4.5 Measurement Errors and their effects

a) Measurement Errors in Y

They get absorbed into residuals

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

ε	biologic/real/unexplained
ε_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; $E[\delta] = 0$; δ 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; $E[\delta] = 0$; δ uncorrelated with X*

  (but necessarily correlated with X:- if told δ, would know X)

Interpreting Var(X) as <u>observed</u> var; Var (δ) 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 ; $E[\delta] = 0$; δ uncorrelated with X

If use naive LS estimator b_1 to estimate β_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...

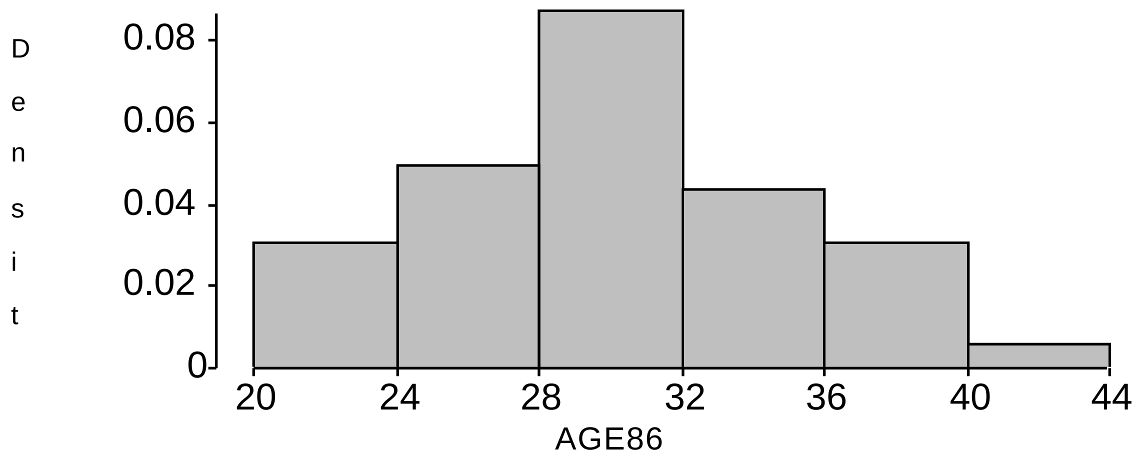
one can DE-ATTENUATE:

"bias-corrected" estimator of β_1 : $b_{1[LS]} \times \frac{1}{ICC}$

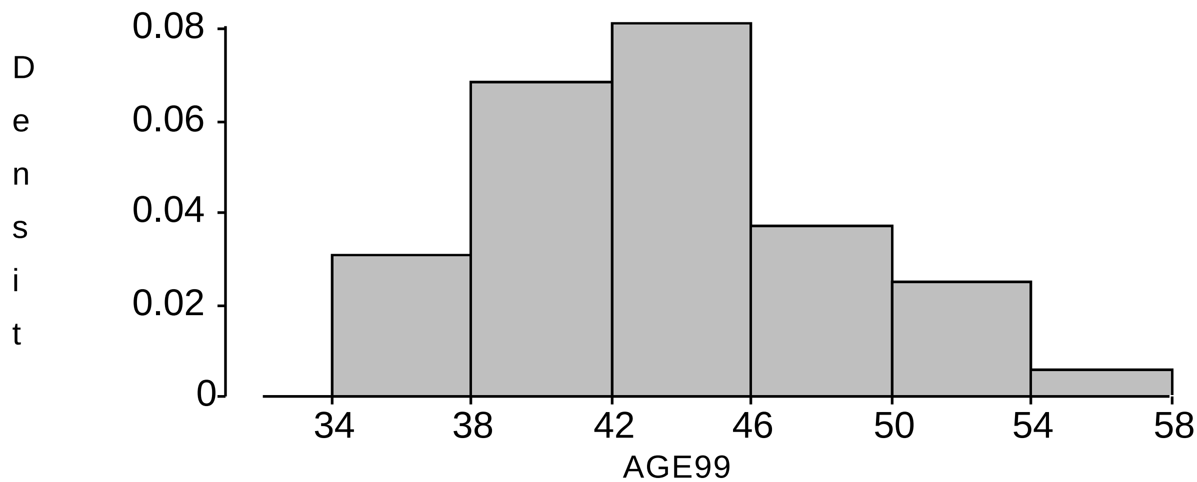
EXAMPLE OF "FLAT" SLOPE ("classical" measurement error model)

