Preliminaries

Brad Jones\textsuperscript{1}

\textsuperscript{1}Department of Political Science
University of California, Davis

January 5, 2010
Today: Preliminaries
Today: Preliminary Concepts
Most of which should seem familiar.
If not, review 211 texts.
Setting the Stage: Basic Statistical Concepts

- Properties of probability

Classic definition implies long-run relative frequency of some event $A$.

Bayesians tell us, this is not always a good definition (they're right).

However, let's walk before we run.

$\Pr(A)$ is a real-valued function defined on a sample space.

Important properties:

1. $0 \leq \Pr(A) \leq 1 \quad \forall A$

2. $\Pr(A + B + C) = 1$

3. $\Pr(A + B + C) = \Pr(A) + \Pr(B) + \Pr(C)$

(2) implies exhaustive events; (3) implies mutual exclusiveness.
Setting the Stage: Basic Statistical Concepts

- Properties of probability
- Classic definition implies long-run relative frequency of some event $A$. 

$\Pr(A) \geq 0$ for all events $A$.
Setting the Stage: Basic Statistical Concepts

- Properties of probability
- Classic definition implies long-run relative frequency of some event $A$.
- Bayesians tell us, this is not always a good definition (they’re right).
Properties of probability

Classic definition implies long-run relative frequency of some event $A$.

Bayesians tell us, this is not always a good definition (they’re right).

However, let's walk before we run.
Setting the Stage: Basic Statistical Concepts

- Properties of probability
- Classic definition implies long-run relative frequency of some event $A$.
- Bayesians tell us, this is not always a good definition (they’re right).
- However, let's walk before we run.
- $Pr(A)$ is real-valued function defined on a sample space. Important properties:

$$0 \leq Pr(A) \leq 1 \quad \forall \ A$$

$$Pr(A + B + C) = 1$$

$$Pr(A + B + C) = Pr(A) + Pr(B) + Pr(C)$$
Setting the Stage: Basic Statistical Concepts

- Properties of probability
- Classic definition implies long-run relative frequency of some event $A$.
- Bayesians tell us, this is not always a good definition (they’re right).
- However, let’s walk before we run.
- $\Pr(A)$ is real-valued function defined on a sample space. Important properties:

$$0 \leq \Pr(A) \leq 1 \quad \forall \ A$$  \hspace{1cm} (1)

$$\Pr(A + B + C) = 1$$  \hspace{1cm} (2)

$$\Pr(A + B + C) = \Pr(A) + \Pr(B) + \Pr(C)$$  \hspace{1cm} (3)

- (2) implies exhaustive events; (3) implies mutual exclusiveness.
Basic Concepts

Random Variable: “a real-valued function defined on a sample space.”
Random Variable: “a real-valued function defined on a sample space.”

It’s an “observable” (with two flavors).
Basic Concepts

- Random Variable: “a real-valued function defined on a sample space.”
- It’s an “observable” (with two flavors).
- Discrete and Continuous
Basic Concepts

- Random Variable: “a real-valued function defined on a sample space.”
- It’s an “observable” (with two flavors).
- Discrete and Continuous
- Probability Density Functions
Basic Concepts

- Random Variable: “a real-valued function defined on a sample space.”
- It’s an “observable” (with two flavors).
- Discrete and Continuous
- Probability Density Functions
- The PDF assigns probabilities to outcomes.
Discrete Random Variable

PMF for a D.R.V.:

\[
f(x) = \Pr(X = x_i) \quad \forall \quad i = 1, 2, \ldots, n \quad (4)
\]

\[
= 0 \quad \forall \quad x \neq x_i. \quad (5)
\]
Discrete Random Variable

- PMF for a D.R.V.:

\[ f(x) = \Pr(X = x_i) \quad \forall \quad i = 1, 2, \ldots, n \]  
\[ = 0 \quad \forall \quad x \neq x_i. \]  

- In words: the probability that \( X \) is equal to some specific (discrete) value.

```python
# Die Rolls in R:
d1 <- sample(1:6, 100000, prob=rep(1/6, 6), replace=TRUE)
d2 <- sample(1:6, 100000, prob=rep(1/6, 6), replace=TRUE)
die.roll <- d1 + d2
hist(die.roll, breaks= seq(0, 12, by=.1), prob=FALSE, las=1, col="red", main="Distribution of 100,000 die rolls", xlab = "Value of Roll")
```
Discrete Random Variable

- PMF for a D.R.V.:

\[
\begin{align*}
  f(x) &= \Pr(X = x_i) \quad \forall \ i = 1, 2, \ldots, n \quad (4) \\
  &= 0 \quad \forall \ x \neq x_i. \quad (5)
\end{align*}
\]

- In words: the probability that \( X \) is equal to some specific (discrete) value.

- Die Rolls in R:

```r
d1 <- sample(1:6, 100000, prob=rep(1/6, 6), replace=TRUE)
d2 <- sample(1:6, 100000, prob=rep(1/6, 6), replace=TRUE)
die.roll <- d1 + d2
hist(die.roll, breaks= seq(0, 12, by=.1), prob=FALSE, las=1, col="red", main="Distribution of 100,000 die rolls", xlab = "Value of Roll")
```
Artwork

PMF for 100,000 die rolls

Frequency

Value of Roll
Discrete Random Variable

PMF in R:

```R
freq <- cbind(table(die.roll))
prval <- freq / 100000
roll <- cbind(2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12)
plot(roll, prval, col = "red", main = "Probability Mass Function",
     xlab = "Value of Roll", ylab = "Probability")
```
Artwork

Probability Mass Function

Value of Roll

Probability
Continuous Random Variable

- Continuous RVs have “density functions.”

\[ f(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}} \]
Continuous Random Variable

- Continuous RVs have “density functions.”
- The density is kind of like a “smoothed out” histogram.

The probability of any specific realization of $X$ is assumed to be 0. (Why?)
∴ we must integrate to define probability (within an infinitesimally small differentiable area).

$f(x)$ in discrete case is easy to define; in continuous case, $f(x)$ may take on a variety of forms.

The PDF, $f(x)$, for the standard normal:

$$f(x) = e^{-x^2/2} \sqrt{2\pi}$$

We use this distribution all the time: $z$-scores, for example.
Continuous Random Variable

- Continuous RVs have “density functions.”
- The density is kind of like a “smoothed out” histogram.
- The probability of any specific realization of $X$ is assumed to be 0. (Why?)
Continuous Random Variable

- Continuous RVs have “density functions.”
- The density is kind of like a “smoothed out” histogram.
- The probability of any specific realization of $X$ is assumed to be 0. (Why?)
- ∴ we must integrate to define probability (within an infinitesimally small differentiable area).
Continuous Random Variable

- Continuous RVs have “density functions.”
- The density is kind of like a “smoothed out” histogram.
- The probability of any specific realization of $X$ is assumed to be 0. (Why?)
- Therefore, we must integrate to define probability (within an infinitesimally small differentiable area).
- $f(x)$ in discrete case is easy to define; in continuous case, $f(x)$ may take on a variety of forms.

The PDF, $f(x)$, for the standard normal:

$$f(x) = e^{-x^2/2} \sqrt{2\pi}$$

We use this distribution all the time: $z$-scores, for example.
Continuous Random Variable

- Continuous RVs have “density functions.”
- The density is kind of like a “smoothed out” histogram.
- The probability of any specific realization of $X$ is assumed to be 0. (Why?)
- \[ \therefore \text{we must integrate to define probability (within an infinitesimally small differentiable area).} \]
- \[ f(x) \text{ in discrete case is easy to define; in continuous case, } f(x) \text{ may take on a variety of forms.} \]
- The PDF, $f(x)$, for the standard normal:

\[
f(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}} \quad (6)
\]
Continuous Random Variable

- Continuous RVs have “density functions.”
- The density is kind of like a “smoothed out” histogram.
- The probability of any specific realization of $X$ is assumed to be 0. (Why?)
- \[ \therefore \text{ we must integrate to define probability (within an infinitesimally small differentiable area).} \]
- $f(x)$ in discrete case is easy to define; in continuous case, $f(x)$ may take on a variety of forms.
- The PDF, $f(x)$, for the standard normal:
  \[
  f(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}
  \] (6)
- We use this distribution all the time: z-scores, for example.
The cumulative distribution function obtains probabilities:

\[ F(x) = \Pr(X \leq x) = \int_{x_{\text{min}}}^{x} f(x) \, dx \quad (7) \]

Here, \( f(x) \) is the PDF; \( F(x) \) is the CDF.

In a sense, the PDF is going to give us the "height" and the CDF gives us the area.

Note that it must be the case all area under the curve must integrate to 1:

\[ F(x) = \Pr(-\infty \leq X \leq \infty) = \int_{-\infty}^{\infty} f(x) \, dx = 1 \quad (8) \]

Also important:

\[ F(b) - F(a) = \Pr(a \leq X \leq b) = \int_{a}^{b} f(x) \, dx \]

Remember this with ordinal logits and probits when you get to 213.
Continuous Random Variable

The cumulative distribution function obtains probabilities:

\[ F(x) = \Pr(X \leq x) = \int_{x_{\text{min}}}^{x} f(x) \, d(x) \]  \hspace{1cm} (7)

Here, \( f(x) \) is the PDF; \( F(x) \) is the CDF.
Continuous Random Variable

- The cumulative distribution function obtains probabilities:

\[ F(x) = \Pr(X \leq x) = \int_{x_{\text{min}}}^{x} f(x) \, d(x) \]  

(7)

- Here, \( f(x) \) is the PDF; \( F(x) \) is the CDF.

- In a sense, the PDF is going to give us the “height” and the CDF gives us the area.
The cumulative distribution function obtains probabilities:

\[ F(x) = \Pr(X \leq x) = \int_{x_{\text{min}}}^{x} f(x) \, dx \]  

Here, \( f(x) \) is the PDF; \( F(x) \) is the CDF.

In a sense, the PDF is going to give us the “height” and the CDF gives us the area.

Note that it must be the case all area under the curve must integrate to 1:

\[ F(x) = \Pr(-\infty \leq X \leq \infty) = \int_{-\infty}^{\infty} f(x) \, dx = 1 \]
Continuous Random Variable

The cumulative distribution function obtains probabilities:

\[ F(x) = \Pr(X \leq x) = \int_{x_{\min}}^{x} f(x) \, dx \]  

Here, \( f(x) \) is the PDF; \( F(x) \) is the CDF.

In a sense, the PDF is going to give us the “height” and the CDF gives us the area.

Note that it must be the case all area under the curve must integrate to 1:

\[ F(x) = \Pr(-\infty \leq X \leq \infty) = \int_{-\infty}^{\infty} f(x) \, dx = 1 \]  

Also important: \( F(b) - F(a) = \Pr(a \leq X \leq b) = \int_{a}^{b} f(x) \, dx \)

Remember this with ordinal logits and probits when you get to 213.
Artwork

**Normal density**

\[ \text{dnorm}(x, \log = \text{FALSE}) \]

**Normal Cumulative**

\[ \text{pnorm}(x, \log = \text{FALSE}) \]
The classic linear model assumes $y$ is continuous.
Continuous Random Variable

- The classic linear model assumes $y$ is continuous.
- It may not be (often, will not be).
Continuous Random Variable

- The classic linear model assumes $y$ is continuous.
- It may not be (often, will not be).
- At “what point” regression fails us is a concern of this class and also 213.
The classic linear model assumes $y$ is continuous.

