DMBA: Statistics

Lecture 2: Simple Linear Regression

Least Squares, SLR properties, Inference, and Forecasting

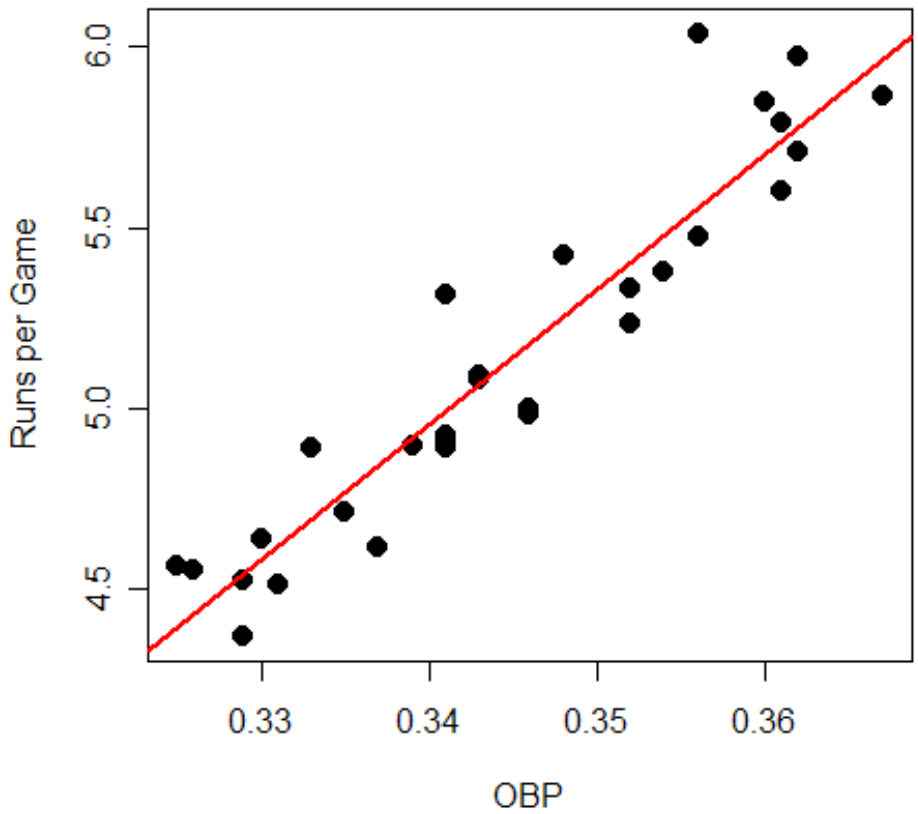
Carlos Carvalho
The University of Texas McCombs School of Business

mccombs.utexas.edu/faculty/carlos.carvalho/teaching

Today's Plan

1. The Least Squares Criteria
2. The Simple Linear Regression Model
3. Estimation for the SLR Model
   - sampling distributions
   - confidence intervals
   - hypothesis testing

Linear Prediction



$$\hat{Y}_i = b_0 + b_1 X_i$$

- b_0 is the intercept and b_1 is the slope
- We find b_0 and b_1 using *Least Squares*

The Least Squares Criterion

The formulas for b_0 and b_1 that minimize the least squares criterion are:

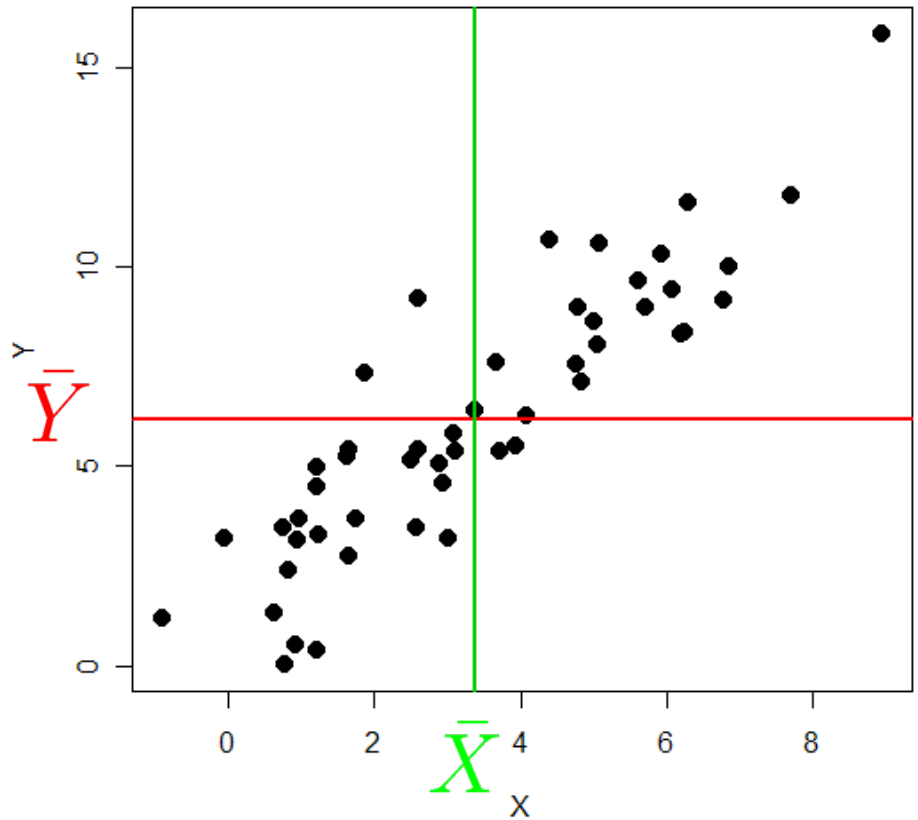
$$b_1 = corr(X, Y) \times \frac{s_Y}{s_X} \qquad b_0 = \bar{Y} - b_1 \bar{X}$$

where,

$$s_Y = \sqrt{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2} \quad \text{and} \quad s_X = \sqrt{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}$$

Correlation and Covariance

Measure the *direction* and *strength* of the linear relationship between variables Y and X



$$Cov(Y, X) = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X})}{n - 1}$$

Correlation and Covariance

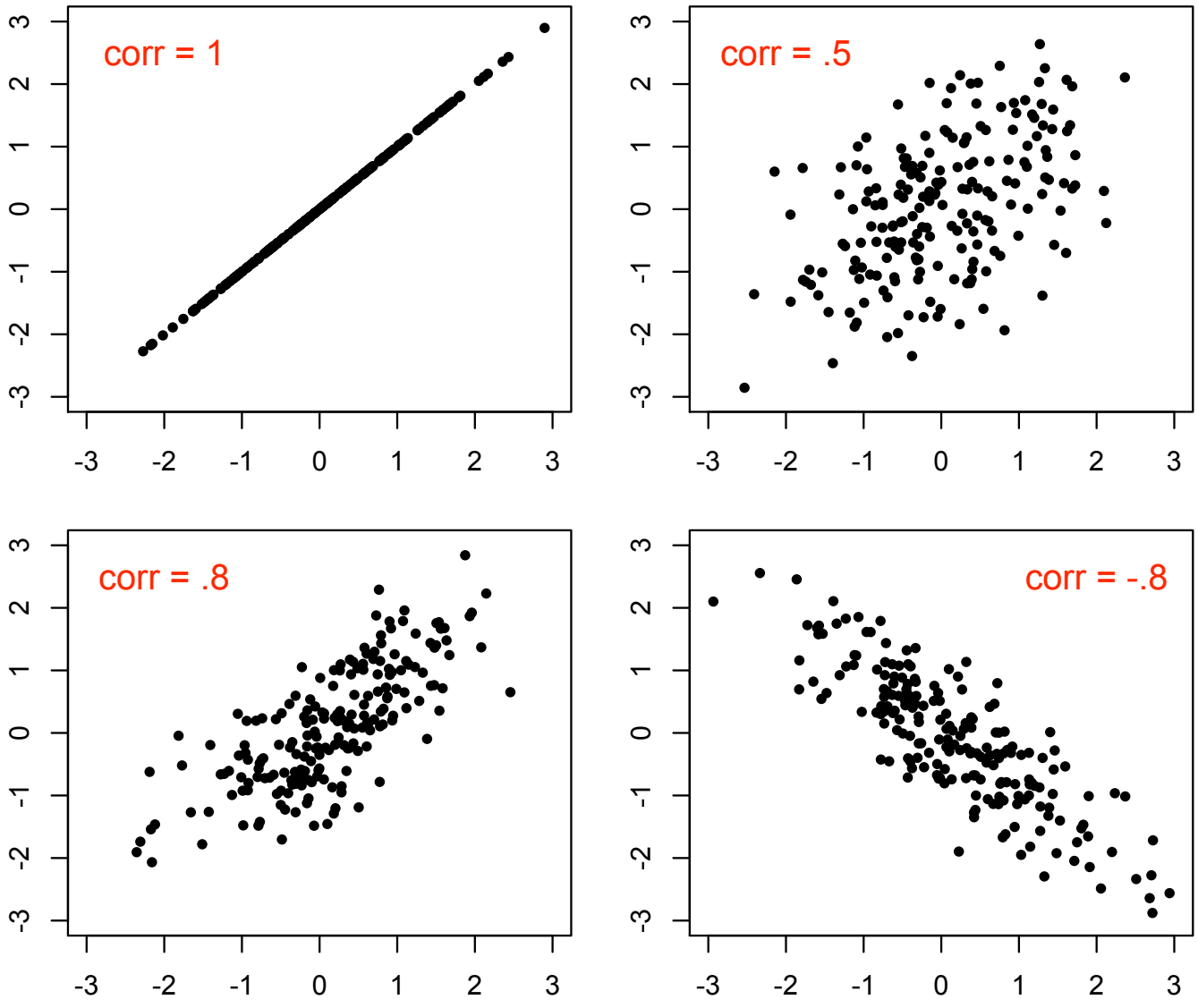
Correlation is the standardized covariance:

$$\mathrm{corr}(X, Y) = \frac{\mathrm{cov}(X, Y)}{\sqrt{\mathrm{var}(X)\mathrm{var}(Y)}} = \frac{\mathrm{cov}(X, Y)}{\mathrm{sd}(X)\mathrm{sd}(Y)}$$

The correlation is scale invariant and the units of measurement don't matter: It is always true that $-1 \leq \mathrm{corr}(X, Y) \leq 1$.

This gives the direction (- or $+$) and strength ($0 \rightarrow 1$)
of the linear relationship between X and Y.

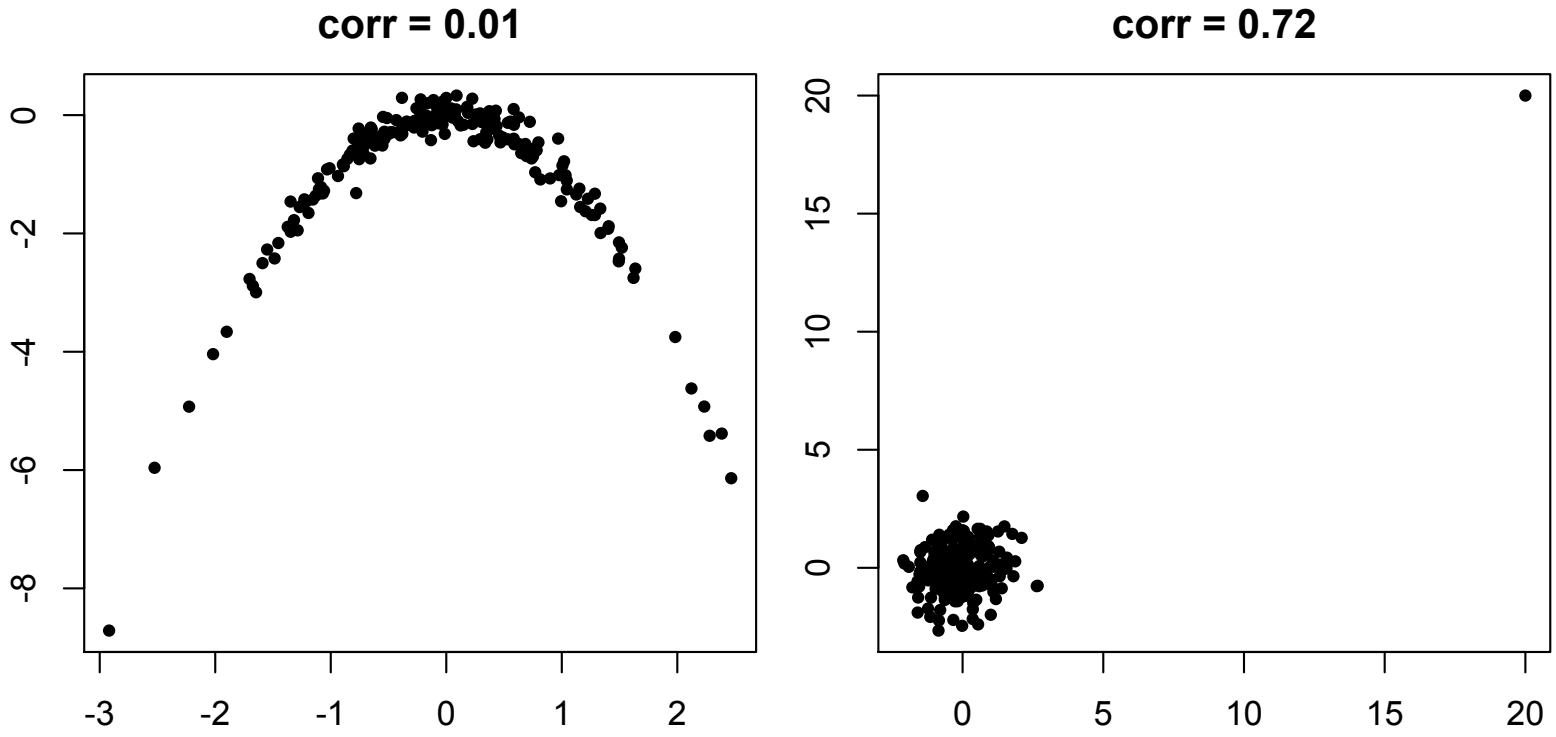
Correlation



Correlation

Only measures linear relationships:

$\text{corr}(X, Y) = 0$ does not mean the variables are not related!



Also be careful with influential observations.

