Section Notes for POLSCI 699: Statistical Methods in Political Research II

Patrick Y. Wu

Winter 2018

Note: I have written, but not necessarily read, these notes. Please let me know if you see any typos.

1 Thursday, January 11

1.1 Welcome / Introduction

- Patrick Wu. E-mail: pywu@umich.edu. Office: Haven Hall 6564.
- I will post section notes right after section. This is not a reason to not come to section.
- Office Hours: I will have office hours on Wednesday from 4 to 5pm and Thursdays from 5 to 7pm. If you can't meet me during these times, just e-mail me and we'll schedule another time to meet.

1.2 LaTeX

- All homeworks have to be completed in LaTeX and uploaded to Canvas. I'll show you how to upload things to Canvas if you need me to.
- I like to use Sublime Text 2 for typing up my LaTeX documents
- Since there are a lot of little exceptions for each individual installations, here are the guides for the Windows and Mac versions
  - Windows: `http://economistry.com/2012/10/first-pdf-sublime-text-2-latex/`
  - Mac: `http://economistry.com/2013/01/installing-and-using-latex-for-mac/`
- Feel free to use another way to type up your LaTeX documents, but this is just how I do it
- If you run into any troubles during installation, just let me know.

1.3 `R`

- Personally, I use RStudio to run all my `R` code.
- Many people, including Rocio and Walter, do not use RStudio.
- To download `R`, go to `https://cran.r-project.org/mirrors.html`

- To download RStudio, go to `https://www.rstudio.com/products/rstudio/download/`
- Basically, RStudio is a nice wrapper over the `R` installation
- I like it because you can run code and explore datasets
- To learn how to use `R`, there are TONS of resources out there
- Here is a good resource: `https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf`

1.4 Basics of `R`: Linear Algebra and Linear Regressions

- Creating a Matrix: `A = matrix(c(1,1,1,2,3,1,1,1,1), nrow = 3, ncol = 3, byrow = TRUE);`
- Element-wise multiplication: `A*B`
- Matrix Multiplication: `A%*%B`
- Cross-Product: `crossprod(A,B)` ($A'B$) or `crossprod(A)` ($A'A$)
- Transpose: `t(A)`
- Inverse: `solve(A)`
- Eigenvalues and Eigenvectors: `y = eigen(A)`. use `y$val` for eigenvalues and `y$vec` for eigenvectors.
- Linear Regressions: `lm` function

1.5 Virtual Sites

- Information about virtual sites: `http://www.itcs.umich.edu/sites/labs/virtual.php`
- I will show you how to log in and access resources

1.6 Submitting Assignments

- You will submit all assignments on Canvas.
- I will show you how to upload an assignment.
- You can also leave comments or questions about your assignment after you get it back on Canvas. Feel free to rage me out there.

1.7 Probability: The Intuition

The probability space is the mathematical description of an experiment. This experiment could be tossing a coin, rolling dice, taking a number in a lottery, or getting up in the morning, opening a window, and checking the weather. In each case what is key is that there is a *certain amount of randomness or unpredictability* in the experiment. As Rocio said in class, the probability space is a triple: $(\Omega, \mathcal{F}, P)$ where Ω is a set, called the sample space; $\mathcal{F}$ is a σ-field of subsets of Ω, and P is a probability measure on $\mathcal{F}$. Elements of $\mathcal{F}$ are called events. We can think of Ω as the set of all

possible outcomes of the experiment: each point of Ω represents an outcome. An event is a set of outcomes. The probability measure P gives the probability of events. For instance, the probability that the outcome falls in a given event A is $P\{A\}$. The collection of all events is the σ-field $\mathcal{F}$. Thus, the triple $(\Omega, \mathcal{F}, P)$ can be thought of as a mathematical description of an experiment: the elements of Ω are the possible outcomes, the elements of $\mathcal{F}$ are the events, and P gives us the probability of all events—and in particular, the probabilities of the outcomes. This means that, informally,

$$P(E) = \frac{|E|}{|\Omega|}$$

We can either think of this as the number of elements of event E over the number of elements of the sample space Ω for discrete cases, or we think of this as the area of E over the area of Ω.

1.7.1 Example 1: Picking Numbers

Suppose I wrote down a number between $[0, 1]$ at random and you also write down a number between $[0, 1]$ at random. The probability space would thus look like a square. [DRAW SQUARE HERE]. Let's say my number is X and your number is Y. (1) What is the probability that you choose a number greater than me? (2) What is the probability that I choose a number greater than you? (3) What is the probability that I pick a number between 0.4 and 0.6 and you pick a number between 0.8 and 1.0? (4) What is the probability that you pick a number that is at least $\frac{1}{3}$ less than my number?

(1) The probability that you choose a number greater than me is formally written as: $E_1 = \{(X, Y) \in \Omega : Y > X\} = \frac{1}{2}$. Visually, we can draw the line through the box and see that all points above the line are eligible. So the area of the triangle is $\frac{1}{2}$ which matches our probability calculation.

(2) The probability that you choose a number less than me is formally written as: $E_2 = \{(X, Y) \in \Omega : Y < X\} = \frac{1}{2}$. Visually, we can draw the line through the box and see that all points below the line are eligible. So the area of the triangle is $\frac{1}{2}$ which matches our probability calculation.

(3) The probability that I pick a number between 0.4 and 0.6 is 0.2. The probability that you pick a number between 0.8 and 1.0 is 0.2. Formally, this event can be written as: $E_3 = \{(X, Y) \in \Omega : 0.4 < X < 0.6, 0.8 < Y < 1.0\}$. Because they are independent events (my picking a number doesn't affect you), the probability of both events occurring is $0.2 \times 0.2 = 0.04$. Visually, we can see this as a square in the larger $[0, 1] \times [0, 1]$ square. The area of this square, which is $[0.4, 0.6] \times [0.8, 1.0]$, is 0.04, which matches our probability calculation.

(4) The probability that you pick a number that is at least $\frac{1}{3}$ less than my number means that you want to pick a number Y such that $Y + \frac{1}{3} < X$. Formally, this is written as: $E_4 = \{(X, Y) \in \Omega : Y < X - \frac{1}{3}\}$. Drawing this out, the area of the triangle is $\frac{2}{3} \times \frac{2}{3} \times \frac{1}{2} = \frac{2}{9}$.

2 Thursday, January 18

2.1 Random Variables: Part I

A popular saying is that a random variable is neither random nor a variable. A random variable X is a function from S to the real numbers. $X : S \to \mathbb{R}$. The distribution function (or cumulative distribution function or c.d.f.) of X is $F_X(x) = \mathbb{P}(X \leq x)$, so $F_X : \mathbb{R} \to [0, 1]$. For continuous random variables, we know that

$$F_X(x) = \int_{-\infty}^{x} f_X(x)dx$$

And by the Fundamental Theorem of Calculus, we have

$$f_X(x) = \frac{d}{dx}F_X(x)$$

This is more consequential than we think. We can apply this definition to answer the following question.