- It may not be (often, will not be).
- At “what point” regression fails us is a concern of this class and also 213.
- With a binary dependent variable, or categorical choice data, regression will fail us in certain kinds of ways (to certain degrees).
Continuous Random Variable

- The classic linear model assumes $y$ is continuous.
- It may not be (often, will not be).
- At “what point” regression fails us is a concern of this class and also 213.
- With a binary dependent variable, or categorical choice data, regression \textit{will} fail us in certain kinds of ways (to certain degrees).
- But before we get to that, let’s go on with a few more preliminaries.
Suppose we have a random variable $\Sigma$ defined on a sample space $\omega$ with probabilities $P$. The expected value

$$E(X) = \int_{\omega} XdP$$

(9)
Characteristics of Distributions

- Suppose we have a random variable $\Sigma$ defined on a sample space $\omega$ with probabilities $P$. The expected value

$$E(X) = \int_{\omega} XdP \quad (9)$$

- For a discrete random variable:

$$E(X) = \sum_{i=1}^{n} x_i p(x_i) \quad (10)$$

where $p(x)$ is the PMF.
Characteristics of Distributions

- Suppose we have a random variable $\Sigma$ defined on a sample space $\omega$ with probabilities $P$. The expected value

$$E(X) = \int_\omega XdP \quad (9)$$

- For a discrete random variable:

$$E(X) = \sum_{i=1}^{n} x_i p(x_i) \quad (10)$$

where $p(x)$ is the PMF.

- For a continuous random variable:

$$E(X) = \int_{-\infty}^{\infty} xf(x)d(x) \quad (11)$$

where $f(x)$ is the PDF.
Expectation

Let $a$ and $b$ denote constants.
Let $a$ and $b$ denote constants.

Then:

\begin{align*}
E(a) &= a \\
E(aX + b) &= E(aX) + E(b) \\
&= aE(x) + b \\
&= a\mu + b \quad (12)
\end{align*}
Moments about the Mean

- Often interest centers on moments about the mean, $\mu$:

$$E(x - \mu)^r = \int_{-\infty}^{\infty} (x - \mu)^r f(x) d(x)$$ (13)

where $r$ is the $r$th moment.
Moments about the Mean

- Often interest centers on moments about the mean, \( \mu \):

\[
E(x - \mu)^r = \int_{-\infty}^{\infty} (x - \mu)^r f(x) d(x) \tag{13}
\]

where \( r \) is the \( r \)th moment.

- Zero Moment: \( E(x - \mu)^0 = 1 \)
Moments about the Mean

Often interest centers on moments about the mean, \( \mu \):

\[
E(x - \mu_x)^r = \int_{-\infty}^{\infty} (x - \mu)^r f(x)dx
\]

(13)

where \( r \) is the \( r \)th moment.

- Zero Moment: \( E(x - \mu_x)^0 = 1 \)
- First Moment:
  \[
  E(x - \mu_x)^1 = E(x - \mu) = E(x) - \mu = \mu - mu = 0 \text{ (why?)}
  \]
Moments about the Mean

Often interest centers on moments about the mean, \( \mu \):

\[
E(x - \mu_x)^r = \int_{-\infty}^{\infty} (x - \mu)^r f(x) \, dx
\]  

where \( r \) is the \( r \)th moment.

- **Zero Moment:** \( E(x - \mu_x)^0 = 1 \)
- **First Moment:**
  \[
  E(x - \mu_x)^1 = E(x - \mu) = E(x) - \mu = \mu - \mu = 0 \quad \text{(why?)}
  \]
- **Second Moment:**
  \[
  E(x - \mu_x)^2 = E[x - E(x)]^2 = \sigma^2 \quad \text{(the variance)}
  \]

Third moment is skewness (deviation from symmetry) and the fourth moment is kurtosis (peakedness).
Moments about the Mean

- Often interest centers on moments about the mean, \( \mu \):

\[
E(x - \mu)^r = \int_{-\infty}^{\infty} (x - \mu)^r f(x) d(x)
\]  

(13)

where \( r \) is the \( r \)th moment.

- Zero Moment: \( E(x - \mu_x)^0 = 1 \)

- First Moment:
  \[
  E(x - \mu_x)^1 = E(x - \mu) = E(x) - \mu = \mu - mu = 0 \text{ (why?)}
  \]

- Second Moment: \( E(x - \mu_x)^2 = E[x - E(x)]^2 = \sigma^2 \) (the variance)

- Third moment is skewness (deviation from symmetry) and the fourth moment is kurtosis (peakedness).
Variance result:

\[
\text{var}(x) = E(x - \mu)^2 \\
= E(x^2 - 2x\mu + \mu^2) \\
= E(x^2) - E(2x\mu) + E(\mu^2) \\
= E(x^2) - 2\mu E(x) + \mu^2 \\
= E(x^2) - 2\mu^2 + \mu^2 \\
= E(x^2) - \mu^2.
\]
Consider two random variables, $x$, $y$. 
Consider two random variables, \( x, y \).
The marginal distribution of \( x \) is \( f(x) \) and of \( y \) is \( f(y) \).
Covariance

- Consider two random variables, $x$, $y$.
- The marginal distribution of $x$ is $f(x)$ and of $y$ is $f(y)$.
- The conditional distributions are:

$$f(x \mid y) = \frac{f(x, y)}{f(y)} \quad (15)$$

for $x$ and

$$f(y \mid x) = \frac{f(x, y)}{f(x)} \quad (16)$$
Covariance

- Consider two random variables, $x$, $y$.
- The marginal distribution of $x$ is $f(x)$ and of $y$ is $f(y)$.
- The conditional distributions are:
  \[ f(x \mid y) = \frac{f(x, y)}{f(y)} \]  
  for $x$ and
  \[ f(y \mid x) = \frac{f(x, y)}{f(x)} \]  
  (15)
  (16)
- Joint distributions and independence.
Covariance

What if \( f(x \mid y) = f(x) \) (or \( f(y \mid x) = f(y) \)?)

- \( x \) and \( y \) are independent \( \forall x, y \).
- The joint PDF is: \( f(x, y) = f(x)f(y) \) (17)
- Or the JPDF = MPFD \( x \) (MPDF \( y \))
- Covariance:
  \[
  \text{cov}(x, y) = \sigma_{xy} = s_{xy} = E[(x - E(x))(y - E(y))] \quad (18)
  \]
- If \( x, y \) are independent, \( \text{cov} = 0 \).
Covariance

- What if \( f(x | y) = f(x) \) (or \( f(y | x) = f(y) \))? 
- \( x \) and \( y \) are independent \( \forall x, y \).
what if \( f(x \mid y) = f(x) \) (or \( f(y \mid x) = f(y) \))? 

\( x \) and \( y \) are independent \( \forall x, y \).

The joint PDF is:

\[
f(x, y) = f(x)f(y) \tag{17}
\]
What if \( f(x \mid y) = f(x) \) (or \( f(y \mid x) = f(y) \))?

\( x \) and \( y \) are independent \( \forall x, y \).

The joint PDF is:

\[
  f(x, y) = f(x)f(y) \quad (17)
\]

Or the \( JPDF = MPFD_x(MPDF_y) \)
Covariance

- What if \( f(x \mid y) = f(x) \) (or \( f(y \mid x) = f(y) \))? 
- \( x \) and \( y \) are independent \( \forall x, y \).
- The joint PDF is:
  \[
  f(x, y) = f(x)f(y)
  \]  \hspace{1cm} (17)

- Or the \( JPDF = MPFD_x(MPDF_y) \)
- Covariance:
  \[
  \text{cov}(x, y) = \sigma_{xy} = s_{xy} = E[x - E(x)][y - E(y)]
  \]  \hspace{1cm} (18)
Covariance

- What if $f(x \mid y) = f(x)$ (or $f(y \mid x) = f(y)$)?
- $x$ and $y$ are independent $\forall x, y$.
- The joint PDF is:
  
  $$f(x, y) = f(x)f(y)$$  \hspace{1cm} (17)

- Or the $JPDF = MPFD_x(MPDF_y)$
- Covariance:
  
  $$cov(x, y) = \sigma_{xy} = s_{xy} = E[x - E(x)][y - E(y)]$$  \hspace{1cm} (18)

- If $x, y$ are independent, $cov=0$. 
Covariance and Correlation

- Zero covariance does not imply statistical independence however!
Zero covariance does not imply statistical independence however!

The correlation coefficient:

\[ r_{xy} = \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x)\text{var}(y)}} \]  

(19)
Covariance and Correlation

- Zero covariance does not imply statistical independence however!

- The correlation coefficient:

  \[ r_{xy} = \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x)\text{var}(y)}} \]  

- Not all uncorrelated \( xy \) are independent.
Sampling Distributions

- The sampling distribution of a statistic is the probability distribution of a statistic obtained from repeated sampling.
Sampling Distributions

- The sampling distribution of a statistic is the probability distribution of a statistic obtained from repeated sampling.
- Suppose our statistic is $\theta$. 

Sampling Distributions

- The sampling distribution of a statistic is the probability distribution of a statistic obtained from repeated sampling.
- Suppose our statistic is $\theta$.
- Let’s review some basic sampling concepts (using R.}

Central Limit Theorem:
The sum of many independently and identically distributed random variables will tend to a normal distribution in the limit.

More specifically, if the sum of independently and identically distributed variables has a mean $\mu$ and a finite variance $\sigma^2$, then it will approximately follow a normal distribution.

To sustain statistical inference, we rely heavily on the central limit theorem.

Importantly, this result holds even if the population distribution is non-normal.
Sampling Distributions

- The sampling distribution of a statistic is the probability distribution of a statistic obtained from repeated sampling.
- Suppose our statistic is \( \theta \).
- Let’s review some basic sampling concepts (using R).
- Central Limit Theorem:
  The sum of many independently and identically distributed random variables will tend to a normal distribution in the limit. More specifically, if the sum of independently and identically distributed variables has a mean \( \mu \) and a finite variance \( \sigma^2 \), then it will approximately follow a normal distribution.
Sampling Distributions

The sampling distribution of a statistic is the probability distribution of a statistic obtained from repeated sampling.

Suppose our statistic is $\theta$.

Let’s review some basic sampling concepts (using R.

Central Limit Theorem:
The sum of many independently and identically distributed random variables will tend to a normal distribution in the limit. More specifically, if the sum of independently and identically distributed variables has a mean $\mu$ and a finite variance $\sigma^2$, then it will approximately follow a normal distribution.

