STP 530
Lecture 2: Inferences

Yi Zheng, Ph.D.
yi.isabel.zheng@asu.edu

ARIZONA STATE UNIVERSITY
SCHOOL OF MATHEMATICAL AND STATISTICAL SCIENCES

August 29, 2023

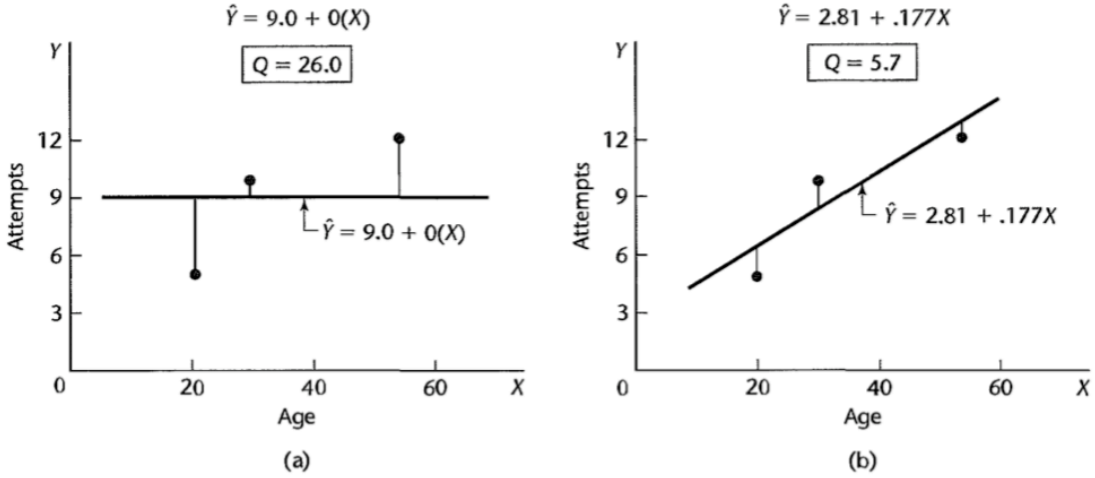
Outline

1. Normal error regression model
2. Sampling distribution
3. Confidence interval for regression coefficients
4. Confidence interval of $E\{Y_h\}$
5. Prediction interval of Y_h
6. Hypothesis testing of regression coefficients

Normal error regression (NER) model

The least square estimation of regression coefficients is NOT based upon any assumption on the **specific distribution of** ε.

FIGURE 1.9 Illustration of Least Squares Criterion Q for Fit of a Regression Line—Persistence Study Example.

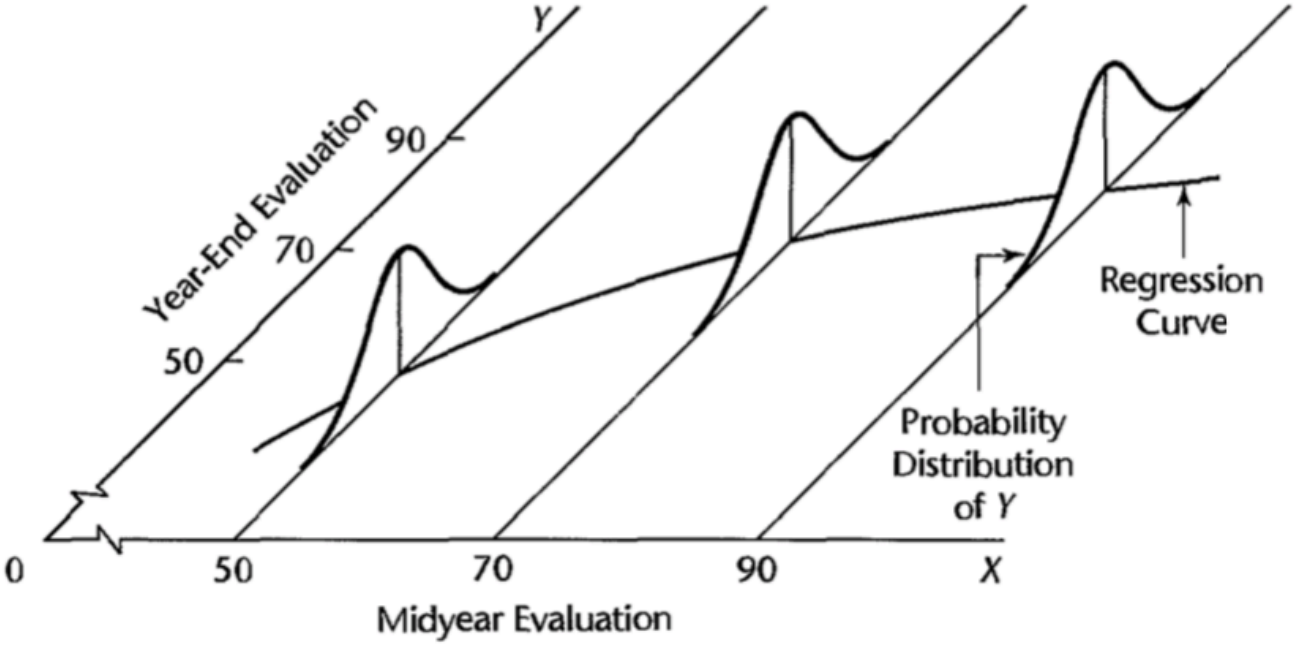


Objective: Minimize Q, the **sum of squared errors**:

$$Q = \text{SSE} = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

Normal error regression (NER) model

- We impose assumptions on the probability distribution of ε so that inferential statistics can be carried out (confidence intervals, hypothesis tests).
- The most common linear regression model assumes ε i.i.d $\sim N(0, \sigma^2)$, a.k.a., the NER model.



Normal error regression model

The **normal error regression model** is defined as follows:

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

where for $i = 1, \cdots, n$,:

- Y_i is the observed response
- X_i is the predictor variable, a known constant
- β_0 and β_1 are parameters to be estimated
- ε_i are independent normal variables with mean 0 and variance σ^2, namely ε_i i.i.d $\sim N(0, \sigma^2)$.

Normal error regression model

The normality assumption for the error terms is justifiable in many situations because the error terms frequently represent the effects of factors omitted from the model that vary at random.

Insofar as these random effects have a degree of mutual independence, the composite error term ε_i representing all these factors would tend to comply with the *central limit theorem* and the error term distribution would approach normality as the number of factor effects becomes large.

Maximum likelihood estimation

Maximum Likelihood Estimation

Refer to textbook pp.27-32.

Understanding probability distribution is essential to understanding MLE.

