Introductory Regression Notes

Terminology

Target Parameter: The parameter that we are trying to learn about. In the case of linear regression, we are typically interested in the coefficients β.

Estimand: A formula that explains how we would compute the target parameter if we had access to the entire population. In the case of linear regression, the estimand is typically the projection coefficient $\beta = \mathbb{E}[XX']^{-1}\mathbb{E}[XY]$.

Estimator: A formula that maps the sample data to an estimate of the target parameter. You can think of this separately from the particular sample that we have access to. Often, we can go directly from estimand to estimator using the analogy principle. In the case of linear regression, the estimator is typically the OLS estimator $\widehat{\beta} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i X_i'^{-1} \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i Y_i = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$.

Estimate: The result of applying the estimator to a particular data set. In the case of linear regression, the estimate is the numeric value of $\widehat{\beta}$ computed from the data.

Properties of Estimators

Goal: We are interested in learning about some target parameter, but we typically do not have access to the entire population—if we did, we could just compute the target parameter directly.

Sampling Distribution: The distribution of an estimator with respect to a repeated sampling thought experiment where we imagine repeatedly drawing new samples of size n from the underlying population and re-computing the estimator for each new sample. For our particular sample, we essentially get one draw from this sampling distribution.

Bias: The difference between the expected value of an estimator and its actual value, i.e., $\text{Bias}(\widehat{\theta}) = \mathbb{E}[\widehat{\theta}] - \theta$, where the expectation is with respect to the sampling distribution of the estimator. An estimator is said to be *unbiased* if $\text{Bias}(\widehat{\theta}) = 0$.

Sampling Variance: The variance of an estimator with respect to its sampling distribution, i.e., $\text{Var}(\widehat{\theta})$.

In general, we prefer estimators with low (or 0) bias and low sampling variance.

Consistency: In large samples, if $\widehat{\theta}$ is consistent, then it is guaranteed to be close to θ if we have enough data, i.e., $\widehat{\theta} \overset{p}{\to} \theta$.

Asymptotic Normality: In large samples, a centered and scaled version of our estimator will follow a normal distribution, typically, $\sqrt{n}(\hat{\theta} - \theta) \overset{d}{\to} \mathcal{N}(0, V)$. Alternatively, $\hat{\theta} \overset{a}{\sim} \mathcal{N}(\theta, V/n)$ (i.e., $\hat{\theta}$ approximately follows a normal distribution with mean θ and variance V/n).

Details for Linear Regression

Continuous Mapping Theorem:

- For convergence in probability: Let $Z_n \in \mathbb{R}^k$ and $g(u) : \mathbb{R}^k \to \mathbb{R}^q$. If $Z_n \overset{p}{\to} c$ as $n \to \infty$ and $g(u)$ is continuous at c, then $g(Z_n) \overset{p}{\to} g(c)$ as $n \to \infty$.
- For convergence in distribution: If $Z_n \overset{d}{\to} Z$ as $n \to \infty$ and $g : \mathbb{R}^k \to \mathbb{R}^q$ has the set of discontinuity points D_g such that $\mathrm{P}(Z \in D_g) = 0$, then $g(Z_n) \overset{d}{\to} g(Z)$ as $n \to \infty$.

These continuous mapping theorems say that continuous functions are limit preserving. Notice that the conditions for the convergence in probability version of the CMT are weaker (they only require g to be continuous at the particular point c) than for the convergence in distribution version (which essentially requires g to be continuous everywhere). The qualification about the set of discontinuity points is a technical one, but comes up enough cases that it is worth including this technical condition.