To sustain statistical inference, we rely heavily on the central limit theorem.
Sampling Distributions

- The sampling distribution of a statistic is the probability distribution of a statistic obtained from repeated sampling.
- Suppose our statistic is $\theta$.
- Let’s review some basic sampling concepts (using R).
- Central Limit Theorem:
  The sum of many independently and identically distributed random variables will tend to a normal distribution in the limit. More specifically, if the sum of independently and identically distributed variables has a mean $\mu$ and a finite variance $\sigma^2$, then it will approximately follow a normal distribution.
- To sustain statistical inference, we rely heavily on the central limit theorem.
- Importantly, this result holds even if the population distribution is non-normal.
Let’s create a world of 100,000 observations. This is our population and this is what it looks like:

\[
> \text{samplesize}<-100000 \\
> \text{dist}<-\text{sample}(\text{rgamma}(\text{samplesize},5,5)) \\
> \text{hist(dist, col="blue1")} \\
> \text{meanX} <- \text{mean(dist)}; \text{meanX} \\
\text{[1]} \ 0.996132
\]

Here, the mean of the distribution is 1.00. Call this \( \mu \). The population distribution looks like this:
Artwork

Histogram of dist

Frequency

0
5000
10000
15000

0
1
2
3
4
dist
The central limit theorem says that the distribution of “draws” of some statistic from the population, even if the population is nonnormal, will tend to a normal distribution. Suppose I took one sample of size 500?

```r
> set.seed(510951)
> nsamp <- 1
> res <- numeric(nsamp)
> for (i in 1:nsamp) res[i] <- mean(sample(dist, 100, replace = TRUE))
> mean(res)
[1] 1.020171
```

The sample estimate, call it $\bar{X}$ is 1.02. It’s “off” from $\mu$. Suppose we were to take 10,000 samples of size 100?
Sampling Distribution

- This distribution is our sampling distribution.
Sampling Distribution

- This distribution is our sampling distribution.
- It gives the distribution of *means* from repeated samples.
Sampling Distribution

- This distribution is our sampling distribution.
- It gives the distribution of *means* from repeated samples.
- The mean of the sampling distribution is 1.00.
Sampling Distribution

- This distribution is our sampling distribution.
- It gives the distribution of means from repeated samples.
- The mean of the sampling distribution is 1.00.
- The variability around the mean is the standard deviation.
Sampling Distribution

- This distribution is our sampling distribution.
- It gives the distribution of means from repeated samples.
- The mean of the sampling distribution is 1.00.
- The variability around the mean is the standard deviation.
- Inferentially, the standard error reported to you in regression output is “the standard deviation of the sampling distribution.”
Sampling Distribution

- This distribution is our sampling distribution.
- It gives the distribution of means from repeated samples.
- The mean of the sampling distribution is 1.00.
- The variability around the mean is the standard deviation.
- Inferentially, the standard error reported to you in regression output is “the standard deviation of the sampling distribution.”
- The problem of estimation is we only have one sample with which to work.
Sampling Distribution

- This distribution is our sampling distribution.
- It gives the distribution of means from repeated samples.
- The mean of the sampling distribution is 1.00.
- The variability around the mean is the standard deviation.
- Inferentially, the standard error reported to you in regression output is “the standard deviation of the sampling distribution.”
- The problem of estimation is we only have one sample with which to work.
- However, the nice thing about the CLT is “it gets us to the normal.” (Or very close to it).
Estimators have no inherent use to us without properties.
Properties of Estimators

- Estimators have no inherent use to us without properties.
- A random guess or mere dead reckoning is an estimator.
Properties of Estimators

- Estimators have no inherent use to us without properties.
- A random guess or mere dead reckoning is an estimator.
- It’s just not very good.
Properties of Estimators

▶ Estimators have no inherent use to us without properties.
▶ A random guess or mere dead reckoning *is* an estimator.
▶ It’s just not very good.
▶ Let’s review some properties of estimators.
Estimators

\( \hat{\theta} \) is what we’re interested in.
Estimators

- $\hat{\theta}$ is what we’re interested in.
- This is a statistic.
Estimators

- \( \hat{\theta} \) is what we’re interested in.
- This is a statistic.
- Often, we’re interested in the “first moment” of the sampling distribution.
Estimators

- $\hat{\theta}$ is what we’re interested in.
- This is a statistic.
- Often, we’re interested in the “first moment” of the sampling distribution.
- That is, the *mean of the sampling distribution.*
Estimators

- \( \hat{\theta} \) is what we’re interested in.
- This is a statistic.
- Often, we’re interested in the “first moment” of the sampling distribution.
- That is, the *mean of the sampling distribution*.
- The question is, what are the desirable properties of an estimator.
Main characteristics of $\hat{\theta}$: The mean is $E(\hat{\theta})$. 
Main characteristics of $\hat{\theta}$: The mean is $E(\hat{\theta})$.

The variance:

$$\text{var}(\hat{\theta}) = E(\hat{\theta}^2) - [E(\hat{\theta})]^2$$  \hspace{1cm} (20)
Main characteristics of $\hat{\theta}$: The mean is $E(\hat{\theta})$.

The variance:

$$\text{var}(\hat{\theta}) = E(\hat{\theta}^2) - [E(\hat{\theta})]^2$$ \hspace{1cm} (20)

Why?
Small Sample Properties: Unbiasedness

“Small sample properties.”
Small Sample Properties: Unbiasedness

▶ “Small sample properties.”
▶ Unbiasedness: \( E(\hat{\theta}) = \theta \).
Small Sample Properties: Unbiasedness

- “Small sample properties.”
- Unbiasedness: $E(\hat{\theta}) = \theta$.
- Equivalently: $E(\hat{\theta}) - \theta = 0$. 
Small Sample Properties: Unbiasedness

▶ “Small sample properties.”
▶ Unbiasedness: $E(\hat{\theta}) = \theta$.
▶ Equivalently: $E(\hat{\theta}) - \theta = 0$.
▶ The estimator is unbiased.
Small Sample Properties: Unbiasedness

- “Small sample properties.”
- Unbiasedness: $E(\hat{\theta}) = \theta$.
- Equivalently: $E(\hat{\theta}) - \theta = 0$.
- The estimator is unbiased.
- In contrast: $E(\hat{\theta}) - \theta \neq 0$
“Small sample properties.”

Unbiasedness: $E(\hat{\theta}) = \theta$.

Equivalently: $E(\hat{\theta}) - \theta = 0$.

The estimator is unbiased.

In contrast: $E(\hat{\theta}) - \theta \neq 0$

Implies biasedness in the estimator.
Small Sample Properties: Unbiasedness

- “Small sample properties.”
- Unbiasedness: $E(\hat{\theta}) = \theta$.
- Equivalently: $E(\hat{\theta}) - \theta = 0$.
- The estimator is unbiased.
- In contrast: $E(\hat{\theta}) - \theta \neq 0$
- Implies biasedness in the estimator.
- Note: this property is a “repeated sampling” property.
Small Sample Properties: Unbiasedness

- “Small sample properties.”
- Unbiasedness: $E(\hat{\theta}) = \theta$.
- Equivalently: $E(\hat{\theta}) - \theta = 0$.
- The estimator is unbiased.
- In contrast: $E(\hat{\theta}) - \theta \neq 0$
- Implies biasedness in the estimator.
- Note: this property is a “repeated sampling” property.
- Note also: $E(\hat{\theta}) - \theta$ is sampling error.
Small Sample Properties: Efficiency

- Unbiasedness only tells us something about the central tendency of the sampling distribution.
Small Sample Properties: Efficiency

- Unbiasedness only tells us something about the central tendency of the sampling distribution.
- An estimator $\hat{\theta}$ is said to have “minimum variance” if $\text{var}(\hat{\theta}) < \text{var}(\tilde{\theta})$.

If $\hat{\theta}$ is a linear function of sample data, then $\hat{\theta}$ is a linear estimator.
Thus, if $\hat{\theta}$ is efficient and linear, then in the class of linear estimators, it is “best unbiased.”

BLUE—best linear unbiased estimator. (More on this later)

If the Gaussian assumptions hold (yet to be covered!), the OLS estimator has this property.
Unbiasedness only tells us something about the central tendency of the sampling distribution.

An estimator $\hat{\theta}$ is said to have “minimum variance” if $\text{var}(\hat{\theta}) < \text{var}(\tilde{\theta})$.

Efficiency: take two estimators, $\hat{\theta}$ and $\tilde{\theta}$. If each are unbiased but $\hat{\theta}$ is a minimum variance estimator, then $\hat{\theta}$ is efficient (or “best unbiased.”).
Unbiasedness only tells us something about the central tendency of the sampling distribution.

An estimator $\hat{\theta}$ is said to have “minimum variance” if $\text{var}(\hat{\theta}) < \text{var}(\tilde{\theta})$.

Efficiency: take two estimators, $\hat{\theta}$ and $\tilde{\theta}$. If each are unbiased but $\hat{\theta}$ is a minimum variance estimator, then $\hat{\theta}$ is efficient (or “best unbiased.”).

If $\hat{\theta}$ is a linear function of sample data, then $\hat{\theta}$ is a linear estimator.
Small Sample Properties: Efficiency

- Unbiasedness only tells us something about the central tendency of the sampling distribution.

- An estimator \( \hat{\theta} \) is said to have “minimum variance” if \( \text{var}(\hat{\theta}) < \text{var}(\tilde{\theta}) \).

- Efficiency: take two estimators, \( \hat{\theta} \) and \( \tilde{\theta} \). If each are unbiased but \( \hat{\theta} \) is a minimum variance estimator, then \( \hat{\theta} \) is efficient (or “best unbiased.”).

- If \( \hat{\theta} \) is a linear function of sample data, then \( \hat{\theta} \) is a linear estimator.

- Thus, if \( \hat{\theta} \) is efficient and linear, then in the class of linear estimators, it is “best unbiased.”
Small Sample Properties: Efficiency

- Unbiasedness only tells us something about the central tendency of the sampling distribution.
- An estimator $\hat{\theta}$ is said to have “minimum variance” if $\text{var}(\hat{\theta}) < \text{var}(\tilde{\theta})$.
- Efficiency: take two estimators, $\hat{\theta}$ and $\tilde{\theta}$. If each are unbiased but $\hat{\theta}$ is a minimum variance estimator, then $\hat{\theta}$ is efficient (or “best unbiased.”).
- If $\hat{\theta}$ is a linear function of sample data, then $\hat{\theta}$ is a linear estimator.
- Thus, if $\hat{\theta}$ is efficient and linear, then in the class of linear estimators, it is “best unbiased.”
- BLUE—best linear unbiased estimator. (More on this later)
Small Sample Properties: Efficiency

- Unbiasedness only tells us something about the central tendency of the sampling distribution.
- An estimator $\hat{\theta}$ is said to have “minimum variance” if $\text{var}(\hat{\theta}) < \text{var}(\tilde{\theta})$.
- Efficiency: take two estimators, $\hat{\theta}$ and $\tilde{\theta}$. If each are unbiased but $\hat{\theta}$ is a minimum variance estimator, then $\hat{\theta}$ is efficient (or “best unbiased.”).
- If $\hat{\theta}$ is a linear function of sample data, then $\hat{\theta}$ is a linear estimator.
- Thus, if $\hat{\theta}$ is efficient and linear, then in the class of linear estimators, it is “best unbiased.”
- BLUE–best linear unbiased estimator. (More on this later)
- If the Gaussian assumptions hold (yet to be covered!), the OLS estimator has this property.
MSE and Variance

- MSE:
  \[
  MSE(\hat{\theta} = E[\hat{\theta} - E(\theta)]^2 + [E(\hat{\theta} - \theta)]^2
  \] (21)
MSE and Variance

MSE:

\[ MSE(\hat{\theta} = E[\hat{\theta} - E(\theta)]^2 + [E(\hat{\theta} - \theta)]^2 \]  

MSE is a combination of sampling variance and bias.
MSE and Variance

- **MSE:**
  \[
  MSE(\hat{\theta} = E[\hat{\theta} - E(\theta)]^2 + [E(\hat{\theta} - \theta)]^2 \tag{21}
  \]

- MSE is a combination of sampling variance and bias.
- The total amount of error comes from two sources. If bias is 0, then \(MSE = \text{var}\).
Large Sample Properties

- Some estimators will not satisfy these properties in small samples.

Asymptotic unbiasedness: \( \lim_{n \to \infty} E(\hat{\theta}_n) = \theta \).

