***Highlights / Key Concepts in NKNW4 Chapter 4***

## 4.1 Joint C.I.'s for $\beta_0$ & $\beta_1$

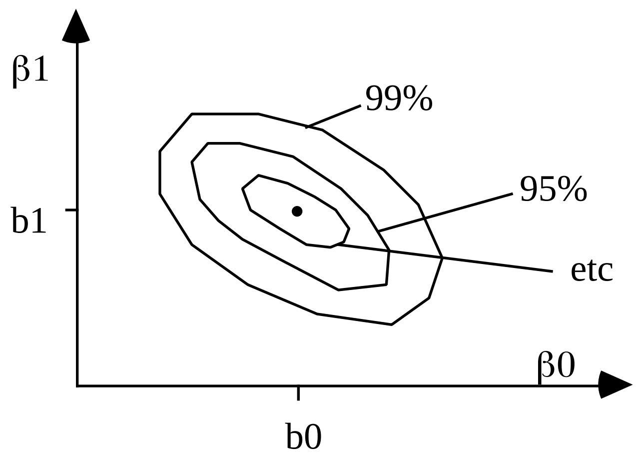
Conservative: For confidence 1-α for Joint Interval...

use (1-α/2) C.I. for each one

e.g. if want 95% CI's for each, use 97.5% CI's for each (true coverage >= 1-α)

More realistic ***(not mentioned in text)***: use Confidence Eclipse

From covar($b_0$,$b_1$), given by software, can calculate:




## 4.2 Simultaneous Interval Estimates of E(Y | $\underline{\underline{X}}_1$), E(Y| $\underline{\underline{X}}_2$), ... ***(means)***

1. Use confidence band for line .. with W= rather than $t_{n-2}$ [ $W^2 = 2F_{(1-\alpha;\, 2,\, n-2)}$ ],

   i.e. $\hat{Y}$ +/- **W** × SE( $\hat{Y}$ ) rather than $\hat{Y}$ +/- **t** × SE( $\hat{Y}$ )

   "W" -- named after Working-Hotelling -- uses one "universal" W, designed for a correct CI for true regression line from X = -∞ to X = +∞ , no matter how many/few X values one is actually interested in.

   Remember **t** = $\sqrt{F_{(1-\alpha;\ \mathbf{1},\ n-2)}}$ , whereas **W** = $\sqrt{2}\sqrt{F_{(1-\alpha;\ \mathbf{2},\ n-2)}}$ ,

2. Use Bonferroni procedure i.e. use g CI's, each with confidence level 1-α/g

   *See comments on section 4.23 , pages 157-158.*

## 4.3 Simultaneous Prediction Intervals for "New" Observations

***[ estimates of*** $Y\,|\,\underline{\underline{X}}_{h1}$ , $Y\,|\,\underline{\underline{X}}_{h2}$ , $Y\,|\,\underline{\underline{X}}_{h3}$ ***... ( individual Y's ) ]***

use $\hat{Y}$ +/- **multiple** × SE( $\hat{Y}$ ): **multiple** based on F (Scheffé) or t (Bonferroni) [both incorporate # of X's in question]

**4.4 Regression through Origin:** $Y \mid X = \beta_1 X + \varepsilon$ ; $E(Y \mid X) = \beta_1 X$ ;

$Y \mid X \sim ??(\beta_1 X, \sigma^2)$ for Least Squares (LS) Estimation

$Y \mid X \sim \text{Gaussian}(\beta_1 X, \sigma^2)$ (or Central Limit Theorem) for t-based inferences

<u>**LS estimates**</u> { or ML under Gaussian $\varepsilon$ 's }

$$b_1 = \hat{\beta}_1 = \frac{\Sigma xy}{\Sigma x^2}$$

$$= \frac{\Sigma \frac{y}{x} x^2}{\Sigma x^2} = \frac{\Sigma \text{ slope} \times x^2}{\Sigma x^2}$$

$$= \frac{\Sigma \text{ slope} \times \text{weight}}{\Sigma \text{ weight}} \quad \text{with slope} = \frac{y}{x}, \ \text{weight} = x^2 .$$

$$\text{Var}(\hat{\text{slope}}_i) = \text{Var}\left(\frac{y_i}{x_i}\right) = \frac{\text{Var}[y_i]}{x_i^2} = \frac{\sigma^2}{x_i^2} .$$

$$\hat{\sigma}^2 = \text{MSE} = \frac{\Sigma e^2}{n-\mathbf{1}}$$

(note: n-**<u>1</u>** free e's)

(**<u>1</u>** constraint: $\Sigma x.e = 0$.)