The asymptotic theory for least squares applies both to linear projection model and to the linear CEF model. Therefore, in this section, we only use the weaker assumptions of the linear projection model. That is, we use the following assumptions throughout this section

Assumption 7.1

1. The variables $\{(Y_i, X_i)\}_{i=1}^n$ are iid
2. $\mathbb{E}[Y^2] < \infty$
3. $\mathbb{E}||X||^2 < \infty$
4. $\mathbb{E}[XX']$ is positive definite

Consistency of Least Squares Estimator

H: 7.2

Step 1: Weak Law of Large Numbers. Recall that

$$\hat{\beta} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} X_i Y_i \tag{1}$$

Next, notice that

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} X_i X_i' &\xrightarrow{p} \mathbb{E}[XX'] \\ \frac{1}{n}\sum_{i=1}^{n} X_i Y_i &\xrightarrow{p} \mathbb{E}[XY] \end{aligned}$$

which holds by the weak law of large numbers (which requires the iid assumption and that $\mathbb{E}[XX'] < \infty$ and $\mathbb{E}[XY] < \infty$, both of which hold by Assumption 7.1)

Step 2: Continuous Mapping Theorem. Next, notice that, we can write

$$\hat{\beta} = g(\hat{\mathbb{E}}[XX'], \hat{\mathbb{E}}[XY])$$

where $g(\mathbf{A}, b) = \mathbf{A}^{-1}b$. This is a continuous function of $\mathbf{A}$ and b at all values of the arguments such that $\mathbf{A}^{-1}$ exists. Assumption 7.1 includes that $\mathbb{E}[XX']$ is positive definite which implies that $\mathbb{E}[XX']^{-1}$ exists. Thus, $g(\mathbf{A}, b)$ is continuous at $\mathbf{A} = \mathbb{E}[XX']$ and we can apply the "convergence in probability" version of the CMT; that is,

$$\begin{aligned} \hat{\beta} &\xrightarrow{p} g(\mathbb{E}[XX'], \mathbb{E}XY) \\ &= \mathbb{E}[XX']^{-1}\mathbb{E}[XY] = \beta \end{aligned}$$

Asymptotic Normality

H: 7.3

For this section, we strengthen Assumption 7.1.

Assumption 7.2

In addition to Assumption 7.1

1. $\mathbb{E}[Y^4] < \infty$
2. $\mathbb{E}||X||^4 < \infty$

Next, we will establish the limiting distribution of $\hat{\beta}$. Plugging $Y_i = X_i'\beta + e_i$ into Equation 1 implies that

$$
\begin{aligned}
\hat{\beta} &= \left(\frac{1}{n} \sum_{i=1}^{n} X_i X_i' \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} \left(X_i (X_i'\beta + e_i) \right) \\
&= \beta + \left(\frac{1}{n} \sum_{i=1}^{n} X_i X_i' \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} X_i e_i
\end{aligned}
$$

Multiplying by $\sqrt{n}$ and re-arranging implies that

$$
\sqrt{n}\left(\hat{\beta} - \beta\right) = \left(\frac{1}{n} \sum_{i=1}^{n} X_i X_i' \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i e_i \tag{2}
$$

Step 1: Central Limit Theorem. First, notice that

$$
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i e_i \xrightarrow{d} \mathcal{N}(0, \mathbf{\Omega})
$$

where $\mathbf{\Omega} = \mathbb{E}[Xe(Xe)'] = \mathbb{E}[XX'e^2]$.

Let's explain carefully why the central limit theorem applies here. First, we have that (Y_i, X_i) are iid, which implies that any function of (Y_i, X_i) is also iid (and this includes $e_i = Y_i - X_i'\beta$ and $X_i e_i$). Also, notice that $\mathbb{E}[Xe] = 0$ so that the inside term of the summation above has mean 0. That $\mathbf{\Omega} = \text{Var}(Xe) = \mathbb{E}[XX'e^2]$ is a direct consequence of applying the central limit theorem.

Step 2: Continuous Mapping Theorem

Combining this with Equation 2, we have that

$$
\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathbb{E}[XX']^{-1}\, \mathcal{N}(0, \mathbf{\Omega}) = \mathcal{N}(0, \mathbf{V}_\beta)
$$

where $\mathbf{V}_\beta = \mathbb{E}[XX']^{-1}\mathbf{\Omega}\mathbb{E}[XX']^{-1}$ and which holds by the continuous mapping theorem.

