Lecture 8: Sampling Distributions

Module 3: part 1

Spring 2024

Logistics

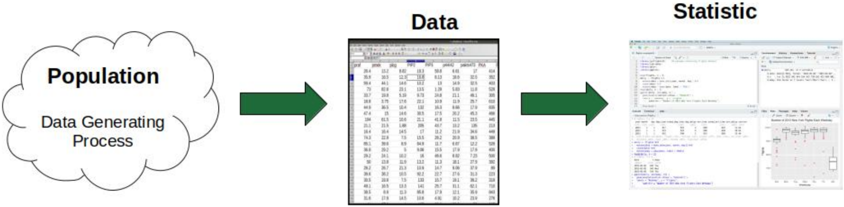
- Start Module 3 on inference and hypothesis testing
- Assessment for Module 2 due 2/20

Sampling Distributions

Sample data vs Population distribution

- In the lab, you fit a model for house prices which included an interaction between quality and age
- $\hat{\beta}_{age} = -0.0045991$
- What would happen if we gathered new data?

Sample data vs Population distribution



Estimator

- Statistic or estimator is a function which takes data as input, and outputs a number
- Examples: Mean, Median, Regression coefficient

Estimator

- Statistic or estimator is a function which takes data as input, and outputs a number
- Examples: Mean, Median, Regression coefficient
- If we have a model for how the data is generated, then we can also describe the distribution of the estimator

Sampling distribution of least squares estimators

Suppose the data is generated from our linear Gaussian model:

$$Y_i = b_0 + b_1 X_i + \varepsilon_i$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$.

Sampling distribution of least squares estimators

Suppose the data is generated from our linear Gaussian model:

$$Y_i = b_0 + b_1 X_i + \varepsilon_i$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$.

Key idea: Understanding how our estimates would vary if we repeated the sampling process.

- High level strategy: Condition on observed covariates (X) and analyze model behavior
- We remain agnostic about covariate generation:
  - Could be drawn from a distribution
  - Could be fixed by experimenter
- We'll focus on $\hat{b}_1$ as our primary coefficient of interest

Sampling distribution intuition

Goal: Derive the sampling distribution of $\hat{b}_1$ step by step.
Starting with our model:

$$Y_i = b_0 + b_1 X_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$$

$$\begin{aligned}
\hat{b}_1 = \frac{s_{xy}}{s_x^2} &= \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} && \text{(OLS formula)} \\
&= \frac{\sum_i (x_i - \bar{x})(b_0 + b_1 x_i + \varepsilon_i)}{\sum_i (x_i - \bar{x})^2} && \text{(Substitute } y_i\text{)} \\
&= b_0 \sum_i k_i + b_1 \sum_i k_i X_i + \sum_i k_i \varepsilon_i && \text{(Rearrange)}
\end{aligned}$$

where $k_i = \frac{x_i - \bar{x}}{\sum_i (x_i - \bar{x})^2}$ are the *standardized weights*.

Sampling distribution intuition

Goal: Derive the sampling distribution of $\hat{b}_1$ step by step.
Starting with our model:

$$Y_i = b_0 + b_1 X_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$$

$$\begin{aligned}
\hat{b}_1 = \frac{s_{xy}}{s_x^2} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} && \text{(OLS formula)} \\
= \frac{\sum_i (x_i - \bar{x})(b_0 + b_1 x_i + \varepsilon_i)}{\sum_i (x_i - \bar{x})^2} && \text{(Substitute } y_i\text{)} \\
= b_0 \sum_i k_i + b_1 \sum_i k_i X_i + \sum_i k_i \varepsilon_i && \text{(Rearrange)}
\end{aligned}$$

where $k_i = \frac{x_i - \bar{x}}{\sum_i (x_i - \bar{x})^2}$ are the *standardized weights*.
Key properties: $\sum_i k_i = 0$ and $\sum_i k_i x_i = 1$, leading to:

$$\hat{b}_1 = b_1 + \sum_i k_i \varepsilon_i$$

Expected Value of $\hat{b}_1$

Key Question: Is our estimator centered at the true value?
Using our simplified form: $\hat{b}_1 = b_1 + \sum_i k_i \varepsilon_i$

$$\begin{aligned}
E(\hat{b}_1 \mid X) &= E(b_1 + \sum_i k_i \varepsilon_i) && \text{(Linearity)} \\
&= b_1 + \sum_i k_i E(\varepsilon_i \mid X) && \text{(Pull out constants)} \\
&= b_1 + \sum_i k_i \cdot 0 && \text{(Key assumption)} \\
&= b_1
\end{aligned}$$