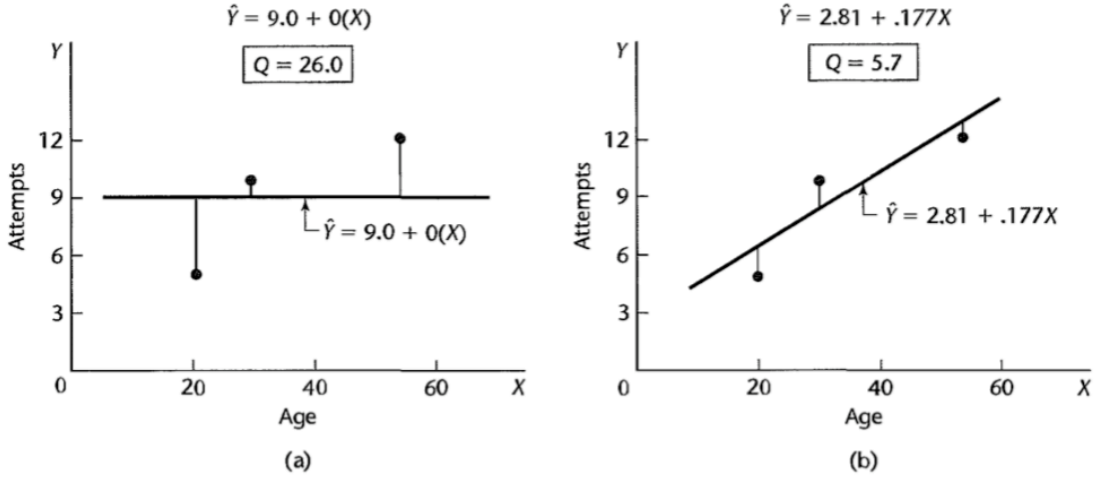
Probability vs. Likelihood

1. Assume a coin is balanced. Toss it 100 times, what is the probability you get 50 heads?
2. You toss a coin a 100 times and get 50 heads, what is the likelihood the coin is balanced?

Maximum likelihood estimation vs. least square estimation

Both MLE and LS seek a combination of b_0 and b_1 values that define the optimal straight line (i.e., the fitted regression model) that best fits the sample data.

FIGURE 1.9 Illustration of Least Squares Criterion Q for Fit of a Regression Line—Persistence Study Example.



LS: Minimize SSE, the sum of squared errors:

$$SSE = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

MLE: Maximize L, the likelihood function of the regression model given the sample data.

For NER models, the formulas for b_0 and b_1 derived from MLE are identical to those derived from the LS method.

Outline

1. Normal error regression model
2. Sampling distribution
3. Confidence interval for regression coefficients
4. Confidence interval of $E\{Y_h\}$
5. Prediction interval of Y_h
6. Hypothesis testing of regression coefficients

Sampling distribution

- A **sample statistic** is a numerical summary of sample data.
- A **population parameter** is a numerical summary of the population.

In practice, we rarely know the values of the parameters. We use sample statistics to estimate population parameters.
e.g., $\bar{x} \to \mu \quad s \to \sigma \quad p \to \pi \quad b \to \beta$

Because the sample is only a subset of the population, our sample statistic is expected to include a random **sampling error**, and the sample statistic value varies by different samples. These varied values form a sampling distribution:

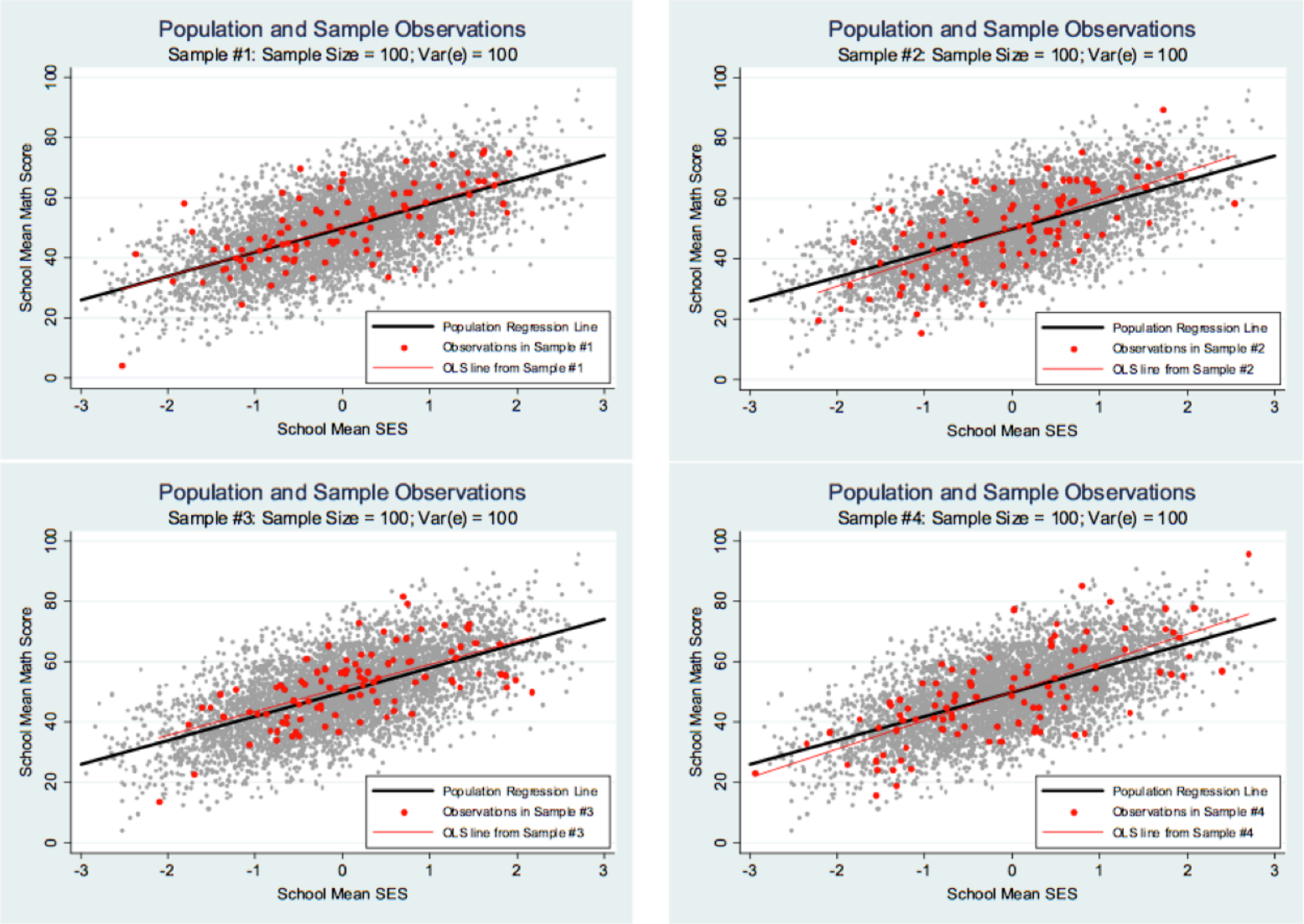
The **sampling distribution** is the probability distribution of a sample statistic.

