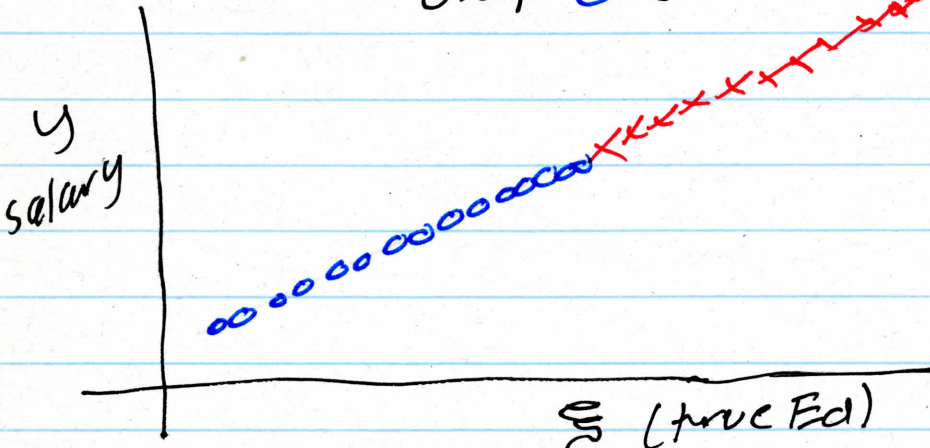


Salary Discrimination Ex (STIGLER)

Group \times \circ



non overlapping
on true Ed
non overlapping
salary

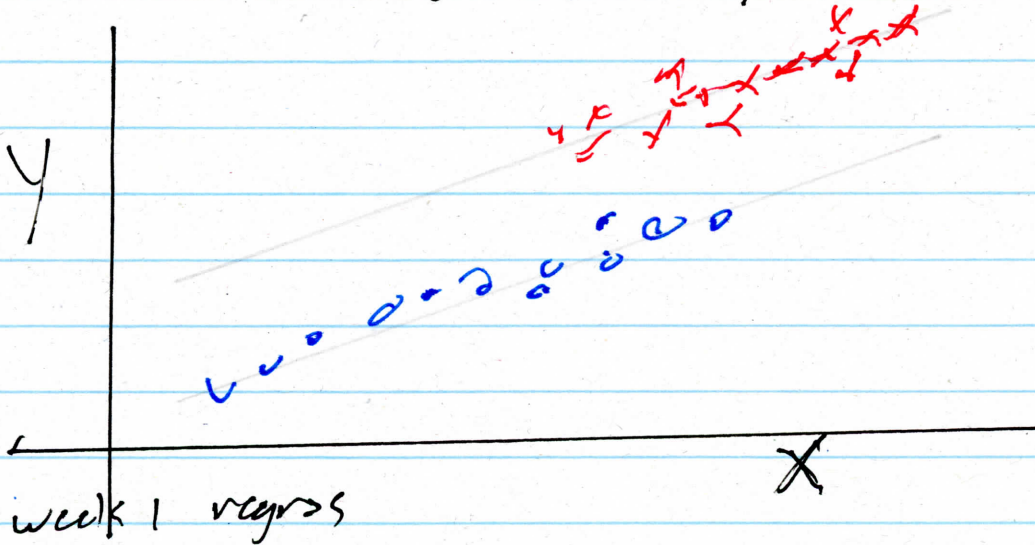
NO Group Difference
on salary setting
(all ca)

week 1
ancova
regression

$$E(Y|G, \epsilon) = \beta_0 + \beta_1 G + \beta_2 \epsilon$$

$$\boxed{\beta_1 = 0}$$

X as a fallible manifest measure of Ed
($X = \epsilon + e, e \sim (0, \sigma^2)$)



see
simulation
plot

week 1 regress

$$E(Y|G, X) = \gamma_0 + \gamma_1 G + \gamma_2 X$$

γ_1 big "advantage" to ~~xxx~~

week 5
ancova
forces
|| lines

results of meas error G measured perfectly.

Maindonald-Braun

sec 6.7 Errors-in-Variables

data ex: Dietary intake

single predictor $y_i = \alpha + \beta x_i + \epsilon_i$

follible
measure

$$w_i = x_i + v_i$$

error

$$v_i \sim (0, \text{Var}(x) \cdot \tau^2)$$

$$\sigma^2 = \text{Var}(\epsilon)$$

p. 205 reliability $\lambda = \frac{\text{Var}(X)}{\text{Var}(w)} = \frac{1}{1 + \tau^2}$