2.1.1 Example 2: Uniform Distributions and Random Variables

The continuous uniform distribution is defined by spreading mass uniformly over an interval $[a, b]$. Its pdf is given by

$$f(x|a, b) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$

Let $X \sim Unif(0, 1)$ and $Y \sim Unif(0, 1)$. Assume X and Y are independent. What is $\mathbb{P}(Y < X - \frac{1}{3})$? We need to set up the limits of integration properly. We need to integrate over all values of Y, which is from 0 to $X - \frac{1}{3}$. And then we need to integrate over the values of X that we are interested in. Is this 0 to 1? No! It is from $\frac{1}{3}$ to 1 because if X is less than $\frac{1}{3}$ then we'll be examining values of Y that are negative. We need to be really, really careful withour bounds of integration. We also know that $f(x, y) = f(x)f(y) = 1$, since the two distributions are independent. Then this becomes

$$\int_{1/3}^{1}\int_{0}^{x-1/3} 1dydx = \int_{1/3}^{1} x - \frac{1}{3}dx = \frac{1}{2}x^2 - \frac{1}{3}x|_{1/3}^{1} = \frac{1}{6} - (\frac{1}{18} - \frac{1}{9}) = \frac{3}{18} - (-\frac{1}{18}) = \frac{4}{18} = \frac{2}{9}$$

2.1.2 Example 3: Finding the Mean and Variance of the Normal Distribution

Sometimes known as the Gaussian distribution, its pdf is

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}exp(-(x - \mu)^2/(2\sigma^2))$$

Now we want to show that the mean is equal to μ and its variance is equal to σ^2. To solve for the mean, we use the Law of the Unconscious Statistician (but some people hate this phrase, like Casella and Berger) and calculate:

$$E[X] = \int_{-\infty}^{\infty} \frac{x}{\sqrt{2\pi}\sigma}exp(-(x - \mu)^2/(2\sigma^2))dx$$

This is actually way trickier than it looks at first! We first let $z = \frac{x-\mu}{\sigma}$. This means that $x = z\sigma + \mu$. This also means that $dx = \sigma dz$. Thus, using the change of variables in integration we get

$$\int_{-\infty}^{\infty} \frac{z\sigma + \mu}{\sqrt{2\pi}\sigma} exp(-z^2/2)dz = \int_{-\infty}^{\infty} \frac{z\sigma}{\sqrt{2\pi}\sigma} exp(-z^2/2)\sigma dz + \int_{-\infty}^{\infty} \frac{\mu}{\sqrt{2\pi}\sigma} exp(-z^2/2)\sigma dz = \mu$$

Notice that the first integral is an odd function—meaning that the function is symmetric around the origin. So therefore, it evaluates to 0. We can pull the μ out of the integral and we're just integrating over the entire support of the density, which is just 1.

Now let's find the variance. We defined variance as $E[(X-\mu)^2]$. Using LOTUS again, we get

$$\int_{-\infty}^{\infty} \frac{(x-\mu)^2}{\sqrt{2\pi}\sigma} exp(-(x-\mu)^2/(2\sigma^2))dx$$

Again, we need a trick to solve this problem! Let $z = \frac{x-\mu}{\sigma}$. So this means that $x = z\sigma + \mu$; we also know that this means that $dx = \sigma dz$. Replacing this into the integral, we get

$$\int_{-\infty}^{\infty} \frac{z^2\sigma^2}{\sqrt{2\pi}\sigma} exp(-z^2/2)\sigma dz$$

Notice first that the σ cancel out. We need σ^2 in the numerator because remember that $z = \frac{x-\mu}{\sigma}$. Now pull out the σ^2.

$$\sigma^2 \int_{-\infty}^{\infty} \frac{z^2}{\sqrt{2\pi}} exp(-z^2/2)dz$$

Now, notice something very peculiar about this integral. This is actually $E[Z^2]$ of the standard normal distribution, or, $N(0,1)$! So that means we know that $E[Z^2] = 1$ because $E[Z^2] - E[Z]^2 = 1$ and $E[Z] = 0$ in the standard normal distribution. So that means the integral equals 1, so that means

$$\sigma^2 \int_{-\infty}^{\infty} \frac{z^2}{\sqrt{2\pi}} exp(-z^2/2)dz = \sigma^2$$

which verifies that this is the variance of the normal distribution.

2.2 Transformation of Random Variables: Distribution Function Technique

Definition of Transformation: We are often interested in the probability distributions or densities of functions of one or more random variables. Suppose we have a set of random variables $X_1, X_2, ..., X_n$ with a known joint probability and/or density function. We may want to know the distribution of some function of these random variables $Y = \phi(X_1, X_2, ..., X_n)$. Realized values of Y will be related to the realized values of the X_i for $i \in \{1, 2, ..., n\}$ as follows:

$$y = \Phi(x_1, x_2, ..., x_n)$$

A simple example might be a single random variable x with the transformation $y = \Phi(x) = log(x)$.

Procedure for Using the Distribution Function Technique: We find the region in the $x_1, x_2, ..., x_n$ space such that $\Phi(x_1, x_2, ..., x_n) \leq \phi$. We can then find the probability that $\Phi(x_1, x_2, ..., x_n) \leq \phi$, i.e. $\mathbb{P}[\Phi(x_1, x_2, ..., x_n) \leq \phi]$ by integrating the density function $f(x_1, x_2, ..., x_n)$ over this region. Of course, $F_\Phi(\phi)$ is just $\mathbb{P}(\Phi \leq \phi)$. Once we have $F_\Phi(\phi)$ we can find the density by integration.

2.2.1 Example 1

Let the probability density funciton of X be given by

$$f(x) = \begin{cases} 6x(1-x) & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

Now let's find the probability density of $Y = X^3$.

$$\begin{aligned} G(y) &= P(Y \leq y) \\ &= P(X^3 \leq y) \\ &= P(X \leq y^{1/3}) \\ &= \int_0^{y^{1/3}} 6x(1-x)dx \\ &= \int_0^{y^{1/3}} 6x - 6x^2 dx \\ &= (3x^2 - 2x^3)\Big|_0^{y^{1/3}} \\ &= 3y^{2/3} - 2y \end{aligned}$$

We can differentiate this to get the density function, which is $g(y) = 2(y^{-1/3} - 1)$ for $0 < y < 1$. Remember to include the bounds when doing these types of problems!

2.3 Transformation of Random Variables: Method of Transformations (Inverse Mappings)

Procedure for Using the Method of Transformations for a Single Random Variable: Let X have pdf $f_X(x)$ and let $Y = g(X)$, where g is a monotone function. Let $\mathcal{X}$ and $\mathcal{Y}$ be defined such that $\mathcal{X} = \{x : f_X(x) > 0\}$ and $\mathcal{Y} = \{y : y = g(x) \text{ for some } x \in \mathcal{X}\}$. Suppose that $f_X(x)$ is continuous on $\mathcal{X}$ and that $g^{-1}(y)$ has a continuous derivative on $\mathcal{Y}$. Then the pdf of Y is given by

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| & \text{for } y \in \mathcal{Y} \\ 0 & \text{otherwise} \end{cases}$$

We proved this theorem in class. Let's do an example.

2.3.1 Example 2: Inverted Gamma PDF

Let $f_X(x)$ be the gamma pdf

$$f(x) = \frac{1}{(n-1)!\beta^n} x^{n-1} e^{-x/\beta}, 0 < x < \infty$$

where β is a positive constant and n is a positive integer. Suppose we want to find the pdf of $g(X) = 1/X$. Note here that the support sets $\mathcal{X}$ and $\mathcal{Y}$ are both the interval $(0, \infty)$. Let $y = g(x)$,

so that means $g^{-1}(x) = 1/y$ and $\frac{d}{dy}g^{-1}(y) = -1/y^2$. Now let's simply apply the theorem above.

$$\begin{aligned} f_Y(y) &= f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| \\ &= \frac{1}{(n-1)!\beta^n} \left(\frac{1}{y}\right)^{n-1} e^{-1/(\beta y)} \frac{1}{y^2} \\ &= \frac{1}{(n-1)!\beta^n} \left(\frac{1}{y}\right)^{n+1} e^{-1/(\beta y)} \end{aligned}$$

This is known as the inverted gamma pdf.