The Central Limit Theorem

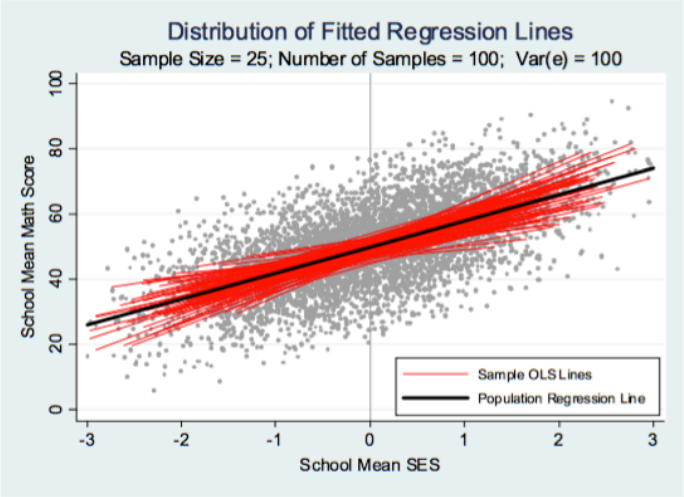
The **Central Limit Theorem** is a theorem about the sampling distribution of the sample mean:

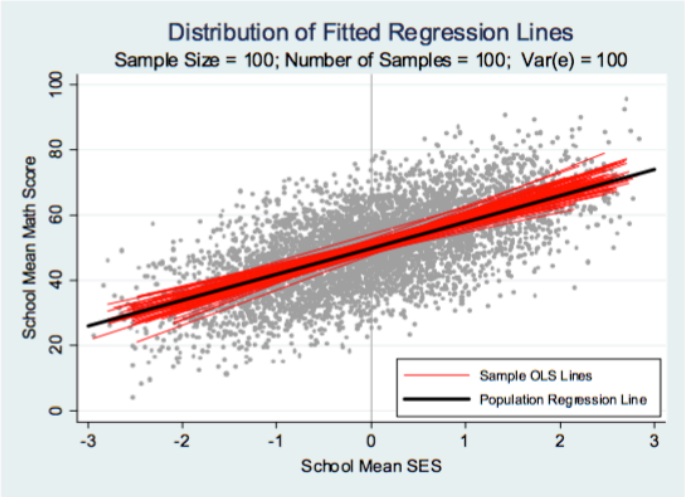
Given a random sample of size n from a population having mean μ and standard deviation σ, as the sample size increases, **the sampling distribution of the sample mean** $\bar{x}$ approaches an approximately normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$, regardless of the probability distribution of the population data.

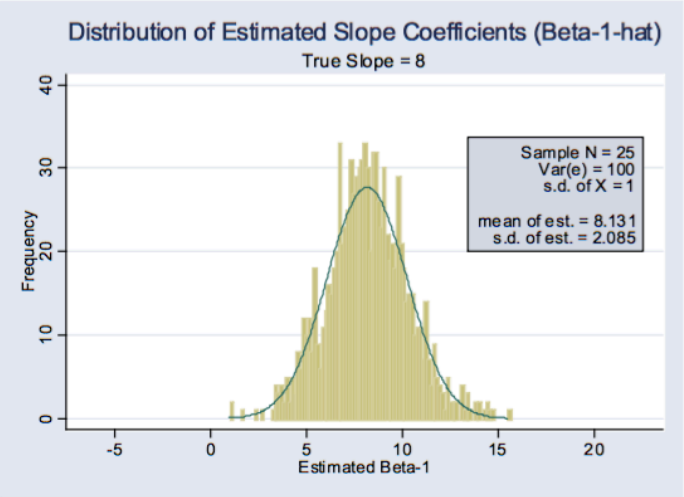
Sampling distribution of regression coefficients

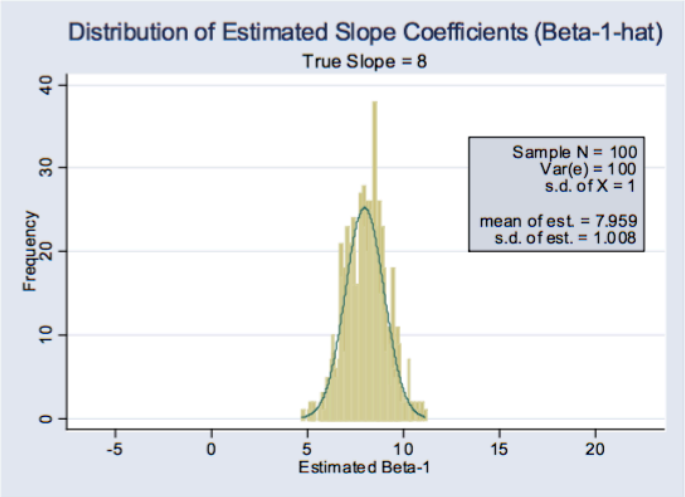


Sampling distribution of regression coefficients









Sampling distribution of b_1

For the normal error regression (NER) model, the sampling distribution of b_1 is **normal**, with mean and variance

$$E\{b_1\} = \beta_1$$

$$\sigma^2\{b_1\} = \frac{\sigma^2}{\sum(X_i - \bar{X})^2}$$

See proof on textbook pp.41-43. Key points to remember:

- b_1 is a linear combination of the Y_i,

$$b_1 = \frac{\sum(X_i - \bar{X})Y_i}{\sum(X_i - \bar{X})^2} = \sum k_i Y_i$$

- X_i are considered known constants in regression models.
- Y_i has constant variance σ^2.

Standard error of b_1

The standard deviation of a sampling distribution is called the **standard error**.

The standard error of the slope estimator b_1 is:

$$\sigma^2\{b_1\} = \frac{\sigma^2}{\sum(X_i - \bar{X})^2} = \frac{\sigma^2}{(n-1) \cdot s^2\{X\}}$$

For model fitting purposes, the **smaller** the standard error, the better the model fits the data.

Factors influencing the standard error of the estimated slope b_1:

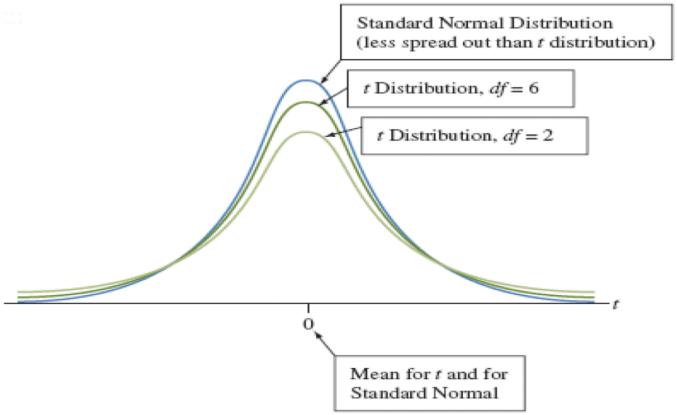
- Error variance σ^2
- Variance (spacing) of the predictor variable $s^2\{X\}$
- Sample size n

Studentized statistic

If X is normally distributed,

Standardized statistic: $(X - \mu_x)/\sigma_x \sim N(0, 1)$

Studentized statistic: $(X - \mu_x)/s_x \sim t(df)$



The **Student's t distribution**:
The distribution takes its name from William Sealy Gosset's 1908 paper in Biometrika under the pseudonym "Student". Gosset worked at the Guinness Brewery in Dublin, Ireland, and was interested in the problems of small samples - for example, the chemical properties of barley where sample sizes might be as few as 3. One version of the origin of the pseudonym is that Gosset's employer preferred staff to use pen names when publishing scientific papers instead of their real name, so he used the name "Student" to hide his identity. Another version is that Guinness did not want their competitors to know that they were using the t-test to determine the quality of raw material. The distribution became well-known through the work of Ronald Fisher, who called the distribution "Student's distribution" and represented the test value with the letter t.

https://en.wikipedia.org/wiki/Student\%27s_t-distribution

Studentized statistic of b_1

Recall the sampling distribution of b_1 is **normal**, with mean and variance

$$E\{b_1\} = \beta_1 \qquad \sigma^2\{b_1\} = \frac{\sigma^2}{\sum(X_i - \bar{X})^2}$$

$\sigma^2\{b_1\}$ is estimated by

$$s^2\{b_1\} = \frac{s^2}{\sum(X_i - \bar{X})^2}$$

Note that $s^2\{b_1\}$ is an unbiased estimator of $\sigma^2\{b_1\}$.

