Error Bounds: Statistics within Algebra/Precalculus

NCSSM
February 7 & 8, 2003

Description

Would you like to integrate statistics and data analysis topics without "sacrificing" present curriculum? Teaching *error bounds* includes traditional skills such as linear, exponential, and logarithmic functions and their transformations and re-expressions. Two other data analysis topics that strengthen algebraic skills will be mentioned and discussed if time allows.

Albert Coons
Department of Mathematics
Buckingham Browne & Nichols School
Cambridge, MA 02183
al_coons@bbns.org
(617) 800-2264

The Need for Error Bounds

- Models are not perfect since they come from samples.
  - ActivStats[1] Demonstration

- Why The Least Squares Regression Model (LSRM) is not THE Model.
  - Debbie Amanti (BB&N '04) Distance versus Loft of Golf Irons Data for 150 golf balls hit with 10 different irons in random order.

- Note: *Calculating LSRM Model by Hand* investigation is wonderful algebra (point-slope, slope-intercept, residuals, identification of variables, squaring binomials, minimizing quadratics). See handout. For a complete generalize approach see [2].

- What is our goal?
  - To predict the value of the dependent variable (y) for a given independent variable (x) in light of the variation of the sample data.

- What type of estimate makes sense for this prediction?
  - Point Estimate or an Interval Estimate?

Observation of Variation

- How do we observe the variation of the sample data from its LSRM?
  - Residuals.
    - The residual of a particular data point =
      (actual y value) – (predicted y-value)
    - Note: the residuals sum to zero.
  - Residual Plot

Quantification of Variation

- What types of intervals estimates make sense?
  - +/- the residual of largest magnitude?
  - +/- the average of the residuals?
  - +/- the average absolute value of the residuals?
  - Standard deviation of the residuals?
    - A measure of the variability of the residuals.
    - Sample Standard Deviation, S_x:

      If $\bar{x} = \frac{\sum x_i}{n}$ then $s_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$
    - With certain assumptions, 95% of the residuals will fall within 2 standard deviations on either side of the average of the residuals, i.e. 0.
    - Check this on the golf data residuals.

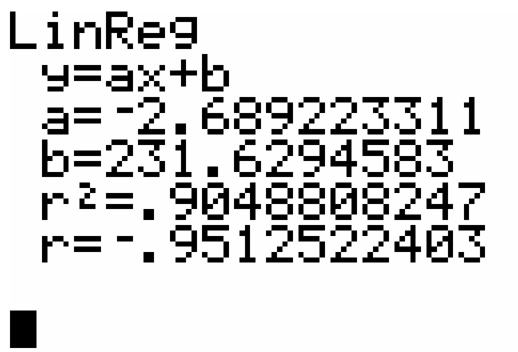
Error Bounds for LSR Model

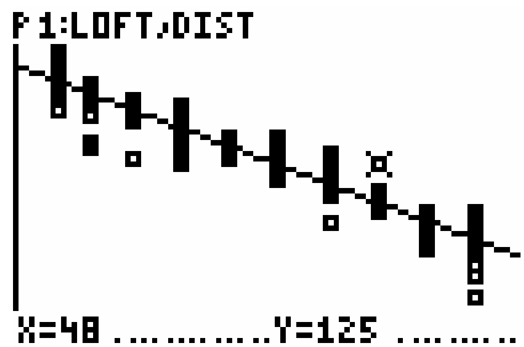
- Caveat. Statisticians have **more formal, more correct, and much more complicated** statistical methods for producing error bounds. They are call *prediction intervals.* (See [3])

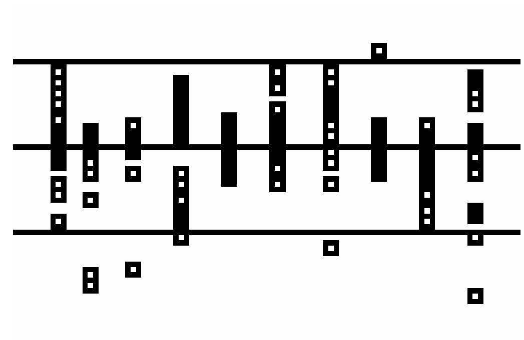
"Our primary purpose in doing error bounds on re-expressed data was to foreshadow the way statisticians work. Granted, error bounds as we do them are not a standard statistical tool ... but we wanted to be as true to the real methods as we felt was reasonable in a course that many students take as pre-statistics rather than pre-calculus." Gloria Barrett, NCSSM, email 2/4/03 (See [4])