2.3.2 Example 3: Multiple Functions of Multiple Random Variables

Theorem: Let $f_{X_1,X_2}(x_1, x_2)$ be the value of the joint probability density of the continuous random variables X_1 and X_2 at (x_1, x_2). If the functions given by $y_1 = u_1(x_1, x_2)$ and $y_2 = u_2(x_1, x_2)$ are partially differentiable with respect to x_1 and x_2 and represent a one-to-one transformation for all values within the range of X_1 and X_2 for which $f_{X_1,X_2}(x_1, x_2) \neq 0$, then, for these values of x_1 and x_2 the equations $y_1 = u_1(x_1, x_2)$ and $y_2 = u_2(x_1, x_2)$ can be uniquely solved for x_1 and x_2 to give $x_1 = w_1(y_1, y_2)$ and $x_2 = w_2(y_1, y_2)$ and for corresponding values of y_1 and y_2 the joint probability density of $Y_1 = u_1(X_1, x_2)$ and $Y_2 = u_2(X_1, X_2)$ is given by

$$f_{Y_1,Y_2}(y_1, y_2) = f_{X_1,X_2}[w_1(y_1, y_2), w_2(y_1, y_2)] \cdot |J|$$

where $|J|$ is the Jacobian of the transformation and is defined as the determinant

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{vmatrix}$$

At all other points, $f_{Y_1,Y_2}(y_1, y_2) = 0$.

Let's do an example.

Let the probability density function of X_1 and X_2 be given by

$$f_{X_1,X_2} = \begin{cases} e^{-(x_1+x_2)} & \text{for } x_1 \geq 0, x_2 \geq 0 \\ 0 & else \end{cases}$$

Consider two random variables Y_1 and Y_2 be defined in the following manner

$$Y_1 = X_1 + X_2$$

$$Y_2 = \frac{X_1}{X_1 + X_2}$$

To find the joint density of Y_1 and Y_2, we first need to solve the system of equations for X_1 and X_2.

$$\begin{aligned} X_1 &= Y_1 - X_2 \\ Y_2 = \frac{Y_1 - X_2}{Y_1 - X_2 + X_2} &= \frac{Y_1 - X_2}{Y_1} \\ Y_1 Y_2 &= Y_1 - X_2 \end{aligned}$$

$$X_2 = Y_1 - Y_1Y_2 = Y_1(1 - Y_2)$$
$$X_1 = Y_1 - (Y_1 - Y_1Y_2) = Y_1Y_2$$

So now we solve the Jacobian

$$\begin{aligned} J &= \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{vmatrix} \\ &= \begin{vmatrix} y_2 & y_1 \\ 1 - y_2 & -y_1 \end{vmatrix} \\ &= -y_2y_1 - y_1(1 - y_2) \\ &= -y_2y_1 - y_1 + y_1y_2 \\ &= -y_1 \end{aligned}$$

Notice that this transformation is one-to-one and maps the domain of X in the (x_1, x_2) plane into the domain of Y in the (y_1, y_2) plane given by $y_1 \geq 0$ and $0 \leq y_2 \leq 1$. Now, plug everything into the theorem.

$$\begin{aligned} f_{Y_1,Y_2}(y_1, y_2) &= f_{X_1,X_2}[w_1(y_1, y_2), w_2(y_1, y_2)] \cdot |J| \\ &= e^{-(y_1y_2 + y_1 - y_1y_2)} | - y_1| \\ &= y_1e^{-y_1} \end{aligned}$$

Considering all possible values of y_1 and y_2 we obtain

$$f_{Y_1,Y_2}(y_1, y_2) = \begin{cases} y_1e^{-y_1} & \text{for } y_1 \geq 0, 0 \leq y_2 \leq 1 \\ 0 & \text{elsewhere} \end{cases}$$

We can then finding the marginal density of Y_1 and Y_2 using techniques covered before.

3 Thursday, January 25

3.1 The Normal Distribution

3.1.1 Show that the height of the normal curve is maximized at $x = \mu$.

We know that

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right]$$

$$f'(x) = \frac{1}{\sqrt{2\pi}\sigma} exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right]\left[-\frac{x-\mu}{\sigma^2}\right]$$

Set this equal to 0. Notice that

$$\frac{1}{\sqrt{2\pi}\sigma} exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right] \neq 0$$

so we can just divide this out. We're left with

$$-\frac{x-\mu}{\sigma^2} = 0 \Rightarrow x^* = \mu$$

Now we have to show that this is a maximum point. Take the second derivative.

$$f''(x) = \frac{1}{\sqrt{2\pi}\sigma} exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right]\left[-\frac{x-\mu}{\sigma^2}\right]^2 + \frac{1}{\sqrt{2\pi}\sigma} exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right]\left[-\frac{1}{\sigma^2}\right]$$

Now if we evaluate this derivative at $x = \mu$, we see that the first part of the second derivative just becomes 0. The second part evaluates to

$$\frac{1}{\sqrt{2\pi}\sigma}\left[-\frac{1}{\sigma^2}\right] < 0$$

because $\frac{1}{\sqrt{2\pi}\sigma} > 0$ always and $-\frac{1}{\sigma^2} < 0$ always. So $x^* = \mu$ is the max point on the normal curve.

3.1.2 Show that the normal curve has inflection points at $\mu \pm \sigma$.

From the previous part, we know that

$$f''(x) = \frac{1}{\sqrt{2\pi}\sigma} exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right]\left[-\frac{x-\mu}{\sigma^2}\right]^2 + \frac{1}{\sqrt{2\pi}\sigma} exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right]\left[-\frac{1}{\sigma^2}\right]$$

Set this equal to 0. Again, we know that

$$\frac{1}{\sqrt{2\pi}\sigma} exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right] \neq 0$$

so this cancels out from both parts of the second derivative. So we're left with

$$\left[-\frac{x-\mu}{\sigma^2}\right]^2 + \left[-\frac{1}{\sigma^2}\right] = 0$$

Simplify to get

$$(x-\mu)^2 = \sigma^2$$

Which implies that

$$x - \mu = \pm\sigma$$

which thus implies that

$$x = \mu \pm \sigma$$

Which means that the normal curve has inflection points at $\mu \pm \sigma$.

3.1.3 Multivariate Normal

We observe a random vector $\boldsymbol{X} = [X_1, X_2, ..., X_n]^T$. Its expectation is $E[\boldsymbol{X}] = \boldsymbol{\mu}$, where $\boldsymbol{\mu} = [E[X_1], E[X_2], ..., E[X_n]]^T$. The covariance matrix (the variance-covariance matrix, or the dispersion matrix) is defined by $\boldsymbol{\Sigma} = E\left[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T\right]$. Then, the random vector $\boldsymbol{X}$ is said to have the (n-variate) multivariate normal distribution with mean $\boldsymbol{\mu}$ and a (positive semi-definite) covariance matrix $\boldsymbol{\Sigma}$ if its probability density function $\boldsymbol{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is given by

$$f(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right)$$

Show that the multivariate normal distribution is well-defined, i.e. for any $\boldsymbol{\mu} \in \mathbb{R}^n$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{n\times n}$ semi-positive definite, the following holds:

$$\frac{1}{(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right) dx_1 dx_2 ... dx_n = 1$$

(Note: I don't bold vectors in this solution)
We want to show that

$$(2\pi)^{-n/2}|\Sigma|^{-1/2} \int_{\mathbb{R}^n} exp(-\frac{1}{2}(x - \mu)^T\Sigma^{-1}(x - \mu))dx = 1 \qquad (\star)$$

Let us focus on the following integral

$$\int_{\mathbb{R}^n} exp(-\frac{1}{2}(x - \mu)^T\Sigma^{-1}(x - \mu))dx$$

Let $z = \Sigma^{-1/2}(x - \mu)$. This means that

$$\Sigma^{1/2}z = x - \mu$$

which means that

$$x = \Sigma^{1/2}z + \mu$$

The Jacobian of the transformation is:

$$\begin{vmatrix} \frac{\partial x_1}{\partial z_1} & \cdots & \frac{\partial x_1}{\partial z_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial z_1} & \cdots & \frac{\partial x_n}{\partial z_n} \end{vmatrix} = |\Sigma^{1/2}| = |\Sigma|^{1/2}$$

Changing the variables of the integration, we get

$$\begin{aligned} &\int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} exp[-\frac{1}{2}(x - \mu)^T\Sigma^{-1}(x - \mu)]dx_1 ... dx_n \\ &= \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} exp[-\frac{1}{2}(\Sigma^{1/2}z)^T\Sigma^{-1}(\Sigma^{1/2}z)]|\Sigma|^{1/2}dz_1 ... dz_n \\ &= |\Sigma|^{1/2} \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} exp[-\frac{1}{2}z^Tz]dz_1 ... dz_n \end{aligned}$$

The integral is just n integrals of independent 1D univariate Gaussians. So:

$$
\begin{aligned}
&= |\Sigma|^{1/2} \prod_{i=1}^{n} \int_{-\infty}^{\infty} exp[-\frac{1}{2} z_i^2] dz_i \\
&= |\Sigma|^{1/2} \prod_{i=1}^{n} (2\pi)^{1/2} \\
&= |\Sigma|^{1/2} (2\pi)^{n/2}
\end{aligned}
$$

This cancels out with the fraction at the front of the integral of ($\star$). Thus, we showed that ($\star$) equals 1.

