450A Political Methodology I

Lecture 4: Simple Regression

3. Inference and Interpretation

Stanford University

Inference with Ordinary Least Squares

Testing and Confidence Intervals

Large Sample Inference

Interpreting OLS Coefficients

Testing and Confidence Intervals

Three ways of making statistical inference out of regression:

1. Point Estimation: Consider the sampling distribution of our point estimator $\hat{\beta}_1$ to infer β_1
2. Hypothesis Testing: Consider the sampling distribution of a test statistic to test hypothesis about β_1 at the α level
3. Interval Estimation: Consider the sampling distribution of an interval estimator to construct intervals that will contain β_1 at least $100(1-\alpha)\%$ of the time.

For 2 and 3, we need to know more than just the mean and the variance of the sampling distribution of $\hat{\beta}_1$. We need to know the full shape of the sampling distribution of our estimators $\hat{\beta}_0$ and $\hat{\beta}_1$.

OLS Assumption VI

Assumption (VI. Normality)

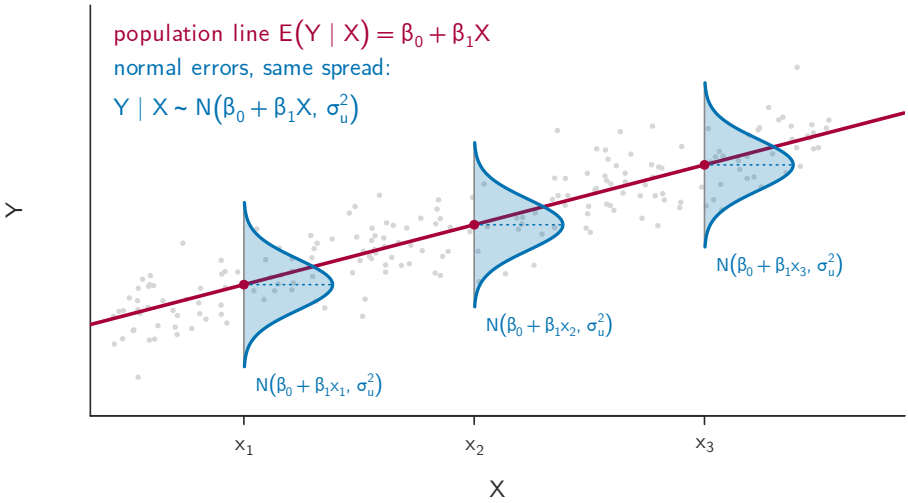
The population error term is independent of the explanatory variable, $u \perp\!\!\!\perp X$, and is normally distributed with mean zero and variance σ_u^2:

$$u \sim N(0, \sigma_u^2), \quad \textit{which implies} \quad Y|X \sim N(\beta_0 + \beta_1 X, \sigma_u^2)$$

Note: This implies homoskedasticity and zero conditional mean.

- Together Assumptions I–VI are the classical linear model (CLM) assumptions.
- The CLM assumptions imply that OLS is BUE (i.e. minimum variance among all linear or non-linear unbiased estimators)
- Non-normality of the errors is a serious concern in small samples. We can *partially* check this assumption by looking at the residuals
- Variable transformations can help to come closer to normality
- We don't need normality assumption in large samples (more on this later)

Normality: $Y|X \sim N(\beta_0 + \beta_1 X, \sigma_u^2)$



Sampling Distribution for $\widehat{\beta}_1$

Theorem (Sampling Distribution of $\widehat{\beta}_1$)

Under Assumptions I–VI, conditional on X,

$$\widehat{\beta}_1 \mid X \sim N\left(\beta_1, Var[\widehat{\beta}_1 \mid X]\right)$$

where

$$Var[\hat{\beta}_1 \mid X] = \frac{\sigma_u^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

which implies

$$\frac{\widehat{\beta}_1 - \beta_1}{\sqrt{Var[\hat{\beta}_1 \mid X]}} = \frac{\widehat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \sim N(0,1)$$

Proof.

Given Assumptions I–VI, $\hat{\beta}_1$ is a linear combination of the i.i.d. normal random variables:

$$\hat{\beta}_1 = \beta_1 + \sum_{i=1}^n \frac{(x_i - \bar{x})}{SST_x} u_i \quad \text{where} \quad u_i \sim N(0, \sigma_u^2).$$

Any linear combination of independent normals is normal, and we can transform/standardize any normal random variable into a standard normal by subtracting off its mean and dividing by its standard deviation. □

The t-Test for Single Population Parameters

- $SE[\hat{\beta}_1] = \frac{\sigma_u}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$ involves the unknown population error variance σ_u^2
- Replace σ_u^2 with its unbiased estimator $\hat{\sigma}_u^2 = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-2}$, and we obtain:

Theorem (Sampling Distribution of t-value)

Under Assumptions I–VI, the t-value for β_1 has a t-distribution with $n-2$ degrees of freedom:

$$T \equiv \frac{\widehat{\beta}_1 - \beta_1}{\widehat{SE[\hat{\beta}_1]}} \sim \tau_{n-2}$$

Proof sketch.