DAAG

errors IN X () function on reverse

$$\tau^2 = 1 \Rightarrow \lambda = 1/2$$

p. 206 Two Predictors

true
score
regr

$$Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

M-B result for predictors w_1 and X_2

i.e. X_1 observed w/ error, X_2 measured perfectly

$E(Y | w_1, X_2)$
pop coeff
of

w_1

$$\beta_1 \left(\frac{1 - \rho^2}{1 - \rho^2 + \tau^2} \right)$$

$$\rho = \text{Cor}(X_1, X_2)$$

pop coeff
of X_2

$$\beta_2 + \beta_1 \left(\frac{1 - \rho^2}{1 - \rho^2 + \tau^2} \right) \beta_{X_1, X_2}$$

p. 207 Fig 6.19 Simulation does a version of Stigler example, creating group diff as a result of measurement error.
R-session on reverse

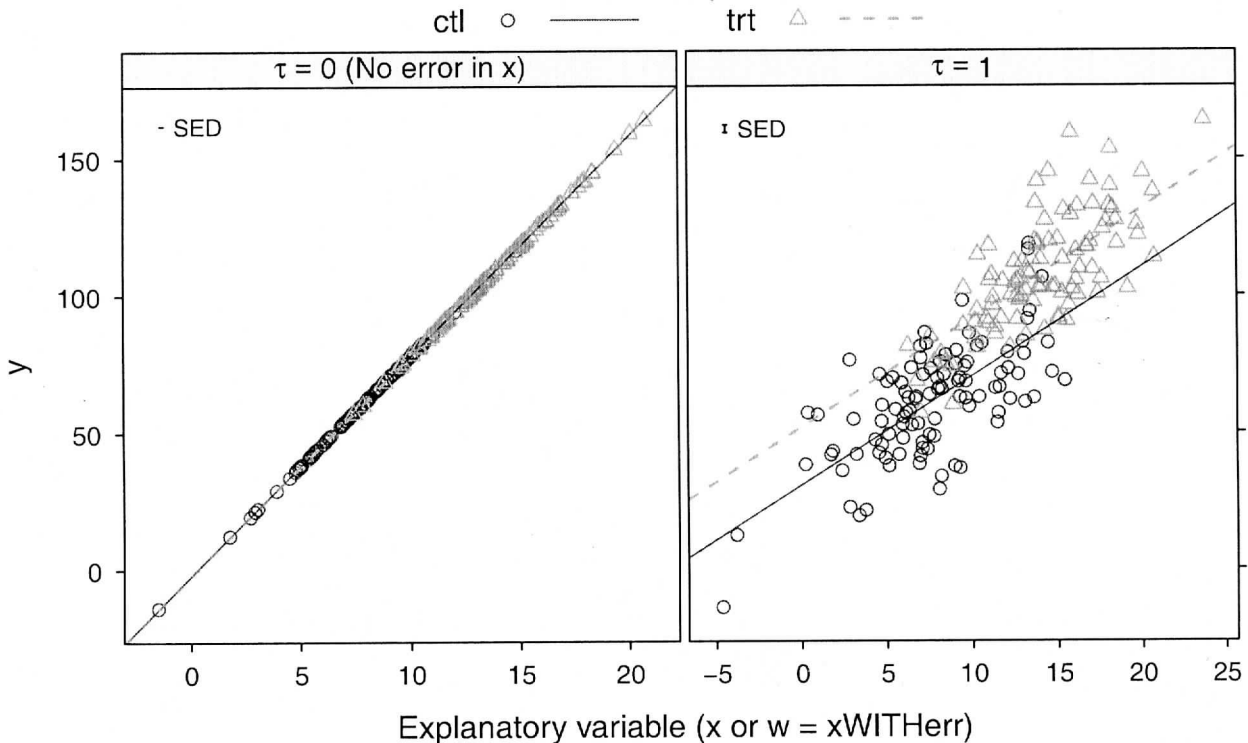
p.2

Stat 209
Week 1

```
-----  
#Version of Stigler example  
#defined salary metric 40K, 80K at hashes, slope 8K/1yr,  
#groups A and B have mean eff ed 6 and 11  
#var(eff ed) 81/12 [uniform var]  
#small resid var salary on eff ed (.06)  
#reliabilty fml ed .5 (low)  
#functions produces plots like Stigler, group diff approx 21K
```

use M-B
function

```
> errorsINx(mu = 8.5, n = 200, a = -2, b = 8, SDx=2.5, SDyerr = .25,  
+          timesSDx=1, gpfactor= TRUE,  
+          gpdiff=5, layout=NULL,  
+          parset = simpleTheme(alpha = 0.75, col = c("black","gray45"),  
+          col.line = c("black","gray45"), lwd=c(1,1.5), pch=c(1,2),  
+          lty=c(1,2)), print.summary=TRUE, plotit=TRUE)  
Intercept:ctl Offset:trt Slope  
No error in x -2.096198 -0.0937266 8.013173  
lsx          30.634011 21.2179763 3.989141  
>
```