Back to Least Squares

1. Intercept:

$$b_0 = \bar{Y} - b_1\bar{X} \Rightarrow \bar{Y} = b_0 + b_1\bar{X}$$

   - The point $(\bar{X}, \bar{Y})$ is on the regression line!
   - Least squares finds the point of means and rotate the line through that point until getting the "right" slope

2. Slope:

$$b_1 = corr(X, Y) \times \frac{s_Y}{s_X}$$

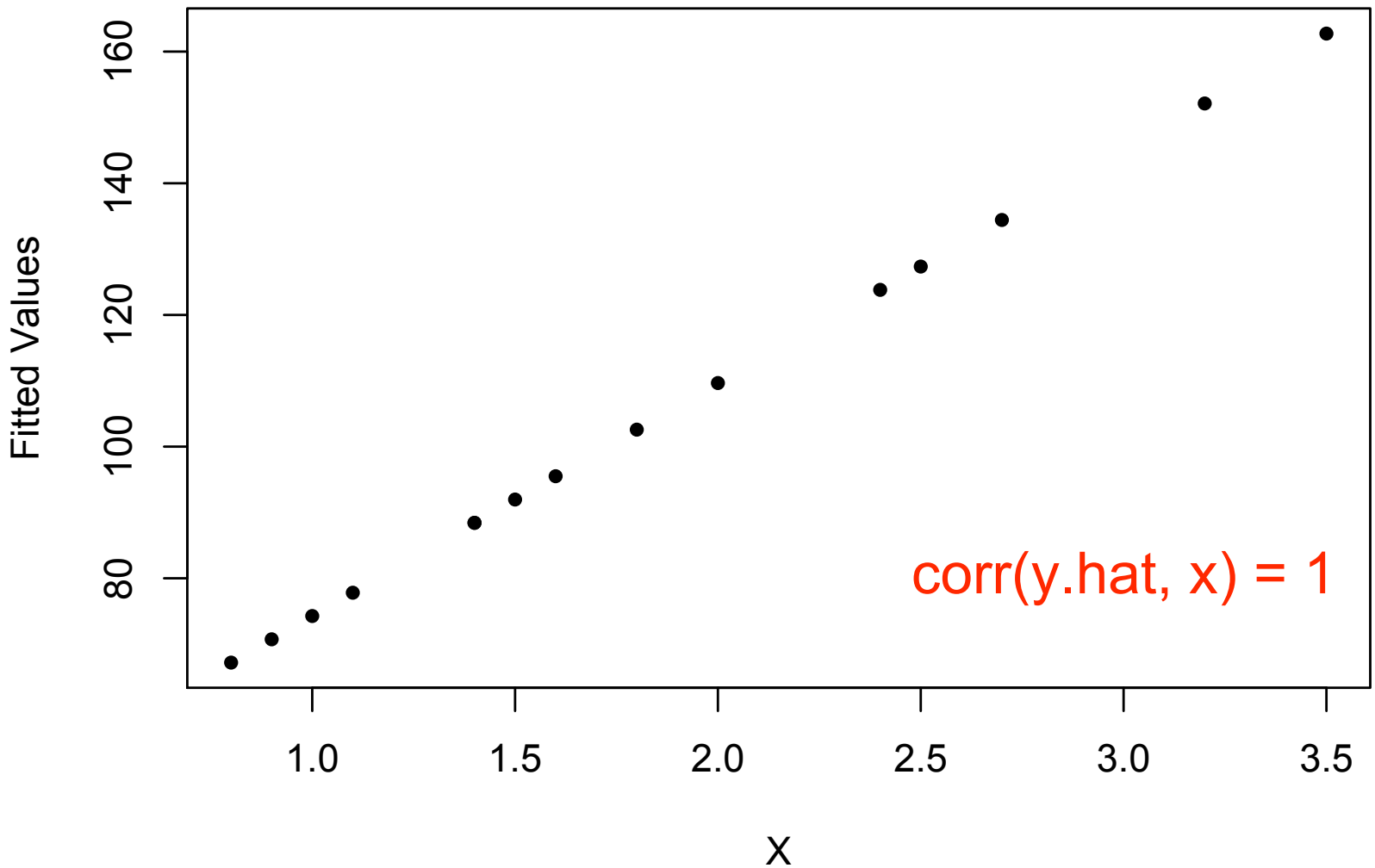
   - So, the right slope is the *correlation coefficient* times a *scaling factor* that ensures the proper units for b_1

More on Least Squares

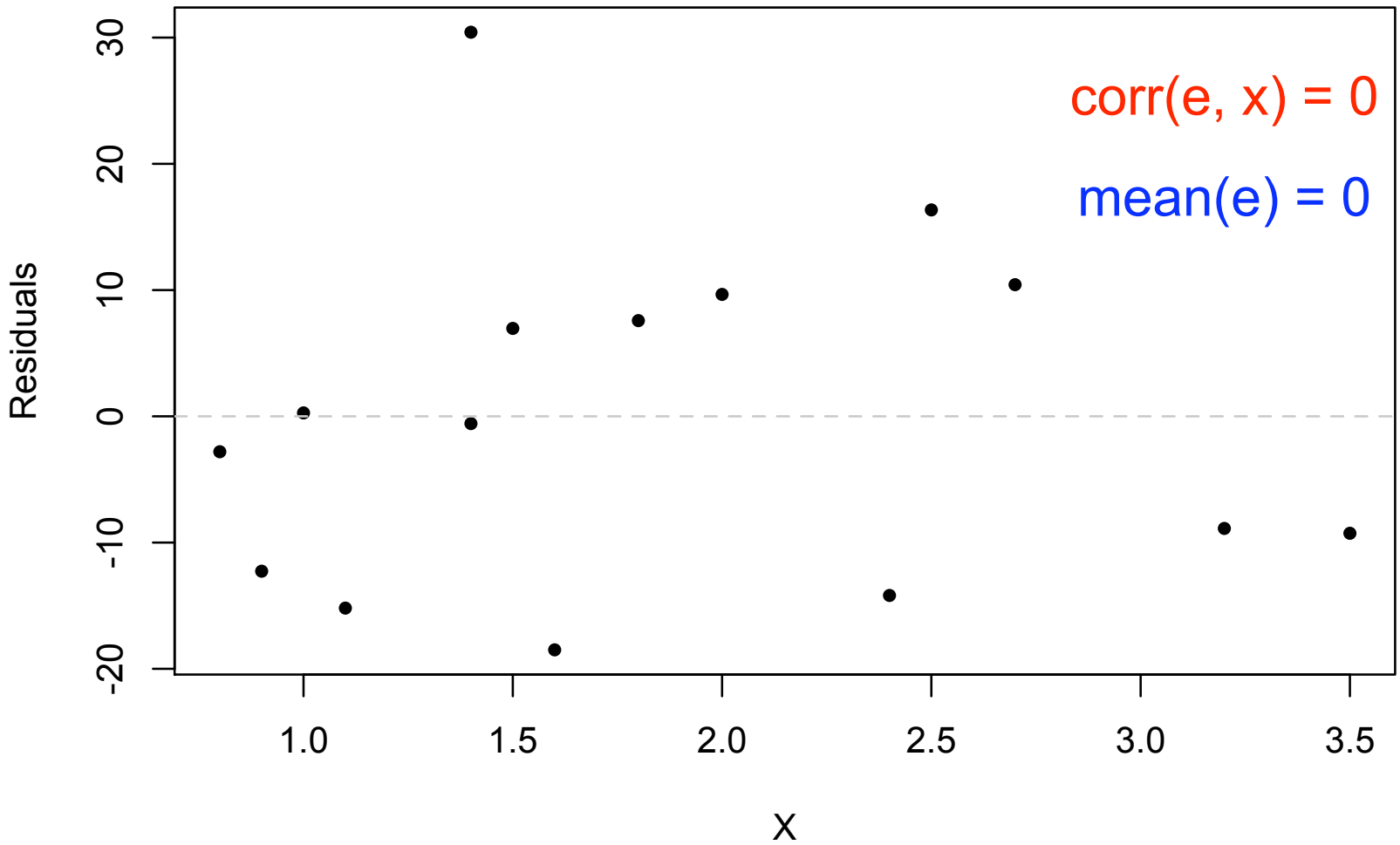
From now on, terms "fitted values" ($\hat{Y}_i$) and "residuals" (e_i) refer to those obtained from the least squares line.

The fitted values and residuals have some special properties. Lets look at the housing data analysis to figure out what these properties are...

The Fitted Values and X



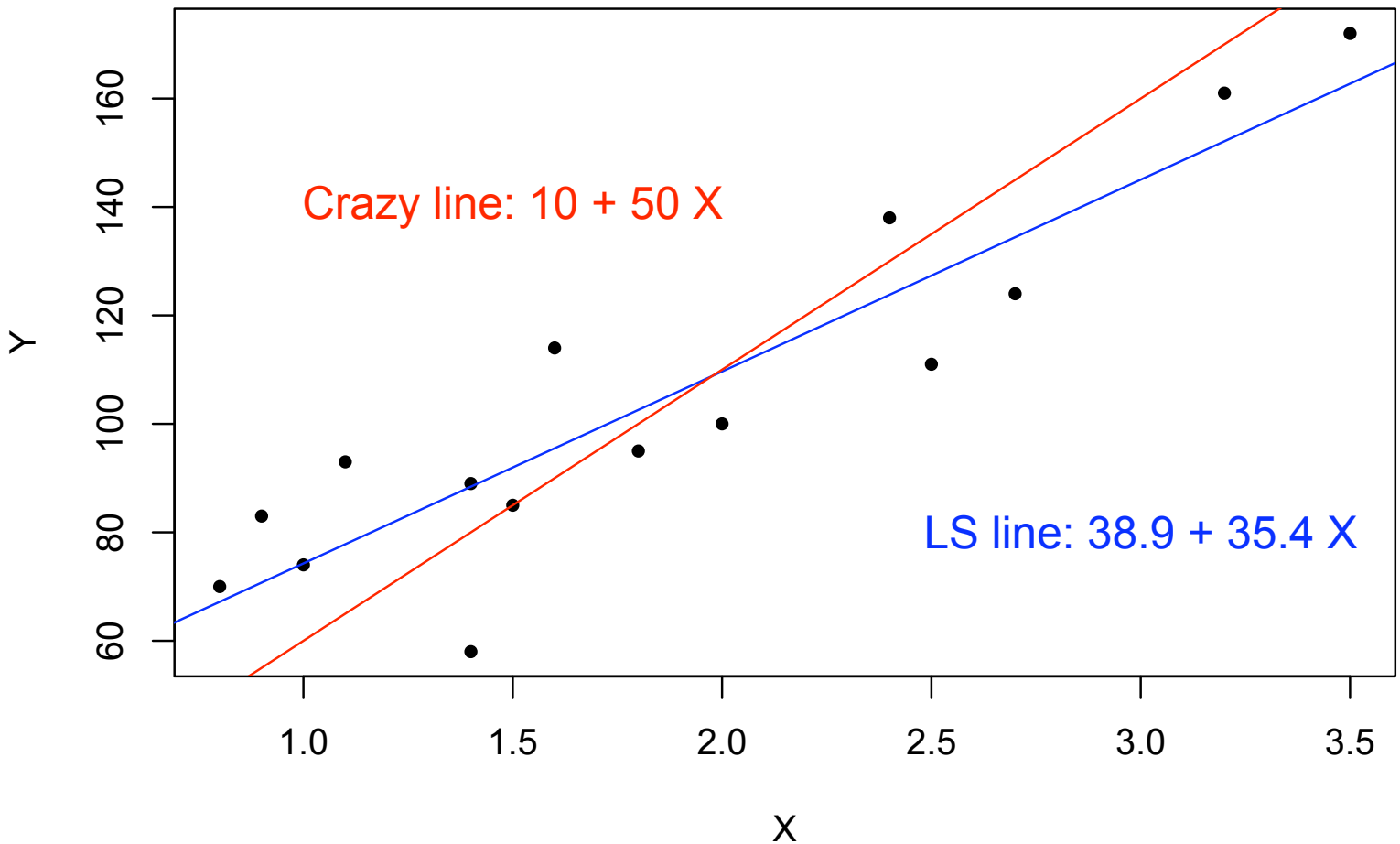
The Residuals and X



Why?

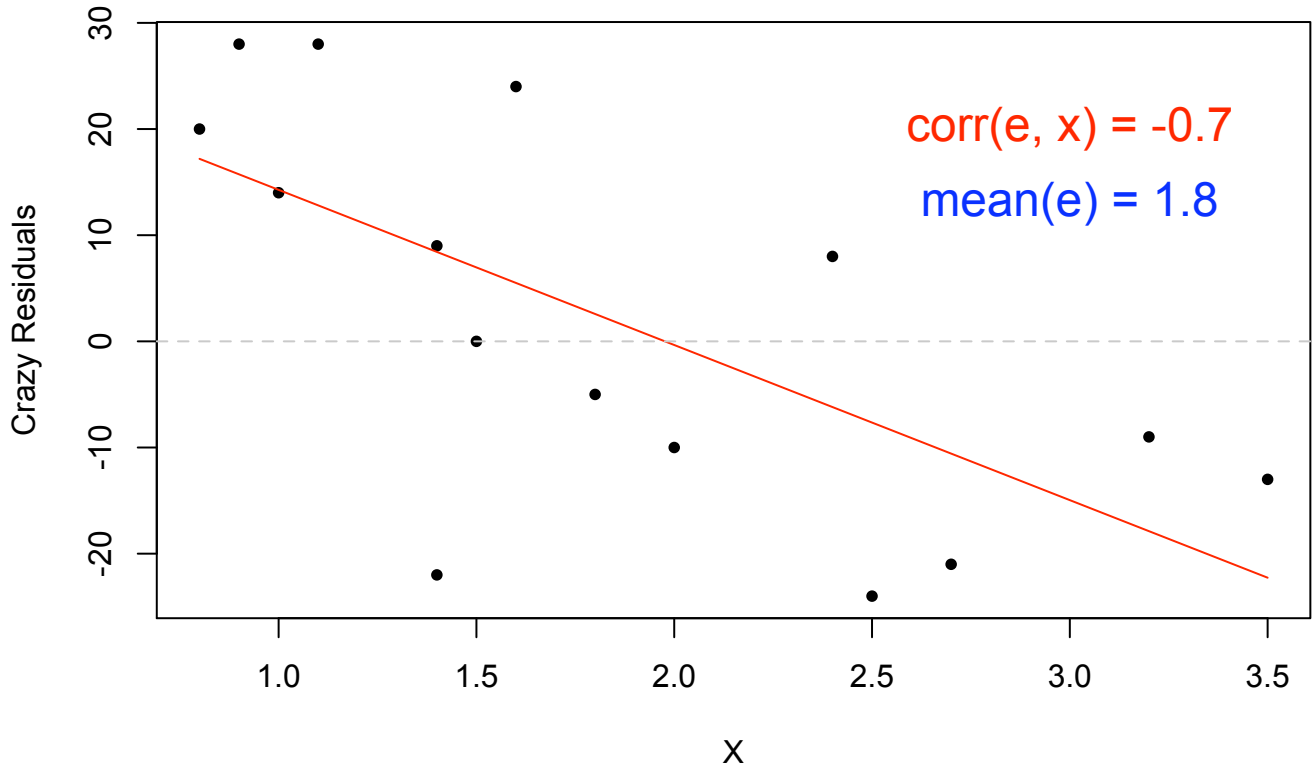
What is the intuition for the relationship between $\hat{Y}$ and e and X?

Lets consider some "crazy" alternative line:



Fitted Values and Residuals

This is a bad fit! We are underestimating the value of small houses and overestimating the value of big houses.



Clearly, we have left some predictive ability on the table!

Fitted Values and Residuals

As long as the correlation between e and X is non-zero, we could always adjust our prediction rule to do better.

We need to exploit all of the predictive power in the X values and put this into $\hat{Y}$, leaving no "*Xness*" in the residuals.

In Summary: $Y = \hat{Y} + e$ where:

- $\hat{Y}$ is "made from X"; $\text{corr}(X, \hat{Y}) = 1$.
- e is unrelated to X; $\text{corr}(X, e) = 0$.

Another way to derive things

The intercept:

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} e_i = 0 \quad &\Rightarrow \quad \frac{1}{n}\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i) \\ &\Rightarrow \quad \bar{Y} - b_0 - b_1\bar{X} = 0 \\ &\Rightarrow \quad b_0 = \bar{Y} - b_1\bar{X} \end{aligned}$$

Another way to derive things

The slope:

$$
\begin{aligned}
\mathrm{corr}(e, X) &= \sum_{i=1}^{n} e_i (X_i - \bar{X}) = 0 \\
&= \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)(X_i - \bar{X}) \\
&= \sum_{i=1}^{n} (Y_i - \bar{Y} - b_1 (X_i - \bar{X}))(X_i - \bar{X}) \\
\Rightarrow \quad b_1 &= \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = r_{xy} \frac{s_y}{s_x}
\end{aligned}
$$

Decomposing the Variance

How well does the least squares line explain variation in Y?

