Data Analysis III

CU- Boulder

CHEM-4181

Instrumental Analysis Laboratory

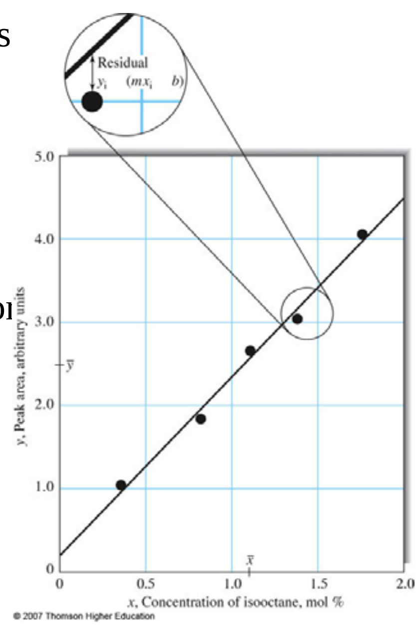
Prof. Jose-Luis Jimenez

Spring 2007

Lecture will be posted on course web page – based on lab manual, Skoog, web links

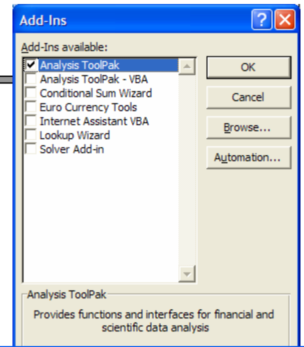
Linear Regression II

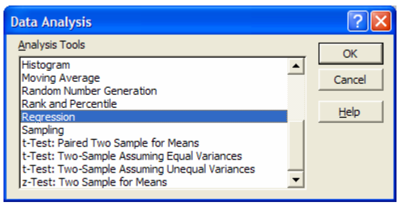
- Standard regression minimizes sum of squared residuals
  - Residual = vertical distance between datapoint and line
- Depending how much scatter there is in the data, the slope and intercept will have more or less error
  - $y = (m \pm s_m) * x + (b \pm s_b)$
  - Not displayed in simple regression in Excel
    - Only gives $y = m * x + b$
  - Need to used advanced reg.

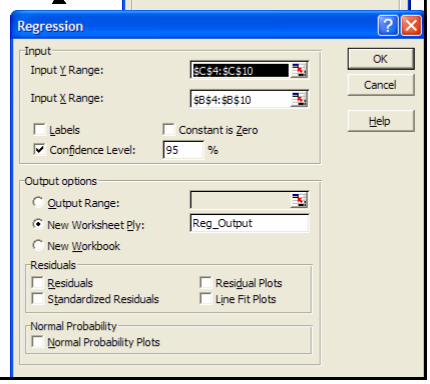


Linear Regression III

- In Excel
  - Tools Menu → Add-Ins
  - Tools Menu → Data Analysis
  - Select data and confidence level

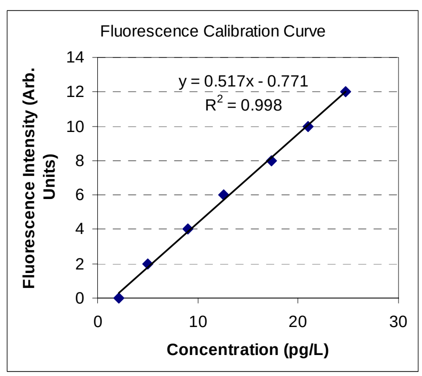






Linear Regression IV

- A wealth of information!
  - (Displayed with excessive sigfigs)



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.998879565
R Square	0.997760386
Adjusted R	0.997312463
Standard E	0.223980696
Observatio	7

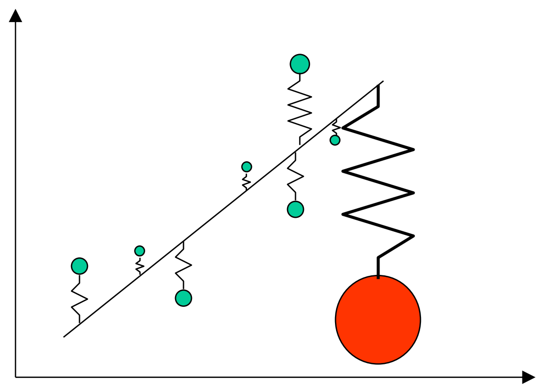
ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regression	1	111.7491632	111.7492	2227.528	8.07E-08
Residual	5	0.25083676	0.050167		
Total	6	112			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*Ipper 95.0%*
Intercept	-0.7711	0.1666	-4.629	0.006	-1.199	-0.343	-1.199	-0.343
X Variable	0.5169	0.0110	47.197	0.000	0.489	0.545	0.489	0.545