Expected Value of $\hat{b}_1$

Key Question: Is our estimator centered at the true value?
Using our simplified form: $\hat{b}_1 = b_1 + \sum_i k_i \varepsilon_i$

$$
\begin{aligned}
E(\hat{b}_1 \mid X) &= E(b_1 + \sum_i k_i \varepsilon_i) && \text{(Linearity)} \\
&= b_1 + \sum_i k_i E(\varepsilon_i \mid X) && \text{(Pull out constants)} \\
&= b_1 + \sum_i k_i \cdot 0 && \text{(Key assumption)} \\
&= b_1
\end{aligned}
$$

- **Key Assumption:** $E(\varepsilon_i \mid X) = 0$
- **Interpretation:** $\hat{b}_1$ is an unbiased estimator
- **Practical meaning:**
  - Each sample gives a different $\hat{b}_1$
  - But they cluster around the true b_1
  - No systematic over/under-estimation

Variance of $\hat{b}_1$

Key Question: How much does our estimator vary around its mean?

$$\begin{aligned}
\mathrm{var}(\hat{b}_1 \mid X) &= \mathrm{var}(\sum_i k_i \varepsilon_i) && \text{(From previous)} \\
&= \sum_i k_i^2 \mathrm{var}(\varepsilon_i \mid X) && \text{(Independence)} \\
&= \sigma_\varepsilon^2 \sum_i k_i^2 && \text{(Homoscedasticity)} \\
&= \frac{\sigma_\varepsilon^2}{\sum_i (x_i - \bar{x})^2} = \frac{\sigma_\varepsilon^2}{(n-1)s_x^2}
\end{aligned}$$

Variance of $\hat{b}_1$

Key Question: How much does our estimator vary around its mean?

$$
\begin{aligned}
\operatorname{var}(\hat{b}_1 \mid X) &= \operatorname{var}(\sum_i k_i \varepsilon_i) && \text{(From previous)} \\
&= \sum_i k_i^2 \operatorname{var}(\varepsilon_i \mid X) && \text{(Independence)} \\
&= \sigma_\varepsilon^2 \sum_i k_i^2 && \text{(Homoscedasticity)} \\
&= \frac{\sigma_\varepsilon^2}{\sum_i (x_i - \bar{x})^2} = \frac{\sigma_\varepsilon^2}{(n-1)s_x^2}
\end{aligned}
$$

- **Key Assumptions:**
  - $\operatorname{var}(\varepsilon_i \mid X) = \sigma_\varepsilon^2$ (constant variance)
  - Independence of errors
- **Normal Case:** If $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$, then: $\hat{b}_1 \mid X \sim \mathcal{N}\left(b_1, \frac{\sigma_\varepsilon^2}{(n-1)s_x^2}\right)$
- **Practical Insights:**
  - Precision increases with sample size (n)
  - More spread in X (larger s_x^2) improves precision
  - Error variance (σ_ε^2) directly affects uncertainty

Summary: Sampling Distribution of $\hat{b}_1$

Key Properties

- Unbiased: $E(\hat{b}_1 \mid X) = b_1$
- Variance: $\text{var}(\hat{b}_1 \mid X) = \frac{\sigma_\varepsilon^2}{(n-1)s_x^2}$

Key Assumptions

- Zero mean errors
- Constant variance
- Independent errors

Practical Implications:

- Larger samples $\rightarrow$ Better precision
- More variable X $\rightarrow$ Better precision
- Noisier data $\rightarrow$ Less precision
- Normal errors $\rightarrow$ Normal sampling distribution