Since $\hat{Y}$ and e are independent (i.e. $\mathrm{cov}(\hat{Y}, e) = 0$),

$$\mathrm{var}(Y) = \mathrm{var}(\hat{Y} + e) = \mathrm{var}(\hat{Y}) + \mathrm{var}(e)$$

This leads to

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} e_i^2$$

Decomposing the Variance – ANOVA Tables

$$\underbrace{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}_{\text{Total Sum of Squares SST}} = \underbrace{\sum_{i=1}^{n}(\hat{Y}_i-\overline{Y})^2}_{\text{Regression SS SSR}} + \underbrace{\sum_{i=1}^{n}e_i^{\,2}}_{\text{Error SS SSE}}$$

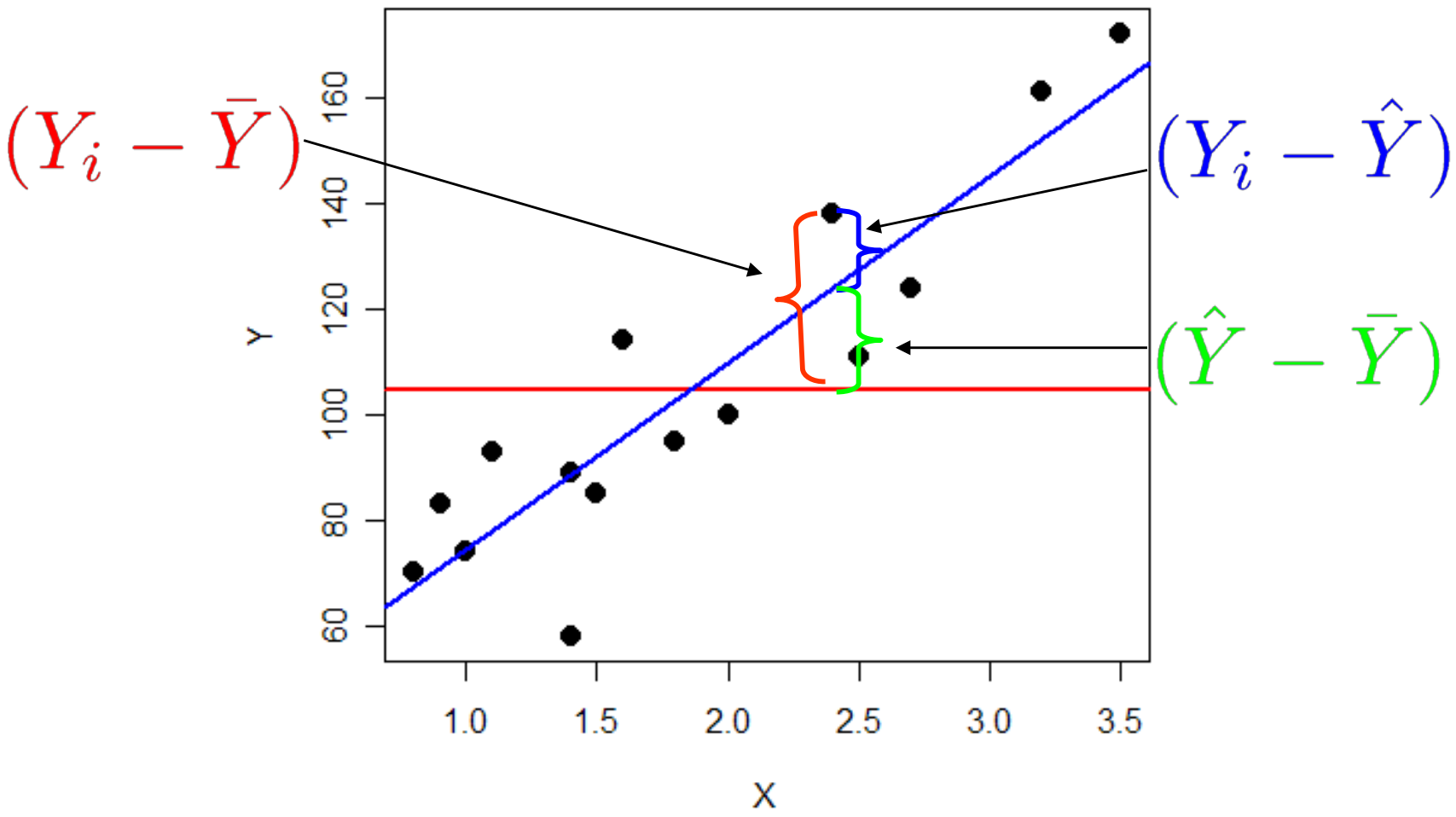
SSR: Variation in Y explained by the regression line.

SSE: Variation in Y that is left unexplained.

$$\text{SSR} = \text{SST} \Rightarrow \text{perfect fit.}$$

Be careful of similar acronyms; e.g. SSR for "residual" SS.

Decomposing the Variance – ANOVA Tables



A Goodness of Fit Measure: R^2

The coefficient of determination, denoted by R^2, measures goodness of fit:

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

- $0 < R^2 < 1$.
- The closer R^2 is to 1, the better the fit.

A Goodness of Fit Measure: R^2

An interesting fact: $R^2 = r_{xy}^2$ (i.e., R^2 is squared correlation).

$$
\begin{aligned}
R^2 &= \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \\
&= \frac{\sum_{i=1}^n (b_0 + b_1 X_i - b_0 - b_1 \bar{X})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \\
&= \frac{b_1^2 \sum_{i=1}^n (X_i - \bar{X})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = \frac{b_1^2 s_x^2}{s_y^2} = r_{xy}^2
\end{aligned}
$$

No surprise: the higher the sample correlation between X and Y, the better you are doing in your regression.

Back to the House Data

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.909209967
R Square	0.826662764
Adjusted R Square	0.81332913
Standard Error	14.13839732
Observations	15

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	12393.10771	12393.10771	61.99831126	2.65987E-06
Residual	13	2598.625623	199.8942787		
Total	14	14991.73333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	38.86468274	9.09390389	4.275906499	0.000902712	19.23849785	58.53086763	19.23849785	58.53086763
X Variable 1	35.38596255	4.494082942	7.873900638	2.65987E-06	25.67708664	45.09483846	25.67708664	45.09483846

SSR

SST

SSE

Prediction and the Modelling Goal

A prediction rule is any function where you input X and it outputs $\hat{Y}$ as a predicted response at X.

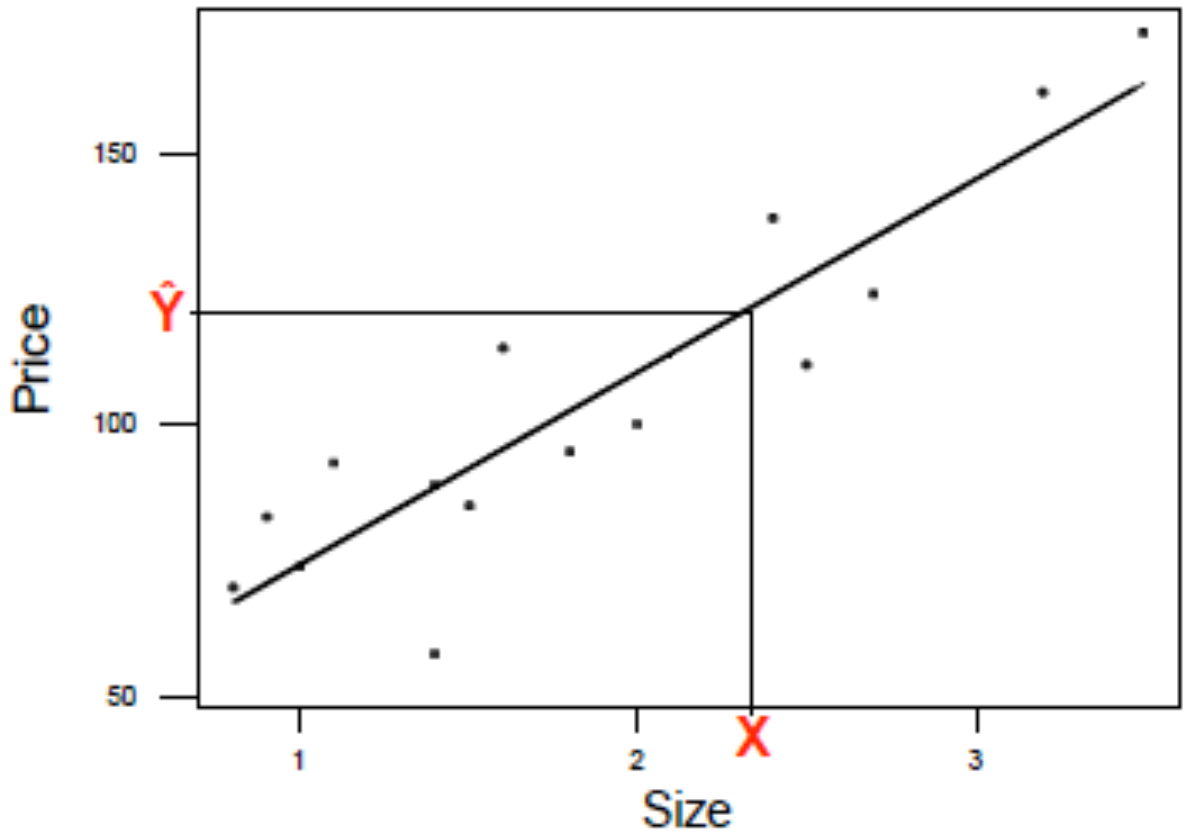
The least squares line is a prediction rule:

$$\hat{Y} = f(X) = b_0 + b_1 X$$

Prediction and the Modelling Goal

$\hat{Y}$ is not going to be a perfect prediction.

We need to devise a notion of forecast accuracy.



Prediction and the Modelling Goal

There are two things that we want to know:

- What value of Y can we expect for a given X?
- How sure are we about this forecast? Or how different could Y be from what we expect?

Our goal is to measure the accuracy of our forecasts or how much uncertainty there is in the forecast. One method is to specify a range of Y values that are likely, given an X value.

Prediction Interval: probable range for Y-values given X

Prediction and the Modelling Goal

Key Insight: To construct a prediction interval, we will have to assess the likely range of residual values corresponding to a Y value that has not yet been observed!

We will build a probability model (e.g., normal distribution).

Then we can say something like "with 95% probability the residuals will be no less than -$28,000 or larger than $28,000".

We must also acknowledge that the "fitted" line may be fooled by particular realizations of the residuals.

The Simple Linear Regression Model

The power of statistical inference comes from the ability to make precise statements about the accuracy of the forecasts.

In order to do this we must invest in a probability model.

Simple Linear Regression Model: $Y = \beta_0 + \beta_1 X + \varepsilon$

$$\varepsilon \sim \mathrm{N}(0, \sigma^2)$$

The error term ε is independent "idosyncratic noise".

Independent Normal Additive Error

Why do we have $\varepsilon \sim \mathrm{N}(0, \sigma^2)$?

- $E[\varepsilon] = 0 \Leftrightarrow E[Y \mid X] = \beta_0 + \beta_1 X$
  ($E[Y \mid X]$ is "conditional expectation of Y given X").
- Many things are close to Normal (central limit theorem).
- MLE estimates for β's are the same as the LS b's.
- It works! This is a very robust model for the world.

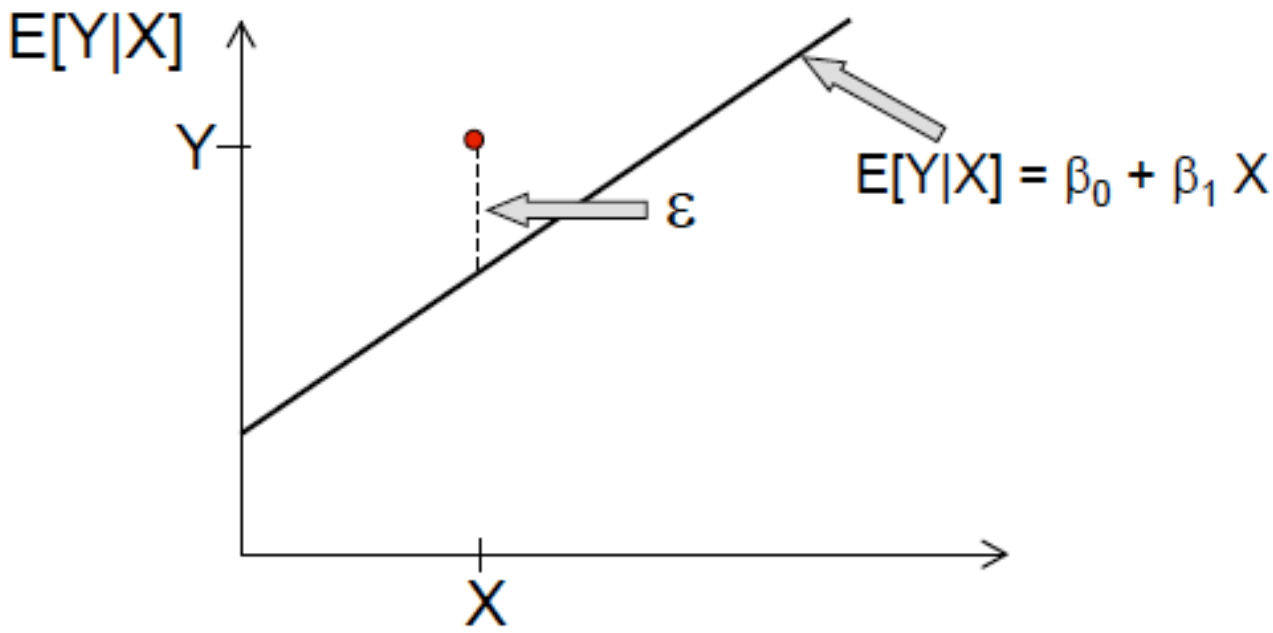
We can think of $\beta_0 + \beta_1 X$ as the "true" regression line.

The Regression Model and our House Data

Think of $E[Y|X]$ as the average price of houses with size X:

Some houses could have a higher than expected value, some lower, and the true line tells us what to expect on average.

The error term represents influence of factors other X.

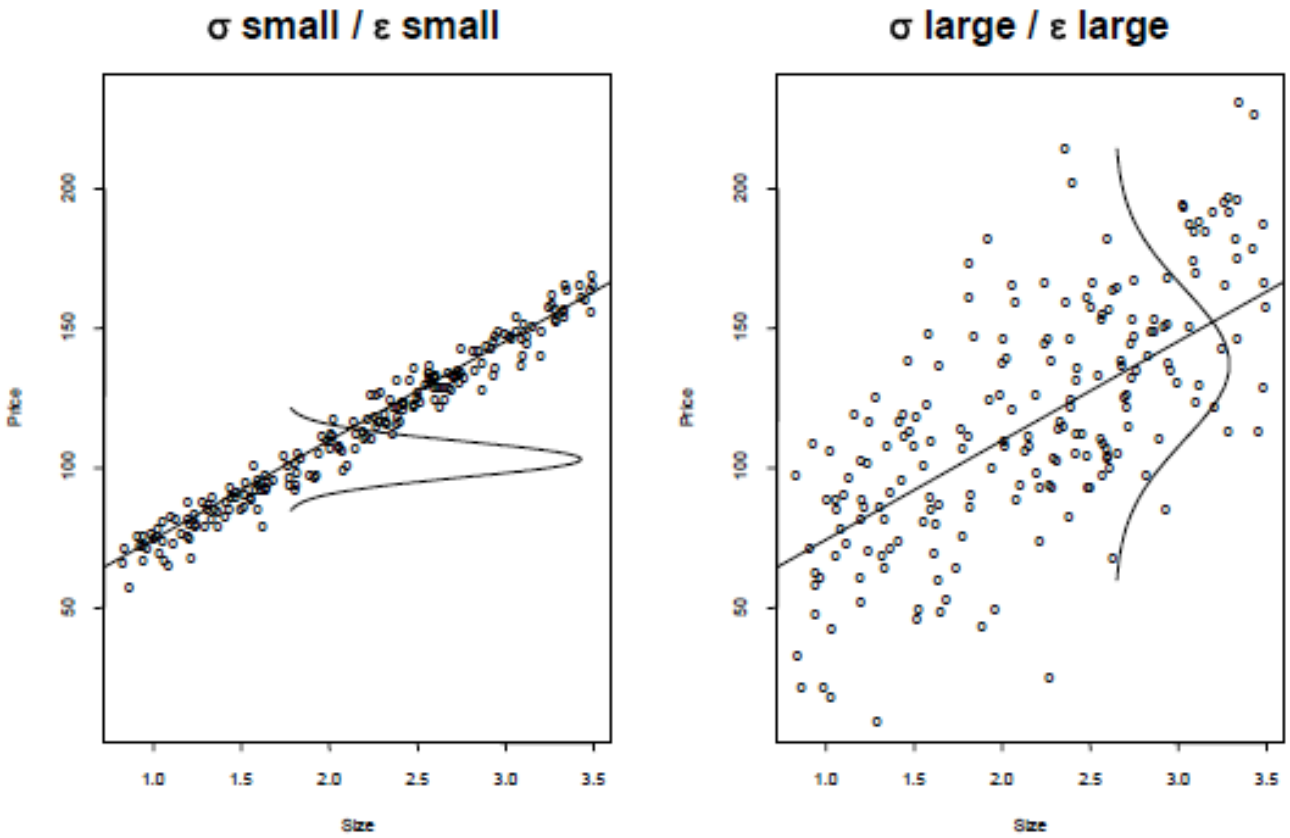


Conditional Distributions

The conditional distribution for Y given X is Normal:

$$Y|X \sim \mathrm{N}(\beta_0 + \beta_1 X, \sigma^2).$$

σ controls dispersion:



Conditional vs Marginal Distributions

More on the conditional distribution:

$$Y|X \sim \mathrm{N}(E[Y|X], \mathrm{var}(Y|X)).$$

- Mean is $E[Y|X] = E[\beta_0 + \beta_1 X + \varepsilon] = \beta_0 + \beta_1 X$.
- Variance is $\mathrm{var}(\beta_0 + \beta_1 X + \varepsilon) = \mathrm{var}(\varepsilon) = \sigma^2$.