$b_1$ is a **weighted average of individual slope estimates,**
with weights that are ***inversely proportional to their variance***s, i.e.,

$$\text{weight}_i \quad \frac{x_i^2}{\sigma^2} \quad x_i^2.$$

$$\textbf{Var}(b_1) = \frac{\sigma^2}{\Sigma x_i^2} = \frac{\sigma^2}{n \times \text{average}[x^2]} \quad \text{so} \quad SE(b_1) = \frac{\text{RMSE}}{\sqrt{n} \times \sqrt{\text{average}[x^2]}}$$

**Cautions**

Formula for $b_1$ is **different** from one for $E(Y \mid X) = \beta_0 + \beta_1 X$ model

(force $\beta_0$ to 0 <u>before</u> estimating $\beta_1$ )

**careful regarding $r^2$**

(see pp 163...)

**To $\beta_0$ = 0 or NOT TO $\beta_0$ =0 ??**

***Do not force line through intercept unless very clear physical model to justify it ...***

***usually, one does not loose much by allowing a non-zero incercept when in fact it is zero.***

***<u>A more serious issue</u> is whether here (and also in the non-zero intercept model) the assumption of the constant variance*** $\sigma^2$ in the $Y \mid X \sim ??(\text{function of } X, \sigma^2)$ ***model is appropriate.***

***For example, if in the zero-intercept model, the variance is proportional to X, then the slope estimator*** $b_1 = \Sigma y / \Sigma x = \Sigma x(y/x) / \Sigma x$ ( i.e., weight of x for individual slope estimate y/x) ***is more efficient than the (also unbiased) estimator derived from the constant variance model above.***

## 4.5 Measurement Errors and their effects

**a) Measurement Errors in Y**

They get absorbed into residuals

$$Y = \beta_0 + \beta_1 X + \varepsilon + \varepsilon_m$$

$\varepsilon$ biologic/real/unexplained

$\varepsilon_m$ measurement error

$$\text{var}(Y \mid X) = \sigma_{\varepsilon}^2 + \sigma_{\varepsilon_m}^2$$

Can average several (k) measurements on same individual to reduce effect of measurement error

$$\text{var}(Y|X) = \sigma_{\varepsilon}^2 + \frac{\sigma_{\varepsilon_m}^2}{k} .$$

**b) Measurement Errors in X**

X real/"true" X

X* observed/recorded value

**2 situations ( difference is quite subtle!!)**

- ***"Classical" Error Model***

  $X^* = X + \delta$

  (X,Y[X]) chosen but (X*,Y[X]) recorded; $E[\delta] = 0$; $\delta$ uncorrelated with X

  so that $\text{var}(X^*) = \text{var}(X) + \text{Var}(\delta)$ #

- ***"Berkson" Error Model***

  $X^* = X + \delta$

  (X*,Y[X*]) targetted but (X*,Y[X]) recorded; $E[\delta] = 0$; $\delta$ uncorrelated with X*

  (but necessarily correlated with X:- if told $\delta$, would know X)

# Interpreting Var(X) as <u>observed</u> var; Var ($\delta$) in sampling variance (repeatable) sense.

## 4.5 Measurement Errors ... b) Measurement Errors in X ...

***-"Classical" Error Model***

**True regression model :** $Y = \beta_0 + \beta_1 X + \varepsilon$

**BUT the "X" values we record are not correct . i.e.**

although X generated Y, we record it as $X^* = X + \delta$

X: true value ; $E[\delta] = 0$; $\delta$ uncorrelated with X

**If use naive LS estimator $b_1$ to estimate $\beta_1$ from the X*'s ...**

**then $b_1$ biased towards null (zero) ("ATTENUATION")**

$$E[b_1] = \beta_1 \frac{var(X)}{var(X^*)} = \frac{var(X)}{var(X) + var(\delta)} < \beta_1 \text{ if } var(\delta) > 0.$$

$$\frac{var(X)}{var(X^*)} = \frac{var(X)}{var(X) + var(\delta)} = \frac{\text{variation in "true" X values}}{\text{variation in observed values}} \leq 1$$

