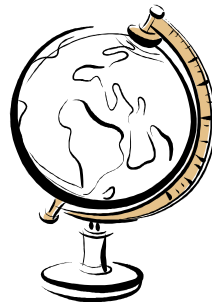


Regression errors in “x” case study

Dave Saville

Bioinformatics, Mathematics and **Statistics**

AgResearch, P O Box 60, **Lincoln, N Z**



Outline of talk

1. Motivation
2. Design of simulation studies
3. Results
4. Conclusions



1. Motivation

- Annual short courses in statistics for researchers, primarily biologists. Hands-on, interactive teaching method.
- Fourth day is devoted to a simple, linear regression workshop.
- Assumptions are discussed during the workshop.
- One assumption is that “the x values have no error”.
- Usual escape route for biometricians: “results are conditional upon the observed x values” (– but what does this mean?)

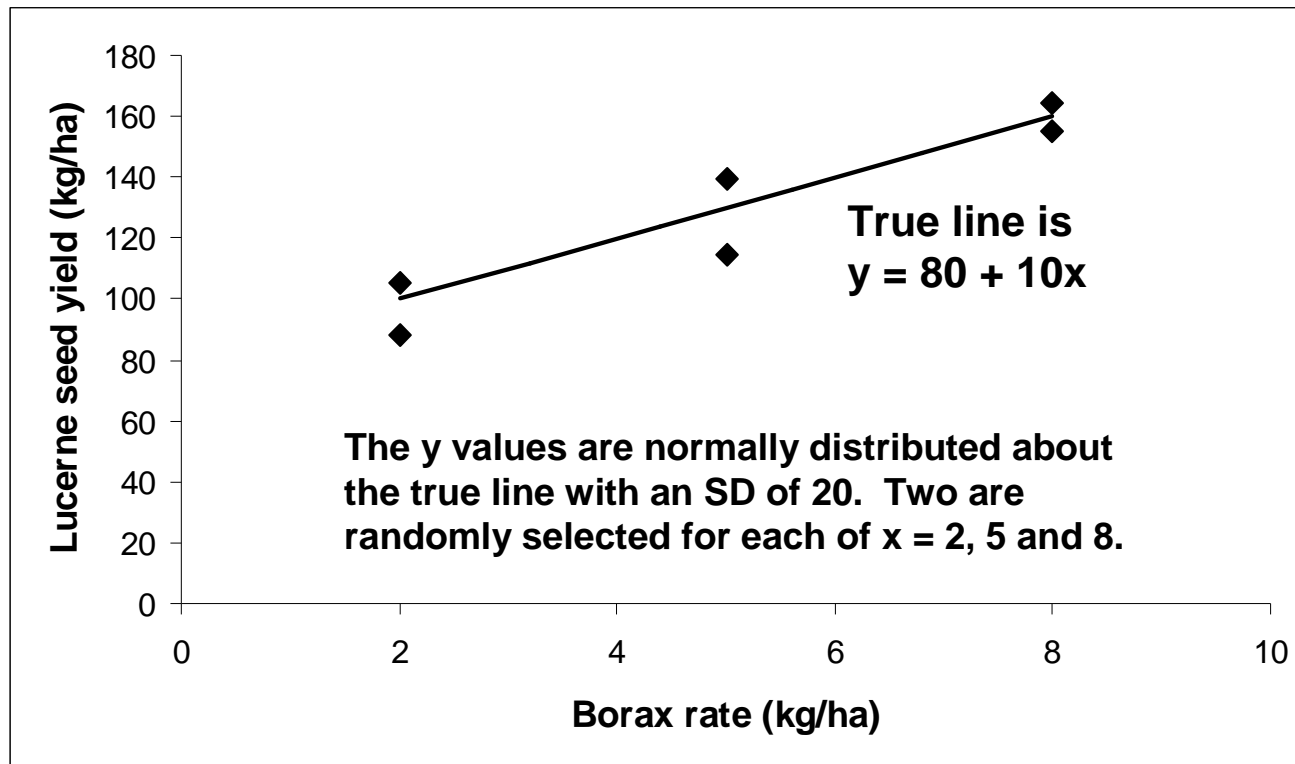


“Results are conditional upon the observed x values” (– but what does this mean?)

- The estimated $se(\text{slope})$, for example, is conceptually based upon the distribution of slopes that you would observe if you did *many identical* studies with the *same* set of x values (with the observed y values normally distributed about the true line).
- This is mathematically very convenient (but does the mind boggle?).
- *Another approach* is to think of “true” (x,y) points lying precisely on a straight line, with both x and y measured imprecisely.
- How does regression behave if the latter is simulated?



Regression class exercise (no error in x)



Regression class exercise (continued)

- Each student effectively simulates an experiment in which the data are normally distributed about a perfectly straight line, with no errors in the x values.
- The power of the test of “true slope = 0” is about 60% (class results were 67%, 48%, 69%, 76%, 60% and 46% in recent workshops of about 20 folk). [Later: 62% in simulations]

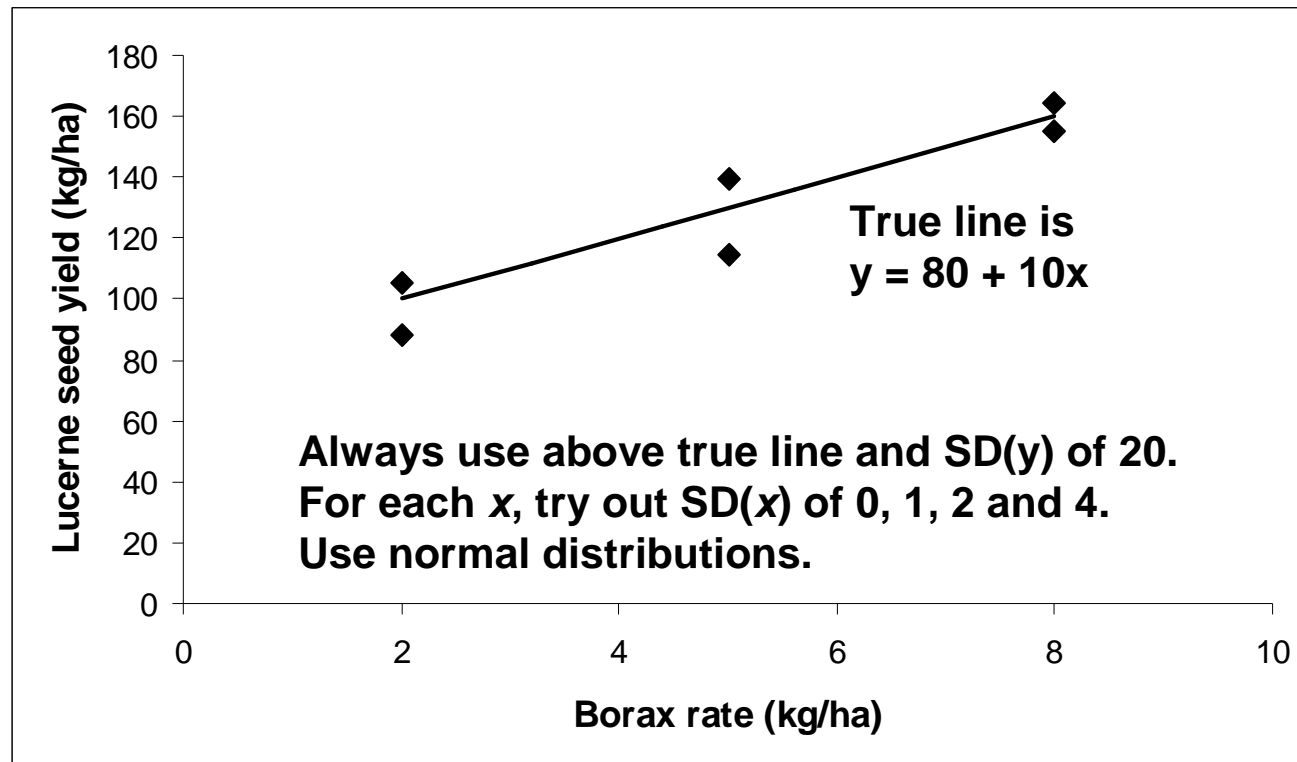
Questions: How would the power change if there were errors in x ?

