Notes on Econometrics I

Grace McCormack

April 28, 2019

Contents

1 Overview

This set of notes is intended to supplement the typical first semester of econometrics taken by PhD students in public policy, economics, and other related fields. It was developed specifically for the first year econometrics sequence at the Harvard Kennedy School of Govenrment, which is designed to provide students with tools necessary for economics and political science research related to policy design. In this vein, I wish us to think of econometrics as a means of using data to understand something about the true nature of the world. The organizing framework for these notes can be seen below. I will be returning to this framework throughout the notes.

1.1 Introduction to a general econometrician framework

1.) We start with a *Population Relationship* or *Population Data-Generating Process (DGP)*, which we can think about as some "law of nature" that is true about the world. The DGP is defined by some population parameter θ.

- **parameter** - a population value or characteristic of the Data-Generating-Process, for example, the mean a distribution or someone's marginal utility of consumption. In this set of notes, I will often use θ to denote a population parameter. The population parameter is what generates data and is what we want to estimate using statistics or econometrics

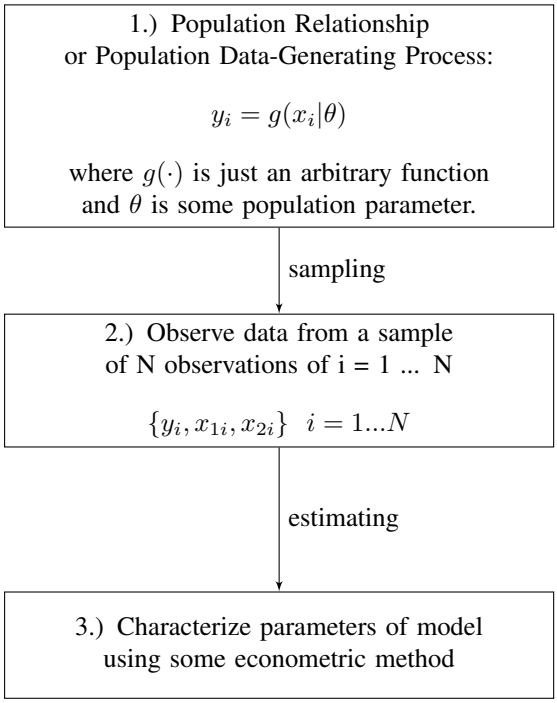
The DGP can be something simple, like the density of a normal distribution in which case θ might be the mean and standard deviation of the distribution. It could also be something quite complicated like the causal effect of education on income, in which case θ might be the financial return to each additional year of education.

2.) This DGP will produce some data from which we will be able to observe a sample of N observations. For example, if the DGP is the normal distribution, we could have a sample of N normally distributed variables. If the DGP is the causal effect of education on income, we could have a sample of N people with information on incomes and education.

3.) We wish to use our data to understand the true population parameter θ. We can characterize the parameter a myriad of ways depending on the context:

- **posterior distribution** - the probability distribution of the parameter θ based on the data that we observed (y, x) and some prior belief of the distribution of the parameter $f(\theta)$. This is what we will learn to be called a *Bayesian* approach.
- **hypothesis test** - we can use our data to see if we can reject various hypothesis about our data (for example, a hypothesis may be that the mean of a distribution is 7 or that education has no effect on income)
- **estimator** - our "best guess" of what the population parameter value is, for example a sample mean or an estimated OLS coefficient. In this set of notes, I will use a "^" to denote an estimator. While the estimator will often be a single value (a so-called "point estimate"), we also typically have to characterize how certain we are that this estimator accurately captures the population parameter, typically with a confidence interval.

We will return to this framework more throughout these notes.



1.2 A rough taxonomy of econometric analyses

Before we get started on the nitty gritties, I would like to take a moment to note how different types of econometric analyses fit broadly into this framework. Unlike microeconomics, which is taught rather similarly across most first year PhD programs, there is some degree of variation in the typical econometric sequence. You might be uncertain about what type of econometric tools that you should be learning or exactly what your choice set is to begin with. I will categorize three broad areas that most econometric courses will fall into (note that this list is not a universally acknowledged taxonomy, but I find it a useful heuristic):

1. Reduced form estimation – This is the type of econometrics that is most often used for Labor Economics and Public Economics. This approach entails linear regression to recover some causal effect of X on Y. It is also usef for "sufficient statistics" approaches. This is likely the type of econometrics that you encountered in your undergraduate courses.

2. Structural estimation – This type of econometrics is much more common in Industrial Organization. This approach requires explicit modeling of the utility function or production function to recover parameters like an individual's price elasticity or risk aversion or a firm's marginal cost of production. In our framework above, we can think of it as requiring the $g(x_i|\theta)$ to be a utility function or some "micro-founded" data-generating process. While often more complicated than reduced form approaches, this approach is useful for modeling counterfactuals – that is, estimating what would happen if we changed something about the world.

3. Machine learning – This is a relatively new tool for economists that is entirely focused on making predictions. That is, unlike reduced form or structural approaches, machine learning is less concerned about recovering the causal impact of X on Y and more just about learning how to predict Y. It typically involves large datasets. In our framework, we may think of machine learning as focusing on estimating $\hat{y}$ and less on $\hat{\theta}$

Don't worry if these distinctions remain somewhat unclear after this brief desciption. The differences will become more clear in taking field courses, attending seminars, and, of course, reading papers if not in introdutory classes alone. While these notes should be useful for all three of these broad categories, I am primarily concerned with providing the fundamentals necessary to take on the first two approaches.

Part I
Probability & Statistics

2 Probability

The first part of the HKS course (and many econometrics courses) is focused on probability. Some students may find the topics tiresome or basic, but they are quite foundational to econometrics and thus important to get right. While you are unlikely to need to have a comprehensive library of distributions memorized to successfully do empirical research, a good working understanding and ability to learn the properties of different distributions quickly is important, especially for more advanced modeling.

We begin with our population data-generating proces $y_i = g(x_i|\theta)$. As mentioned before, this can be something complicated like a causal relationship or it can be a simple distribution. Even if the population DGP is just a simple distribution, we must have a healthy grasp on probability and the properties of distributions and expectations in order to have hope of proceeding to sampling and estimation. After all, if we cannot understand the properties of the distributions that could underly the population DGP, how could we ever hope to estimate its parameters?

For this section, it probably makes sense to think of the probability generating process as a distribution, i.e. $x_i \sim f(x_i)$.

I will not spend a lot of time on probability, given that most people have some background in it already by the time they take a PhD course and that there are several textbooks and online resources that treat it in much greater detail than I could. I will instead focus on a few concepts that you might have not seen in detail before that are going to be useful in more complex probability problems. Specifically, we will be studying:

- Moment-generating-functions (MGF's): this is merely a transformation of our probability distribution that makes the "moments" (i.e. mean, variance) of very complicated distributions easier to calculate
- Convolutions : this is a way of deriving the distribution of *sums* of random variables

2.1 Moment Generating Functions

One way to understand a population DGP is to characterize its mean, variance and other so-called "moments" of the distribution that help us understand the distribution's shape. Unfortunately, we are often interested in estimating parameters θ for quite complicated distributions $f(x_i|\theta)$, and often, these distributions are too complicated for us to recover mean and variance using the simple equations that we learned in undergrad.

Instead, we can use a moment-generating-function (MGF). An MGF is just a tool used to recover means and variances from complicated distributions by exploiting the properties of derivatives of exponential functions.

For distribution $x \sim f(x)$

- We define the Moment-Generating-Function as

  $$M_x(t) = E[exp(tx)]$$

  where we are taking the expectation over all the possible values of x. The variable t is just some arbitrary variable, which we will use to pull out the actual moments from this distribution

- We define the following derivatives:
  * $M_x(t)' = \frac{\partial}{\partial t} M_x(t)$
  * $M'_x(t) = \frac{\partial^2}{\partial x t^2} M_x(t)$

- Moments
  * $E(x) = lim_{t \to 0} M'_x(t)$
  * $Var(x) = lim_{t \to 0} \{M''_x(t) - M'_x(t)^2\}$

Many students find MGF's non-intuitive or difficult to visualize and try to understand if there is something more significant going on here. However, at the end of the day, we are just exploiting that the derivative of the exponential function is self-replicating. Thus, you should consider MGF's just a tool that is useful for recovering different statistics about our population DGP, nothing more.

Example: Moment Generating Function
Consider a uniform distribution $f(x) = \frac{1}{10}$, $x \in [0, 10]$, find the mean using a MGF

Solve:

First, we find the MGF
$M_x(t) = E[exp(tx)]$
$M_x(t) = \int_0^{10} exp(tx)\frac{1}{10}dx$
$M_x(t) = \frac{1}{t}exp(tx)\frac{1}{10}|_0^{10}$
$M_x(t) = \frac{1}{t}exp(t10)\frac{1}{10} - \frac{1}{t}exp(0)\frac{1}{10}$
$M_x(t) = \frac{1}{t}exp(t10)\frac{1}{10} - \frac{1}{t}\frac{1}{10}$

Now, we find the derivative
$M_x(t)' = \frac{\partial}{\partial t}M_x(t)$
$M_x(t)' = \frac{exp(10t)}{t} - \frac{exp(10t)}{10t^2} + \frac{1}{10t^2}$
$M_x(t)' = \frac{10texp(10t) - exp(10t) + 1}{10t^2}$

Finally, we are ready to take the limit to find the mean
$E(x) = lim_{t\to 0}M_x'(t)$
$E(x) = lim_{t\to 0}\frac{10texp(10t) - exp(10t) + 1}{10t^2}$ We see that both the numerator and the denomenator go to zero. Thus, we have to use L'Hopital's rule. And take the derivative of the numerator and denomenator
$E(x) = lim_{t\to 0}\frac{10exp(10t) + 100texp(10t) - 10exp(10t)}{20t}$
$E(x) = lim_{t\to 0}5exp(10t)$
$E(x) = 5$

2.2 Convolutions

Convolutions are used when we want to know the pdf $f(y)$ of some variable Y, which is equal to the sum of some variables ($Y = X_1 + X_2$). It's useful when we are aggregating multiple observations X_1, X_2 or when we are getting multiple signals, for example if we wanted to know the distribution of a small sample mean.

Simple discrete example: Before we get to the generic form of continuous convolutions, let us start with a simple discrete example. Consider if $X_i = 1\{\text{coin flip i is a heads}\}$ and $Y = X_1 + X_2$. That is, Y is merely the total number of heads we get in two flips. What if we wanted to calulate the pdf?

$P(Y = 0) = P(X_1 = 0) * P(X_2 = 0) = \frac{1}{2}\frac{1}{2} = \frac{1}{4}$
$P(Y = 1) = P(X_1 = 0) * P(X_2 = 1) + P(X_1 = 1) * P(X_2 = 0) = \frac{1}{2}\frac{1}{2} + \frac{1}{2}\frac{1}{2} = \frac{1}{2}$
$P(Y = 2) = P(X_1 = 1) * P(X_2 = 1) = \frac{1}{2}\frac{1}{2} = \frac{1}{4}$

While the above approach looks good, we may instead want to represent our pdf in summation notation. Our first trick will be to observe that since $Y = X_1 + X_2$, for any Y value y and X_1 value a, we already know X_2. Thus, we could write the probability as below:
$P(Y = y) = \sum P(X_1 = a)P(X_2 = y - a)$

However, we run into a problem – we don't know the limits of integration! For a given y value, we may not be able to fix X_1 to be 0 or 1. Consider if $Y = 0$, we clearly cannot allow X_1 to equal 1, since no possible value of X_2 (which is also constrained to be 0 or 1) will be able to satisfy the condition that $Y = X_1 + X_2$ (that is, X_2 cannot equal -1).

Instead, we will have to break this into a piece-wise function:

$$P(Y = y) = \begin{cases} \sum_{a=0}^{0} P(X_1 = a)P(X_2 = y - a) & \text{if } y = 0 \\ \sum_{a=0}^{1} P(X_1 = a)P(X_2 = y - a) & \text{if } y = 1 \\ \sum_{a=1}^{1} P(X_1 = a)P(X_2 = y - a) & \text{if } y = 2 \end{cases}$$

Notice that we have different limits of integration for the different y values. While this might seem like an unnecesary step in such a simple discrete example, we will see that for continuous distributions, it is less clear what the limits of integration should be.

Continuous distributions: For continuous functions, the generic formula for convolution is $\boxed{f(y) = \int f_{X_1}(a) f_{X_2}(y - a) da}$

As in the discrete example, we are integrating over different values of X_1 to achieve different values of y, we have to be careful about our limits of integration and usually end up with a piece wise function

$$f(y) = \begin{cases} \int_{c1}^{d1} f_{X_1}(a) f_{X_2}(y - a) da & y \in [b_1, b_2] \\ \int_{c2}^{d2} f_{X_1}(a) f_{X_2}(y - a) da & y \in [b_2, b_3] \end{cases}$$

I have a few general steps to solve a convolution of continuous distributions, that is when I want to find $f(y)$ when $y = x_1 + x_2$ and $x_1, x_2 \sim f(x)$, where $f(x)$ is continuous

1. Find range of y (b_1 to b_3 in above example)
2. Find potential "break-points" where we are going to want to break up our piece-wise functions (b_2 in the above example) using the ranges of the underlying variables X_1 and X_2
3. Within each "sub-range," identify limits of integration for X_1
   (a) Check actual min or max of X_1
   (b) If that doesn't work, use $Y = X_1 + X_2$
      - Go back and check range and make sure that implied limit is within X_1 range
4. Once we have our sub-ranges of the piece-wise function and the limits of integration within each range, plug in our distribution function and integrate. Construct piece-wise function

---

Example: Convolution
Suppose $X_1 \sim U(-1,1) \Rightarrow f_{X_2}(x) = \frac{1}{2}$
Suppose $X_2 \sim U(0,1) \Rightarrow f_{X_1}(x) = 1$
Using the method of convolution, find $f(y)$ where $Y = X_1 + X_2$

Solve:

1. Y can range from -1 to 2
2. Using the range of X_1 and X_2, we have two possible "break points" coinciding with lower and upper bounds of our underlying variables $\{0,1\}$. Thus, we have three plausible regions to investigate:
   - $y \in [-1,0]$
   - $y \in [0,1]$
   - $y \in [1,2]$
3. We now must find the limits of x_1 within each region.
   - $y \in [-1,0]$

     We have to figure out the limits of integration: c and d in $f(y) = \int_c^d f_{X1}(a) f_{X2}(y-a) da$

     Lower limit (c): say we have some arbitrary $y \in [-1,0]$, say -0.5, what is the minimum value x_1 could take on?
     (a) If $X_1 = -1$ (its actual minimum), then we can let $X_2 = 0.5$, so this should work.

     Testing higher numbers in the range (e.g. -0.00001) and lower numbers (e.g. -0.9), we see that X_2 can be sufficiently high to handle when $X_1 = -1$. So $c = -1$

     Upper limit (d): say we have some arbitrary $y \in [-1,0]$, say -0.5, what is the maximum value x_1 could take on?
     (a) If $X_1 = 1$ (its actual maximum), we run into a problem. X_2 cannot take on negative values, so X_1 cannot equal -1 when $y = -0.5$. Considering other values $y \in [-1,0]$, we see this is still a problem.
     (b) We now can note that for a given y, X_1 will be maximized when X_2 is minimized. The minimum of X_2 is 0, so the maximum of X_1 must be y in this range
        – Can X_1 take on these values? When $y \in [-1,0]$, its entire range is a subset of the range of X_1, so we should be good to go.

     Now, we plug in to the convolution function:

     $f(y) = \int_{-1}^{y} 1 * \frac{1}{2} da$
     $f(y) = \frac{1}{2} a|_{-1}^{y} da$
     $f(y) = (y+1)\frac{1}{2}$

- $y \in [0, 1]$

  We have to figure out the limits of integration: c and d in $f(y) = \int_c^d f_{X1}(a) f_{X2}(y - a) da$

  Lower limit (c) : Suppose $y = 0.5$

  (a) Could $X_1 = -1$ (its actual minimum)? No, we see that X_2 cannot be sufficiently large (1.5) to make up the difference for $y = 0.5$, so we should not allow X_1 to dip so low.