**alias:** "Intra-Class Correlation Coefficient" or "Reliability Coefficient"

**is the "ATTENUATION" factor**

***If pilot studies or literature can furnish an estimate of ICC...***

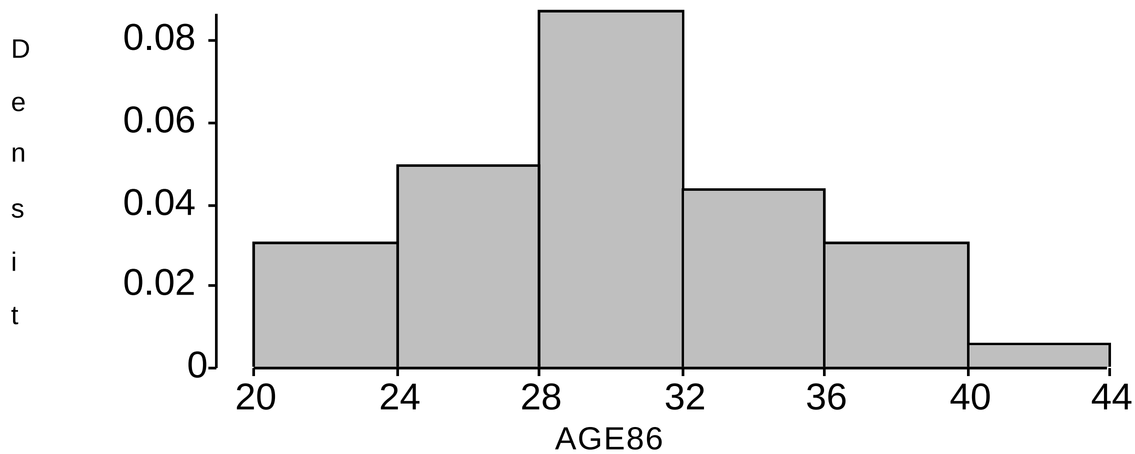
***one can DE-ATTENUATE:***

**"bias-corrected" estimator of $\beta_1$ :** $b_{1[LS]} \times \frac{1}{ICC}$

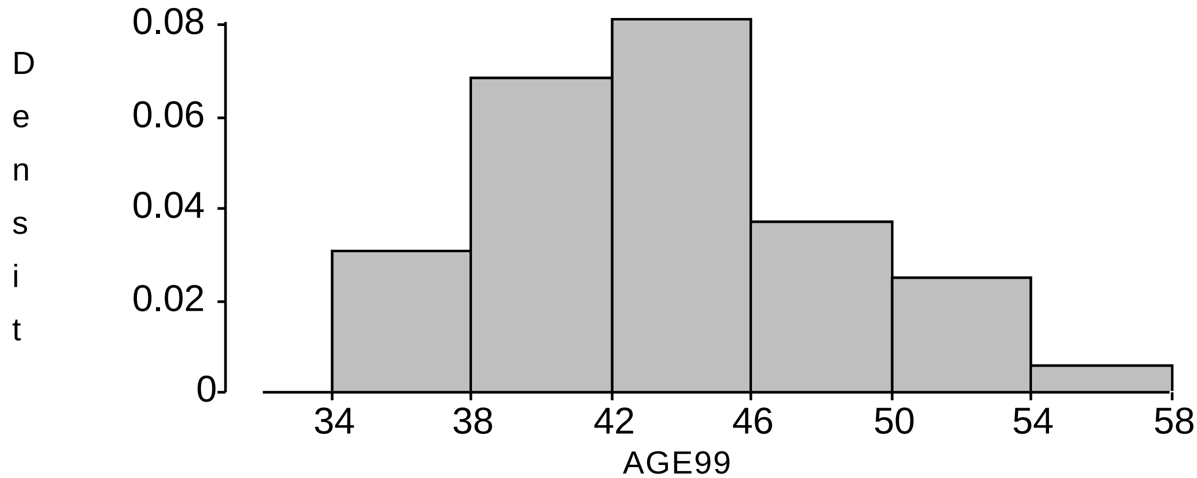
## EXAMPLE OF "FLAT" SLOPE ("classical" measurement error model)