Would biases creep into the estimates?



2. Design of simulation studies

(case true slope=10)



Design of simulations – more details

- Four cases were simulated: SD-x (about true values) of 0, 1, 2 and 4.
- Why choose these?
- With a true slope of 10, the range of “true” y values on the line was $160 - 100 = 60$, while the range of x values was $8 - 2 = 6$. As a proportion of the range, the SD-y of 20 was $20/60 = 0.333$. Similarly, the third SD-x was $2/6 = 0.333$. That is, the SD-x values chosen were *zero, half, same and twice* that of SD-y (%-wise).
- What are these in terms of CV%s? The CV%-y is $20/130 * 100 = 15\%$. For x, the CV%-x's are 0, 20% ($=1/5 * 100$), 40% and 80% respectively.

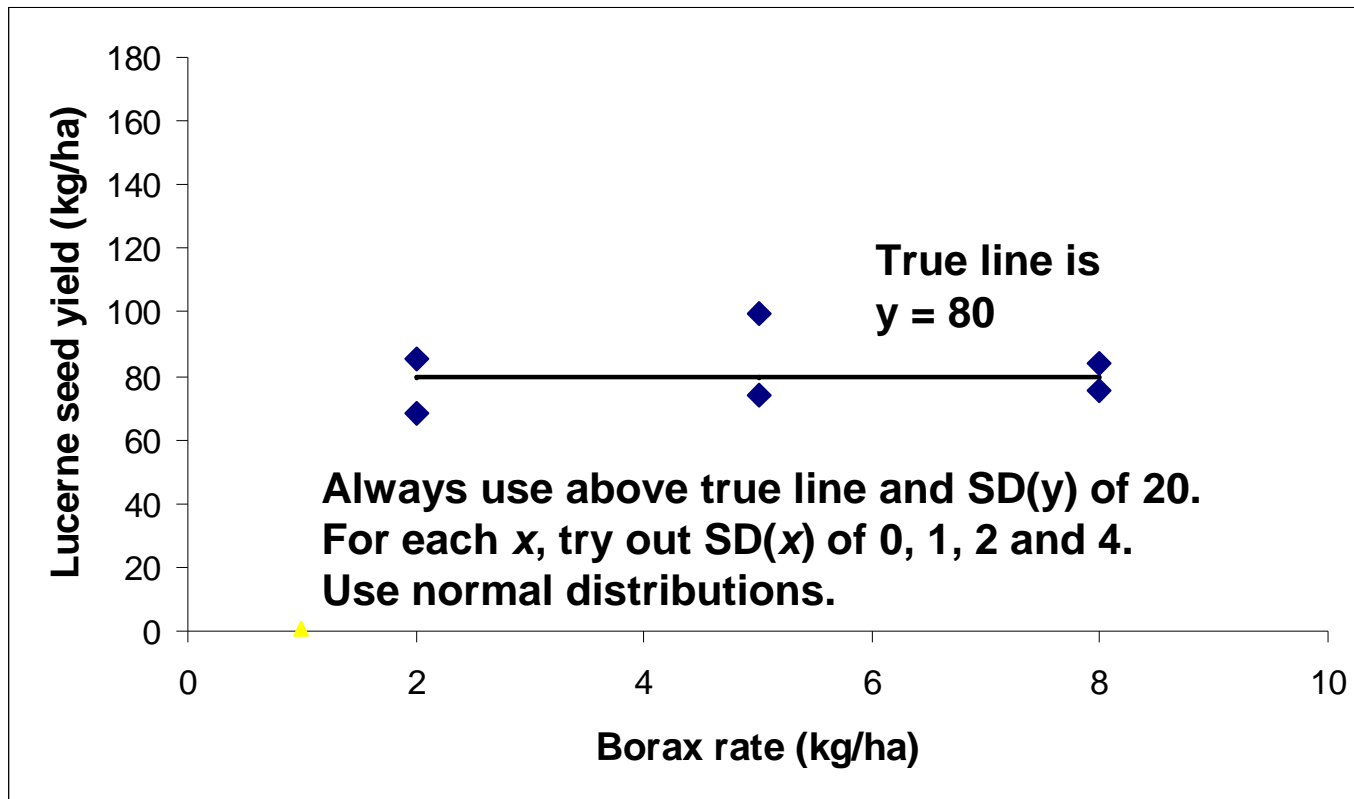


Design of simulations – more details (contd)

- For each simulation of a study with 6 experimental units, GenStat's random number generator was used to generate two sets of 6 values from a standard normal distribution, $N(0,1)$.
- These were then rescaled as appropriate (six by 20 and six by 0, 1, 2 or 4) and added to the appropriate "true" values for y and x .
- GenStat was used to do 10,000 simulations for each of the 4 cases.
- For each simulation, the slope, estimated elevation of the line for the mean x ($= 5$), variance about the line, and significance of the slope were stored, and used in averaging and making histograms.



Design of simulations (case true slope=0)



Design of simulations – more details

- Details are the same as for the case “true slope = 10”, except the following logic changes:
- With a true slope of 0, the range of “true” y values on the line was 0, so we can’t sensibly think of the SD- y as a percentage of the range. So the first logic breaks down.
- What are the SDs in terms of CV%s? The CV%- y is $20/80*100 = 25\%$. The SD- x ’s of 0, 1, 2 and 4 in terms of CV%- x ’s are 0, 20% ($=1/5*100$), 40% and 80% respectively.



3. Results

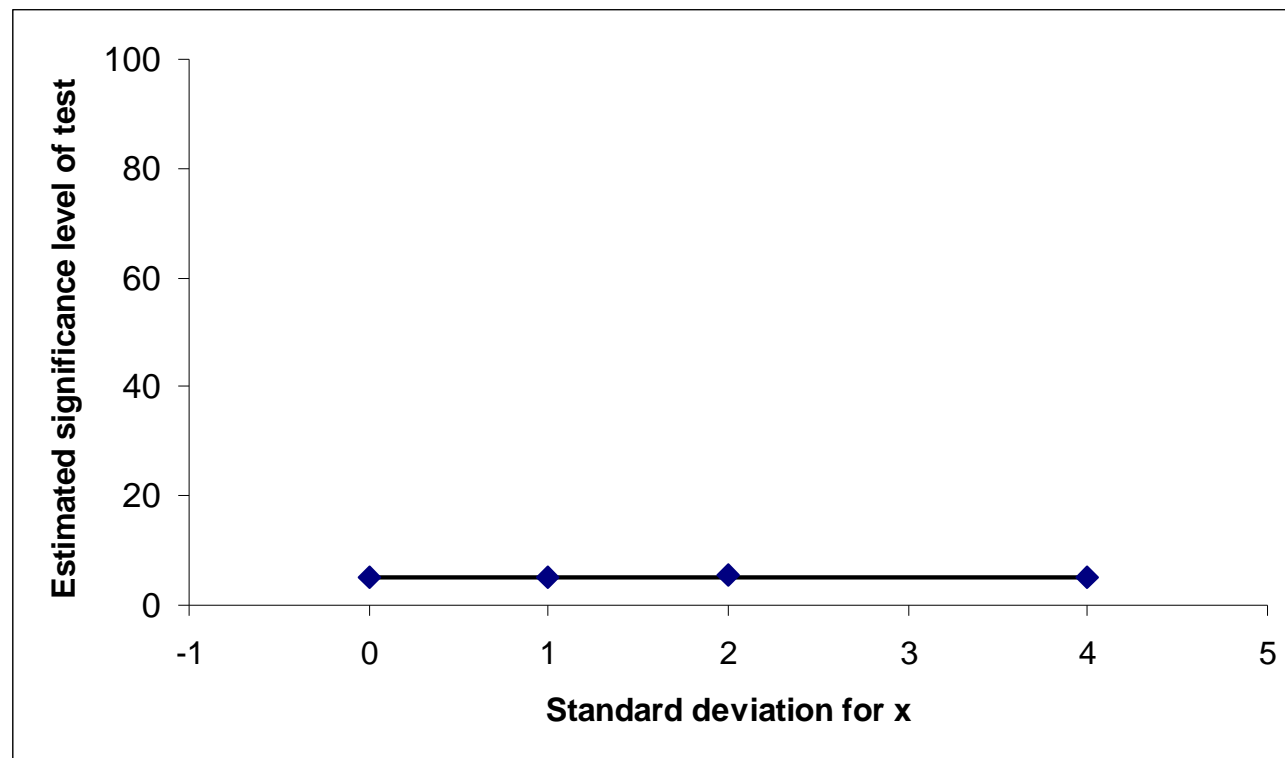
First, the good news!

