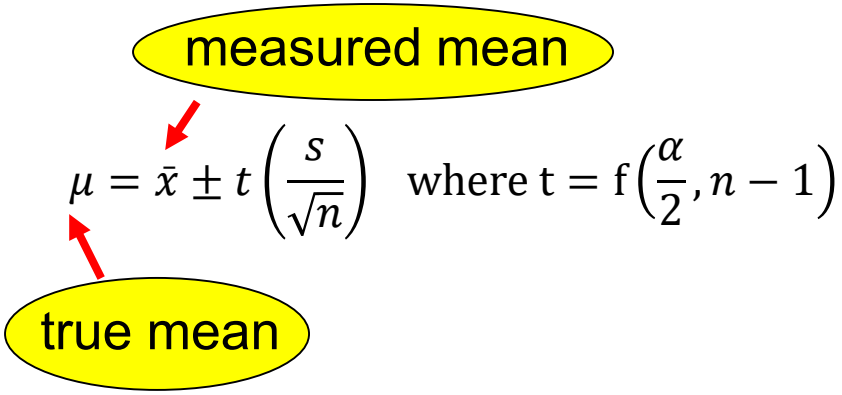
ChE 475

Winter 2020

Dr. Fletcher

1. Repeated Data Points

- **Use t-test based on measured st dev (*s*)**



In Excel,

=T.INV(α,r) for one-tailed test (α =0.025 for 95% confidence interval)

=T.INV.2T(α,r) for two-tailed test (α =0.05 for 95% confidence interval)

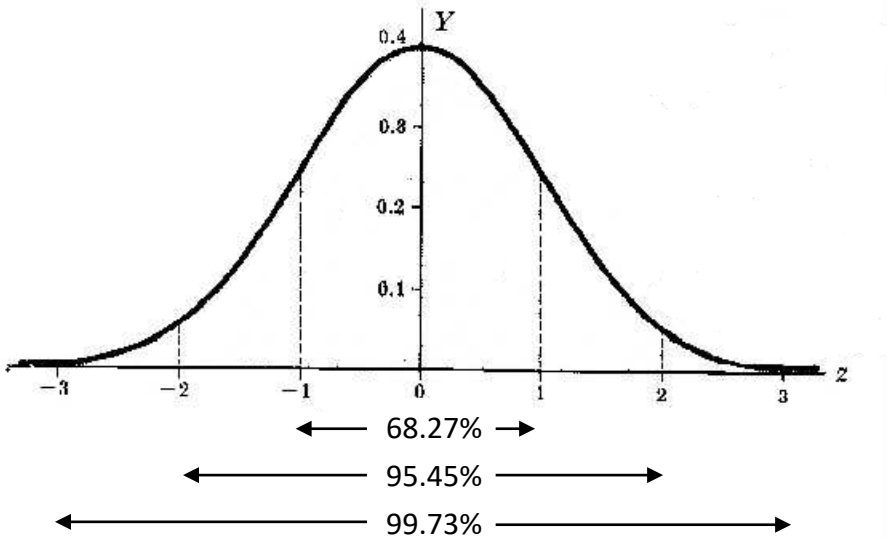
r = n-1

Alternate:

=CONFIDENCE.T(α,STDEV,n) for two-tailed test (α =0.05 for 95% confidence interva

(use n, not n-1)

Gaussian Distribution



- 68.27% of distribution lies within one σ
- 95.45% of distribution lies within two σ
- 99.73% of distribution lies within three σ

- *t*-test is used when we do not have enough data points

2. Comparing averages of measured variables

New formula:

Step 1

$$T = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_{x1}-1)s_{x1}^2 + (n_{x2}-1)s_{x2}^2}{n_{x1}+n_{x2}-2}\left(\frac{1}{n_{x1}}+\frac{1}{n_{x2}}\right)}}$$

For this example, $T = 2.5$

Larger $|T|$: More likely different

Step 2

r	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$	$\alpha = 0.005$
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921

$r = n_{x1}+n_{x2}-2$

Step 3

At a given confidence level (e.g. 95% or α=0.05), there is a difference if:

$$|T| > t(\alpha, r)$$

2-tail

3. Linear Regression

- $y = mx + b$
- Fit m and b, get r^2
  - r^2 is a measure of goodness of fit
  - Best fit as r^2 approaches 1
- Find confidence intervals for m and b
  - $m = 3.56 \pm 0.02$, etc.
  - Use standard error and t-statistic
  - Excel add-on (or Igor)
- Find **confidence intervals** for line
  - Use standard error around mean
  - Narrow waisted curves around line
  - Depends on n
  - Meaning: **How many ways can I draw a line through data**
- Find **prediction band** for line
  - Meaning: **Where are the bounds of where the data lie**

