MAS 420: Statistics, Linear Algebra, and Regression Review

Adam Rohde
Department of Statistics
University of California, Los Angeles

October 3, 2022

Recommended resources

The previous set of slides focused on probability, random variables, and relationships between random variables. In these slides, we will deal with data, estimation, and statistics. We will also review linear algebra and regression.

Useful resources

1. Foundations of Agnostic Statistics - Aronow and Miller (2019) [link] - part 2
2. Elements of Statistical Learning - Hastie et al (2009) - chapter 3 [link]
3. Matrix Algebra - Gentle (2007) [link]
4. Causal Inference: the Mixtape - Cunningham (2021) [link] - chapter 2
5. Statistical Rethinking (Videos) - McElreath (2022) [link]

These slides draw heavily from 1.

Table of Contents

Motivating statistics

Statistics aims to learn features of populations of units or people despite not having access to the full population.

We aim to estimate things like expected values, variance, correlation, and best predictors for the population using the sample of data we do have access to.

Since we do not have access to the underlying random variables directly we must infer from the data we do have what these quantities might be. This involves uncertainty and the quantification of uncertainty.

Have in your mind that our data is drawn from some joint probability distribution over random variables that represents the real-world process we are studying. In this way we build on the probability theory we saw last time.

IID Random Variables

We are studying some set of units. Each unit might have multiple attributes. We consider each attribute for each unit to be a random variable.

Example

For UCLA students, each person's height is a random variable as is each person's weight. There are different probabilities for height and weight for each person. So we have two random variables for each person. Height and weight *for each person* are likely dependent. What about two people's height? How do these relate?

We often make some simplifying assumptions: that different people's heights are independent and that they are identically distributed (IID).

Would this make sense for an attribute for whether or not each person has COVID? Probably not - one person having COVID likely depends on how many people around them have COVID. So the IID assumption has limitations.

IID Random Variables

The discussion today will assume that we have IID observations.

Definition (Independent and identically distributed (IID))

Let $X_1, \ldots, X_n$ be random variables with CDFs $F_1, \ldots, F_n$. Let F_A denote the joint CDF of the RVs with indices in the set A. Then $X_1, \ldots, X_n$ are IID if

- Mutually independent: $F_A((x_i)_{i \in A}) = \prod_{i \in A} F_i(x_i)$, $\forall A \subset \{1, \ldots, n\}, \forall (x_1, \ldots, x_n) \in \mathbb{R}^n$.
- Identically distributed: $F_i(x) = F_j(x)$, $\forall i, j \subset \{1, \ldots, n\}, \forall x \in \mathbb{R}$.

We take a draw from a random variable X and then take another draw and another etc. in such a way that the different draws are independent, but they all come from an identical random process.

The sample mean

Assuming IID random variables, as our sample of observations of draws from the random variable X grows larger, we are better able to estimate features of the distribution of X.

Definition (Sample mean)

For IID RVs $X_1, \ldots, X_n$, the sample mean is $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$.

Note that since $\bar{X}$ is a function of RVs, it is itself an RV.

Expected value of sample mean is population mean

For IID RVs $X_1, \ldots, X_n$, $\mathbb{E}[\bar{X}] = \mathbb{E}[X]$.

Sampling variance of sample mean

For IID RVs $X_1, \ldots, X_n$, the sampling variance (how much variation we can expect in the sample mean across different hypothetical draws of n observations) is $V[\bar{X}] = \frac{V[X]}{n}$.

$V[\bar{X}]$ decreases as n increases.

Law of large numbers

Definition (Convergence in probability)

Let $(T_{(1)}, T_{(2)}, \dots)$ be a sequence of RVs and let $c \in \mathbb{R}$. $T_{(n)}$ converges in probability to c if, $\forall \epsilon > 0$, $\lim_{n\to\infty} Pr\left[|T_{(n)} - c| \leq \epsilon\right] = 1$. Write $\bar{T}_{(n)} \xrightarrow{p} c$.
(As n goes to infinity, it becomes extremely likely that $T_{(n)}$ will be extremely close to c.)

Weak law of large numbers

For IID RVs $X_1, \dots, X_n$, let $\bar{X}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then $\bar{X}_{(n)} \xrightarrow{p} \mathbb{E}[X]$.

As n gets large, $\bar{X}$ becomes more likely to approximate $\mathbb{E}[X]$ to arbitrary precision.
[R demo]

Estimation

Let θ be some feature (e.g., mean or variance) of the RV X. θ is called an **estimand**.
We observe n IID draws of X: $\mathbf{X} = (X_1, \ldots, X_n)^\top$.
An **estimator** of θ is an RV $\hat{\theta} = h(\mathbf{X}) = h(X_1, \ldots, X_n)$. (e.g., sample mean)
Estimators take values called **estimates**.