Mean results for the four cases with *true slope* = 0
(with SD-x of 0, 1, 2 and 4).



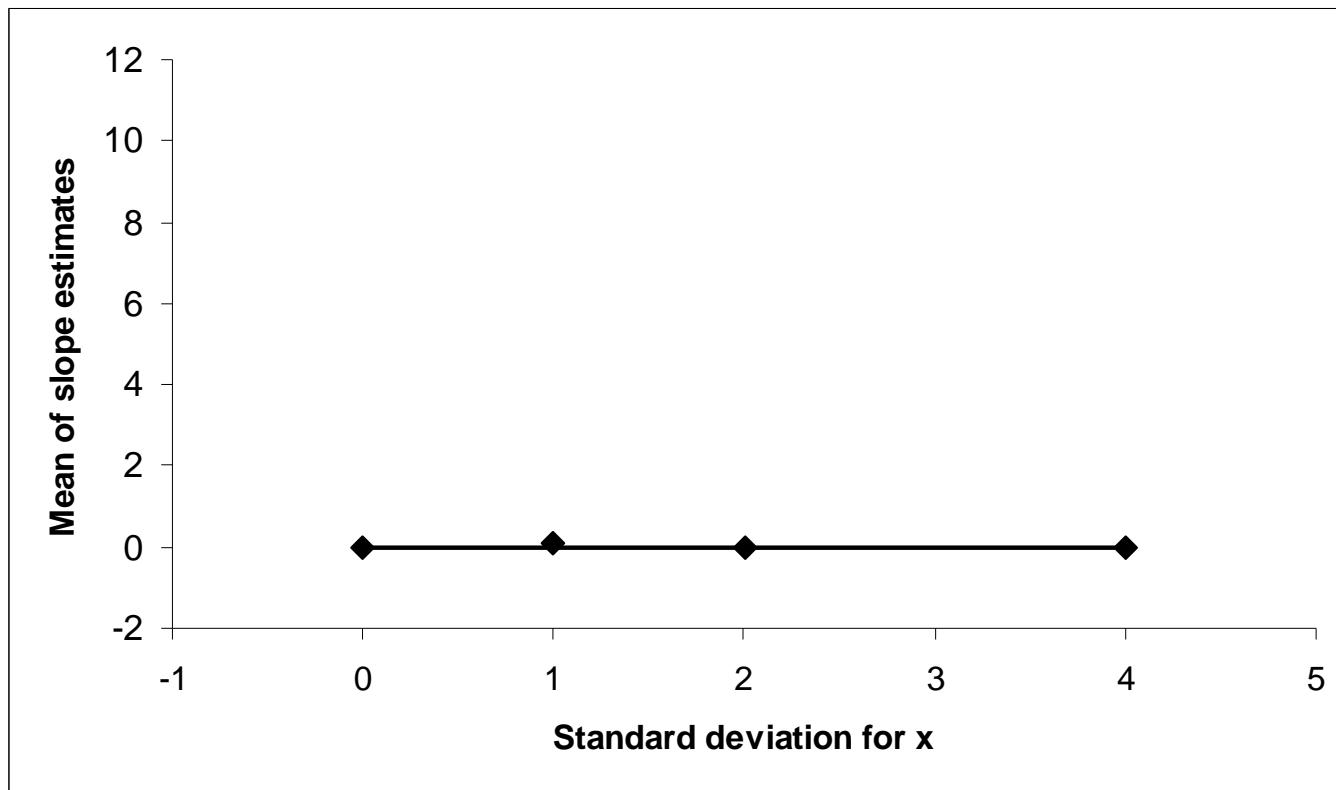
(A) Four cases with true slope = 0

- Estimated *significance level of test* was very close to 5.0% regardless of SD-x (values were 5.1, 5.1, 5.4 and 5.2%). That is, no drift.



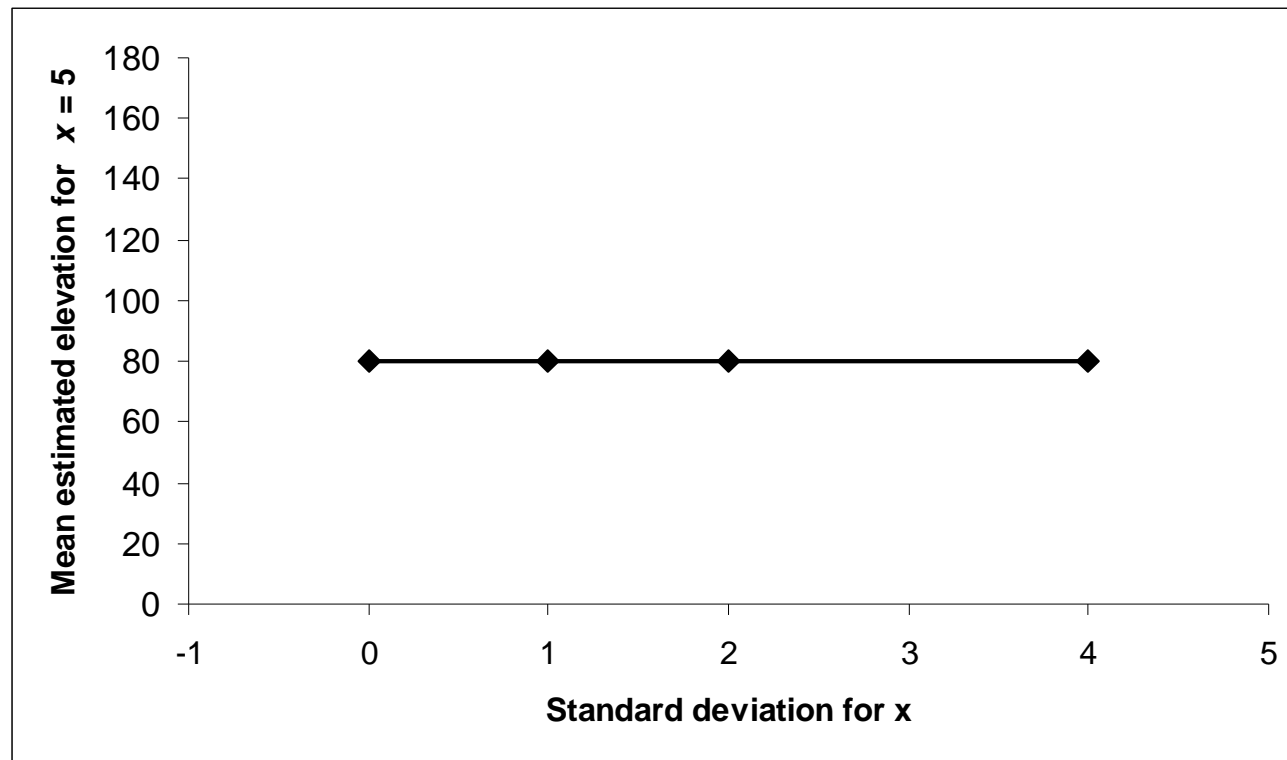
(A) Four cases with true slope = 0

- Mean *estimated slope* was very close to zero regardless of SD-x (values were 0, 0.1, 0 and 0). That is, no bias.