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
;
```



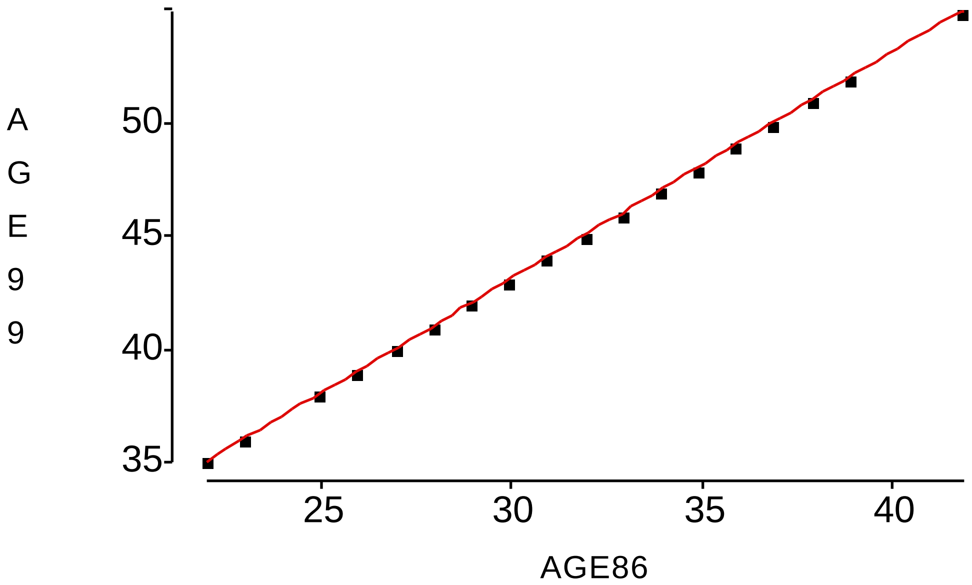
These 40 students 13 years later ... in 1999



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

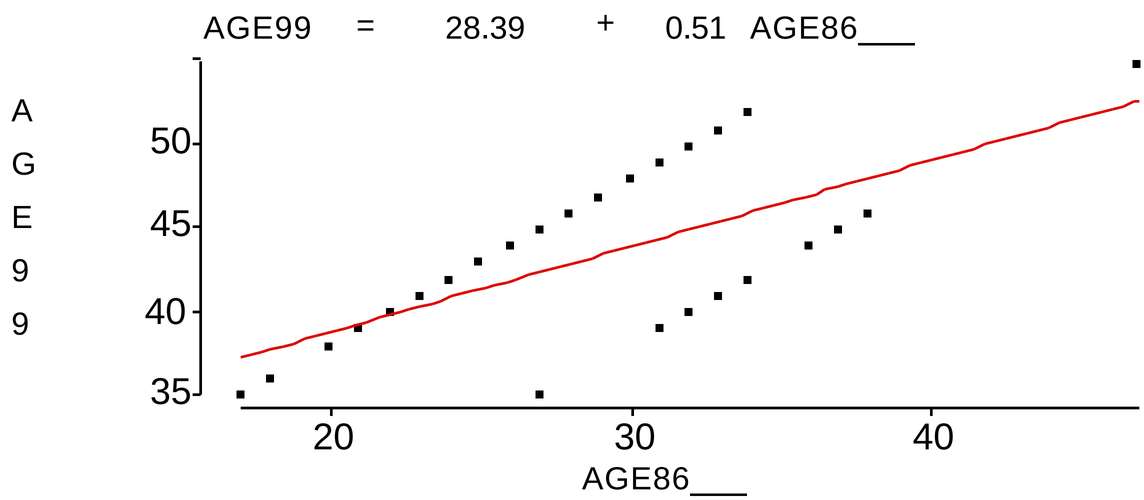


What if these 40 students had given their ages as true age +/- 5 years (with the + or - determined at random, without regard to true age)?

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

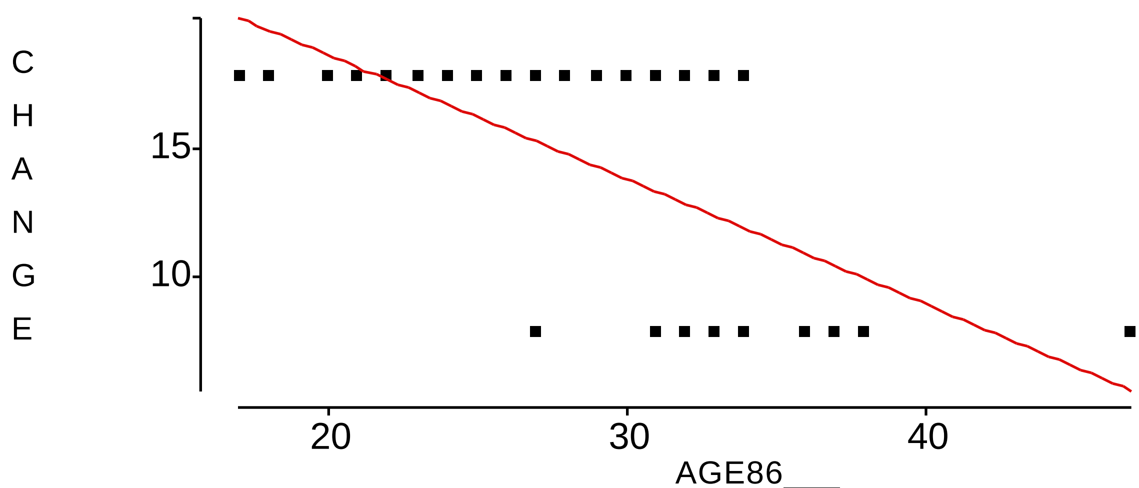
change__ = Age99 - Age86___;



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

If use naive LS estimator b_1 to estimate β_1 from the X*'s ...
then b_1 un<u>biased</u>

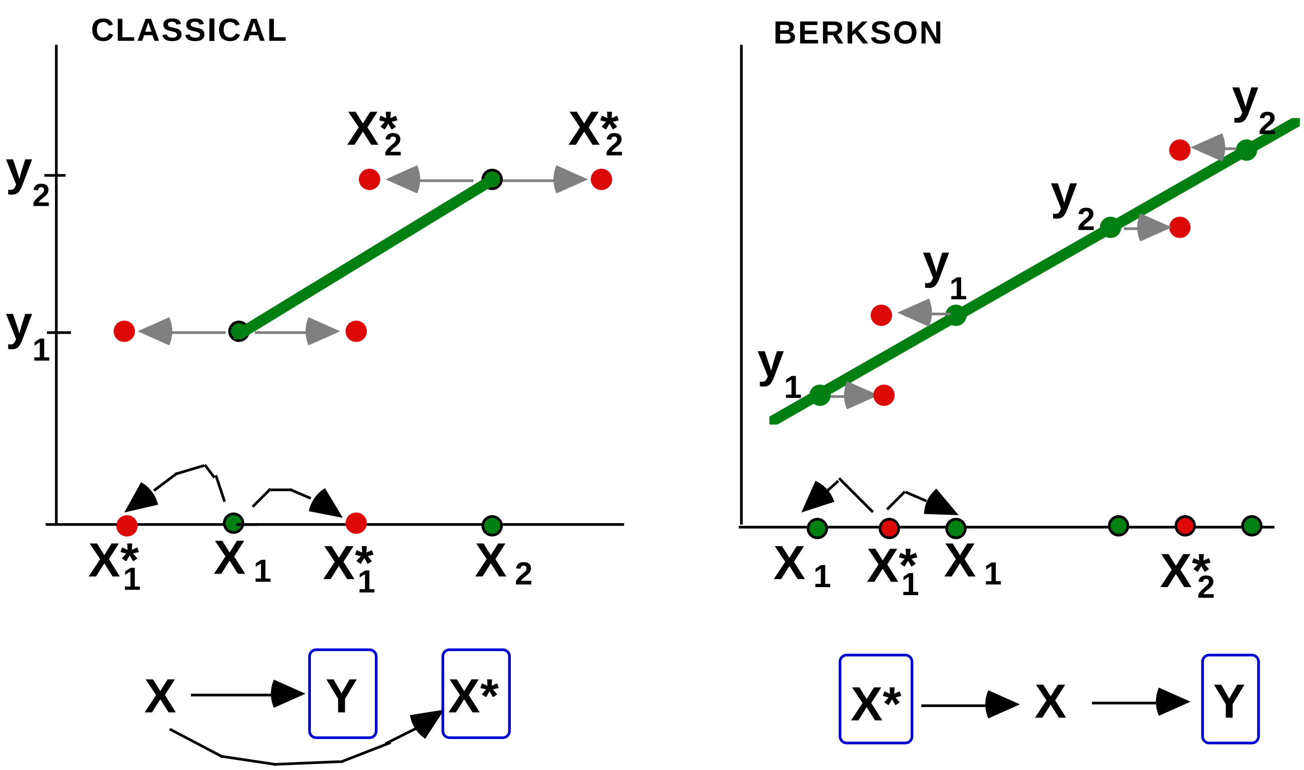