$\mathbf{V}_\beta$ is called the **asymptotic variance matrix** of $\hat{\beta}$. $\mathbb{E}[XX']^{-1}\mathbf{\Omega}\mathbb{E}[XX']^{-1}$ is called a "sandwich form". It is called this because $\mathbf{\Omega}$ is sandwiched by $\mathbb{E}[XX']^{-1}$ (sometimes $\mathbf{\Omega}$ is called the "meat" and $\mathbb{E}[XX']^{-1}$ is called the "bread"). Many asymptotic variance matrices have a similar form.

The previous result is the basis for hypothesis testing/inference, constructing confidence intervals, etc. To operationalize it, though, we need to construct an estimator of $\mathbf{V}_\beta$. The natural estimator is

$$
\hat{\mathbf{V}}_\beta = \left(\frac{1}{n} \sum_{i=1}^{n} X_i X_i' \right)^{-1} \hat{\mathbf{\Omega}} \left(\frac{1}{n} \sum_{i=1}^{n} X_i X_i' \right)^{-1}
$$

where $\hat{\mathbf{\Omega}}$ is an estimate of $\mathbf{\Omega}$ given by

$$\hat{\mathbf{\Omega}} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i' \hat{e}_i^2$$

Inference

H: 7.11-7.13, 7.16, 9.7, 9.9

Inference and hypothesis testing were covered in detail in 8070. This section provides a brief review along with (brief) explanations in the context of regression.

Next, let us return to our results on asymptotic normality of $\hat{\beta}$ in order to see how this is useful for hypothesis testing.

We define the **standard error** of the $\hat{\beta}_j$ (the jth element of $\hat{\beta}$) as

$$\text{se}(\hat{\beta}_j) = \frac{\sqrt{\hat{\mathbf{V}}_{\beta,jj}}}{\sqrt{n}}$$

where $\hat{\mathbf{V}}_{\beta,jj}$ is the (j,j) element of $\hat{\mathbf{V}}_{\beta}$. As a side-comment, I define this slightly differently from the book. I use $\hat{\mathbf{V}}_{\beta}$ (the asymptotic variance matrix) rather than $\mathbf{V}_{\hat{\beta}} = \text{var}(\hat{\beta}|\mathbf{X})$. These are different from each other by a factor of n; in particular, $\hat{\mathbf{V}}_{\beta}$ does not go to 0 as $n \to \infty$.

It is common in applications in economics to report $\hat{\beta}$ along with standard errors for each estimated parameter.

t-statistic

The t-statistic for testing the null hypothesis $\text{H}_0 : \beta_j = 0$ is given by

$$t = \frac{\hat{\beta}_j - 0}{\text{se}(\hat{\beta}_j)} = \frac{\sqrt{n}(\hat{\beta}_j - 0)}{\sqrt{\hat{\mathbf{V}}_{\beta,jj}}}$$

Then, we most often follow the decision rule to reject the null hypothesis if $|t| > c_{1-\alpha}$ where c is a critical value that depends on the significance level α (e.g., 5%). The reason this works is that, under H_0,

$$\begin{aligned} t &= \frac{\sqrt{n}(\hat{\beta}_j - 0)}{\sqrt{\hat{\mathbf{V}}_{\beta,jj}}} \\ &= \frac{\sqrt{n}(\hat{\beta}_j - \beta_j)}{\sqrt{\hat{\mathbf{V}}_{\beta,jj}}} \\ &= \frac{\sqrt{n}(\hat{\beta}_j - \beta_j)}{\sqrt{\mathbf{V}}_{\beta,jj}} + o_p(1) \\ &\xrightarrow{d} \mathcal{N}(0,1) \end{aligned}$$

where the second equality holds under H_0 and the third equality holds by the consistency of $\hat{\mathbf{V}}_\beta$ for $\mathbf{V}_\beta$, and the last equality holds by the asymptotic normality of $\hat{\beta}$. This implies that, under H_0, t should behave like a draw from a standard normal distribution.

