Statistics Review Part 3

Hypothesis Tests,

Regression

The Importance of Sampling Distributions

- Why all the fuss about sampling distributions?
  - Because they are fundamental to **hypothesis testing**.
- Remember that our goal is to learn about the population distribution.
  - In practice, we are interested in things like the population mean, population variance, some population conditional mean, etc.
  - We estimate these quantities in a random sample taken from the population.
  - This is done with sample estimators like the sample mean, the sample variance, or some sample conditional mean.
- Knowing the sampling distribution of our sample estimator (e.g., the sampling distribution of the sample mean), gives us a way to assess whether particular values of population quantities (e.g., the population mean) are likely or unlikely.
- E.g., suppose we calculate a sample mean of 5. If the true population mean is 6, is 5 a “likely” or “unlikely” occurrence?

Hypothesis Testing

- We use hypothesis tests to evaluate claims like:
  - the population mean is 5
  - the population variance is 16
  - some population conditional mean is 3
- When we know the sampling distribution of a sample statistic, we can evaluate whether its observed value in the sample is "likely" **when the above claim is true**.
- We formalize this with two **hypotheses.**
- For now, we'll focus on the case of hypotheses about the population mean, but we can generalize the approach to **any** population quantity.

Null and alternative hypotheses

- Suppose we're interested in evaluating a specific claim about the population mean. For instance:
  - "the population mean is 5"
  - "the population mean is positive"
- We call the claim that we want to evaluate the **null hypothesis**, and denote it H_0.
  - $H_0 : \mu = 5$
  - $H_0 : \mu > 0$
- We compare the null hypothesis to the **alternative hypothesis**, which holds **when the null is false**. We will denote it H_1.
  - $H_1 : \mu \neq 5$ (a "two-sided" alternative hypothesis)
  - $H_1 : \mu \leq 0$ (a "one-sided" alternative hypothesis)

How tests about the population mean work

- Step 1: Specify the null and alternative hypotheses.
- Step 2a: Compute the sample mean and variance
- Step 2b: Use the estimates to construct a new statistic, called a **test statistic**, that has a **known sampling distribution *when the null hypothesis is true*** ("under the null")
  - the sampling distribution of the test statistic depends on the sampling distribution of the sample mean and variance
- Step 3: Evaluate whether the calculated value of the test statistic is "likely" when the null hypothesis is true.
  - We **reject** the null hypothesis if the value of the test statistic is "unlikely"
  - We **do not reject** the null hypothesis if the value of the test statistic is "likely"
  - (Note: thanks to Popper, we never "accept" the null hypothesis)

Example: the t-test

- Suppose we have a random sample of *n* observations from a $N(\mu,\sigma^2)$ distribution.
- Suppose we're interested in testing the null hypothesis:

  $$H_0 : \mu = \mu_0$$

  against the alternative hypothesis:

  $$H_1 : \mu \neq \mu_0$$

- A natural place to start is by estimating the sample mean, $\overline{X}$
- We know that **if the null hypothesis is true**, then the sampling distribution of $\overline{X}$ is normal with mean μ_0 and variance σ^2/n.
  - We say: $\overline{X} \sim N(\mu_0,\sigma^2/n)$ **under the null**
  - (draw a picture)

Example: the t-test (continued)

- Because $\overline{X}$ ~ N(μ_0,σ^2/*n)* under the null, we know that

$$Z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}} \sim N(0,1) \text{ under the null}$$

  (recall we can transform any normally distributed RV to have a standard normal distribution by subtracting off its mean and dividing by its standard deviation)

- If we knew σ^2, we could compute *Z*, and this would be our test statistic:
  - If *Z* is “far” from zero, it is unlikely that the null hypothesis is true, and we would reject it.
  - If *Z* is “close” to zero, it is likely that the null hypothesis true, and we would not reject it.
  - Why *Z*? Because we can look up its critical values in a table.
- Problems with this approach:
  - we don’t know σ^2
  - how do we quantify “close” and “far”?