- Given X, $\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})u_i}{SST_x}$ is a linear combination of normal errors, so $Z \equiv \frac{\hat{\beta}_1 - \beta_1}{SE[\hat{\beta}_1]} \sim N(0,1)$
- $Q \equiv \frac{(n-2)\hat{\sigma}_u^2}{\sigma_u^2} = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{\sigma_u^2} \sim \chi^2_{n-2}$: the residuals lose two degrees of freedom to $\hat{\beta}_0$ and $\hat{\beta}_1$
- Under normal errors, $\hat{\beta}_1$ is independent of the residuals (fitted values and residuals are orthogonal), so Z and Q are independent
- Hence $T = \frac{Z}{\sqrt{Q/(n-2)}} \sim \tau_{n-2}$, by the definition of the t-distribution

□

Using the t-Value as a Test Statistic

$$T \equiv \frac{\widehat{\beta}_1 - c}{\widehat{SE}[\hat{\beta}_1]} \sim \tau_{n-2} \quad \text{if} \quad \beta_1 = c$$

1. Assuming some null value c (e.g. $c = 0$) and given the number of observations we know that the sampling distribution of this test statistic is τ_{n-2}

2. Given the null value, we can derive the probability that the value of our test statistics falls into a given interval:
$$P\left(-t_{\alpha/2}(n-2) \leq T \leq t_{\alpha/2}(n-2)\right) = 1 - \alpha$$

3. Since we can also compute $\widehat{\beta}_1$ and $\widehat{SE}[\hat{\beta}_1]$ we can easily compute the realized t-value in our application.

4. We can then check if the realized value is "unusual" given the known distribution of the test statistic.

The Testing Workflow

1. Specify the null hypothesis and alternative hypothesis. This is equivalent to setting the null value c.

Consider our regression example where Y: Trump's share of the two-party vote in 2024 (%) and X: % of adults 25 and older with a bachelor's degree, in the 50 states.

Suppose we are interested in testing $H_0 : \beta_1 = c = 0$, versus $H_1 : \beta_1 \neq c$. In words we test whether the population slope is zero.

- The null hypothesis is typically an assumed state of the world, or a strawman that we would like to knock down.
- The alternative hypothesis is typically a claim we would like to test. We call this particular alternative hypothesis a two-sided alternative, because evidence on either side of the hypothesized null value will help us to reject the null.
- Notice we can use other values for c than zero.

The Testing Workflow

1. Specify the null hypothesis and alternative hypothesis
2. Specify the test level α (= our tolerance for a type I error)

Type I error = rejecting the null hypothesis when it is true; α is the probability of doing so that we accept

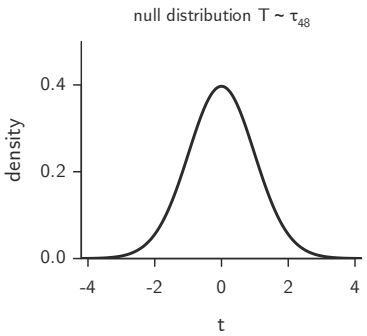
What is a good level of α?

- We typically want strong evidence to reject the null, so we set α to a small number like .05.

The Testing Workflow

1. Specify the null hypothesis and alternative hypothesis.
2. Specify the test level α.
3. Examine the null distribution of the test statistic $T \sim \tau_{n-2}$

We have 50 observations so the null distribution is $T \sim \tau_{48}$



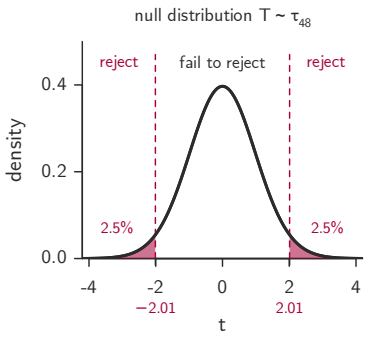
The Testing Workflow

1. Specify the null hypothesis and alternative hypothesis.
2. Specify the test level α.
3. Examine the null distribution of the test statistic
4. Determine the rejection region

At $\alpha = .05$ the critical value $t_{\alpha/2}(48)$ is?
2.01

So we reject H_0 if:

$$|\text{our realized t-value}| > 2.01$$

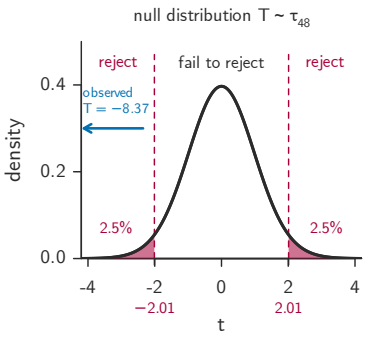


The Testing Workflow

1. Specify the null hypothesis and alternative hypothesis.
2. Specify the test level α.
3. Examine the null distribution of the test statistic
4. Determine the rejection region
5. Compute the observed value of T and compare it to the critical value

Here $\hat{\beta}_1 = -1.4537$ and $\widehat{SE}(\hat{\beta}_1) = 0.17376$ so $T = \frac{-1.4537-0}{0.17376} = -8.37$.
So what do we conclude?

We reject H_0 at the .05 level



College Degrees and the 2024 Trump Vote, 50 States

By default, R provides the t-values for the test against $H_0 : \beta = 0$ for slope and intercept

R Code

```
> lm.out <- lm(trump ~ ba, data = states)
> summary(lm.out)

~~~~ output omitted ~~~~

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 102.8376     5.9818  17.192  < 2e-16 ***
ba           -1.4537     0.1738  -8.366  6.2e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

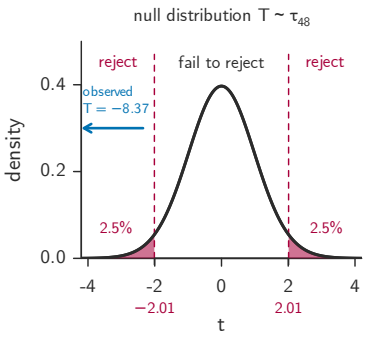
Residual standard error: 6.589 on 48 degrees of freedom
Multiple R-squared:  0.5932,        Adjusted R-squared:  0.5847
F-statistic: 69.99 on 1 and 48 DF,  p-value: 6.199e-11
```

The p-value

- The p-value summarizes our evidence against H_0: it's the probability of observing a t-value (over repeated samples) at least as extreme as one we observe assuming that H_0 is true.
- p-value is $P(|T_0| > |T_{obs}|)$, the probability that the t-statistic under the null is larger in absolute value than the observed t value.
- If p-value $\leq \alpha$ we reject H_0 at the α level. It is therefore the smallest α for which we can reject H_0.

Given that $T_{obs} = -8.37$ will the p-value be small or large?

p-value is 6.2×10^{-11} so we have strong evidence against H_0.



College Degrees and the 2024 Trump Vote, 50 States

R Code