  (b) We can now note that the for a given value of y, the minimum value of X_1 corresponds to the maximum value of X_2, which is 1. If $X_2 = 1$, then we can solve that $X_1 = y - 1$. Thus, for any y, our lower bound for $X_1 = y - 1$
  - Can X_1 take on all these values of $y - 1$? For $y \in [0, 1]$, $y - 1 \in [-1, 0]$, which is a subset of X_1's range so we should be good to go

  Upper Limit (d) : Suppose $y = 0.5$

  (a) Could $X_1 = 1$? Its actual maximum value? Clearly not, since X_2 can't be negative

  (b) For a given y, we can max X_1 by minimizing X_2. Thus, max(X_1) = $y - min(X_2) = y - 0 = y$
  - Can X_2 take on these values? the range of y for this region is 0,1, a subset of X_1's range, so we're good to go.

  Now, we plug in to the convolution function:

  $f(y) = \int_{y-1}^{y} \frac{1}{2} da$
  $f(y) = \frac{1}{2} a|_{y-1}^{y} da$
  $f(y) = \frac{1}{2}(y - (y - 1))$
  $f(y) = \frac{1}{2}$

- $y \in [1, 2]$

  We have to figure out the limits of integration: c and d in $f(y) = \int_c^d f_{X1}(a) f_{X2}(y - a) da$

  Lower limit (c): Suppose $y = 1.5$

  (a) Could $X_1 = -1$ (its actual minimum)? Clearly not

  (b) For a given y, the minimum $X_1 = y - max(X_2) = y - 1$.
  - Can X_1 take on all these values? For $y \in [1, 2]$, $y - 1 \in [0, 1]$, so X_1 can take on any of these values.

  Upper limit (d) : Suppose $y = 1.5$

  (a) Could X_1 take on its actual maximum value (1)? Yes! We see that even when X_1 is maximized, X_2 can take on sufficiently small numbers to rationalize any y in this range.

  Now, we plug in to the convolution function:

  $f(y) = \int_{y-1}^{1} \frac{1}{2} da$
  $f(y) = a\frac{1}{2}|_{y-1}^{1} da$
  $f(y) = (1 - (y - 1))\frac{1}{2}$
  $f(y) = (2 - y)\frac{1}{2}$

SOLUTION: $f(y) = \begin{cases} (y+1)\frac{1}{2} & y \in [-1, 0] \\ \frac{1}{2} & y \in (0, 1] \\ (2-y)\frac{1}{2} & y \in (1, 2] \end{cases}$

3 Bayesian statistics

Bayesian statistics is a branch of statistics that extends the logic of Bayes Rule to characterize distributions of parameter values. While this branch of statistics is less common for economic analysis, it's an important area of which to have understanding. The concept also comes up in many areas of economic theory including statistical discrimination and adverse selection. We will briefly discuss the distinction between Bayesian and Classical (sometimes known as Frequentist) Statistics. Then, we shall discuss Bayesian Updating, the formal updating of priors using bayes rule, and Decision Theory, an application of Bayes Rule.

3.1 Bayesian vs. Classical Statistics

Let us quickly discuss a distinction between two schools of thought in statistics: bayesian and classical (or frequentist) statistics. While there has been reams of paper written comparing these two approaches and parsing out the exact differences (or lack thereof), I will just give the briefest explanation below:

(a) Bayesian statistics – This branch of statistics allows the econometrician (or statitician) to have some belief about the value of θ prior to observing the data (called a *prior distribution*). This prior will affect our final guess $\hat{\theta}$. Decision analysis will be in this camp of statistics. While this branch of statistics is not as commonly used in economic studies, its ideas underly many analyses nonetheless.

(b) Classical statistics – This branch of statistics does not allow a prior. Instead, it takes the observed data as the only type of information from which one can draw conclusions. Conventional hypothesis testing falls in this camp.

We can fit both of these into our framework as below.

(a) Bayesian statistics framework

1.) Population Relationship
or Population Data-Generating Process:

$$x_i \sim g(x_i|\theta)$$

where $g(\cdot)$ is just an arbitrary density function and θ is some population parameter.

↓ sampling

2.) Observe data from a sample
of N observations of i = 1 ... N

$$x = \{y_i, x_{1i}, x_{2i}\} \quad i = 1...N$$

↓ updating our prior

3.) Calculate a Posterior Probability
Density Function of parameter θ based
off of our prior $f(\theta)$ and the data x

$$f(\theta|x) = \frac{g(x|\theta)f(\theta)}{\int g(x|\theta)f(\theta)d\theta}$$

(b) Classical statistics framework

1.) Population Relationship
or Population Data-Generating Process:

$$x_i \sim g(x_i|\theta)$$

where $g(\cdot)$ is just an arbitrary density function and θ is some population parameter.

↓ sampling

2.) Observe data from a sample
of N observations of i = 1 ... N

$$x = \{y_i, x_{1i}, x_{2i}\} \quad i = 1...N$$

↓ testing / estimating

3.) Use data to test hypotheses about population parameter θ OR calculate estimator $\hat{\theta}$ of population parameter about which we can calculate a confidence interval

3.2 Bayesian updating and conjugate prior distributions

Considering our framework, imagine that we are uncertain about the parameter θ. We know that it could take on any number of values, but we don't know which one. However, we do have some prior beliefs about which θ values are more likely than others. For example, if you see someone eating a sandwich, you cannot be certain what type of filling the sandwich has, but you can have some sense before getting any more information from the person that the probability that the filling is peanutbutter is probably higher than the probability that the filling is jello. This is the idea behind a *prior distribution*.

- A **prior distribution** $f(\theta)$ - a pdf that defines the relative likelihood of observing different values of θ prior to observing any data

Depending on the context, this prior distribution could come from some information we know about the world ex-ante or could just be some arbitrary belief one has – either way, the mechanics are the same.

Bayesian updating is the process of using data x to generate a *posterior distribution*, which characterizes the relative likelihood of observing different θ values, accounting for the new information that we have learned from the data.

- **posterior distribution** $g(\theta)$ - a pdf that defines the relative likelihood of observing different values of θ, formed using bayes rule from the observed data x and the prior distribution $f(\theta)$

As a reminder, *Bayes Rule* can take on the following forms:

Bayes Rule: Continuous θ
$f(\theta|x) = \frac{f(x|\theta)f(z)}{\int f(x|\theta)f(\theta)d\theta}$

Bayes Rule: Discrete θ
$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{\sum f(x|\theta)f(\theta)}$

You should think of Bayesian updating as a way of better understanding our DGP that incorporates both our data that we are observing in a given experiment or study *and* some external information or belief which is defining our prior.

Note: A common related concept is that of a *conjugate prior*. A conjugate prior is a prior $f(\theta)$ such that if it has a distributional form in a certain "family" of distributions, then our updated posterior $g(\theta)$ will also follow a form in that family. Examples include gaussian priors, which produce posteriors also in the gaussian family.[1]

On the next page, we will iterate through a few simple examples of Bayesian updating and see how our posterior distribution is informed by both our prior and our data.

[1] The normal distribution is part of the Gaussian family.

Consider if we have a coin with unknown probability $\theta \in [0, 1]$ of flipping heads. We want to know the probability that the coin unfairly favors heads (i.e. $\theta > /frac12$). We have some prior belief about the likelihood of different θ values being the true θ, but we want to use data to update this belief. To do so, we will be flipping coins and updating our prior to generate a posterior distribution.

Example 1: flat prior

Set up	Prior	$f(\theta) = 1$
	Experiment	1 coin flip
	Result	H
$f(\theta\|x)$	$f(\theta\|x) = \frac{f(x\|\theta)f(\theta)}{\int_0^1 f(x\|\theta)f(\theta)d\theta} = \frac{\theta}{\int \theta d\theta} = 2\theta$	
$Pr(\textit{unfair coin})$	$Pr(\textit{unfair coin}) = \int_{1/2}^1 2\theta d\theta = \frac{3}{4}$	

Example 2: H-favored prior

Set up	Prior	$f(\theta) = 2\theta$
	Experiment	1 coin flip
	Result	H
$f(\theta\|x)$	$f(\theta\|x) = \frac{f(x\|\theta)f(\theta)}{\int_0^1 f(x\|\theta)f(\theta)d\theta} = \frac{2\theta^2}{\int 2\theta^2 d\theta} = 3\theta^2$	
$Pr(\textit{unfair coin})$	$Pr(\textit{unfair coin}) = \int_{1/2}^1 23\theta^2 d\theta = \frac{7}{8}$	

Example 3: Flat prior, more data

Set up	Prior	$f(\theta) = 2\theta$
	Experiment	1 coin flip
	Result	HH
$f(\theta\|x)$	$f(\theta\|x) = \frac{f(x\|\theta)f(\theta)}{\int_0^1 f(x\|\theta)f(\theta)d\theta} = \frac{\theta^2}{\int \theta^2 d\theta} = 3\theta^2$	
$Pr(\textit{unfair coin})$	$Pr(\textit{unfair coin}) = \int_{1/2}^1 23\theta^2 d\theta = \frac{3}{4}$	

3.3 Decision theory

Decision theory studies how agents make choices under uncertainty and has two main branches: a normative philosophical side and a more formal mathematical side. We shall be giving the briefest introduction to the latter. You might notice that it is closely related to what you are already learning in your first semester of microeconomics.

The main idea:

- We consider an individual who may take action $\alpha_1 ... \alpha_j$ (For example α_1 = invest, α_2 = not invest)
- Each of these actions is associated with some distribution of utility values associated with possible outcomes (For example, $f_1(u)$ might be the distribution of possible returns on investment)
- Individuals are *expected utility maximizers* - that is, when faced with multiple options, they will choose the option that will give them the highest value in expectation *given the information that htey have.*
  - The chosen action $\alpha^* = \textit{argmax}_j \{\int u f_j(u)\}$
- In our example, even though an individual may "win" or "lose" on an investment, they will still invest if the investment will pay off in expectation.

The value of information

- New information can change your behavior; for example, if the individual from the previous example has a machine that perfectly predicts if the investment pays off, they will only invest in "winners" and never invest in "losers."
- Consider information I that changes the distribution of possible utilities f_j^I so that we have more certainty over what shall occur
  - Under *perfect information*, f_j^I will be degenerate, giving you a perfect prediction of what will happen
  - Under *imperfect information*, f_j^I will not give you perfect omniscience, but will give you a more accurate picture of the future than no information
- If information changes your behavior, it also changes your expected utility; we can calculate the expected utility that individuals get under different levels of information
  - EV^{PI} – the expected value that an individual will get under perfect information
  - EV^{II} – the expected value that an individual will get under imperfect information
  - EV^{NI} – the expected value that an individual will get under no information (this is the baseline in most situations)
- We can use these expected values to derive the incremental value of different pieces of information
  - $V^{PI} = EV^{PI} - EV^{NI}$, the additional value that individuals get from having perfect infromation relative to no information. You can think of this as the willingness to pay for that information
  - $V^{II} = EV^{II} - EV^{NI}$, the additional value that individuals get from having imperfect infromation relative to no information.
  - Typically, $V^{PI} \geq V^{II}$

4 Classical statistics

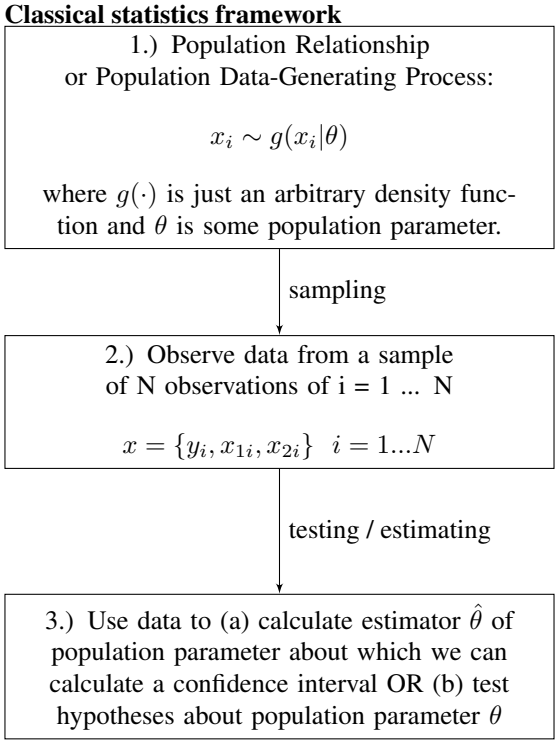
While Bayesian Statistics makes a lot of sense, it does present some difficulties as an empirical framework. For one, it requires us to take a stance on what prior to use. While some sitations might elicit a natural prior, researchers will often end up arguing over the appropriate prior. A researcher hungry for citations might be tempted to use whatever prior that will generate a more dramatic finding. Further, because Bayesian statistics typically results in posterior distributions of a parameter θ, not necessarily a single "point estimate," it is less clear as a way to present material (though this limiation can be overcome for some Bayesian Estimators).

Classical Statistics (sometimes called Frequentist Statistics) is the dominant type of statistical framework for the sciences and social sciences. The main distinction between Bayesian and Classical is that we no longer have to choose a prior. The only thing that matters for estimation and inference is the data we observe.

There are a couple inter-related concepts with which we will have to become comfortable as we dive into ClassicCal Statistics:

1. Estimators - our "best guess" $\hat{\theta}$ of a parameter θ
2. Hypothesis test - a process of deciding whether or not to reject hypotheses at certain "confidence levels" based on the data we observe
3. Power and MDE - given a desired hypothesis test, a way to assess how we want to design our experiment to make sure we will be able to draw meaningful conclusions from our data
4. Confidence intervals - related to hypothesis testing, a range of values which we can be confident at a certain level contains the true value
5. Chi-squared tests - a specific type of hypothesis test for categorical variables

Classical statistics framework



4.1 Estimators

As defined earlier in the notes overview, we can think of an estimator $\hat{\theta}$ as a "best guess" of the population parameter θ based off of the data that we are observing. Notice that when we are talking about estimators, the convention is to put a "hat" on top of the estimator's symbol. Unlike in Bayesian Statistics, Classical Statistics will often arrive at a single "point estimate" of the population parameter value, instead of a distribution of possible θ values.

Examples of estimators:

- Consider if we have normally distributed data $x \sim N(\theta_1, \theta_2)$, with mean θ_1 and standard deviation θ_2. We could define estimators $\hat{\theta_1}$ as the sample mean ($\frac{1}{n}\sum x_i$) and $\}\hat{theta_2}$ as the sample variance ($\frac{1}{n}\sum(x_i - \bar{x})^2$)
- Consider if we believe that people with a bachelor's degree get β more dollars per hour than people with just a high school degree. We could define an estimator $\hat{beta}$ as the difference in average wages between people with bachelors and people with high school degrees.

What one should understand is that we have quite expansive freedom in defining our estimators however we want. However, there are going to be some characteristics that we would want our estimator to have.

 # 1: Unbiased

- def'n: estimator $\hat{\theta}$ is unbiased iff $E(\hat{\theta}) = \theta$
- interpretation: we expect that an unbiased estimator will be right on average over all the samples we could have from a given distribution

 # 2: Consistent

- def'n: estimator $\hat{\theta}$ is consistent iff $lim_{n\to\inf}\hat{\theta} = \theta$
- interpretation: as sample size grows, a consistent estimator will have an increasing probability of being close to the population parameter. As the smaple size approaches infinity, the estimator will be arbitrarily close to the parameter.
- Another important related term is *efficiency*, which characterizes how quickly a consistent estimator will approach the true parameter as N increases.

You will often hear that we want to find *the most efficient unbiased estimator*.

4.2 Hypothesis testing

In the previous section, we learned how to derive estimators $\hat{\theta}$ of population parameters θ. In this section, we will learn how to test specific hypotheses about population parameters. The mechanics of this section will also be useful for constructing "confidence intervals."

The typical presentation is that we have a "null hypothesis" (H_0) and an "alternative hypothesis" (H_1). It may be useful throughout this section to think about the material in the context of an experiment.

Examples of hypotheses:

- Consider $x \sim N(\theta, 1)$. We may have
  - $H_0 : \theta = 4$
  - $H_1 : \theta = 9$
- Consider if we believe that people with a bachelor's degree get β more dollars per hour than people with just a high school degree.
  - $H_0 : \beta = 0$
  - $H_1 : \beta \neq 0$ (i.e. education affects earnings)

In the next subsection, will use the *Neyman Pearson Lemma* to derive the form of the hypothesis test, which will tell us whether we can reject the null hypothesis.[2] We will then proceed to a discussion of *power* and *MDE* calculations which we can use to determine what type of experiment settings (i.e. sample size or differences in null and alternatives) that are necessary to reject the null.