Asymptotic properties are directly tied to sample sizes: small samples, they will not hold.

You've seen this before:

\[ s^2 = \frac{\sum (X_i - \bar{X})^2}{n} \]

In small samples, this estimator for the variance is biased; in large samples, the bias tends to 0.

Thus, the estimator is asymptotically unbiased.
Large Sample Properties

- Some estimators will not satisfy these properties in small samples.
- Only in large samples do approximately equivalent properties hold.
Large Sample Properties

- Some estimators will not satisfy these properties in small samples.
- Only in large samples do approximately equivalent properties hold.
- These kinds of properties are *asymptotic properties* or large sample properties.

Asymptotic unbiasedness: \( \lim_{n \to \infty} E(\hat{\theta}_n) = \theta \).

Asymptotic properties are directly tied to sample sizes: small samples, they will not hold.

You've seen this before: \( s^2 = \frac{1}{n} \sum (X_i - X)^2 \).

In small samples, this estimator for the variance is biased; in large samples, the bias tends to 0.

Thus, the estimator is asymptotically unbiased.
Large Sample Properties

- Some estimators will not satisfy these properties in small samples.
- Only in large samples do approximately equivalent properties hold.
- These kinds of properties are *asymptotic properties* or large sample properties.
- Asymptotic unbiasedness: \( \lim_{n \to \infty} E(\hat{\theta}_n) = \theta \).
Large Sample Properties

- Some estimators will not satisfy these properties in small samples.
- Only in large samples do approximately equivalent properties hold.
- These kinds of properties are *asymptotic properties* or large sample properties.
- Asymptotic unbiasedness: \( \lim_{n \to \infty} E(\hat{\theta}_n) = \theta \).
- Asymptotic properties are directly tied to sample sizes: small samples, they will not hold.
Large Sample Properties

- Some estimators will not satisfy these properties in small samples.
- Only in large samples do approximately equivalent properties hold.
- These kinds of properties are *asymptotic properties* or large sample properties.
- Asymptotic unbiasedness: \( \lim_{n \to \infty} E(\hat{\theta}_n) = \theta \).
- Asymptotic properties are directly tied to sample sizes: small samples, they will not hold.
- You’ve seen this before:
  \[
  s^2 = \frac{\sum(X_i - \bar{X})^2}{n}
  \]
Large Sample Properties

- Some estimators will not satisfy these properties in small samples.
- Only in large samples do approximately equivalent properties hold.
- These kinds of properties are *asymptotic properties* or large sample properties.
- Asymptotic unbiasedness: \( \lim_{n \to \infty} E(\hat{\theta}_n) = \theta \).
- Asymptotic properties are directly tied to sample sizes: small samples, they will not hold.
- You’ve seen this before: 
  \[
  s^2 = \frac{\sum(X_i - \bar{X})^2}{n}
  \]
- In small samples, this estimator for the variance is biased; in large samples, the bias tends to 0.
Large Sample Properties

- Some estimators will not satisfy these properties in small samples.
- Only in large samples do approximately equivalent properties hold.
- These kinds of properties are *asymptotic properties* or large sample properties.
- Asymptotic unbiasedness: \( \lim_{n \to \infty} E(\hat{\theta}_n) = \theta \).
- Asymptotic properties are directly tied to sample sizes: small samples, they will not hold.
- You’ve seen this before:
  \[
  s^2 = \frac{\sum (X_i - \bar{X})^2}{n}
  \]
- In small samples, this estimator for the variance is biased; in large samples, the bias tends to 0.
- Thus, the estimator is *asymptotically unbiased.*
Large Sample Properties: Consistency

- Consistency is a probabilistic statement:

\[
\lim_{n \to \infty} \Pr\{ | \hat{\theta} - \theta | < \delta \} = 1 \quad \delta > 0. \tag{22}
\]
Large Sample Properties: Consistency

- Consistency is a probabilistic statement:

\[
\lim_{n \to \infty} \Pr\{|\hat{\theta} - \theta| < \delta\} = 1 \quad \delta > 0. 
\] (22)

- Or \(\text{plim}\hat{\theta} = \theta\).

Note that unbiasedness can hold for any sample size; consistency is purely an asymptotic property. A sufficient condition for consistency is that the bias and variance both tend toward 0 as \(n\) increases. Note that the MSE criteria is not used in the OLS context because it is biased. In large samples, this bias diminishes. The central limit theorem is an "asymptotic theorem." That is, asymptotic normality holds if the sampling distribution of \(\hat{\theta}\) \(\to N\) as the sample size increases.
Consistency is a probabilistic statement:

$$\lim_{n \to \infty} \Pr\{ | \hat{\theta} - \theta | < \delta \} = 1 \quad \delta > 0.$$  \hfill (22)

Or $\text{plim} \hat{\theta} = \theta$.

This is the consistency condition.
Large Sample Properties: Consistency

- Consistency is a probabilistic statement:

\[
\lim_{n \to \infty} \Pr\{|\hat{\theta} - \theta| < \delta\} = 1 \quad \delta > 0.
\] (22)

- Or \(\text{plim} \hat{\theta} = \theta\).

- This is the consistency condition.

- Note that unbiasedness can hold for any sample size; consistency is purely an asymptotic property.
Large Sample Properties: Consistency

- Consistency is a probabilistic statement:

\[
\lim_{n \to \infty} \Pr\{ | \hat{\theta} - \theta | < \delta \} = 1 \quad \delta > 0.
\] (22)

- Or \( \text{plim} \hat{\theta} = \theta \).
- This is the consistency condition.
- Note that unbiasedness can hold for any sample size; consistency is purely an asymptotic property.
- A sufficient condition for consistency is that the bias and variance both tend toward 0 as \( n \) increases.
Large Sample Properties: Consistency

- Consistency is a probabilistic statement:
  \[
  \lim_{n \to \infty} \Pr\{| \hat{\theta} - \theta | < \delta \} = 1 \quad \delta > 0.
  \]
  \[\text{(22)}\]

- Or \( \text{plim} \hat{\theta} = \theta \).
- This is the consistency condition.
- Note that unbiasedness can hold for any sample size; consistency is purely an asymptotic property.
- A sufficient condition for consistency is that the bias and variance both tend toward 0 as \( n \) increases.
- Note that the \( MSE \) criteria is not used in the OLS context because it is biased. In large samples, this bias diminishes.
Large Sample Properties: Consistency

Consistency is a probabilistic statement:

$$\lim_{n \to \infty} \Pr\{|\hat{\theta} - \theta| < \delta\} = 1 \quad \delta > 0.$$  \hfill (22)

Or $\text{plim}\hat{\theta} = \theta$.

This is the consistency condition.

Note that unbiasedness can hold for any sample size; consistency is purely an asymptotic property.

A sufficient condition for consistency is that the bias and variance both tend toward 0 as $n$ increases.

Note that the $MSE$ criteria is not used in the OLS context because it is biased. In large samples, this bias diminishes.

The central limit theorem is an “asymptotic theorem.”
Large Sample Properties: Consistency

- Consistency is a probabilistic statement:
  \[
  \lim_{n \to \infty} \Pr\{|\hat{\theta} - \theta| < \delta\} = 1 \quad \delta > 0. \tag{22}
  \]

- Or \( \text{plim} \hat{\theta} = \theta \).
- This is the consistency condition.
- Note that unbiasedness can hold for any sample size; consistency is purely an asymptotic property.
- A sufficient condition for consistency is that the bias and variance both tend toward 0 as \( n \) increases.
- Note that the MSE criteria is not used in the OLS context because it is biased. In large samples, this bias diminishes.
- The central limit theorem is an “asymptotic theorem.”
- That is, asymptotic normality holds if the sampling distribution of \( \hat{\theta} \to N \) as the sample size increases.
Some Matrix Basics

What is a matrix?
Some Matrix Basics

- What is a matrix?
- “A rectangular array of elements arranged in rows and columns.”

\[
\begin{pmatrix}
55 & 900 & 0 \\
67 & 1112 & 1 \\
73 & 525 & 0
\end{pmatrix}
\]
Some Matrix Basics

- What is a matrix?
- “A rectangular array of elements arranged in rows and columns.”

\[
\begin{pmatrix}
55 & 900 & 0 \\
67 & 1112 & 1 \\
73 & 525 & 0
\end{pmatrix}
\]

- All data are really matrices.
Some Matrix Basics

- What is a matrix?
- “A rectangular array of elements arranged in rows and columns.”

\[
\begin{pmatrix}
55 & 900 & 0 \\
67 & 1112 & 1 \\
73 & 525 & 0 \\
\end{pmatrix}
\]

- All data are really matrices.
- Suppose we have “votes,” “money,” and “PID.”
Some Matrix Basics

▶ What is a matrix?
▶ “A rectangular array of elements arranged in rows and columns.”

\[
\begin{pmatrix}
55 & 900 & 0 \\
67 & 1112 & 1 \\
73 & 525 & 0 \\
\end{pmatrix}
\]

▶ All data are really matrices.
▶ Suppose we have “votes,” “money,” and “PID.”
▶ Dimension of a matrix is $r \times c : 3 \times 3$. 
Some Matrix Basics

- Dimensionality
Some Matrix Basics

- Dimensionality
- If we had 435 MCs and 3 variables:

\[
\begin{pmatrix}
55 & 900 & 0 \\
67 & 1112 & 1 \\
73 & 525 & 0 \\
\vdots & \vdots & \vdots \\
67 & 874 & 1
\end{pmatrix}
\]
Some Matrix Basics

- Dimensionality

- If we had 435 MCs and 3 variables:

\[
\begin{bmatrix}
55 & 900 & 0 \\
67 & 1112 & 1 \\
73 & 525 & 0 \\
\vdots & \vdots & \vdots \\
67 & 874 & 1 \\
\end{bmatrix}
\]

- Dimensionality: 435 × 3
Some Matrix Basics

- Dimensionality

- If we had 435 MCs and 3 variables:

\[
\begin{pmatrix}
55 & 900 & 0 \\
67 & 1112 & 1 \\
73 & 525 & 0 \\
\vdots & \vdots & \vdots \\
67 & 874 & 1
\end{pmatrix}
\]

- Dimensionality: 435 × 3

- Applied issue: always a good idea to know the dimensions of your research problem.
Some Matrix Basics

▶ Symbolism
Some Matrix Basics

- Symbolism
- We can use symbols to denote $rc$:

$$
\begin{pmatrix}
a_{11} & a_{12} & a_{13} \\
a_{21} & a_{22} & a_{23} \\
a_{31} & a_{32} & a_{33}
\end{pmatrix}
$$
Some Matrix Basics

- Symbolism
- We can use symbols to denote $rc$:

$$
\begin{pmatrix}
a_{11} & a_{12} & a_{13} \\
a_{21} & a_{22} & a_{23} \\
a_{31} & a_{32} & a_{33}
\end{pmatrix}
$$

- Often will denote matrices with bold-faced symbols:

$$
A =
\begin{pmatrix}
a_{11} & a_{12} & a_{13} \\
a_{21} & a_{22} & a_{23} \\
a_{31} & a_{32} & a_{33}
\end{pmatrix}
$$
Vectors

- A matrix with a single column is a “column vector.”

\[
\begin{pmatrix}
a_{11} \\
a_{21} \\
a_{31}
\end{pmatrix}
\]
Vectors

- A matrix with a single column is a “column vector.”
  \[
  \begin{pmatrix}
  a_{11} \\
  a_{21} \\
  a_{31}
  \end{pmatrix}
  \]

- A matrix with a single row is a “row vector.”
  \[
  \begin{pmatrix}
  a_{11} & a_{12} & a_{13}
  \end{pmatrix}
  \]

Dimensions of the column vector?
Dimensions of the row vector?