3. Linear Regression
(Confidence Interval)

- Confidence Interval for each $(X_0, \hat{Y}_0)$

$$\hat{Y}_0 \pm S_{\hat{Y}} t_{n-2,1-\alpha} \qquad \text{(using 2-tailed t table)}$$

$$S_{\hat{Y}} = \left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right)^{0.5} s_{yx}$$

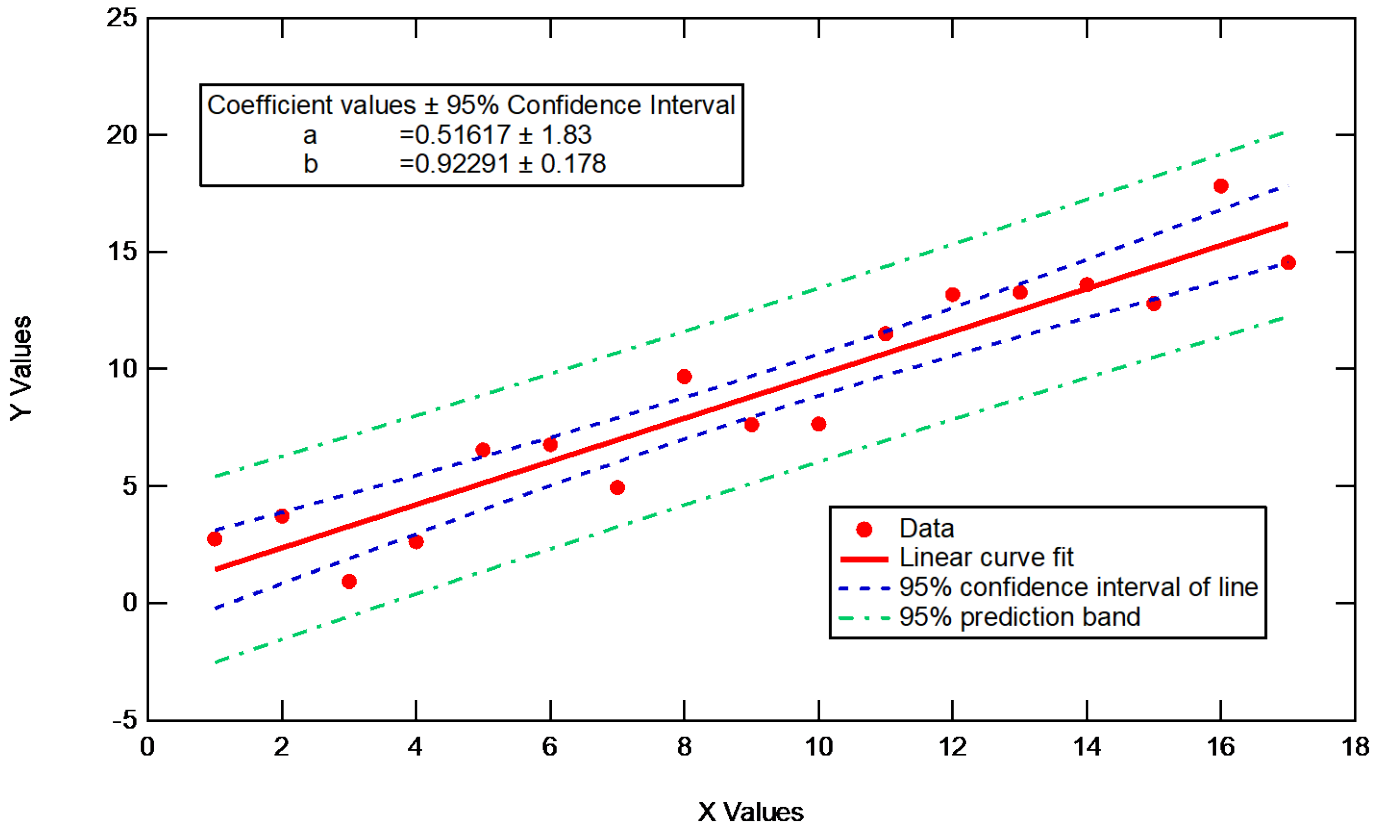
$$s_{yx} = \left(\frac{1}{n-2} \left[\sum_{i=1}^{n} (Y_i - \bar{Y})^2 - \frac{[\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})]^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right] \right)^{0.5}$$