[2] Note that we can never *accept* the Null or Alternative, even with enough data.

4.2.1 The Neyman Pearson Lemma

Main idea: We have two possible population states of the world(H_0 and H_1). We can typically think of this as two different possible θ values. We want to come up with some sort of "rule" or "test" that will characterize whether or not we can reject H_0. The general form of this test will be:

Reject H_0 iff $\hat{\theta} \in R$, where R is the "rejection region", which we can define as a...

- threshold: $\{R|x > \eta\}$
- finite interval : $\{R|x \in (\eta_1, \eta_2)\}$
- union of many intervals: $\{R|x \in (\eta_1, \eta_2) \cup ... \cup (\eta_{n-1}, \eta_n)\}$

What type of rule do we want? The Neyman-Pearson Lemma states that the optimal rejection rule will minimize type 2 error, subject to a fixed constraint of type 1 error. We define these errors below:

- Type 1 : reject H_0 when it is true (set $Pr(\text{type 1 error}) = \alpha$)
- Type 2 : accept H_0 when it is not true (minimize $Pr(\text{Type 2 error})$

Steps to find optimal rejection region:

1. Calculate likelihood ratio $\Lambda(x)$ using distributions implied by different hypotheses
   - The likelihood ratio is defined as $\Lambda(x) = \frac{L(x|\theta_A)}{L(x|\theta_0)} = \frac{f(x|\theta_A)}{f(x|\theta_0)}$
   - Intuition: for datapoint x, $\Lambda(x)$ is the relative likelihood of observing datapoint x under the alternative relative to the null. The higher $\Lambda(x)$, the more likely that that observation x was produced by θ_A instead of θ_0.
2. Use likelihood ratio to derive the form of the test (i.e. if the rejection region is a threshold, finite interval, or set of intervals). The rejection region will correspond to the area where $\Lambda(x)$ is highest.
3. Use type 1 constraint that $Pr(\text{type 1 error}) = \alpha$ to find the actual numeric values of the limits of the rejection regoin, whether that region be a threshold or a finite interval. Notice that $Pr(\text{type 2 error})$ only matters in determining the form (threshold or finite region) and not the actual limits of the rejection region.

We shall now run through two examples of derivations of optimal rejection regions.

Example 1 : Apply Neyman-Pearson to an arbitrary pdf, only one observation

Consider if we have a random variable $X \sim f(x) = \theta x^{\theta-1}$, $x \in (0, 1]$

We are presented with two states of the world:
Null $H_0 : \theta = 3$
Alternative $H_1 : \theta = 2$

We observe only one draw x from the distribution and we want to derive a rule to indicate when we think we are in Null or Alternative state of the world. We shall construct the rule to minimize probability of type 1 error subject to fixing probability of type 2 error $\alpha = 0.05$

Solve

1. Calculate likelihood ratio

   Let's fill in the generic likelihood function (aka the distribution function) $L(x|\theta) = \theta x^{\theta-1}$, for our two potential values of the population parameter θ

   $\Lambda(x) = \frac{L(x|\theta_1)}{L(x|\theta_0)}$
   $\Lambda(x) = \frac{\theta_1 x^{\theta_1-1}}{\theta_0 x^{\theta_0-1}}$
   $\Lambda(x) = \frac{2x^{2-1}}{3x^{3-1}}$
   $\Lambda(x) = \frac{2}{3x}$

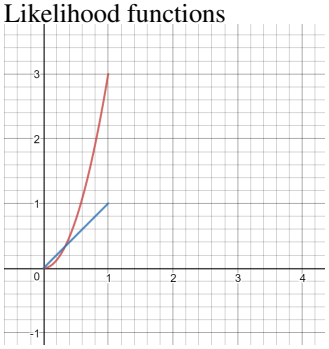
2. Derive form of rejection region using either formal math or graphical intuition

   - We want to reject null if the likelihood of the alternative compared to the null is too high. Thus, we reject null if

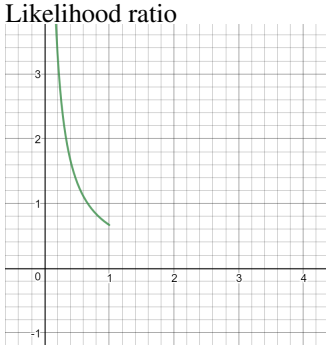
     $\Lambda(x) \geq \eta$, where η is some arbitrary value to be determined. We will then use this relationship to derive a new threshold specifically for our test statistic (x)
     $\frac{2}{3x} \geq \eta$
     $\underbrace{\frac{2}{3\eta}}_{\eta^*} \geq x$

     Thus, reject H_0 iff $x \leq \eta^*$
   - Because our likelihood function is in a nice form, we can also use our graphical intuition to get the threshold value $\Lambda(x) = \frac{2}{3x}$. Graphing below, we see that the likelihood function is decreasing over the range of x, so we know that we are looking to define a rejection region between 0 and η^*

     Likelihood functions

     Likelihood ratio

     Using either approach, we know that our rejection region will be defined as $\{R|x \in (0, \eta^*]\}$

3. Use type 1 constraint to get specific value of η^*

   $\alpha =$ Pr(type 1 error)
   $\alpha =$ Pr(reject H_0 even though H_0 true)

$\alpha = \Pr(x \leq \eta^* | \theta_0)$
$\alpha = \int_0^{\eta^*} \theta_0 x^{\theta_0 - 1} dx$
$\alpha = \int_0^{\eta^*} 3x^{3-1} dx$
$\alpha = x^3 |_0^{\eta^*}$
$\sqrt[3]{\alpha} = \eta^*$

We are working with a 95 % CI, so we set $\alpha = 0.05 \rightarrow$ Optimal test: Reject$H_0(\theta_0 = 3)$in favor of $H_1(\theta_1 = 2)$ iff $x \leq \sqrt[3]{0.05}$

Example 2 : Normal pdf ; many observations
Consider if we have a random variable $X \sim N(\theta, 1)$

We are presented with two states of the world:
Null $H_0 : \theta = 5$
Alternative $H_1 : \theta = 3$

We observe a series of draws $x = \{x_1, ...x_n\}$ from the distribution and we want to derive a rule to indicate when we think we are in Null or Alternative state of the world. We shall construct the rule to minimize probability of type 1 error subject to fixing probability of type 2 error $\alpha = 0.05$

Solve

1. Calculate likelihood ratio

   Unlike before, we are now observing multiple x values, so we have to incorporate all of these observations into our likelihood

   Likelihood of one observation x_i: $L(x_i|\theta) = (2\pi)^{-1/2}\text{exp}\{\frac{-(x_i-\theta)^2}{2}\}$
   Likelihood of vector $x = \{x_1, ...x_n\} : L(x|\theta) = \prod_{i=1}^{n} L(x_i|\theta)$
   $L(x|\theta) = \prod_{i=1}^{n}(2\pi)^{-1/2}\text{exp}\{\frac{-(x_i-\theta)^2}{2}\}$
   $L(x|\theta) = (2\pi)^{-n/2}\text{exp}\{\frac{-\sum(x_i-\theta)^2}{2}\}$
   $L(x|\theta) = (2\pi)^{-n/2}\text{exp}\{\frac{-\sum(x_i-\bar{x}+\bar{x}-\theta)^2}{2}\}$
   $L(x|\theta) = (2\pi)^{-n/2}\text{exp}\{\frac{-\sum\{(x_i-\bar{x})^2+(\bar{x}-\theta)^2+(x_i-\bar{x})(\bar{x}-\theta)\}}{2}\}$
   $L(x|\theta) = (2\pi)^{-n/2}\text{exp}\{\frac{-\sum(x_i-\bar{x})^2-\sum(\bar{x}-\theta)^2-\sum(x_i-\bar{x})(\bar{x}-\theta)}{2}\}$
   $L(x|\theta) = (2\pi)^{-n/2}\text{exp}\{\frac{-\sum(x_i-\bar{x})^2-n(\bar{x}-\theta)^2-\sum(x_i\bar{x}-x_i\theta-\bar{x}^2+\bar{x}\theta)}{2}\}$
   $L(x|\theta) = (2\pi)^{-n/2}\text{exp}\{\frac{-\sum(x_i-\bar{x})^2-n(\bar{x}-\theta)^2-(n\bar{x}^2-n\bar{x}\theta-n\bar{x}^2+n\bar{x}\theta)}{2}\}$
   $L(x|\theta) = (2\pi)^{-n/2}\text{exp}\{\frac{-\sum(x_i-\bar{x})^2-n(\bar{x}-\theta)^2}{2}\}$

   $\frac{L(x|\theta_1)}{L(x|\theta_0)} = \frac{(2\pi)^{-n/2}\text{exp}\{\frac{-\sum(x_i-\bar{x})^2-n(\bar{x}-\theta_1)^2}{2}\}}{(2\pi)^{-n/2}\text{exp}\{\frac{-\sum(x_i-\bar{x})^2-n(\bar{x}-\theta_0)^2}{2}\}}$
   $\frac{L(x|\theta_1)}{L(x|\theta_0)} = \text{exp}\{\frac{-n(\bar{x}-\theta_1)^2+n(\bar{x}-\theta_0)^2}{2}\}$
   $\frac{L(x|\theta_1)}{L(x|\theta_0)} = \text{exp}\{\frac{-n}{2}((\bar{x}-\theta_1)^2-(\bar{x}-\theta_0)^2)\}$

   Plug in for $\theta_0 = 5$ and $\theta_1 = 3$

   $\frac{L(x|\theta_1)}{L(x|\theta_0)} = \text{exp}\{\frac{-n}{2}((\bar{x}-3)^2-(\bar{x}-5)^2)\}$
   $\frac{L(x|\theta_1)}{L(x|\theta_0)} = \text{exp}\{\frac{-n}{2}((\bar{x}^2-6\bar{x}+9)-(\bar{x}^2-10\bar{x}+25))\}$
   $\frac{L(x|\theta_1)}{L(x|\theta_0)} = \text{exp}\{\frac{-n}{2}(\bar{x}^2-6\bar{x}+9-\bar{x}^2+10\bar{x}-25)\}$
   $\frac{L(x|\theta_1)}{L(x|\theta_0)} = \text{exp}\{\frac{-n}{2}(4\bar{x}-16)\}$
   $\frac{L(x|\theta_1)}{L(x|\theta_0)} = \text{exp}\{-n(2\bar{x}-8)\}$

2. Derive form of rejection region using either formal math or graphical intuition

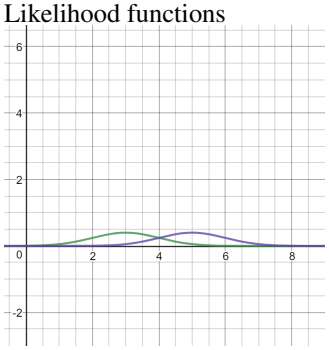
   - We want to reject null if the likelihood of the alternative compared to the null is too high. Because we have set the likelihood to be the likelihood of the alternative over the null, it is highest when the alternative is relatively more likely. Thus, we reject the null when the likelihood ratio is higer. We can write this rule as follows:

     Reject the null (H_0) if....
     $\frac{L(x|\theta_1)}{L(x|\theta_0)} \geq \eta$
     $\text{exp}\{-n(2\bar{x}-8)\} \geq \eta$
     $-n(2\bar{x}-8) \geq ln(\eta)$
     $2\bar{x} \leq \frac{1}{-n}ln(\eta)+8$

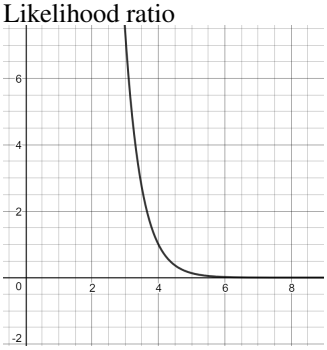
$$\bar{x} \leq \underbrace{\frac{1}{-2n} ln(\eta) + 4}_{\eta^*}$$
$$\bar{x} \leq \eta^*$$

- Suppose we did not want to do the math, but instead wanted to just use graphical intuition. This equation is slightly less nice than our earlier example, but if we are familiar with exponential functions, we can still make it work:

Likelihood functions



Likelihood ratio



Beautiful! So either way we go about determining the shape of our likelihood function, we know that the form our test will take is some threshold over which we will reject our null.

3. Use type 1 error constraint to get specific value of η^*

$\alpha =$ Pr(type 1 error)
$\alpha =$ Pr(reject H_0 even though H_0 true)
$\alpha =$ Pr($\bar{x} \leq \eta^* | \theta_0$)

To do this last step, we must draw on our knowledge of (1) sums of normal variables and (2) the standard normal distribution.

(1) if $x \sim N(5, 1) \Rightarrow \bar{x} \sim N(5, \frac{1}{n})$

(2) if $\bar{x} \sim N(5, \frac{1}{n}) \Rightarrow \frac{\bar{x}-5}{\frac{1}{\sqrt{n}}} \sim N(0, 1)$

Thus,
$\alpha =$ Pr($\bar{x} \leq \eta^* | \theta_0$)
$\alpha =$ Pr($\sqrt{n}(\bar{x} - 5) \leq \sqrt{n}(\eta^* - 5)$)

We can then use the standard normal table and find the associate z statistic for $\alpha = 0.05$ (as we see in the table below, it is 1.64; because we are dealing with a left-sided test, we use the negative side, *-1.64*)

α	z_α	α	z_α	α	z_α	α	z_α	α	z_α
.50	0.00	.050	1.64	.030	1.88	.020	2.05	.010	2.33
.45	0.13	.048	1.66	.029	1.90	.019	2.07	.009	2.37
.40	0.25	.046	1.68	.028	1.91	.018	2.10	.008	2.41
.35	0.39	.044	1.71	.027	1.93	.017	2.12	.007	2.46
.30	0.52	.042	1.73	.026	1.94	.016	2.14	.006	2.51
.25	0.67	.040	1.75	.025	1.96	.015	2.17	.005	2.58
.20	0.84	.038	1.77	.024	1.98	.014	2.20	.004	2.65
.15	1.04	.036	1.80	.023	2.00	.013	2.23	.003	2.75
.10	1.28	.034	1.83	.022	2.01	.012	2.26	.002	2.88
.05	1.64	.032	1.85	.021	2.03	.011	2.29	.001	3.09

Now we plug into the rejection inequality
$\sqrt{n}(\bar{x} - 5) \leq -1.64$

Optimal test: Reject H_0 iff $\bar{x} \leq \frac{-1.64}{n} + 5$

4.3 Confidence Intervals

Confidence intervals are a useful rearrangement of a normal distribution hypothesis test. They become quite useful when we get to regressions and establishing ranges of values that our β coefficients could satisfy. Before we dive in, let us remind ourselves, of some definitions:

- θ - some true parameter about the world that we do not know
- $\hat{\theta}$ - an estimator for this true parameter
- Z_{crit} - the critical value necessary to reject a null under a normal distribution, 1.96 under a two-sided hypothesis test

Now that we have these terms, let us consdider a new object:

- **confidence interval (CI)** - range of possible values in which the true θ value will fall a certain percentage of the time. That is, for a 95 % CI, 95 % of the time, the true θ will in that interval

Deriving the 95 % CI: We start with a two-sided hypothesis test at the 95 % confidence level, which has critical value 1.96. As a reminder, we can interpret the below expression as stating that 95 % of the times that θ is the true parameter value, we will have an estimator $\hat{\theta}$ that satisfies the condition below.

$$Pr(|\frac{\hat{\theta}-\theta}{SE_\theta}| \leq 1.96) = 0.95$$

We can rearrange this to get an interval of values that θ will be in 95 % of the time...

$$\begin{aligned}
|\frac{\hat{\theta}-\theta}{SE_\theta}| &\leq 1.96 \\
\frac{\hat{\theta}-\theta}{SE_\theta} \leq 1.96 \quad &\& \quad \frac{\hat{\theta}-\theta}{SE_\theta} \geq -1.96 \\
\frac{\hat{\theta}-\theta}{SE_\theta} \leq 1.96 \quad &\& \quad \frac{\hat{\theta}-\theta}{SE_\theta} \geq -1.96 \\
\hat{\theta}-\theta \leq 1.96 * SE_\theta \quad &\& \quad \hat{\theta}-\theta \geq -1.96 * SE_\theta \\
\hat{\theta}-1.96 * SE_\theta \leq \theta \quad &\& \quad \hat{\theta}+1.96 * SE_\theta \geq \theta
\end{aligned}$$

$$\boxed{95\ \%\ \text{CI}\ :\ \theta \in (\hat{\theta}-1.96 * SE_\theta, \hat{\theta}+1.96 * SE_\theta)}$$

How do we interpret the above range of values?