Example: the t-test (continued)

- Since we don't know σ^2, why not estimate it? We know that the sample variance s^2 is an unbiased estimator of σ^2.
- Unfortunately,

$$Q = \frac{\overline{X} - \mu_0}{s / \sqrt{n}} \text{ does } \textbf{not} \text{ have a } N(0,1) \text{ distribution under the null}$$

- Some facts about sampling from a N(μ,σ^2):
  - Fact 1: $(n - 1)s^2/ \sigma^2 \sim \chi^2_{n\text{-}1}$
  - Fact 2: s^2 is independent of $\overline{X}$
- We know already that if $Z \sim$ N(0,1) and $W \sim \chi^2_v$ and Z and W are independent, then

$$\boxed{T = \frac{Z}{\sqrt{W / v}} \sim t_v}$$

- We can put all this together to compute a test statistic that has a t sampling distribution with n-1 degrees of freedom.

Example: the t-test (concluded)

- Putting together the things we know, under the null :

$$T = \frac{Z}{\sqrt{\left[(n-1)s^2/\sigma^2\right]/(n-1)}} = \frac{\left(\bar{X}-\mu_0\right)/\left(\sigma/\sqrt{n}\right)}{\sqrt{\left[(n-1)s^2/\sigma^2\right]/(n-1)}} \sim t_{n-1}$$

- This is a bit of a mess, and it still involves an unknown quantity (σ^2). But we can simplify it with a bit of algebra:

$$T = \frac{\left(\bar{X}-\mu_0\right)/\left(\sigma/\sqrt{n}\right)}{\sqrt{\left[(n-1)s^2/\sigma^2\right]/(n-1)}} = \frac{\left(\bar{X}-\mu_0\right)/\left(\sigma/\sqrt{n}\right)}{\sqrt{s^2/\sigma^2}}$$

$$= \frac{\left(\bar{X}-\mu_0\right)/\left(\sigma/\sqrt{n}\right)}{s/\sigma} = \frac{\left(\bar{X}-\mu_0\right)\sigma}{s\left(\sigma/\sqrt{n}\right)} = \frac{\left(\bar{X}-\mu_0\right)}{s/\sqrt{n}} \sim t_{n-1}$$

- Notice that this test statistic has two crucial properties:
  - it doesn’t contain any unknowns (so we can compute it)
  - we know its sampling distribution (so we can look in a table and see if a particular value is “likely” or “unlikely” under the null)

This is how **ALL** hypothesis testing works

1. We form the null and alternative hypotheses that we care about.
2. Then we use the sample data to construct a test statistic that has a known sampling distribution when the null hypothesis is true.
3. Then we decide whether or not to reject the null on the basis of whether the value of the test statistic that we observe is “likely” or “unlikely” under the null hypothesis.

- The **right way** to think about hypothesis tests:
  - A hypothesis test is a **rule for using the data to decide whether or not to reject the null hypothesis.**

A note about Normality

- When we derived the sampling distribution for our *t* statistic, we relied on some facts that are **only** true when sampling from a normal distribution.
  - This gives us a normal sampling distribution for the sample mean under the null, independence of the sample mean and variance, and the Chi-square result for s^2.
- What do we do if we're not sampling from a normal distribution? (the usual case)
  - Usually, we rely on an **asymptotic approximation**.
- As the sample size gets "big" (technically, as $n \to \infty$), then s^2 gets very close to σ^2 (it's consistent). Thus our *T* statistic gets very close to our *Z* statistic:

$$\boxed{\text{as } n \to \infty,\ T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \to \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = Z}$$

- Furthermore, we know from the central limit theorem that as $n \to \infty$, under the null hypothesis the sampling distribution of *Z* is approximately N(0,1)
  NO MATTER WHAT THE POPULATION DISTRIBUTION IS!
- So if our sample is "big enough" (**how big**?) we can compute the *T* statistic as usual, and use values from the standard normal distribution to decide whether a specific value is "likely" or "unlikely" under the null hypothesis.

How do we know if a particular value of the test statistic is “likely”?

