The Least Squares Assumptions
(SW Section 4.4)

What, in a precise sense, are the properties of the OLS estimator? We would like it to be unbiased, and to have a small variance. Does it? Under what conditions is it an unbiased estimator of the true population parameters?

To answer these questions, we need to make some assumptions about how Y and X are related to each other, and about how they are collected (the sampling scheme)

These assumptions – there are three – are known as the Least Squares Assumptions.

The Least Squares Assumptions

$$Y_i = \beta_0 + \beta_1 X_i + u_i,\ i = 1,\ldots, n$$

1. The conditional distribution of u given X has mean zero, that is, $E(u|X = x) = 0$.

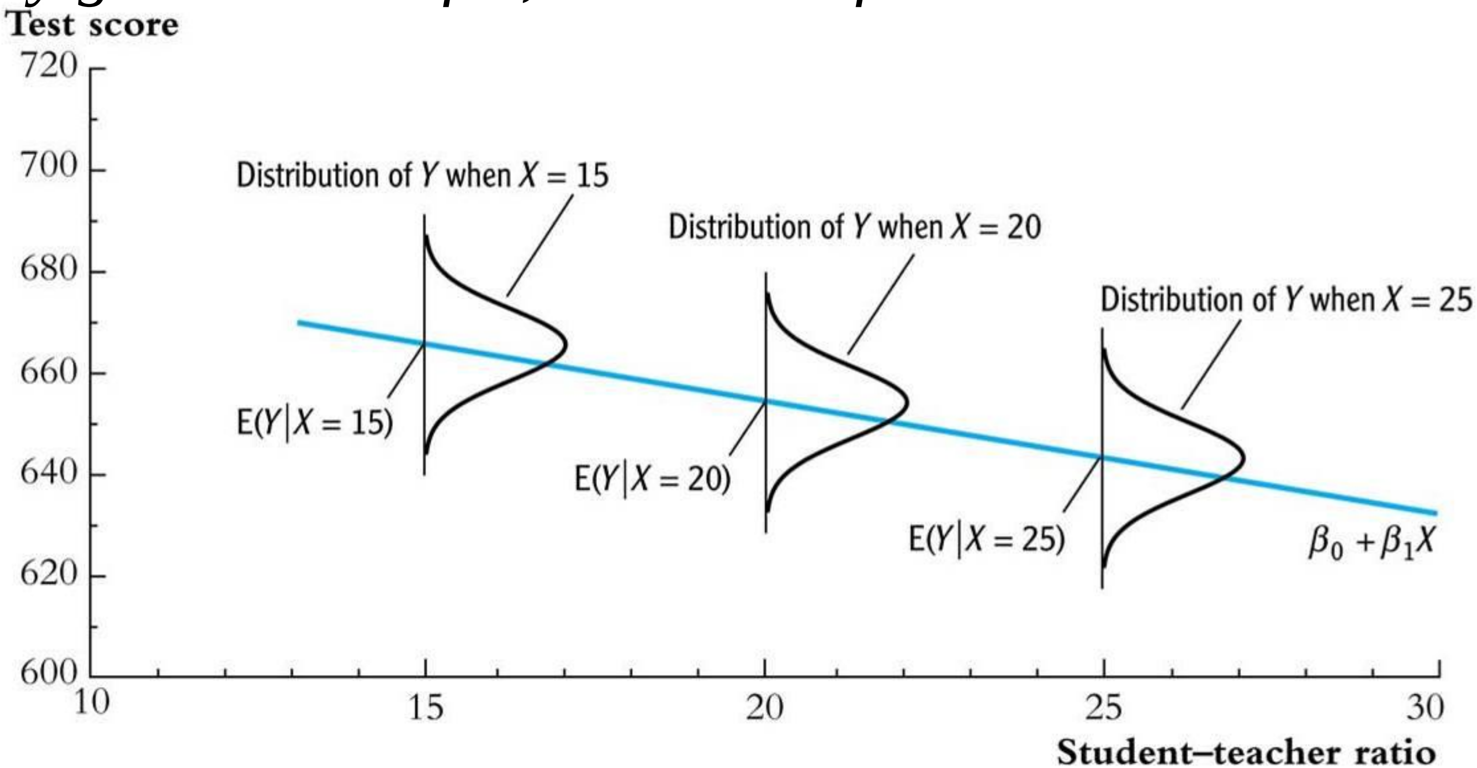
   This implies that $\hat{\beta}_1$ *is unbiased*