```
> summary(lm.out)

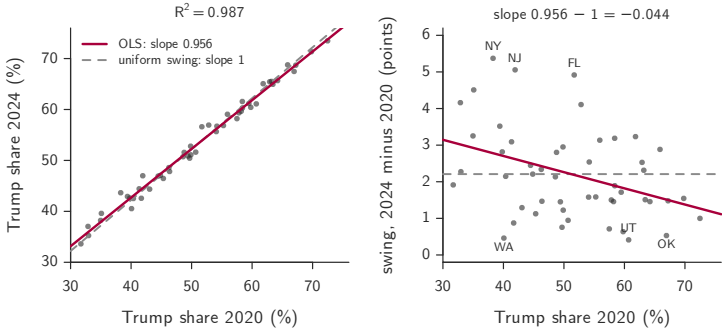
~~~~ output omitted ~~~~

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 102.8376     5.9818  17.192  < 2e-16 ***
ba           -1.4537     0.1738  -8.366  6.2e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.589 on 48 degrees of freedom
Multiple R-squared:  0.5932,        Adjusted R-squared:  0.5847
F-statistic: 69.99 on 1 and 48 DF,  p-value: 6.199e-11
```

A Null That Matters: Uniform Swing

Regress Trump's 2024 share on his 2020 share across the 50 states. Testing $\beta_1 = 0$ here tells us nothing new, because 2020 shares obviously predict 2024 shares. The interesting null is $H_0 : \beta_1 = 1$. It says every state moved by the same number of points (a uniform swing).



Dashed: the best line with slope 1 (every state moves 2.21 points). Solid: OLS. Right: the same two lines, with the swing (2024 minus 2020) on the vertical axis.

Source: MIT Election Data and Science Lab, 2020 and 2024 presidential returns by state. DC excluded.

A Null That Matters: Testing $H_0 : \beta_1 = 1$

R Code

```
> summary(lm(trump24 ~ trump20, data = states))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.47307    0.81741   5.472 1.59e-06 ***
trump20      0.95578    0.01564  61.115  < 2e-16 ***
Residual standard error: 1.164 on 48 degrees of freedom
Multiple R-squared:  0.9873
```

- R's t value (61.1) tests $\beta_1 = 0$. For $H_0 : \beta_1 = 1$, set $c = 1$. Then $T = \frac{0.9558-1}{0.0156} = -2.83$. With 48 df, $p = .007$. We reject uniform swing at the .05 level.
- Every state moved toward Trump. Because $\hat{\beta}_1 < 1$, states that were more Republican in 2020 moved a little less. A state 10 points more Republican in 2020 has a fitted swing 0.44 points smaller.
- $R^2 = .987$ is not the point. It says 2020 shares predict 2024 shares almost perfectly. It does not tell us whether the slope is 1. The uniform-swing line fits almost as well ($R^2 = .985$).

One-Sided Versus Two-Sided Test

- Using H_0: $\beta_1 = 0$ and H_1: $\beta_1 \neq 0$ is called a two-sided test because the null hypothesis can be rejected in both directions ($\beta_1 < 0$ and $\beta_1 > 0$)

- What is a one-sided test? H_0: $\beta_1 = 0$ and H_1: $\beta_1 < 0$ (say we were sure, before seeing the data, that states with more college graduates cannot give Trump a larger share). The null hypothesis can be rejected only in one direction.

- Given that we keep $\alpha = .05$ constant. What does this imply for the acceptance region? All of the 5% rejection region now sits in the lower tail, so the critical value is closer to zero: -1.68 instead of ± 2.01 with 48 degrees of freedom. We reject more easily when $\hat{\beta}_1 < 0$, and never when $\hat{\beta}_1 > 0$.

- When is a one-sided test appropriate? If the null can be wrong in only one direction and the other direction is completely inconceivable *a priori*. In other words, rarely in social sciences.

Confidence Intervals

Suppose instead of a point estimator, we want to be more conservative, and specify an interval as our estimate for the location of the true slope β_1.

This is straightforward now since we know the sampling distribution for our t-value:

$$T = \frac{\widehat{\beta}_1 - \beta_1}{\widehat{SE}[\widehat{\beta}_1]} \sim \tau_{n-2}$$

So we also know the probability that the value of our test statistics falls into a given interval:

$$P\left(-t_{\alpha/2}(n-2) \leq \frac{\widehat{\beta}_1 - \beta_1}{\widehat{SE}[\widehat{\beta}_1]} \leq t_{\alpha/2}(n-2)\right) = 1 - \alpha$$

In our example with $n = 50$ and $\alpha = .05$ we know that:

$$P\left(-2.01 \leq \frac{\widehat{\beta}_1 - \beta_1}{\widehat{SE}[\widehat{\beta}_1]} \leq 2.01\right) = .95$$

Recall: $t_{\alpha/2}(n-2)$ refers to the $100(1-\alpha/2)$ percentile in the τ_{n-2} distribution.