Remember our sliced boxplots:

- $\sigma^2 < \mathrm{var}(Y)$ if X and Y are related.

Prediction Intervals with the True Model

You are told (without looking at the data) that

$$\beta_0 = 40;\ \beta_1 = 45;\ \sigma = 10$$

and you are asked to predict price of a 1500 square foot house.

What do you know about Y from the model?

$$\begin{aligned} Y &= 40 + 45(1.5) + \varepsilon \\ &= 107.5 + \varepsilon \end{aligned}$$

Thus our prediction for price is

$$Y \sim \mathrm{N}(107.5, 10^2)$$

Prediction Intervals with the True Model

The model says that the mean value of a 1500 sq. ft. house is \$107,500 and that deviation from mean is within $\approx$ \$20,000.

We are 95% sure that

- $-20 < \varepsilon < 20$
- $\$87,500 < Y < \$127,500$

In general, the 95 % Prediction Interval is $\text{PI} = \beta_0 + \beta_1 X \pm 2\sigma$.

Summary of Simple Linear Regression

Assume that all observations are drawn from our regression model and that errors on those observations are independent.

The model is

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

where ε is independent and identically distributed $\mathrm{N}(0, \sigma^2)$.

The SLR has 3 basic parameters:

- β_0, β_1 (linear pattern)
- σ (variation around the line).

Key Characteristics of Linear Regression Model

- Mean of Y is linear in X.
- Error terms (deviations from line) are normally distributed (very few deviations are more than 2 sd away from the regression mean).
- Error terms have constant variance.

Break

Back in 15 minutes...

Recall: Estimation for the SLR Model

SLR assumes every observation in the dataset was generated by the model:

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

This is a model for the conditional distribution of Y given X.

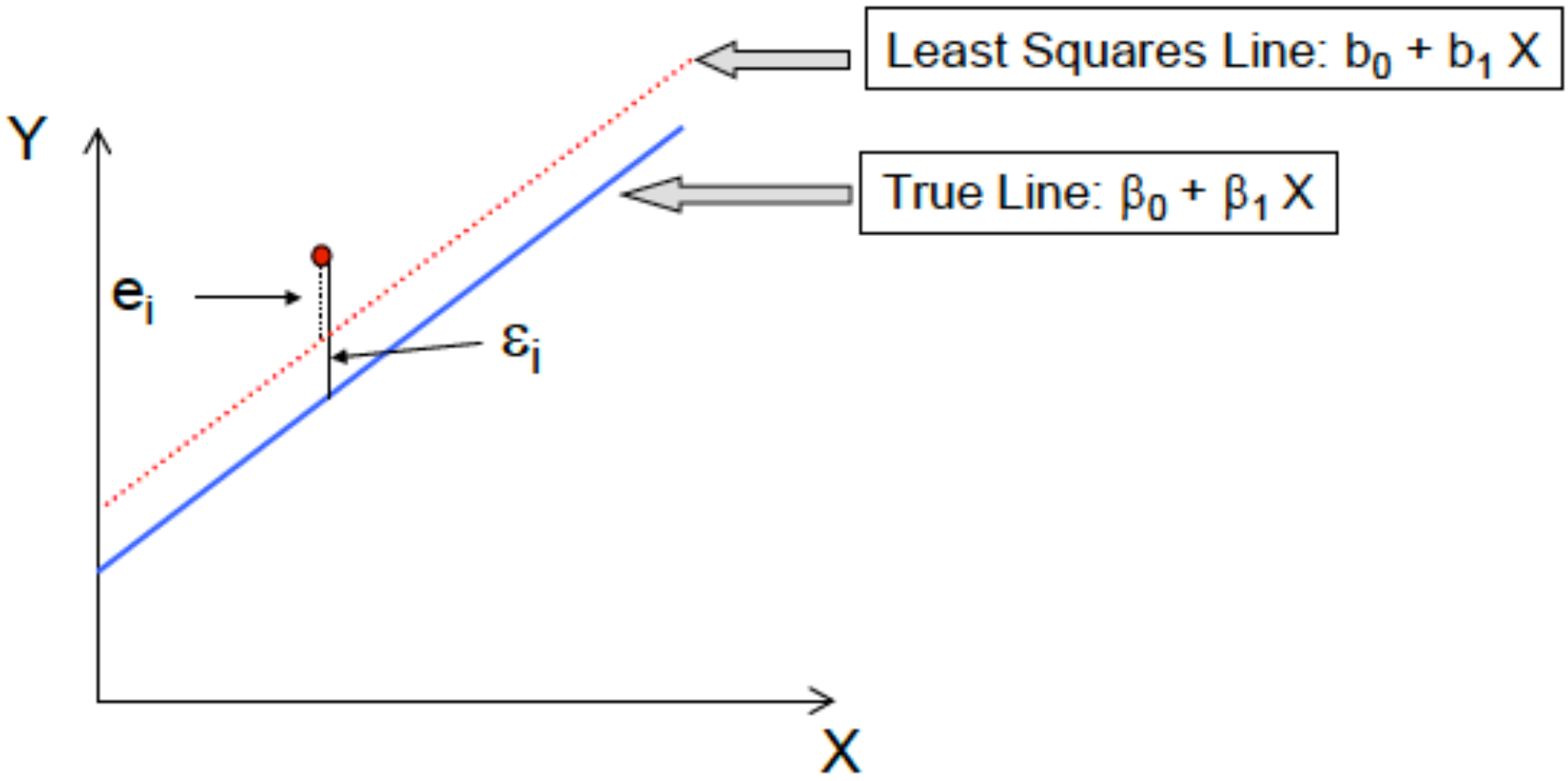
We use Least Squares to estimate β_0 and β_1:

$$\hat{\beta}_1 = b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = b_0 = \bar{Y} - b_1 \bar{X}$$

Estimation for the SLR Model

NOTE!!: β_0 **is not** b_0, β_1 **is not** b_1 **and** ε_i **is not e**



Estimation of Error Variance

Recall that $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, and that σ drives the width of the prediction intervals:

$$\sigma^2 = \text{var}(\varepsilon_i) = E[(\varepsilon_i - E[\varepsilon_i])^2] = E[\varepsilon_i^2]$$

A sensible strategy would be to estimate the average for squared errors with the sample average squared residuals:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2$$

Estimation of Error Variance

However, this is not an unbiased estimator of σ^2. We have to alter the denominator slightly:

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2 = \frac{SSE}{n-2}$$

(2 is the number of regression coefficients; i.e. 2 for $\beta_0 + \beta_1$).

We have $n-2$ degrees of freedom because 2 have been "used up" in the estimatiation of b_0 and b_1.

We usually use $s = \sqrt{SSE/(n-2)}$, in the same units as Y.

Degrees of Freedom

Degrees of Freedom is the number of times you get to observe useful information about the variance you're trying to estimate.

For example, consider $SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$:

- If $n = 1$, $\bar{Y} = Y_1$ and $SST = 0$: since Y_1 is "used up" estimating the mean, we haven't observed any variability!
- For $n > 1$, we've only had $n - 1$ chances for deviation from the mean, and we estimate $s_y^2 = SST/(n-1)$.

In regression with p coefficients (e.g., $p = 2$ in SLR), you only get $n - p$ real observations of variability $\Rightarrow DoF = n - p$.

Estimation of Error Variance

Where is s in the Excel output?

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.909209967
R Square	0.826662764
Adjusted R Square	0.81332913
Standard Error	14.13839732
Observations	15

s

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	12393.10771	12393.10771	61.99831126	2.65987E-06
Residual	13	2598.625623	199.8942787		
Total	14	14991.73333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	38.88468274	9.09390389	4.275906499	0.000902712	19.23849785	58.53086763	19.23849785	58.53086763
X Variable 1	35.38596255	4.494082942	7.873900638	2.65987E-06	25.67708664	45.09483846	25.67708664	45.09483846

Remember that whenever you see "standard error" read it as estimated standard deviation: σ is the standard deviation.

Sampling Distribution of Least Squares Estimates

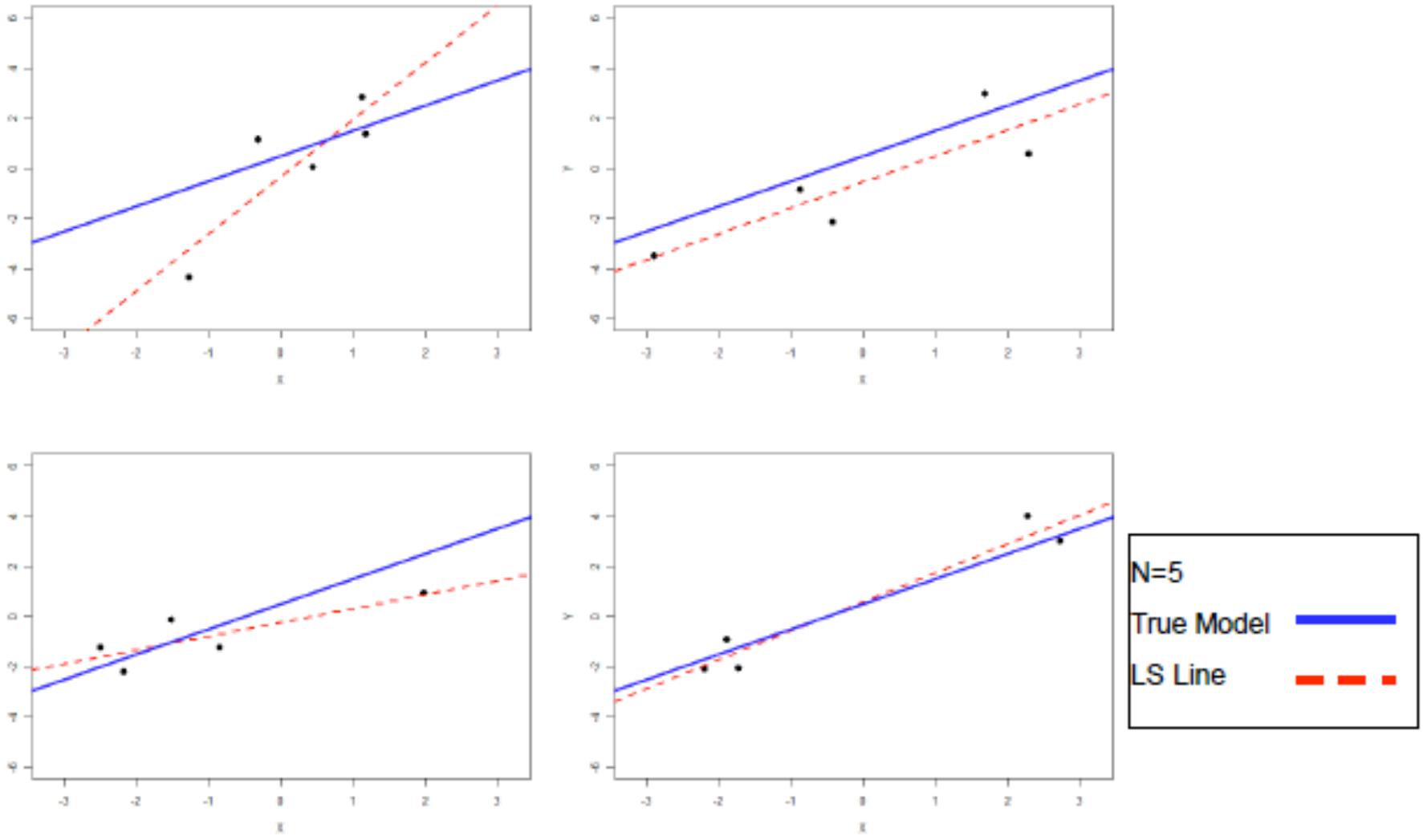
How much do our estimates depend on the particular random sample that we happen to observe? Imagine:

- Randomly draw different samples of the same size.
- For each sample, compute the estimates b_0, b_1, and s.

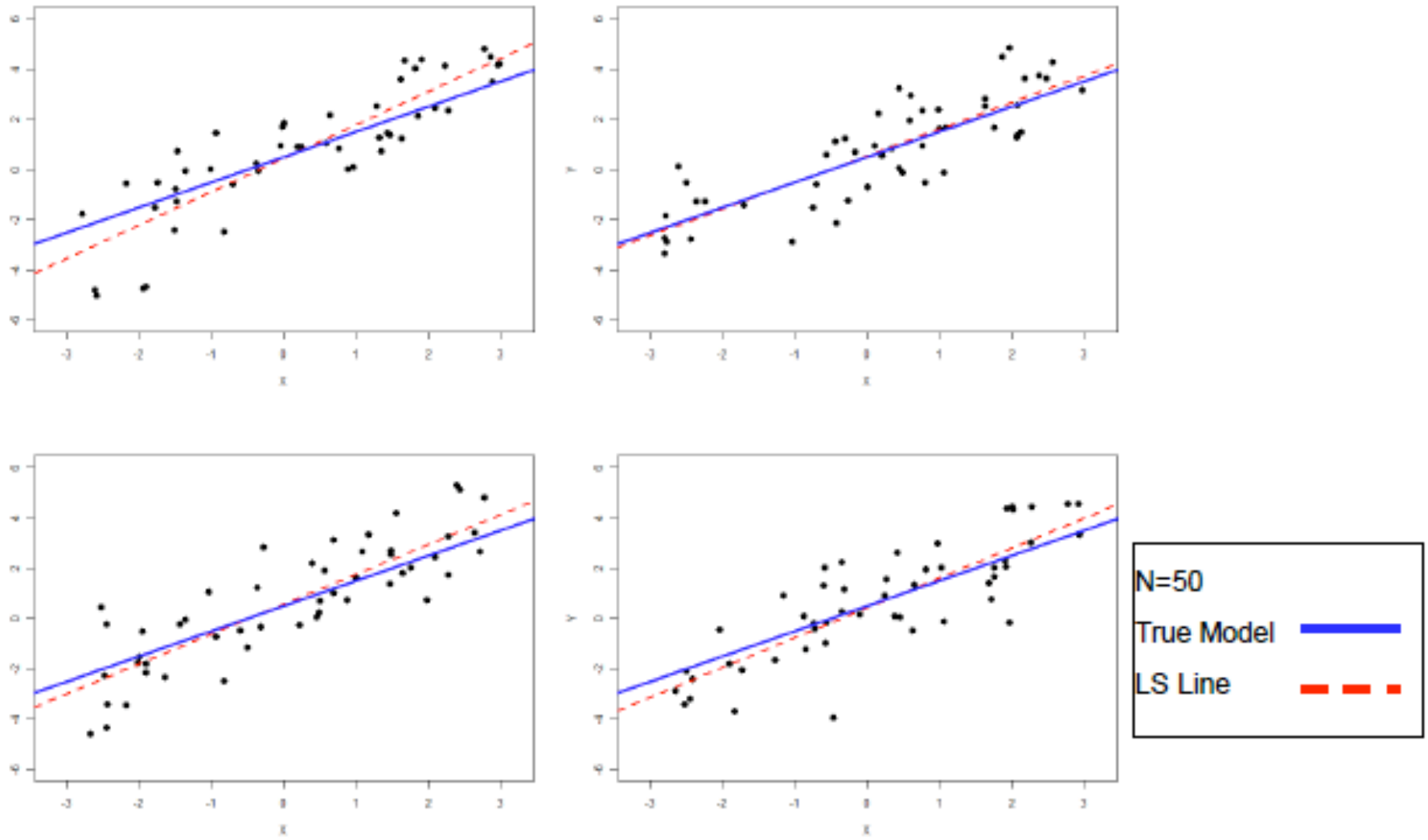
If the estimates don't vary much from sample to sample, then it doesn't matter which sample you happen to observe.