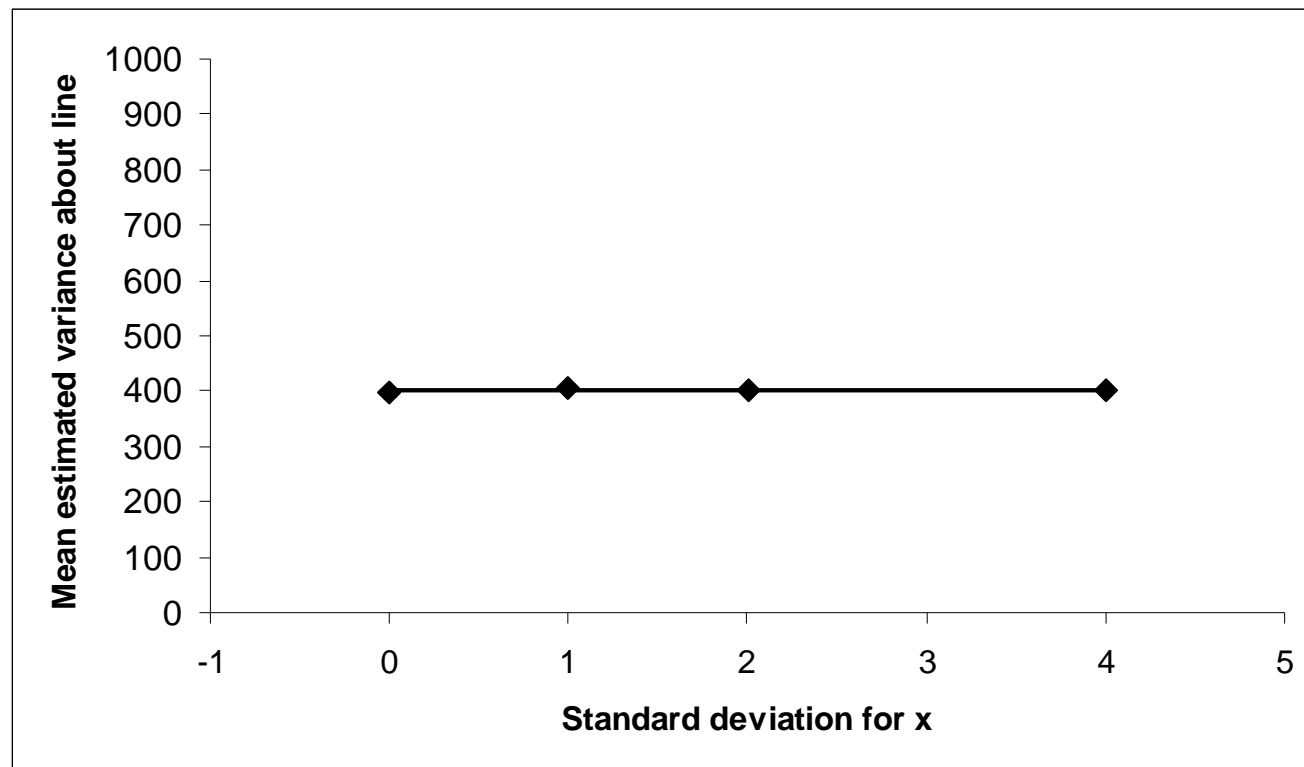
(A) Four cases with true slope = 0

- Mean *estimated elevation of the true line at the mean x value of 5* was very close to the true value of 80 regardless of SD-x (values were 80.1, 80, 80 and 80.2). That is, no bias.



(A) Four cases with true slope = 0

- Mean *estimated variance about the true line* was very close to the true value of 400 regardless of SD- x (values were 396, 404, 402 and 400). That is, no bias.



(A) Four cases with true slope = 0

MORALS:

- When the true slope is zero – that is, no relationship between the x and y variables – errors in the measurement of the x values have no effect on the operating characteristics of the regression procedure.
- The significance level of the test is unaffected.
- The estimates of the elevation and slope are still unbiased.
- The estimated variance about the line is still unbiased.

I guess this makes sense..... (were you thinking, this is obvious....?)



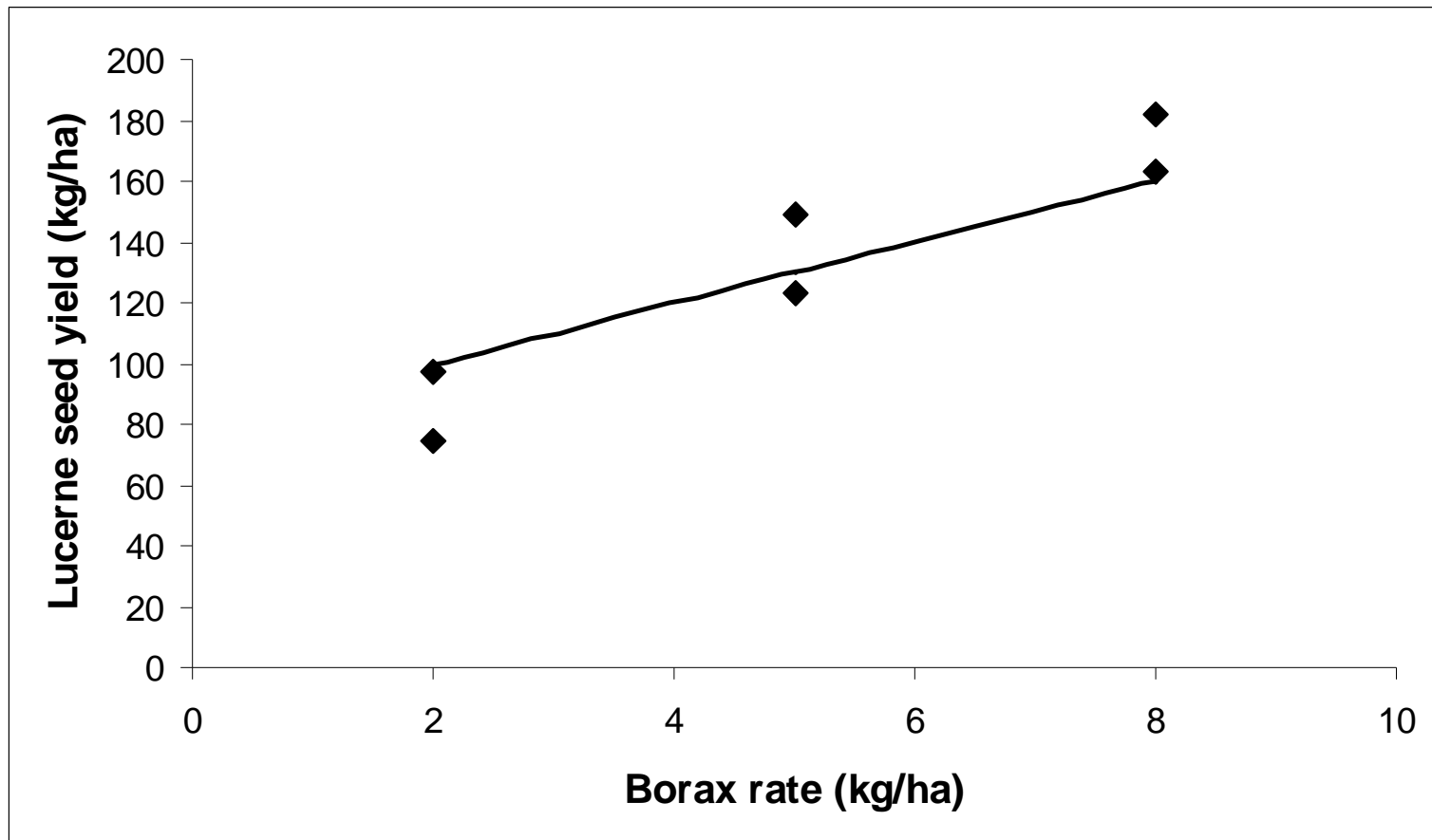
3. Results (continued)

Now, the bad news!