4 Thursday, February 1

4.1 Some Inequalities

Jensen's Inequality: For any random variable X, if $g(x)$ is a convex function then

$$E[g(X)] \geq g(E[X])$$

Equality holds if and only if, for every line $a + bx$ that is tangent to $g(x)$ at $x = E[X]$, $P(g(X) = a + bX) = 1$.

Proof: Definition of Convexity: A function $g(x)$ is convex if $g(\lambda x + (1-\lambda)y) \leq \lambda g(x) + (1-\lambda)g(y)$ for all x and y and $0 < \lambda < 1$. The function $g(x)$ is concave if $-g(x)$ is convex. So now that we have this definition, let $l(x)$ be a tangent line to $g(x)$ at the point $g(E[X])$. Write $l(x) = a + bx$ for some a and b. Now by the convexity of g we have $g(x) \geq a + bx$. Since expectations preserve inequalities,

$$\begin{aligned} E[g(X)] &\geq E[a + bX] \\ &= a + bE[X] \\ &= l(E[X]) \\ &= g(E[X]) \end{aligned}$$

Can you show the only if part?

Chebychev's Inequality: For any random variable Y and any constants $a > 0$ and c,

$$E[(Y - c)^2] \geq a^2 P[|Y - c| \geq a]$$

4.2 Convergence in Probability: Proofs

4.2.1 Example 1: Simple Distribution

Let

$$Y_n = \begin{cases} 1 & \text{with probability } 1 - p_n \\ n & \text{with probability } p_n \end{cases}$$

First, show that

$$Y_n \xrightarrow{p} 1 \text{ if } p_n \to 0$$

To show convergence in probability, we use Chebychev's and find that

$$\mathbb{P}(|Y_n - 1| \geq \varepsilon) \leq \frac{\mathbb{E}[(Y_n - 1)^2]}{\varepsilon^2}$$

Calculating $\mathbb{E}[(Y_n - 1)^2]$ we get

$$\mathbb{E}[Y_n^2 - 2Y_n + 1] = \mathbb{E}[Y_n^2] - 2\mathbb{E}[Y_n] + 1 = 1 - p_n + n^2 p_n - 2 + 2p_n - 2np_n + 1 = n^2 p_n - 2np_n$$

So as $p_n \to 0$ we can see that this tends to 0. Therefore, with this as an upperbound this means that

$$\mathbb{P}(|Y_n - 1| \geq \varepsilon) \leq \frac{\mathbb{E}[(Y_n - 1)^2]}{\varepsilon^2} = \frac{n^2 p_n - 2np_n}{\varepsilon^2} \to 0 \text{ as } p_n \to 0$$

Second, show that

$$E[(Y_n - 1)^2] \to \infty \text{ if } p_n = 1/n$$

Using the previous part, we see that

$$\mathbb{E}[Y_n^2 - 2Y_n + 1] = \mathbb{E}[Y_n^2] - 2\mathbb{E}[Y_n] + 1 = 1 - p_n + n^2 p_n - 2 + 2p_n - 2np_n + 1 = n^2 p_n - 2np_n$$

So substituting in $p_n = 1/n$,

$$\mathbb{E}[Y_n^2 - 2Y_n + 1] = n - 2$$

And we see that

$$n - 2 \to \infty \text{ as } n \to \infty$$

4.2.2 Example 2: Exponential Distribution

Let $X_n \sim Exponential(n)$. Show that $X_n \to 0$ in probability.

Proof:

$$\begin{aligned}
\lim_{n\to\infty} P(|X_n - 0| \leq \varepsilon) &= \lim_{n\to\infty} P(X_n \leq \epsilon) \\
&= \lim_{n\to\infty} 1 - e^{-n\varepsilon} \\
&= 1 - \lim_{n\to\infty} \frac{1}{e^{n\varepsilon}} \\
&= 1
\end{aligned}$$

4.2.3 Example 3: Cauchy Distribution

We know that the Cauchy Distribution won't converge to the mean; yet, we saw that if we divide the mean each time by n, it does converge to 0. Why?

Proof: The trick here is that the "sample mean" of the Cauchy distribution is itself Cauchy, which is why it does not converge. So let's call our "sample mean" $\bar{X}_n$. This means that

$$P\left(\left|\frac{\bar{X}_n}{n} - 0\right| \leq \varepsilon\right) = P\left(\left|\bar{X}_n\right| \leq n\varepsilon\right)$$

which is equivalent to

$$P\left(\bar{X}_n \leq n\varepsilon\right) - P\left(\bar{X}_n \leq -n\varepsilon\right)$$

So using the fact that the CDF of the Cauchy distribution is

$$\frac{1}{\pi} arctan(x) + \frac{1}{2}$$

We can see that this just amounts to

$$\lim_{n\to\infty} \frac{1}{\pi} arctan(n\varepsilon) + \frac{1}{2} - (\frac{1}{\pi} arctan(-n\varepsilon) + \frac{1}{2}) = 1$$

5 Thursday, February 8

5.1 Convergence in Probability But Infinite Variance

Today we discuss an example of a sequence of random variable that converges in probability but whose variance goes to infinity. Consider the sequence of random variables $X_1, X_2, ...$, where the pdf of X_n is equal to

$$f_n(x) = \begin{cases} \frac{n-1}{2} & \text{if } -\frac{1}{n} < x < \frac{1}{n} \\ \frac{1}{n} & \text{if } n < x < n+1 \\ 0 & \text{if otherwise} \end{cases}$$

The mean of X_n is (can you show this is true?)

$$E[X_n] = 1 + \frac{1}{2n}$$

Notice that the mean of the random variable goes to 1 as $n \to \infty$. Now we show convergence in probability. Let $\varepsilon > n + 1$. Then this means that

$$\mathbb{P}(|X_n - 0| < \varepsilon) = \int_{-1/n}^{1/n} \frac{n-1}{2} dx + \int_{n}^{n+1} \frac{1}{n} dx = \frac{n-1}{n} + \frac{n+1}{n} - 1 \to 1$$

as $n \to \infty$. Now if $\varepsilon > \frac{1}{n}$ and $\varepsilon < n$, then we see that

$$\mathbb{P}(|X_n - 0| < \varepsilon) = \int_{-1/n}^{1/n} \frac{n-1}{2} dx = \frac{n-1}{n} \to 1$$

as $n \to \infty$. The last case we have to check is if $n < \varepsilon < n + 1$. Notice that in this case, the first integral still equals 1 as $n \to \infty$. Notice that the second integral is still bound by zero, because ϵ must always be greater than n. If $\varepsilon = n$, then notice that the second integral becomes $n/n - 1 = 0$. If $\varepsilon = n + 1$, then notice the second integral becomes $(n + 1)/n - 1 = 0$ as $n \to \infty$. Since it is bound by zero, that means it still integrates to one and convergence in probability still holds.

Now the variance is

$$Var(X_n) = \frac{n-1}{3n^3} + \frac{(n+1)^3 - n^3}{3n} - (1 + \frac{1}{2n})^2 \to \infty$$

as $n \to \infty$. Can you show the above variance holds?

Thus, notice that even though the variance goes to ∞ convergence in probability still holds.