Answers: 3 \times 1; 1 \times 3
Vectors

- A matrix with a single column is a “column vector.”

\[
\begin{pmatrix}
  a_{11} \\
  a_{21} \\
  a_{31}
\end{pmatrix}
\]

- A matrix with a single row is a “row vector.”

\[
\begin{pmatrix}
  a_{11} & a_{12} & a_{13}
\end{pmatrix}
\]

- Dimensions of the column vector?

Answers: 3 \times 1; 1 \times 3
Vectors

- A matrix with a single column is a “column vector.”
  \[
  \begin{pmatrix}
  a_{11} \\
  a_{21} \\
  a_{31}
  \end{pmatrix}
  \]

- A matrix with a single row is a “row vector.”
  \[
  \begin{pmatrix}
  a_{11} & a_{12} & a_{13}
  \end{pmatrix}
  \]

- Dimensions of the column vector?
- Dimensions of the row vector?
Vectors

- A matrix with a single column is a “column vector.”
  \[
  \begin{pmatrix}
  a_{11} \\
  a_{21} \\
  a_{31}
  \end{pmatrix}
  \]

- A matrix with a single row is a “row vector.”
  \[
  (a_{11} \ a_{12} \ a_{13})
  \]

- Dimensions of the column vector?
- Dimensions of the row vector?
- Answers: $3 \times 1; 1 \times 3$
Vectors

- A matrix with a single column is a "column vector."

\[
\begin{pmatrix}
a_{11} \\
a_{21} \\
a_{31}
\end{pmatrix}
\]

- A matrix with a single row is a "row vector."

\[
\begin{pmatrix}
a_{11} & a_{12} & a_{13}
\end{pmatrix}
\]

- Dimensions of the column vector?
- Dimensions of the row vector?
- Answers: $3 \times 1$; $1 \times 3$
- Linear models, $y$ is usually an $n \times 1$ matrix.
The transpose of matrix $A$ is another matrix, $A'$. 

$A' = 
\begin{pmatrix}
 a_{11} & a_{21} & a_{31} \\
 a_{12} & a_{22} & a_{32} \\
 a_{13} & a_{23} & a_{33}
\end{pmatrix}$

That is, the first column of $A$ is now the first row of $A'$. 

Transposing Matrices

- The transpose of matrix $A$ is another matrix, $A'$. 
- It is obtained by interchanging corresponding columns and rows.
Transposing Matrices

- The transpose of matrix $\mathbf{A}$ is another matrix, $\mathbf{A}'$.  
- It is obtained by interchanging corresponding columns and rows.  
- So if $\mathbf{A}$ is:  

\[
\begin{pmatrix}
a_{11} & a_{12} & a_{13} \\
a_{21} & a_{22} & a_{23} \\
a_{31} & a_{32} & a_{33}
\end{pmatrix}
\]

Then $\mathbf{A}'$ is:  

\[
\begin{pmatrix}
a_{11} & a_{21} & a_{31} \\
a_{12} & a_{22} & a_{32} \\
a_{13} & a_{23} & a_{33}
\end{pmatrix}
\]

That is, the first column of $\mathbf{A}$ is now the first row of $\mathbf{A}'$. 

Transposing Matrices

- The transpose of matrix $A$ is another matrix, $A'$. It is obtained by interchanging corresponding columns and rows.
- So if $A$ is:
  $$
  \begin{pmatrix}
  a_{11} & a_{12} & a_{13} \\
  a_{21} & a_{22} & a_{23} \\
  a_{31} & a_{32} & a_{33}
  \end{pmatrix}
  $$
- Then $A'$ is:
  $$
  \begin{pmatrix}
  a_{11} & a_{21} & a_{31} \\
  a_{12} & a_{22} & a_{32} \\
  a_{13} & a_{23} & a_{33}
  \end{pmatrix}
  $$
- That is, the first column of $A$ is now the first row of $A'$. 
Transposing Matrices

- The transpose of a column vector is a row vector (and vice versa).
Transposing Matrices

- The transpose of a column vector is a row vector (and vice versa).
- Note that the dimensions of the input matrix $\mathbf{A}$ is $r \times c$.
- For the transpose matrix, $\mathbf{A}'$, the dimension is $c \times r$. 
Transposing Matrices

- The transpose of a column vector is a row vector (and vice versa).
- Note that the dimensions of the input matrix $A$ is $r \times c$.
- For the transpose matrix, $A'$, the dimension is $c \times r$.
- Transposing matrices is really important in regression-like settings.
Generating “data” in R as vectors:

\[ x_0 \leftarrow c(1,1,1,1,1) \]
\[ x_1 \leftarrow c(2,4,6,8,10) \]
\[ x_2 \leftarrow c(7.3,8.6,9.2,3.2,1.2) \]
\[ y \leftarrow c(32,14,71,81,101) \]
Creating a matrix $\mathbf{X}$:

```r
xmat <- cbind(x0, x1, x2); xmat
```

```
x0  x1  x2
[1,] 1  2  7.3
[2,] 1  4  8.6
[3,] 1  6  9.2
[4,] 1  8  3.2
[5,] 1 10  1.2
```

```r
xmat[1,3]
```

```
[1] 7.3
```

```r
xmat[1,1]
```

```
[1] 1
```

```r
xmat[1,]
```

```
x0  x1  x2
1.0 2.0  7.3
```
Basic Manipulations: Multiplication by a scalar:

\[ 2 \times \text{xmat} \]

\[
\begin{array}{ccc}
 x0 & x1 & x2 \\
[1,] & 2 & 4 & 14.6 \\
[2,] & 2 & 8 & 17.2 \\
[3,] & 2 & 12 & 18.4 \\
[4,] & 2 & 16 & 6.4 \\
[5,] & 2 & 20 & 2.4 \\
\end{array}
\]
Basic Manipulations: Transposing $X$ to create $X'$:

```r
xprime<-t(xmat); xprime
x0  1.0  1.0  1.0  1.0  1.0
x1  2.0  4.0  6.0  8.0 10.0
x2  7.3  8.6  9.2  3.2  1.2
```

Transpose $Y$ to create $Y'$:

```r
y
[1] 32 14 71 81 101
```

```r
yprime<-t(y); yprime
[1,]  32  14  71  81 101
```

Jones
POL 212: Research Methods
Multiplying Matrices

Multiplying matrices will be important for us in order to derive variance and covariances.
Multiplying Matrices

- Multiplying matrices will be important for us in order to derive variance and covariances.
- Multiplication of matrices having the same dimensions is not problematic.
Multiplying Matrices

- Multiplying matrices will be important for us in order to derive variance and covariances.
- Multiplication of matrices having the same dimensions is not problematic.
- Suppose \( A \) is:
  
  $$
  \begin{pmatrix}
  2 & 5 \\
  4 & 1 
  \end{pmatrix}
  $$

- Suppose \( B \) is:
  
  $$
  \begin{pmatrix}
  4 & 6 \\
  5 & 8 
  \end{pmatrix}
  $$

  The product is a matrix, \( AB \). Its elements are obtained by finding the cross-products of rows of \( A \) with columns of \( B \).
Multiplying Matrices

- Multiplying matrices will be important for us in order to derive variance and covariances.
- Multiplication of matrices having the same dimensions is not problematic.
- Suppose $A$ is:
  \[
  \begin{pmatrix}
  2 & 5 \\
  4 & 1 
  \end{pmatrix}
  \]
- Suppose $B$ is:
  \[
  \begin{pmatrix}
  4 & 6 \\
  5 & 8 
  \end{pmatrix}
  \]
- The product is a matrix, $AB$. Its elements are obtained by finding the cross-products of rows of $A$ with columns of $B$. 
For these, the product matrix is obtained by:

\[
\begin{align*}
2(4) + 5(5) &= 33 \\
2(6) + 5(8) &= 52 \\
4(4) + 1(5) &= 21 \\
4(6) + 1(8) &= 32
\end{align*}
\]
Multiplying Matrices

- For these, the product matrix is obtained by:
  \[ 2(4) + 5(5) = 33 \]
  \[ 2(6) + 5(8) = 52 \]
  \[ 4(4) + 1(5) = 21 \]
  \[ 4(6) + 1(8) = 32 \]

- Product matrix \(AB\):
  \[
  \begin{pmatrix}
  33 & 52 \\ 
  21 & 32 
  \end{pmatrix}
  \]
Multiplying Matrices

For these, the product matrix is obtained by:

\[
\begin{align*}
2(4) + 5(5) &= 33 \\
2(6) + 5(8) &= 52 \\
4(4) + 1(5) &= 21 \\
4(6) + 1(8) &= 32
\end{align*}
\]

Product matrix \( AB \):

\[
\begin{pmatrix}
33 & 52 \\
21 & 32
\end{pmatrix}
\]

Above, we say we postmultiplied \( A \) by \( B \) or premultiplied \( B \) by \( A \).
Multiplying Matrices

For these, the product matrix is obtained by:

\[
\begin{align*}
2(4) + 5(5) &= 33 \\
2(6) + 5(8) &= 52 \\
4(4) + 1(5) &= 21 \\
4(6) + 1(8) &= 32
\end{align*}
\]

Product matrix \( AB \):

\[
\begin{pmatrix}
33 & 52 \\
21 & 32
\end{pmatrix}
\]

Above, we say we postmultiplied \( A \) by \( B \) or premultiplied \( B \) by \( A \).

Why the language? Multiplication is not always defined on matrices.
Scalar multiplication has the commutative property:

\[ xy = yx \]
Multiplying Matrices

- Scalar multiplication has the commutative property:
  
  \[ xy = yx \]

- However, \( AB \) may not equal \( BA \).
Multiplying Matrices

- Scalar multiplication has the commutative property:
  \[ xy = yx \]

- However, \( AB \) may not equal \( BA \).

- The product matrix is defined when the number of columns in \( A \) equals the number of rows in \( B \).
Multiplying Matrices

- Scalar multiplication has the commutative property:
  \[ xy = yx \]

- However, \( AB \) may not equal \( BA \).

- The product matrix is defined when the number of columns in \( A \) equals the number of rows in \( B \).

- When this condition holds, we say the matrices are conformable for multiplication.
Multiplying Matrices

- Scalar multiplication has the commutative property:
  \[ xy = yx \]

- However, \( AB \) may not equal \( BA \).

- The product matrix is defined when the number of columns in \( A \) equals the number of rows in \( B \).

- When this condition holds, we say the matrices are conformable for multiplication.

- If \( A \) is \( 2 \times 2 \) and \( B \) is \( 2 \times 2 \), \( AB \) is conformable (and would have dimension \( 2 \times 2 \)).

- If \( A \) is \( 2 \times 3 \) and \( B \) is \( 3 \times 1 \), \( AB \) is conformable (and would have dimension \( 2 \times 1 \)).