Confidence Intervals

We rearrange:

$$P\left(-t_{\alpha/2}(n-2) \leq \frac{\widehat{\beta}_1 - \beta_1}{\widehat{SE}[\hat{\beta}_1]} \leq t_{\alpha/2}(n-2)\right) = 1-\alpha$$

$$P\left(\widehat{\beta}_1 - t_{\alpha/2}(n-2)\widehat{SE}[\hat{\beta}_1] \leq \beta_1 \leq \widehat{\beta}_1 + t_{\alpha/2}(n-2)\widehat{SE}[\hat{\beta}_1]\right) = 1-\alpha$$

So we construct the confidence interval using:

$$\hat{\beta}_1 \pm t_{\alpha/2}(n-2) \cdot \widehat{SE}[\hat{\beta}_1]$$

In our example:

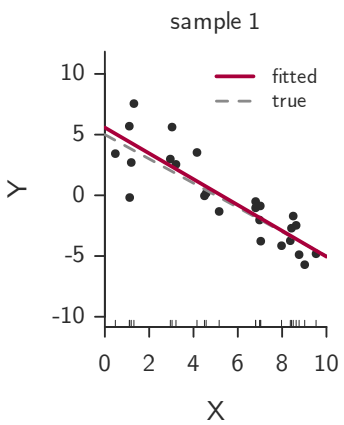
$$-1.45 \pm 2.01 \cdot 0.174 \text{ so } [-1.80\,,\ -1.10]$$

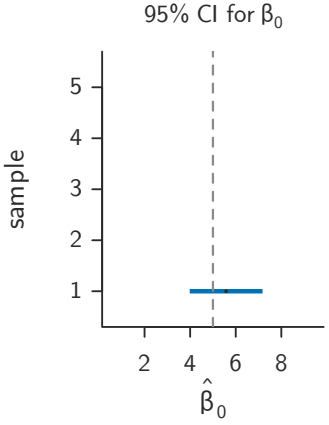
Note that the random quantities in this statement are $\widehat{\beta}_1$ and $\widehat{SE}[\hat{\beta}_1]$. Therefore, once the data are observed, nothing else is random!

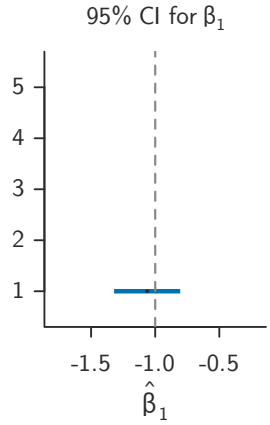
In R: `confint(lm.out)`

CIs Simulation Example

Returning to the simulation example of Lecture 4-2, we can simulate the sampling distributions of the 95% confidence intervals for β_1 and β_0

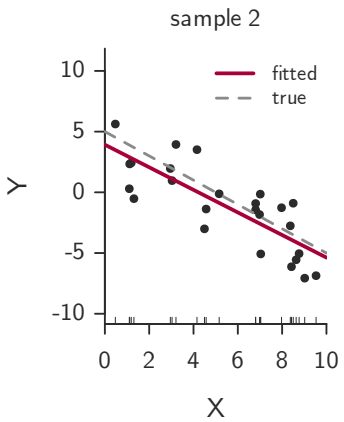


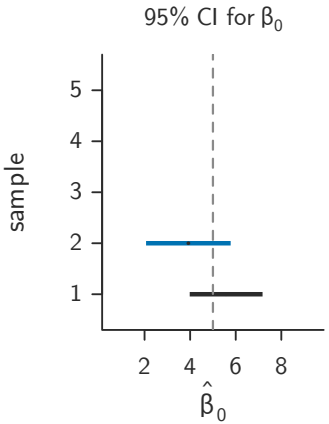


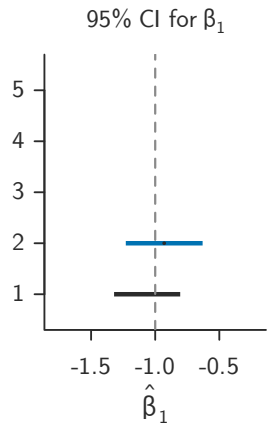


CIs Simulation Example

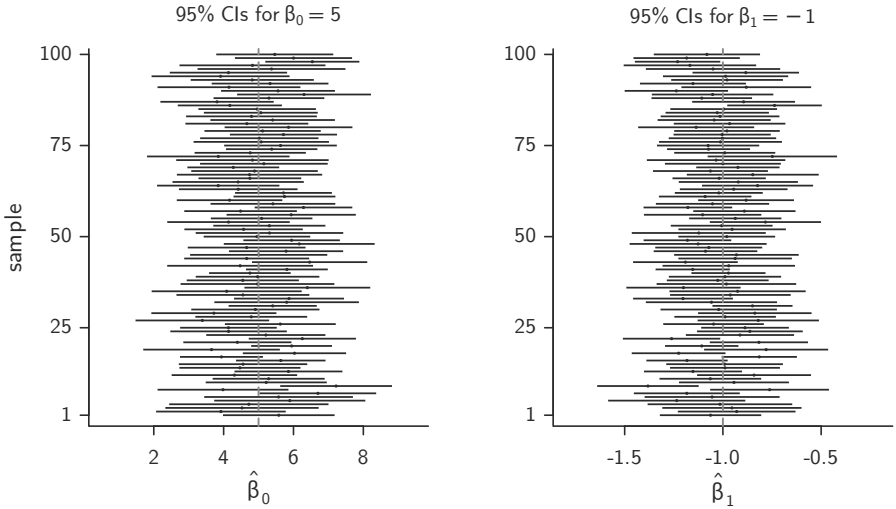
Returning to the simulation example of Lecture 4-2, we can simulate the sampling distributions of the 95% confidence intervals for β_1 and β_0