## Ages of 40 students in 1986 class 513-607 (Inferential Statistics)

```
DATA ages;  keep age86 age86___ age99;
  INPUT Age86 @@;                      /* @@ : multiple observations on 1 line */
  age99 = Age86+13;

  b =    int(ranuni(7534567)+0.5) ;  /* b    ~ Bernoulli( 0,1), prob 0.5 each */
  sign = 2 * b - 1;                  /* sign ~ Bernoulli(-1,1)  prob 0.5 each */
  d =    sign * 5    ;               /* d    ~ Bernoulli(-5,5)  prob 0.5 each */

  age86___ = Age86 + d;

  LINES;

  22 22 22 22     23     25     26 26 26     27 27 27 27     28 28 28
  29 29     30 30 30 30 30     31 31 31 31     32 32     33 33     34 34
  35     36     37     38 38     39     42
;
```



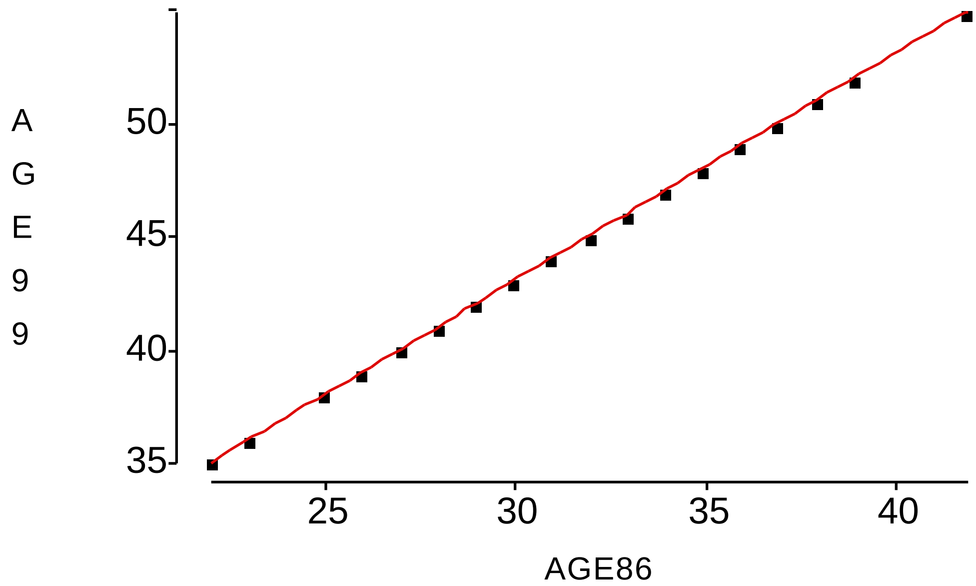

## These 40 students 13 years later ... in 1999



**How much, and at what rate, did they age in these 13 years?**

| AGE99 | = | AGE86 |
|---|---|---|
| Response Distribution: | | Normal |
| Link Function: | | Identity |

| Model Equation | | | | |
|---|---|---|---|---|
| AGE99 = | 13 | + | 1.0 | AGE86 |



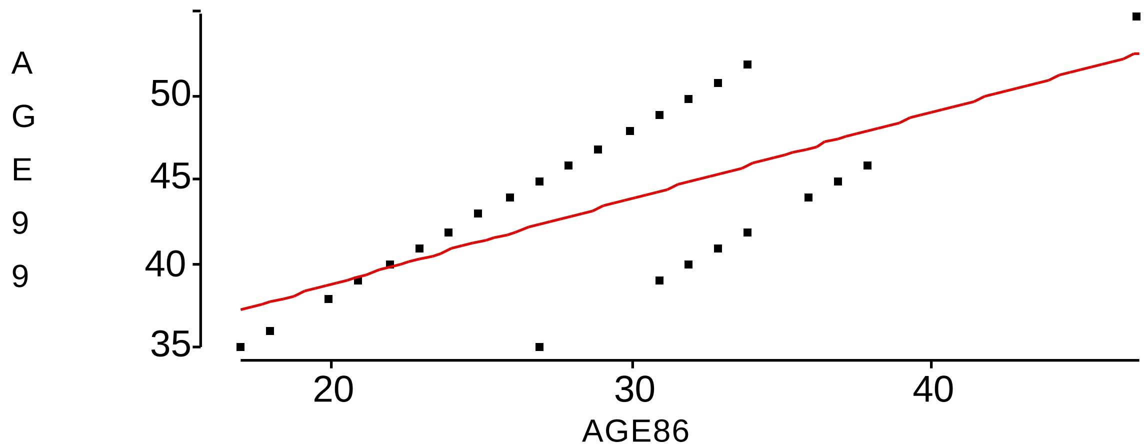

**What if these 40 students had given their ages as true age +/- 5 years (with the + or - determined at random, without regard to true age)?**