Example

Consider a coin flip represented by a Bernoulli RV. We want to estimate the probability of heads (our estimand). We might consider the following estimators. Not all of these provide equally good estimates.

- $[\max(X_1, \ldots, X_n) + \min(X_1, \ldots, X_n)]/2$
- $\sum X_i$
- $1/n \sum X_i$

Bias

Definition (Unbiasedness)

An estimator $\hat{\theta}$ is unbiased for θ if $\mathbb{E}[\hat{\theta}] = \theta$.

Definition (Bias)

The bias of $\hat{\theta}$ for estimating θ is $\mathbb{E}[\hat{\theta}] - \theta$.

Bias tells us the difference between the average values of the estimator and the true value. Unbiased estimators have zero bias. Unbiasedness is a useful property but it is not always the most important element of an estimator.

Features of distributions of estimators

Definition (Sampling variance of an estimator)

The sampling variance of $\hat{\theta}$ is $V[\hat{\theta}]$. This is the variance in estimates of $\hat{\theta}$ from repeated samples. Recall that the sampling variance of the sample mean is $V[\bar{X}] = \frac{V[X]}{n}$.

Definition (Standard error of an estimator)

The standard error of $\hat{\theta}$ is $\sqrt{V[\hat{\theta}]}$.

Definition (Mean squared error (MSE) of an estimator)

The MSE of $\hat{\theta}$ in estimating θ is $\mathbb{E}[(\hat{\theta} - \theta)^2] = V[\hat{\theta}] = (\mathbb{E}[\hat{\theta}] - \theta)^2 = \text{Variance} + (\text{Bias})^2$.

Definition (Consistency)

An estimator $\hat{\theta}$ is consistent for θ if $\hat{\theta} \xrightarrow{P} \theta$.
(As n gets large, $\hat{\theta}$ becomes more likely to approximate θ to arbitrary precision.)

Sample variance

Definition (Sample variance)

We might try to estimate the variance of X using $\bar{X^2} - \bar{X}^2$ (just plug in sample means into $V[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$). It turns out that $\mathbb{E}[\bar{X^2} - \bar{X}^2] = \frac{n-1}{n} V[X]$ and is biased. Instead we can use $\hat{V}[X] = \frac{n}{n-1}\left(\bar{X^2} - \bar{X}^2\right)$, which is unbiased.

Properties of sample variance

The sample variance is unbiased and consistent for the variance of X.

- $\mathbb{E}\left[\hat{V}[X]\right] = V[X]$
- $\hat{V}[X] \xrightarrow{p} V[X]$

Sample variance ($\hat{V}[X]$, an estimator of variance of X) is different from the **sampling variance of an estimator** ($V[\hat{\theta}]$, variance of the distribution of $\hat{\theta}$). We can also have estimators for the sampling variance of an estimator: $\hat{V}[\hat{\theta}]$. [Really try to understand this.]

Convergence in distribution and standardized sample mean

Definition (Convergence in distribution)

Let $(T_{(1)}, T_{(2)}, \dots)$ be a sequence of RVs with CDFs $(F_{(1)}, F_{(2)}, \dots)$. Let T be a RV with CDF F. Then $T_{(n)}$ converges in distribution to T if, $\forall t \in \mathbb{R}$ at which F is continuous, $\lim_{n\to\infty} F_{(n)}(t) = F(t)$. Write $T_{(n)} \xrightarrow{d} T$.
(As n goes to infinity, the distribution of $T_{(n)}$ converges to the distribution of T.)

Definition (Standardized sample mean)

For IID RVs $X_1, \dots, X_n$ with $\mathbb{E}[X] = \mu$ and $V[X] = \sigma^2 > 0$, the standardized sample mean is $Z = \frac{\bar{X} - \mathbb{E}[\bar{X}]}{\sqrt{V[\bar{X}]}} = \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}$. It's possible to show that $\mathbb{E}[Z] = 0, v[Z] = 1$.

Central limit theorem

Central limit theorem (CLT)

For IID RVs $X_1, \ldots, X_n$ with $\mathbb{E}[X] = \mu$ and $V[X] = \sigma^2 > 0$ and Z the standardized sample mean, then $Z \xrightarrow{d} \mathcal{N}(0, 1)$ or, equivalently, $\sqrt{n}(\bar{X} - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$.