- When we know the sampling distribution of the test statistic, we know the probability that the value of the test statistic will fall in a given interval **when the null hypothesis is true**.
  - Just the area under the pdf of the test statistic’s sampling distribution.
  - If T is our test statistic, for any α we can find L,U such that

    $$\Pr[L \leq T \leq U] = 1 - \alpha \quad \text{when } H_0 \text{ is true}$$
- So we can define a range of “acceptable” values of the test statistic.
  - e.g., we can define an interval $[L,U]$ such that **if the null hypothesis is true**, the test statistic falls in the interval with probability 0.95 (α=0.05)
  - that is, if the null is true and we draw 100 random samples, the test statistic will fall in this interval about 95 times.
  - if we observe a value of the test statistic outside this interval, we know it is “unlikely” that the null is true (e.g., it would only happen in 5% of random samples) and we can therefore comfortably reject the null.
- We call *L* and *U* **critical values at the 100α% significance level,** and we can look them up in a table.

Making mistakes: Type I and Type II errors

- When testing a hypothesis, there’s **always** the possibility that we make a mistake.
- There are two kinds of mistakes:
  - **Type I error**: we erroneously reject the null when it’s true.
  - **Type II error**: we fail to reject the null when it’s false
- We call the probability of making a Type I error the **significance level** of the test, and denote it α.
- We use β to denote the probability of making a Type II error.
  - We call $1 - \beta$ the **power** of a test. It is the probability that we correctly reject the null when it is false.
- We usually choose a significance level α that we’re comfortable with (0.05 is most common), and look for a powerful test (small β)
- There’s a tradeoff: as α gets smaller, β must get bigger.
- (draw a picture)

An example

- Suppose you survey 121 randomly selected Canadians about the number of times they went to the movies last year.
- Using the survey data, you calculate $\overline{X} = 7.5, \quad s^2 = 75$
- You want to test the hypothesis that the population mean of movie attendance is 9:

$$H_0 : \mu = 9 \qquad H_1 : \mu \neq 9$$

- You figure movie attendance is probably normally distributed in the population, and construct the T statistic:

$$T = \frac{\overline{X} - \mu_0}{s / \sqrt{n}} = \frac{7.5 - 9}{\sqrt{75} / \sqrt{121}} = -1.90526$$

- You know that $T \sim t_{120}$ if H_0 is true. In a table of critical values for the t_{120} distribution, you find that the critical value at the 10% level of significance (α = 0.1) is 1.658; at the 5% level of significance (α = 0.05) the critical value is 1.980.
- This means that if H_0 is true,

$$\Pr[-1.658 \leq T \leq 1.658] = 0.90$$
$$\Pr[-1.980 \leq T \leq 1.980] = 0.95$$

- Therefore t = -1.90526 is quite unlikely. You would reject H_0 at the 10% level of significance, but would not reject it at the 5% level of significance.

p-values

- There’s another (related) way to decide whether or not to reject the null hypothesis.
- Suppose we construct a test statistic called *W* that has a known sampling distribution when H_0 is true. Suppose that in our sample, the value we compute for the test statistic is *w*.
- We can ask “what is the probability of observing a value of the statistic *W* as big as *w* **when the null hypothesis is true**?” i.e.,
  - $\Pr[-w \leq W \leq w] = 1 - p^*$ (for a two-sided alternative)
  - $\Pr[W \leq w] = 1 - p^*$ (for a one-sided alternative)
- The probability p^* is called a **p-value**. It is a tail probability of the sampling distribution of *W*. It is the probability of observing a value of *W* that is more extreme (unusual) than *w*.
- It is also the probability of making a type I error if we reject the null.
- If the p-value is “small” (say ≤ 0.05) we can confidently reject H_0.
- Computers routinely report p-values for common hypothesis tests, so you can save yourself some time by looking at the p-value rather than looking up critical values in a table.
- The p-value for the example above is between 5% and 10%

Interval Estimation

- We’re done talking about hypothesis testing for now – but it will come up again soon in the context of linear regression.
- We talked earlier about estimators – statistics that we use to estimate a population quantity.
- The examples we saw (the sample mean, sample variance, sample covariance, etc.) are all called **point estimators** because they give us a single value for the population quantity.
- An alternative to a point estimator is an **interval estimator.**
- This is an interval that contains a population quantity with a known probability.
- An interval estimator of a population quantity Q takes the form $[L,U]$, where L and U are functions of the data (they’re statistics).
- We use the interval estimator $[L,U]$ to make statements like:

  $\Pr[L \leq Q \leq U] = 1 - \alpha$ (look familiar yet?)