The studentized statistic of b_1:

$$\frac{b_1 - \beta_1}{s\{b_1\}} \sim t(n-2)$$

Outline

1. Normal error regression model
2. Sampling distribution
3. Confidence interval for regression coefficients
4. Confidence interval of $E\{Y_h\}$
5. Prediction interval of Y_h
6. Hypothesis testing of regression coefficients

Confidence interval

- **Estimation**: We use sample statistics to estimate population parameters. e.g., $\bar{x} \to \mu \quad s \to \sigma \quad p \to \pi \quad b \to \beta$
- A **point estimate** is a single number that is the "best guess" for the parameter.
  - e.g., The best point estimate of μ is $\bar{x}$.
  - A point estimate doesn't tell us how close the estimate is likely to be to the parameter. Any one point estimate may or may not be close to the parameter it estimates.
- A **confidence interval** (interval estimate) is an interval containing the most plausible values for a parameter.

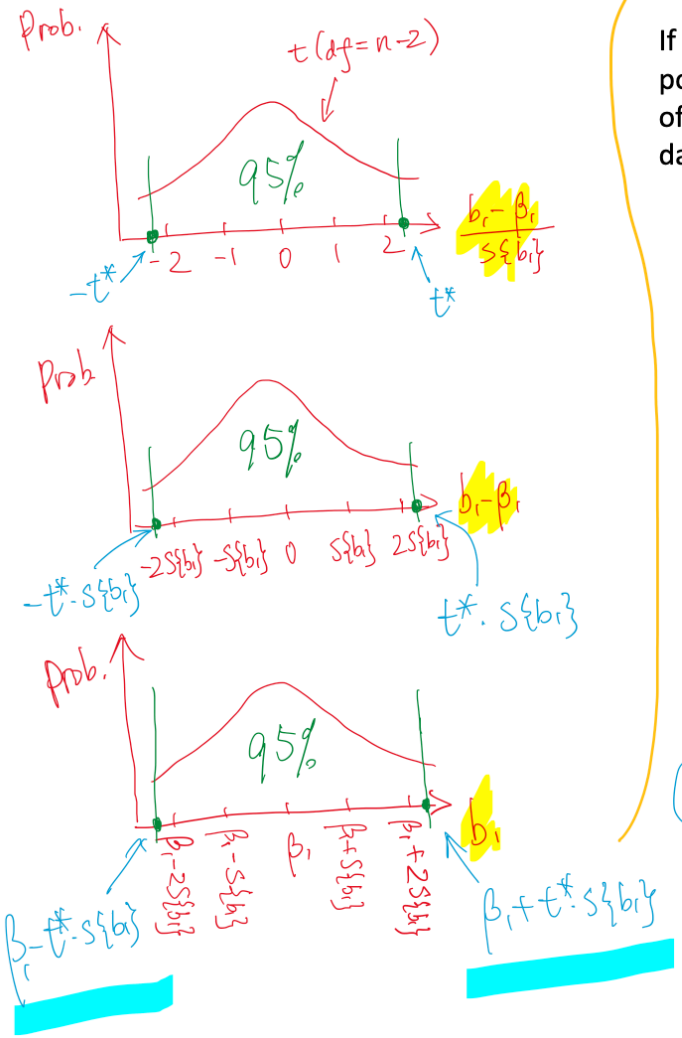
Confidence Interval = Point Estimate $\pm$ Margin of Error

= Point Estimate $\pm$ Distribution Multiplier * Standard error of estimate

An interval estimate is more useful than a point estimate. The margin of error helps us gauge the accuracy of the point estimate.

Confidence interval of β_1

Because $(b_1 - \beta_1)/s\{b_1\}$ follows a t distribution...



If we draw a random sample repeatedly from the same population and estimate the regression coefficients, 95% of the times the estimated slope obtained from the sample data falls between the two limits.

$$P(\beta_1 - t^*\cdot s\{b_1\} < b_1 < \beta_1 + t^*\cdot s\{b_1\}) = .95$$

$$\beta_1 < b_1 + t^*\cdot s\{b_1\} \qquad \beta_1 > b_1 - t^*\cdot s\{b_1\}$$

$$P(b_1 - t^*\cdot s\{b_1\} < \beta_1 < b_1 + t^*\cdot s\{b_1\}) = .95$$

The 95% confidence interval for β_1 is:

$$(b_1 - t^*\cdot s\{b_1\},\ b_1 + t^*\cdot s\{b_1\})$$

Confidence interval of β_1

The $1-\alpha$ **confidence interval for** β_1 is given by the following confidence limits:

$$b_1 \pm t(1-\alpha/2, df = n-2) \cdot s\{b_1\}$$

where $s\{b_1\} = s/\sqrt{\sum(X_i - \bar{X})^2}$ and $t(1-\alpha/2, df = n-2)$ can be obtained by R code: `qt(p=(1-alpha/2),df=(n-2))`.

Notes:

- For a 95% confidence interval, $1-\alpha = .95$, where $\alpha = .05$ corresponds to the "significance level" or Type I error rate in hypothesis testing.
- This is the formula for the simple linear regression model (1 predictor). For multiple regression with more than one predictors, the df of the t-distribution is $n-p$, where p is the number of parameters being estimated, i.e., the number predictors in the model plus 1.

Confidence interval of β_1

The $1-\alpha$ confidence interval for β_1 is given by the following confidence limits (i.e., the lower limit or LL and the upper limit or UL):

$$b_1 \pm t(1-\alpha/2, df = n-2) \cdot s\{b_1\}$$

Interpretation:

- We are $(1-\alpha)100\%$ sure that the true population slope falls within this interval.
- With a $(1-\alpha)100\%$ confidence, we estimate that expected Y value changes by somewhere between LL and UL for each 1-unit increase in X.

Confidence interval of β_1

Example: What's the 95% CI for the slope?

```
Call:
lm(formula = sales_y ~ adv_x)

Residuals:
         1          2          3          4          5
 4.000e-01 -3.000e-01 -3.886e-16 -7.000e-01  6.000e-01

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.1000     0.6351  -0.157   0.8849
adv_x         0.7000     0.1915   3.656   0.0354 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6055 on 3 degrees of freedom
Multiple R-squared:  0.8167, Adjusted R-squared:  0.7556
F-statistic: 13.36 on 1 and 3 DF,  p-value: 0.03535
```

Get $t(1-\alpha/2, df = n-2)$ in **R** by `qt(p=.975, df=3)`

Confidence interval of β_0

The $1-\alpha$ confidence interval for β_0 is given by the following confidence limits:

$$b_0 \pm t(1-\alpha/2, df = n-2) \cdot s\{b_0\}$$

where

$$s\{b_0\} = s\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}}$$

Interpretation: We are $(1-\alpha)100\%$ sure that the expected Y value when $X = 0$ falls within this interval.