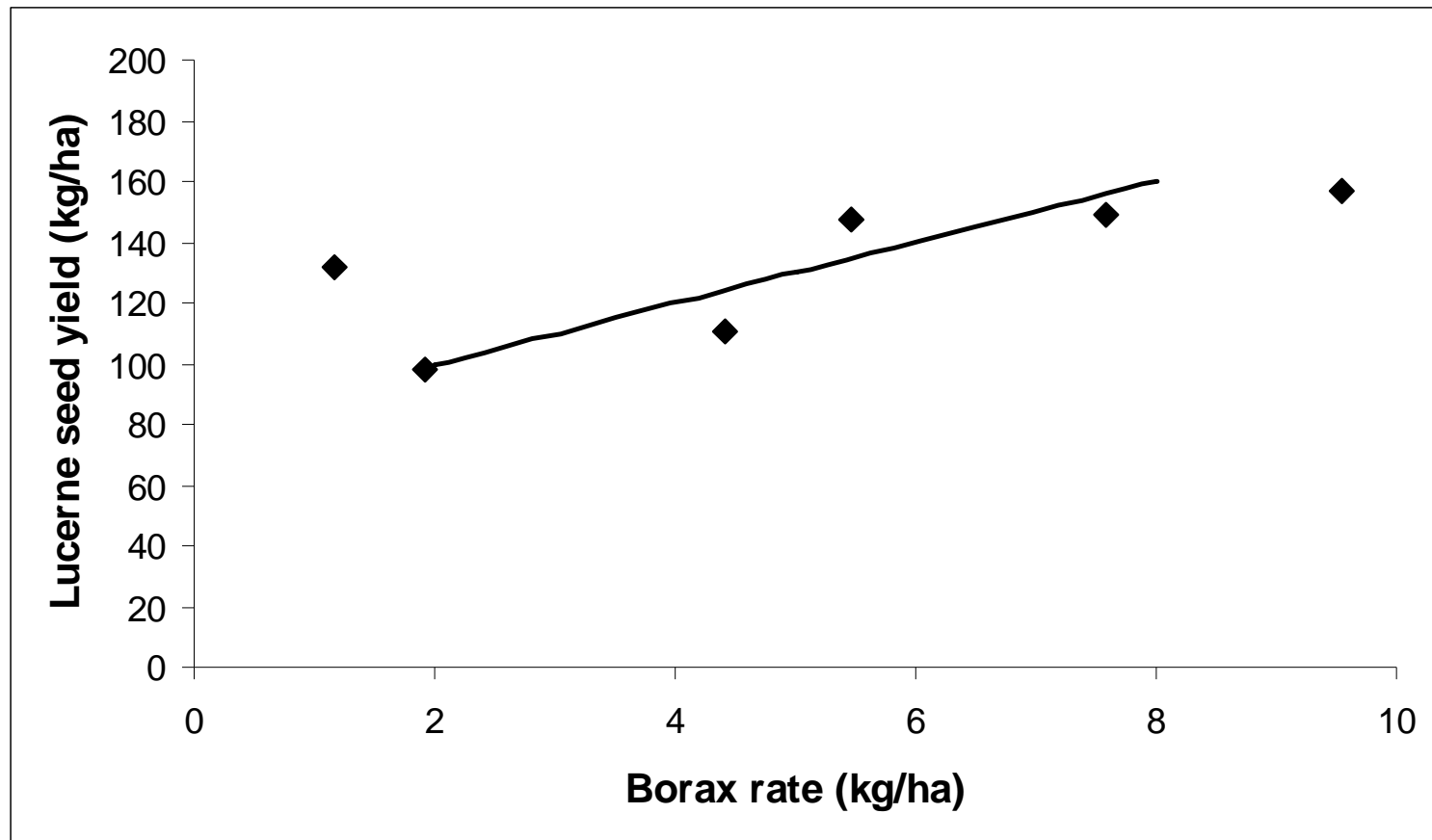
Mean results for the four cases with *true slope* = 10
(with SD-x of 0, 1, 2 and 4).



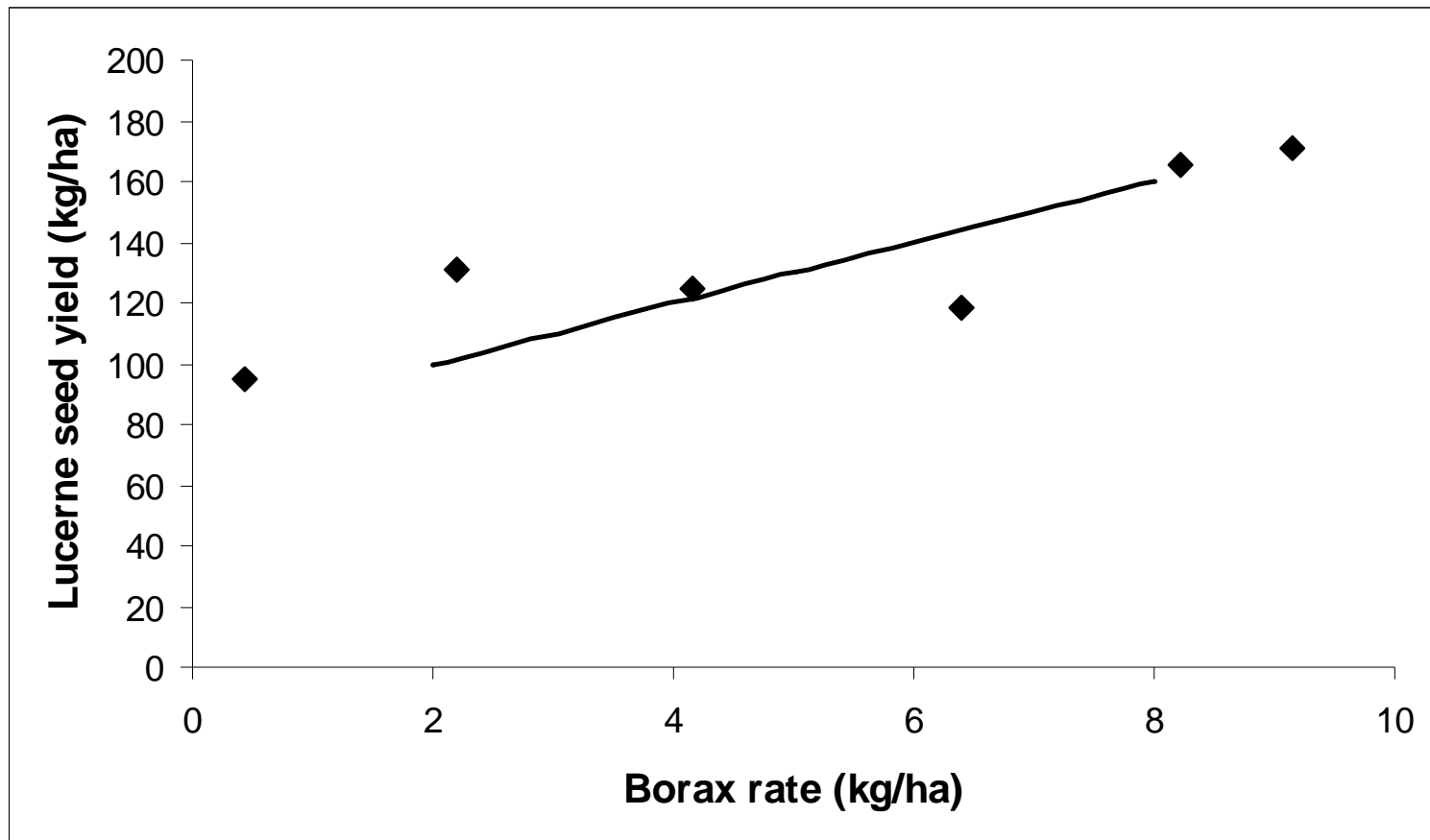
Example of data when $SD-x = 0$ (true line shown)