**Age86___ = Age86 +/- 5**

| | Age86 | Age99 | Age86___ |
|---|---|---|---|
| Mean | 30.0 | 43.0 | 28.5 |
| Std Dev | 4.9 | 4.9 | 6.5 |
| Variance | 24.4 | 24.4 | 41.9 |
| Minimum | 22 | 35 | 17 |
| Maximum | 42 | 55 | 47 |

**change__ = Age99 - Age86___;**

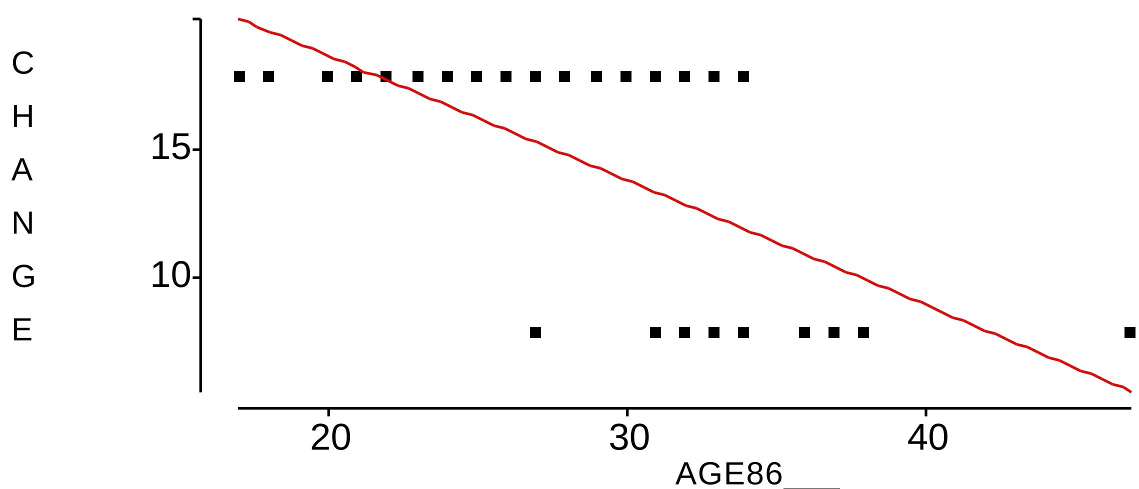
AGE99 = 28.39 + 0.51 AGE86___




| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| Source | DF | Sum of Squares | Mean Square | F Stat | Prob > F |
| Model | 1.00 | 430.76 | 430.76 | 31.35 | 0.0001 |
| Error | 38.00 | 522.21 | 13.74 | | |
| C Total | 39.00 | 952.98 | | | |

==========================================================

| Model Equation | | | | | |
|---|---|---|---|---|---|
| CHANGE__ | = | 28.39 | - | 0.49 | AGE86___ |



## 4.5 Measurement Errors ... b) Measurement Errors in X ...

***-"Berkson" Error Model***

**True regression model :** $Y = \beta_0 + \beta_1 X + \varepsilon$

**<u>BUT</u> the "X" values we record are not correct . i.e.**

we targetted (and recorded) X* (e.g. thermometer set to X* = 22 C)
but actual X is different from targetted/recorded X*
i.e. true value $X = X^* + \delta$ ; $E[\delta] = 0$; $\delta$ uncorrelated with X*

> **If use naive LS estimator $b_1$ to estimate $\beta_1$ from the X*'s ...**
>
> **then $b_1$ un<u>biased</u>**