Notes:

- See derivation of $\sigma^2\{b_0\}$ on p.53 of textbook (Comment #1).
- This is the formula for the simple linear regression model (1 predictor). For multiple regression with more than one predictors, the df of the t-distribution is $n-p$, where p is the number of parameters being estimated, i.e., the number predictors in the model plus 1.

Confidence interval of β_0

Example: What's the 95% CI for the intercept?

```
Call:
lm(formula = sales_y ~ adv_x)

Residuals:
         1          2          3          4          5
 4.000e-01 -3.000e-01 -3.886e-16 -7.000e-01  6.000e-01

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.1000     0.6351  -0.157   0.8849
adv_x         0.7000     0.1915   3.656   0.0354 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6055 on 3 degrees of freedom
Multiple R-squared:  0.8167, Adjusted R-squared:  0.7556
F-statistic: 13.36 on 1 and 3 DF,  p-value: 0.03535
```

Get $t(1-\alpha/2, df = n-2)$ in **R** by `qt(p=.975, df=3)`

Outline

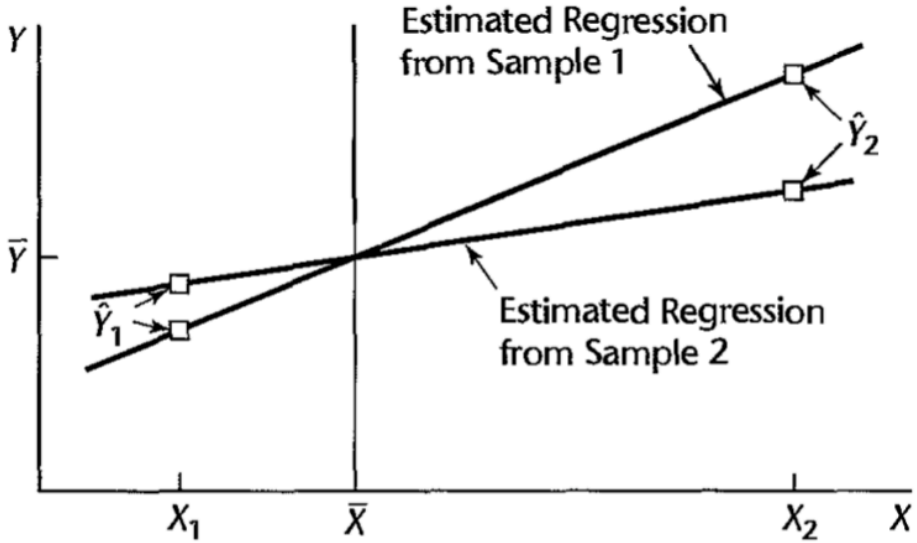
1. Normal error regression model
2. Sampling distribution
3. Confidence interval for regression coefficients
4. Confidence interval of $E\{Y_h\}$
5. Prediction interval of Y_h
6. Hypothesis testing of regression coefficients

$E\{Y_h\}$ vs. $\hat{Y}_h$

$E\{Y_h\}$ is the mean of the probability distribution of Y when $X = X_h$.

$\hat{Y}_h = b_0 + b_1 X_h$ is the predicted outcome when $X = X_h$, also a point estimate of $E\{Y_h\}$.

The confidence interval of $E\{Y_h\}$ reflects the variability of the **regression line** caused by sampling variability.



Confidence interval of $E\{Y_h\}$

The $1-\alpha$ **confidence interval for** $E\{Y_h\}$ is given by the following confidence limits:

$$\hat{Y}_h \pm t(1-\alpha/2, df = n-2) \cdot s\{\hat{Y}_h\}$$

where

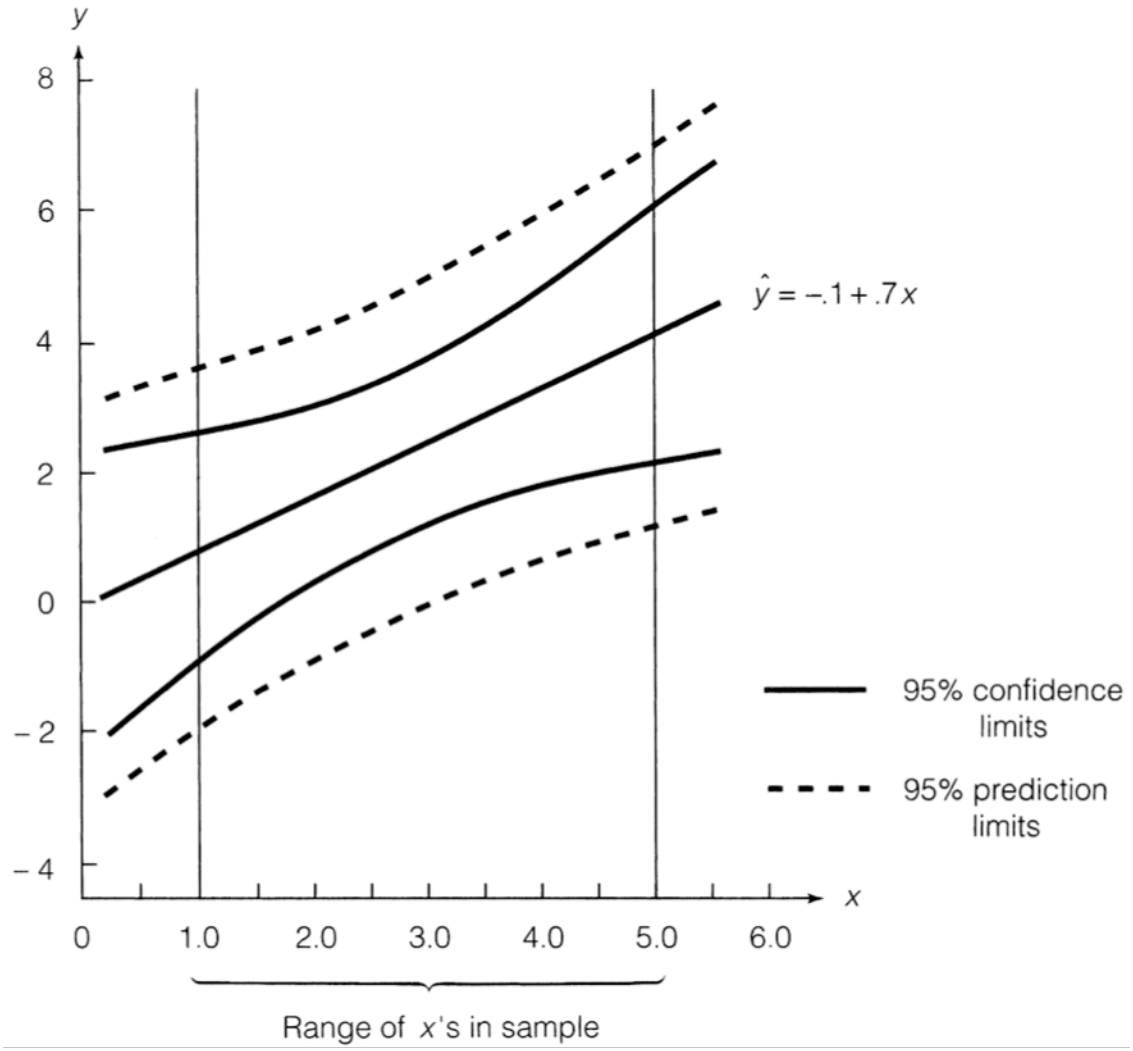
$$s\{\hat{Y}_h\} = s\sqrt{\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}}$$

Interpretation: We are $(1-\alpha)100\%$ sure that the mean value of Y for all observations with $X = X_h$ falls within this interval.