- Prediction Band for each $(X_0, \hat{Y}_0)$

$$Y_0 = \hat{Y}_0 \pm t_{n-2,1-\alpha} \left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right)^{0.5} s_{yx}$$

- In Excel, s_{yx} can be solved using
  =STEYX(Ydata,Xdata)
- In Excel, $\sum_{i=1}^{n} (X_i - \bar{X})^2$ can be solved using
  =DEVSQ(Xdata)
- For 95% confidence interval with 15 data points, get t from
  =T.INV.2T(.05,13)

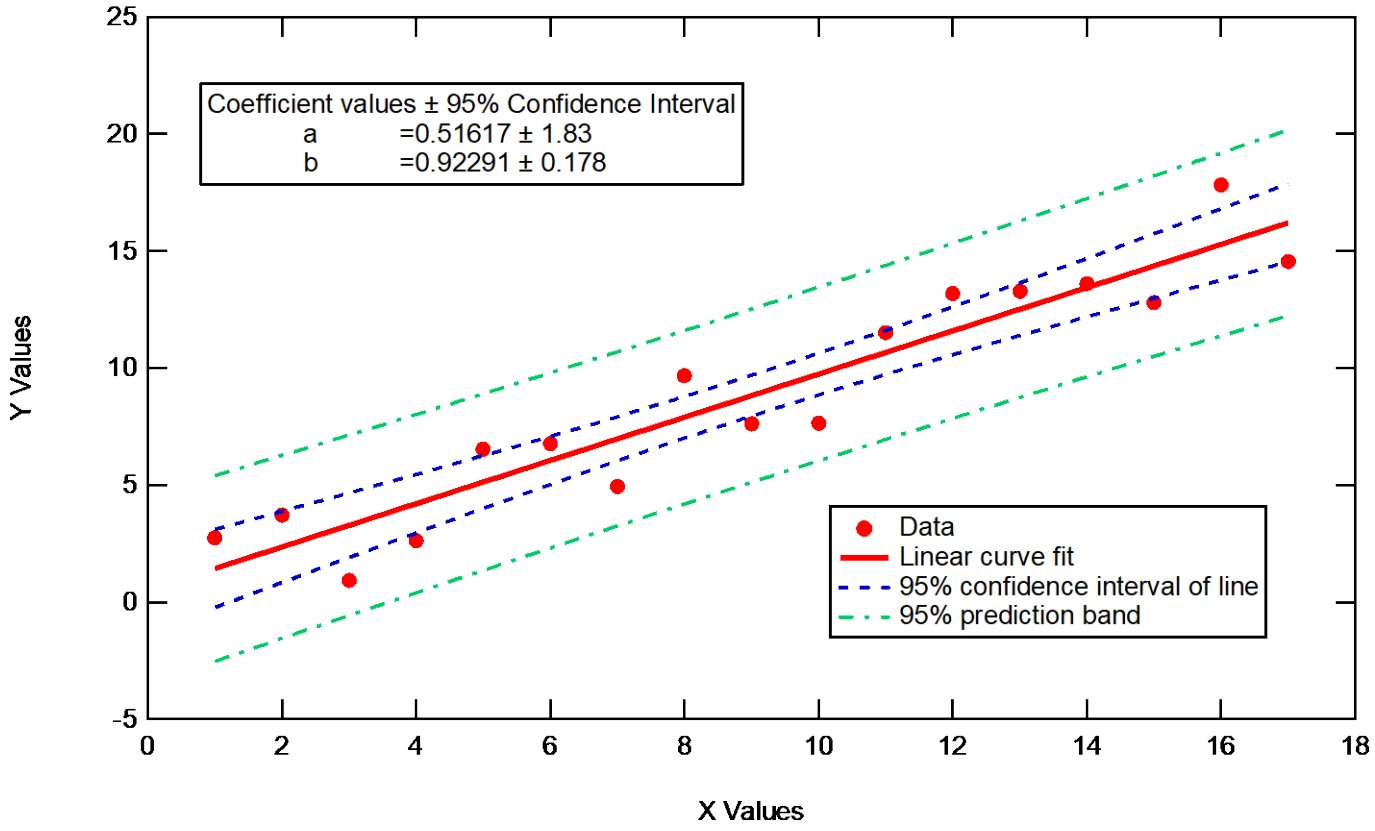
Example


Coefficient values ± 95% Confidence Interval
a =0.51617 ± 1.83
b =0.92291 ± 0.178
Data
Linear curve fit
95% confidence interval of line
95% prediction band
Y Values
X Values
25
20
15
10
5
0
-5
0
2
4
6
8
10
12
14
16
18