If the estimates do vary a lot, then it matters which sample you happen to observe.

Sampling Distribution of Least Squares Estimates



Sampling Distribution of Least Squares Estimates



Sampling Distribution of Least Squares Estimates

LS lines are much closer to the true line when $n = 50$.

For $n = 5$, some lines are close, others aren't:

we need to get "lucky"

Review: Sampling Distribution of Sample Mean

Step back for a moment and consider the mean for an *iid* sample of n observations of a random variable $\{X_1, \ldots, X_n\}$

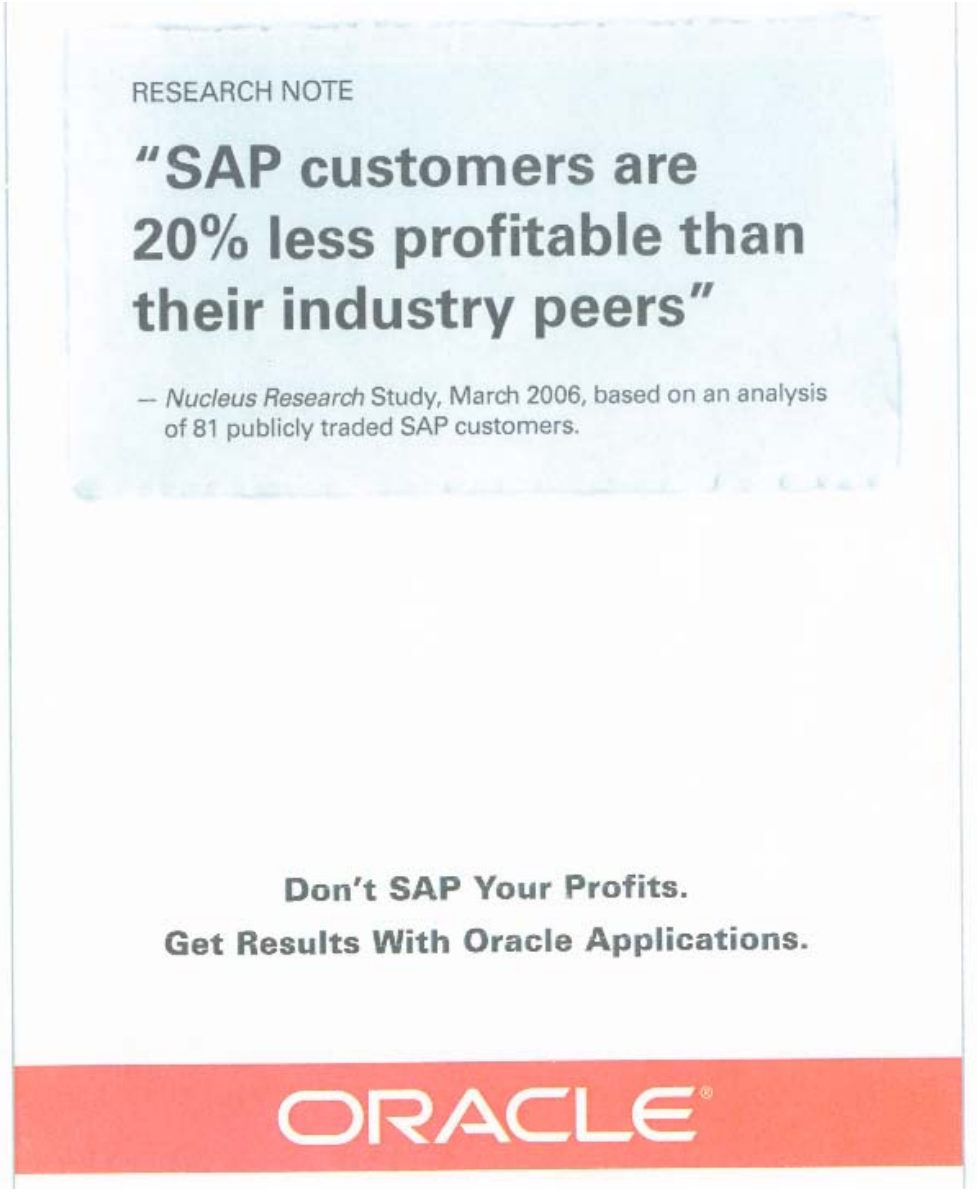
Suppose that $E(X_i) = \mu$ and $\text{var}(X_i) = \sigma^2$

- $E(\bar{X}) = \frac{1}{n} \sum E(X_i) = \mu$
- $\text{var}(\bar{X}) = \text{var}\left(\frac{1}{n} \sum X_i\right) = \frac{1}{n^2} \sum \text{var}\left(X_i\right) = \dfrac{\sigma^2}{n}$

If X is normal, then $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.

If X is not normal, we have the central limit theorem (more in a minute)!

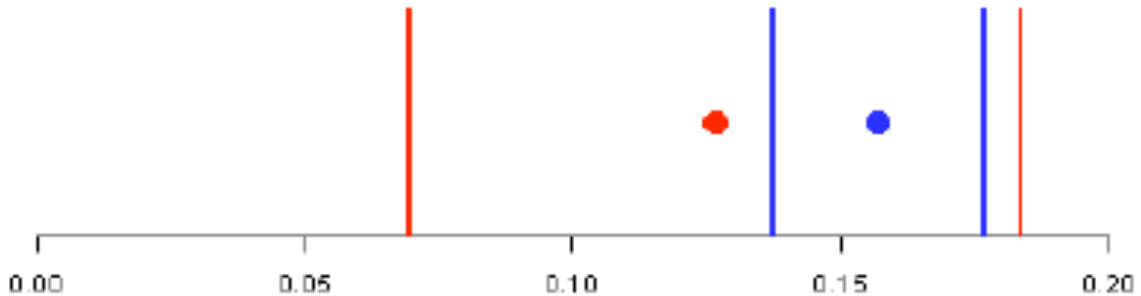
Oracle vs SAP Example (understanding variation)



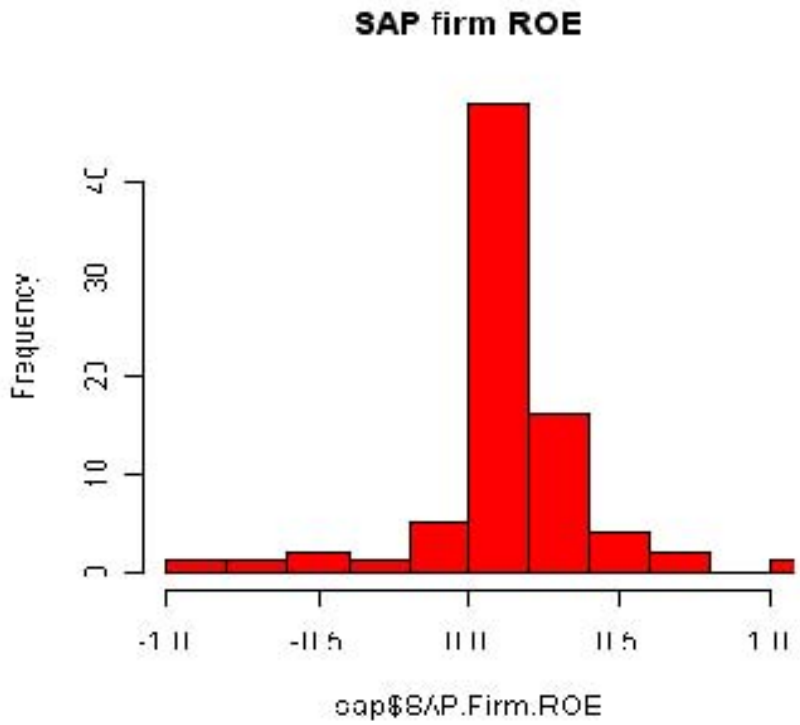
Oracle vs SAP

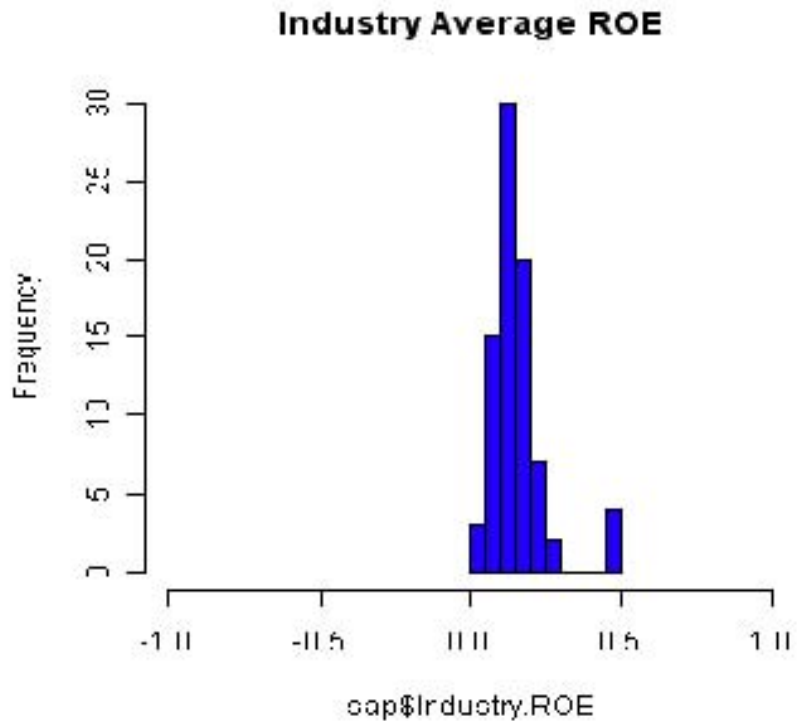
SAP Firm ROE		Industry ROE	
Mean	0.12637037	Mean	0.156987654
Standard Error	0.028509439	Standard Error	0.009905767
Median	0.134	Median	0.14
Mode	0.083	Mode	0.195
Standard Deviation	0.256584949	Standard Deviation	0.089151906
Sample Variance	0.065835836	Sample Variance	0.007948062

$$\frac{0.126}{0.157} \approx 0.8$$ OK, but: $$2(0.0099) = 0.0198$$ $$2(0.0285) = 0.057$$



Oracle vs SAP





Do you really believe that SAP affects ROE?

How else could we look at this question?

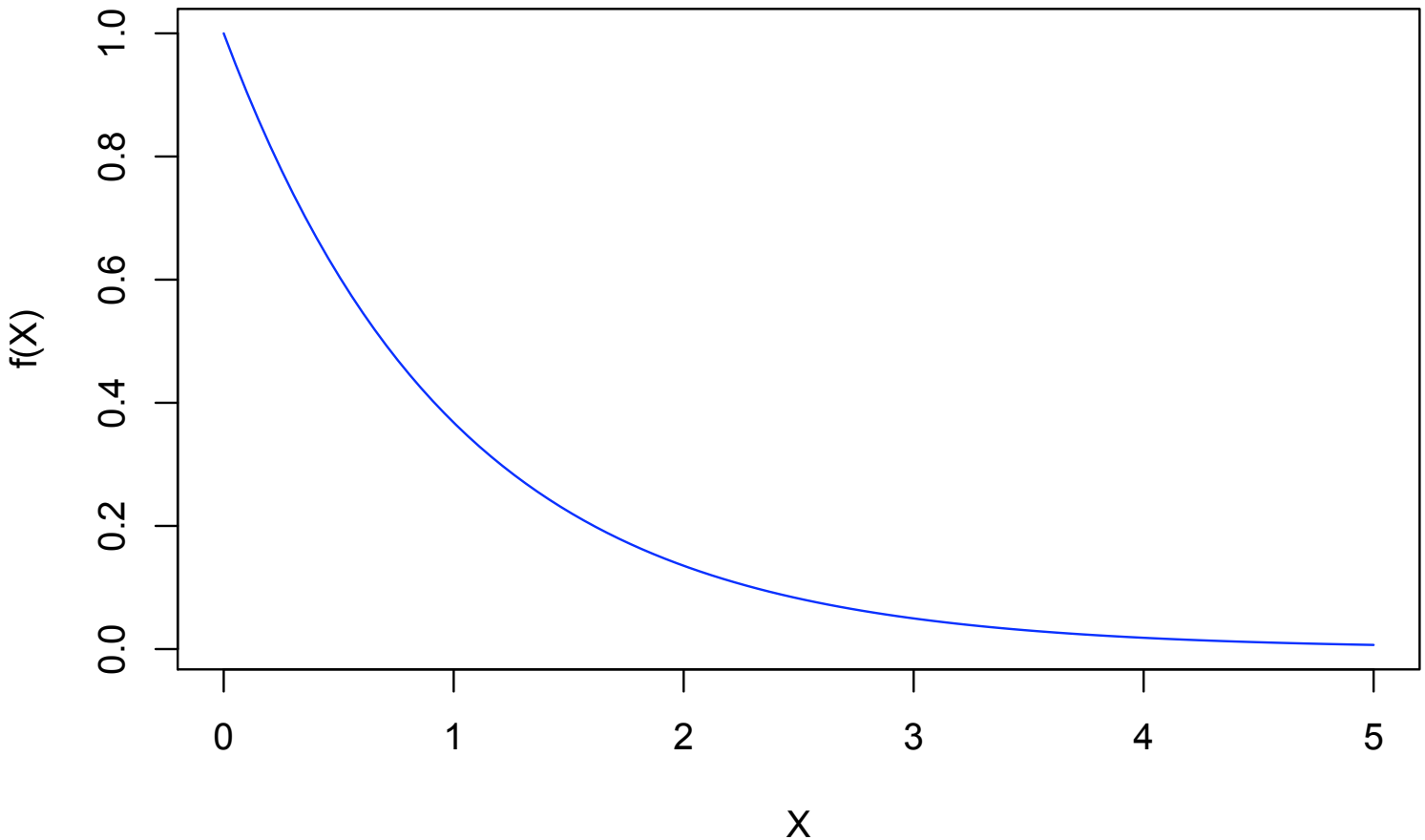
Central Limit Theorem

Simple CLT states that for *iid* random variables, X, with mean μ and variance σ^2, the distribution of the sample mean becomes normal as the number of observations, n, gets large.