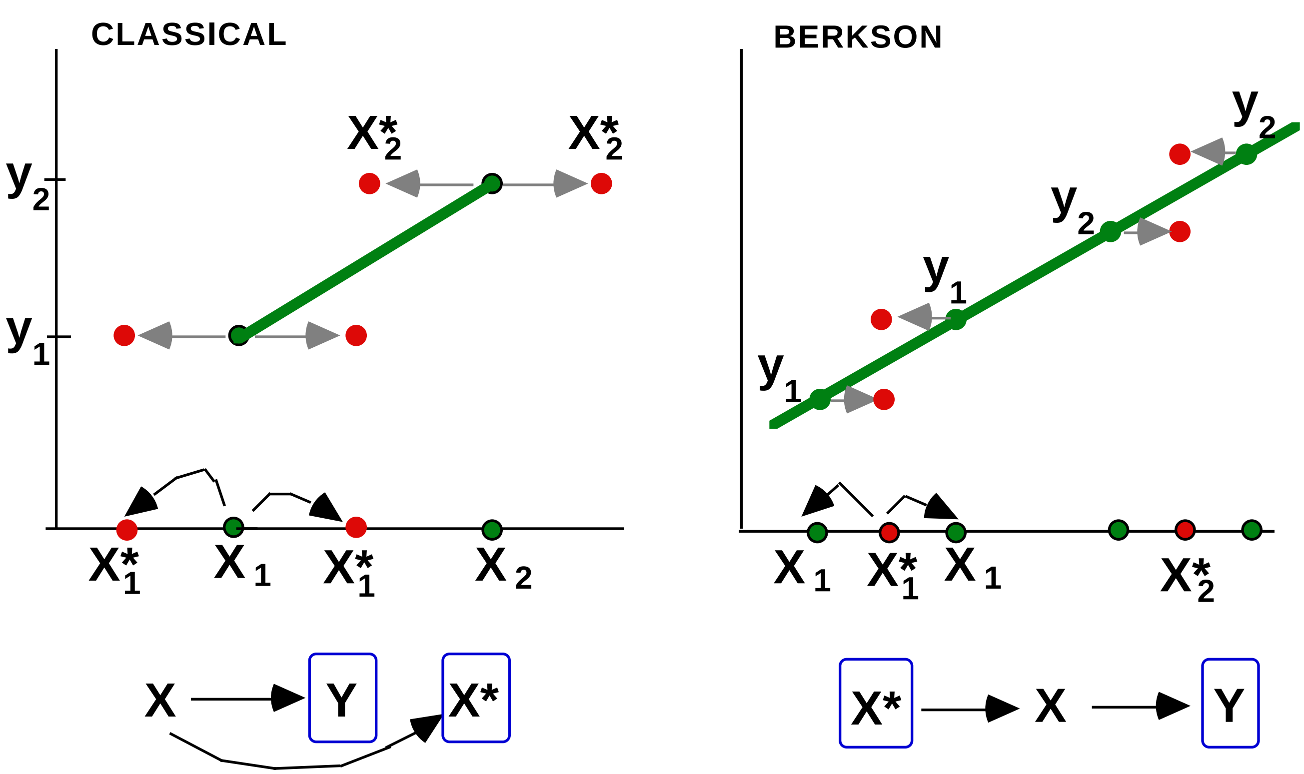
## The "Classical" vs. "Berkson" difference ...