The "Classical" vs. "Berkson" difference ...

Assume

- **No Biologic Variation (i.e. all ε '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 $\sigma^2(\varepsilon)=0$

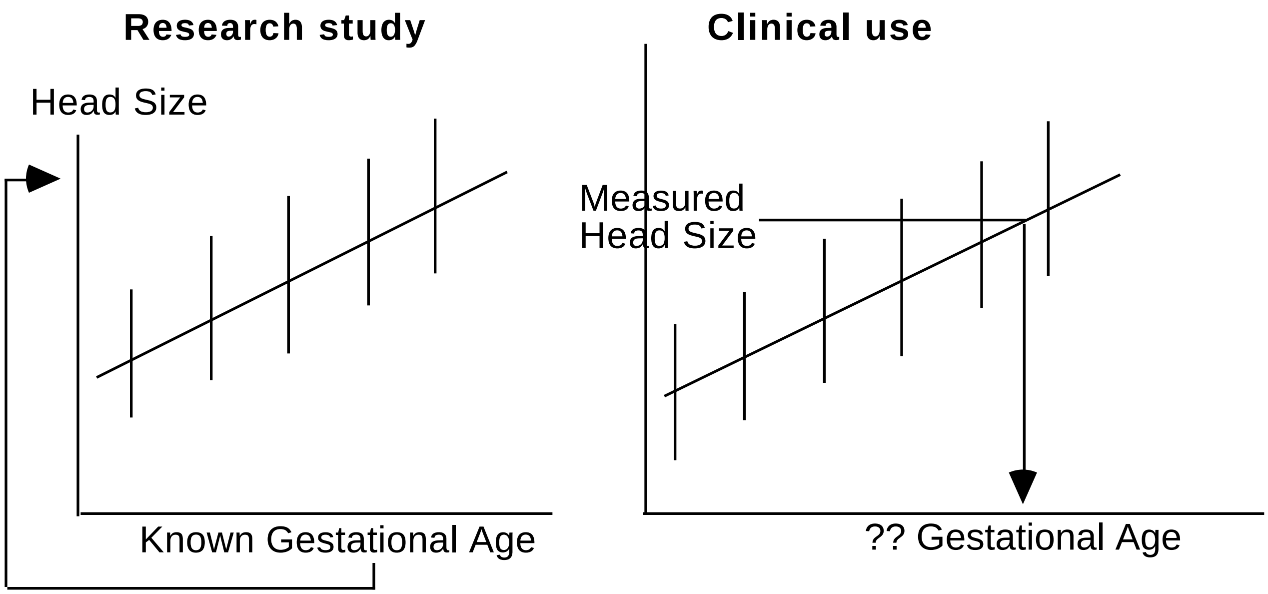
"Classical" Error Model	*"Berkson" Error Model*
$\dfrac{y_2 - y_1}{x^*_2 - x^*_1}$	$\dfrac{y_2 - y_1}{x^*_2 - x^*_1}$
$\dfrac{\beta\{x_2 - x_1\}}{[x_2 + \delta_2] - [x_1 + \delta_1]}$	$\dfrac{\beta\{x^*_2 + \delta_2\} - \beta\{x^*_1 + \delta_1\}}{x^*_2 - x^*_1}$
$\dfrac{\beta\{x_2 - x_1\}}{[x_2 - x_1] + [\delta_2 - \delta_1]}$	$\dfrac{\beta\{x^*_2 - x^*_1\} + \beta\{\delta_2 - \delta_1\}}{x^*_2 - x^*_1}$
$\dfrac{\beta}{1 + \dfrac{\delta_2 - \delta_1}{x_2 - x_1}}$	$\beta\left(1 + \dfrac{\delta_2 - \delta_1}{x^*_2 - x^*_1}\right)$
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($\hat{X}_h$) ???

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

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

4.7 Choice of X levels

Well explained in book, pp 169-170

Would simply emphasize a different way of viewing the terms

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

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

namely

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

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

This way, for example, $\mathbf{SD(b_1)} = \dfrac{\sigma}{\sqrt{n}\ \mathrm{SD}(X)}$

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.