Confidence Intervals and Prediction Bands

- What good is the confidence interval for a line?
  - Shows how many ways the line can fit the points
  - Let's you state the confidence region for any predicted point
  - r^2 still helps determine how good the fit is
- What good is the prediction band?
  - Shows where the data should lie
  - Helps identify outlying data points to consider discarding

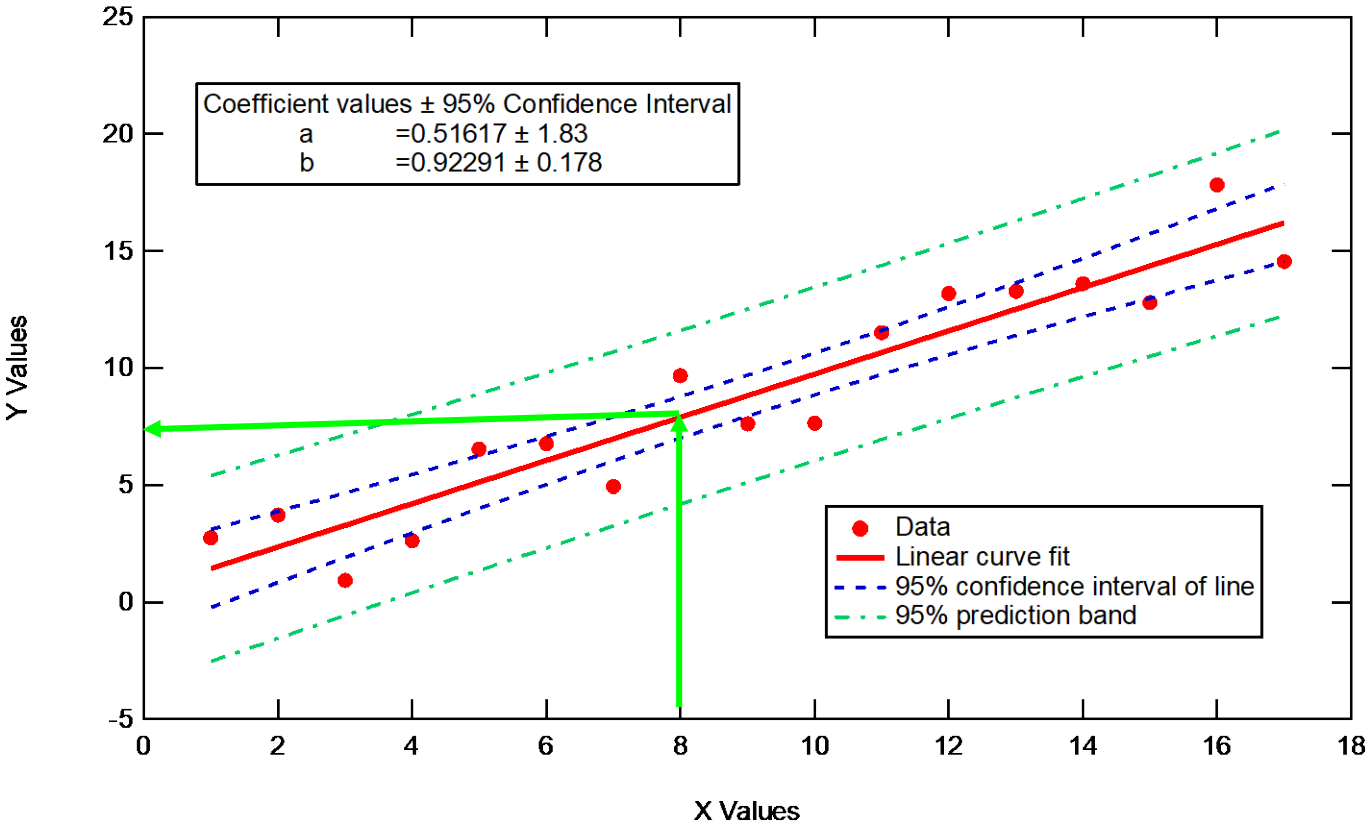
Example:
Here is a curve fit of some data



A. If the X value is 8.0, what is the predicted value of Y?
B. What is the confidence interval for that value of Y?

Solution to Part A
(predicted value if y when x = 8.0)