That is, $\bar{X} \rightarrow_n N(\mu, \frac{\sigma^2}{n})$, and sample averages tend to be normally distributed in large samples.

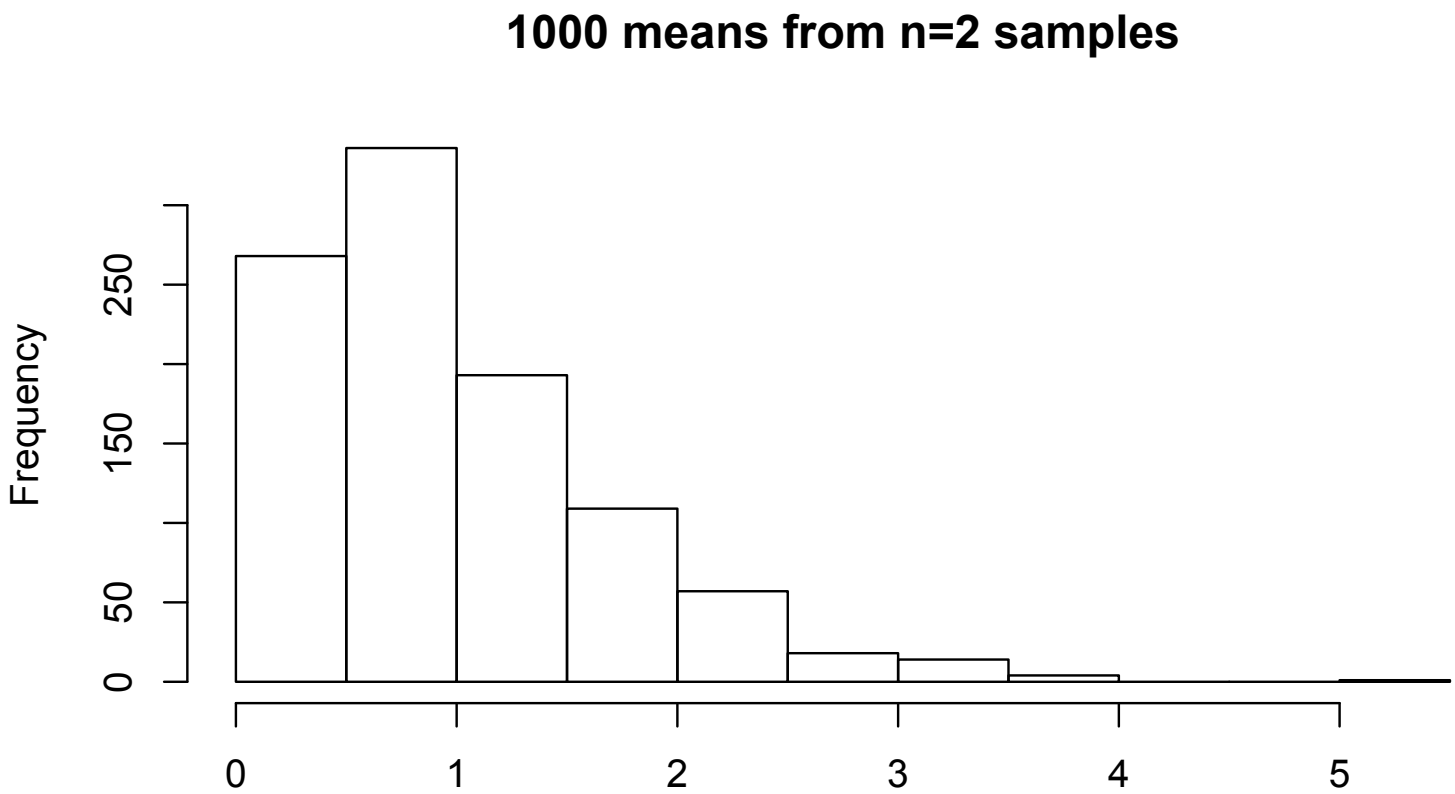
Central Limit Theorem

Exponential random variables don't look very normal:

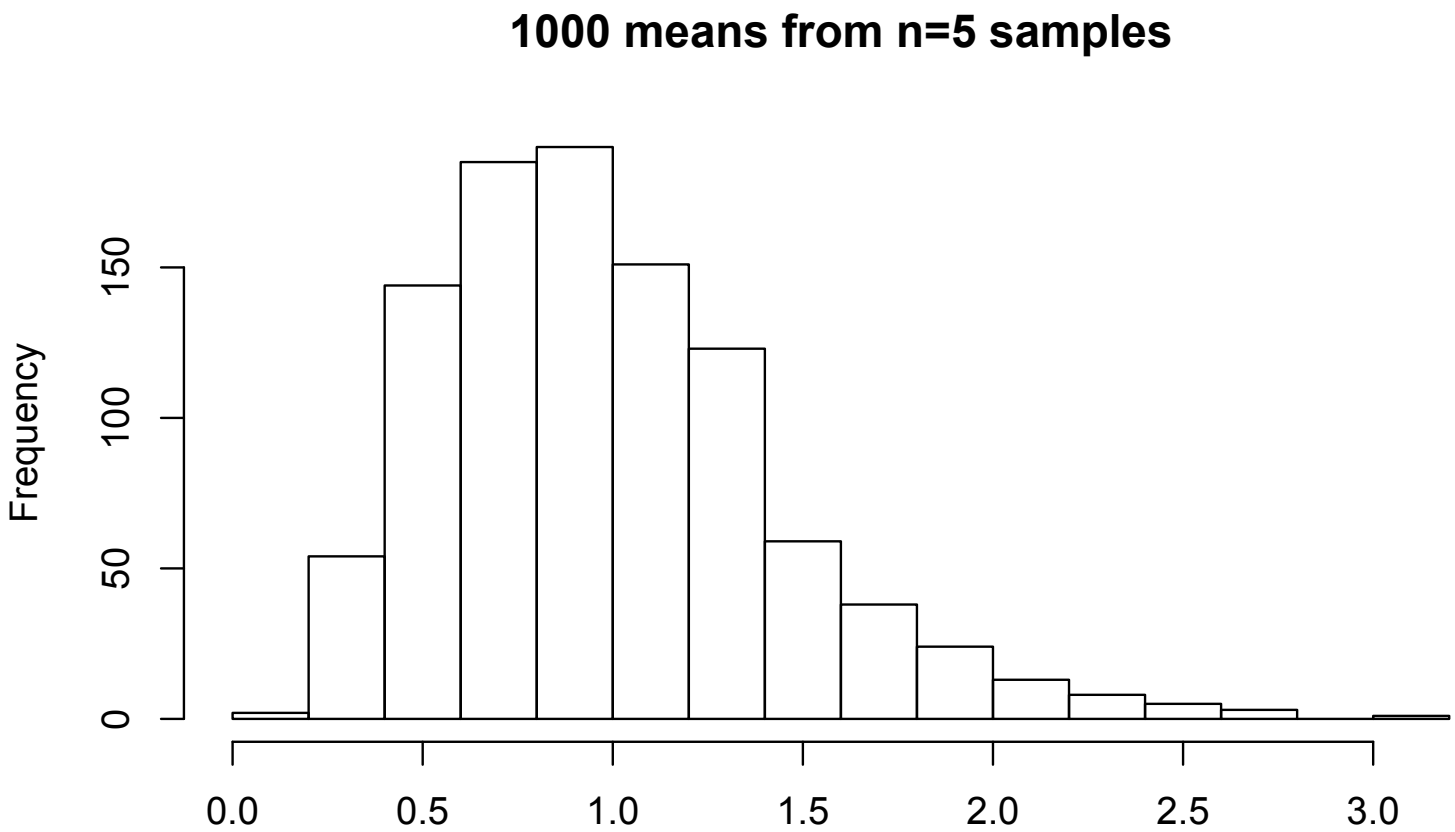


$E[X] = 1$ and $\mathrm{var}(X) = 1$.