5.2 Sampling from a $N(\mu, \sigma^2)$

$$L = \prod_{i=1}^{N} \frac{1}{(2\pi)^{1/2}\sigma} exp\left[-\frac{1}{2}\frac{(x_i - \mu)^2}{\sigma^2}\right]$$

$$logL = \sum_{i=1}^{N} log(1) - \frac{1}{2}log(2\pi) - log\sigma - \frac{1}{2}\left[\frac{(x_i - \mu)^2}{\sigma^2}\right]$$

$$= -\frac{n}{2}log(2\pi) - \frac{n}{2}log[\sigma^2] - \frac{1}{2}\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma^2}$$

Now differentiate with respect to μ

$$\frac{\partial lnL}{\partial \mu} = -\sum_{i=1}^{N} \frac{(x_i - \mu)(-1)}{\sigma^2} = 0$$

This means that

$$n\mu = \sum_{i=1}^{N} x_i \Rightarrow \mu^* = \frac{1}{n}\sum_{i=1}^{N} x_i$$

Now differentiate with respect to σ^2 and get

$$\frac{\partial lnL}{\partial \sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2}\sum_{i=1}^{N} \frac{(x_i - \mu)^2}{\sigma^4} = 0$$

$$\frac{n}{\sigma^2} = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{\sigma^4} \Rightarrow \sigma^2 = \frac{1}{n}\sum_{i=1}^{N} (x_i - \mu)^2$$

Notice that this is a biased estimate of the sample variance.

5.3 Fisher Information

Fisher Information is a way to measure the amount of information that an observed random variable X carries about an unknown parameter θ upon which the probability of X depends. Two definitions of Fisher Information, I, are used. Mathematically, they are

$$I(\theta) = E\left[\left(\frac{\partial}{\partial \theta} log f(X;\theta)\right)^2 |\theta\right]$$

the expected value of the slope of the log-likelihood function, and

$$I(\theta) = -E\left[\frac{\partial^2}{\partial \theta^2} log f(X;\theta)|\theta\right]$$

which is the curvature of the log-likelihood; it is equivalent for those cases where the second partial exists.

$$I(\theta) = E\left[\left(\frac{\partial}{\partial \theta} log f(X;\theta)\right)^2 |\theta\right] = \int \left(\frac{\partial}{\partial \theta} log f(X;\theta)\right)^2 f(X;\theta) dx$$

As Rocio showed in class, these two definitions are equivalent. So the Fisher Information is the negative of the expectation of the 2nd derivative with respect to θ of the natural logarithm of f. Information is seen to be the "curvature" of the support curve near the MLE of θ. A "blunt" support curve (one with a shallow max) would have a low negative expected derivative, and thus low information, while a sharp one would have a high negative expected second derivative and thus high information.

5.4 Cramer-Rao Lower Bound

The CRLB is the inverse of the Fisher Information, and is a lower bound on the variance of any unbiased estimator of θ. As a reminder, the likelihood function $f(X;\theta)$ describes the probability that we observe a given sample X when given a known value of θ. If f is sharply peaked with

respect to changes in θ, it is easy to indicate the correct value of θ from data, or equivalently, that data X provides a lot of information about parameter θ. If likelihood function f is flat and spread out, then it will take many samples like X to estimate the actual "true" value of θ that would be obtained using the entire population being sampled. Thus, we would intuit that the data contains much less information about the parameter. The curvature is given by its second derivative with respect to θ, hence our interest in how large $\frac{\partial^2 logf}{\partial \theta^2}$ could be.

Consider an unbiased estimator (with sample size $n = 1$). Then,

$$E[\hat{\theta}(x) - \theta|\theta] = \int [\hat{\theta}(x) - \theta] f(x;\theta) dx = 0$$

$$\frac{\partial}{\partial \theta} \int [\hat{\theta}(x) - \theta] f(x;\theta) dx = \int [\hat{\theta}(x) - \theta] \frac{\partial f}{\partial \theta} dx - \int f dx = 0$$

We're going to use 3 facts: (1) likelihood f is the probability of data given the parameter; (2) $\int f dx = 1$; (3) $\frac{\partial f}{\partial \theta} = f \frac{\partial logf}{\partial \theta}$. Using these three facts in the above, write

$$\int [\hat{\theta} - \theta] f \frac{\partial logf}{\partial \theta} dx = 1$$

Factor this expression such that

$$\int (\hat{\theta} - \theta) \sqrt{f} \sqrt{f} \frac{\partial logf}{\partial \theta} dx = 1$$

Then we can use the Cauchy-Schwarz inequality to write

$$[\int (\hat{\theta} - \theta)^2 f dx][\int (\frac{\partial logf}{\partial \theta})^2 f dx] \geq \left([\int [\hat{\theta} - \theta] \sqrt{f}][\sqrt{f} \frac{\partial logf}{\partial \theta}] \right)^2 = 1$$

Notice that the second bracket is the Fisher Information and the first bracket is the MSE of the estimator $\hat{\theta}$, also known as the variance of the estimator, since

$$E[(\hat{\theta}(x) - \theta)^2|\theta] = \int (\hat{\theta} - \theta)^2 f dx$$

So that means

$$E[(\hat{\theta} - \theta)^2|\theta] \geq \frac{1}{I(\theta)}$$

And using the fact that the MSE of the estimator $\hat{\theta}$ is also just the variance of $\hat{\theta}$, so:

$$Var(\hat{\theta}) \geq \frac{1}{I\theta}$$

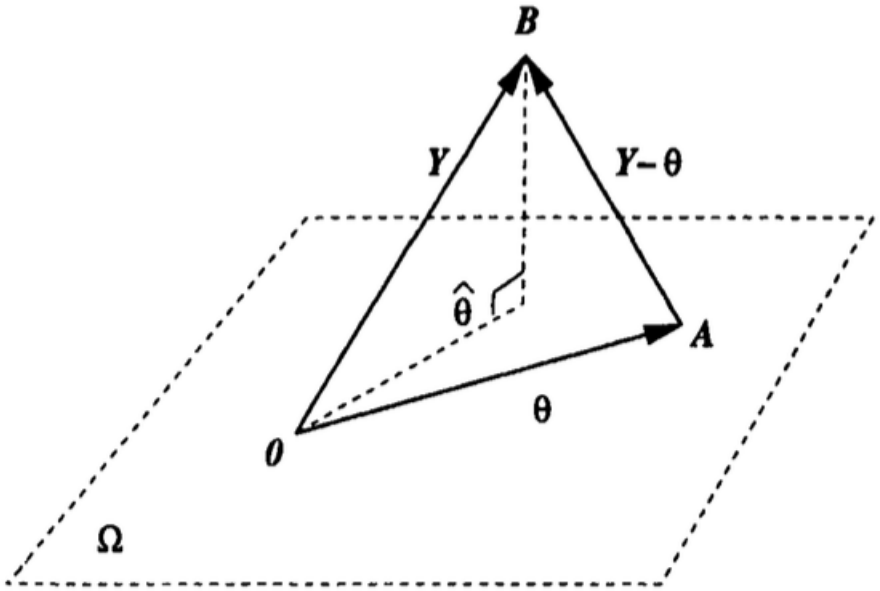
So the precision by which we can estimate θ is fundamentally limited by Fisher information of the likelihood function. To get into the form from class, just use multiple integrals and we can easily intuit that we get

$$nVar(\hat{\theta}) \geq \frac{1}{I(\theta)} \Rightarrow Var(\hat{\theta}) \geq \frac{1}{nI(\theta)}$$

6 Thursday, March 15

6.1 Linear Regression, Visualized

I tend to favor the linear algebra approach to OLS over the algebraic version. As a review, we want to minimize $\sum_i \varepsilon_i^2$ with respect to β. That is, if $\theta = X\beta$ then we minimize $\varepsilon^T\varepsilon = \|Y-\theta\|^2$, subject to the fact that $\theta \in C(X) = \Omega$, where Ω is the column space of X. If we let θ vary in Ω, then $\|Y-\theta\|^2$ will be a minimum for $\theta = \hat{Y}$, when $(Y-\hat{Y}) \perp \Omega$. We can show this both geometrically and we can show it algebraically as follows. Geometrically, it looks as the following:



First, note that $\hat{Y}$ can be obtained via a symmetric idempotent projection matrix P, namely, $\hat{Y} = PY$, where P represents the orthogonal projection onto Ω. Then,

$$Y - \theta = (Y - \hat{Y}) + (\hat{Y} - \theta)$$

Let's pause here and review projection matrices.