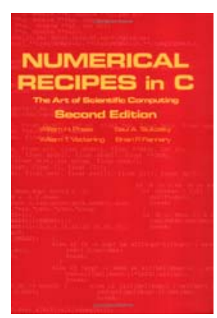
The Trouble w/ Standard Regression

- Every point pulls the line towards itself
  - With a weight equal to the squared residual
  - Noisy points, outliers, can seriously distort fit



Even More Complete Regression

- Nonparametric regression
  - Does not assume a distribution
    - Typical linear regression assumes no errors on X, Gaussian errors on Y
  - More robust in the presence of outliers
  - http://www.chem.uoa.gr/Applets/AppletTheil/Appl_Theil2.html
- Regression with errors in X and Y
- Weighted linear regression
  - Different points have more or less error
- Numerical recipes for explanations
  - Chapters 14 & 15
  - http://www.nr.com/
- Different regressions in many programs



Confidence Intervals

- In most situations μ cannot be determined
  - Can't afford to make lots and lots of measurements
  - We *will never know* the true value
  - Cannot make deterministic statements:
    - "Pb concentration is 4.7 ppb"
  - Can and need to make probabilistic statements
    - We can say "the probability that the Pb concentration is between 4.5 and 4.9 ppb is 95%"
    - Known as "confidence intervals"
      - Confidence: 95%
      - Interval: 4.5 to 4.9
        - Also expressed as 4.7 ± 0.2

Determining Confidence Intervals

- Width of interval is related to precision (s, σ)
  - If measurements are:
    - Highly precise: small interval
      - 0.482, 0.479, 0.488…
    - Very imprecise: large interval
      - 0.482, 0.310, 0.650…
- Confidence interval when σ is known
  - Just use the distrib. of $\bar{x}$, N($\bar{x}, \sigma_m$)
  - CI for $\mu = \bar{x} \pm \frac{z\sigma}{\sqrt{N}}$

TABLE a1-3 Confidence Levels for Various Values of z

Confidence Level, %	z
50	0.67
68	1.00
80	1.28
90	1.64
95	1.96
95.4	2.00
99	2.58
99.7	3.00
99.9	3.29

From Skoog

Size of CI vs. Number of Measurements

TABLE a1-4 Size of Confidence Interval as a Function of the Number of Measurements Averaged

From Skoog

Number of Measurements Averaged	Relative Size of Confidence Interval
1	1.00
2	0.71
3	0.58
4	0.50
5	0.45
6	0.41
10	0.32

$$\rightarrow as \ \frac{1}{\sqrt{N}}$$

- Greatest benefit with first few measurements, then diminishing returns

Example

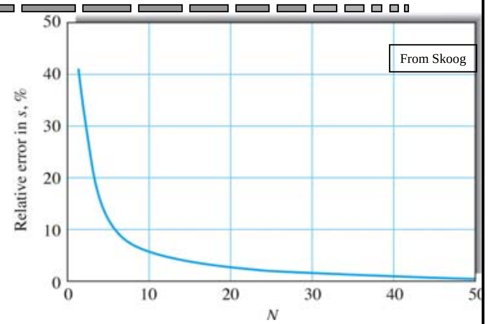
- From 10 measurements, we determine that the 68% CI of average glucose in the blood of CU students is 1100 ± 9 mg/L
  - Assuming that we have a good estimate of σ
- CQ: how many measurements do we need for the size of the 95% CI to be 4.5 mg/L?

A. 25
B. 100
C. 160
D. 225
E. I don't know

Which Confidence Interval to Report?

- Various confidence intervals
  $\pm 1\sigma$ (67%) $\pm 2\sigma$, 95% CI, 99% CI...
- You have to choose
  - Statistics doesn't answer this question, it depends on the value and use of the information
- E.g.
  - You are a chemist in a steel factory, analyzing for Mn (related to hardness). You add very expensive elements to steel based on this analysis. You get a raise based on how small the confidence interval is $\Rightarrow$ choose +/-s
  - If you are wrong, you are fired $\Rightarrow$ choose 99% CI
  - Uncertainty in temperature rise for a given increase of CO_2 emissions $\Rightarrow$ depends on evaluation of risks vs. costs

How to Estimate σ



- Perform preliminary experiments
  - Repeat exp. When developing method, just to estimate σ
  - E.g. COD, do one sample 15 times, then do other samples 3 times
- Pooling data