- When we have estimator value $\hat{\theta}$, the true value of θ will fall into this range of values 95 % of the time.

4.4 Statistical power and MDE

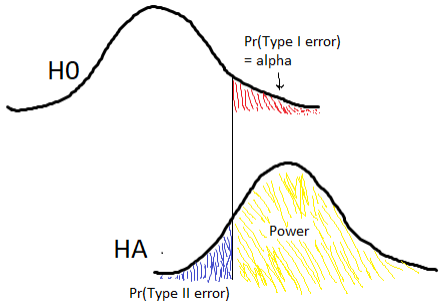
As discussed previously, the cannonical form of the optimal hypothesis test per Neyman-Pearson is:

$$\text{Minimize Pr(Type II error) subject to Pr(Type I error)} = \alpha$$

Where

- Type I error is rejecting the null when the null is true
- Type II error is failing to reject the null when the alternative is true

We will now emphasize a related concept: power. Power can be thought of the ability of our hypothesis test to identify if the alternative is true. Formally, **Power = 1 - Pr(Type II error)**, thus when we are minimizing Type II eror in our optimal hypothesis test, we are maximizing power. We can visualize the power associated with a one-sided test of a normal distribution as below:



There are two types of thought experiments for statistical power:

1. **Minimum detectable effect:** We want to test the hypothesis that two distributions are different. For a given sample size and significance level α, how different do our two distributions have to be to get a target power?
2. **Predetermining the sample size:** We want to test the hypothesis that two distributions are different. For a given minimal detectable effect, what is the sample size we need to get a target power?

Tips to solve: The steps to solving these types of problems is pretty similar. The goal is usually to get all of the components to the following equality:

$$\boxed{Equation\ 1 : z_A^{crit} = \frac{\bar{x}^* - \mu_A}{SE_A(n)} = \frac{(\mu_0 - \mu_A) + Z_0^{crit} * SE_0(n)}{SE_A(n)}}$$

and then solve for your parameter of interest: $(\mu_0 - \mu_A)$ if a MDE problem or n if a sample size problem. I write the steps generically below on how to get the individual components of the formula, but solving the problem will obviously depend on what parameters you are given and what parameters you solve for. I will be writing these steps for the case of normal distributions, which is by far the most common setting for these calculations to be conducted.

1. Find z_0^{crit}
   - interpretation: z_0^{crit} is the critical value necessary to reject the null hypothesis α percent of the time that the null is true
   - input: α confidence level
   - process: Use the normal distribution tables (e.g. $z_0^{crit} = 1.96$ when $\alpha = 0.05$, two-sided test)
2. Find $\bar{x}^*$
   - interpretation: $\bar{x}^*$ is the test statistic necessary to observe in order to reject the null
   - input: z_0^{crit}, null mean μ_0, and standard error of the test statistic under the null distribution $SE_0(n)$–which I have written as a function of n
   - process : use identity $\bar{x}^* = \mu_0 + z_0^{crit} * SE_0(n)$
3. Find z_A^{crit}
   - interpretation: z_A^{crit} is the critical value necessary to detect the alternative the desired percent of the time (power)
   - input : target power$(1 - \beta)$, typically 80%
   - process : Use the normal distribution tables (e.g. $z_A^{crit} = 0.84$ when $1 - \beta = 0.8$, one-sided test)
4. Plug all these components into Equation 1 and solve for the desired parameter, $\mu_0 - \mu_A$ or n

---

Example: Power calculations

We have a x's which are drawn from a distribution with two possible means (μ_0 and μ_1) and a variance of 4. We observe N = 81 draws from the distribution, how different do our null and alternative mean have to be in order to be able to conduct a two-sided test with significance of 0.05 and power equal to 80 %? Assume $\mu_A > \mu_0$

Solve

1. Find z_0^{crit}
   - When $\alpha = 0.05$, $Z_0^{crit} = 1.96$
2. Find $\bar{x}^*$
   - To reject null, $\frac{\bar{x}-\mu_0}{SE_0(n)} > 1.96$
     We rearrange and get $\bar{x} > 1.96 * SE_0(n) + \mu_0$
3. Find Z_A^{crit}
   - When $1 - \beta = 0.8$, $Z_A^{crit} = 0.84$, because we will reject the null for larger values of $\bar{x}$, the area associated with type II error will be to the left, so we use $Z_A^{crit} = -0.84$
4. Use our relationship $z_A^{crit} = \frac{\bar{x}^*-\mu_A}{SE_A(n)}$
   - First, let's solve for the standard errors of our sample means
     $SE_A(n) = SE_0(n) = \sqrt{var(\frac{1}{n}\sum x_i)}$
     $SE_A(n) = SE_0(n) = \sqrt{\frac{1}{n^2}\sum var(x_i)}$
     $SE_A(n) = SE_0(n) = \sqrt{\frac{1}{n^2}4n}$
     $SE_A(n) = SE_0(n) = \frac{2}{\sqrt{n}}$
     $SE_A(n) = SE_0(n) = \frac{2}{\sqrt{81}}$
     $SE_A(n) = SE_0(n) = \frac{2}{9}$
   - Now, we are ready to plug in everything
     $z_A^{crit} = \frac{\bar{x}^*-\mu_A}{SE_A(n)}$
     $-0.84 = \frac{1.96*\frac{2}{9}+\mu_0-\mu_A}{\frac{2}{9}}$
     $\frac{2}{9}(-0.84) - 1.96 * \frac{2}{9} = \mu_0 - \mu_A$

     Or equivalently, $\mu_A - \mu_0 = \frac{2}{9}(0.84 + 1.96)$
     $\mu_A - \mu_0 = \frac{2}{9}2.8 = \frac{5.6}{9}$

     Solution: Our alternative mean must be at minimum $\frac{5.6}{9}$ greater than the null mean in order to satisfy the significance and power criteria in the setup given our sample size and individual observation standard errrors

4.5 Chi-Squared Tests

You are probably familiar with binomial distributed variables, which consider N trials each of which has one of two outcomes (often phrased as "success" or "failure"). The binomial distribution corresponds to the number of "successes" in N trials.

Consider now if each of our trials can result in more than two outcomes, say there are k total outcomes or categories. Such a variable has a multinomial distribution. More formally:

- If Y is a multinomial variable, it can take an of the k values in the set $\{outcome_1, outcome_2, ...outcome_k\}$. Each outcome r_j has a probability p_j of occurring. You often hear multinomial variables being referred to as *categorical variables*

Chi-Squared Test: We have some categorical variable that can take on k values $\{outcome_1, outcome_2, ...outcome_k\}$, each with some probability $\{p_1, p_2, ...p_k\}$. We have two possible states of the world:

- Null: $\{p_1 = r_{10}, p_2 = r_{20}, ..., p_k = r_{k0}\}$
- Alternative: $\{p_1 \neq r_{10}, p_2 \neq r_{20}, ..., p_k \neq r_{k0}\}$

Per usual, we want to derive a rule or test to determine whether or not to reject the null with a given confidence level α. This process has the following steps:

1. Calculate test statistic: $\chi^2 = \sum_{j=1}^{k} \frac{(x_j - np_j)^2}{np_j}$
2. Determine "degrees of freedom: " k-1
3. Look up this combination of α and degrees of freedom in a chi-squared table to get critical value χ^2_{crit}
4. If $\chi^2 > \chi^2_{crit}$, reject the null

Importance in statistics: While the chi-squared independence test can seem somewhat strange, it is actually connected to many other more familiar statistical concepts

- Normalcy: The sum of k iid squared normal variables is distributed according to the chi-squared distribution. That is, if $Z \sim N(0, 1)$ and $Y = \sum_{i=1}^{;} Z_i^2$, $y \sim \chi_k^2$
- The F-test: A useful test when we turn to regression analysis is the F-test, which is used to test the overall significance of the model. That is, if in a linear regression, we have coefficients, $\beta_0, \beta_1, ...\beta_k$, the f-test can test the joint hypothesis that $\beta_0 = 0, \beta_1 = 0, ...\beta_k = 0$. The F test is essentially testing a variable $\gamma = \frac{b}{c}$ where b and c are both chi-squared distributed.

Example: Chi-Squared test

We divide educational attainment into three categories: (1) less than high school, (2) high school, and (3) more than high school. Suppose we know that in the United States, an individual has the following probability of attaining a certain level of education:
$p_1 = \frac{1}{4}, p_2 = \frac{1}{2}, p_3 = \frac{1}{4}$

We go take a survey of 36 people in Illinois individuals in Illinois, and we observe the following break-down of educational attainment: $num_1 = 12$, $num_2 = 12$, $num_3 = 12$

Can we reject the following null at significance level $\alpha = 0.05$ significance level?

- Null : $p_1^{Illinois} = \frac{1}{4}, p_2^{Illinois} = \frac{1}{2}, p_3^{Illinois} = \frac{1}{4}$
- Alternative: $p_1^{Illinois} \neq \frac{1}{4}, p_2^{Illinois} \neq \frac{1}{2}, p_3^{Illinois} \neq \frac{1}{4}$

Solve

1. $\chi^2 = \sum_{j=1}^{k} \frac{(x_j - np_j)^2}{np_j}$
   $\chi^2 = \frac{(x_1 - np_1)^2}{np_1} + \frac{(x_2 - np_2)^2}{np_2} + \frac{(x_3 - np_3)^2}{np_3}$
   $\chi^2 = \frac{(12 - (36)\frac{1}{4})^2}{(36)\frac{1}{4}} + \frac{(12 - (36)\frac{1}{2})^2}{(36)\frac{1}{2}} + \frac{(12 - (36)\frac{1}{4})^2}{(36)\frac{1}{4}}$
   $\chi^2 = \frac{(12-9)^2}{9} + \frac{(12-18)^2}{18} + \frac{(12-9)^2}{9}$
   $\chi^2 = \frac{(3)^2}{9} + \frac{(6)^2}{18} + \frac{(3)^2}{9}$
   $\chi^2 = 1 + 2 + 1$
   $\chi^2 = 4$
2. Degrees of freedom: 3 - 1 = 2
3. For $\alpha = 0.05$ and dof = 2, we look at the following table under $\alpha = 0.05$ and 2 degrees of freedom. We get critical value equal to 5.99, that is, 95% of the times that the null is true, our χ^2 statistic will be less than 5.99.

Percentage Points of the Chi-Square Distribution

Degrees of Freedom	Probability of a larger value of x^2								
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.01
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84	6.63
2	0.020	0.103	0.211	0.575	1.386	2.77	4.61	5.99	9.21
3	0.115	0.352	0.584	1.212	2.366	4.11	6.25	7.81	11.34
4	0.297	0.711	1.064	1.923	3.357	5.39	7.78	9.49	13.28
5	0.554	1.145	1.610	2.675	4.351	6.63	9.24	11.07	15.09
6	0.872	1.635	2.204	3.455	5.348	7.84	10.64	12.59	16.81
7	1.239	2.167	2.833	4.255	6.346	9.04	12.02	14.07	18.48
8	1.647	2.733	3.490	5.071	7.344	10.22	13.36	15.51	20.09
9	2.088	3.325	4.168	5.899	8.343	11.39	14.68	16.92	21.67
10	2.558	3.940	4.865	6.737	9.342	12.55	15.99	18.31	23.21

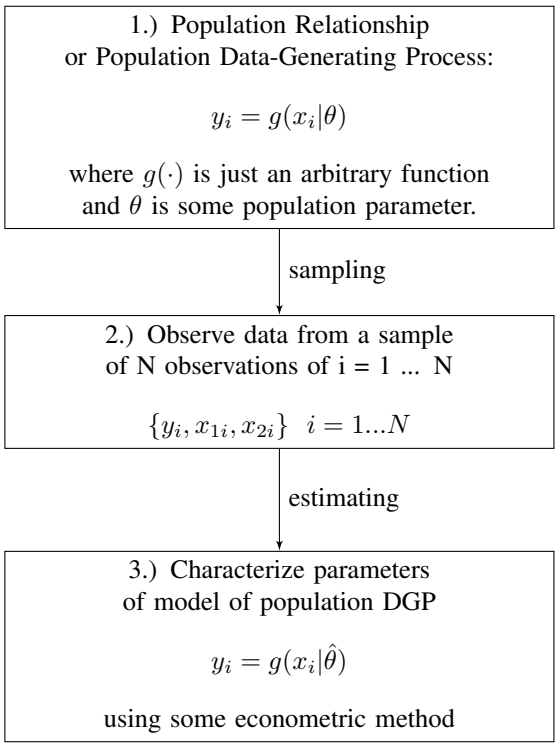
4. We have $\chi^2 = 4$ and $\chi^2_{crit} = 5.99$, so $\chi^2 < \chi^2_{crit}$, thus...

Solution: We cannot reject the null at $\alpha = 0.05$ level

Part II
Econometrics

The notes up to this point have been devoted to learning basic probability and statisics that are used to recover either distributions of possible values of population parameter θ (if we are Bayesians) or point estimators $\hat{\theta}$ of our "best guesses" of the parameter values (if we are Frequentists). We will be drawing from both of these schools of thought (though initially, much more so from the latter) as we take our discussion into more "pure econometrics."

While the distinction between econometrics and plain statistics is quite squishy, we can think of this part of the notes as helping us relate actual relationships between variables X and Y. Our ultimate goal is typically to characterize some sort of causal relationship betwen x and y where we can say x causes y. This fits well into our econometric framework at right, where our usual goal is to recover $\hat{\theta}$ using some econometric method.



There are a number of econometric methods that we could use:

- Ordinary Least Squares
- Weighted Least Squares
- Maximum Likelihood Estimation
- Generalized Method of Moments

Many of these methods are closely related, and we can often get very similar (or even identical) estimators from different methodologies. All of these methodologies also have ways of characterizing confidence intervals around $\hat{\theta}$.

5 Linear regression

Linear regression is in many ways the work-horse methodology for econometric analysis. For individual i, we assume that indpendent variables $x_1 ... x_j$ are *linearly related* to outcome variable y_i. We typically assume that there is some unobserved component ϵ_i that also influences y_i.

A simple example:

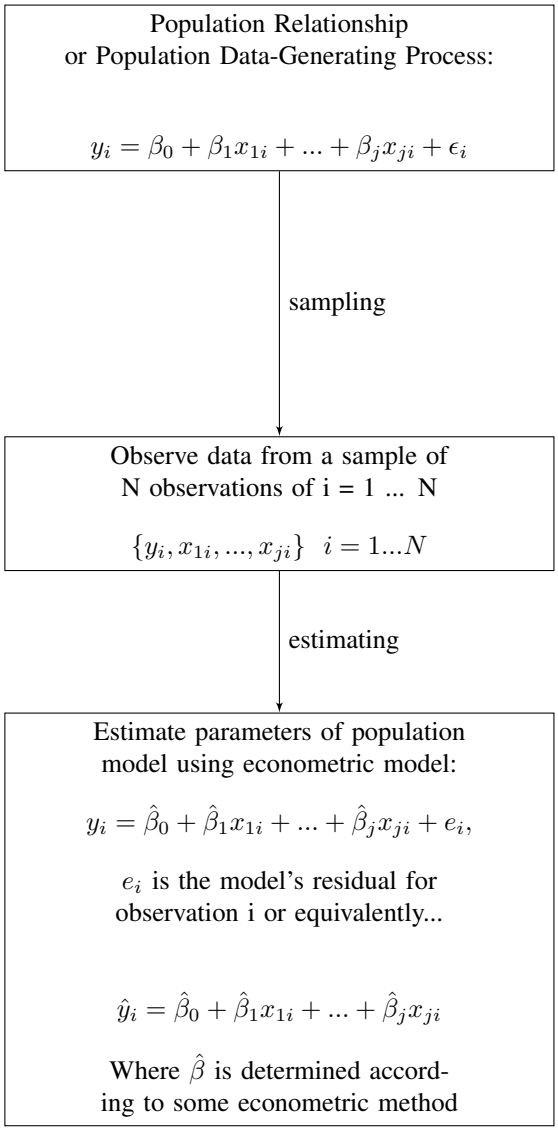
$$wage_i = \beta_0 + \beta_1 education_i + \epsilon_i$$

If this is indeed the true relationship between wages and education, we would expect that a person with no education would recieve β_0 in wages, while a person with γ years of education would recieve $\beta_0 + \beta_1 \gamma$. However, there is some "unexplained component" of wage ϵ_i that does not come from education (for instance intelligence or networking acumen) that could also impact wages.