Example: Confidence Interval for the Population Mean

- A 95% confidence interval for the population mean μ is an interval $[L,U]$ such that:

$$\Pr[L \leq \mu \leq U] = 0.95$$

- How do we find the interval $[L,U]$ such that this is true?
- An illustrative (but impossible) way:

1. Pick a random value μ_1 and construct the T statistic to test $H_0 : \mu = \mu_1$ vs. $H_1 : \mu \neq \mu_1$.
2. If we reject H_0, then μ_1 **is not** in the interval. If we do not reject H_0, then μ_1 **is** in the interval.
3. Pick another value μ_2 and repeat.
4. Do this for all possible values of μ (this is why it's impossible).

- Thankfully, there's an easier way.

The 95% confidence interval for μ

- The easier way is to make use of the sampling distribution of our *T* statistic. We know that when sampling from the normal distribution,

$$T = \frac{\overline{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

- So we can **always** look up the critical value $t_{n\text{-}1,\alpha/2}$ such that:

$$\Pr[-\, t_{n\text{-}1,\alpha/2} \le T \le t_{n\text{-}1,\alpha/2}] = 1 - \alpha$$

- For a 95% confidence interval, we have

$$\Pr[-\, t_{n\text{-}1,0.025} \le T \le t_{n\text{-}1,0.025}] = 0.95$$

- Now we just plug in the formula for our *T* statistic, and rearrange things as necessary (next slide)

Confidence Interval Algebra

$$
\begin{aligned}
0.95 &= \Pr\left[-t_{n\text{-}1,0.025} \leq T \leq t_{n\text{-}1,0.025}\right] \\
&= \Pr\left[-t_{n\text{-}1,0.025} \leq \frac{\bar{X}-\mu}{s/\sqrt{n}} \leq t_{n\text{-}1,0.025}\right] \\
&= \Pr\left[-\left(\frac{s}{\sqrt{n}}\right)t_{n\text{-}1,0.025} \leq \bar{X}-\mu \leq \left(\frac{s}{\sqrt{n}}\right)t_{n\text{-}1,0.025}\right] \\
&= \Pr\left[-\left(\frac{s}{\sqrt{n}}\right)t_{n\text{-}1,0.025} - \bar{X} \leq -\mu \leq \left(\frac{s}{\sqrt{n}}\right)t_{n\text{-}1,0.025} - \bar{X}\right] \\
&= \Pr\left[\bar{X}-\left(\frac{s}{\sqrt{n}}\right)t_{n\text{-}1,0.025} \leq \mu \leq \bar{X}+\left(\frac{s}{\sqrt{n}}\right)t_{n\text{-}1,0.025}\right]
\end{aligned}
$$

The Last Word on Confidence Intervals (for now)

- So now we know how to build a 95% confidence interval for μ when sampling from the normal distribution.
- It's easy to generalize this to 90% or 99% etc. confidence intervals – just change the critical value you use.
- When we're not comfortable assuming that we're sampling from the normal distribution, we can replace critical values from the *t* distribution with critical values from the standard normal distribution if the sample is "big enough"
- We can build similar intervals for other population quantities – the idea is always the same. We just need a sample quantity (usually a test statistic) whose sampling distribution we know & we can use it to build the interval.
- Typically, a confidence interval for a population quantity *Q* looks like

  $$[L,U] = [q - \textit{something}, q + \textit{something}]$$

  where *q* is a point estimator of *Q*, and *something* depends on the variance of the sampling distribution of *q*.