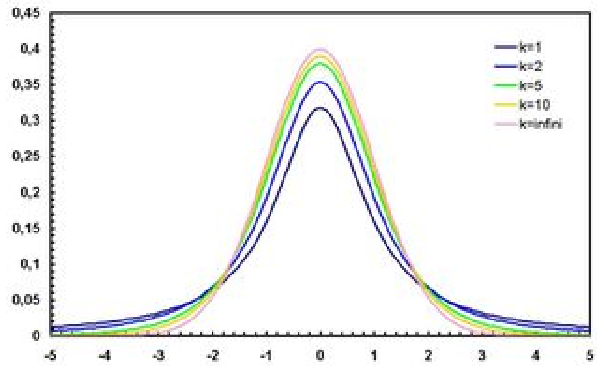
$$s_{pooled} = \sqrt{\frac{\sum_{i=1}^{N_1}(x_i - \bar{x}_1)^2 + \sum_{j=1}^{N_2}(x_i - \bar{x}_2)^2 + \ldots \sum_{p=1}^{N_{n_t}}(x_i - \bar{x}_{n_t})^2}{N_1 + N_2 + \ldots N_p - n_t}}$$

CIs when σ is not known

- Often we only have e.g. 3 measurements
  - More common situation
    - Limitation of time, of available sample, etc.
  - All we know about σ is s estimated from 3 meas.
    - Can be very uncertain
    - Confidence intervals will be LARGER
- In this situation, we will use t
  - For a single measurement → $t = \frac{x-\mu}{s}$
  - For the mean of N measurements → $t = \frac{\bar{x}-\mu}{s/\sqrt{N}}$
  - Look up in table, or use Excel
  - $t \to z$ as $N \to \infty$
  - Comparison: http://www.econtools.com/jevons/java/Graphics2D/tDist.html

Student's t vs Normal Distribution

- The t distribution has wider tails
  - We are less sure about CI, because we don't really know σ
- As N increases, we know more and more about σ, and $t \to N$



From Wikipedia
http://en.wikipedia.org/wiki/Student%27s_t_distribution (linked on class web page)

Table for Student's *t* Distribution

- TDIST(t,v,2) in Excel
  - t from previous page
  - v is degrees of freedom
    - = N-1
  - "2" means prob. of both tails
  - TDIST(1.89,2,2) = 20%
  - TDIST(2.36,7,2) = 5%
- Also TINV(prob, v)
  - TINV(0.2,2) = 1.89
  - TINV(0.05,7) = 2.36

From Skoog

TABLE a1-5 Values of t for Various Levels of Probability

Degrees of Freedom	80%	90%	95%	99%	99.9%
1	3.08	6.31	12.7	63.7	637
2	1.89	2.92	4.30	9.92	31.6
3	1.64	2.35	3.18	5.84	12.9
4	1.53	2.13	2.78	4.60	8.61
5	1.48	2.02	2.57	4.03	6.87
6	1.44	1.94	2.45	3.71	5.96
7	1.42	1.90	2.36	3.50	5.41
8	1.40	1.86	2.31	3.36	5.04
9	1.38	1.83	2.26	3.25	4.78
10	1.37	1.81	2.23	3.17	4.59
15	1.34	1.75	2.13	2.95	4.07
20	1.32	1.73	2.09	2.84	3.85
40	1.30	1.68	2.02	2.70	3.55
60	1.30	1.67	2.00	2.62	3.46
∞	1.28	1.64	1.96	2.58	3.29

Example

- Three measurements give
  - $\bar{x} = 1000$
  - $s = 17.3$
- CQ: The 99% CI for μ is:

A. 1000 ± 100
B. 1000 ± $17.3/\sqrt{2}$
C. 1000 ± 34.6
D. 1000 ± 50
E. I don't know

Outlier Rejection

- Is this data point reasonable?
  - It may seem too large or too small compared to the others
  - You CANNOT just remove it because "it looks wrong"
  - Use statistical test to check whether it can be rejected as an "outlier"
    - Include this in lab report
- Dixon's Q test
  - Q = gap / range
    - Gap: |outlier – next closest value|
    - Range: max – min
    - $Q > Q_{crit} \Rightarrow$ datapoint can be reject with 95% confidence

From p. 17-18 of lab manual

Outlier Rejection Example

- You've measured the following 8 values for Pb in soil (ppb):
  - 3.07 3.00 3.03 3.05 3.10 3.20 3.11 3.02
- CQ: Can you reject the 3.20 datapoint?

A. Yes

B. No

C. It depends

D. I don't know

N	Q_{crit} (CL:90%)	Q_{crit} (CL:95%)	Q_{crit} (CL:99%)
3	0.941	0.970	0.994
4	0.765	0.829	0.926
5	0.642	0.710	0.821
6	0.560	0.625	0.740
7	0.507	0.568	0.680
8	0.468	0.526	0.634
9	0.437	0.493	0.598
10	0.412	0.466	0.568

Table of critical values of Q