For large enough n, the sampling distribution of the sample mean is approximately a normal distribution.
[R demo]

Confidence intervals

We've looked at how we can estimate features, θ, of the distribution of random variable X. But how certain are we that these estimates reflect the true value of θ?
A confidence interval for θ is an estimated interval that covers the true value of θ with at least some given probability.

Definition (Valid confidence interval)

A valid confidence interval for θ with coverage $(1-\alpha)$ is a random interval $\mathrm{CI}_{(1-\alpha)}(\theta)$ such that $Pr(\theta \in \mathrm{CI}_{(1-\alpha)}(\theta)) \geq 1-\alpha$.

A 95% confidence interval for $\mathbb{E}[X]$ should contain $\mathbb{E}[X]$ with probability 0.95. [R demo]

Normal approx. based CI

Let $\hat{\theta}$ be an asymptotically normal estimator of θ, $\hat{V}[\hat{\theta}]$ be a consistent estimator of the sampling variance and $\alpha \in (0,1)$. An asymptotically valid normal approx. based CI for θ is $\mathrm{CI}_{(1-\alpha)}(\theta) = \left(\hat{\theta} - z_{1-\frac{\alpha}{2}}\sqrt{\hat{V}[\hat{\theta}]}, \hat{\theta} + z_{1-\frac{\alpha}{2}}\sqrt{\hat{V}[\hat{\theta}]}\right)$, z_c is the cth quantile of $\mathcal{N}(0,1)$.

Hypothesis testing

Hypothesis testing is closely related to confidence intervals.
Suppose we want to test the null hypothesis that $\theta = \theta_0$.
We have an estimate $\hat{\theta}$ and want to ask: if θ were really equal to θ_0, what is the probability that we would have obtained an estimate $\hat{\theta}$ at least as far from θ_0 as we did?

Definition (p-value)

Let $\hat{\theta}$ be an estimator of θ and let $\hat{\theta}^*$ be the observed value of $\hat{\theta}$. A (two tailed) p-value under the null hypothesis $\theta = \theta_0$ is $p = Pr_{\theta_0}\left[|\hat{\theta} - \theta_0| \geq |\hat{\theta}^* - \theta_0|\right]$, where Pr_{θ_0} denotes the probability under the null hypothesis $\theta = \theta_0$.

A low p-value means, under the null, we would infrequently encounter an estimate as large as we observed. We therefore might reject the null hypothesis on this basis.

Hypothesis testing

Definition (t-statistic)

Let $\hat{\theta}$ be an asymptotically normal estimator of θ, $\hat{V}[\hat{\theta}]$ be a consistent estimator of the sampling variance, and $\hat{\theta}^*$ be the observed value of $\hat{\theta}$. Then the t-statistic is $t = \frac{\hat{\theta}^* - \theta_0}{\sqrt{\hat{V}[\hat{\theta}]}}$.

Normal approx. based p-values

An asymptotically valid (two-tailed) p-value under the null hypothesis that $\theta = \theta_0$ is $p = 2\left(1 - \Phi\left(\frac{|\hat{\theta}^* - \theta_0|}{\sqrt{\hat{V}[\hat{\theta}]}}\right)\right) = 2\left(1 - \Phi\left(t\right)\right)$.

(Asymptotically valid CIs and p-values have no guarantees in finite samples.)
We often reject the null hypothesis when the p-value is smaller than some threshold (e.g., 0.05). This leads to statements like "our result is statistically significant at the 0.05 level." [R demo]

Table of Contents

Vectors

Linear algebra is a rich branch of mathematics that deals with vectors and matrices. We provide a very limited review. See Matrix Algebra - Gentle (2007) [link] for more.

An n-dimensional vector is an array with n entries: $x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ with $x_1, \ldots, x_n \in \mathbb{R}$.

We will consider all vectors to be column vectors, not row vectors.
So the transpose of x is $x^\top = \begin{pmatrix} x_1 & \cdots & x_n \end{pmatrix}$.
Scalar multiplication, vector addition, and vector subtraction are all elementwise.

Definition (Inner product)

An inner product between two n-dimensional vectors x, y is $\langle x, y \rangle = \sum_{i=1}^{n} x_i y_i$.

We also write this as $x^\top y = \begin{pmatrix} x_1 & \cdots & x_n \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = x_1 y_1 + \cdots + x_n y_n$.

Vectors and sample statistics

The L_2 (Euclidean) norm or length of a vector x is $\|x\|_2 = \sqrt{\sum_i x_i^2} = \sqrt{x^\top x}$.

We call the vector $\mathbf{1}^\top = \begin{pmatrix} 1 & \cdots & 1 \end{pmatrix}$ the one vector.