Moving on

- To this point we've focused on a single random variable, *X*.
- We've talked about population quantities (e.g., μ and σ^2).
- We've discussed how to compute sample statistics to estimate population quantities, the properties of estimators (e.g., bias and efficiency), and how to test hypotheses about the population.
- Things get much more interesting when we consider two or more random variables.
  - we care about the relationship between variables
  - we can use one or more variables to predict a variable of interest, etc.
- We can get a lot of mileage out of studying the conditional expectation of a variable of interest (*Y*) given another variable (or group of variables) *X*. That is, studying E($Y|X$).
- Recall that another name for E($Y|X$) is the **regression function**.
- We'll spend the rest of the semester talking about regression analysis, which is a very powerful tool for analyzing economic data.
- Regression analysis is based on E($Y|X$).

What is Regression Analysis?

- Regression analysis is a very common statistical/econometric technique
- We use it to measure/explain relationships between economic variables
- Example: casual observation will reveal that more educated individuals tend to have higher incomes.
  - regression methods can be used to **measure the rate of return** of an extra year of education
  - or, use regression methods to **estimate the relationship** between income and education, gender, labour market experience, etc.
- Example: economic theory tells us that if the price of a good increases, individuals will consume less of it.
  - that is, demand curves slope down
  - but economic theory doesn’t predict **how big** the change in consumption will be for a given price change
  - we can use regression analysis to **measure how much** individuals reduce their consumption in response to a price increase (i.e., we can estimate the elasticity of demand)

The Regression Model

- Regression is a tool to conveniently learn about *conditional means*.
- The goal of regression analysis is to **explain** the value of one variable of interest (the **dependent variable**) as a function of the values of other variables (the **independent** or **explanatory variables**)
  - Usually, the dependent variable is denoted Y
  - The independent variables are X_1, X_2, X_3 etc.
  - Sometimes we say we want to explain "movement" in Y as a function of "movement" in the X variables. That is, how much does Y change when the X variables change? In this context, a better word for "movement" is **variation**.
- We use an equation (sometimes more than one) to specify the relationship between Y and the X variables.
- This equation is called the **regression model**:

$$\mathrm{E}(Y|X_1,X_2,X_3) = f(X_1, X_2, X_3)$$

- Example: Y is income, X_1 is years of education, X_2 is gender, X_3 is years of labour market experience, and f is some function ...
- Note: we are not saying that the X variables **cause** Y

Simple Linear Regression

- The simplest example of a regression model is the case where the regression function *f* is a line, and where there is only one *X*

$$E[Y|X] = \beta_0 + \beta_1 X$$

- This specification of the regression function says that the dependent variable *Y* is a linear function of the independent variable *X*
- This is just the equation of a line
- We call β_0 and β_1 the regression **coefficients**
- β_0 is called the **intercept** or **constant term**. It tells us the value of *Y* when *X* is zero
- β_1 is the **slope coefficient**. It measures the amount that *Y* changes for a unit change in *X*--it is the slope of the line relating *X* and *Y*:

$$\frac{dY}{dX} = \beta_1$$

sometimes we call β_1 the **marginal effect** of *X* on *Y*.

About Linearity

- There are **two** kinds of linearity present in the regression model

  $$Y = \beta_0 + \beta_1 X$$

  from the previous slide.
- This regression function is **linear in *X*.**
  - *counter-example:* $Y = \beta_0 + \beta_1 X^2$
- This regression function is **linear in the coefficients** β_0 and β_1
  - *counter-example:* $Y = \beta_0 + X^{\beta}$
- In general, neither kind of linearity is necessary.
- However, we will focus our attention mostly on what econometricians call the **linear regression model**.
- The linear regression model **requires** linearity in the coefficients, but **not** linearity in *X*.
  - When we say “linear regression model” we mean a model that is *linear in the coefficients*.

The Stochastic Error Term

- Econometricians recognize that the regression function is never an **exact** representation of the relationship between dependent and independent variables.
    - e.g., there is no exact relationship between income (*Y*) and education, gender, etc., because of things like luck
- There is **always** some variation in *Y* that cannot be explained by the model.
- There are many possible reasons: there might be “important” explanatory variables that we leave out of the model; we might have the wrong functional form (*f*), variables might be measured with error, or maybe there’s just some randomness in outcomes.
- These are all sources of **error**. To reflect these kinds of error, we include a **stochastic (random) error term** in the model.
- The error term reflects all the variation in *Y* that cannot be explained by *X*.
- Usually, we use epsilon (ε) to represent the error term.