6.1.1 Projection Matrices

- Given Ω, a vector *subspace* of $\mathbb{R}^n$, every $n \times 1$ vector y can be expressed uniquely in the form $y = u + v$, where $u \in \Omega$, $v \in \Omega^{\perp}$
- $u = P_\Omega y$, then P_Ω is unique
- The matrix P_Ω can be expressed in the form $P_\Omega = TT^T$, where the columns of T form an orthogonal basis for Ω
  Proof: Let $T = (\alpha_1, \alpha_2, ..., \alpha_r)$, where r is the dimension of Ω. Expand the set α_i to give an orthonormal basis for $\mathbb{R}^n$, namely, $\alpha_1, ..., \alpha_r, \alpha_{r+1}, ..., \alpha_n$. Then,

$$y = \sum_{i=1}^{n} c_i\alpha_i = \sum_{i=1}^{r} c_i\alpha_i + \sum_{i=r+1}^{n} c_i\alpha_i = u + v$$

where $u \in \Omega$ and $v \in \Omega^{\perp}$. But $\alpha_i'\alpha_j = \delta_{ij}$ and $\alpha_i'\alpha_j = 1$, so that $\alpha_i'y = c_i$. Hence,

$$u = (\alpha_1, ..., \alpha_r) \begin{bmatrix} \alpha_1^T y \\ ... \\ \alpha_r^T y \end{bmatrix} = TT^T y$$

So this means that $P_\Omega = TT^T$

- P_Ω is symmetric and idempotent. TT' is obviously symmetric; $P_\Omega^2 = TT'TT' = TI_rT' = TT' = P_\Omega$
- $C(P_\Omega) = \Omega$
- $I_n - P_\Omega$ represnts an orthogonal projection onto $\Omega^{\perp}$

6.1.2 Now Back to the Main Show

Now from the above notions of projection, we know that $P\theta = \theta$, $P' = P$, and $P^2 = P$, and we have, preparing to square both sides,

$$(Y - \hat{Y})'(\hat{Y} - \theta) = (Y - PY)'(PY - P\theta) = Y'(I_n - P)P(Y - \theta) = 0$$

Hence

$$\|Y - \theta\|^2 = \|Y - \hat{Y}\|^2 + \|\hat{Y} - \theta\|^2 \geq \|Y - \hat{\theta}\|^2$$

which results in equality if and only if $\theta = \hat{\theta}$. Since $Y - \hat{\theta}$ is perpendicular to θ,

$$X'(Y - \hat{Y}) = 0$$

Or

$$X'\hat{Y} = X'Y$$

Here, $\hat{Y}$ is uniquely determined, being the unique orthogonal projection of Y onto Ω (see previous section on projections). Now, assume that the columns of X are linearly independent so that there exists a unique vector $\hat{\beta}$ such that $\hat{Y} = X\hat{\beta}$. So that means we have

$$X'X\hat{\beta} = X'Y$$

Now, since X has rank p, $X'X$ is positive definite and therefore nonsingular. So that means this has a unique solution, namely,

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

We denote the fitted values as $X\hat{\beta}$ by $\hat{Y} = (\hat{Y}_1, ..., \hat{Y}_n)'$. The elements of the vector

$$Y - \hat{Y} = Y - X\hat{\beta} = (I_n - P)Y$$

are called the residuals and are denoted by e. The minimum value of $\varepsilon'\varepsilon$ is, namely

$$\begin{aligned} e'e &= (Y - X\hat{\beta})'(Y - X\hat{\beta}) \\ &= Y'Y - 2\hat{\beta}'X'Y = \hat{\beta}'X'X\hat{\beta} \\ &= Y'Y - \hat{\beta}'X'Y + \hat{\beta}'[X'X\hat{\beta} - X'Y] \\ &= Y'Y - \hat{\beta}'X'Y \\ &= Y'Y - \hat{\beta}X'X\hat{\beta} \end{aligned}$$

Notice that $X'X\hat{\beta} - X'Y = 0$, which is why we get that above result.

6.1.3 Example 1

Let Y_1 and Y_2 be independent random variables with means α and 2α respectively. We will now find the least squares estimate of α and the residual sum of squares using what we did above. Then, we would write

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \alpha + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}$$

Notice that this is in the form $Y = X\beta + \epsilon$, where $X = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ and $\beta = \alpha$. Hence, from what we derived above

$$\hat{\alpha} = (X'X)^{-1}X'Y = \left[(1,2)\begin{bmatrix} 1 \\ 2 \end{bmatrix}\right]^{-1}(1,2)Y = \frac{1}{5}(1,2)(Y_1, Y_2)^T = \frac{1}{5}(Y_1 + 2Y_2)$$

and

$$e'e = Y'Y - \hat{\beta}'X'Y = Y'Y - \hat{\alpha}(Y_1 + 2Y_2) = Y_1^2 + Y_2^2 - \frac{1}{5}(Y_1 + 2Y_2)^2$$

6.1.4 A Theorem About Projections

Suppose that X is $n \times p$ of rank p, so that $P = X(X'X)^{-1}X'$. Then the following holds.

- P and $I_n - P$ are symmetric and idempotent
- $rank(I_n - P) = tr(I_n - P) = n - p$. Notice that this is why we talk about degrees of freedom as $n - p$, because this is the dimension of the space that we don't know anything about (the part that is NOT explained by data!)
- $PX = X$

7 Thursday, March 22

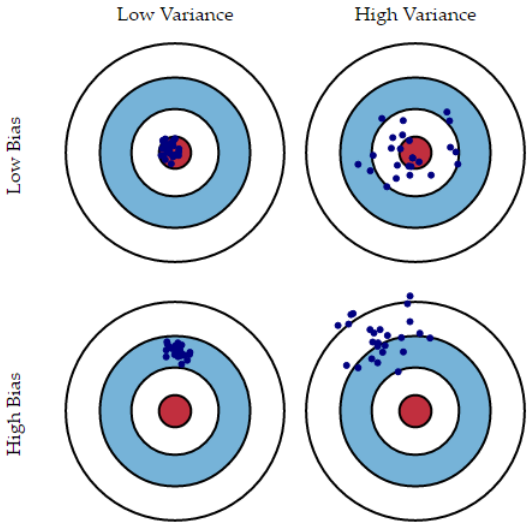
7.1 Problems with OLS

7.1.1 Bias vs. Variance

Bias: error from erroneous assumptions in the learning algorithm. Over or under estimate the value of a population parameter. High bias can cause an algorithm to miss the relevant relations between features and target outputs (underfitting).

Variance: error from sensitivity to small fluctuations in the training set. High variance can cause overfitting: modeling the random noise in the training data, rather than the intended outputs. How widely our estimates vary.

Here is a graphical depiction of the difference between bias and variance.