7.9



If you used Igor, then the equation for the line is y = a + bx

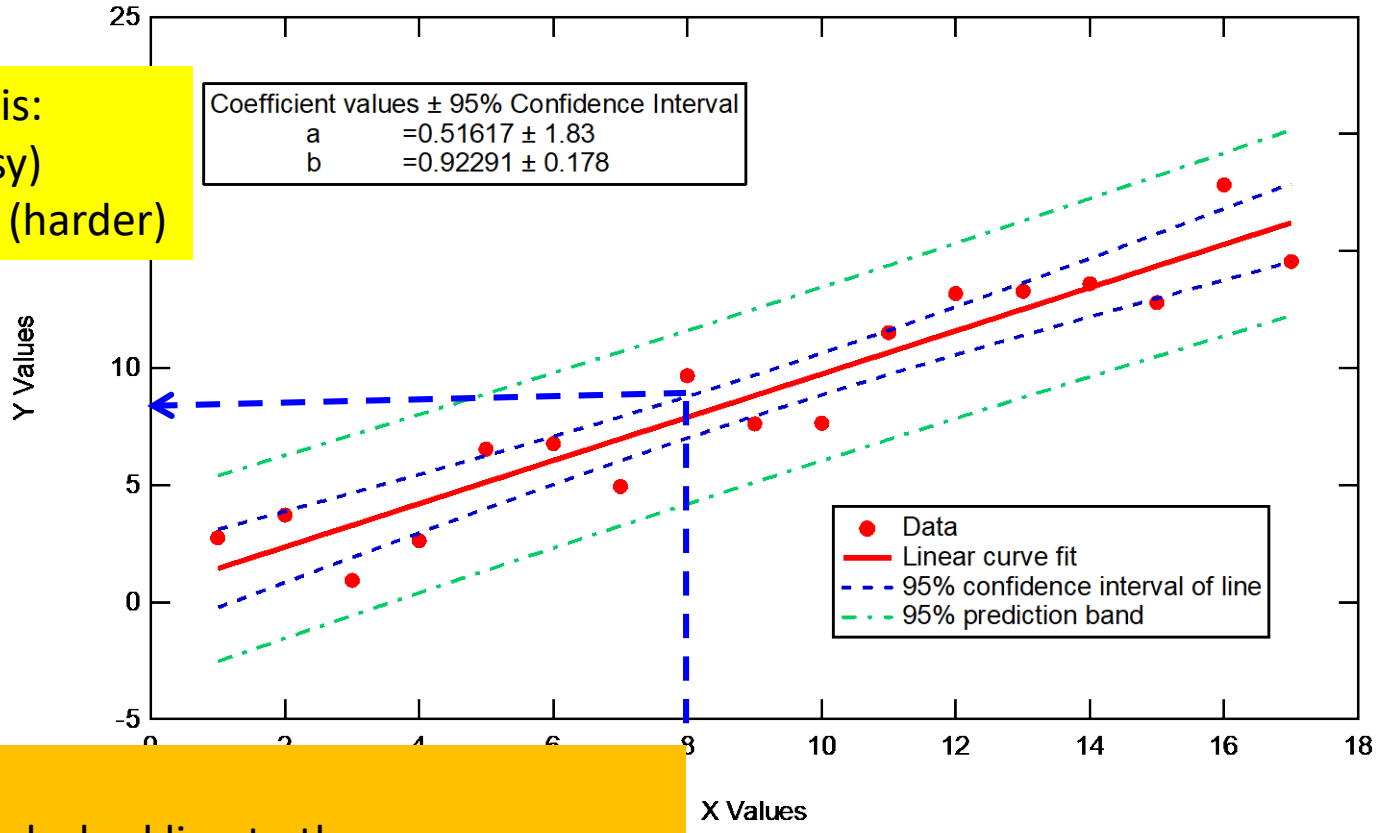
$Y(x=8.0) = 0.51617 + 0.92291 \times 8.0 = 7.8995$

The vertical green line on the graph shows the solution

Solution to Part B
(confidence interval for y when x = 8.0)

Two ways to do this:
1. Use graph (easy)
2. Use equations (harder)

8.786



Using graph,
1. Follow the blue dashed line to the upper confidence interval line and read the value
2. Read the corresponding y value from the left axis
3. Subtract the mean value found in part A from the upper confidence interval value

UC(8) = 8.786, so (8.786-7.8995) = ± 0.89

Obviously you cannot read this many significant figures from the graph!

Rules for Statistics Quiz

- Closed book, closed online notes
- You can bring one 8.5 $\times$ 11 paper with handwritten equations and notes (one-sided)
- You can use Excel and Igor
- Expected time: 15 minutes
- Maximum time: 30 minutes
- 20 points