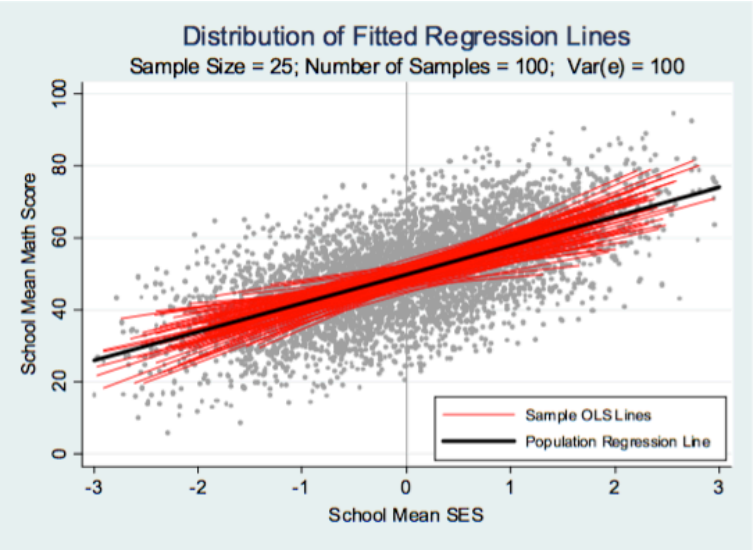
Notes:

- See derivation of $\sigma^2\{\hat{Y}_h\} = \sigma^2[\frac{1}{n} + \frac{(X_h-\bar{X})^2}{\sum(X_i-\bar{X})^2}]$ in textbook (p.53, Comment #2).
- This is the formula for the simple linear regression model (1 predictor). For multiple regression with more than one predictors, the df of the t-distribution is $n-p$, where p is the number of parameters being estimated, i.e., the number predictors in the model plus 1.

Confidence interval of $E\{Y_h\}$



← The confidence interval for $E\{Y_h\}$ is the narrowest when $X_h = \bar{X}$.



Outline

1. Normal error regression model
2. Sampling distribution
3. Confidence interval for regression coefficients
4. Confidence interval of $E\{Y_h\}$
5. Prediction interval of Y_h
6. Hypothesis testing of regression coefficients

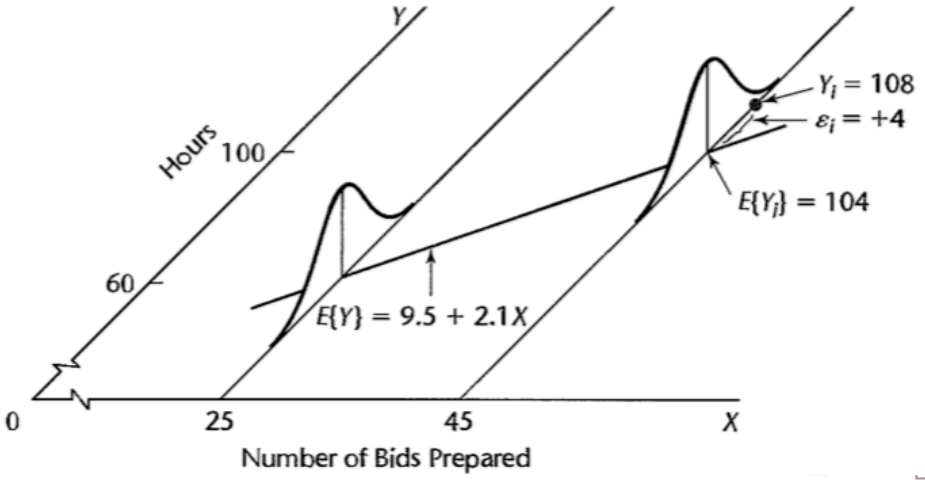
Estimating $E\{Y_h\}$ vs. predicting Y_h

Estimating $E\{Y_h\}$ is estimating the **overall trend** of Y given X_h. It's the regression line itself.

Predicting Y_h is predicting **an individual outcome** drawn from the probability distribution of Y_h given X_h. It's a point near the regression line. Using $\hat{Y}_h$ to predict a single case involves two steps:

$$\hat{Y}_h \longrightarrow E\{Y_h\} \longrightarrow Y_h$$

Both use $\hat{Y}_h$ as the point estimator, but the purposes are different.



Prediction interval of Y_h (when regression parameters are estimated)

The $1-\alpha$ **prediction interval** for Y_h is given by the following prediction limits:

$$\hat{Y}_h \pm t(1-\alpha/2, df = n-2) \cdot s_{\text{pred}}\{\hat{Y}_h\}$$

where

$$s_{\text{pred}}\{\hat{Y}_h\} = s\sqrt{1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}}$$

Interpretation: With a $(1-\alpha)100\%$ confidence, we predict that the value of Y for the next observation with a $X = X_h$ falls within this interval.

Note: This is the formula for the simple linear regression model (1 predictor). For multiple regression with more than one predictors, the df of the t-distribution is $n-p$, where p is the number of parameters being estimated, i.e., the number predictors in the model plus 1.