Normal Distribution

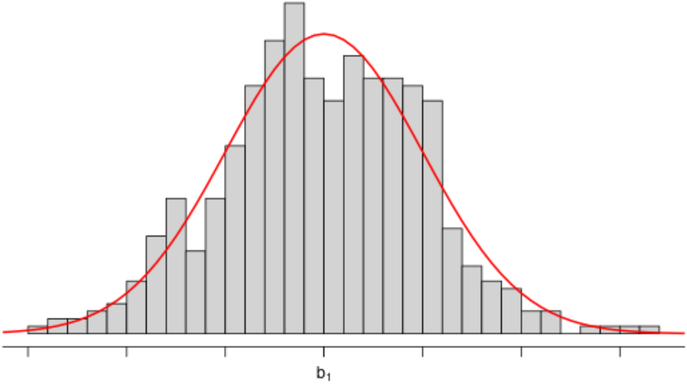


Figure: Distribution of $\hat{b}_1$

Estimating the Variance: The Challenge

Recall: Variance of our slope estimator is

$$\text{var}(\hat{b}_1 \mid X) = \frac{\sigma_\varepsilon^2}{\sum_i (x_i - \bar{x})^2}$$

- $\sum_i (x_i - \bar{x})^2$ is known from our data
- But σ_ε^2 is unknown and must be estimated

Estimating the Variance: The Challenge

Recall: Variance of our slope estimator is

$$\text{var}(\hat{b}_1 \mid X) = \frac{\sigma_\varepsilon^2}{\sum_i (x_i - \bar{x})^2}$$

- $\sum_i (x_i - \bar{x})^2$ is known from our data
- But σ_ε^2 is unknown and must be estimated

Strategy: Use residuals to estimate σ_ε^2

$$\begin{aligned} \text{True errors:} \quad & \varepsilon_i = y_i - (b_0 + b_1 x_i) \\ \text{True variance:} \quad & \sigma_\varepsilon^2 = E(\varepsilon_i^2) \approx \frac{1}{n} \sum_i \varepsilon_i^2 \end{aligned}$$

Estimating the Variance: The Solution

Step 1: Replace true errors with residuals

$$\hat{\varepsilon}_i = y_i - (\hat{b}_0 + \hat{b}_1 x_i)$$

Step 2: Initial estimate using residuals

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n} \sum_i \hat{\varepsilon}_i^2 = \frac{1}{n} RSS(\hat{b}_0, \hat{b}_1)$$

Estimating the Variance: The Solution

Step 1: Replace true errors with residuals

$$\hat{\varepsilon}_i = y_i - (\hat{b}_0 + \hat{b}_1 x_i)$$

Step 2: Initial estimate using residuals

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n}\sum_i \hat{\varepsilon}_i^2 = \frac{1}{n} RSS(\hat{b}_0, \hat{b}_1)$$

Problem: This estimate is biased downward because

$$\frac{1}{n} RSS(\hat{b}_0, \hat{b}_1) \leq \frac{1}{n} RSS(b_0, b_1)$$

Solution: Adjust degrees of freedom

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n-2} RSS(\hat{b}_0, \hat{b}_1)$$

Properties of the Variance Estimator

Key Result:

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n-2} RSS(\hat{b}_0, \hat{b}_1)$$

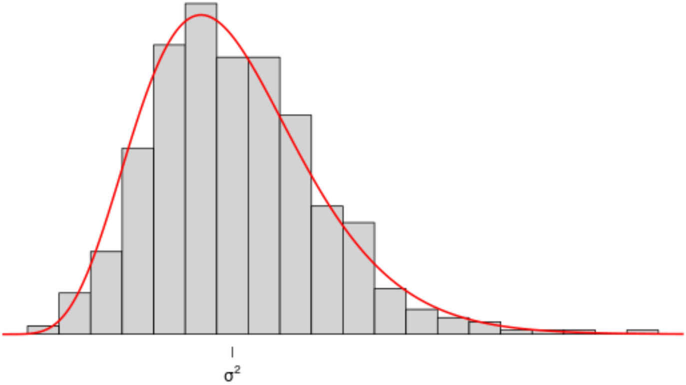
- $\hat{\sigma}_\varepsilon^2$ is a random variable (depends on data)
- It is unbiased: $E(\hat{\sigma}_\varepsilon^2) = \sigma_\varepsilon^2$
- Under normality:

$$\hat{\sigma}_\varepsilon^2 \sim \frac{\sigma_\varepsilon^2}{n-2} \chi^2(n-2)$$

Intuition:

- $n-2$ appears because we estimated two parameters (b_0, b_1)
- Compare to $n-1$ when estimating mean only

Distribution of $\hat{\sigma}_\varepsilon^2$



Multiple Linear Regression

Sampling distribution for MLR

For multiple linear regression, a similar but more complex calculation shows:

$$
\begin{aligned}
E(\hat{\mathbf{b}} \mid X) &= \mathbf{b} \\
\text{var}(\hat{\mathbf{b}} \mid X) &= \begin{bmatrix} \text{var}(\hat{b}_0) & \text{cov}(\hat{b}_0, \hat{b}_1) & \text{cov}(\hat{b}_0, \hat{b}_2) & \dots & \text{cov}(\hat{b}_0, \hat{b}_p) \\ \text{cov}(\hat{b}_0, \hat{b}_1) & \text{var}(\hat{b}_1) & \text{cov}(\hat{b}_1, \hat{b}_2) & \dots & \text{cov}(\hat{b}_1, \hat{b}_p) \\ \dots & & \dots & & \end{bmatrix} \\
&= \sigma_\varepsilon^2 (\mathbf{X}'\mathbf{X})^{-1}
\end{aligned}
$$

- Estimates of coefficients are still unbiased!
- If $\bar{X} = 0$, then $(\mathbf{X}'\mathbf{X})$ is the covariance of $\mathbf{X}$ where

$$(\mathbf{X}'\mathbf{X})_{u,v} = \sum_{i=1}^{n} x_{i,u} x_{i,v}.$$

- Variance decreases as $(\mathbf{X}'\mathbf{X})$ is "larger" i.e., covariates have more variability
- The results hold regardless of the distribution of ε_i. But, if ε_i is normally distributed, then $\hat{\mathbf{b}}$ follows a multivariate normal distribution
- In general, each estimated coefficient is not independent of the other estimated coefficients
- Roughly speaking, dependence between coefficients will depend on how correlated the corresponding covariates are

From Simple to Multiple Linear Regression

Key Results for Multiple Linear Regression:

$$E(\hat{b}_k \mid X) = b_k$$

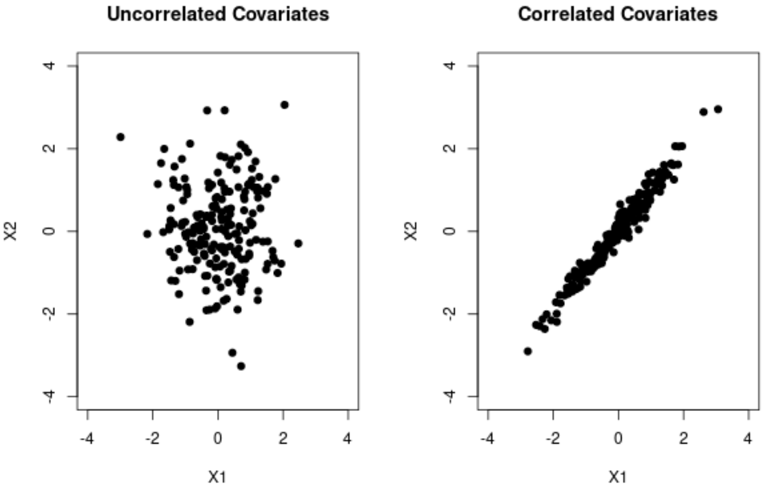
$$\text{var}(b_k \mid X) = \sigma_\varepsilon^2 \left[(\mathbf{X}'\mathbf{X})^{-1}\right]_{kk} \neq \frac{\sigma_\varepsilon^2}{\sum_i (x_{i,k} - \bar{x}_k)^2}$$