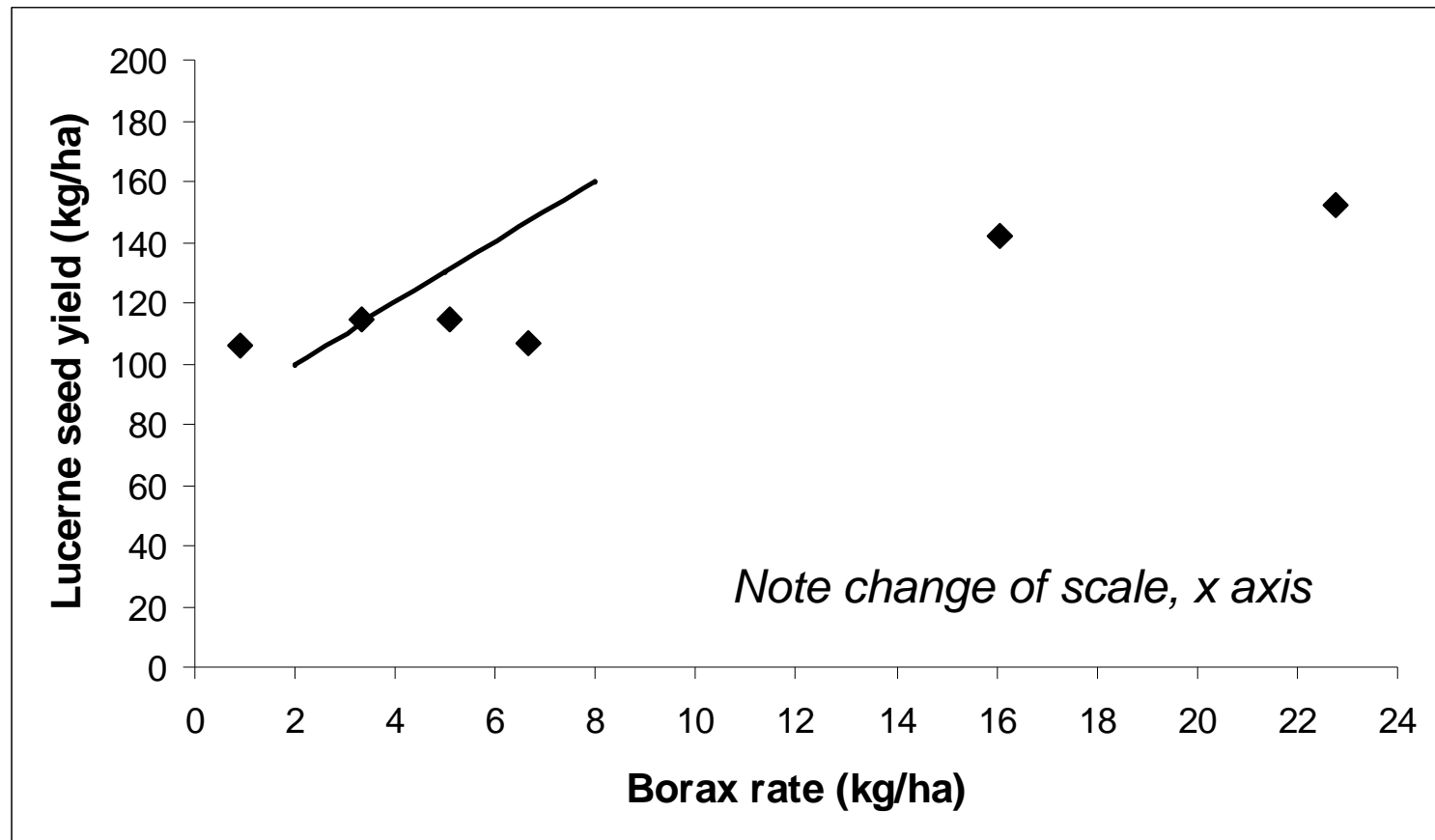
Example of data when $SD-x = 1$ (true line shown)



Example of data when $SD-x = 2$ (true line shown)

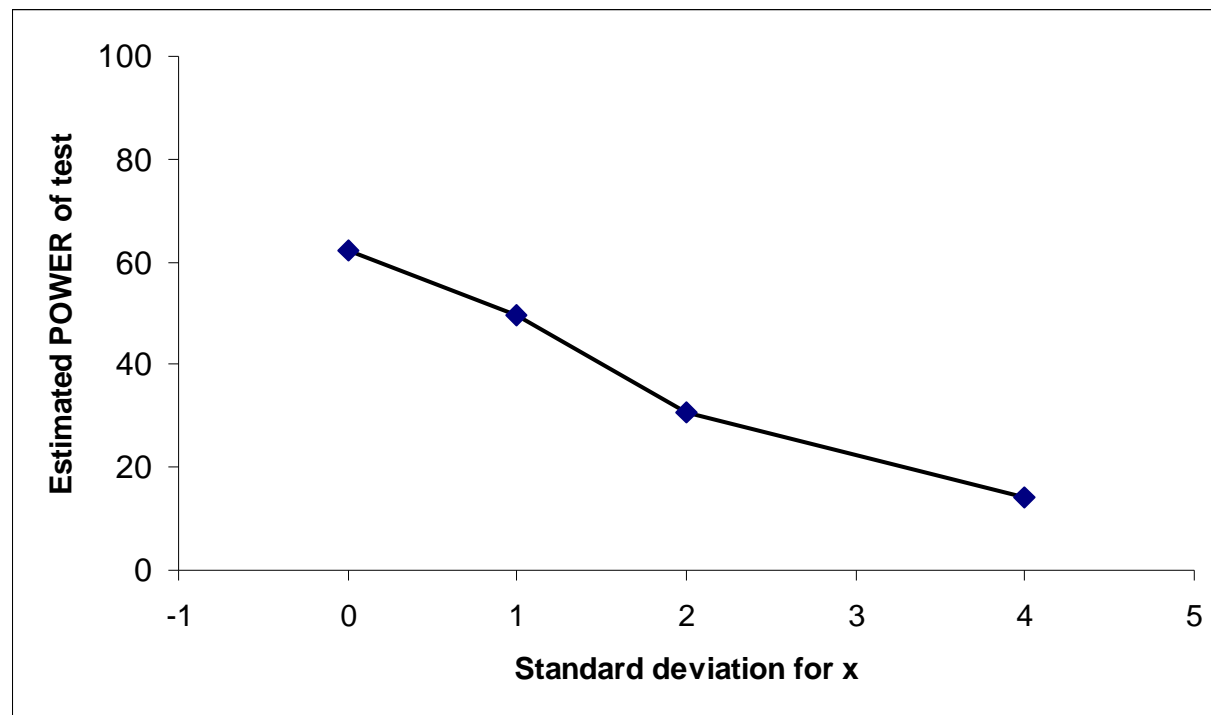


Example of data when $SD-x = 4$ (true line shown)



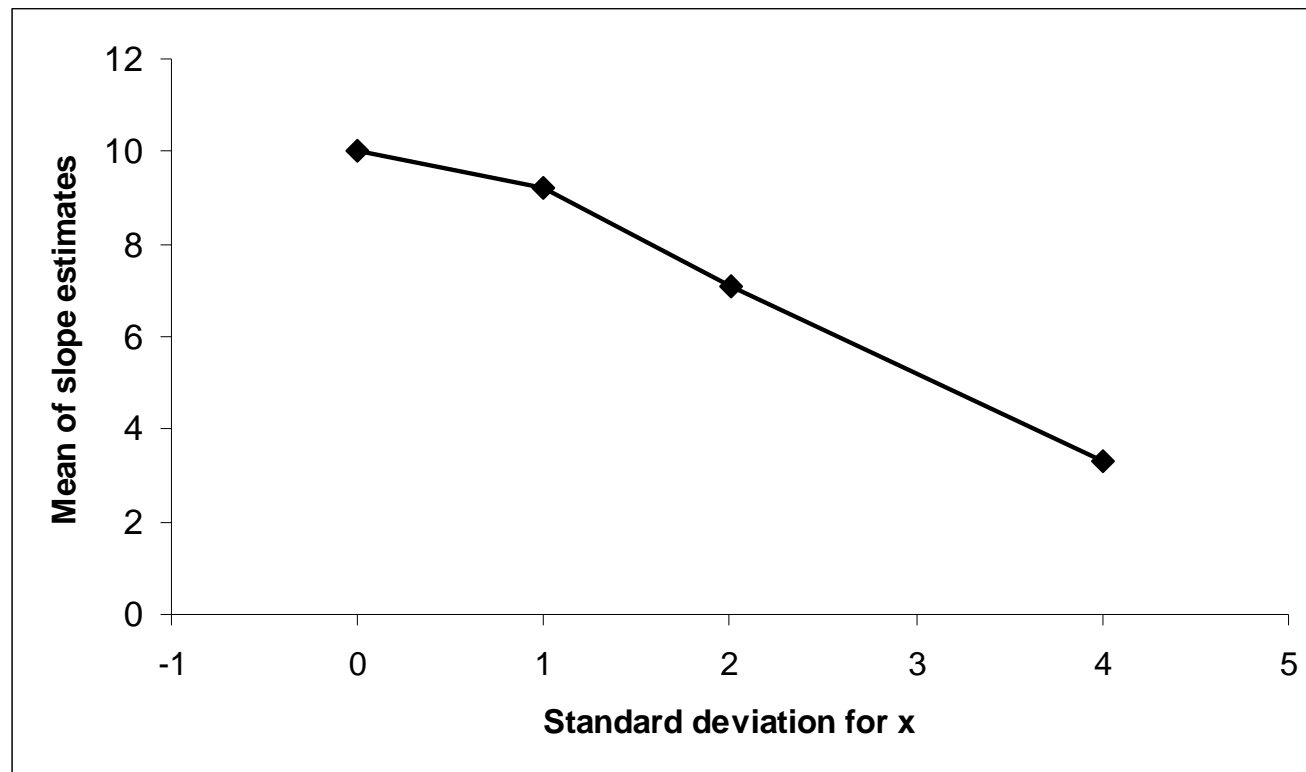
(B) Four cases with true slope = 10

- Estimated *power of test* (probability of rejecting the idea that the true slope = 0) was 62% when there were no errors in x , *but decreased steadily* with increasing errors in x (values were 62%, 50%, 31% and 14%).