Before we proceed further, just some points to keep in mind:

- while the example above is a *univariate* regression, with only one indpendent variable $education_i$, we can include more independent variables to produce a *multivariate* regression and estimate an individual $\hat{\beta}$ for each independent variable
- the $\hat{\beta}$'s that we estimate for our linear regressions are still estimators in the same sense of our classical statistics estimators
- the imposition of a linearity assumption underlies both OLS and WLS, methodologies that we will be discussing. This will be too strong an assumption in some cases, but can also be quite flexible, depending on the variables or transformation of variables that we add to our model
- Matrix algebra is particularly useful for this area since is well suited for linearity and is a compact way of presenting many variables

Generic linear problem framework



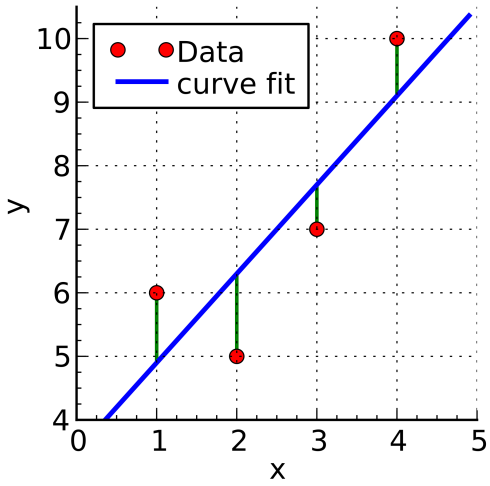
5.1 OLS

Ordinary least squares (OLS) is a methodology for estimating the parameters $\hat{\beta}$ of a linear model. Its logic is based on minimizing the "mistakes" that our model makes in predicting y values. Specifically consider....
population DGP : $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
we estimate model : $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i$
This model predicts the following "fitted values" for our y's: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i$

Because our model does not perfectly fit the data, it produces an errror: $e_i = y_i - \hat{y}_i$.

OLS chooses $\hat{\beta}$ to minimize $\sum_i (y_i - \hat{y}_i)^2$, the Sum of Squared Errors (SSE)

Visualization: Consider if we collected the four data points at right then calculated an OLS line. The red dots are our data; these would be our x_i, y_i values. The blue line is defined by $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Thus, for each x_i, we have a predicted $\hat{y}$. For example, for x = 1, our model would predict $\hat{y} = 6$. Finally, the green vertical line is our error e_i. For example, the error for $x = 1$ is $y - \hat{y} = -1$.



1) Calculus derivation of OLS coefficients (Univariate Case)
Suppose that the population D.G.P. is $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$.

We are given a sample of N y_i, x_i observations and want to estimate the population parameters β_0, β_1. We do this by setting up the following minimization problem:

$$\hat{\beta}_0^{OLS}, \hat{\beta}_1^{OLS} = \text{argmin}_{\hat{\beta}_0, \hat{\beta}_1} \{\sum_{i=1}^{N} (e_i)^2\}$$
$$\hat{\beta}_0^{OLS}, \hat{\beta}_1^{OLS} = \text{argmin}_{\hat{\beta}_0, \hat{\beta}_1} \{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2\}$$
$$\hat{\beta}_0^{OLS}, \hat{\beta}_1^{OLS} = \text{argmin}_{\hat{\beta}_0, \hat{\beta}_1} \{\sum_{i=1}^{N} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2\}$$

Take FOC's:

- $FOC_{\hat{\beta}_1}$:
  $0 = -2\sum_{i=1}^{N}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i$
  $0 = \sum_{i=1}^{N} y_i x_i - \sum_{i=1}^{N} \hat{\beta}_0 x_i - \sum_{i=1}^{N} \hat{\beta}_1 x_i^2$
  $\hat{\beta}_0 = \frac{\sum_{i=1}^{N} y_i x_i - \sum_{i=1}^{N} \hat{\beta}_1 x_i^2}{\sum_{i=1}^{N} x_i}$
  $\hat{\beta}_0 = \frac{\sum_{i=1}^{N} y_i x_i - \sum_{i=1}^{N} \hat{\beta}_1 x_i^2}{N\bar{x}}$
- $FOC_{\hat{\beta}_0}$:
  $0 = -2\sum_{i=1}^{N}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$
  $0 = \sum_{i=1}^{N} y_i - N\hat{\beta}_0 - \sum_{i=1}^{N} \hat{\beta}_1 x_i$
  $\hat{\beta}_0 = \frac{\sum_{i=1}^{N} y_i - \sum_{i=1}^{N} \hat{\beta}_1 x_i}{N}$
  $\hat{\beta}_0 = \frac{N\bar{y} - \sum_{i=1}^{N} \hat{\beta}_1 x_i}{N}$
  $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

  This result is known as the "grand mean" result. It means that our regression line will go through the mean value of $\bar{x}$ and $\bar{y}$.

Combine and solve:

$\frac{\sum_{i=1}^{N} y_i x_i - \sum_{i=1}^{N} \hat{\beta}_1 x_i^2}{N\bar{x}} = \bar{y} - \hat{\beta}_1 \bar{x}$
$\sum_{i=1}^{N} y_i x_i - \sum_{i=1}^{N} \hat{\beta}_1 x_i^2 = \bar{y}\bar{x} - \hat{\beta}_1 \bar{x}^2$
$\bar{x}\sum_{i=1}^{N} \hat{\beta}_1 x_i - \sum_{i=1}^{N} \hat{\beta}_1 x_i^2 = \bar{x}\bar{y} - \sum_{i=1}^{N} y_i x_i$
$\hat{\beta}_1(\bar{x}\sum_{i=1}^{N} x_i - \sum_{i=1}^{N} x_i^2) = \bar{x}\bar{y} - \sum_{i=1}^{N} y_i x_i$
$\hat{\beta}_1 = \frac{\bar{x}\bar{y} - \sum_{i=1}^{N} y_i x_i}{\bar{x}^2 - \sum_{i=1}^{N} x_i^2}$
$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$
$\hat{\beta}_1 = \frac{cov(x,y)}{var(x)}$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
$\hat{\beta}_0 = \bar{y} - \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}\bar{x}$

Univariate OLS Coefficients: $\hat{\beta}_0 = \bar{y} - \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}\bar{x} \quad \hat{\beta}_1 = \frac{cov(x,y)}{var(x)}$

2) Matrix derivations of OLS coefficients (Multivariate Case)
Matrix algebra makes multivariate OLS more compact, but we do have to incur some fixed costs in learning how to manipulate matrices and vectors.

Suppose our Population Data Generating Process is $y_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_J x_{Ji} + \epsilon_i$, which produces each i observation our of N data points.

We can represent this in matrix notation as follows:

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ ... \\ y_N \end{bmatrix}}_{Y} = \underbrace{\begin{bmatrix} 1 & x_{11} & ... & x_{J1} \\ 1 & x_{12} & ... & x_{J2} \\ ... & ... & ... & ... \\ 1 & x_{1N} & ... & x_{JN} \end{bmatrix}}_{X} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ ... \\ \beta_J \end{bmatrix}}_{\beta} + \underbrace{\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ ... \\ \epsilon_N \end{bmatrix}}_{\epsilon} \leftrightarrow Y = X\beta + \epsilon$$

We estimate the following model:

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ ... \\ y_N \end{bmatrix}}_{Y} = \underbrace{\begin{bmatrix} 1 & x_{11} & ... & x_{J1} \\ 1 & x_{12} & ... & x_{J2} \\ ... & ... & ... & ... \\ 1 & x_{1N} & ... & x_{JN} \end{bmatrix}}_{X} \underbrace{\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ ... \\ \hat{\beta}_J \end{bmatrix}}_{\hat{\beta}} + \underbrace{\begin{bmatrix} e_1 \\ e_2 \\ ... \\ e_N \end{bmatrix}}_{e} \leftrightarrow Y = X\hat{\beta} + e$$

This model produces "fitted values" for each x_i as follows:

$$\underbrace{\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ ... \\ \hat{y}_N \end{bmatrix}}_{\hat{Y}} = \underbrace{\begin{bmatrix} 1 & x_{11} & ... & x_{J1} \\ 1 & x_{12} & ... & x_{J2} \\ ... & ... & ... & ... \\ 1 & x_{1N} & ... & x_{JN} \end{bmatrix}}_{X} \underbrace{\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ ... \\ \hat{\beta}_J \end{bmatrix}}_{\hat{\beta}} \leftrightarrow \hat{Y} = X\hat{\beta}$$

Using this more compact notation where e is a vector of each observation's errors, we can derive the usual minimization for OLS:

$\hat{\beta}^{OLS} = \text{argmin}_{\hat{\beta}} \ \sum_i e_i^2$
$\hat{\beta}^{OLS} = \text{argmin}_{\hat{\beta}} \ e^T e$
$\hat{\beta}^{OLS} = \text{argmin}_{\hat{\beta}} \ [Y - X\hat{\beta}]^T [Y - X\hat{\beta}]$
$\hat{\beta}^{OLS} = \text{argmin}_{\hat{\beta}} \ [Y^T - \hat{\beta}^T X^T][Y - X\hat{\beta}]$
$\hat{\beta}^{OLS} = \text{argmin}_{\hat{\beta}} \ [Y^T Y - \hat{\beta}^T X^T Y - Y^T X\hat{\beta} + \hat{\beta}^T X^T X\hat{\beta}]$

- Note: $\hat{beta}^T X^T Y$ is a scalar (i.e. only 1 row and 1 column). Thus, it is equal to its inverse, so we can rewrite:

$\beta^{OLS} = \text{argmin}_{\hat{\beta}} \ [Y^T Y - 2\hat{\beta}^T X^T Y + \hat{\beta}^T X^T X\hat{\beta}]$

Take FOC's:

- Note: $\frac{d}{da} a^T b = \frac{d}{da} b^T a = b$
- Note: $\frac{d}{da} a^T b^T ba = 2b^T ba$

$0 = -2X^T Y + 2X^T X\hat{\beta}$

OLS coefficients : $\hat{\beta} = [X^T X]^{-1} X^T Y$

5.2 Confidence intervals

In hypothesis testing from classical statisics, we set up our problem as accepting or rejecting if a certain hypothesis was true. Often in a regression context, we are testing whether a given coefficient $\beta_j = 0$ with the alternative hypothesis being that $\beta_j \neq 0$

Consider the population data generating process: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. We run a linear regression and get $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i$, where $y_i - \hat{y} = e_i$.

We will construct a two sided test at the α level, where $Z_{\alpha/2}$ is the critical value.[3] It takes the form $|\frac{\hat{\beta}_1 - \beta_1^{null}}{SE_{\beta_1}}| \leq Z_{\alpha/2}$

We can rearrange this to get a $1-\alpha\%$ Confidence Interval:

$$
\begin{aligned}
|\frac{\hat{\beta}_1 - \beta_1^{null}}{SE_{\beta_1}}| &\leq Z_{\alpha/2} \\
\frac{\hat{\beta}_1 - \beta_1^{null}}{SE_{\beta_1}} \leq Z_{\alpha/2} \quad &\& \quad \frac{\hat{\beta}_1 - \beta_1^{null}}{SE_{\beta_1}} \geq -Z_{\alpha/2} \\
\frac{\hat{\beta}_1 - \beta_1^{null}}{SE_{\beta_1}} \leq Z_{\alpha/2} \quad &\& \quad \frac{\hat{\beta}_1 - \beta_1^{null}}{SE_{\beta_1}} \geq -Z_{\alpha/2} \\
\hat{\beta}_1 - \beta_1^{null} \leq Z_{\alpha/2} * SE_{\beta_1} \quad &\& \quad \hat{\beta}_1 - \beta_1^{null} \geq -Z_{\alpha/2} * SE_{\beta_1} \\
\hat{\beta}_1 - Z_{\alpha/2} * SE_{\beta_1} \leq \beta_1^{null} \quad &\& \quad \hat{\beta}_1 + Z_{\alpha/2} * SE_{\beta_1} \geq \beta_1^{null}
\end{aligned}
$$

$$
\boxed{1-\alpha\%\ CI\ :\ \beta_1^{null} \in (\hat{\beta}_1 - Z_{\alpha/2} * SE_{\hat{\beta}_1}, \hat{\beta}_1 + Z_{\alpha/2} * SE_{\hat{\beta}_1})}
$$

This is equivalent to saying that $(1-\alpha)$ % of the time that β^{null} is true, it should be in the above interval.

For more details on $SE_{\hat{\beta}_1}$, check out the section on the Variance-Covariance Matrix section.

[3] For example, $Z_{\alpha/2} = 1.96$ for $\alpha = 0.05$

Sample Stata Output

```
. regress cholesterol time_tv

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) =   17.47
       Model |  5.04902329     1  5.04902329            Prob > F      =  0.0001
    Residual |  28.3220135    98  .289000137           R-squared     =  0.1513
-------------+------------------------------           Adj R-squared =  0.1426
       Total |  33.3710367    99  .337081179           Root MSE      =  .53759

------------------------------------------------------------------------------
 cholesterol |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     time_tv |   .0440691   .0105434     4.18   0.000     .0231461    .0649921
       _cons |  -2.134777   1.813099    -1.18   0.242    -5.732812    1.463259
------------------------------------------------------------------------------
```

Above we have regressed cholesterol on the variable time_tv and an intercept:

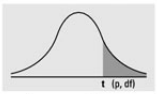
$$cholesterol = \hat{\beta}_0 + \hat{\beta}_1 time\_tv$$

Let's go to the horizontal line next to time_tv and see if we can intepret the results.

- First, we have coefficent 0.0440691. This is $\hat{\beta}_1$
- Next, we have the standard error of our coefficient: $SE_{\hat{\beta}_1} = \frac{s^2}{\sum(x_i - \bar{x})^2} = 0.0105434$
- the t column is the t-statistic (analogous to our z-statistic in our rule – see my brief note on z vs. t statistics above). Notice that $t = \frac{\hat{\beta}_1 - 0}{SE_{\hat{\beta}_1}} = \frac{0.0440691}{0.0105434} = 4.18$
- Next, we have the p-value associated with this t-statistic.

The p-value comes from finding the smallest p-value associated with our t-statistic (4.18) and degrees of freedom (100-2 = 98). If we were doing this by hand, we would go down to the bottom of the table below to the line that says "z" (since the number of d.o.f only matters up to n-k = 30 and we have 98, so we can just use the Z-score), and then we would move rightward until we found a critical value greater than or equal to our number that would mean that we could not reject the null. Then, we would see what p-value is associated with that column. (In this case, the coefficient is so strongly significant that the table does not provide a p-value small enough).

Numbers in each row of the table are values on a *t*-distribution with (*df*) degrees of freedom for selected right-tail (greater-than) probabilities (*p*).



df/p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	43178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0.256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0.256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.684850	1.317836	1.710882	2.06390	2.49216	2.79694	3.7454
25	0.256060	0.684430	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251
26	0.255955	0.684043	1.314972	1.705618	2.05553	2.47863	2.77871	3.7066
27	0.255858	0.683685	1.313703	1.703288	2.05183	2.47266	2.77068	3.6896
28	0.255768	0.683353	1.312527	1.701131	2.04841	2.46714	2.76326	3.6739
29	0.255684	0.683044	1.311434	1.699127	2.04523	2.46202	2.75639	3.6594
30	0.255605	0.682756	1.310415	1.697261	2.04227	2.45726	2.75000	3.6460
z	0.253347	0.674490	1.281552	1.644854	1.95996	2.32635	2.57583	3.2905
CI	———	———	80%	90%	95%	98%	99%	99.9%

- The final columns are the 95% confidence interval of our β_1 coefficient. It gives the region of values that contains our actual β_1 value with a confidence of 95%. We could check that this is consistent with the equation that we derived already in the previous section by plugging in the relevant values into our formula...