More About the Error Term

- Add an error term, and our simple linear regression model is:
  $$Y = \beta_0 + \beta_1 X + \varepsilon$$
- It is helpful to think of the model as having two components:
  1. a *deterministic* (non-random) component $\beta_0 + \beta_1 X$
  2. a *stochastic* (random) component ε
- Basically, we are decomposing Y into the part that we can explain using X (i.e., $\beta_0 + \beta_1 X$) and the part that we cannot explain using X (i.e., the error ε)
- The **right way** to think about it:
  $\beta_0 + \beta_1 X$ is the **conditional mean of *Y* given *X*.** That is,
  $$Y = \mathrm{E}(Y|X) + \varepsilon$$
  where $\mathrm{E}(Y|X) = \beta_0 + \beta_1 X$
- Remember "regression function" **means** $\mathrm{E}(Y|X)$
- This gives us another way to think about errors: $\varepsilon = Y - \mathrm{E}(Y|X)$
- (draw some pictures)

An example

- Think about starting salaries for new university graduates (Y).
- There is **a lot** of variation in starting salaries between individuals.
- Some of this variation is predictable:
  - starting salary depends on university, field of study, industry, occupation/title, firm size, etc.
  - Call all of this X
  - The predictable part of starting salary goes into the deterministic component of the regression: $\mathrm{E}(Y|X)$.
    - We don’t need to impose that X enters linearly, but we will require $\mathrm{E}(Y|X)$ to be linear in the βs.
    - We choose the specific **functional form** of $\mathrm{E}(Y|X)$ when we build the model.
- Much of the variation in starting salary is unpredictable:
  - starting salary also depends on luck, nepotism, interview skill, etc.
  - We can’t measure these things, so we can’t include them in $\mathrm{E}(Y|X)$.
  - The unpredictable part ends up in the error term, ε.

Even Simpler Regression

- Suppose we have a linear regression model with one independent variable and NO INTERCEPT: $\boxed{Y_i = \beta X_i + \varepsilon_i}$

- Suppose also that $\boxed{E[\varepsilon_i] = 0 \; and \; E[(\varepsilon_i)^2] = \sigma^2}$
  For all *i*.

- Now, define an estimator as the number $\hat{\beta}$ that minimises the sum of the squared prediction error
  $\boxed{e_i = Y_i - \hat{\beta} X_i}$

- $\underset{\hat{\beta}}{\text{Min}}$ $\boxed{\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{\beta} X_i\right)^2 = \sum_{i=1}^{n} Y_i^2 - \sum_{i=1}^{n} \left(2Y_i \hat{\beta} X_i\right) + \sum_{i=1}^{n} \left(\hat{\beta} X_i\right)^2}$

Minimisation

- The squared *Y* leading term doesn't have $\hat{\beta}$
- Min $\underset{\hat{\beta}}{}$ $\boxed{-\sum_{i=1}^{n}\left(2Y_i\hat{\beta}X_i\right)+\sum_{i=1}^{n}\left(\hat{\beta}X_i\right)^2}$
- First-Order Condition

$$\boxed{\begin{gathered}-2\sum_{i=1}^{n}\left(X_iY_i\right)+2\hat{\beta}\sum_{i=1}^{n}\left(X_i^{\,2}\right)=0\\ -\sum_{i=1}^{n}\left(X_iY_i\right)+\hat{\beta}\sum_{i=1}^{n}\left(X_i^{\,2}\right)=0\\ \hat{\beta}\sum_{i=1}^{n}\left(X_i^{\,2}\right)=\sum_{i=1}^{n}(Y_iX_i)0\\ \hat{\beta}=\frac{\sum_{i=1}^{n}X_iY_i}{\sum_{i=1}^{n}\left(X_i^{\,2}\right)}\end{gathered}}$$

OLS Coefficients are Sample Means

- The estimated coefficient is a weighted average of the *Y*’s:

$$\boxed{\begin{aligned} \hat{\beta} &= \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} \left(X_i^{\,2} \right)} = \sum_{i=1}^{n} w_i Y_i \\ w_i &= \frac{X_i}{\sum_{i=1}^{n} \left(X_i^{\,2} \right)} \end{aligned}}$$

- It is a function of the data (a special kind of sample mean), and so it is a *statistic*.
- It can be used to estimate something we are interested in: the population value of β
- Since it is a statistic, it has a sampling distribution that we can evaluate for bias and variance.