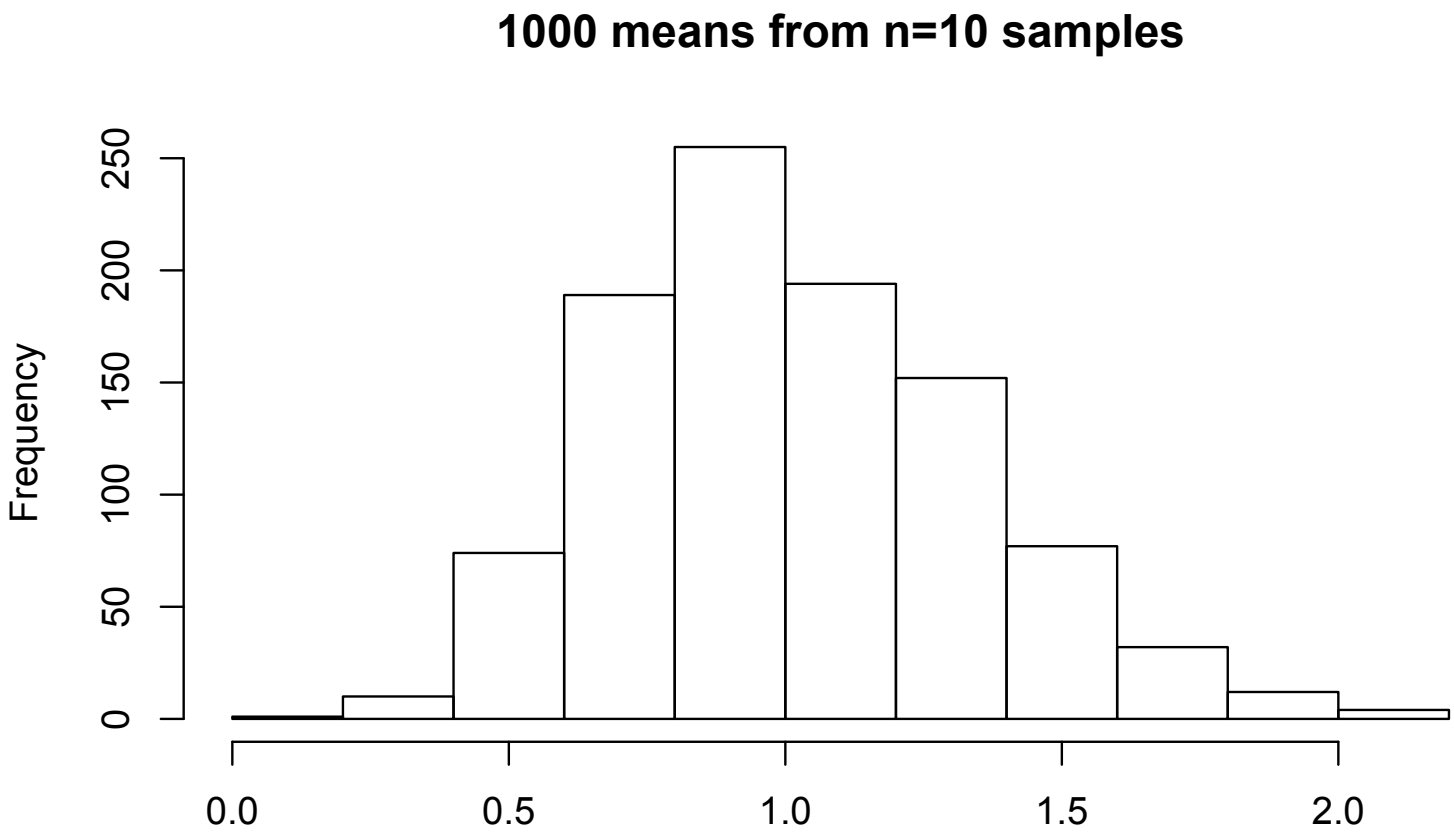
Central Limit Theorem



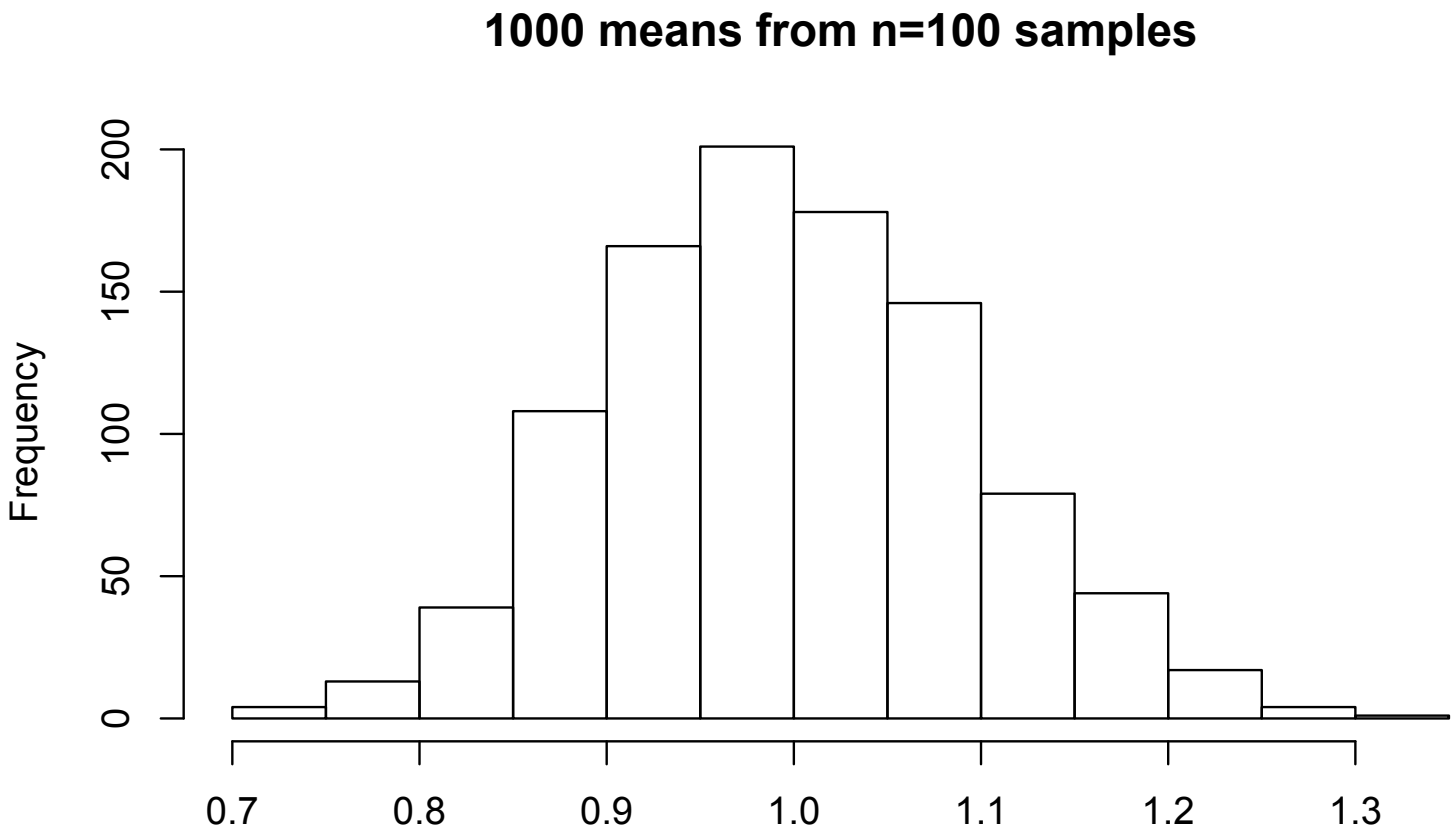
Central Limit Theorem



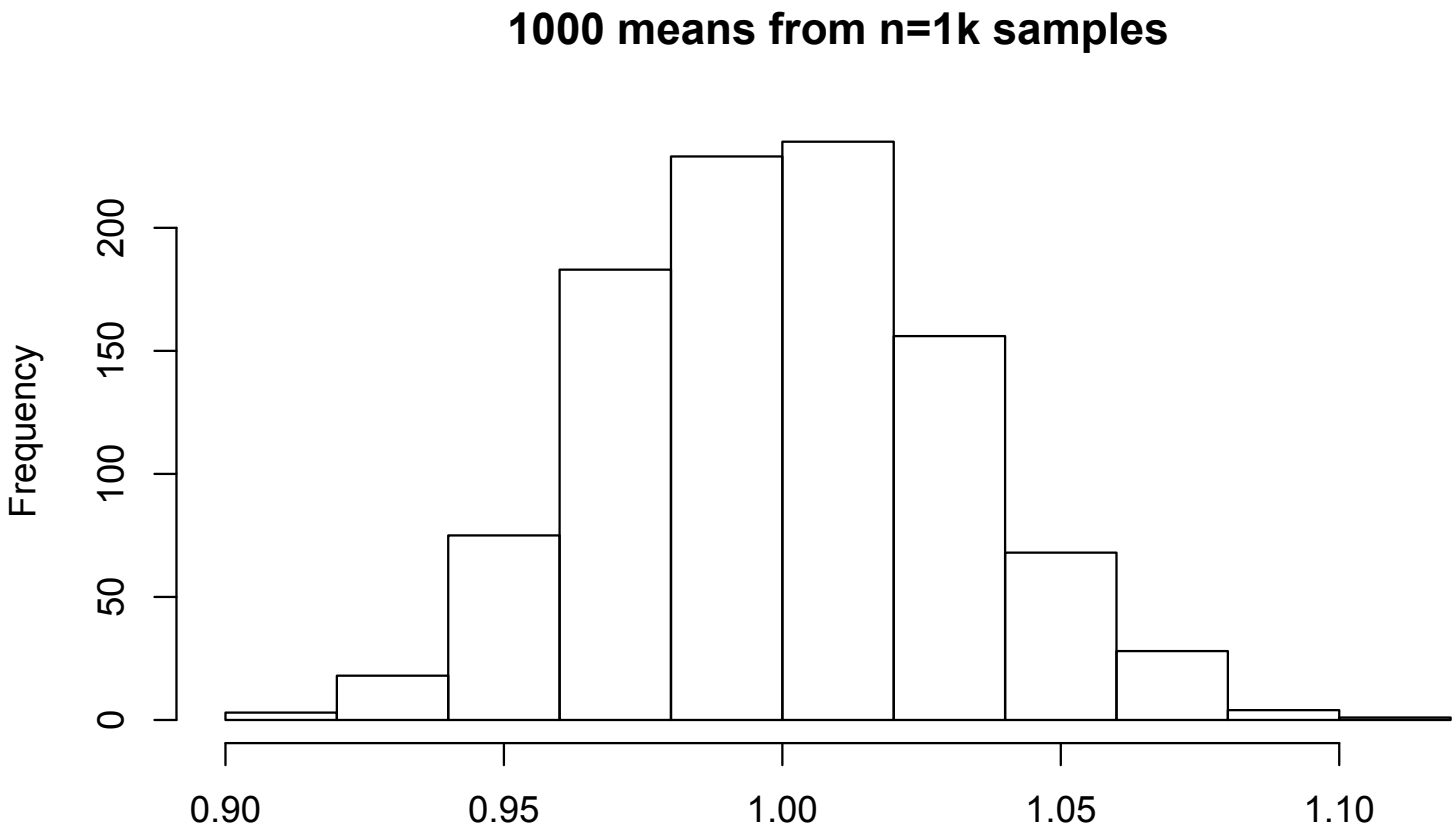
Central Limit Theorem



Central Limit Theorem



Central Limit Theorem



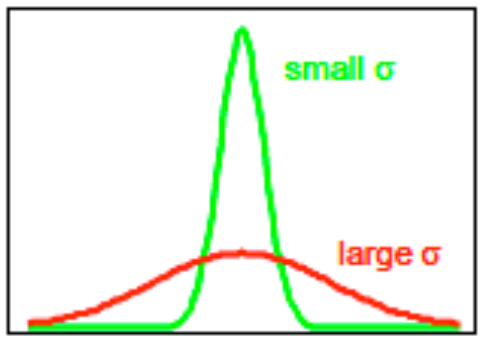
Sampling Distribution of b_1

The sampling distribution of b_1 describes how estimator $b_1 = \hat{\beta}_1$ varies over different samples with the X values fixed.

It turns out that b_1 is normally distributed: $b_1 \sim N(\beta_1, \sigma^2_{b_1})$.

- b_1 is unbiased: $E[b_1] = \beta_1$.
- Sampling sd σ_{b_1} determines precision of b_1.

The variance term determines how close the estimate will be to the true value.
Remember: large σ is bad!



Sampling Distribution of b_1

Can we intuit what should be in the formula for σ_{b_1}?

- How should σ figure in the formula?
- What about n?
- Anything else?

$$\text{var}(b_1) = \frac{\sigma^2}{\sum(X_i - \bar{X})^2} = \frac{\sigma^2}{(n-1)s_x^2}$$

Three Factors:

sample size (n), error variance ($\sigma^2 = \sigma_\varepsilon^2$), and X-spread (s_x).

Sampling Distribution of b_0

The intercept is also normal and unbiased: $b_0 \sim N(\beta_0, \sigma^2_{b_0})$.

$$\sigma^2_{b_0} = \mathrm{var}(b_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{(n-1)s_x^2} \right)$$

What is the intuition here?

The Importance of Understanding Variation

When estimating a quantity, it is vital to develop a notion of the precision of the estimation; for example:

- estimate the slope of the regression line
- estimate the value of a house given its size
- estimate the expected return on a portfolio
- estimate the value of a brand name
- estimate the damages from patent infringement

Why is this important?

We are making decisions based on estimates, and these may be very sensitive to the accuracy of the estimates!

The Importance of Understanding Variation

Example from "everyday" life:

- When building a house, we can estimate a required piece of wood to 1/4"?
- When building a fine cabinet, the estimates may have to be accurate to 1/16" or even 1/32".

The standard deviations of the least squares estimators of the slope and intercept give a precise measurement of the accuracy of the estimator.

However, these formulas aren't especially practical since they involve the unknown quantity: σ

Estimated Variance

We estimate variation with "sample standard deviations":

$$s_{b_1} = \sqrt{\frac{s^2}{(n-1)s_x^2}} \quad s_{b_0} = \sqrt{s^2\left(\frac{1}{n} + \frac{\bar{X}^2}{(n-1)s_x^2}\right)}$$

Recall that $s = \sqrt{\sum e_i^2/(n-2)}$ is the estimator for $\sigma = \sigma_\varepsilon$.
Hence, $s_{b_1} = \hat{\sigma}_{b_1}$ and $s_{b_0} = \hat{\sigma}_{b_0}$ are estimated coefficient sd's.

A high level of info/precision/accuracy means small s_b values.

Normal and Student's t

Recall what *Student* discovered:

If $\theta \sim N(\mu, \sigma^2)$, but you estimate $\sigma^2 \approx s^2$ based on $n - p$ degrees of freedom, then $\theta \sim t_{n-p}(\mu, s^2)$.

For example:

- $\bar{Y} \sim t_{n-1}(\mu, s_y^2/n)$.
- $b_0 \sim t_{n-2}\left(\beta_0, s_{b_0}^2\right)$ and $b_1 \sim t_{n-2}\left(\beta_1, s_{b_1}^2\right)$

The t distribution is just a fat-tailed version of the normal. As $n - p \longrightarrow \infty$, our tails get skinny and the t becomes normal.

Standardized Normal and Student's t

We'll also usually standardize things:

$$\frac{b_j - \beta_j}{\sigma_{b_j}} \sim N(0,1) \Longrightarrow \frac{b_j - \beta_j}{s_{b_j}} \sim t_{n-2}(0,1)$$

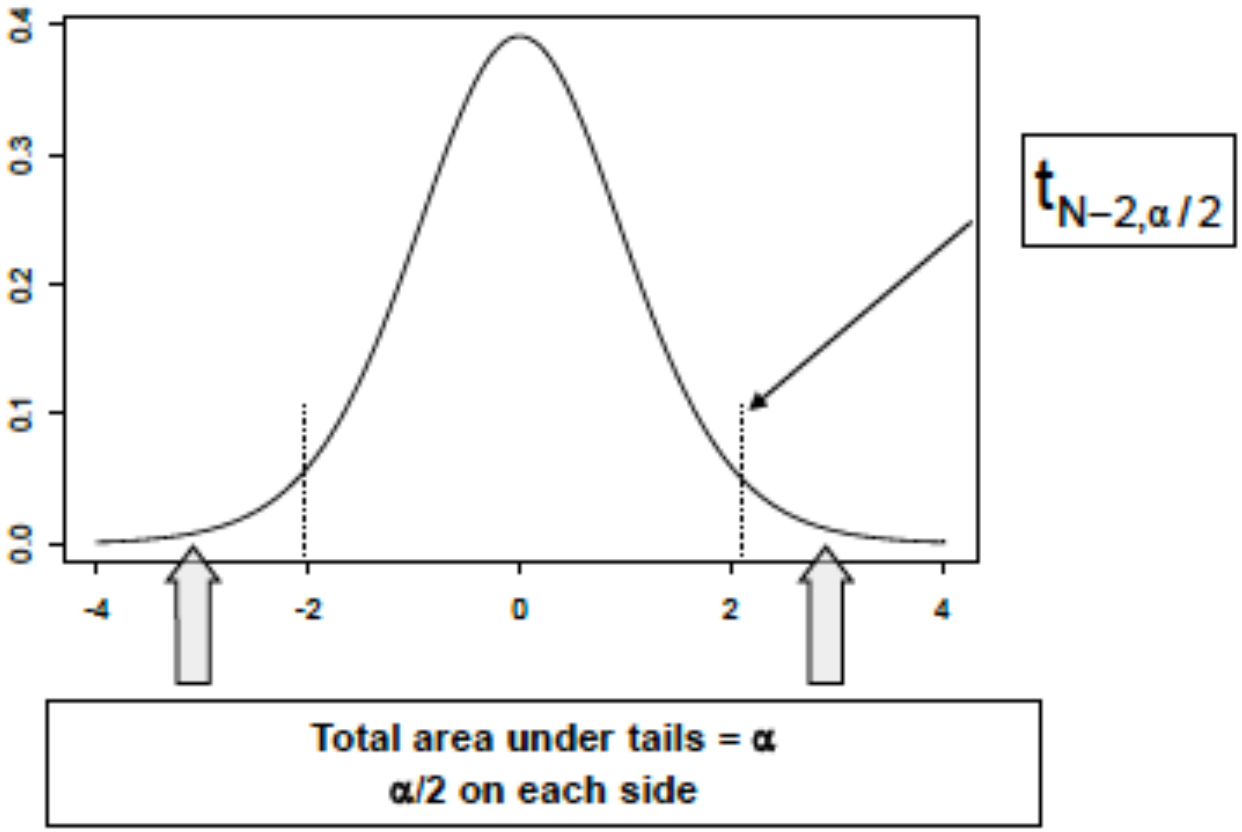
We use $Z \sim N(0,1)$ and $Z_{n-p} \sim t_{n-p}(0,1)$ to represent standard random variables.

Notice that the t and normal distributions depend upon assumed values for β_j: this forms the basis for confidence intervals, hypothesis testing, and p-values.

Testing and Confidence Intervals (in 3 slides)

Suppose Z_{n-p} is distributed $t_{n-p}(0,1)$. A centered interval is

$$\mathrm{P}(-t_{n-p,\alpha/2} < Z_{n-p} < t_{n-p,\alpha/2}) = 1-\alpha$$



Confidence Intervals

Since $b_j \sim t_{n-p}(\beta_j, s_{b_j})$,

$$\begin{aligned} 1-\alpha &= \mathrm{P}(-t_{n-p,\alpha/2} < \frac{b_j - \beta_j}{s_{b_j}} < t_{n-p,\alpha/2}) \\ &= \mathrm{P}(b_j - t_{n-p,\alpha/2} s_{b_j} < \beta_j < b_j + t_{n-p,\alpha/2} s_{b_j}) \end{aligned}$$

Thus $(1-\alpha)$*100% of the time, β_j is within the Confidence Interval: $b_j \pm t_{n-p,\alpha/2} s_{b_j}$

Testing

Similarly, suppose that assuming $b_j \sim t_{n-p}(\beta_j, s_{b_j})$ for our sample b_j leads to (recall $Z_{n-p} \sim t_{n-p}(0,1)$)

$$\mathrm{P}\left(Z_{n-p} < -\left|\frac{b_j - \beta_j}{s_{b_j}}\right|\right) + \mathrm{P}\left(Z_{n-p} > \left|\frac{b_j - \beta_j}{s_{b_j}}\right|\right) = \varphi.$$

Then the "p-value" is $\varphi = 2\mathrm{P}(Z_{n-p} > |b_j - \beta_j|/s_{b_j})$.

You do this calculation for $\beta_j = \beta_j^0$, an assumed null/safe value, and only reject β_j^0 if φ is too small (e.g., $\varphi < 1/20$).

In regression, $\beta_j^0 = 0$ almost always.

More Detail... Confidence Intervals

Why should we care about Confidence Intervals?

- The confidence interval captures the amount of information in the data about the parameter.
- The center of the interval tells you what your estimate is.
- The length of the interval tells you how sure you are about your estimate.

More Detail... Testing

Suppose that we are interested in the slope parameter, β_1.

For example, is there any evidence in the data to support the existence of a relationship between X and Y?

We can rephrase this in terms of competing hypotheses.

$H_0 : \beta_1 = 0$. Null/safe; implies "no effect" and we ignore X.

$H_1 : \beta_1 \neq 0$. Alternative; leads us to our best guess $\beta_1 = b_1$.

Hypothesis Testing

If we want statistical support for a certain claim about the data, we want that claim to be the alternative hypothesis.

Our hypothesis test will either reject or not reject the null hypothesis (the default if our claim is not true).

If the hypothesis test rejects the null hypothesis, we have statistical support for our claim!

Hypothesis Testing

We use b_j for our test about β_j.

- Reject H_0 when b_j is far from β_j^0 (ususally 0).
- Assume H_0 when b_j is close to β_j^0.

An obvious tactic is to look at the difference $b_j - \beta_j^0$.

But this measure doesn't take into account the uncertainty in estimating b_j: What we really care about is how many standard deviations b_j is away from β_j^0.

Hypothesis Testing

The *t*-statistic for this test is

$$z_{b_j} = \frac{b_j - \beta_j^0}{s_{b_j}} = \frac{b_j}{s_{b_j}} \text{ for } \beta_j^0 = 0.$$

If H_0 is true, this should be distributed $z_{b_j} \sim t_{n-p}(0, 1)$.