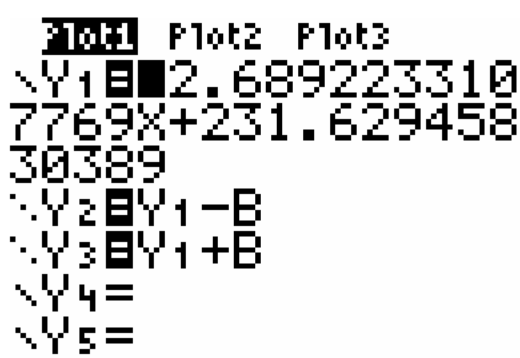
- The following simplified methods serve as a sensible introduction to the topic:
  - Compute error bounds for the golf data as LRSM +/- 2 stdev(residuals).
  - Make predictions from the Error Bounds - use value/trace.

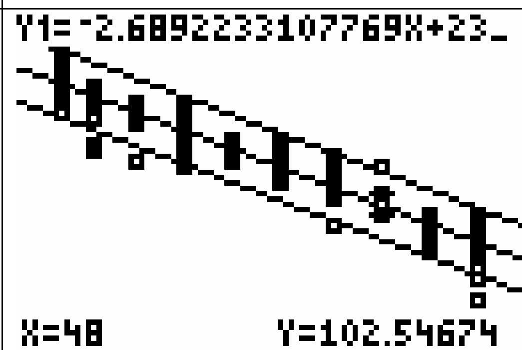
Error Bounds on Apparently Linear Data
Debbie Amanti (BB&N ‘04) Distance versus Loft of Golf Iron





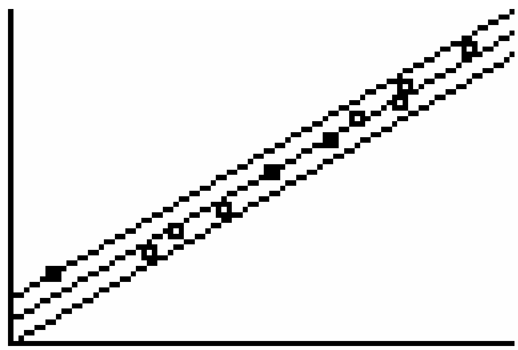






Error Bounds: An In Class Investigation

- Give *Error Bounds for Linear Models & Reasonable Predictions* Investigation
- Show pictures of circular objects.
- Results



Error Bounds in Re-expressed Data

What do error bounds look like if the data is non-linear?

- View TVLIFE data with a Power Model with error bounds. Data represents 22 countries n 1993. Source: Workshop Statistics (Angola: people/TV = 200; Haiti: people/TV = 234)
- Any concerns?
- Look at the residuals of the model. Any observations?
- Solution? Add the error bounds to the re-expressed model.
- Do the mathematics by hand.

Computing Error Bounds for Power Model for TVLIFE Data

$$y_{\text{power model}} = 82.35394597 x^{-.0931221773}$$

Log – Log Re - Expression

$$\ln y \approx \ln 82.3x^{-.09} \approx \ln 82.3 + \ln x^{-.09}$$

$$\ln y \approx \ln 82.3 - .09 \ln x$$

Check above with LSRM from TI.

View Re - expressed Residuals.

Compute Error Bounds

$$\text{Determine } 2 * \text{stdev}(\text{residuals}_{\text{LSRM}}) = .1443071017$$

$$\ln y \approx \ln 82.3 - .09 \ln x \pm .144$$

$$\ln y \approx \ln 82.3 \pm .144 - .09 \ln x$$

$$y \approx e^{\ln 82.3 \pm .144 - .09 \ln x} \approx e^{\ln 82.3} e^{\pm .144} e^{\ln x^{-.09}}$$

$$y \approx 82.3 e^{\pm .144} x^{-.09} \approx e^{\pm .144} \left(82.3 x^{-.09}\right)$$

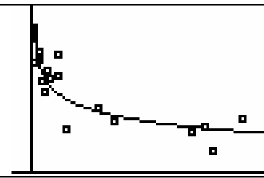
$$y_{lower} \approx e^{-.1443071017} y_{\text{power model}} = .8656 y_{\text{power model}}$$

$$y_{upper} \approx e^{+.1443071017} y_{\text{power model}} = 1.155 y_{\text{power model}}$$

Plot the EBs along with the previous EBs.