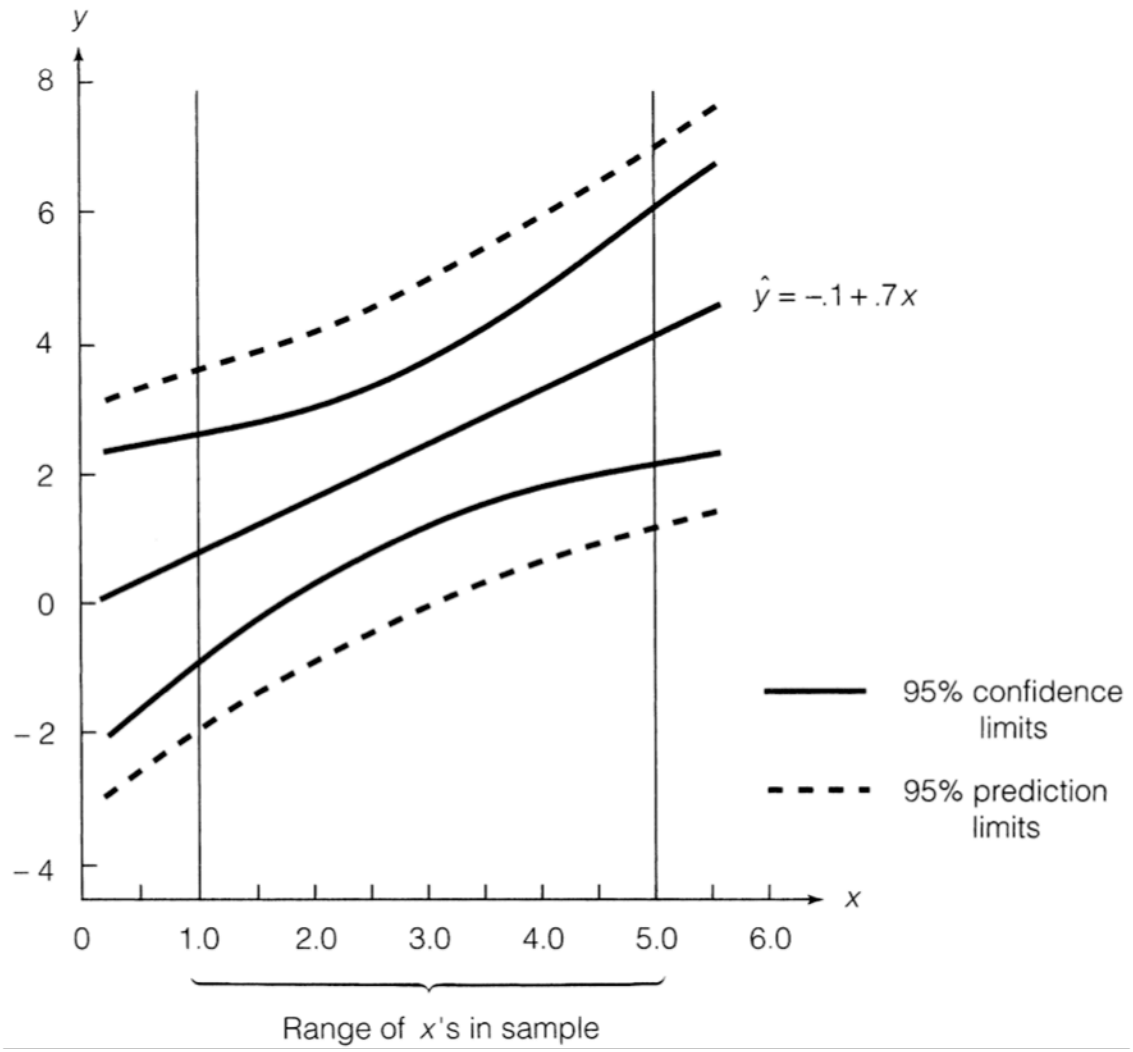
Derivation of $\sigma^2_{\text{pred}}\{\hat{Y}_h\}$

The formula of the prediction variance $\sigma^2_{\text{pred}}\{\hat{Y}_h\}$ is formed by considering two independent sources of errors:

1. The **sampling variance** when using $\hat{Y}_h$ to predict $E\{Y_h\}$ (i.e., the variation of the fitted line), that is, $\sigma^2\{\hat{Y}_h\}$.
2. The **residual variance** when using $E\{Y_h\}$ to predict Y_h (i.e., the variation from the line to the point), that is, σ^2.

Because these two sources of errors are **independent**, the variance of the summation is the summation of the variances:

$$\begin{aligned}\sigma^2_{\text{pred}}\{\hat{Y}_h\} &= \sigma^2\{\hat{Y}_h\} + \sigma^2 \\ &= \sigma^2\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right] + \sigma^2 \\ &= \sigma^2\left[1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right]\end{aligned}$$



- For a given $X = X_h$, the prediction interval is always wider than the confidence interval for $E\{Y_h\}$.
- The prediction interval for Y_h is the narrowest when $X_h = \bar{X}$. In other words, our prediction for cases close to the boundary of X is less accurate.

Outline

1. Normal error regression model
2. Sampling distribution
3. Confidence interval for regression coefficients
4. Confidence interval of $E\{Y_h\}$
5. Prediction interval of Y_h
6. Hypothesis testing of regression coefficients

Hypothesis test of β_1

Assume the significance level $\alpha = .05$. Conduct a two-sided t-test for $H_0 : \beta_1 = 0 \quad vs. \quad H_1 : \beta_1 \neq 0$.

```
Call:
lm(formula = sales_y ~ adv_x)

Residuals:
         1          2          3          4          5
 4.000e-01 -3.000e-01 -3.886e-16 -7.000e-01  6.000e-01

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.1000     0.6351  -0.157   0.8849
adv_x         0.7000     0.1915   3.656   0.0354 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6055 on 3 degrees of freedom
Multiple R-squared:  0.8167, Adjusted R-squared:  0.7556
F-statistic: 13.36 on 1 and 3 DF,  p-value: 0.03535
```

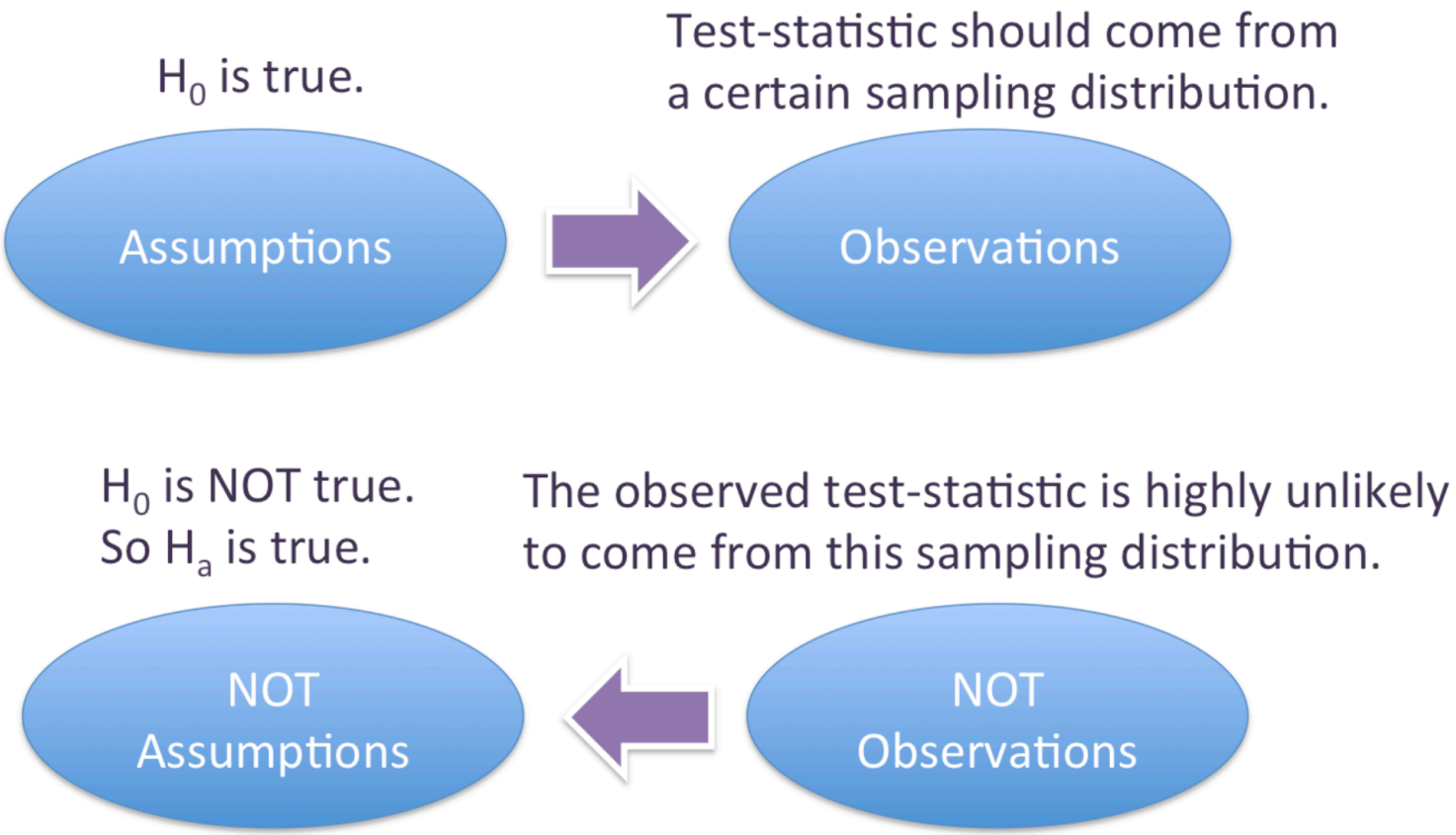
Hypothesis testing of β_1 (the 5 steps)