$Pr(\hat{\beta}_1 \in CI | \beta_1) = 0.95$
$Pr(\beta_1^{null} \in (\hat{\beta}_1 - Z_{\alpha/2} * SE_{\beta_1}, \hat{\beta}_1 + Z_{\alpha/2} * SE_{\beta_1})) = 0.95$

Example: Regression coefficients and confidence interval
Suppose we have n = 100 observations of tuples $\{x_i, y_i\}$, with $\{x_1, ...x_{20}\} = 0$, $\{x_{21}, ...x_{80}\} = 1$, $\{x_{81}, ...x_{100}\} = 2$

We run the linear regression: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

We observe in our sample

- Conditional sample means : $\bar{y_j} = \frac{\sum_{i \in \{x_i = j\}} y_i}{\sum_{i \in \{x_i = j\}} 1}$
- Sample variance of y : $s_y^2 = \frac{1}{n-2} \sum_i (y_i - \hat{y})^2 = \frac{1}{n-2} \sum_i e_i^2$

How would you construct a 95% confidence interval around β_1 in terms of conditional sample means and variances: $\bar{y_0}, \bar{y_1}, \bar{y_2}, s^2$?
Where would you reject $H_0 : \beta_1 = 0$

Solve

Per our derivation above, we the following formula for our confidence interval:

$Pr(\beta_1^{null} \in (\hat{\beta}_1 - Z_{\alpha/2} * SE_{\beta_1}, \hat{\beta}_1 + Z_{\alpha/2} * SE_{\beta_1})) = 0.95$

Or equivalently, reject $\beta_1 = 0$ iff $|\frac{\hat{\beta}_1}{\sqrt{var(\hat{\beta}_1)}}| \leq 1.96$

Let's solve for $\hat{\beta}_1$ and $SE_{\hat{\beta}_1}$

- $\hat{\beta}_1 = \frac{\sum_{i=1}^{100}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{100}(x_i - \bar{x})^2}$
  $\hat{\beta}_1 = \frac{\sum_{i \in \{x_i=0\}}(x_i - \bar{x})(y_i - \bar{y}) + \sum_{i \in \{x_i=1\}}(x_i - \bar{x})(y_i - \bar{y}) + \sum_{i \in \{x_i=2\}}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i \in \{x_i=0\}}(x_i - \bar{x})^2 + \sum_{i \in \{x_i=1\}}(x_i - \bar{x})^2 + \sum_{i \in \{x_i=2\}}(x_i - \bar{x})^2}$
  $\bar{x} = \frac{1}{100}(20(0) + 60(1) + 20(2)) = 1$
  $\hat{\beta}_1 = \frac{\sum_{i \in \{x_i=0\}}(0-1)(y_i - \bar{y}) + \sum_{i \in \{x_i=1\}}(1-1)(y_i - \bar{y}) + \sum_{i \in \{x_i=2\}}(2-1)(y_i - \bar{y})}{\sum_{i \in \{x_i=0\}}(0-1)^2 + \sum_{i \in \{x_i=1\}}(1-1)^2 + \sum_{i \in \{x_i=2\}}(2-1)^2}$
  $\hat{\beta}_1 = \frac{\sum_{i \in \{x_i=0\}}(-1)(y_i - \bar{y}) + \sum_{i \in \{x_i=2\}}(y_i - \bar{y})}{\sum_{i \in \{x_i=0\}} 1 + \sum_{i \in \{x_i=2\}}(1)}$
  $\hat{\beta}_1 = \frac{20(-1)(\bar{y_0} - \bar{y}) + 20(y_2 - \bar{y})}{20+20}$
  $\hat{\beta}_1 = \frac{\bar{y_2} - \bar{y_0}}{2}$
- $SE_{\hat{\beta}_1} = SE(\hat{\beta}_1)$
  $SE_{\hat{\beta}_1} = \sqrt{var(\hat{\beta}_1)}$
  $SE_{\hat{\beta}_1} = \sqrt{\frac{\sigma^2}{\sum(x_i - \bar{x})^2}}$
  $SE_{\hat{\beta}_1} = \sqrt{\frac{\frac{1}{n-2}\sum e_i^2}{\sum(x_i - \bar{x})^2}}$
  $SE_{\hat{\beta}_1} = \sqrt{\frac{s^2}{40}}$
  $SE_{\hat{\beta}_1} = \frac{s}{2\sqrt{10}}$

SOLUTION:

We can have 95 % level of confidence that our true β_1 is in the range..

$\beta_1 \in [\hat{\beta} - Z_{\alpha/2,n-2} SE(\hat{\beta}_1) \ , \ \hat{\beta}_1 + Z_{\alpha/2,n-2} SE(\hat{\beta}_1)]$

$\beta_1 \in [\frac{\bar{y_2} - \bar{y_0}}{2} - (1.96)\frac{s}{2\sqrt{10}} \ , \ \frac{\bar{y_2} + \bar{y_0}}{2} + (1.96)\frac{s}{2\sqrt{10}}]$

We reject $H_0 : \beta_0 = 0$ iff $\frac{\bar{y_2} - \bar{y_0}}{2} \geq (1.96)\frac{s}{2\sqrt{10}}$ or $\frac{\bar{y_2} - \bar{y_0}}{2} \leq -(1.96)\frac{s}{2\sqrt{10}}$

5.3 Variance-Covariance Matrix of OLS Coefficients

- **Definition of Variance-Covariance Matrix**

The goal of variance covariance matrix is to get something that looks like this...

$$E\{[\hat{\beta}-\beta][\hat{\beta}-\beta]^T\} = \begin{bmatrix} \text{var}(\hat{\beta}_0) & \text{cov}(\hat{\beta}_0,\hat{\beta}_1) & ... & \text{cov}(\hat{\beta}_0,\hat{\beta}_J) \\ \text{cov}(\hat{\beta}_0,\hat{\beta}_1) & \text{var}(\hat{\beta}_1,\hat{\beta}_1) & ... & \text{cov}(\hat{\beta}_1,\hat{\beta}_J) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(\hat{\beta}_0,\hat{\beta}_J) & \text{cov}(\hat{\beta}_1,\hat{\beta}_J) & ... & \text{var}(\hat{\beta}_J) \end{bmatrix}$$

- **Derivations**

Let us derive the variance-covariance matrix. We begin by cleaning up our vector of estimators $\hat{\beta}$:
$\hat{\beta} = [X^T X]^{-1} X^T Y$
$\hat{\beta} = [X^T X]^{-1} X^T [X\beta + \epsilon]$
$\hat{\beta} = \beta + [X^T X]^{-1} X^T \epsilon$

We can notice that there are two components to this estimator–the constant population parameter vector β and a stochastic component $[X^T X]^{-1} X^T \epsilon$ (where the stochasticity is coming from the ϵ vector). Let us put this into our definition of the variance-covariance matrix and simplify in terms of X and ϵ:

$$\begin{aligned} E\{[\hat{\beta}-\beta][\hat{\beta}-\beta]^T\} &= E\{[[X^T X]^{-1} X^T \epsilon][[X^T X]^{-1} X^T \epsilon]^T\} \\ &= E\{[X^T X]^{-1} X^T \epsilon\epsilon^T X[[X^T X]^{-1}]^T\} \\ &= [X^T X]^{-1} X^T \underbrace{E\{\epsilon\epsilon^T\}}_{\Omega} X[X^T X]^{-1} \end{aligned}$$

Wonderful! The above expression characterizes how our β coefficients are correlated together. Let us focus as well $\Omega = \begin{bmatrix} E[\epsilon_1^2] & ... & ... & E[\epsilon_1,\epsilon_N] \\ E[\epsilon_1,\epsilon_2] & E[\epsilon_2,\epsilon_2] & ... & E[\epsilon_N,\epsilon_2] \\ \vdots & \vdots & \ddots & \vdots \\ E[\epsilon_1,\epsilon_N] & E[\epsilon_2,\epsilon_N] & ... & E[\epsilon_N,\epsilon_N] \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{12}^2 & ... & \sigma_{1N}^2 \\ \sigma_{12}^2 & \sigma_2^2 & ... & \sigma_{N2}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1N}^2 & \sigma_{2N}^2 & ... & \sigma_N^2 \end{bmatrix}$

If our sample is iid, the off-diagonal elements are 0...
$\Omega = \begin{bmatrix} \sigma_1^2 & 0 & ... & 0 \\ 0 & \sigma_2^2 & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & ... & \sigma_N^2 \end{bmatrix}$

5.4 Gauss-Markov and BLUE OLS

Under the following assumptions, known as the Gauss-Markov Assumptions, the ordinary least squares line is the best linear unbiased estimator for a relationship between as set of x's and a set of y's, where "best" means "most efficient[4]."

1. Y is a linear function of X
   - algabraic notation - $Y_i = \sum_{i=0}^{j} \beta_{ji} x_{ji} + \epsilon_i$, where i is the individual observation and j is the independent variable
   - matrix notation - $Y = X\beta + \epsilon$
2. Strict exogeneity
   - $E[\epsilon_i | X] = 0$, that is our error and our independent variables are orthogonal to one another. This would be violated under endogeneity or simultaneity or other forms of OMVB.
3. No multicollinearity
   - This assumption is essentially stating that each new x variable we add must add variation that is not explained by the x variables aleady included. We need variation in X in order to see any sort of relationship between changes in X and changes in Y.
   - algabraic notation - We see this assumption in our $\hat{\beta}$ coefficient equations. For example, in the univariate case:

     $$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

     . If there's no variation in x, $\bar{x} = x_i \forall i$, so the denominator is 0, thus the $\hat{\beta}_1$ is undefined.
   - matrix algebra notation - this is often described as the requirement that X'X is "full rank," which is to say that no column of X can be a linear transform of another (i.e. no independent variable is just a linear transform of another). This allows us to take the inverse of $X^T X$ in our $\hat{\beta} = [X^T X]^{-1} X^T Y$
4. Homoskedastic standard errors
   - We'll discuss this in greater detail in the next section, but the assumption is essentially that $\sigma_i^2 = \sigma^2$ for all i

[4] remember back to our disccussion of estimators: most efficient means the estimator that approaches the true value the quickest as sample size increases

5.5 Heteroskedasticity

Let us dive into the assumption that our errors are *homoskedastic*.

- **homoskedasticity** - $\sigma_i^2 = \sigma^2$ for all observations i. Visually, it means that if we were to draw our OLS regression line through our data points (for example, through the distribution at right), the resulting residuals would be distributed roughly evenly throughout our x distribution. We can rewrite our Ω matrix as below:

$$\Omega = \begin{bmatrix} \sigma_1^2 & 0 & ... & 0 \\ 0 & \sigma_2^2 & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & ... & \sigma_N^2 \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & ... & 0 \\ 0 & \sigma^2 & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & ... & \sigma^2 \end{bmatrix} = \sigma^2 I$$

This means we can simplify our variance-covariance matrix as follows:

$$\begin{aligned} E\{[\hat{\beta} - \beta][\hat{\beta} - \beta]^T\} &= [X^T X]^{-1} X^T \Omega X [X^T X]^{-1} \\ &= [X^T X]^{-1} X^T \sigma^2 I X [X^T X]^{-1} \\ &= \sigma^2 [X^T X]^{-1} X^T X [X^T X]^{-1} \\ &= \sigma^2 [X^T X]^{-1} \end{aligned}$$

We can visualize this at right. We can imagine if we ran a regression line through the cloud of data, the distance between y_i and $\hat{y_i}$ does not seem to change systematically with the value of x.

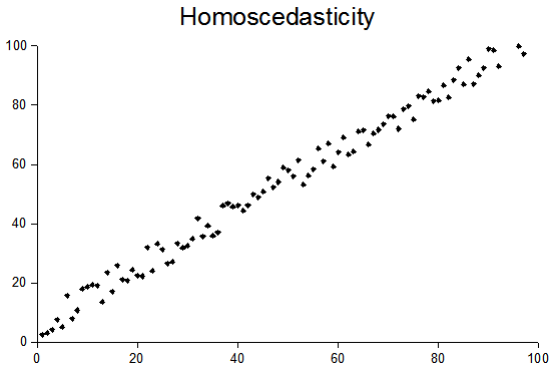


Image lifted from wikipedia article on the subject

We can contrast this with *heteroskedasticity*.

- **heteroskedasticity** - when $\sigma_i^2 \neq \sigma^2 \; \forall \; i$. That is, we have some values of x that have a noisier relationship with y than others. In particular, at right we see a situation where the noise is increasing as x increases. That is, the "cloud" around the regression line is increasing as we increase x. We must keep our Ω matrix as below:

$$\Omega = \begin{bmatrix} \sigma_1^2 & 0 & ... & 0 \\ 0 & \sigma_2^2 & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & ... & \sigma_N^2 \end{bmatrix}$$

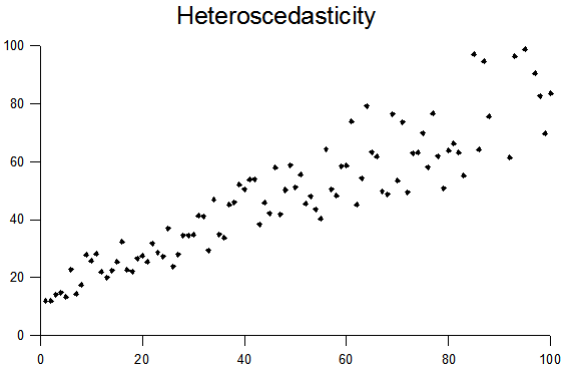


Image lifted from wikipedia article on the subject

The issue(s) with heteroskedasticity:
When we run OLS in the presence of heteroskedasticity, we run into two main issues.

1. Standard errors are not accurate (and often, too small), thus our confidence intervals around $\hat{\beta}$ will no longer be accurate.
2. OLS is no longer BLUE (specifically, it is not "best" in the sense that it is not the most efficient out of all the unbiased estimators)

To see why these issues arise, let's focus on a univariate regression: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i$, which results in the following equation for $\hat{\beta}_1$:

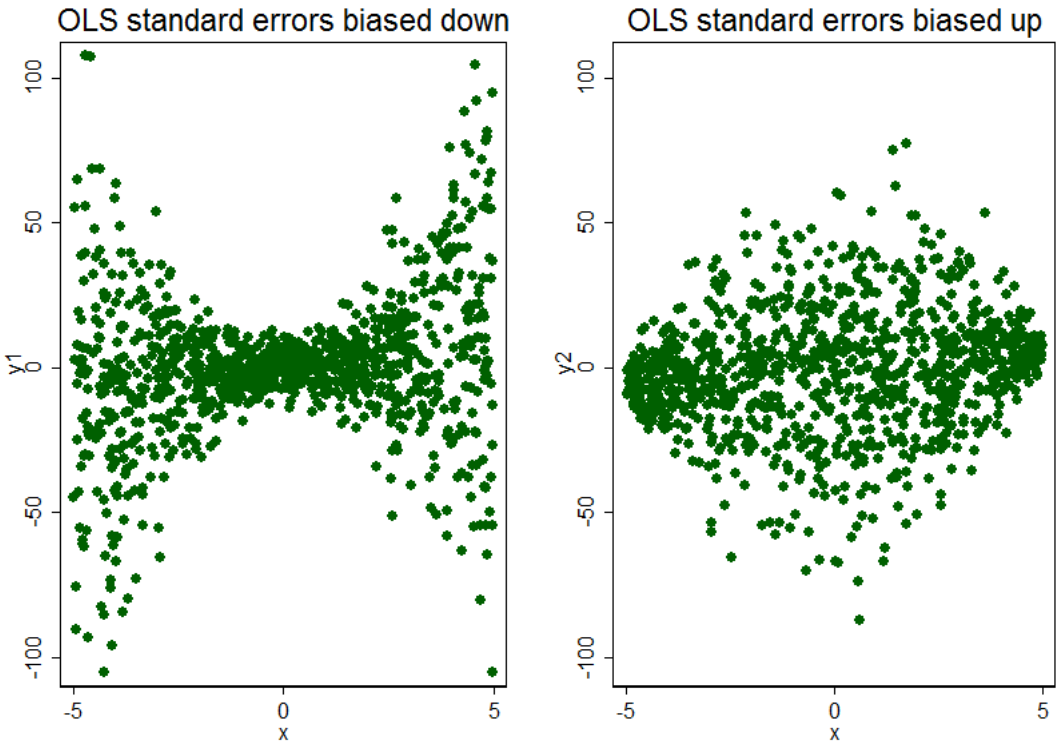
$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

Let us expand this expression so that it is written as a weighted sum of $y_i - \bar{y}$, where each "weight" is $\frac{x_i - \bar{x}}{(x_i - \bar{x})^2}$

$$\hat{\beta}_1 = \frac{(x_1 - \bar{x})}{\sum(x_i - \bar{x})^2}(y_1 - \bar{y}) + ... + \frac{(x_N - \bar{x})}{\sum(x_i - \bar{x})^2}(y_N - \bar{y})$$

We can observe that for a given $(y_i - \bar{y})$ value, the value of $\hat{\beta}_1$ is going to be more influenced by observations further from $\bar{x}$ (i.e. $x_i - \bar{x}$ is larger in magnitude, either positive or negative) because such observations have a "weight" $\frac{x_i - \bar{x}}{(x_i - \bar{x})^2}$ of greater magnitude. Let us now discuss how this contributes to our two problems:

1. Standard errors: As noted above, observations that have values of x that are further away from $\bar{x}$ are going to be more important in determining the value of $\hat{\beta}_1$. If our data looks like the left image below, we might have a problem. Specifically, we have the problem that the important points in our distribution (the x's on the far left and right) are very noisy. Under homoskedastic assumptions, we would just take the mean of the noise across the distribution of x, so we would not take into account the fact that the very points that are the most important for determining the value of $\hat{\beta}_1$ are also the least certain. Thus, in this case, *our standard errors are going to be too small typically*. The opposite could be the case in the right picture.