On the other hand, under the alternative hypothesis $\mathrm{H}_1 : \beta_j \neq 0$, $\hat{\beta}_j - 0 \xrightarrow{p} \beta_j \neq 0$, which implies that $\sqrt{n}\hat{\beta}_j$ diverges, and, hence, t also diverges. This implies that, under H_1, t should be large in absolute value with high probability.

The difference in the behavior of t under H_0 and H_1 is what allows us to use the decision rule of rejecting H_0 if $|t| > c_{1-\alpha}$.

Confidence Interval The set of values of β_j that are “compatible” with the observed data. A 95% confidence interval can be constructed as $\hat{\beta}_j \pm 1.96 \times \mathrm{se}(\hat{\beta}_j)$.

p-value

This is the probability of getting an estimate as “extreme’ ’ as the one we got if the null hypothesis were true. In the case of testing $H_0 : \beta_j = 0$, the p-value is given by $2(1 - \Phi(|t|))$ where Φ is the CDF of a standard normal distribution and t is the t-statistic for testing $H_0 : \beta_j = 0$.

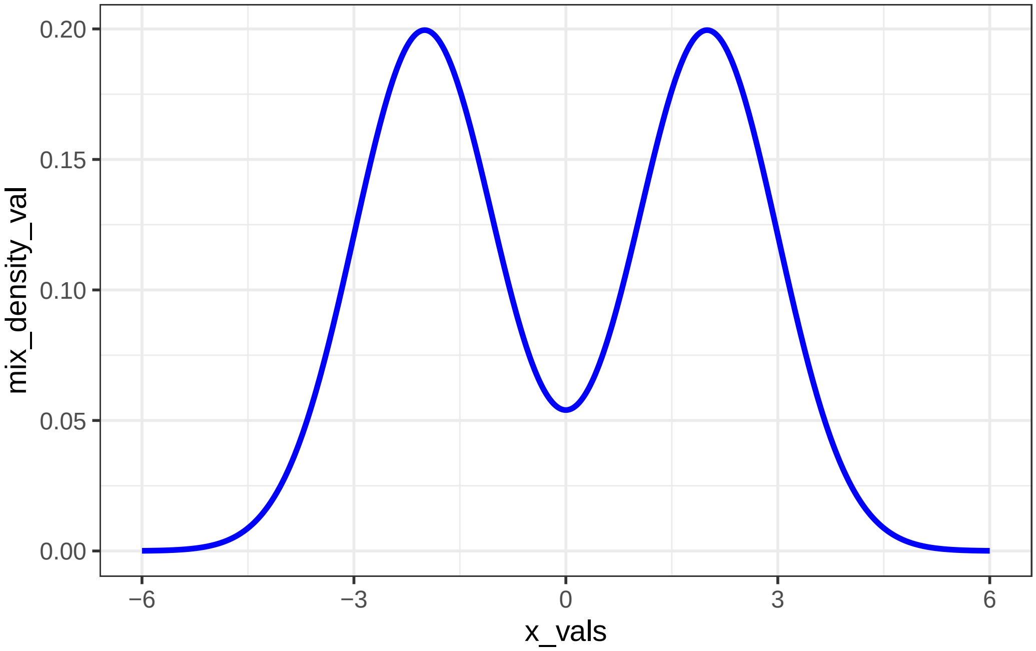
Monte Carlo Simulations

These are Monte Carlo simulations used for understanding properties of the sampling distribution of $\hat{\beta}$ under different scenarios. You can change the values of the simulation parameters (e.g., `n`, `b1`, `H0`).

```r
library(estimatr)
library(mixtools)
library(ggplot2)

# lm_robust
n <- 100
b0 <- 0
b1 <- 1
mix_probs <- c(0.5, 0.5)
mix_means <- c(-2, 2)

# plot mixture distribution
x_vals <- seq(-6, 6, length.out = 1000)
mix_density_val <- mix_probs[1] * dnorm(x_vals, mean = mix_means[1]) +
    mix_probs[2] * dnorm(x_vals, mean = mix_means[2])
ggplot(
    data = data.frame(x_vals, mix_density_val),