2. (X_i,Y_i), $i = 1,\ldots,n$, are i.i.d.
   - *This is true if X, Y are collected by simple random sampling*
   - *This delivers the sampling distribution of* $\hat{\beta}_0$ *and* $\hat{\beta}_1$

3. Large outliers in X and/or Y are rare.
   - *Technically, X and Y have finite fourth moments*
   - *Outliers can result in meaningless values of* $\hat{\beta}_1$

Least squares assumption #1: $E(u|X = x) = 0$.

For any given value of X, the mean of u is zero:



Example: $Test\ Score_i = \beta_0 + \beta_1 STR_i + u_i$, u_i = other factors

- What are some of these "other factors"?
- Is $E(u|X=x) = 0$ plausible for these other factors?

Least squares assumption #1, ctd.

A benchmark for thinking about this assumption is to consider an <u>ideal randomized controlled experiment</u>:

- X is randomly assigned to people (students randomly assigned to different size classes; patients randomly assigned to medical treatments). Randomization is done by computer – using no information about the individual.
- Because X is assigned randomly, all other individual characteristics – the things that make up u – are independently distributed of X
- Thus, in an ideal randomized controlled experiment, $\boldsymbol{E(u|X = x) = 0}$ (that is, LSA #1 holds)
- In actual experiments, or with observational data, we will need to think hard about whether $E(u|X = x) = 0$ holds.

Least squares assumption #2: (X_i, Y_i), $i = 1,\ldots,n$ are i.i.d.

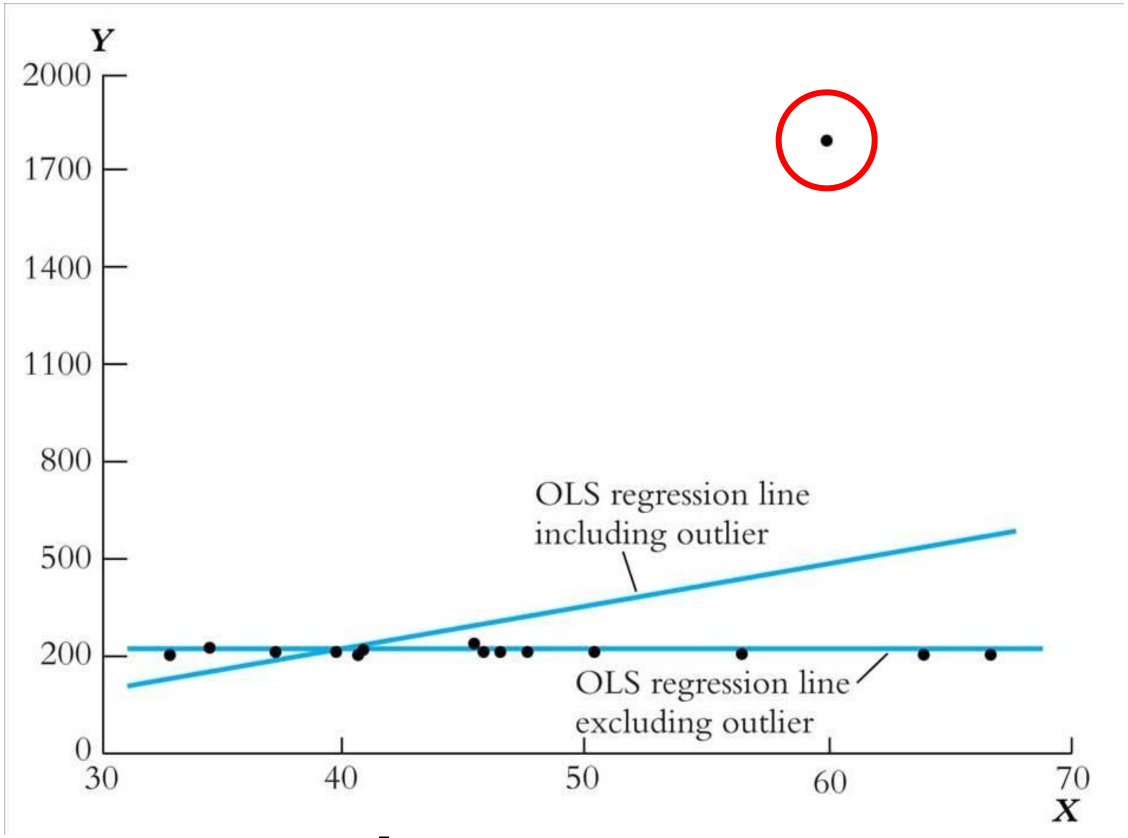
This arises automatically if the entity (individual, district) is sampled by simple random sampling: the entity is selected then, for that entity, X and Y are observed (recorded).