- Note, \( BA \) is not conformable (why?).
The OLS model can be written as:

\[ y = X\beta + \epsilon. \]  (23)
Regression

- The OLS model can be written as:

\[ y = X\beta + \epsilon. \]  \hspace{1cm} (23)

- What are the dimensions of each matrix?
Regression

- The OLS model can be written as:

\[ y = X\beta + \epsilon. \]  \hspace{1cm} (23)

- What are the dimensions of each matrix?
- They are:

\[
\begin{align*}
  y & : n \times 1 \\
  X & : n \times k \\
  \beta & : k \times 1 \\
  \epsilon & : n \times 1
\end{align*}
\]  \hspace{1cm} (24)

- Note that \( X\beta \) is conformable for multiplication (i.e. we have one parameter for each column vector in \( X \)).
Basic Manipulations: Multiplying Matrix $X'$ by $X$ to create $X'X$.

```r
xprime_x <- xprime %*% xmat
>
xprime_x
   x0  x1  x2
x0  5.0 30.0 29.50
x1  30.0 220.0 141.80
x2  29.5 141.8 223.57
```

$X'$ had dimension $3 \times 5$.
$X$ had dimension $5 \times 3$.
Thus, $X'X$ is $3 \times 3$ (and is obviously conformable for multiplication (i.e. $n$ columns of $X'$ equal $n$ rows of $X$; i.e. 5=5).
One of the most important manipulations we need to perform is matrix inversion.
One of the most important manipulations we need to perform is matrix inversion.

In matrix algebra, inversion is “kind of like” division.
One of the most important manipulations we need to perform is matrix inversion.

In matrix algebra, inversion is “kind of like” division.

For a scalar, $k^{-1} = 1/k$. 
Inversion

- One of the most important manipulations we need to perform is matrix inversion.
- In matrix algebra, inversion is “kind of like” division.
- For a scalar, $k^{-1} = 1/k$.
- Multiplication by an inverse has the property:

$$kk^{-1} = 1$$
Inversion

- One of the most important manipulations we need to perform is matrix inversion.
- In matrix algebra, inversion is “kind of like” division.
- For a scalar, $k^{-1} = 1/k$.
- Multiplication by an inverse has the property:
  \[ kk^{-1} = 1 \]
- The question for us is this: can we find a matrix such that
  \[ BA = AB = I \]
One of the most important manipulations we need to perform is matrix inversion.

In matrix algebra, inversion is “kind of like” division.

For a scalar, \( k^{-1} = 1/k \).

Multiplication by an inverse has the property:

\[
k k^{-1} = 1
\]

The question for us is this: can we find a matrix such that

\[
BA = AB = I
\]

What is \( I \)?
Identity Matrix

- Suppose $A$ is:
  $\begin{pmatrix}
  1 & 1 \\
  3 & 4
  \end{pmatrix}$

- Suppose $B$ is:
  $\begin{pmatrix}
  4 & -1 \\
  -3 & 1
  \end{pmatrix}$

- Multiplying $A \times B$ yields:
  $\begin{pmatrix}
  1 & 0 \\
  0 & 1
  \end{pmatrix}$

- Call this $I$: an "identity matrix."

- Here, $B$ is the inverse matrix and $A$ is said to be invertible.

- BTW, verify you could obtain $I$. 
Identity Matrix

- Suppose $A$ is:
  \[
  \begin{pmatrix}
  1 & 1 \\ 3 & 4
  \end{pmatrix}
  \]

- Suppose $B$ is:
  \[
  \begin{pmatrix}
  4 & -1 \\ -3 & 1
  \end{pmatrix}
  \]

Multiplying $A \times B$ yields:
\[
\begin{pmatrix}
1 & 0 \\ 0 & 1
\end{pmatrix}
\]

Call this $I$: an "identity matrix." Here, $B$ is the inverse matrix and $A$ is said to be invertible.

BTW, verify you could obtain $I$. 
Identity Matrix

- Suppose \( A \) is:
  \[
  \begin{pmatrix}
  1 & 1 \\
  3 & 4 \\
  \end{pmatrix}
  \]

- Suppose \( B \) is:
  \[
  \begin{pmatrix}
  4 & -1 \\
  -3 & 1 \\
  \end{pmatrix}
  \]

- Multiplying \( A \times B \) yields:
  \[
  \begin{pmatrix}
  1 & 0 \\
  0 & 1 \\
  \end{pmatrix}
  \]

- Call this \( I \): an “identity matrix.”
- Here, \( B \) is the inverse matrix and \( A \) is said to be invertible.
- BTW, verify you could obtain \( I \).
Identity Matrix

- Suppose $A$ is:
  $$\begin{pmatrix} 1 & 1 \\ 3 & 4 \end{pmatrix}$$

- Suppose $B$ is:
  $$\begin{pmatrix} 4 & -1 \\ -3 & 1 \end{pmatrix}$$

- Multiplying $A \times B$ yields:
  $$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

- Call this $I$: an “identity matrix.”
Identity Matrix

▷ Suppose \( A \) is:
\[
\begin{pmatrix}
1 & 1 \\
3 & 4
\end{pmatrix}
\]

▷ Suppose \( B \) is:
\[
\begin{pmatrix}
4 & -1 \\
-3 & 1
\end{pmatrix}
\]

▷ Multiplying \( A \times B \) yields:
\[
\begin{pmatrix}
1 & 0 \\
0 & 1
\end{pmatrix}
\]

▷ Call this \( I \): an “identity matrix.”

▷ Here, \( B \) is the inverse matrix and \( A \) is said to be invertible.
Identity Matrix

- Suppose \( A \) is:

\[
\begin{pmatrix}
1 & 1 \\
3 & 4
\end{pmatrix}
\]

- Suppose \( B \) is:

\[
\begin{pmatrix}
4 & -1 \\
-3 & 1
\end{pmatrix}
\]

- Multiplying \( A \times B \) yields:

\[
\begin{pmatrix}
1 & 0 \\
0 & 1
\end{pmatrix}
\]

- Call this \( I \): an “identity matrix.”

- Here, \( B \) is the inverse matrix and \( A \) is said to be invertible.

- BTW, verify you could obtain \( I \).
Identity Matrix

- A diagonal matrix is a matrix with all 0s on the off-diagonal.
Identity Matrix

- A diagonal matrix is a matrix with all 0s on the off-diagonal.
- An identity matrix is a diagonal matrix with 1s down the main diagonal.
Identity Matrix

- A diagonal matrix is a matrix with all 0s on the off-diagonal.
- An identity matrix is a diagonal matrix with 1s down the main diagonal.
- In matrix algebra, the identity matrix serves the role of a “1.”
Determinants

- Determinants are a scalar: a number associated with a matrix.
Determinants

- Determinants are a scalar: a number associated with a matrix.
- They are useful in many mathematical endeavors.
Determinants

- Determinants are a scalar: a number associated with a matrix.
- They are useful in many mathematical endeavors.
- wrt matrix algebra, the determinant is useful in informing us as to whether or not a matrix is invertible.
Determinants

- Determinants are a scalar: a number associated with a matrix.
- They are useful in many mathematical endeavors.
- wrt matrix algebra, the determinant is useful in informing us as to whether or not a matrix is invertible.
- A matrix with a determinant of 0 is a **singular matrix**.
Determinants

- Determinants are a scalar: a number associated with a matrix.
- They are useful in many mathematical endeavors.
- wrt matrix algebra, the determinant is useful in informing us as to whether or not a matrix is invertible.
- A matrix with a determinant of 0 is a **singular matrix**.
- Important: a singular matrix is NOT invertible. This will cause us trouble!
Determinants

- Determinants are a scalar: a number associated with a matrix.
- They are useful in many mathematical endeavors.
- wrt matrix algebra, the determinant is useful in informing us as to whether or not a matrix is invertible.
- A matrix with a determinant of 0 is a **singular matrix**.
- Important: a singular matrix is NOT invertible. This will cause us trouble!
- A matrix with a non-zero determinant is a nonsingular matrix.
Determinants

- If all the elements of any row of a matrix are 0, its determinant is 0 and the matrix cannot be inverted.
Determinants

- If all the elements of any row of a matrix are 0, its determinant is 0 and the matrix cannot be inverted.
- If matrix is not of full rank, its determinant is 0.
Determinants

- If all the elements of any row of a matrix are 0, its determinant is 0 and the matrix cannot be inverted.
- If matrix is not of full rank, its determinant is 0.
- Finding a determinant is a bit complicated beyond the $2 \times 2$ setting (not mathematically complicated, just computationally tedious).
Determinants

- If all the elements of any row of a matrix are 0, its determinant is 0 and the matrix cannot be inverted.
- If matrix is not of full rank, its determinant is 0.
- Finding a determinant is a bit complicated beyond the $2 \times 2$ setting (not mathematically complicated, just computationally tedious).
- Consider this $2 \times 2$ matrix:

\[
\begin{pmatrix}
a_{11} & a_{21} \\
a_{12} & a_{22}
\end{pmatrix}
\]
Determinants

- If all the elements of any row of a matrix are 0, its determinant is 0 and the matrix cannot be inverted.
- If matrix is not of full rank, its determinant is 0.
- Finding a determinant is a bit complicated beyond the $2 \times 2$ setting (not mathematically complicated, just computationally tedious).
- Consider this $2 \times 2$ matrix:

\[
\begin{pmatrix}
  a_{11} & a_{21} \\
  a_{12} & a_{22}
\end{pmatrix}
\]

- In this setting, the determinant is found by cross-multiplication of elements on the main diagonal and subtracting them from cross-product elements on the off-diagonal.
Determinants

- If all the elements of any row of a matrix are 0, its determinant is 0 and the matrix cannot be inverted.
- If matrix is not of full rank, its determinant is 0.
- Finding a determinant is a bit complicated beyond the $2 \times 2$ setting (not mathematically complicated, just computationally tedious).
- Consider this $2 \times 2$ matrix:

\[
\begin{pmatrix}
a_{11} & a_{21} \\
a_{12} & a_{22}
\end{pmatrix}
\]

- In this setting, the determinant is found by cross-multiplication of elements on the main diagonal and subtracting them from cross-product elements on the off-diagonal.
- $|| A ||= a_{11}a_{22} - a_{21}a_{12}$
Determinants

Beyond the $2 \times 2$ setting, finding the determinant is time-consuming (why? LOTS of cross-product manipulations).
Determinants

- Beyond the $2 \times 2$ setting, finding the determinant is time-consuming (why? LOTS of cross-product manipulations).
- Easy to do computationally. The process is sometimes called matrix expansion.

∴ The rank of a matrix is the dimension (or order) of the largest square submatrix whose determinant is not 0.