1. **Assumptions:** ε i.i.d. $\sim \mathrm{N}(0, \sigma^2)$
2. **Hypotheses:** $H_0 : \beta_1 = 0 \qquad H_1 : \beta_1 \neq 0$
3. **Test-statistic:**

$$t_{obs} = \frac{b_1 - \beta_1 | H_0}{s\{b_1\}} = \frac{b_1 - 0}{s\{b_1\}} \sim t(df = n - p)$$

4. **P-value:** The two-tail probability in the above sampling distribution beyond t_{obs} — If the true population slope $\beta_1 = 0$, what is the probability that we get a b_1 this far from 0 or even more extreme?
   R code: `(1-pt(q=`$|t_{obs}|$`, df=`$(n-p)$`))*2`
5. **Conclusion:**
   - If p-value $< \alpha$ (the pre-determined significance level, typically 0.05), reject H_0 and conclude that the true population slope is not 0 — there is a linear association between X and Y.
   - If p-value $> \alpha$, we fail to reject H_0 and conclude there is not enough evidence to prove a linear association between X and Y.

See separate document for the logic of hypothesis testing.



Type I error, type II error, and power

Hypothesis Testing		Decision	
		Fail to reject H_0	**Reject H_0**
Truth	**H_0 is true**	Correctly fail to reject H_0	Falsely reject H_0 (Type I error)
	H_1 is true	Falsely fail to reject H_0 (Type II error)	Correctly reject H_0

Legal Trials		Decision	
		Innocent	**Guilty**
Truth	**Innocent**	Correctly acquit the defendant	Falsely convict the defendant (Type I error)
	Guilty	Falsely acquit the defendant (Type II error)	Correctly convict the defendant

Medical Diagnostic Tests		Decision	
		Negative	**Positive**
Truth	**Disease absent**	Correctly clear the diagnosis	Falsely diagnose the disease (Type I error, False Positive, False Alarm)
	Disease present	Falsely clear the diagnosis (Type II error, False Negative, Miss)	Correctly detect the disease

Type I error, type II error, and power

Type I error

- If we reject H_0 when in fact H_0 is true, we make a Type I error.
- The probability of making a Type I error is called the Type I error rate, and is equal to the significance level, α.
- If H_0 is true and $\alpha = 0.05$:
  - About 5% of all samples from this population will lead us to incorrectly reject H_0.
  - In this case, the extremity of the test statistic is completely due to chance.
- We can control the Type I error rate by our choice of the significance level.
- The more serious the consequences of a Type I error, the smaller α should be.

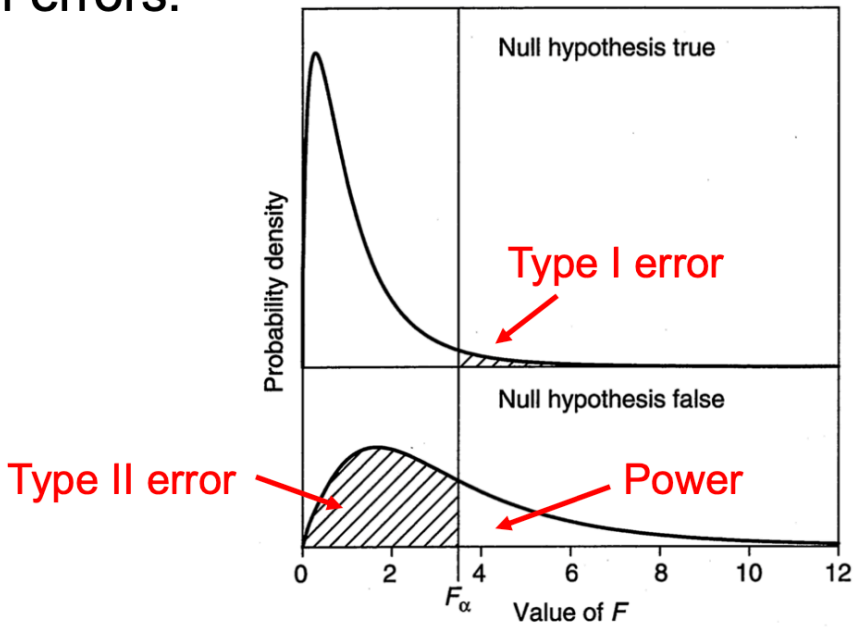
Type I error, type II error, and power

Type II error

- If we fail to reject H_0 when in fact H_1 is true, we make a Type II error.
- The probability of making a Type II error is called the Type II error rate, denoted by β.
- $Power = 1 - \beta$.
- We strive to find a test with high power.

Things that affect the power of a test

- For a fixed testing scenario, as the Type I error rate (α) goes down, the Type II error rate (β) goes up --- a trade-off between the two types of errors.



$\alpha \downarrow \rightarrow$ Power $\downarrow$

- For a fixed significance level, power increases:
  - as the true effect gets larger (i.e., the parameter value moves farther away from the H_0 value and into the H_1 values.)
  - as the sample size increases.

Statistical significance vs. practical significance

Statistical significance does not mean practical significance

- A hypothesis test only gives us information about **whether** the parameter differs from the H_0 value, but it does NOT tell us **how much** difference there is in a practical sense.
- The **statistical significance** only tells us whether the observed effects should be taken seriously, or more likely just random noise.
- We need to assess the **effect size** to indicate the **practical significance** of the findings.

Statistical significance vs. practical significance

"The fallacy of the obese N"

- When the sample size is very large, tiny deviations from the null hypothesis (with little practical consequence) may be statistically significant.
- When the sample size is very small, large deviations from the null hypothesis (of great practical importance) might go undetected (statistically insignificant).
- Effect size measures can be used to reflect the practical significance and are not affected by sample size.

See separate document for controversies over hypothesis testing.

- Why is $p = 0.49$ significant but $p = 0.51$ is not?
- Why $\alpha = 0.05$?
- The publication bias
- The file drawer effect
- p-hacking
- Study registration
- The neoFisherian Framework
- The worshiping of original research and the premature generalization of single studies
- The alternatives to HT: descriptive statistics, confidence interval, model comparison