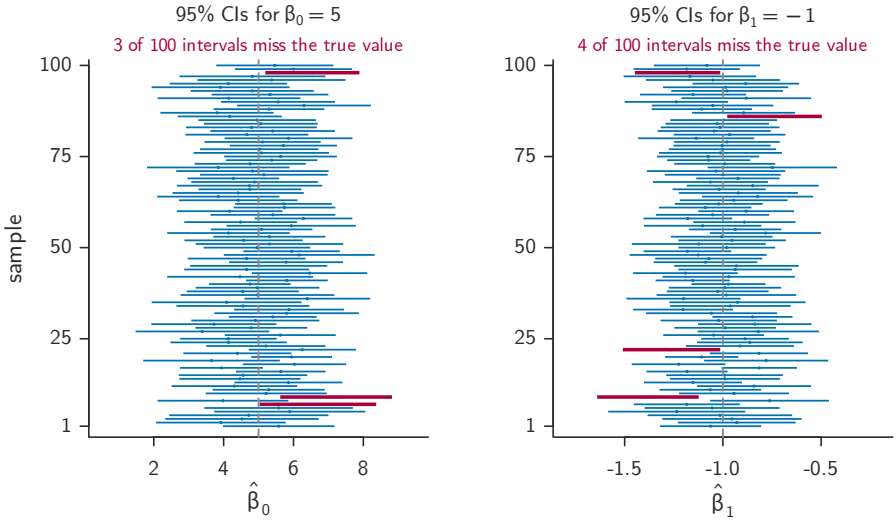




CIs Simulation Example



CIs Simulation Example



Testing with Confidence Intervals

- Notice that the boundary of the upper critical region is 2.01
- Recall that this was the value of $t_{\alpha/2}(n-2)$ from our 95% CI
- This is not a coincidence, and it suggests a method for hypothesis testing with a confidence interval.

Procedure:

1. Pick $H_0 : \beta_1 = c$ and a significance level α
2. Form a $100(1-\alpha)\%$ confidence interval
3. Reject the null if the CI doesn't contain c

Our 95 % CI was $[-1.80\,,\ -1.10]$ so we reject H_0: $\beta_1 = 0$ at the $\alpha = 0.05$ level.

Inference with Ordinary Least Squares

Testing and Confidence Intervals

Large Sample Inference

Interpreting OLS Coefficients

Large Sample Properties: Consistency

- So far we have studied the small sample properties of the OLS estimator, i.e., how $(\hat{\beta}_0, \hat{\beta}_1)$ behaves *in repeated samples* of a given n.
- Now, we look at their large sample properties, i.e., how $(\hat{\beta}_0, \hat{\beta}_1)$ behaves *when* $n \to \infty$.

Theorem (Consistency of OLS Estimator)

Given Assumptions I–IV, the OLS estimator $\widehat{\beta}_1$ *is consistent for* β_1 *as* $n \to \infty$*:*

$$\operatorname*{plim}_{n\to\infty} \widehat{\beta}_1 = \beta_1$$

Proof.

Worksheet Q1 □

Large Sample Properties: Consistency

- Technical note: We can slightly relax Assumption IV:

$$E[u|X] = 0 \quad \text{(any function of } X \text{ is uncorrelated with } u\text{)}$$

to its implication:

$$\text{Cov}[u, X] = 0 \quad (X \text{ is uncorrelated with } u)$$

for consistency to hold (but not unbiasedness).

- OLS is inconsistent (and biased) unless $\text{Cov}[X, u] = 0$
- If $\text{Cov}[u, X] > 0$ then asymptotic bias is upward; if $\text{Cov}[u, X] < 0$ asymptotic bias is downwards

Large Sample Properties: Asymptotic Normality

- For statistical inference, we need to know the sampling distribution of $\hat{\beta}$ when $n \to \infty$.

Theorem (Asymptotic Normality of OLS Estimator)

Given Assumptions I–V, the OLS estimator $\widehat{\beta}_1$ is asymptotically normally distributed:

$$\frac{\hat{\beta}_1 - \beta_1}{\widehat{SE}[\hat{\beta}_1]} \overset{approx.}{\sim} N(0,1)$$

where $\widehat{SE}[\hat{\beta}_1] = \frac{\hat{\sigma}_u}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$ with the consistent estimator for the error variance: $\hat{\sigma}_u^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2 \xrightarrow{p} \sigma_u^2$ (dividing by n also works)

Worksheet Q2

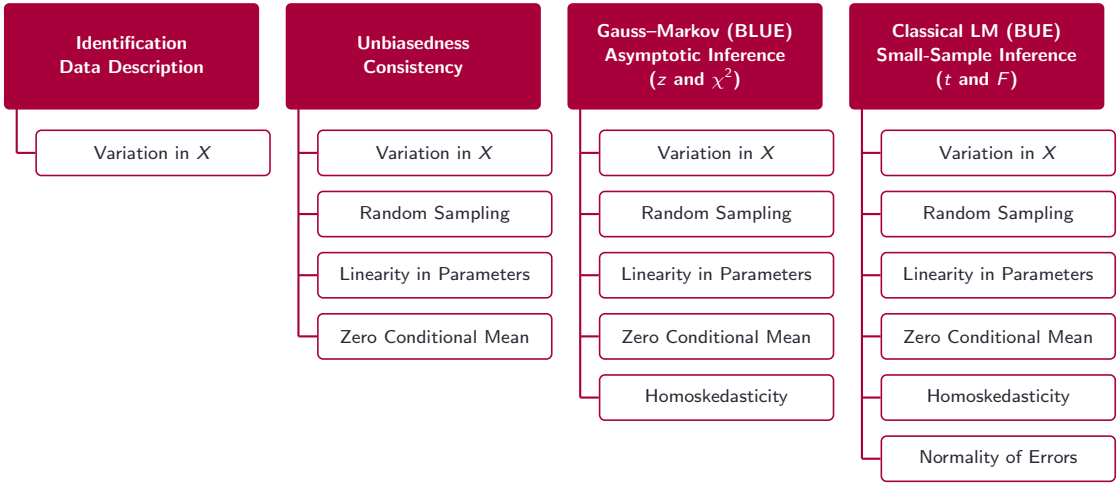
Large Sample Inference

- We need homoskedasticity (Assumption V) for this result, but we do not need normality (Assumption VI).
- This result implies that asymptotically our usual standard errors, t-values, p-values, and CIs remain valid even without the normality assumption! We just proceed as in the small sample case where we assume normality.
- It turns out that, given Assumptions I–V, the OLS asymptotic variance is also the lowest in class (asymptotic Gauss-Markov).