R version 2.11.1 (2010-05-31)
Copyright (C) 2010 The R Foundation for Statistical Computing
ISBN 3-900051-07-0

```
> install.packages("DAAG")
```

```
Warning in install.packages("DAAG") :  
  argument 'lib' is missing: using 'C:\Users\rag\Documents\R\win64-library/2.11'  
Warning: unable to access index for repository http://www.stats.ox.ac.uk/pub/RWin/bin/w  
also installing the dependency 'randomForest'
```

```
trying URL 'http://cran.cnr.Berkeley.edu/bin/windows64/contrib/2.11/randomForest_4.5-36'  
Content type 'application/zip' length 146152 bytes (142 Kb)  
opened URL  
downloaded 142 Kb
```

```
trying URL 'http://cran.cnr.Berkeley.edu/bin/windows64/contrib/2.11/DAAG_1.02.zip'  
Content type 'application/zip' length 2343699 bytes (2.2 Mb)  
opened URL  
downloaded 2.2 Mb
```

```
package 'randomForest' successfully unpacked and MD5 sums checked  
package 'DAAG' successfully unpacked and MD5 sums checked
```

```
The downloaded packages are in  
  C:\Users\rag\AppData\Local\Temp\Rtmp52dMzb\downloaded_packages
```

```
> library(DAAG)
```

```
Loading required package: MASS  
Loading required package: rpart  
Loading required package: randomForest  
randomForest 4.5-36  
Type rfNews() to see new features/changes/bug fixes.
```

```
Attaching package: 'DAAG'
```

```
The following object(s) are masked from 'package:MASS':
```

```
  hills
```

```
> ?errorsINx  
errorsINx                package:DAAG                R Documentation
```

```
Simulate data for straight line regression, with "errors in x".
```

```
Description:
```

```
  Simulates  $y$ - $x$  and  $x$ - $y$  values for the straight line regression  
  model, but with  $x$ - $y$  values subject to random measurement error,  
  following the classical "errors in  $x$ " model. Optionally, the  
   $x$ -values can be split into two groups, with one group shifted  
  relative to the other
```

```
Usage:
```

```
errorsINx(mu = 12.5, n = 200, a = 15, b = 1.5, SDx=2, SDyerr = 1.5,  
          timesSDx=(1:5)/2.5, gpfactor=if(missing(gpdiff))FALSE else TRUE,  
          gpdiff=if(gpfactor) 1.5 else 0, layout=NULL,  
          parset = simpleTheme(alpha = 0.75, col = c("black","gray45"),  
                                col.line = c("black","gray45"), lwd=c(1,1.5), pch=c(1,2),
```

```
lty=c(1,2)), print.summary=TRUE, plotit=TRUE)
```

Arguments:

mu: Mean of z

n: Number of points

a: Intercept in model where z is measured without error

b: Slope in model where z is measured without error

SDx: SD of z -values, measured without error

SDyerr: SD of error term in 'y' where z is measured without error

timesSDx: SD of measurement error is 'timesSDx', as a multiple of 'SDx'

gpfactor: Should x-values be split into two groups, with one shifted relative to the other?

gpdiff: Amount of shift of one group of z-values relative to the other

layout: Layout for lattice graph, if requested

parset: Parameters to be supplied to the lattice plot, if any

print.summary: Print summary information on fits?

plotit: logical: plot the data?

Details:

The argument 'timesSDx' can be a numeric vector. One set of x -values that are contaminated with measurement error is simulated for each element of 'timesSDx'.

Value:

A matrix, with 'length(timesSDx)+2' columns. Values of z are in the first column. There is one further column (x with error) for each element of 'timesSDx', followed by a column for y . If there is a grouping variable, a further column identifies the groups.

Author(s):

John Maindonald

References:

Data Analysis and Graphics Using R, 2nd edn, Section 6.8.1

```
-----  
#Version of Stigler example  
#defined salary metric 40K, 80K at hashes, slope 8K/lyr,  
#groups A and B have mean eff ed 6 and 11  
#var(eff ed) 81/12 [uniform var]
```

```

#small resid var salary on eff ed (.06)
#reliabilty fml ed .5 (low)
#functions produces plots like Stigler, group diff approx 21K

> errorsINx(mu = 8.5, n = 200, a = -2, b = 8, SDx=2.5, SDyerr = .25,
+          timesSDx=1, gpfactor= TRUE,
+          gpdiff=5, layout=NULL,
+          parset = simpleTheme(alpha = 0.75, col = c("black","gray45"),
+          col.line = c("black","gray45"), lwd=c(1,1.5), pch=c(1,2),
+          lty=c(1,2)), print.summary=TRUE, plotit=TRUE)
      Intercept:ctl Offset:trt      Slope
No error in x      -2.096198 -0.0937266  8.013173
lsx                30.634011 21.2179763  3.989141
>

```