7.1.2 The Bias-Variance Decomposition of Squared Error

Suppose that we have a training set $x_1, x_2, ..., x_n$ and real values y_i associated with each observation. Of course, there is noise: $y = f(x) + \varepsilon$, where the noise ε has zero mean and variance σ^2. Now we want to find a function $\hat{f}(x)$ that approximates the true function $f(x)$ as well as possible. This may be because we're interested in using $\hat{f}(x)$ to make predictions on a dataset that has no values associated with y_i. Formally, the idea of "to be as close as possible" means making $(y - \hat{f}(x))^2$ as small as possible, both for points $x_1, ..., x_n$ *and for points outside of our dataset.* On any given sample, then,

$$E\left[(y - \hat{f}(x))^2\right] = Bias[\hat{f}(x)]^2 + Var[\hat{f}(x)] + \sigma^2$$

where

$$Bias[\hat{f}(x)] = E[\hat{f}(x) - f(x)]$$
$$Var[\hat{f}(x)] = E[\hat{f}(x)^2] - E[\hat{f}(x)]^2$$

We can think of these three terms as follows:

- The square of the bias of the learning method, which can be thought of as the error caused by the simplifying assumptions built into the method, such as assuming linearity
- Variance of the learning method intuitively tells us how much $\hat{f}(x)$ will move about its mean
- The irreducible error σ^2. Since all three terms are non-negative, it forms a lower bound on the expected error on unseen examples.

The derivation of the bias-variance decomposition is as follows. Here, $y = f + \varepsilon$, where f is deterministic.

$$\begin{aligned} E[(y - \hat{f})^2] = E[y^2 + \hat{f}^2 - 2y\hat{f}] \\ &= E[y^2] + E[\hat{f}^2] - E[2y\hat{f}] \\ &= Var[y] + E[y]^2 + Var[\hat{f}] + E[\hat{f}]^2 - 2fE[\hat{f}] \\ &= Var[y] + Var[\hat{f}] + (f^2 - 2fE[\hat{f}] + E[\hat{f}]^2) \\ &= Var[y] + Var[\hat{f}] + (f - E[\hat{f}])^2 \\ &= Var[y] + Var[\hat{f}] + E[f - E[\hat{f}]]^2 \\ &= \sigma^2 + Var[\hat{f}] + Bias[\hat{f}]^2 \end{aligned}$$

7.1.3 Two Shortcomings of OLS

Predictive Ability: the linear regression fit often has low bias but high variance. Recall that when we are estimating $\hat{\beta}$, we argue that it is an unbiased estimate of β. But this implies that our methods are driving bias to 0 while not controlling for variance at all.

Interpretive Ability: linear regression "freely" assigns a coefficient to each predictor variable. When the number of variables p is very large, which is the case in many modern datasets, we may sometimes seek a smaller set of important variables. Thus, we may want a procedure that only makes a subset of coefficients large, and others very small or even zero.

7.1.4 $p > n$

The above shortcomings become major problems in high-dimensional regression settings, where the number of predictors p rivals or even exceeds the number of of observations n. In fact, when $p > n$, the linear regression β estimates are actually not well-defined. This is a common problem in the world of big data nowadays, like with text analysis. In text analysis, p are the words and each observation is a 1 or 0 indicating whether the word is in a sentence or not (bag of words).

Why is this a problem? Simple. Recall that $\hat{\beta} = (X'X)^{-1}X'y$. But if X has $p > n$, this means that X cannot have independent columns which means $X'X$ is not a symmetric positive definite matrix (recall that n forms our space so if $p > n$ we can't form a subspace!). Thus, we have the problem if we have way more variables than actual observations.

Even if we somehow got a huge n that someone exceeded our p (which we are getting to in the world of really big data), there are still many problems. Fitting the full model without any sort of penalization will result in incredibly large prediction intervals; the LS regression estimators may not exist uniquely either, especially if $p > n$.

7.1.5 Ill-Conditioned X

Because the least-squares estimate depends on $(X'X)^{-1}$, we would have a problem computing β if $X'X$ was singular or *nearly* singular. In those cases, even smallest changes in X can lead to large changes in $(X'X)^{-1}$. The least square estimator β may provide a good fit for training data, but it will not fit sufficiently well for test data.

7.1.6 Ridge Regression

I won't talk about the derivation of the ridge regression, but I will talk about the intuition. So one way out of our quandary illustrated above is to actually drop the requirement for an unbiased estimator. That is, we first assume that our X and Y are centered, so we can forget about the constant term. Now, we can maybe get out of our situation above if we do the following:

$$\hat{\beta}_{ridge} = (X'X + \lambda I_p)^{-1} X'Y$$

Ridge regression places a particular form of constraint on the parameters β's: $\hat{\beta}_{ridge}$ is chosen to minimize the penalized sum of squares:

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

This is equivalent to minimizing

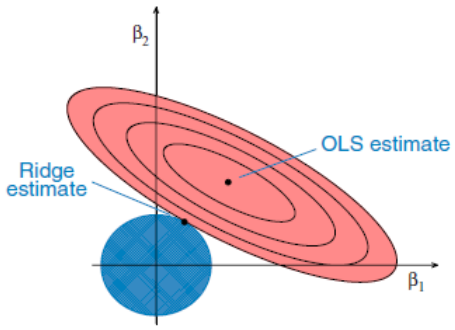
$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

subject to, for some $c > 0$,

$$\sum_{j=1}^{p} \beta_j^2 < c$$

i.e. constraining the sum of squared coefficients. You can solve this above using a Lagrangian. It is not trivial to show that the two formulations are equivalent; however, if you are really interested in these kinds of problems then STATS 600 is for you! But intuitively, what we're doing here is that we're imposing a pre-chosen λ that is penalizing us every time we estimate a really large β_j. We would prefer to take smaller β_j or β_j that are close to zero to drive the penalty term smaller.

We can examine ridge regression geometrically. So for $p = 2$, we see that the ellipses correspond to the contours of the residual sum of squares: the inner ellipse has a smaller RSS, and the RSS is minimized by the least squares (OLS) estimates. The constraint for the ridge regression, on the other hand, corresponds to a circle: $\sum_{j=1}^{p} \beta_j^2 < c$. We are trying to minimize the ellipse size and circle simultaneously in the ridge regression. The ridge estimate is given by the point at which the ellipse and the circle touch. Notice that this means β would never truly be zero, since you wouldn't ever touch at a point where $\beta_1 = 0$ or $\beta_2 = 0$. In LASSO regression, this can be the case—in LASSO, you would use a $|\beta|$ penalty term instead of a β^2 penalty term.


β2
OLS estimate
Ridge
estimate
β1

8 Thursday, March 29

8.1 Question 4 on the Midterm

8.2 Estimation with Linear Restrictions

We're going to use Lagrange multipliers to show how to estimate our new $\hat{\beta}$. Let $Y = X\beta + \varepsilon$, where X is $n \times k$ of full rank k. Suppose that we wish to find the minimum of $\varepsilon'\varepsilon$ subject to the linear restrictions $R\beta = q$, where R is a known $j \times p$ matrix of rank j and q is a known $j \times 1$ vector. One method of solving this problem is to use Lagrange multipliers, one for linear constraint $r_i'\beta = q_i$, $(i = 1, 2, ..., j)$, where r_i' is the ith row of R. So as a first step, we note that

$$\sum_{i=1}^{j} \lambda_i(r_i'\beta - q_i) = \lambda'(R\beta - q) = (\beta'R' - q')\lambda$$

Now we can set up the Lagrangian.

$$\mathcal{L} = \varepsilon'\varepsilon + (\beta'R \ - q') = (y - Xb)'(y - Xb) + (b - \beta)'X'X(b - \beta) + (\beta'R' - q')\lambda$$

Now taking $\partial\mathcal{L}/\partial\beta = 0$ we get

$$-2X'y + 2X'X\beta + R'\lambda = 0$$

Then for future reference we now denote the solutions to these two equations by $\hat{\beta}_H$ and $\hat{\lambda}_H$. Then we find:

$$\hat{\beta}_H = (X'X)^{-1}X'y - \frac{1}{2}(X'X)^{-1}R'\hat{\lambda}_H = \hat{\beta} - \frac{1}{2}(X'X)^{-1}R'\hat{\lambda}_H$$

So this means that

$$q = R\hat{\beta}_H = R\hat{\beta} - \frac{1}{2}R(X'X)^{-1}R'\hat{\lambda}_H$$

And since $(X'X)^{-1}$ is positive definite, being the inverse of a positive-definite matrix, $R(X'X)^{-1}R'$ is also positive-definite and therefore nonsingular. Hence,

$$-\frac{1}{2}\hat{\lambda}_H = [R(X'X)^{-1}R']^{-1}(q - R\hat{\beta})$$

Which means that substituting this into the above we have

$$\hat{\beta}_H = \hat{\beta} + (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(q - R\hat{\beta})$$

8.3 Bootstrapping

See `R` code on bootstrapping.