Interpreting OLS Coefficients

- Predictive vs. Causal Interpretations
- Practical Advice
- Incorporating Nonlinearities using Log

OLS Assumptions Summary



What Do the Regression Coefficients Mean Substantively?

- So far, we have learned the statistical properties of the OLS estimator
- However, these properties do not tell us what types of inference we can draw from the estimates

Three types of inference:

1. Descriptive inference:
   - Summarizing sample data by drawing the "best fitting" line
   - No inference about the underlying population intended
   - Assumption required: III (variation in X) only
2. Predictive inference:
   - Inference about a new observation coming from the same population
   - Example: Wage (Y) and education (X):
     "What's my best guess about the wage of a new worker who only has high school education?"
   - Assumptions required: III and II (random sampling)
   - Assumptions desired: I (linearity)

What Do the Regression Coefficients Mean Substantively?

3. Causal inference:
   - Inference about counterfactuals, i.e. hypothetical interventions to the same units
   - Example: Wage (Y) and education (X):
     "What would my current wage be if I only had high school education?"
   - Assumptions required (under the current framework): I, II, III and IV (zero conditional mean)
   - In 450B, we will introduce a different framework which provides more nuanced views about causal identification assumptions

- Notice in the wage example, how the omission of unobserved ability from the equation does or does not affect each type of inference

- Implications:
  - When Assumptions I–IV are all satisfied, we can estimate the structural parameters β without bias and thus make causal inference.
  - However, we can make predictive inference even if some assumptions are violated.

OLS as a Best Linear Predictor

- Suppose that we want to predict the values of Y given observed X values
- Suppose further that we've decided to *use* a linear predictor $\hat{\beta}_0 + \hat{\beta}_1 X$ (but not necessarily *assume* a true linear relationship in the population)
- How to choose a good predictor? A popular criterion is mean squared error:

$$MSE \;=\; E\left[(Y - b_0 - b_1 X)^2\right]$$

  for a candidate line $b_0 + b_1 X$. The smaller a predictor makes MSE, the better.
- Now, note that the sample version of MSE is $\frac{1}{n}\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2$
- Recall from Lecture 4-1 how we got the OLS estimator; we *minimized* $\sum_{i=1}^{n}(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)^2$ over $(\tilde{\beta}_0, \tilde{\beta}_1)$!
- This implies that OLS gives the best linear predictor *in the sample*, and it estimates the population best linear predictor, the line that minimizes $E[(Y - b_0 - b_1 X)^2]$
- Which assumptions did we use to get this result?
  - Needed: Assumptions II (random sampling) and III (variation in X)
  - *Not* needed: Assumptions I (linearity) and IV (zero cond. mean)
- Note that if $E[Y|X]$ is linear (Assumptions I and IV), the population line is the best, not just best linear, predictor, so linearity is certainly desired

OLS as an Estimator of Predicted Difference

- Recall the bias expression for $\hat{\beta}_1$ under Assumptions I–III (and not IV) from Lecture 4-2 for a binary explanatory variable:

$$E[\hat{\beta}_1] \;=\; \beta_1 \;+\; \{E[u \mid X = 1] - E[u \mid X = 0]\}$$

- Suppose that the association between u and X is linear:

$$u \;=\; \gamma_0 + \gamma_1 X + \eta$$

where $E[\eta \mid X] = 0$ (Note that this is consistent with Assumption I)

- Then, $\hat{\beta}_1$ estimates a different parameter β_1^* without bias:

$$\begin{aligned} E[\hat{\beta}_1] &= \beta_1 + \{E[u|X = 1] - E[u|X = 0]\} \\ &= \beta_1 + \{(\gamma_0 + \gamma_1 \cdot 1 + 0) - (\gamma_0 + \gamma_1 \cdot 0 + 0)\} \\ &= \beta_1 + \gamma_1 \;\equiv\; \beta_1^* \end{aligned}$$

- β_1^* has a predictive interpretation, i.e., the expected difference between two randomly-drawn units that happen to have $X = 1$ and $X = 0$ respectively

Interpreting OLS Coefficients

- Predictive vs. Causal Interpretations
- Practical Advice
- Incorporating Nonlinearities using Log

College Degrees and the 2024 Trump Vote, 50 States

R Code

```
> summary(lm(trump ~ ba, data = states))