We can use $\mathbf{1}$ in an inner product to sum the elements of a vector x: $\mathbf{1}^\top x = \sum x_i$.

We can then calculate the mean value of x as $\bar{x} = \frac{1}{n}\mathbf{1}^\top x = \frac{1}{n}\sum x_i$.

We can create a vector containing n copies of the mean as $\bar{x}\mathbf{1}$, since $\bar{x}$ is a scalar.

For a vector x, we can create its centered counterpart as $x_c = x - \bar{x}\mathbf{1}$.

A centered vector has mean zero: $\mathbf{1}^\top x_c = 0$.

We can also calculate variance using an inner product:

$s_x^2 = \frac{x_c^\top x_c}{n-1} = \frac{\|x_c\|_2^2}{n-1} = \frac{1}{n-1}\left(\sqrt{\sum_i (x_i - \bar{x})^2}\right)^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2$.

And so standard deviation is $s_x = \sqrt{\frac{x_c^\top x_c}{n-1}} = \frac{\|x_c\|_2}{\sqrt{n-1}}$.

Covariance between x and y is $\mathrm{Cov}(x, y) = \frac{x_c^\top y_c}{n-1}$, which is similar to variance.

Correlation between x and y is $\mathrm{Cor}(x, y) = \frac{x_c^\top y_c}{\sqrt{x_c^\top x_c}\sqrt{y_c^\top y_c}} = \frac{x_c^\top y_c}{\|x_c\|_2 \|y_c\|_2}$, which takes the familiar form of covariance over SDs.

Matrices

An $n \times m$ dimensional matrix, A, is an array with n rows and m columns where each element $a_{i,j} \in \mathbb{R}$. We write $A \in \mathbb{R}^{n \times m}$. Vectors are special cases of matrices in $\mathbb{R}^{n \times 1}$.

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,m} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,m} \end{pmatrix}$$

The transpose of $A = (a_{ij})$ is $A^\top = (a_{ji})$. If $A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$ then $A^\top = \begin{pmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{pmatrix}$.

For $c \in \mathbb{R}$, $cA = (c \times a_{ij})$.
$(A^\top)^\top = A$
$(cA)^\top = cA^\top$
$(A + B)^\top = A^\top + B^\top$

Matrix multiplication

We can multiply matrices. For matrices $A_{n\times m}, B_{m\times p}$, their product $AB = C_{n\times p}$. Matrix products are not commutative. So typically $AB \neq AB$. Moreover, two matrices can only be multiplied if their dimensions are compatible - they must have only dimension that is the same size. For our $A_{n\times m}, B_{m\times p}$, BA is actually not computable.

$A_{n\times m}B_{m\times p} = C_{n\times p}$ where $c_{ij} = \sum_k a_{ik}b_{kj} = a_i^\top b_j$
where a_i is the vector for the ith row of A and b_j is the vector for the jth column of B

$$C = \quad A \; B$$
$$\begin{pmatrix} \quad \end{pmatrix}_{n\times p} = \begin{pmatrix} \quad \end{pmatrix}_{n\times m} \begin{pmatrix} \quad \end{pmatrix}_{m\times p}$$

$(AB)^\top = B^\top A^\top$; $A(BC) = (AB)C$; $A(B + C) = AB + AC$; $(B + C)A = BA + CA$

Systems of linear equations

We often want to solve systems of linear equations like $Ax = b$ for x, where $A_{n \times m}$ is a matrix, x is a $m \times 1$ vector and b is a $n \times 1$ vector.

Does it make sense to do the following?

$$Ax = b \iff A^{-1}Ax = A^{-1}b \iff x = A^{-1}b$$

(Note that we're only looking at the linear algebra that is going to be useful for understanding OLS. Again, see the references for more depth.)

Matrix inverses

We can take the inverse of $n \times n$ matrices. $AA^{-1} = A^{-1}A = \mathbb{I}_n$. $\mathbb{I}_n$ is the identity matrix that has all zeroes except for ones on the diagonal. $\mathbb{I}_n A = A$

The matrix inverse is the solution to the equations $AX = \mathbb{I}_n$ and $XA = \mathbb{I}_n$.

Matrix inverses only exist for matrices that are square (think about why; has to do with conformability) and also have some other attributes. One of these is that the matrix cannot be rank deficient, which essentially means that no column is a linear combination of the other columns in the matrix ($a_i \neq \gamma_1 a_1 + \cdots + \gamma_n a_n$). This is why OLS does not work under "perfect multicollinearity" - OLS involves taking a matrix inverse.