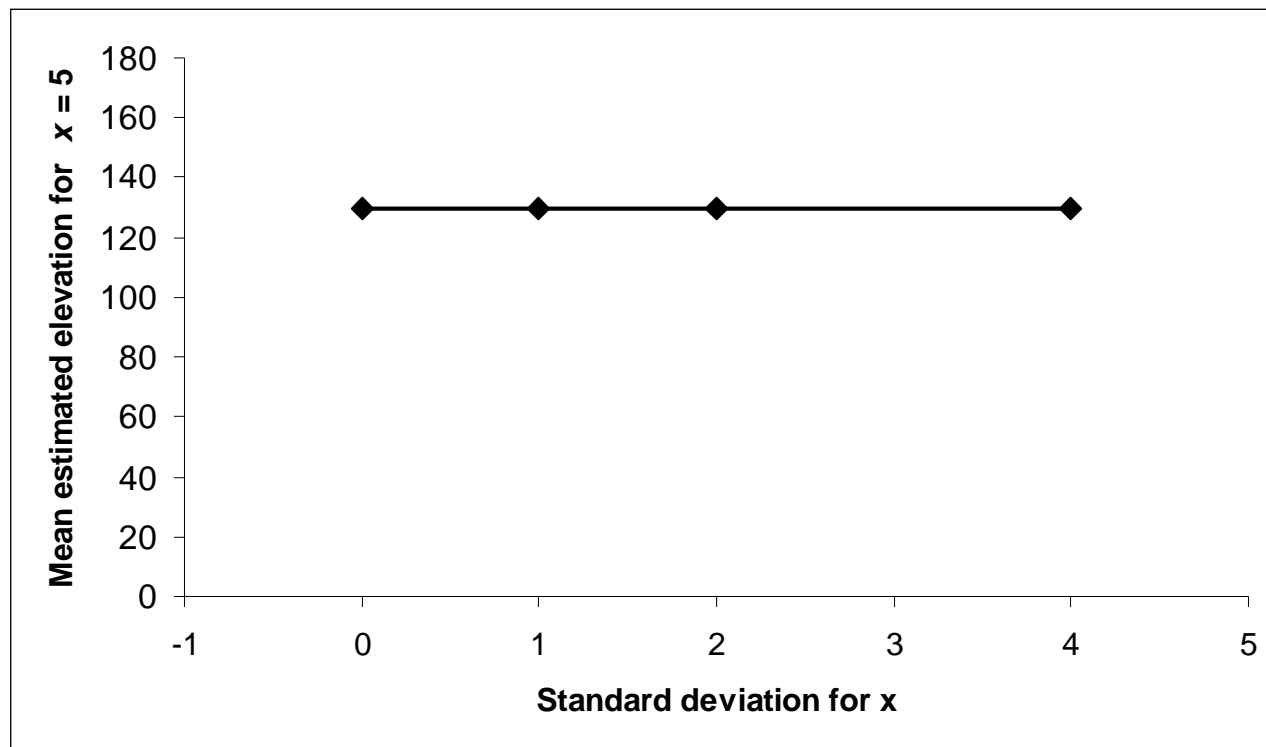
(B) Four cases with true slope = 10

- Mean *estimated slope* was very close to 10 when there were no errors in x , but *decreased steadily* with increasing errors in x (values were 10.0, 9.2, 7.1 and 3.3).



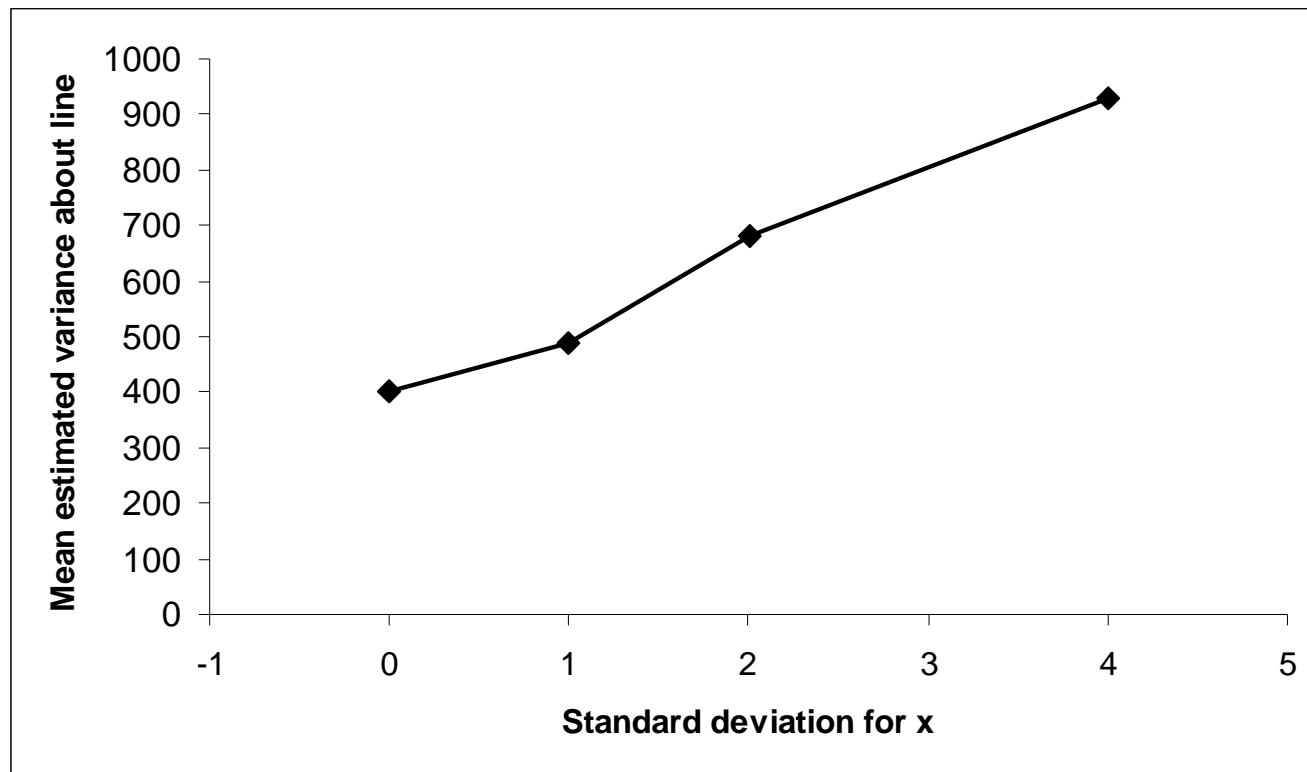
(B) Four cases with true slope = 10

- Mean *estimated elevation of the true line at the mean x value of 5* was very close to the true value of 130 regardless of SD-x (values were 129.9, 130, 130 and 129.8). That is, *no bias*.



(B) Four cases with true slope = 10

- Mean *estimated variance about the true line* was very close to the true value of 400 when there were no errors in x , but *increased steadily* with increasing errors in x (values were 402, 488, 682 and 928).



(B) Four cases with true slope = 10

MORALS:

- When the true slope is non-zero (10) – that is, a positive relationship between the x and y variables – errors in the measurement of the x values can have *substantial effects* on the operating characteristics of the regression procedure.
- The *power* of the test is *reduced*.
- The estimates of the *slope* are *biased downwards*.
- The estimates of the *elevation at the mean of the x values* (5 used here) are still *unbiased*.
- The estimated *variance* about the true line is *increased*.