For a matrix to be invertible, an $n \times n$ matrix must be of rank $n$. 
Determinants

- Beyond the $2 \times 2$ setting, finding the determinant is time-consuming (why? LOTS of cross-product manipulations).
- Easy to do computationally. The process is sometimes called matrix expansion.
- The important idea: if a determinant is 0, “stop!”
Determinants

- Beyond the $2 \times 2$ setting, finding the determinant is time-consuming (why? LOTS of cross-product manipulations).
- Easy to do computationally. The process is sometimes called matrix expansion.
- The important idea: if a determinant is 0, “stop!”
- The matrix is not invertible.
Determinants

- Beyond the $2 \times 2$ setting, finding the determinant is time-consuming (why? LOTS of cross-product manipulations).
- Easy to do computationally. The process is sometimes called matrix expansion.
- The important idea: if a determinant is 0, “stop!”
- The matrix is not invertible.
- As noted before, if a matrix is not of full rank, it will not be invertible.
Determinants

- Beyond the $2 \times 2$ setting, finding the determinant is time-consuming (why? LOTS of cross-product manipulations).
- Easy to do computationally. The process is sometimes called matrix expansion.
- The important idea: if a determinant is 0, “stop!”
- The matrix is not invertible.
- As noted before, if a matrix is not of full rank, it will not be invertible.
- $\therefore$ the rank of a matrix is the dimension (or order) of the largest square submatrix whose determinant is not 0.
Determinants

- Beyond the $2 \times 2$ setting, finding the determinant is time-consuming (why? LOTS of cross-product manipulations).
- Easy to do computationally. The process is sometimes called matrix expansion.
- The important idea: if a determinant is 0, “stop!”
- The matrix is not invertible.
- As noted before, if a matrix is not of full rank, it will not be invertible.
- \[ \therefore \] the rank of a matrix is the dimension (or order) of the largest square submatrix whose determinant is not 0.
- For a matrix to be invertible, an $n \times n$ matrix must be of rank $n$. 
Basic Manipulations: Finding the determinant:

> Define matrix q
>
> q1<-c(3,6)
> q2<-c(5,2)
> qmat<-cbind(q1,q2);qmat
   q1  q2
[1,]  3 5
[2,]  6 2

> #The Determinant
>
> #Finding determinant in 2x2 case
>
> det<-qmat[1,1]*qmat[2,2]-qmat[2,1]*qmat[1,2]; det

[1] -24

The determinant is $-24$ (i.e. it’s not 0; therefore, the matrix must be invertible.)
Basic Manipulations: Finding the inverse:

```r
> qmat
 q1  q2
[1,] 3 5
[2,] 6 2
>
> #Solving by hand
>
> qmat[1,1]/det
[1] -0.125
> qmat[2,2]/det
[1] -0.0833333
> -qmat[1,2]/det
[1] 0.2083333
> -qmat[2,1]/det
[1] 0.25
```

This would produce the inverse matrix. I could tell R to do this for me:

```r
> #Using R to do it for us
>
> qinverse<-solve(qmat); qinverse

[,1]     [,2]
q1 -0.08333333 0.2083333
q2  0.25000000 -0.1250000
```

These cells are identical to what we computed above.
Recall from before, if $BA = AB = I$, then from above is we multiply matrix $Q$ by $Q^{-1}$, we should obtain an identity matrix. Let’s check:

```r
qmat %*% qinverse
[,1] [,2]
[1,] 1 0
[2,] 0 1
```

Formally, this is $QQ^{-1}$ (and it’s an identity matrix).
At long last, regression

- The model:

\[ y = X\beta + \epsilon \]  

(25)
At long last, regression

- The model:
  \[ y = X\beta + \epsilon \]  
  (25)

- The dimensions of each matrix are as before.
At long last, regression

- The model:
  \[ y = X\beta + \epsilon \]  
  \( (25) \)

- The dimensions of each matrix are as before.

- The residual sums of squares (in scalar form):
  \[
  \sum_{i}^{n} \hat{\epsilon}_{i}^{2} = \sum_{i}(Y_{i} - \hat{\beta}_{0} - \hat{\beta}_{1}X_{1i} - \hat{\beta}_{2}X_{2i} - \ldots - \hat{\beta}_{k}X_{ki})^{2}
  \]  
  \( (26) \)
At long last, regression

- The model:
  \[ y = X\beta + \epsilon \]  
  (25)

- The dimensions of each matrix are as before.
- The residual sums of squares (in scalar form):
  \[ \sum_{i}^{n} \hat{\epsilon}_{i}^{2} = \sum_{i} (Y_{i} - \hat{\beta}_{0} - \hat{\beta}_{1}X_{1i} - \hat{\beta}_{2}X_{2i} - \ldots - \hat{\beta}_{k}X_{ki})^{2} \]  
  (26)

- In matrix form: \( \hat{\epsilon}'\hat{\epsilon} \). (A bit simpler).
At long last, regression

- The model:
  \[ y = X\beta + \epsilon \]  
  (25)

- The dimensions of each matrix are as before.
- The residual sums of squares (in scalar form):
  \[ \sum_{i} \hat{\epsilon}_{i}^2 = \sum (Y_{i} - \hat{\beta}_{0} - \hat{\beta}_{1}X_{1i} - \hat{\beta}_{2}X_{2i} - \ldots - \hat{\beta}_{k}X_{ki})^2 \]  
  (26)

- In matrix form: \( \hat{\epsilon}'\hat{\epsilon} \). (A bit simpler).
- But why?
At long last, regression

- The model:
  \[ y = X\beta + \epsilon \]  \hspace{1cm} (25)

- The dimensions of each matrix are as before.

- The residual sums of squares (in scalar form):
  \[ \sum_{i} \hat{\epsilon}_{i}^2 = \sum (Y_{i} - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} - \ldots - \hat{\beta}_k X_{ki})^2 \]  \hspace{1cm} (26)

- In matrix form: \( \hat{\epsilon}'\hat{\epsilon} \). (A bit simpler).

- But why?

- What are the dimensions of this matrix?
At long last, regression

- The model:
  \[ y = X\beta + \epsilon \] (25)

- The dimensions of each matrix are as before.

- The residual sums of squares (in scalar form):
  \[ \sum_{i}^{n} \hat{\epsilon}_{i}^{2} = \sum (Y_{i} - \hat{\beta}_{0} - \hat{\beta}_{1}X_{1i} - \hat{\beta}_{2}X_{2i} - \ldots - \hat{\beta}_{k}X_{ki})^{2} \] (26)

- In matrix form: \( \hat{\epsilon}'\hat{\epsilon} \). (A bit simpler).

- But why?

- What are the dimensions of this matrix?

- Multiply a column vector, \( \hat{\epsilon}' \), by a row vector, \( \hat{\epsilon} \), and you will get a scalar.
At long last, regression

- The model:
  \[ y = X\beta + \epsilon \]  
  (25)

- The dimensions of each matrix are as before.

- The residual sums of squares (in scalar form):
  \[ \sum_{i}^{n} \hat{\epsilon}_{i}^2 = \sum_{i}^{n} (Y_{i} - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} - \ldots - \hat{\beta}_k X_{ki})^2 \]  
  (26)

- In matrix form: \( \hat{\epsilon}'\hat{\epsilon} \). (A bit simpler).

- But why?

- What are the dimensions of this matrix?

- Multiply a column vector, \( \hat{\epsilon}' \), by a row vector, \( \hat{\epsilon} \), and you will get a scalar.

- i.e. dimension of the first matrix is \([1 \times n]\); the second is \([n \times 1]\); the dimension of the product: \([1]\).
Thus, $\hat{\epsilon}'\hat{\epsilon} = SSError$
Thus, $\hat{\epsilon}'\hat{\epsilon} = SSE_{\text{error}}$

Note in matrix terms what the residual is:

$$\epsilon = y - X\hat{\beta}.$$  \hspace{1cm} (27)
Thus, $\hat{\epsilon}'\hat{\epsilon} = SSE_{\text{error}}$

Note in matrix terms what the residual is:

$$\epsilon = y - X\hat{\beta}.$$  \hspace{1cm} (27)

The least squares solution is the one that minimizes the sum of the squared errors (hence “least” and “squares.”).
Regression

Thus, $\hat{\epsilon}'\hat{\epsilon} = SSE\text{rror}$

Note in matrix terms what the residual is:

$$\epsilon = y - X\hat{\beta}.$$  \hspace{1cm} (27)

The least squares solution is the one that minimizes the sum of the squared errors (hence “least” and “squares.”).

To solve for $\beta$, you partially differentiate the SSE\text{rror}...
For two variable regression, differentiate

\[ \sum \hat{e}_i^2 = \sum (Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i})^2 \]

partially with respect to the three unknown parameter estimates.
For two variable regression, differentiate

$$\Sigma \hat{e}_i^2 = \Sigma (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i})^2$$

partially with respect to the three unknown parameter estimates.

This gives

$$\frac{\partial \Sigma \hat{e}_i^2}{\partial \hat{\beta}_0} = 2\Sigma (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i})(-1),$$

$$\frac{\partial \Sigma \hat{e}_i^2}{\partial \hat{\beta}_1} = 2\Sigma (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i})(-X_{1i}),$$

$$\frac{\partial \Sigma \hat{e}_i^2}{\partial \hat{\beta}_2} = 2\Sigma (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i})(-X_{2i}),$$
Regression

When set to 0 and rearranging terms produces the normal equations:

\[
\begin{align*}
\bar{Y} &= \hat{\beta}_0 + \hat{\beta}_1 \bar{X}_{1i} + \hat{\beta}_2 \bar{X}_{2i} \\
\sum Y_i X_{1i} &= \hat{\beta}_0 \sum X_{1i} + \hat{\beta}_1 \sum X_{1i}^2 + \hat{\beta}_2 \sum X_{1i} X_{2i} \\
\sum Y_i X_{2i} &= \hat{\beta}_0 \sum X_{2i} + \hat{\beta}_1 \sum X_{2i}^2 + \hat{\beta}_2 \sum X_{1i} X_{2i}.
\end{align*}
\]
When set to 0 and rearranging terms produces the *normal equations*:

\[
\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}_1 + \hat{\beta}_2 \bar{X}_2 \\
\Sigma Y_i X_{1i} = \hat{\beta}_0 \Sigma X_{1i} + \hat{\beta}_1 \Sigma X_{1i}^2 + \hat{\beta}_2 \Sigma X_{1i} X_{2i} \\
\Sigma Y_i X_{2i} = \hat{\beta}_0 \Sigma X_{2i} + \hat{\beta}_1 \Sigma X_{2i}^2 + \hat{\beta}_2 \Sigma X_{1i} X_{2i}.
\]

These equations can be rewritten yet again in terms of the parameter estimates:

\[
\hat{\beta}_1 = \frac{\Sigma (X_1 - \bar{X}_1)(Y_i - \bar{Y})(\Sigma (X_2 - \bar{X})^2) - \Sigma (X_2 - \bar{X}_2)(Y_i - \bar{Y})(\Sigma (X_1 - \bar{X}_1)(X_2 - \bar{X}_2))}{\Sigma (X_1 - \bar{X}_1)^2 \Sigma (X_2 - \bar{X}_2)^2 - (\Sigma (X_1 - \bar{X}_1)(X_2 - \bar{X}_2))^2} \\
\hat{\beta}_2 = \frac{\Sigma (X_2 - \bar{X}_2)(Y_i - \bar{Y})(\Sigma (X_1 - \bar{X})^2) - \Sigma (X_1 - \bar{X}_1)(Y_i - \bar{Y})(\Sigma (X_1 - \bar{X}_1)(X_2 - \bar{X}_2))}{\Sigma (X_1 - \bar{X}_1)^2 \Sigma (X_2 - \bar{X}_2)^2 - (\Sigma (X_1 - \bar{X}_1)(X_2 - \bar{X}_2))^2} \\
\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2
\]
In matrix terms, the normal equations are summarized as:

\[(X'X)\hat{\beta} = X'y\]  \hspace{1cm} (28)
Regression

- In matrix terms, the normal equations are summarized as:

\[(X'X)\hat{\beta} = X'y\] (28)

- What is \(X'X\)?
Regression

- In matrix terms, the normal equations are summarized as:

\[(X'X)\hat{\beta} = X'y\] (28)

- What is \(X'X\)?
- It is a matrix of the raw sums of squares and cross-products for the \(X\) variables (include a constant term, which is a vector of 1s).
Remember from before, we created $X'X$ from our simulated data:

```r
xprime_x<-xprime %*% xmat
>
xprime_x
  x0  x1  x2
x0  5.0 30.0 29.50
x1  30.0 220.0 141.80
x2  29.5 141.8 223.57
```