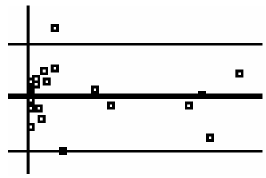
Error Bounds on a Power Model
Data from <u>Workshop Statistics</u>. Alan Rossman.

```
PwrReg
 y=a*x^b
 a=82.35394597
 b=-.0931221773
```

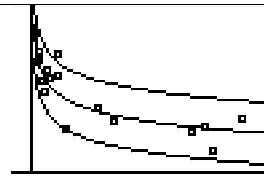


```
2stdDev(ʟRESID)→
B
     8.698625762

 Plot1  Plot2  Plot3
\Y1=■2.353945974
695X^-.093122177
28623
\Y2=-B
\Y3=B
```

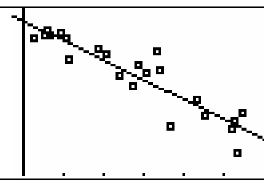


```
 Plot1  Plot2  Plot3
\Y1=■2.353945974
695X^-.093122177
28623
\Y2=Y1-B
\Y3=Y1+B
\Y4=
\Y5=
```

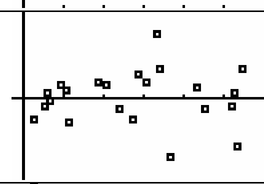


Re-expressed the data and then form. Error Bounds

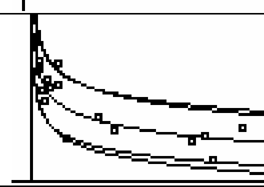
```
LinReg
 y=ax+b
 a=-.0931221773
 b=4.411026373
```



```
2stdDev(ʟRESID)→
L
     .1443071017
```



```
 Plot1  Plot2  Plot3
\Y1=■2.353945974
695X^-.093122177
28623
\Y2=Y1-B
\Y3=Y1+B
\Y4=e^(L)Y1
\Y5=e^(-L)Y1
```



Note how the Error Bounds from the re-expressed data (Y5) never are < 0 while those from the power model (Y2) are.

X	Y2	Y5
50	48.511	49.522
100	44.935	46.427
150	42.948	44.706
200	41.583	43.525
500	37.47	39.965
5000	28.56	32.252
1E99	-8.699	4.3E-8

X=1E99

Prove the Following Generalizations:

Letting s = stdev(residuals$_{\text{re-expressed}}$)

Model	Equation	Show its Error Bounds Are
Power	$y = ax^m$	$y = e^{\pm 2s} ax^m$
Exponential	$y = ae^x$	$y = e^{\pm 2s} ae^x$
Either	$y = ax^m$ or $y = ae^x$	$y = e^{\pm 2s}\left(y_{\text{original-model}}\right)$

Answer: Generalize Errors Bounds for Power Model

$$y_{power-model} = ax^m$$

$$\ln(y) = \ln(ax^m)$$

Let s = stdev(residuals$_{\text{re-expressed}}$)

$$\ln(y_{error-bounds}) = \ln(ax^m) \pm 2s$$

$$y_{error-bounds} = e^{\ln(ax^m) \pm 2s}$$

$$y_{error-bounds} = e^{\ln(ax^m)} e^{\pm 2s}$$

$$y_{error-bounds} = ax^m e^{\pm 2s} = e^{\pm 2s} ax^m$$

$$y_{error-bounds} = e^{\pm 2s}\left(y_{power-model}\right)$$

Challenge questions

- Does working in base 10 rather than base e make any difference? If so, what difference does it make? What nuance should be noted in the general proof?

- Can you come up with generalized formulas for error bounds for other types of models such as quadratics, reciprocals, square roots...?

- If we are adding and subtracting twice the standard deviation of the re-expressed residuals to the re-expressed model, why do we end up multiply the original models by a constant?

- We multiply the original models by a constant to create each error bound. Why are the error bounds converging if these constants do not depend on x?

---

[1] ActivStats. Paul Velleman. Addison Wesley.

[2] Contemporary Precalculus Through Applications. 2nd Edition. The North Carolina School of Science and Mathematics. Everyday Learning Corporation. p 545-5.

[3] For an introduction to Prediction Intervals see: Introduction to Statistics and Data Analysis. Peck, Olsen, Devore. Duxbury. P 673-678.

[4] This presentation was motivated by material developed in: Contemporary Precalculus Through Applications. 2nd Edition. The North Carolina School of Science and Mathematics. Everyday Learning Corporation. Section 1.8: Error Bounds and the Accuracy of Prediction. Section 3.14: Error Bounds in Re-expressed Data