4. Conclusions

As a simple case study, I found this quite enlightening!
(though not original, and presumably producing answers that are well known to at least some of you).



4. Conclusions

- (A) If there is truly *no* relationship between x and y (true slope = 0), then errors in x have *no effect* on the significance level of the test of the slope and introduce *no biases* into the estimates of slope, elevation and variance.
- (B) If there is truly a *positive* relationship between x and y (true slope > 0), then errors in x *reduce the power* of the test of the slope, *bias the slope* estimates downwards, and *increase y -variance* estimates. Estimates of *elevation* at the mean x value are unbiased, but are biased for other x values.



The End

The logo for 'agresearch' features a solid green square on the left. The word 'agresearch' is written in white, lowercase, sans-serif font across the bottom portion of the square.

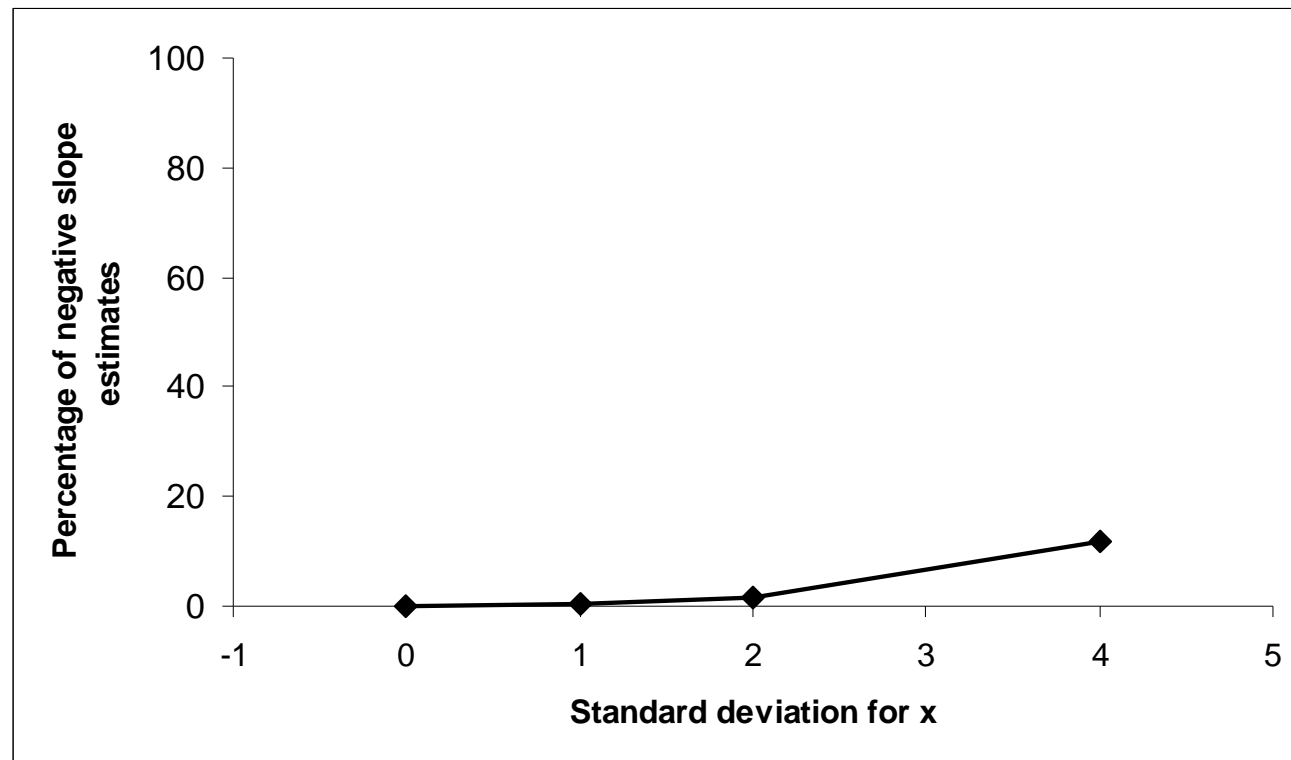
agresearch

Farming, Food and Health. **First**

*Te Ahuwhenua, Te Kai me te Whai Ora. **Tuatahi***

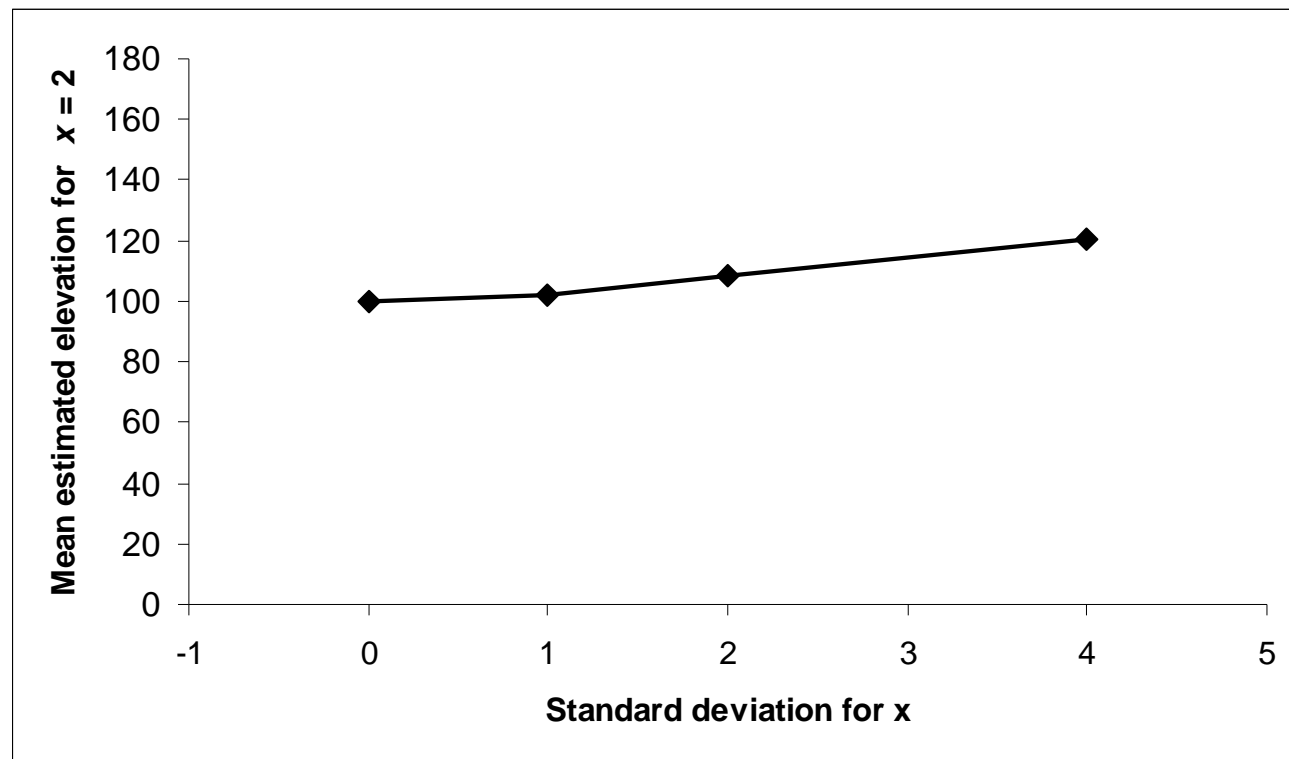
(B) Four cases with true slope = 10

- *How often would a researcher be misled into thinking the relationship is negative instead of positive? The answer is very seldom when there were no errors in x , but increasingly often with increasing errors in x (values were 0.1%, 0.2%, 1.7% and 11.9%).*



(B) Four cases with true slope = 10

- Mean *estimated elevation of the true line at a low x value of 2* was very close to the true value of 100 when there were no errors in x , *but became more and more upwardly biased* with increasing errors in x (values were 100, 102.4, 108.6 and 120.2).



(B) Four cases with true slope = 10

- Mean *estimated elevation of the true line at a high x value of 8* was very close to the true value of 160 when there were no errors in x , *but became more and more downwardly biased* with increasing errors in x (values were 160, 157.6, 151.3 and 140.3).