The main place we will encounter non-i.i.d. sampling is when data are recorded over time ("time series data") – this will introduce some extra complications.

Least squares assumption #3: *Large outliers are rare Technical statement:* $E(X^4) < \infty$ and $E(Y^4) < \infty$

- A large outlier is an extreme value of *X* or *Y*
- On a technical level, if *X* and *Y* are bounded, then they have finite fourth moments. (Standardized test scores automatically satisfy this; *STR*, family income, etc. satisfy this too).
- However, the substance of this assumption is that a large outlier can strongly influence the results

OLS can be sensitive to an outlier:



- *Is the lone point an outlier in X or Y?*
- In practice, outliers often are data glitches (coding/recording problems) – so check your data for outliers! The easiest way is to produce a scatterplot.

The Sampling Distribution of the OLS Estimator

(SW Section 4.5)

The OLS estimator is computed from a sample of data; a different sample gives a different value of $\hat{\beta}_1$. This is the source of the "sampling uncertainty" of $\hat{\beta}_1$. We want to:

- quantify the sampling uncertainty associated with $\hat{\beta}_1$
- use $\hat{\beta}_1$ to test hypotheses such as $\beta_1 = 0$
- construct a confidence interval for β_1
- All these require figuring out the sampling distribution of the OLS estimator. Two steps to get there…
    - Probability framework for linear regression
    - Distribution of the OLS estimator

Probability Framework for Linear Regression

The probability framework for linear regression is summarized by the three least squares assumptions.

Population

The group of interest (ex: all possible school districts)

Random variables*: *Y*, *X

Ex: (*Test Score*, *STR*)

Joint distribution of (Y, X)

The population regression function is linear

$E(u|X) = 0$ (1st Least Squares Assumption)

X, Y have finite fourth moments (3rd L.S.A.)

***Data Collection by simple random sampling*:**

$\{(X_i, Y_i)\}$, $i = 1,\ldots, n$, are i.i.d. (2nd L.S.A.)

The Sampling Distribution of $\hat{\beta}_1$

Like $\bar{Y}$, $\hat{\beta}_1$ has a sampling distribution.

- What is $E(\hat{\beta}_1)$? (where is it centered?)
  - If $E(\hat{\beta}_1) = \beta_1$, then OLS is unbiased – a good thing!
- What is $\text{var}(\hat{\beta}_1)$? (measure of sampling uncertainty)
- What is the distribution of $\hat{\beta}_1$ in small samples?
  - It can be very complicated in general
- What is the distribution of $\hat{\beta}_1$ in large samples?
  - It turns out to be relatively simple – in large samples, $\hat{\beta}_1$ is normally distributed.