Lifted from the excellent blog post on the topic at chrisauld.com

2. Inefficiency: Under OLS, the "importance" of a data point (x_i, y_i) is determined by how much it differs from the mean since points the large changes in our variables will offer more information on the relationship between these changes than points with little changes. However, there is another dimension of information quality within a single datapoint that we can take into account: the degree of variation within that x value. Thus, under heteroskedasticity, there are other estimators that could take this additional dimension of quality into account.

The remedies to heteroskedasticity:
There are two main ways of addressing heteroskedasticity.

- Robust standard errors – this modification to our standard error equation will make our confidence interval for our $\hat{\beta}$ coefficient more accurate; it will not change our estimates of $\hat{\beta}$ (thus it addresses issue # 1 but not issue # 2). Effectively, we weight each error by the degree to which it influenced our $\hat{\beta}$ estimates, i.e. weighing by our distance $x_i - \bar{x}$. For both versions of standard errors, we can use the sample variation s_i^2 to approximate σ_i^2, so long as we take into account the degrees of freedom.
  - Under the homoskedastic assumption, we have the following standard error formulae, which will not be accurate if our data actually exhibits heteroskedasticity:
    - algebraic notation (uniariate case) - $SE_{\hat{\beta}_1}^{homo} = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}$
    - matrix algebra notation - $SE_{\hat{\beta}}^{homo} = \sigma^2[X^T X]^{-1}$
  - Standard errors under the heteroskedastic assumption will allow our errors to be heteroskedastic and further allow these errors to differentialy inform the overall error of our $\hat{\beta}$ estimator.
    - algebraic notation (uniariate case) - $SE_{\hat{\beta}_1}^{het} = \frac{\sum(x_i - \bar{x})^2 \sigma_i^2}{(\sum(x_i - \bar{x})^2)^2}$
    - matrix algebra notation - $SE_{\hat{\beta}_1}^{het} = [X^T X]^{-1} X^T \Omega X [X^T X]^{-1}$, where we allow each diagnoal Ω element to be unique
  - O
- Generalized least squares[5] – this modification to our estimation of the $\hat{\beta}$ coefficients themselves will make our $\hat{\beta}$'s more consistent and the resulting standard errors more accurate *if we have all the proper information to implement* (thus it addresses both issue # 1 and issue # 2). We will address WLS into more detail later.

[5] GLS is a class of estimators that includes weighted least squares (WLS)

5.6 Weighted Least Squares

Weighting is an important tool in empirical analyses to satisfy the various objectives we have for our estimators, in this case, our objective for creating an estimator that is not only unbiased but as efficient as possible (i.e. consistent and approaching the true value as quickly as possible as we increase sample size) in the presence of heteroskedasticity. Essentially, in weighted least squares we will be constructing our estimator $\hat{\beta}_1$ to rely more on data-points that we think have a stronger signal of the true parameter β_1. Points that are less noisy will be stronger signals.

There are other contexts for weighting aside from heteroskedasticity specifically (though this is probably the most common motivating example). One should also be aware that weighting enters into the least-squares framework as well as general methods of moments and other estimation frameworks[6]. Since Generalized Least Squares will be covered next semester, I will not go into much detail here, except to make the note that Weighted Least Squares is a special case of Generalize Least Squares and that we often have to rely on a method called Feasible Generalized Least Squares to actually estimate the parameters in such a framework.

Weighted Least Squares: Let's return to our OLS set up. OLS solved the following minimization problem

$$\hat{\beta}_0^*, \hat{\beta}_1^* = \text{argmin}_{\hat{\beta}_0,\hat{\beta}_1}\{\sum_{i=1}^{N}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2\}$$

This minimization problem produced the following estimates:

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{(x_1 - \bar{x})}{\sum(x_i - \bar{x})^2}(y_1 - \bar{y}) + ... + \frac{(x_N - \bar{x})}{\sum(x_i - \bar{x})^2}(y_N - \bar{y})$$

As noted a few times before at this point, for a given y-residual value ($y_i - \bar{y}$), points that have x values further away from $\bar{x}$ will have greater influence on our actual estimator $\hat{\beta}_1$. If we are in a setting with heteroskedasticity, we might want to take into account that some points have more noise than others and thus provide a poorer signal to the relationship between the x and y.

That is, if our population relationship is $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ and we are in a region of the x distribution where ϵ_i varies a ton, we can conclude less from an instance of a really high $y - \bar{y}$ value or a really low $y - \bar{y}$ value because these deviations from the mean y value could have just been produced by a really high or low ϵ_i.

Weighted least squares takes this into account by weighting each observation by its variance.

OLS

algabreic notation (univariate)

$$\hat{\beta}_0^{OLS}, \hat{\beta}_1^{OLS} = \text{argmin}_{\hat{\beta}_0,\hat{\beta}_1} \sum_{i=1}^{N}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

matrix notation

$$\hat{\beta}^{OLS} = \text{argmin}_{\hat{\beta}}[Y - X\hat{\beta}]^T[Y - X\hat{\beta}]$$

WLS

algabreic notation (univariate)

$$\hat{\beta}_0^{WLS}, \hat{\beta}_1^{WLS} = \text{argmin}_{\hat{\beta}_0,\hat{\beta}_1} \sum_{i=1}^{N} \frac{1}{\sigma_i^2}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

matrix notation

$$\hat{\beta}^{WLS} = \text{argmin}_{\hat{\beta}}[Y - X\hat{\beta}]^T\Omega^{-1}[Y - X\hat{\beta}]$$

In the presence of heteroskedasticity, WLS will produce more efficient estimates *if we know* σ_i^2. This is a major limitation of WLS. It is rare to know σ_i^2 ex-ante, and if we are wrong about σ_i^2, bad things can occur:

- We can end up with estimates that are *less consistent* than those that would have been produced under OLS
- Depending how we are parameterizing σ_i^2, we can also end up in the situation where our estimates are also biased.
- Our standard errors can also be incorrect

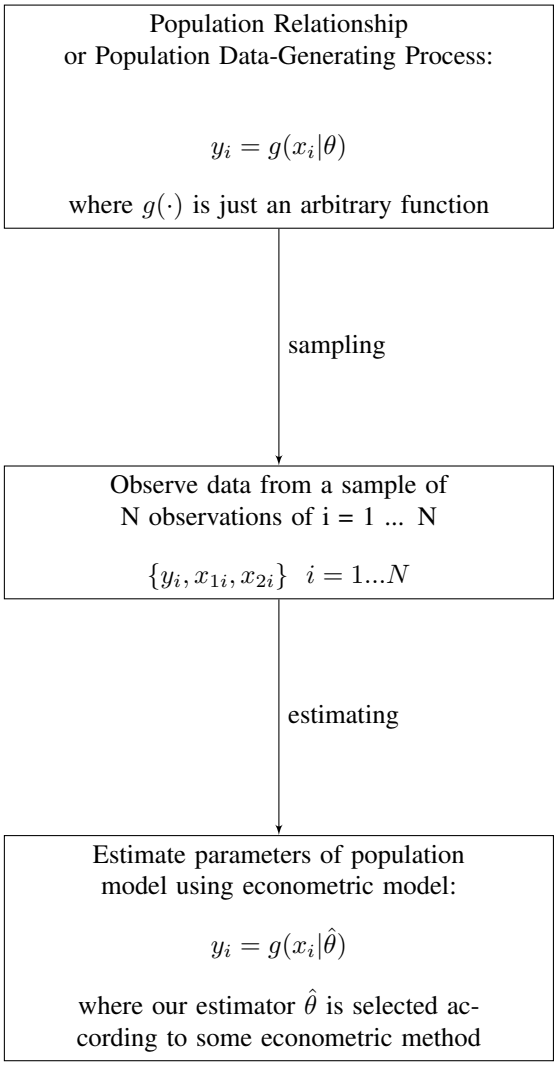
Regardless, WLS still makes sense in many contexts and it's part of a class of estimators (GLS) and deeply connected to an important concept (signal value and weighting) that one should understand as an applied economist.

[6] As we learn about other frameworks for motivating estimators, this will (hopefully) not be too surprising since different motivating frameworks can produce equivalent estimators for β

6 Maximum Likelihood Estimation

The econemtrician's task is to come up with parameter estimators ($\hat{\theta}$) from a sample (X_i, Y_i) that was generated according to some population data-generating-process $Y_i = g(X_i|\theta)$, where $g(\cdot)$ is just some arbitrary function. OLS and WLS methodology imposes the assumption that $g(\cdot)$ is linear. While in some sense OLS is the workhorse of much reduced form empirical work, by restricting our model selection to unbiased linear models, we also restricted the set of data-generating-processes we could accurately capture and thus parameters we could reasonably estimate. Maximum likelihood estimation is a method that allows us to estimate parameters for a broader set of potential data-generating-processes.

General econometric problem framework



6.1 General MLE framework

Let us start with an admittedly convoluted analogy. Consider if there are three possible flavors of ice-cream in the world: chocolate, vanilla, and pistachio. We see a puddle of melted green ice-cream, and we wish to guess what flavor it is without tasting it because of sanitary reasons. One method of arriving at a guess (or estimator) is to guess the flavor which has the highest likelihood of producing a green puddle out of all the possible flavors. That is, for each ice-cream flavor, we think about the probability that that specific ice-cream flavor could melt into a green puddle. We would choose the ice-cream flavor that is associated with the highest conditional probability of producing a green puddle. In this example, because the ice-cream puddle is green, we would guess pistachio. That is, out of the possible puddle-generating-icecream-flavors, pistachio maximizes the probability of observing the actual data that we have collected (the puddle)[7].

In MLE, we are going to choose the estimator value $\hat{\theta}$ that is associated with the highest probability of observing the data that we have in our sample, conditional on that $\hat{\theta}$. For example, consider if our $x_i \sim f(x_i|\theta)$ for some unknown θ and we observe *one observation of* x_i....

$$\hat{\theta}^{MLE} = argmax_{\hat{\theta}} \ f(x_i, y_i|\hat{\theta})$$

More often, we will have many observations of x_i, so the MLE estimator will be the maximizing input for the joint distribution of all these independent x draws

$$\hat{\theta}^{MLE} = argmax_{\hat{\theta}} \underbrace{\prod_{i=1..N} f(x_i, y_i|\hat{\theta})}_{L(\hat{\theta})}$$

We can call $L(\hat{\theta})$ the likelihood function. This is the same object that we used to construct likelihood ratios earlier this semester. Note that *it is not a distribution!!!* That is, it is not a PDF – it would not make sense to integrate it out to get an expected $\hat{\theta}$!!! It is a function of $\hat{\theta}$ that takes the observed data as given. This is distinct from a distribution that characterizes probabilities of random outcomes x_i given a parameter value.[8]

The logic here is that our data was generated using some data-generating-process with parameters θ. It seems reasonable that a good estimator $\hat{\theta}$ of these population parameters would be the ones that are the most likely out of all possible parameters to actually produce the data that we observe.

Why do we need other methods of deriving estimators (i.e. why isn't OLS enough)?

- The simplest answer is that the world is complicated and the data-generating processes that we are trying to understand might not be appropriately characterized by a linear function. For instance, our "y variable" may be a binary variable for which a linear regression might not be appropriate (more on this when we get to probit and logit).
- In some sense, maximum likelihood estimators are also more appealing than a mindless application of OLS since they require one to be careful about specifying what one believes to be the statistical nature of the data generating process.[9]

Is Maximum likelihood estimation a bayesian estimator?

- This is a reasonable point of confusion since likelihood functions can seem quite like conditional probabilities, which are of course a very bayesian concept. However, MLE is a firmly frequentist or classical econometric method and could indeed only be considered a bayesian method as an edge case where the prior over possible parameters is uniform. The equivalent of MLE in a bayesian framework is called Maximum a Posteriori estimation (MAP).
- To see the difference more clearly, let's look at the bayesian approach to deriving an estimator. If we were bayesians, we would derive a distribution function $Pr(\theta|data) = \frac{Pr(data|\theta)Pr(\theta)}{Pr(data)}$ (bayes rule). One typically would get a distribution of possible θ 's with different associated probabilities. We could get a point estimate of $\hat{\theta}$ by just picking $\hat{\theta}^{Bayesian} = argmax_{\hat{\theta}} \ Pr(\hat{\theta}|data)$. Notice that this bayesian approach requires us to take a stand on a prior $Pr(\theta)$.

[7] In some ways this is a stupidly obvious process that our brain does all the time without us consciously thinking about it. If we need to make a guess based off of imperfect information, we try to guess the option that's most consistent with what information we have been given. MLE is just formalizing this.

[8] There is actually a fair bit of writing on this point. If you are interested, a quick google search of " MLE likelihood vs. probability" will result in many forums on the topic. In the interest of brevity, I will leave those further explorations to the reader to do on their own time.

[9] What happens if we get the statistical nature of the DGP wrong? For example, what should we think of a MLE estimator when we have assumed our data is normal when it is actually poisson? The short answer is that MLE will give us the likelihood maximizing estimator *within the class of the assumed statistical distribution* we have assumed for our model, but depending on which statistical distribution the data is generated from and which model we have actually used, our MLE estimator might actually not be so great. In fact, we might want to take a step back and select our model to minimize the impact of mis-specification itself. This motivates a literature on model selection criteria, the most common of which is the Kullbach-Leibler information criterion. A more thorough discussion is outside the scope of this particular set of notes

- MLE instead maximizes $Pr(data|\theta)$. This should produce a point estimate of θ. We can see that our Bayesian $\hat{\theta}^{Bayesian}$ would be the same as our $\hat{\theta}^{MLE}$ if and only if $\frac{Pr(\theta)}{Pr(data)}$ were constant, which would be equivalent with a flat prior (i.e. we have an equal probability of any θ)

What does this actually look like?

- **Setup:** Consider if we observe data points $x_i, i = 1....N$ and we believe that each draw is generated iid from a certain distribution $f(x|\theta)$ and want to estimate $\hat{\theta}$. Let's use MLE

- **Implement MLE:** We choose the estimator $\hat{\theta}$ that maximizes the probability of observing all of the x's that we observe. Remember your basic probability rules that for independent events A and B, $Pr(A \cap B) = Pr(A) * Pr(B)$ or if we're considering continuous distributions, $f(A \cap B) = f(A) * f(B)$.