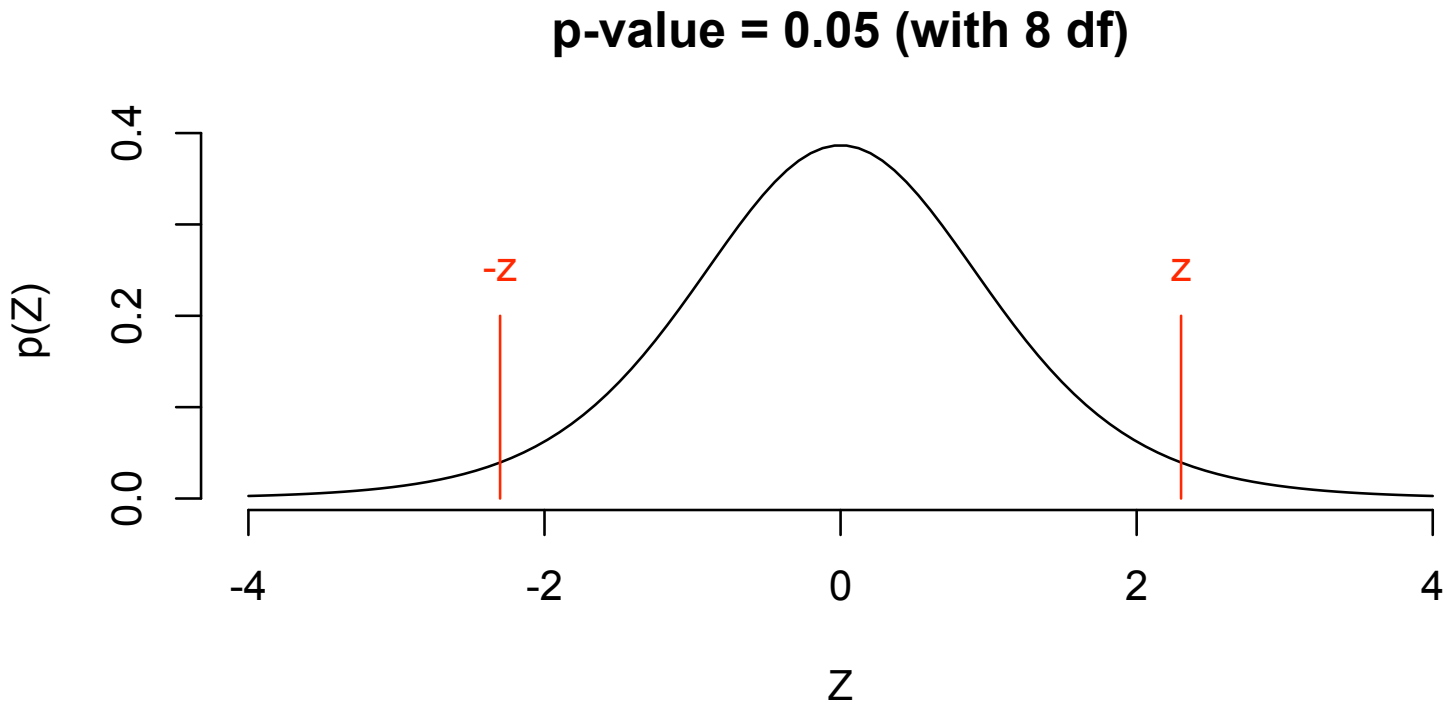
- Small $|z_{b_j}|$ leaves us happy with the null β_j^0.
- Large $|z_{b_j}|$ (i.e., $>$ about 2) should get us worried!

Hypothesis Testing

We assess the size of z_{b_j} with the p-value :

$$\varphi = \mathrm{P}(|Z_{n-p}| > |z_{b_j}|) = 2\mathrm{P}(Z_{n-p} > |z_{b_j}|)$$

(once again, $Z_{n-p} \sim t_{n-p}(0, 1)$).



Hypothesis Testing

The p-value is the probablity, assuming that the null hypothesis is true, of seeing something more extreme
(further from the null) than what we have observed.

You can think of $1 - \varphi$ (inverse p-value) as a measure of distance between the data and the null hypothesis. In other words, $1 - \varphi$ is the strength of evidence against the null.

Hypothesis Testing

The formal 2-step approach to hypothesis testing

- Pick the significance level α (often $1/20 = 0.05$),
  our acceptable risk (probability) of rejecting a true null hypothesis (we call this a type 1 error).
  This α plays the same role as α in CI's.
- Calculate the p-value, and reject H_0 if $\varphi < \alpha$
  (in favor of our best alternative guess; e.g. $\beta_j = b_j$).
  If $\varphi > \alpha$, continue working under null assumptions.

This is equivilent to having the rejection region $|z_{b_j}| > t_{n-p,\alpha/2}$.

Example: Hypothesis Testing

Consider again a CAPM regression for the Windsor fund.

Does Windsor have a non-zero intercept?
(i.e., does it make/lose money independent of the market?).

H_0 : $\beta_0 = 0$ and there is no-free money.

H_1 : $\beta_0 \neq 0$ and Windsor is cashing regardless of market.

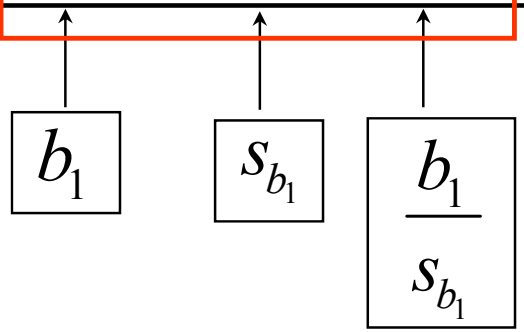
Example: Hypothesis Testing

Regression Statistics	
Multiple R	0.923417768
R Square	0.852700374
Adjusted R Square	0.851872848
Standard Error	0.018720015
Observations	180

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	0.3611	0.361099761	1030.421266	6.0291E-76
Residual	178	0.062378	0.000350439		
Total	179	0.423478			

	Coefficients	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*Upper 95.0%*
Intercept	0.003646881	0.001409	2.587596412	0.010462425	0.000865657	0.006428	0.000866	0.006428
X Variable 1	0.935717012	0.02915	32.10017549	6.0291E-76	0.878193151	0.993241	0.878193	0.993241



It turns out that we reject the null at $\alpha = .05$ ($\varphi = .0105$). Thus Windsor does have an "alpha" over the market.

Example: Hypothesis Testing

Looking at the slope, this is a very rare case where the null hypothesis is not zero:

H_0 : $\beta_1 = 1$ Windsor is just the market (+ alpha).

H_1 : $\beta_1 \neq 1$ and Windsor softens or exaggerates market moves.

We are asking whether or not Windsor moves in a different way than the market (e.g., is it more conservative?).

Now,

$$t = \frac{b_1 - 1}{s_{b_1}} = \frac{-0.0643}{0.0291} = -2.205$$

$$t_{n-2,\alpha/2} = t_{178,0.025} = 1.96$$

Reject H_0 at the 5% level

Forecasting

The conditional forecasting problem: Given covariate X_f and sample data $\{X_i, Y_i\}_{i=1}^n$, predict the "future" observation y_f.

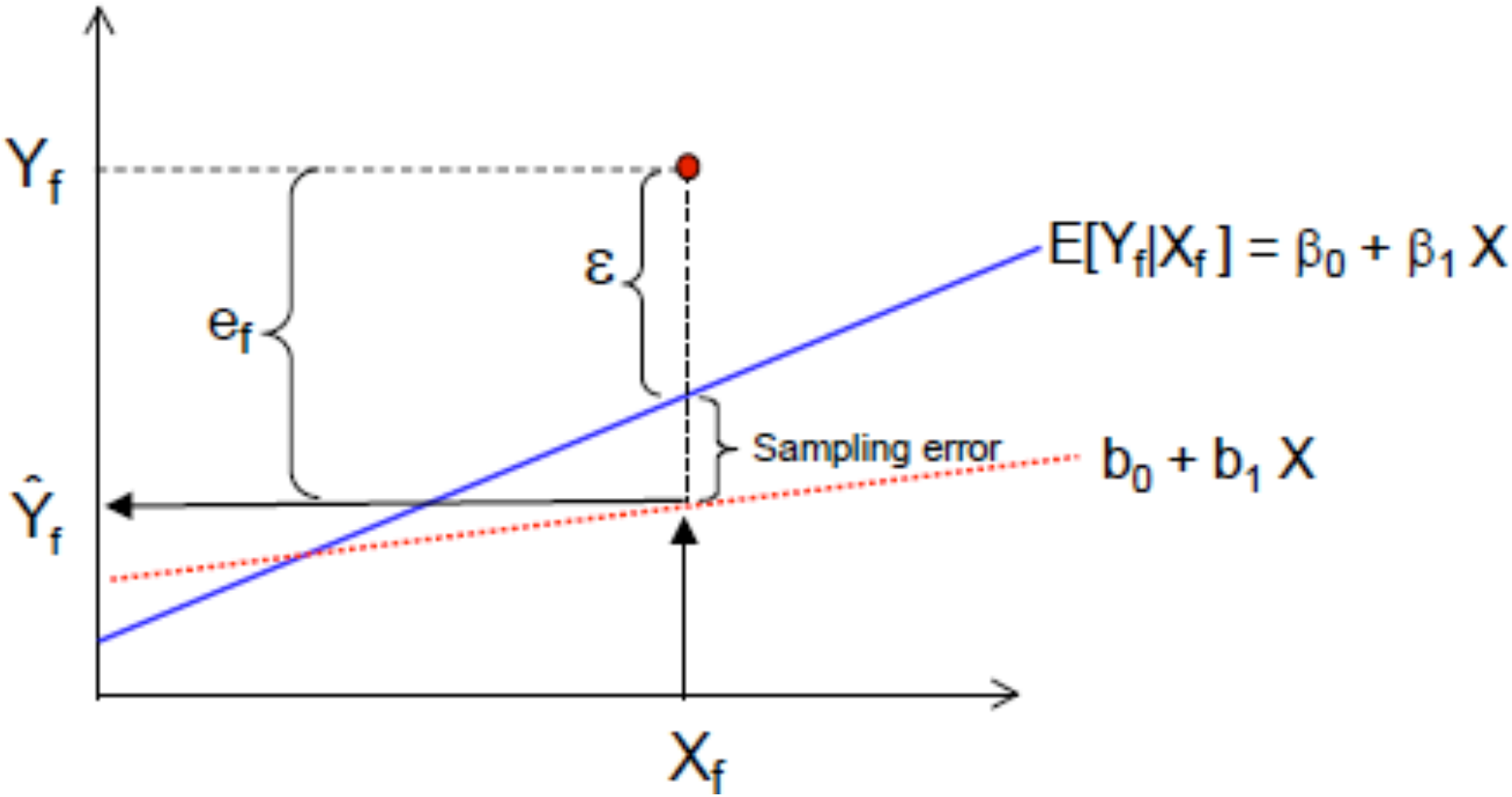
The solution is to use our LS fitted value: $\hat{Y}_f = b_0 + b_1 X_f$.

This is the easy bit. The hard (and very important!) part of forecasting is assessing uncertainty about our predictions.

Forecasting

If we use $\hat{Y}_f$, our prediction error is

$$e_f = Y_f - \hat{Y}_f = Y_f - b_0 - b_1 X_f$$



Forecasting

This can get quite complicated! A simple strategy is to build the following $(1-\alpha)100\%$ prediction interval:

$$b_0 + b_1 X_f \pm t_{n-2,\alpha/2} s$$

A large predictive error variance (high uncertainty) comes from

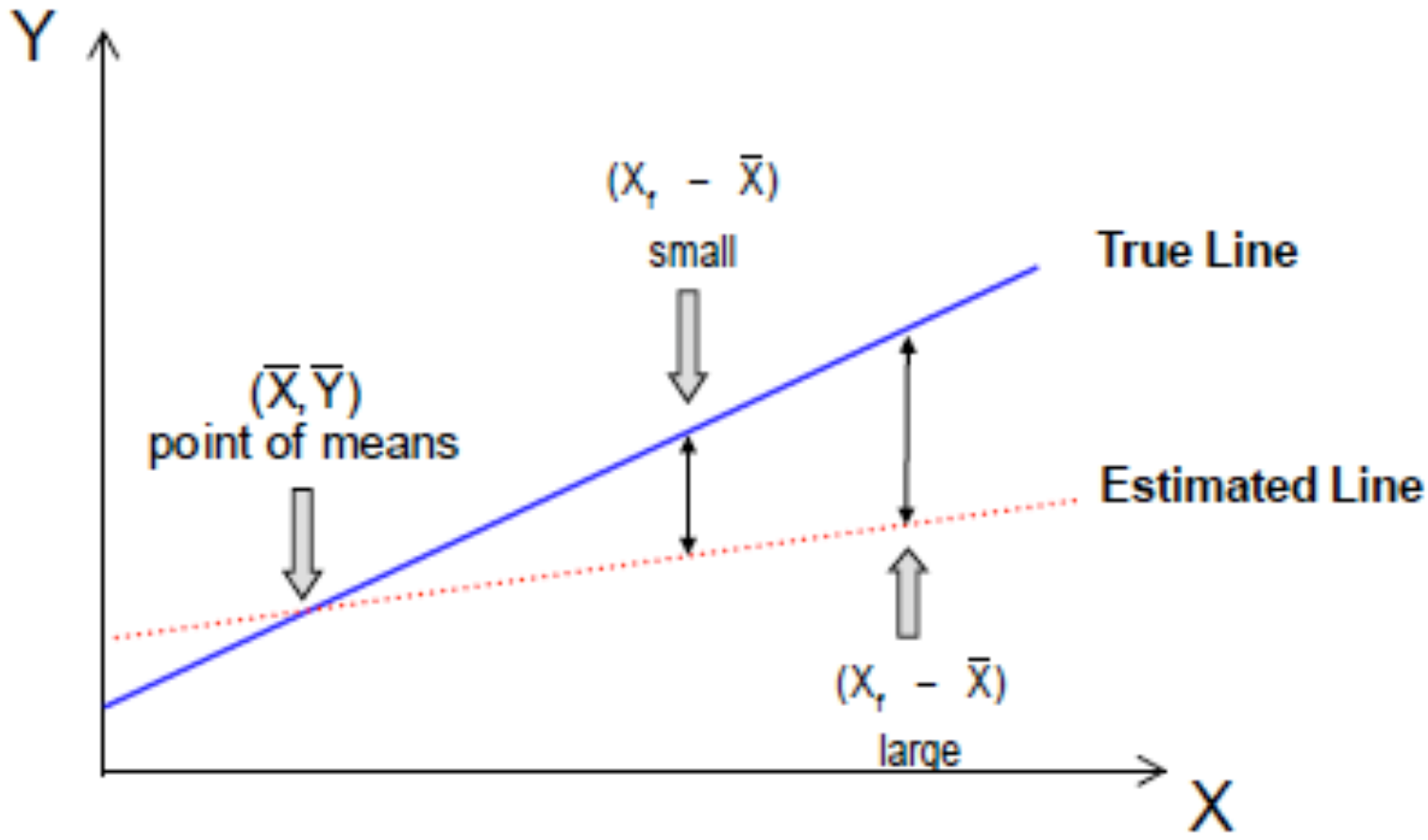
- Large s (i.e., large ε's).
- Small n (not enough data).
- Small s_x (not enough observed spread in covariates).
- Large difference between X_f and $\bar{X}$.

Just remember that you are uncertain about b_0 and b_1!

Reasonably inflating the uncertainty in the interval above is always a good idea... as always, this is problem dependent.

Forecasting

For X_f far from our $\bar{X}$, the space between lines is magnified...



Glossary and Equations

- $\hat{Y}_i = b_0 + b_1 X_i$ is the ith fitted value.
- $e_i = Y_i - \hat{Y}_i$ is the ith residual.
- s: standard error of regression residuals ($\approx \sigma = \sigma_\varepsilon$).

$$s^2 = \frac{1}{n-2} \sum e_i^2$$

- s_{b_j}: standard error of regression coefficients.

$$s_{b_1} = \sqrt{\frac{s^2}{(n-1)s_x^2}} \qquad s_{b_0} = s\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{(n-1)s_x^2}}$$

Glossary and Equations

- α is the significance level (prob of type 1 error).
- $t_{n-p,\alpha/2}$ is the value such that for $Z_{n-p} \sim t_{n-p}(0,1)$,

$$\mathrm{P}(Z_{n-p} > t_{n-p,\alpha/2}) = \mathrm{P}(Z_{n-p} < -t_{n-p,\alpha/2}) = \alpha/2.$$

- $z_{b_j} \sim t_{n-p}(0,1)$ is the standardized coefficient t-value:

$$z_{b_j} = \frac{b_j - \beta_j^0}{s_{b_j}} \quad (= b_j/s_{b_j} \textit{ most often})$$

- The $(1-\alpha) * 100\%$ for β_j is $b_j \pm t_{n-p,\alpha/2} s_{b_j}$.