The mean and variance of the sampling distribution of $\hat{\beta}_1$

Some preliminary algebra:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

$$\bar{Y} = \beta_0 + \beta_1 \bar{X} + \bar{u}$$

so $$Y_i - \bar{Y} = \beta_1(X_i - \bar{X}) + (u_i - \bar{u})$$

Thus,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$= \frac{\sum_{i=1}^{n}(X_i - \bar{X})[\beta_1(X_i - \bar{X}) + (u_i - \bar{u})]}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\hat{\beta}_1 = \beta_1 \frac{\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} + \frac{\sum_{i=1}^{n}(X_i - \bar{X})(u_i - \bar{u})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

so $$\hat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(u_i - \bar{u})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

Now $$\sum_{i=1}^{n}(X_i - \bar{X})(u_i - \bar{u}) = \sum_{i=1}^{n}(X_i - \bar{X})u_i - \left[\sum_{i=1}^{n}(X_i - \bar{X})\right]\bar{u}$$

$$= \sum_{i=1}^{n}(X_i - \bar{X})u_i - \left[\left(\sum_{i=1}^{n}X_i\right) - n\bar{X}\right]\bar{u}$$

$$= \sum_{i=1}^{n}(X_i - \bar{X})u_i$$

Substitute $\sum_{i=1}^{n}(X_i - \bar{X})(u_i - \bar{u}) = \sum_{i=1}^{n}(X_i - \bar{X})u_i$ into the expression for $\hat{\beta}_1 - \beta_1$:

$$\hat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(u_i - \bar{u})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

so

$$\boxed{\hat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

Now we can calculate $E(\hat{\beta}_1)$ *and* $\mathrm{var}(\hat{\beta}_1)$*:*

$$E(\hat{\beta}_1) - \beta_1 = E\left[\frac{\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]$$

$$= E\left\{E\left[\left.\frac{\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right| X_1,\ldots,X_n\right]\right\}$$

$= 0$ **because $E(u_i|X_i=x) = 0$ by LSA #1**

- Thus LSA #1 implies that $E(\hat{\beta}_1) = \beta_1$
- That is, $\hat{\beta}_1$ **is an unbiased estimator of β_1.**
- For details see App. 4.3

Next calculate var($\hat{\beta}_1$):

write

$$\hat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\frac{1}{n}\sum_{i=1}^{n} v_i}{\left(\frac{n-1}{n}\right)s_X^2}$$

where $v_i = (X_i - \bar{X})u_i$. If n is large, $s_X^2 \approx \sigma_X^2$ and $\frac{n-1}{n} \approx 1$, so

$$\hat{\beta}_1 - \beta_1 \approx \frac{\frac{1}{n}\sum_{i=1}^{n} v_i}{\sigma_X^2},$$

where $v_i = (X_i - \bar{X})u_i$ (see App. 4.3). Thus,

$$\hat{\beta}_1 - \beta_1 \approx \frac{\frac{1}{n}\sum_{i=1}^{n} v_i}{\sigma_X^2}$$

so
$$\operatorname{var}(\hat{\beta}_1 - \beta_1) = \operatorname{var}(\hat{\beta}_1)$$

$$= \frac{\operatorname{var}(v)/n}{(\sigma_X^2)^2}$$

so

$$\boxed{\operatorname{var}(\hat{\beta}_1 - \beta_1) = \frac{1}{n} \times \frac{\operatorname{var}[(X_i - \mu_x)u_i]}{\sigma_X^4}}\ .$$

<u>Summary so far</u>

- $\hat{\beta}_1$ is unbiased: $E(\hat{\beta}_1) = \beta_1$ – just like $\bar{Y}$!
- $\operatorname{var}(\hat{\beta}_1)$ is inversely proportional to n – just like $\bar{Y}$!

What is the sampling distribution of $\hat{\beta}_1$?

The exact sampling distribution is complicated – it depends on the population distribution of (Y, X) – but when n is large we get some simple (and good) approximations:

(1) Because var($\hat{\beta}_1$) $\propto$ 1/n and $E(\hat{\beta}_1) = \beta_1$, $\hat{\beta}_1 \xrightarrow{p} \beta_1$

(2) When n is large, the sampling distribution of $\hat{\beta}_1$ is well approximated by a normal distribution (CLT)

Recall the ***CLT***: suppose $\{v_i\}$, $i = 1,\ldots, n$ is i.i.d. with $E(v) = 0$ and var(v) = σ^2. Then, when n is large, $\frac{1}{n}\sum_{i=1}^{n} v_i$ is approximately distributed $N(0,\sigma_v^2/n)$.