~~~~ output omitted ~~~~

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 102.8376     5.9818  17.192  < 2e-16 ***
ba           -1.4537     0.1738  -8.366  6.2e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.589 on 48 degrees of freedom
Multiple R-squared:  0.5932,        Adjusted R-squared:  0.5847
F-statistic: 69.99 on 1 and 48 DF,  p-value: 6.199e-11
```

"Across states, a share of adults with a bachelor's degree one percentage point higher is associated with a Trump share of the two-party vote 1.45 percentage points lower ($p < 0.001$)."

Ground Rules: Interpretation of the Slope

1. Give a short, but precise interpretation of the exact meaning of the value of the slope coefficient referring to the concepts, units, direction, and magnitude.
   - Estimate suggests that one additional hour of reading the textbook is associated with 10 additional points on the exam.

2. Do not resort to unwarranted causal language: Say "predicts", "associated with", "expected difference" or "correlated with" instead of "causes", "leads" or "affects."

3. Give a short, but precise interpretation of statistical significance.

4. Give a short, but precise interpretation of practical significance. You want to discuss the magnitude of the slope in your particular application.

Reporting Statistical Significance

- A reasonable way to think about statistical significance is to think about the precision of the estimates.
- If the slope is large substantively but just barely fails to reach conventional levels of significance it may still be interesting.
- Examples:
  - We reject the null hypothesis that the slope is zero at the .05 level.
  - The slope coefficient suggests that a one unit change in X is associated with a 10 unit change in Y ($p < .02$).
  - The slope coefficient is fairly precisely estimated, the 95 % confidence interval ranging from 8 to 12.

Reporting Substantive Significance

- Statistical significance and substantive significance are not the same: with a large enough sample size even truly microscopic differences can be statistically significant!
- To comment on substantive magnitude you should set up a "plausible" contrast keeping in mind (1) the distributions of variables and (2) the substantive context
- Examples (the numbers in the first and third are made up for illustration):

  Earnings on Schooling: The slope estimate is $4,000 per year of schooling. The standard deviation is 2.5 years for schooling and $50,000 for annual earnings. So a one standard deviation increase in schooling is associated with $10,000 higher earnings, a .2 standard deviation increase.

Interpreting OLS Coefficients

- Predictive vs. Causal Interpretations
- Practical Advice
- Incorporating Nonlinearities using Log

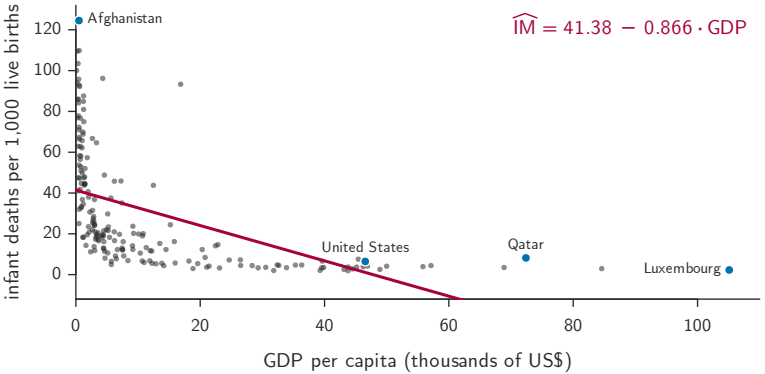
Incorporating Nonlinearities in X

- The linear regression model *can* accommodate non-linearity in X (but not in β)
- We do this by first transforming X appropriately

- A useful transformation when variables are positive and right-skewed is the (natural) logarithm

- The log transformation changes the interpretation of β_1:
  - Regress $\log(Y)$ on $X \longrightarrow 100\beta_1$ approximates the percent change in Y associated with a one-unit increase in X
  - Regress Y on $\log(X) \longrightarrow \beta_1/100$ approximates the change in Y associated with a one percent increase in X