Bias

- Pretend *X* is not random. Remember assumptions from above:

$$\boxed{Y_i = \beta X_i + \varepsilon_i}$$

- $\boxed{E[\varepsilon_i] = 0 \; and \; E[\left(\varepsilon_i\right)^2] = \sigma^2}$

- Substitute into the estimator and take an expectation:

$$\boxed{\begin{aligned} E\left[\hat{\beta}\right] &= E\left[\frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n}\left(X_i{}^2\right)}\right] = E\left[\frac{\sum_{i=1}^{n} X_i\left(\beta X_i + \varepsilon_i\right)}{\sum_{i=1}^{n}\left(X_i{}^2\right)}\right] \\ &= \beta E\left[\frac{\sum_{i=1}^{n} X_i\left(X_i\right)}{\sum_{i=1}^{n}\left(X_i{}^2\right)}\right] + E\left[\frac{\sum_{i=1}^{n} X_i \varepsilon_i}{\sum_{i=1}^{n}\left(X_i{}^2\right)}\right] = \beta + 0 = \beta \end{aligned}}$$

Variance

- $$V\left[\hat{\beta}\right] = E\left[\left(\hat{\beta} - E\left[\hat{\beta}\right]\right)^2\right] = E\left[\left(\frac{\sum_{i=1}^{n} X_i \varepsilon_i}{\sum_{i=1}^{n}\left(X_i{}^2\right)}\right)^2\right] = \frac{1}{\left(\sum_{i=1}^{n}\left(X_i{}^2\right)\right)^2} E\left[\left(\sum_{i=1}^{n} X_i \varepsilon_i\right)^2\right]$$

$$= \frac{1}{\left(\sum_{i=1}^{n}\left(X_i{}^2\right)\right)^2} E\left[X_1 X_1 \varepsilon_1 \varepsilon_1 + X_1 X_2 \varepsilon_1 \varepsilon_2 + \ldots + X_{n-1} X_n \varepsilon_{n-1} \varepsilon_n + X_n X_n \varepsilon_n \varepsilon_n\right]$$

$$= \frac{1}{\left(\sum_{i=1}^{n}\left(X_i{}^2\right)\right)^2} E\left[\sum_{i=1}^{n}\left(X_i\right)^2\left(\varepsilon_i\right)^2\right] = \frac{\sum_{i=1}^{n}\left(X_i\right)^2}{\left(\sum_{i=1}^{n}\left(X_i{}^2\right)\right)^2} E\left[\left(\varepsilon_i\right)^2\right] = \frac{1}{\sum_{i=1}^{n}\left(X_i{}^2\right)} \sigma^2$$

Inference

- You have everything you need to do make probability statements about OLS regression coefficients when there's just 1 variable (and no intercept).
- It is basically the same when you have more variables:
  - The OLS coefficient is a weighted sum of *Y*'s
  - It is a statistic with a sampling distribution
  - The sampling distribution depends on the data, and you can thus use the data to make statements about the population parameter.

Extended Notation

- We need to extend our notation of the regression function to reflect the number of observations.
- As usual, we'll work with an iid random sample of *n* observations.
- If we use the subscript *i* to indicate a particular observation in our sample, our regression function with one independent variable is:

$$\boxed{Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad \text{for } i = 1,2,\ldots,n}$$

- So really we have *n* equations (one for each observation):

$$\boxed{\begin{aligned} Y_1 &= \beta_0 + \beta_1 X_1 + \varepsilon_1 \\ Y_2 &= \beta_0 + \beta_1 X_2 + \varepsilon_2 \\ &\vdots \\ Y_n &= \beta_0 + \beta_1 X_n + \varepsilon_n \end{aligned}}$$

Notice that the coefficients β_0 and β_1 are **the same** in each equation. The only thing that varies across equations is the data (Y_i, X_i) and the error ε_i.