    aes(x = x_vals, y = mix_density_val)
) +
    geom_line(color = "blue", linewidth = 1) +
    theme_bw()
```



```
# consistency - try different values of n
X <- rnorm(n)
e <- mixtools::rnormmix(n,
    lambda = mix_probs,
    mu = mix_means
)
Y <- b0 + b1 * X + e

reg <- lm_robust(Y ~ X)
summary(reg)
```

```

Call:
lm_robust(formula = Y ~ X)

Standard error type:  HC2

Coefficients:
            Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper DF
```

```
(Intercept)  -0.1722     0.2165 -0.7953 4.284e-01  -0.6017    0.2574 98
X             1.1403     0.2147  5.3118 6.793e-07   0.7143    1.5663 98

Multiple R-squared:  0.1924 ,   Adjusted R-squared:  0.1842
F-statistic: 28.22 on 1 and 98 DF,  p-value: 6.793e-07
```

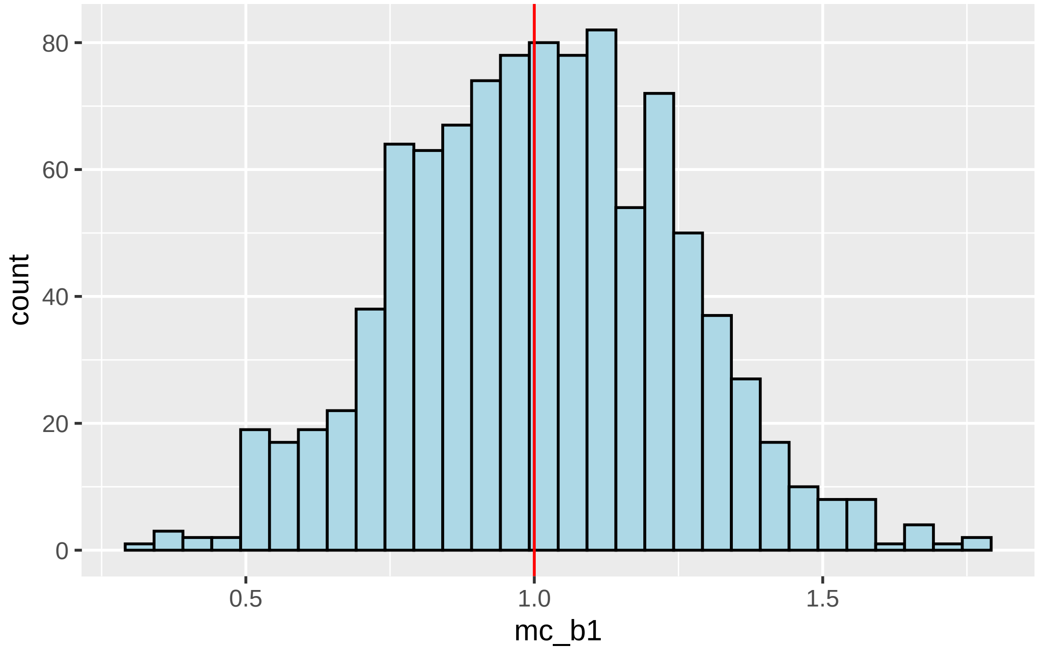
```r
# monte carlo simulations

# create some variables to hold results
n <- 100
nsims <- 1000
b1 <- 1
H0 <- 0
mc_b1 <- c()
mc_var <- c()
mc_se <- c()
mc_tstat <- c()
mc_ciL <- c()
mc_ciU <- c()

for (i in 1:nsims) {
    # generate data
    X <- rnorm(n)
    e <- mixtools::rnormmix(n,
        lambda = mix_probs,
        mu = mix_means
    )
    Y <- b0 + b1 * X + e

    # run regression
    reg <- lm_robust(Y ~ X)
    mc_b1[i] <- coef(reg)[2]
    mc_var[i] <- vcov(reg)[2, 2]
    mc_se[i] <- sqrt(mc_var[i])
    mc_tstat[i] <- (mc_b1[i] - H0) / mc_se[i]
    mc_ciL[i] <- mc_b1[i] - 1.96 * mc_se[i]
    mc_ciU[i] <- mc_b1[i] + 1.96 * mc_se[i]
}
```

```r
# plot sampling distribution
ggplot(
    data = data.frame(mc_b1),
    aes(x = mc_b1)
) +
    geom_histogram(
        bins = 30, color = "black",
        fill = "lightblue"
    ) +
    geom_vline(xintercept = b1, color = "red")
```