  - Note that these approximations work only for small changes
  - In particular, they can fail for a dummy D moving from 0 to 1 (Worksheet Q3)

Example: Infant Mortality and National Income

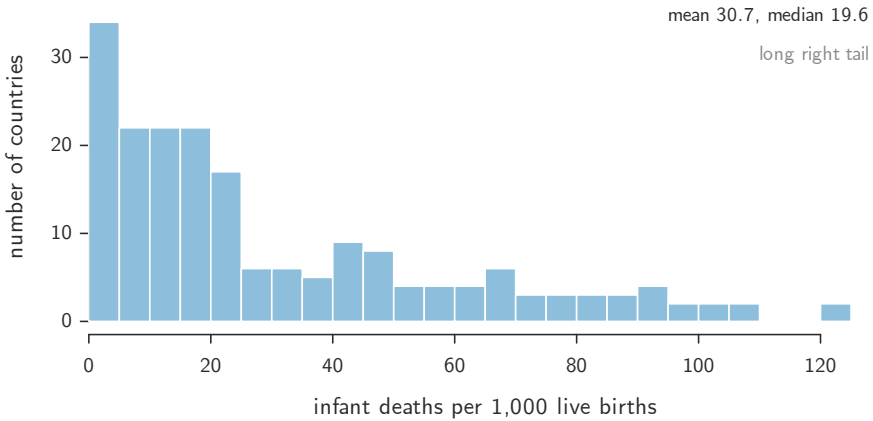


$\hat{\beta}_1 = -0.866 \longrightarrow$ An additional \$1,000 of GDP per capita predicts 0.866 fewer infant deaths per 1,000 live births on average

But the line crosses zero at \$47,800, so it predicts negative infant mortality for the 8 countries richer than that.

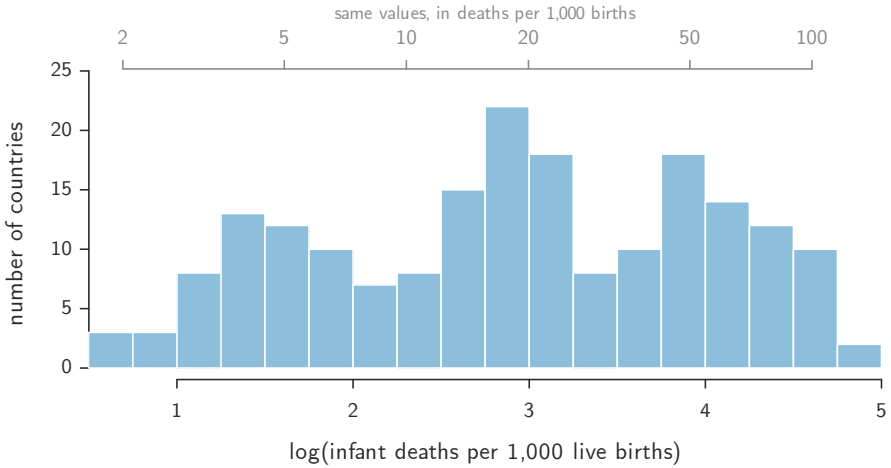
Source: United Nations social indicators, mostly 2009–2011; 193 of the 213 countries and territories, those with both values (R package `carData`, data set `UN`)

Infant Mortality (Scale in Levels)



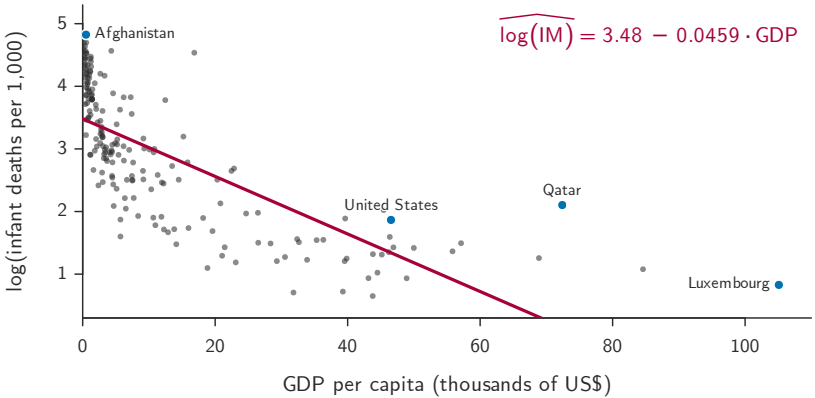
Source: United Nations social indicators, mostly 2009–2011; 193 of the 213 countries and territories, those with both values (R package `carData`, data set `UN`)

Infant Mortality (Logarithmic Scale)



Source: United Nations social indicators, mostly 2009–2011; 193 of the 213 countries and territories, those with both values (R package `carData`, data set `UN`)

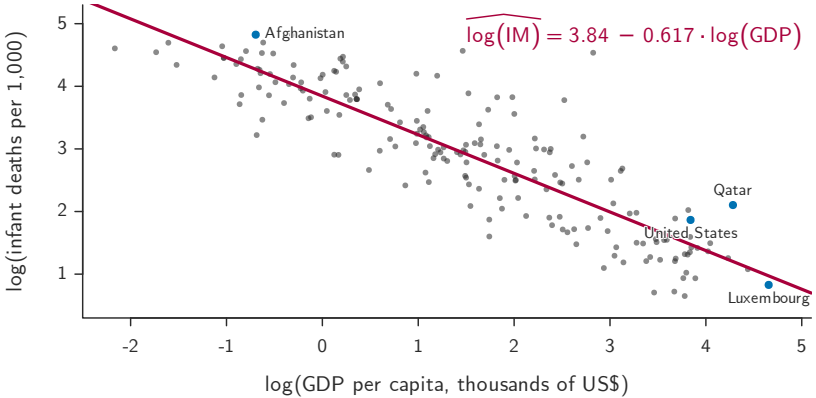
Regression: Log-Level



$\hat{\beta}_1 = -0.0459 \longrightarrow$ An additional \$1,000 of GDP per capita predicts an infant mortality rate about 4.6 percent lower on average

Source: United Nations social indicators, mostly 2009–2011; 193 of the 213 countries and territories, those with both values (R package `carData`, data set `UN`)

Regression: Log-Log



$\hat{\beta}_1 = -0.617 \longrightarrow$ A one percent increase in GDP per capita predicts a 0.617 percent decrease in infant mortality on average

Source: United Nations social indicators, mostly 2009–2011; 193 of the 213 countries and territories, those with both values (R package `carData`, data set `UN`)

Four Most Commonly Used Models

Model	Equation	β_1 Interpretation
Level-Level	$Y = \beta_0 + \beta_1 X$	$\Delta Y = \beta_1 \Delta X$
Log-Level	$log(Y) = \beta_0 + \beta_1 X$	$\%\Delta Y \approx 100\beta_1 \Delta X$
Level-Log	$Y = \beta_0 + \beta_1 log(X)$	$\Delta Y \approx (\beta_1/100)\%\Delta X$
Log-Log	$log(Y) = \beta_0 + \beta_1 log(X)$	$\%\Delta Y \approx \beta_1 \%\Delta X$

- **Level–Level:** One additional year of education (X) is associated with an increase of β_1 thousand dollars in income (Y).
- **Log–Level:** One additional year of education (X) is associated with approximately $100\beta_1$ percent higher income (Y).
- **Level–Log:** A 1% increase in population size (X) is associated with approximately a $\beta_1/100$ unit increase in government spending (Y).
- **Log–Log:** A 1% increase in income (X) is associated with approximately a β_1% increase in consumption (Y).

Be Careful: Log-Level with binary X

Worksheet Q3

Consider $\log(Y) = \beta_0 + \beta_1 D + u$ with $D \in \{0, 1\}$, assuming $D \perp u$.

1. Derive the exact percent change in $E[Y \mid D]$ when D moves from 0 to 1.
2. Explain why the rule "$100 \cdot \beta_1$%" is generally invalid here, and relate the sign and magnitude of the approximation error to β_1.
3. For $\beta_1 = 0.30$, compute the exact percent change and the approximation error in percentage points.
4. Find the value of $\beta_1 > 0$ above which the absolute approximation error exceeds 1 percentage point.