Extending the Notation Further

- If we have more (say *k*) independent variables, then we need to extend our notation further.
- We could use a different letter for each variable (i.e., *X*, *Z*, *W*, etc.) but instead we usually just introduce another subscript on the *X*.
- So now we have two subscripts: one for the variable number (first subscript) and one for the observation number (second subscript).

$$\boxed{Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + \varepsilon_i}$$

- What do the regression coefficients measure now? They are **partial derivatives**. That is,

$$\boxed{\beta_1 = \frac{\partial Y_i}{\partial X_{1i}} \quad \beta_2 = \frac{\partial Y_i}{\partial X_{2i}} \quad \cdots \quad \beta_k = \frac{\partial Y_i}{\partial X_{ki}}}$$

So, β_1 measures the effect on Y_i of a one unit increase in X_{1i} **holding all the other independent variables X_{2i}, X_{3i}, ... , X_{ki} constant.**

What is Known, What is Unknown, and What is Assumed

- It is useful to summarize what is known, what is unknown, and what is hypothesized.
- **Known:** Y_i and X_{1i} , X_{2i} , ... , X_{ki} (the data)
- **Unknown:** β_0 , β_1 , β_2 , ... , β_k and ε_i (the coefficients and errors)
- **Hypothesized:** the form of the regression function, e.g.,
$$E(Y_i \mid X_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_k X_{ki}$$
- We use the observed data to learn about the unknowns (coefficients and errors), and then we can test the hypothesized form of the regression function.
- We can hope to learn a lot about the βs because they are the same for each observation.
- We can’t hope to learn much about the ε_i because there is only one observation on each of them.

Estimated Regression Coefficients

- Think of the regression function we've developed to this point as the **population regression function**.
- As always in econometrics, we collect a sample of data to learn about the population.
- We don't know the (population) regression coefficients (*β*), so we estimate them. We'll discuss the details of how to do this next day.
- For now, we'll just introduce a notation for the estimated coefficients. The estimated coefficients are:

$$\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_k$$

- The estimated coefficients are sample statistics that we can compute from our data.
- Because they are sample statistics, they are RVs, have sampling distributions, etc., just like all sample statistics.

Predicted Values and Residuals

- Once we have estimated the regression coefficients, we can calculate the **predicted value** of Y_i.
- It is a sample estimate of the conditional expectation of Y_i given all the Xs*:*

$$\boxed{\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_k X_{ki}}$$

- It is our “best guess” of the value of Y_i given the value of the Xs.
- Predicted values lie on the estimated regression line.
- Of course, Y_i and its predicted value are rarely equal.
- We call the difference between Y_i and its predicted value the **residual** e_i*:*

$$\boxed{e_i = Y_i - \hat{Y}_i}$$

- Residuals are the sample counterpart to the (unknown) errors

$$\varepsilon_i = Y_i - \mathrm{E}(Y_i|X_i).$$

- We can write the **estimated regression function** as:

$$\boxed{Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_k X_{ki} + e_i}$$

- (draw a picture – true & estimated regression lines, residuals, and errors)

What we do with regression analysis

- We use regression models for lots of different things.
- Sometimes we care most about predicted values & use the estimated regression to predict (forecast) things.
  - e.g., estimate a regression of stock prices (Y_i) on “leading indicators” (unemployment rate, cpi, etc.) to forecast future stock prices.
- Sometimes we care most about the coefficient values & use the estimated regression to develop policy, etc.
  - e.g., estimate a regression of labour earnings on years of education, experience, gender, etc. (“the kitchen sink”)
  - the estimated coefficient on years of education gives an estimate of the rate of return to an extra year of education *ceteris paribus*
  - the estimated coefficient on gender gives an estimate of the male-female wage differential *ceteris paribus* (see lecture 1)
  - both of these are important for designing government policy