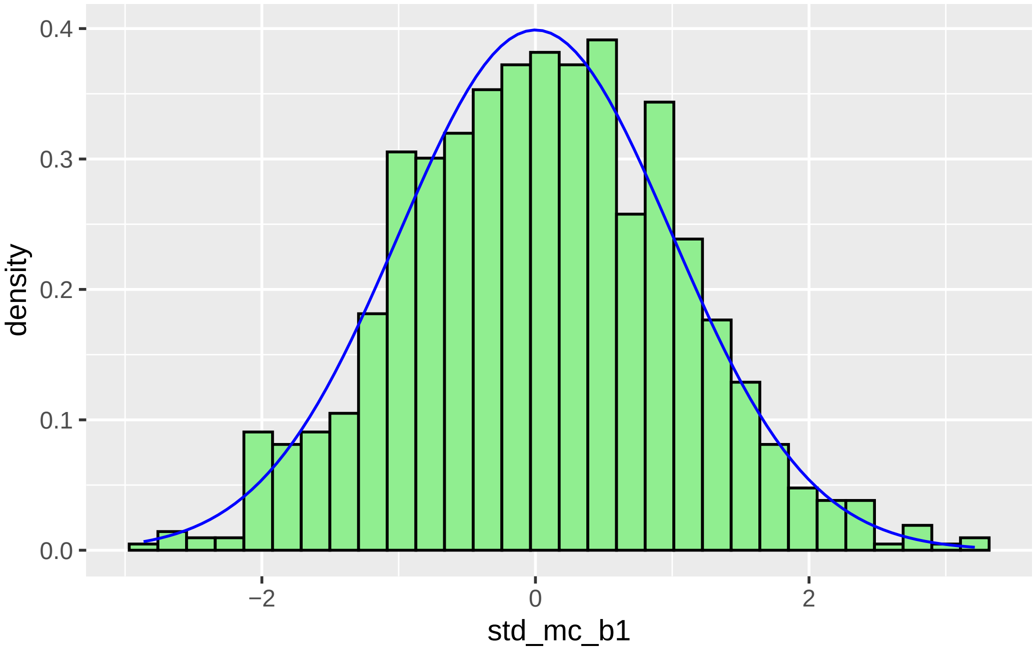
```r
# compute bias
bias <- mean(mc_b1) - b1
bias
```

```
[1] 0.01028695
```

```r
# compute sampling variance
sampling_variance <- var(mc_b1)
sampling_variance
```

```
[1] 0.05702511
```

```
# asymptotic normality
std_mc_b1 <- (mc_b1 - b1) / sqrt(sampling_variance)
ggplot(
    data = data.frame(std_mc_b1),
    aes(x = std_mc_b1)
) +
    geom_histogram(aes(y = after_stat(density)),
        bins = 30, color = "black",
        fill = "lightgreen"
    ) +
    stat_function(fun = dnorm, color = "blue")
```



```
# rejection rates
rej <- mean(abs(mc_tstat) > 1.96)
rej
```

```
[1] 0.995
```

```
# coverage probability
cover <- (mc_ciL <= b1) & (mc_ciU >= b1)
mean(cover)
```

```
[1] 0.936
```