Large-*n* approximation to the distribution of $\hat{\beta}_1$:

$$\hat{\beta}_1 - \beta_1 = \frac{\frac{1}{n}\sum_{i=1}^{n} v_i}{\left(\frac{n-1}{n}\right)s_X^2} \approx \frac{\frac{1}{n}\sum_{i=1}^{n} v_i}{\sigma_X^2}, \text{ where } v_i = (X_i - \bar{X})u_i$$

- When n is large, $v_i = (X_i - \bar{X})u_i \approx (X_i - \mu_X)u_i$, which is i.i.d. (*why*?) and $\text{var}(v_i) < \infty$ (*why*?). So, by the CLT, $\frac{1}{n}\sum_{i=1}^{n} v_i$ is approximately distributed $N(0, \sigma_v^2/n)$.
- Thus, for n large, $\hat{\beta}_1$ is approximately distributed

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma_v^2}{n\sigma_X^4}\right), \text{ where } v_i = (X_i - \mu_X)u_i$$

The larger the variance of *X*, the smaller the variance of $\hat{\beta}_1$

<u>The math</u>

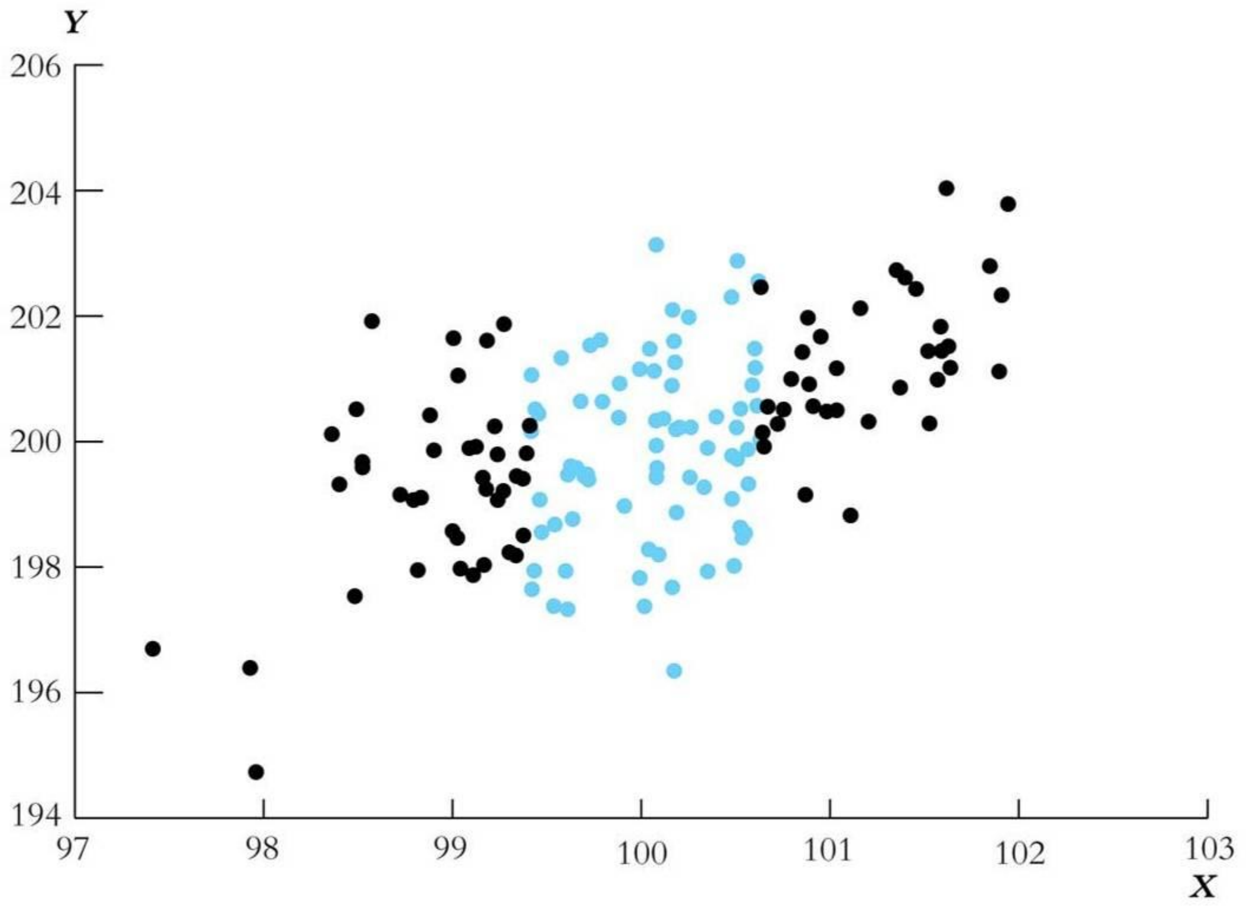
$$\text{var}(\hat{\beta}_1 - \beta_1) = \frac{1}{n} \times \frac{\text{var}[(X_i - \mu_x)u_i]}{\sigma_X^4}$$