Note several things: $[1, 1]$ is simply $n$; $[2, 1] = \sum X_{1i}$; $[3, 1] = \sum X_{2i}$; $[2, 2] = \sum X_{1i}^2$; $[3, 2] = \sum X_{2i} X_{1}$; $[3, 3] = \sum X_{1i}^2$. That is, these are simply functions of the data, nothing more.
From the normal equations, it is clear we have known quantities in \((X'X)\) and \(X'y\).
Regression

- From the normal equations, it is clear we have known quantities in \((X'X)\) and \(X'y\).
- We of course, do not know what \(\hat{\beta}\) are.
Regression

- From the normal equations, it is clear we have known quantities in \((X'X)\) and \(X'y\).
- We of course, do not know what \(\hat{\beta}\) are.
- We solve for them.
Regression

- From the normal equations, it is clear we have known quantities in \((X'X)\) and \(X'y\).
- We of course, do not know what \(\hat{\beta}\) are.
- We solve for them.
- Let’s take advantage of the inversion result: recall \(QQ^{-1} = I\) (if the inverse exists; it may not).
Regression

- From the normal equations, it is clear we have known quantities in $\text{(X}'\text{X})$ and $\text{X}'\text{y}$.
- We of course, do not know what $\hat{\beta}$ are.
- We solve for them.
- Let’s take advantage of the inversion result: recall $\text{QQ}^{-1} = \text{I}$ (if the inverse exists; it may not).
- So, consider this. Let us premultiply equation (28) by $(\text{X}'\text{X})^{-1}$:

$$
\begin{align*}
\text{(X}'\text{X})^{-1}(\text{X}'\text{X})\hat{\beta} &= \text{(X}'\text{X})^{-1}\text{X}'\text{y} \\
\text{I}\hat{\beta} &= \text{(X}'\text{X})^{-1}\text{X}'\text{y} 
\end{align*}
$$

(29)
Regression

- From the normal equations, it is clear we have known quantities in \((X'X)\) and \(X'y\).
- We of course, do not know what \(\hat{\beta}\) are.
- We solve for them.
- Let’s take advantage of the inversion result: recall \(QQ^{-1} = I\) (if the inverse exists; it may not).
- So, consider this. Let us premultiply equation (28) by \((X'X)^{-1}\):

\[
(X'X)^{-1}(X'X)\hat{\beta} = (X'X)^{-1}X'y
\]

\[
I\hat{\beta} = (X'X)^{-1}X'y
\]  

(29)

- The existence of an inverse is important! Without it, we cannot obtain the OLS estimator:

\[
\hat{\beta} = (X'X)^{-1}X'y
\]  

(30)
From the normal equations, it is clear we have known quantities in $(X'X)$ and $X'y$.

We of course, do not know what $\hat{\beta}$ are.

We solve for them.

Let's take advantage of the inversion result: recall $QQ^{-1} = I$ (if the inverse exists; it may not).

So, consider this. Let us premultiply equation (28) by $(X'X)^{-1}$:

$$(X'X)^{-1}(X'X)\hat{\beta} = (X'X)^{-1}X'y$$

$$I\hat{\beta} = (X'X)^{-1}X'y$$  \hspace{1cm} (29)

The existence of an inverse is important! Without it, we cannot obtain the OLS estimator:

$$\hat{\beta} = (X'X)^{-1}X'y$$  \hspace{1cm} (30)

This is the OLS estimator: the regression parameters are a function of the data.
Continuing with our example, let’s create \((X'X)^{-1}\):

\[
xxinverse<-\text{solve}(xprime_x); xxinverse
\]

\[
\begin{array}{ccc}
  x0 & x1 & x2 \\
  x0 & 7.8403149 & -0.6805436 & -0.6028904 \\
  x1 & -0.6805436 & 0.06676014 & 0.0474547 \\
  x2 & -0.6028904 & 0.04745470 & 0.0539258 \\
\end{array}
\]

Ok. Now let’s verify that \((X'X)^{-1}(X'X)\) is an identity matrix:

\[
> \text{solve}(xprime_x)%*%xprime_x
\]

\[
\begin{array}{ccc}
  x0 & x1 & x2 \\
  x0 & 1.000000e+00 & -8.701373e-15 & -4.009293e-14 \\
  x1 & 1.783513e-16 & 1.000000e+00 & 5.793976e-16 \\
  x2 & -2.123952e-16 & -2.896988e-16 & 1.000000e+00 \\
\end{array}
\]

Looks like an identity matrix \(I\). We can solve for \(\hat{\beta}\) by first creating \(X'y\):

\[
> xprimey=xprime %*% y; xprimey
\]

\[
\begin{array}{c}
  [,1] \\
  x0 & 299.0 \\
  x1 & 2204.0 \\
  x2 & 1387.6 \\
\end{array}
\]
The least squares estimates are obtained by $\hat{\beta} = (X'X)^{-1}X'y$:

```r
> b = xxinverse %*% xprimey; b
   [,1]
  x0  7.765380
  x1  9.504961
  x2 -0.846635
```

Are we correct?

```r
> model1 <- lm(y ~ x1 + x2); summary(model1)
```

Call:
```
lm(formula = y ~ x1 + x2)
```

Residuals:
```
            1            2            3            4            5
11.40513 -24.50416  13.99390  -0.09584  -0.79903
```

Coefficients:
```
            Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  7.76541   60.28281    0.129    0.909
   x1       9.50500    5.56271    1.709    0.230
   x2      -0.84664    4.99955   -0.169    0.881
```

Residual standard error: 21.53 on 2 degrees of freedom
Multiple R-Squared: 0.8197, Adjusted R-squared: 0.6395
F-statistic: 4.548 on 2 and 2 DF, p-value: 0.1803

Which, by the way, is how you run regression in R.
Regression

- The “hat” matrix is a useful matrix (you used in a lot in POL 213).
Regression

- The “hat” matrix is a useful matrix (you used in a lot in POL 213).
- Predicted values:

\[ \hat{y} = X(X'X)^{-1}X'y \]  

(31)
The “hat” matrix is a useful matrix (you used in a lot in POL 213).

Predicted values:

\[ \hat{y} = X(X'X)^{-1}X'y \]  

This isn’t mysterious: it is \( X \) times \( \hat{\beta} \).
Regression

- The “hat” matrix is a useful matrix (you used in a lot in POL 213).
- Predicted values:

\[ \hat{y} = X(X'X)^{-1}X'y \]  

(31)

- This isn’t mysterious: it is \( X \) times \( \hat{\beta} \).
- Let \( H = X(X'X)^{-1}X' \); we call \( H \) the “hat” matrix.
Regression

The “hat” matrix is a useful matrix (you used in a lot in POL 213).

Predicted values:

\[ \hat{y} = X(X'X)^{-1}X'y \]  \hspace{1cm} (31)

This isn’t mysterious: it is \( X \) times \( \hat{\beta} \).

Let \( H = X(X'X)^{-1}X' \); we call \( H \) the “hat” matrix.

It puts the “hat” on \( y \).
The “hat” matrix is a useful matrix (you used in a lot in POL 213).

Predicted values:

\[ \hat{y} = X(X'X)^{-1}X'y \]  

This isn’t mysterious: it is \( X \) times \( \hat{\beta} \).

Let \( H = X(X'X)^{-1}X' \); we call \( H \) the “hat” matrix.

It puts the “hat” on \( y \).

Thus, \( \hat{y} = Hy \).
Let’s compute the hat matrix:

\[
\text{hat} \leftarrow x \text{mat} \% \% (xx \text{inverse}) \% \% x \text{prime}
\]

And now, let’s get the fitted values \( \hat{y} = Hy \):

\[
> \text{yhat} \leftarrow \text{hat} \% \% \text{y}; \text{yhat}
\]

\[
\begin{array}{c}
[1,] \quad 20.59487 \\
[2,] \quad 38.50416 \\
[3,] \quad 57.00610 \\
[4,] \quad 81.09584 \\
[5,] \quad 101.79903 \\
\end{array}
\]

Now the residuals, \( y - \hat{y} \):

\[
\text{residual} \leftarrow y - \text{yhat}; \text{residual}
\]

\[
\begin{array}{c}
[1,] \quad 11.40513374 \\
[2,] \quad -24.50416307 \\
[3,] \quad 13.99389560 \\
[4,] \quad -0.09583693 \\
[5,] \quad -0.79902934 \\
\end{array}
\]
Since $\sum e_i^2 = SSE_{Error}$, we can compute this quantity using matrix operations:

```r
> res<-cbind(residual)
> eprime<-t(res); eprime
[1,] 11.40513 -24.50416 13.99390 -0.09583693 -0.7990293

> eprimee<-eprime %*% res; eprimee
     [,1]
[1,] 927.0078
```

... or the old-fashioned way:

```r
> SSE_{Error}=sum(residual^2); SSE_{Error}
[1] 927.0078
```

Other quantities? How about the $R^2$:

```r
> meany<-cbind(rep(c(mean(y)), 5))
> explained<-yhat-meany; explained
     [,1]
[1,] -39.205134
[2,] -21.295837
[3,]  -2.793896
[4,]  21.295837
[5,]  41.999029
> SS_{Regress}<-(explained^2)
> r2=SS_{Regress}/(eprimee+SS_{Regress}); r2
     [,1]
[1,] 0.8197465
```
This is what we want. So now that we have the variance components, we can rule the world.

Mean Square Error:
> MSE<-eprimee/(2); MSE
  [,1]        
[1,] 463.5039

Root Mean Square Error (s.e. of the estimate):
MSE^(.5)
  [,1]        
[1,] 21.52914

Note that the variance-covariance matrix of $\beta = MSE(X'X)^{-1}$; ::

> varcov<-463.5039 * xxinverse; varcov
   x0    x1    x2
x0 363.0165 -315.4346 -279.4421
x1 -315.4346  30.9436  21.9954
x2 -279.4421  21.9954  24.9948
>
> #Now we can obtain the standard errors:
>
> s.e._int<-sqrt(varcov[1,1])
> s.e._b1<-sqrt(varcov[2,2])
> s.e._b2<-sqrt(varcov[3,3])
> s.e<-cbind(s.e._int, s.e._b1, s.e._b2); s.e.
  s.e._int  s.e._b1  s.e._b2
[1,] 60.2828  5.562696  4.999482
Recall:

```r
> model1<-lm(y~x1+ x2); summary(model1)
```

Call:
```
  lm(formula = y ~ x1 + x2)
```

Residuals:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>11.40513</td>
<td>-24.50416</td>
<td>13.99390</td>
<td>-0.09584</td>
<td>-0.79903</td>
</tr>
</tbody>
</table>

Coefficients:

|         | Estimate | Std. Error | t value | Pr(>|t|) |
|---------|----------|------------|---------|----------|
| (Intercept) | 7.7654   | 60.2828    | 0.129   | 0.909    |
| x1       | 9.5050   | 5.5627     | 1.709   | 0.230    |
| x2       | -0.8466  | 4.9995     | -0.169  | 0.881    |

Residual standard error: 21.53 on 2 degrees of freedom
Multiple R-Squared: 0.8197, Adjusted R-squared: 0.6395
F-statistic: 4.548 on 2 and 2 DF, p-value: 0.1803

We’ve basically just replicated everything here (or could replicate everything here).