- Good news: Each coefficient remains unbiased
- Important change: Variance formula becomes more complex
  - Now depends on all covariates, not just x_k
  - Other variables affect precision of $\hat{b}_k$
- Interpretation of b_k changes: "effect holding other variables constant"

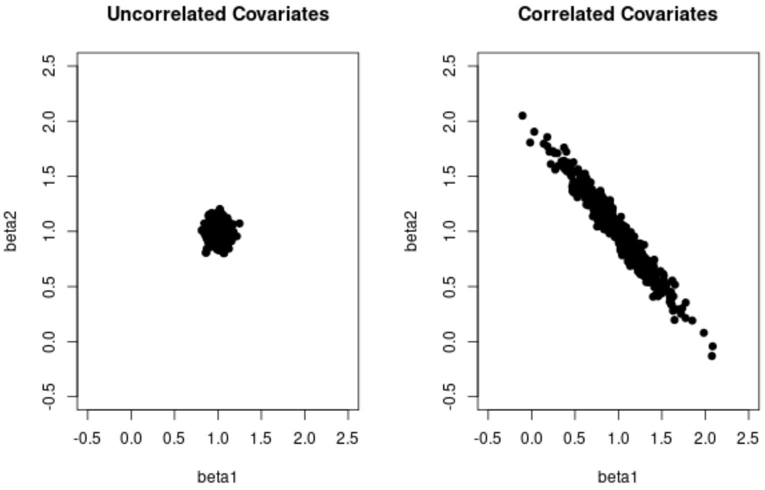
Variance of Estimates: Independent Predictors

We simulate from:

$$Y_i = X_{i,1} + X_{i,2} + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0,1)$$



Variance of estimates



Understanding Collinearity

Definition: High correlation between predictor variables

Extreme Case: Perfect correlation ($\rho = 1$)

- When $X_{i,1} = X_{i,2}$:
  - Cannot separate effects of variables
  - Multiple solutions give identical predictions
- Example: These models are equivalent

$$\begin{aligned} Y_i &= b_0 + b_1 X_{i,1} + b_2 X_{i,2} + \varepsilon_i \\ &= b_0 + (b_1 + c) X_{i,1} + (b_2 - c) X_{i,2} + \varepsilon_i \end{aligned}$$

Practical Impact:

- Estimates become highly sensitive to random errors
- Large changes in coefficients from sample to sample
- Standard errors increase dramatically

Estimating Variance in Multiple Regression

Key Idea: Adjust for model complexity

Since $\hat{\mathbf{b}}$ minimizes RSS:

$$\frac{1}{n} RSS(\hat{\mathbf{b}}) \leq \frac{1}{n} RSS(\mathbf{b})$$

Variance Estimator:

$$\hat{\sigma}_{\varepsilon}^2 = \frac{1}{n - (p + 1)} RSS(\hat{\mathbf{b}})$$

where:

- $p + 1 =$ number of coefficients (including intercept)
- $n - (p + 1) =$ residual degrees of freedom

Estimating Variance in Multiple Regression

Key Idea: Adjust for model complexity

Since $\hat{\mathbf{b}}$ minimizes RSS:

$$\frac{1}{n} RSS(\hat{\mathbf{b}}) \leq \frac{1}{n} RSS(\mathbf{b})$$

Variance Estimator:

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n - (p+1)} RSS(\hat{\mathbf{b}})$$

where:

- $p + 1 =$ number of coefficients (including intercept)
- $n - (p+1) =$ residual degrees of freedom

Properties:

- Unbiased: $E(\hat{\sigma}_\varepsilon^2) = \sigma_\varepsilon^2$
- Under normality: $\hat{\sigma}_\varepsilon^2 \sim \frac{\sigma_\varepsilon^2}{n-p-1} \chi^2(n-p-1)$

Key Takeaways: Multiple Regression

Properties

- Coefficients are unbiased
- Variance depends on:
  - Error variance
  - Predictor spread
  - Predictor correlation

Practical Implications

- Watch for collinearity
- More variables $\rightarrow$ More complexity
- Need larger samples for precise estimation

Design Principles:

- Collect enough data relative to model complexity
- Consider whether highly correlated predictors are both needed
- Balance model complexity against estimation precision