9 Thursday, April 5

9.1 Robust Regression

Long-tailed errors may prove to be more troublesome than short-tailed errors. In class, we talked about heteroskedasticity. But what if you have a few outliers? You could just remove them, but then it may not be random that certain observations have massive errors. One way of dealing with it is *M-estimation*.

M-estimates modify the least squares idea so we choose a β that minimizes:

$$\sum_{i=1}^{n} \rho(y_i - x_i^T \beta)$$

Some possible choices for ρ are

$$\rho(x) = x^2$$

$$\rho(x) = |x|$$

$$\rho(x) = \begin{cases} x^2/2 & \text{if } |x| \leq c \\ c|x| - c^2/2 & \text{otherwise} \end{cases}$$

The last one is called Huber's method and is a compromise between least-squares and least absolute deviation (LAD). c should be a robust estimate of σ. A value proportional to the median of $|\hat{\epsilon}|$ is suitable.

M-estimation is related to weighted least square. Normal equations tell us that

$$X^T(y - X\hat{\beta}) = 0$$

Putting this with weights and non-matrix form, this becomes

$$\sum_{i=1}^{n} w_i x_{ij} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right) = 0$$

Now, differentiating the M-estimate criterion with respect to β_j and setting it equal to zero, we get

$$\sum_{i=1}^{n} \rho' \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right) x_{ij} = 0$$

for $j = 1, ..., p$. Let $u_i = y_i - \sum_{j=1}^{p} x_{ij}\beta_j$, and we get

$$\sum_{i=1}^{n} \frac{\rho'(u_i}{u_i} x_{ij} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right) = 0$$

for $j = 1, ..., p$. That means we can make the identification of a weight function as

$$w(u) = \frac{\rho'(u)}{u}$$

Using Huber's method, we see that

$$w(u) = \begin{cases} 1 & \text{if } |u| \leq c \\ c/|u| & \text{otherwise} \end{cases}$$

9.2 Galapagos Island Example

Notice that the estimates haven't changed too much, which means that robust regression can often be used to confirm

10 Thursday, April 12

10.1 Probit

Suppose that the latent variable y_i^* follows

$$y_i^* = x_i\beta + \varepsilon_i$$

Here, ε_i is independent of x_i, which is a $1 \times K$ vector with the first element equal to unity for all i, θ is a $K \times 1$ vector of parameters, and $\varepsilon_i \sim Normal(0, 1)$. Instead of observing y_i^*, however, we only observe the binary variable indicating the *sign* of y_i^*.

$$y_i = \begin{cases} 1 & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* \leq 0 \end{cases}$$

Would be easier to write this as an indicator function:

$$y_i = \mathbb{I}[y_i^* > 0]$$

Because $\varepsilon_i \sim N(0, 1)$, it does not matter whether the strict inequality is in the first or second case. Thus, we can see that the distribution of y_i given x_i is

$$\begin{aligned} P(y_i = 1|x_i) &= P(y_i^* > 0|x_i) \\ &= P(x_i\beta + \varepsilon_i > 0|x_i) \\ &= P(\varepsilon_i > -x_i\beta|x_i) \\ &= 1 - \Phi(-x_i\beta) \\ &= \Phi(x_i\beta) \end{aligned}$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function. Thus, this means that

$$P(y_i = 0|x_i) = 1 - \Phi(x_i\beta)$$

This means that

$$f(y|x_i) = [\Phi(x_i\beta)]^y[1 - \Phi(x_i\beta)]^{1-y}$$

Now, notice that interpretation is not very straightforward. Think about OLS for a second. If we think about OLS:

$$E[Y_i|X_i; \beta] = \beta_0 + \sum_{k=1}^{K} \beta_k X_{ki}$$

Then, we can see that

$$\frac{\partial E[Y_i|X_i; \beta]}{\partial X_{ki}} = \beta_k$$

which gives us our usual interpretation of OLS coefficients. But notice that if we carry out this same first derivative on probit, we get

$$\frac{\partial P(Y_i = 1|X_i; \beta)}{\partial X_{ki})} = \beta_k \phi \left(\sum_{k=1}^{K} \beta_k X_{ki} \right)$$

Notice that there's no great interpretation of this! What a lot of people do is fix the X's at their mean, and then give an estimate of the change in probability that way. But it's kind of a sketchy way of interpreting the coefficients...

10.2 Linear Probability Model

If interpretation is of utmost importance, maybe consider something like the linear probability model, which is quite straightforward:

$$P(y = 1|x) = \beta_0 + \beta_1 x_1 + ... + \beta_K x_k$$

What we can do is to first run OLS over the observed 0-1 y's and the explanatory variables. Then, provided that $0 < \hat{y}_i < 1$ for all observations i, define the standard deviation as $\hat{\sigma}_i = \sqrt{\hat{y}_i(1 - \hat{y}_i)}$. Then, the WLS estimator, β^*, is obtained from the OLS regression

$$y_i/\hat{\sigma}_i \text{ on } 1/\hat{\sigma}_i, x_{i1}/\hat{\sigma}_i, ..., \hat{x}_{iK}/\hat{\sigma}_i$$

for $i = 1, ..., n$. This will then give you standard errors, etc. based on the theory of weighted least squares. But no guarantees that the coefficients make sense given the constraints operated here—this approach requires a bit of improvisation.

10.3 General Approach Using Latent Variable Interpretation

The latent variable interpretation makes the most sense to me, personally. This actually also works for logit as well. That is, assuming

$$y^* = x\beta + \varepsilon$$

$$y = 1[y^* > 0]$$

Assume that ε is a continuously distributed variable independent of x and the distribution of ε is symmetric around 0. If G is the cdf of ε, then, because the pdf of ε is symmetric about zero, that means that $1 - G(-z) = G(z)$ for all real numbers z. Then,

$$P(y = 1|x) = P(y^* > 0|x) = P(\varepsilon > -x\beta|x) = 1 - G(-x\beta) = G(x\beta)$$

The logit model, thus, is simply a special case of the equation above with

$$G(z) = \Lambda(z) = \frac{exp(z)}{1 + exp(z)}$$

Notice that, under this condition,

$$\frac{\partial P(y = 1|x)}{\partial x_j} = g(x\beta)\beta_j$$

which is what we found with probit above too. So with the logit model, it seems like we're stuck in the same quagmire, right? No! What we can do is as follows. We know that, if $p(x)$ is the probability of a success, then, the odds are defined as

$$\frac{p(x)}{1 - p(x)}$$

We also know that we can define log-odds, which is just

$$ln\left(\frac{p(x)}{1 - p(x)}\right)$$

Given that

$$p(x) = \frac{exp(x\beta)}{1 + exp(x\beta)} = \frac{1}{1 + exp(-x\beta)}$$

we see that

$$1 - p(x) = \frac{e^{-x\beta}}{1 + e^{-x\beta}}$$

which means that

$$\frac{p(x)}{1 - p(x)} = \frac{1}{e^{-x\beta}} = e^{x\beta}$$

which means that

$$ln\left(\frac{p(x)}{1 - p(x)}\right) = x\beta = \beta_0 + \beta_1 x_1 + ... + \beta_K x_K$$

Boom! So that means we can interpret the coefficients as changes in the log-odds of success!! This is why logistic regression is so much more popular than probit. So why would you ever use probit? Well, the probit is based on the normal distribution, which means there is a multivariate normal distribution...which allows you to jointly model many response variables, adjusting the covariance matrix! There is no multivariate logit distribution...which means you cannot model jointly many response variables.