where $\sigma_X^2 = \text{var}(X_i)$. The variance of X appears in its square in the denominator – so increasing the spread of X decreases the variance of β_1.

<u>The intuition</u>

If there is more variation in X, then there is more information in the data that you can use to fit the regression line. This is most easily seen in a figure…

The larger the variance of X, the smaller the variance of $\hat{\beta}_1$



There are the same number of black and blue dots – using which would you get a more accurate regression line?

Summary of the sampling distribution of $\hat{\beta}_1$:

If the three Least Squares Assumptions hold, then

- The exact (finite sample) sampling distribution of $\hat{\beta}_1$ has:
  - $E(\hat{\beta}_1) = \beta_1$ (that is, $\hat{\beta}_1$ is unbiased)
  - $\text{var}(\hat{\beta}_1) = \frac{1}{n} \times \frac{\text{var}[(X_i - \mu_x)u_i]}{\sigma_X^4} \propto \frac{1}{n}$.
- Other than its mean and variance, the exact distribution of $\hat{\beta}_1$ is complicated and depends on the distribution of (X,u)
- $\hat{\beta}_1 \xrightarrow{p} \beta_1$ (that is, $\hat{\beta}_1$ is consistent)
- When n is large, $\frac{\hat{\beta}_1 - E(\hat{\beta}_1)}{\sqrt{\text{var}(\hat{\beta}_1)}} \sim N(0,1)$ (CLT)
- *This parallels the sampling distribution of* $\bar{Y}$.

LARGE-SAMPLE DISTRIBUTIONS OF $\hat{\beta}_0$ AND $\hat{\beta}_1$

If the least squares assumptions in Key Concept 4.3 hold, then in large samples $\hat{\beta}_0$ and $\hat{\beta}_1$ have a jointly normal sampling distribution. The large-sample normal distribution of $\hat{\beta}_1$ is $N(\beta_1, \sigma^2_{\hat{\beta}_1})$, where the variance of this distribution, $\sigma^2_{\hat{\beta}_1}$, is

$$\sigma^2_{\hat{\beta}_1} = \frac{1}{n}\frac{\operatorname{var}[(X_i - \mu_X)u_i]}{[\operatorname{var}(X_i)]^2}. \tag{4.21}$$

The large-sample normal distribution of $\hat{\beta}_0$ is $N(\beta_0, \sigma^2_{\hat{\beta}_0})$, where

$$\sigma^2_{\hat{\beta}_0} = \frac{1}{n}\frac{\operatorname{var}(H_i u_i)}{[E(H_i^2)]^2}, \text{ where } H_i = 1 - \left(\frac{\mu_X}{E(X_i^2)}\right)X_i. \tag{4.22}$$

We are now ready to turn to hypothesis tests & confidence intervals…