**Assume**

- **No Biologic Variation ( i.e. all $\varepsilon$ 's = 0)**

  **i.e. $Y = \beta_0 + \beta_0 X + 0$**

- **2-point regression $(x^*_1, y_1)$ and $(x^*_2, y_2)$**



**Without loss of generality, assume $\beta_0 = 0$ and $\sigma^2(\varepsilon)=0$**

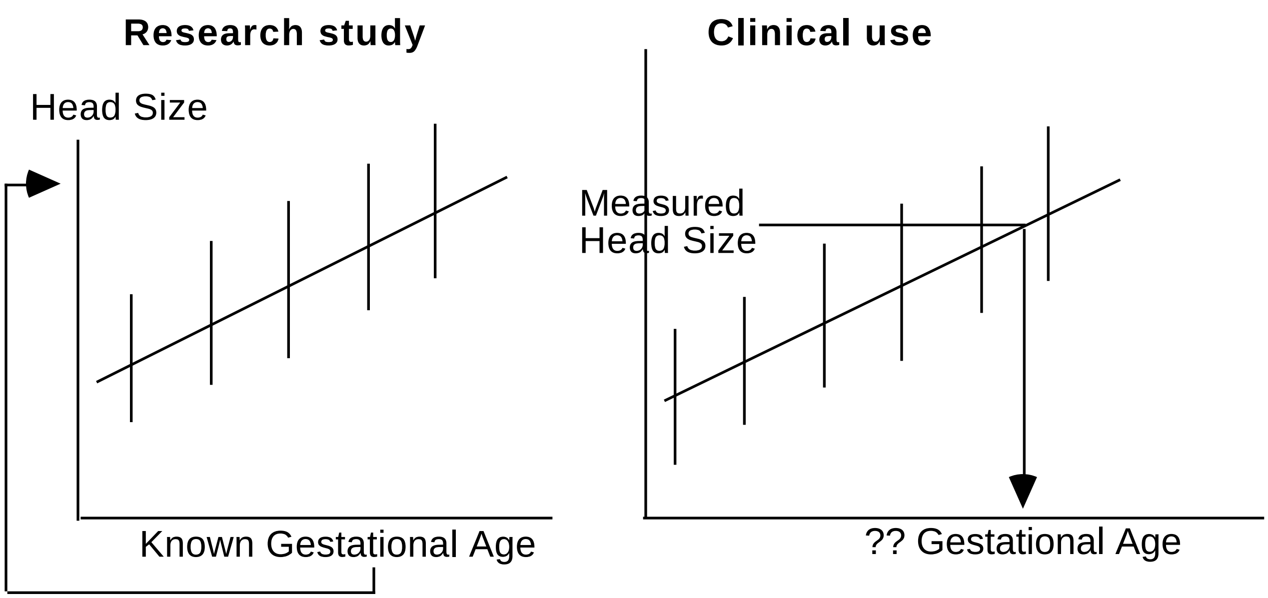
| *"Classical" Error Model* | *"Berkson" Error Model* |
|---|---|
| $\dfrac{y_2 - y_1}{x^*_2 - x^*_1}$ | $\dfrac{y_2 - y_1}{x^*_2 - x^*_1}$ |
| $\dfrac{\beta\{x_2 - x_1\}}{[x_2 + \delta_2] - [x_1 + \delta_1]}$ | $\dfrac{\beta\{x^*_2 + \delta_2\} - \beta\{x^*_1 + \delta_1\}}{x^*_2 - x^*_1}$ |
| $\dfrac{\beta\{x_2 - x_1\}}{[x_2 - x_1] + [\delta_2 - \delta_1]}$ | $\dfrac{\beta\{x^*_2 - x^*_1\} + \beta\{\delta_2 - \delta_1\}}{x^*_2 - x^*_1}$ |
| $\dfrac{\beta}{1 + \dfrac{\delta_2 - \delta_1}{x_2 - x_1}}$ | $\beta\left(1 + \dfrac{\delta_2 - \delta_1}{x^*_2 - x^*_1}\right)$ |
| ***random component*** $\delta_2 - \delta_1$ ***is in denominator*** | ***random component*** $\delta_2 - \delta_1$ ***is in numerator*** |

**Replacing subjects' ages (X) with X* = average age for subjects in an age category, generates Berkson type measurement errors.**

## 4.6 Inverse Predictions (Use of regression for "calibration": see comments p 169)

***Example:***

***Estimation of Gestational Age from Ultrasound Measurements of Fetal Head Size***




n (X,Y) pairs
with known X's
==> ($b_0$, $b_1$, MSE, $X_{bar}$)

$$Y_h \implies \hat{X}_h = \frac{Y_h - b_0}{b_1}$$

Exact Var( $\hat{X}_h$ ) ???

$$\hat{X}_h = \frac{Y_h - RV_0}{RV_1}$$

$$\text{(Approx) est. of Var}(\hat{X}_h) : \frac{MSE}{b_1^2}\left[1 + \frac{1}{n} + \frac{(\hat{X}_h - X_{bar})^2}{\Sigma(X - X_{bar})^2}\right]$$

## 4.7 Choice of X levels

***Well explained in book, pp 169-170***

***Would simply emphasize a different way of viewing the terms***

$$\frac{\sigma^2}{\Sigma(X - X_{bar})^2} ,$$

$$\frac{1}{n} + \frac{(X_h - X_{bar})^2}{\Sigma(X - X_{bar})^2} , \text{ etc}$$

***namely***

$$\frac{\sigma^2}{n \text{ Var}(X)} ,$$

$$\frac{1}{n} + \frac{(X_h - X_{bar})^2}{n \text{ Var}(X)} , \text{ etc}$$

**This way, for example, $SD(b_1) = \frac{\sigma}{\sqrt{n}\, SD(X)}$**

**Here, don't fuss about Var(X) being defined with divisor of n vs. n-1.**
**If we have the choice of which X's to study, we are using our defintion of variance, namely**

$$\text{"Var"}(X) = \frac{1}{n} \Sigma(X - X_{bar})^2$$

**as a measure of the spread of the chosen X's.**