  Thus, we can write our problem as....
  $\hat{\theta}^{MLE} = argmax_{\hat{\theta}}$ joint likelihood of observing all the x's
  $\hat{\theta}^{MLE} = argmax_{\hat{\theta}}\ f(x_1|\hat{\theta}) * f(x_2|\hat{\theta}) * .... * f(x_N|\hat{\theta})$
  $\hat{\theta}^{MLE} = argmax_{\hat{\theta}}\ \prod_{i=1...N} f(x_i|\hat{\theta})$

  Math trick: A common trick at this point is to take the log of the joint likelihood which turns the product into a summation:

  $$ln(A * B * C) = ln(A) + ln(B) + ln(C)$$

  Because natural log is a *monotonically increasing* function of its arguments, maximizing ln(f(x)) is the same as maximizing f(x)

  Thus, we can write our problem as...
  $\hat{\theta}^{MLE} = argmax_{\hat{\theta}}\ \sum_{i=1...N} ln\{f(x_i|\hat{\theta})\}$

  This is the way to approach most of the analytically tractable MLE problems you will encounter.

Example: MLE of a normal distribution

Consider if we know $x \sim N(\mu, \sigma)$
Suppose that we observe $\{x\} = \{5, 3, 9, 3\}$.
Let $\theta = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix}$.

Use MLE to estimate θ

Implement MLE:

$\hat{\theta}^{MLE} = argmax_{\hat{\theta}} L(\{x\}|\hat{\theta})$
$\hat{\theta}^{MLE} = argmax_{\hat{\theta}} \prod_{i=1,2,3,4} L(x_i|\hat{\theta})$
$\hat{\theta}^{MLE} = argmax_{\hat{\theta}} \prod_{i=1,2,3,4} \frac{1}{\sqrt{\hat{\sigma^2} 2\pi}} exp\{-\frac{(x_i-\hat{\mu})^2}{2\hat{\sigma^2}}\}$
$\hat{\theta}^{MLE} = argmax_{\hat{\theta}} \sum_{i=1,2,3,4} ln[\frac{1}{\sqrt{\hat{\sigma^2} 2\pi}} exp\{-\frac{(x_i-\hat{\mu})^2}{2\hat{\sigma^2}}\}]$
$\hat{\theta}^{MLE} = argmax_{\hat{\theta}} \sum_{i=1,2,3,4} [ln(1) - ln(\sqrt{\hat{\sigma^2} 2\pi}) + -\frac{(x_i-\hat{\mu})^2}{2\hat{\sigma^2}}]$
$\hat{\theta}^{MLE} = argmax_{\hat{\theta}} \sum_{i=1,2,3,4} [-\frac{1}{2} ln(\hat{\sigma^2} 2\pi) + -\frac{(x_i-\hat{\mu})^2}{2\hat{\sigma^2}}]$
$\hat{\theta}^{MLE} = argmax_{\hat{\theta}} \sum_{i=1,2,3,4} [-ln(\hat{\sigma}) - \frac{1}{2} ln(2) - \frac{1}{2} ln(\pi) - \frac{(x_i-\hat{\mu})^2}{2\hat{\sigma^2}}]$

Great, now let's take FOC:

$\frac{\partial}{\partial \hat{\mu}} = \frac{\sum_{i=1,2,3,4} -2(x_i-\hat{\mu})(-1)}{2\hat{\sigma^2}}$
$0 = \frac{\sum_{i=1,2,3,4} 2(x_i-\hat{\mu})}{2\hat{\sigma^2}}$
$0 = \sum_{i=1,2,3,4} (x_i - \hat{\mu})$
$\hat{\mu} = \frac{1}{4} \sum_{i=1,2,3,4} x_i$ (We can recognize this as the mean formula!)
$\hat{\mu} = 5$

$\frac{\partial}{\partial \hat{\mu}} = \sum_{i=1,2,3,4} [-\frac{1}{\hat{\sigma}} + 2\frac{(x_i-\hat{\mu})^2}{2\hat{\sigma^3}}]$
$0 = \sum_{i=1,2,3,4} [-\frac{1}{\hat{\sigma}} + \frac{(x_i-\hat{\mu})^2}{\hat{\sigma^3}}]$
$\sum_{i=1,2,3,4} \frac{1}{\hat{\sigma}} = \sum_{i=1,2,3,4} \frac{(x_i-\hat{\mu})^2}{\hat{\sigma^3}}$
$\sum_{i=1,2,3,4} \hat{\sigma^2} = \sum_{i=1,2,3,4} (x_i - \hat{\mu})^2$
$\hat{\sigma^2} = \frac{1}{4} \sum_{i=1,2,3,4} (x_i - \hat{\mu})^2$ (We can recognize this as the variance formula!)
$\hat{\sigma^2} = \frac{1}{4}((5-5)^2 + (3-5)^2 + (9-5)^2 + (3-5)^2)$
$\hat{\sigma^2} = \frac{1}{4}(0 + 4 + 16 + 4)$
$\hat{\sigma^2} = \frac{1}{4}(24)$
$\hat{\sigma^2} = 6$

Perhaps comfortingly, we see that the MLE estimates of $\hat{\mu}$ and $\hat{\sigma}$ correspond to the mean and variance of our sample in this case! This seems reasonable given what MLE is doing: the distribution that is most likely to give us data with mean 5 and variance 6 is a distribution that also has mean 5 and variance 6. You probably would have guessed this even before you knew MLE.

An important concept to recognize at this point is that often a given estimator can be derived using different methods (i.e. MLE or some other approach). You should think Maximum Likelihood Estimation as just another tool in your tool-belt as an empiricist trying to characterize Data-Generating-Processes world.

6.2 Logit and Probit

Two of the most common econometric models that you will encounter are logit and probit. These are used often in the context of *discrete outcomes* or *discrete choice*, such as when we only have some number "N" outcomes or are a consumer choosing between N products. Notice that this nests binary outcomes, which are super common in economic analysis. Logit is the workhorse model in industrial organization, where the seminal work is Berry, Levinsohn, and Pakes (1994) or BLP (which also includes random coefficients but is based on the logit econometric model).[10]

Some random thoughts before we talk more about probit and logit

- Binary or categorical? For the remainder of these notes, I will be focusing on models of binary outcome variables, but these versions of the logit and probit models can also be adapted to handle categorical outcome variables that can take on more than two values. Check out Kenneth Train's textbook (referenced in the appendix) for more info on this.
- Why do we like probit or logit better than OLS for binary variables? There are actually a few reasons, but some of the most common reasons cited are as follows:
  - probit and logit models don't impose an assumption that a given independent variable has a constant impact on an outcome. When we run a linear regression
    $$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$$
    where y is a binary variable, we want to interpret $\hat{\beta}_1$ as the change in probability that Y = 1 given a value of x_1. If x_1 is continuous, a change of 1 unit change of x_1 from 1 to 2 has the same impact as a 1 unit change from 899 to 900. There are plenty of examples where this is probably not the case.
  - Linear models can also produce fitted values $haty$ greater than 1 or less than 0. Clearly, we can't have a probability greater than 1 or less than 0, so this is not super desirable. Because of their shape (graphs in following pages), probit and logit models do not have this problem.
- Estimation method? MLE is a natural approach to estimating the parameters of probit and logit models, but they could also be estimated using what is called Generalized Method of Moments, which will be covered later in the semester, and I believe next semester, too.

[10] For a nice summary on demand estimation, please see excellent write-up by Harvard Economics PhD candidate Frank Pinter here

6.2.1 Binary Logit

(c) Logit problem framework

Population Relationship
or Population Data-Generating Process:

$$y_i = \begin{cases} 1 & \beta_0 + \beta_1 x_{1i} + ... + \beta_j x_{ji} + \epsilon_i > 0 \\ 0 & otherwise \end{cases}$$

Where $\epsilon_i \sim$ logistic distribution

sampling

Observe data from a sample of
N observations of i = 1 ... N

$\{y_i, x_{1i}, x_{2i}\} \quad i = 1...N$

estimating

Estimate parameters of population
model using econometric model:

$$\hat{y}_i = \begin{cases} 1 & \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + ... + \hat{\beta}_J x_{ji} + \epsilon_i \\ 0 & otherwise \end{cases}$$

Where $\epsilon_i \sim$ logistic distribution where our estimator $\hat{\beta}$ is selected according to some econometric method (often MLE)

Let's assume that we are considering a binary variable y, which we believe to be determined by independent variables $x_1, ...x_j$. Instead assuming a constant impact of variable x on the likelihood of y being 1, we will use a bit more complicated specification that will allow for a non-constant impact.

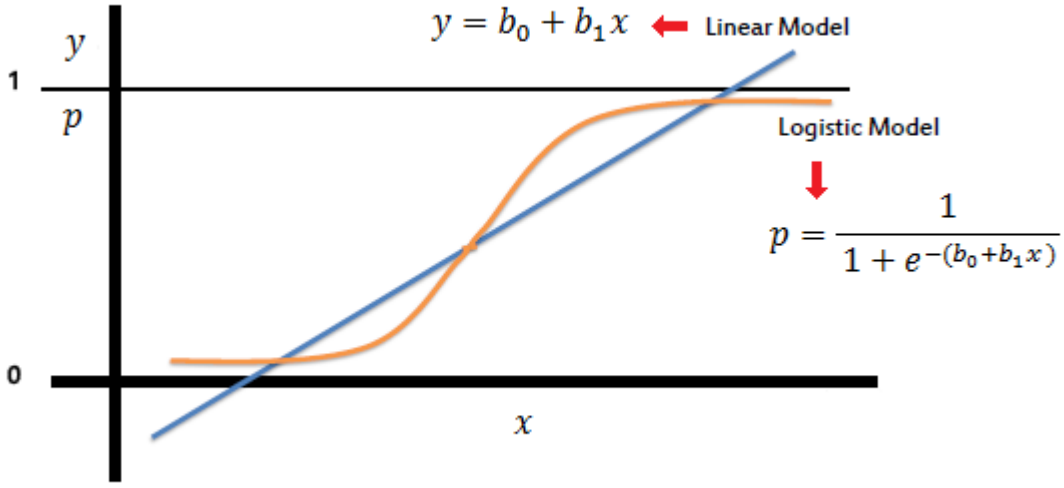
We now have a condition that must be satisfied for y to be equal to 1. We could think of it as some threshold. For instance, if y is a dummy variable that indicates whether or not our company invests in something, $\beta_0 + \beta_1 x_{1i} + ... + \beta_J x_{ji} + \epsilon$ could indicate our profit function, so that we are only investing if our profit from doing so is greater than zero.

Let us rearrange this threshold condition to get a nice expression for $Pr(Y_i = 1|x_i)$

$Pr(y_i = 1|x_i) = Pr(\beta_0 + \beta_1 x_{1i} + ... + \beta_J x_{ji} + \epsilon_i > 0)$
$Pr(y_i = 1|x_i) = Pr(X\beta + \epsilon_i > 0)$
$Pr(y_i = 1|x_i) = Pr(\epsilon_i > -X\beta)$
$Pr(y_i = 1|x_i) = 1 - Pr(\epsilon_i < -X\beta)$
$Pr(y_i = 1|x_i) = 1 - F(-X\beta)$ where F is the logistic CDF
$Pr(y_i = 1|x_i) = 1 - \frac{1}{1+exp\{X\beta\}}$
$Pr(y_i = 1|x_i) = \frac{1+exp\{X\beta\}}{1+exp\{X\beta\}} - \frac{1}{1+exp\{X\beta\}}$
$Pr(y_i = 1|x_i) = \frac{exp\{X\beta\}}{1+exp\{X\beta\}}$
$Pr(y_i = 1|x_i) = \frac{1}{1+exp\{-X\beta\}}$

Great! You will often see the logit model presented this way.

When it comes to estimating, we will observe our x_{ji}'s and y_i's per usual and use an econometric method in order to generate estimators $\hat{\beta}_j$ of the population parameters β_j (assuming our population model is accurate). MLE is a natural approach here. Stata has a convenient command 'logit' which you can use just like the 'reg' command. If you have a categorical outcome that can take on more than two values, then you can use 'ologit', which stands for "ordered logistic" regression.



6.2.2 Binary Probit

(D) Probit problem framework

Population Relationship
or Population Data-Generating Process:

$$y_i = \begin{cases} 1 & \beta_0 + \beta_1 x_{1i} + ... + \beta_J x_{ji} + \epsilon_i > 0 \\ 0 & otherwise \end{cases}$$

Where $\epsilon_i \sim N(0,1)$

↓ sampling

Observe data from a sample of
N observations of i = 1 ... N

$\{y_i, x_{1i}, x_{2i}\} \quad i = 1...N$

↓ estimating

Estimate parameters of population
model using econometric model:

$y_i = g(x_i|\hat{\theta})$

where our estimator $\hat{\theta}$ is selected according to some econometric method

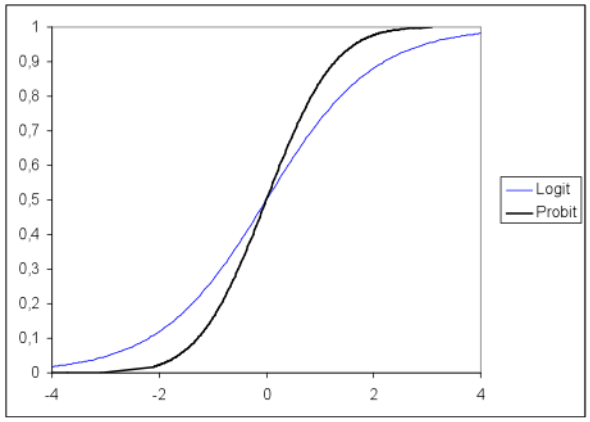
Let's still assume that we are considering a binary variable y, which we believe to be determined by some threshold rules based off of independent variables $x_1, ...x_j$. The only difference between the logit and probit here is that we are now assuming the ϵ_i is normally distributed.

Once again, let us rearrange our threshold to derive a probability
$Pr(y_i = 1|x_i) = Pr(\beta_0 + \beta_1 x_{1i} + ... + \beta_J x_{ji} + \epsilon_i > 0)$
$Pr(y_i = 1|x_i) = Pr(X\beta + \epsilon_i > 0)$
$Pr(y_i = 1|x_i) = Pr(\epsilon_i > -X\beta)$
$Pr(y_i = 1|x_i) = Pr(\epsilon_i < X\beta)$ (because the normal distribution is symmetric)
$Pr(y_i = 1|x_i) = \Phi(X\beta)$, where $\Phi(\cdot)$ is the Normal CDF

Again, you will often see the probit model motivated in this way.

As before, we will observe our x_{ji}'s and y_i's and use an econometric model in order to generate estimators $\hat{\beta}_j$ of the population parameters β_j (assuming our population model is accurate). Once again, MLE would be a good choice here. We can use the 'probit' command in stata for binary outcomes (oprobit works for multiple outcomes, standing for an "ordered probit" regression).

Probit v. logit: We can plot the probit and logit functions below and see that they are in a sense very similarly shaped. In many reduced form applications, there will not be a big difference between logit and probit and it usually makes sense to just choose one and roll with it. However, there are other situations where a certain model will have better properties than others or ones that make more sense than others. In particular, for structural models, the choice of the error term can make an enormous difference, not just for tractability but also for internal consistency and reasonability of parameter estimates. Please see Kenneth Train's text on Probit and Logit (referenced in appendix) for more information on this.



A Additional Resources

A.1 General notes

These are front-to-back excellent resources to have around to 'ctrl+f' rough topics.

- Ben Lambert's course in econometerics
- William Greene's Econometrics Course. His textbook is also excellent, so if you have access to it, I would highly reccomend giving it a look.
- Suhasini Subba Rao's advanced statistical inference course
- Cosma Shalizi's course on linear statistical models

A.2 Notes on specific topics

- Probability
  - PennState's public probability notes
- Bayesian statistics
  - Larry Wasserman's notes
- Decision Theory
  - Richard Bradley's book, Decision Theory with a Human Face (this probably would not be an efficient resource for the purposes of this class, but I found it quite interesting and relates to much that we discuss!)
- Maximum Likelihood Estimation
  - Ben Lambert's MLE video series
  - Ryan Martin's notes on Likelihood and Maximum Likelihood Estimation, which I have personally found incredibly clear and helpful.
  - Suhasini Subba Rao's notes on the Kullbach-Leibler information criterion (her website also has an excellent set of statistics notes on a broad range of topics, if you are interested).
- Structural modeling and demand systems
  - Kenneth Train's generously public textbook on discrete choice methods using simulation. This is a quite advanced text, but the introduction sections on logit and probit are written in a pretty understandable way that can provide some context.
  - Frank Pinter's overview of demand estimation in the context of industrial organization. This is helpful for people who want to understand what is involved in structural estimation, particularly BLP demand estimation.