$(A^{-1})^{-1} = A$
$(cA)^{-1} = c^{-1}A^{-1}$
$(A^\top)^{-1} = (A^{-1})^\top$

Vector calculus

Very informally, we can think about taking derivatives with vectors in a way that looks similar to taking regular derivatives. For vector x, a and matrix A,[1]

Scalar	Vector
$f(x) \to \frac{df}{dx}$	$f(x) \to \frac{df}{dx}$
$ax \to a$	$x^\top A \to A$
$ax \to a$	$x^\top a \to a$
$x^2 \to 2x$	$x^\top x \to 2x$
$ax^2 \to 2ax$	$x^\top Ax \to 2Ax$
$(a - bx)^2 \to -2b(a - bx)$	$(a - Ax)^\top (a - Ax) \to -2A^\top (a - Ax)$

[1] This slide pulls from Kristy McNaught's "Matrix derivatives cheat sheet" [link] and Petersen and Pedersen's "The Matrix Cookbook" [link]

Table of Contents

What is regression for?

Suppose we have data on two variables X and Y that are drawn from a joint distribution.

There are different things we might want to do with this data:

1. use X to *predict* values for Y for units that we do not have data on
2. understand how Y changes as X changes *statistically*
3. understand the *causal* effect that X has on Y

It is important to recognize that these are different goals and require different things.

Recall that "Correlation (i.e., statistical association) does not imply causation."

We'll discuss 1 and 2 briefly here. See Elements of Statistical Learning - Hastie et al (2009) [link] for a thorough introduction. The rest of the course will focus on 3. See Causal Inference: the Mixtape - Cunningham (2021) [link] for a useful reference.

CEF and BLP Review

For every value of x, $\mathbb{E}[Y|X = x]$ maps x to the conditional mean of Y (a single value). So we can consider this a type of function that takes in values of x and outputs the conditional expectation of Y when $X = x$. We call this the conditional expectation function (CEF). We usually write $\mathbb{E}[Y|X]$ to denote the CEF. The CEF is a feature of the joint distribution of X, Y.

The CEF, $\mathbb{E}[Y|X]$, is the best (minimum MSE) predictor of Y given X.

Restricting to *linear* functions of X, the best (minimum MSE) *linear* predictor of Y given X is $g(X) = \alpha + \beta X$, where

$$\alpha = \mathbb{E}[Y] - \beta\mathbb{E}[X] \text{ and } \beta = \frac{\text{Cov}[X, Y]}{V[X]}$$

We call this the best linear predictor (BLP).

Regression, BLP, and CEF

To predict Y using X or to understand how Y changes with X statistically, we might hope know the CEF. But the CEF is typically unknown; so we must estimate it using data.

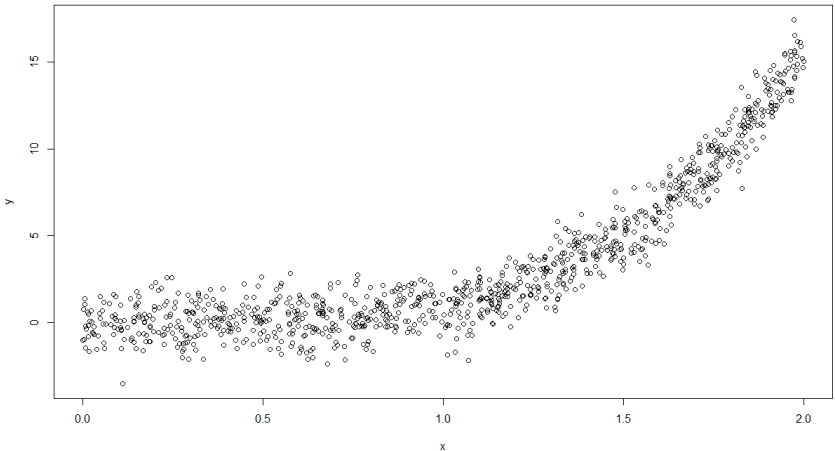
How might we do this?
One option would be to try to *directly estimate the CEF* using some flexible ML method. This can be a good option for prediction. But it doesn't always provide a simple summary of how Y changes with X (though some approaches do; e.g., KRLS [link]).

Another approach is to try to estimate the BLP as a *linear approximation of the CEF* and then inspect the coefficients in the estimated equation $\hat{g}(X) = \hat{\alpha} + \hat{\beta}X$. This estimation can be done using ordinary least squares (OLS) regression.

OLS regression estimates of the BLP can provide simple, transparent, and useful summaries of and approximations to the CEF.

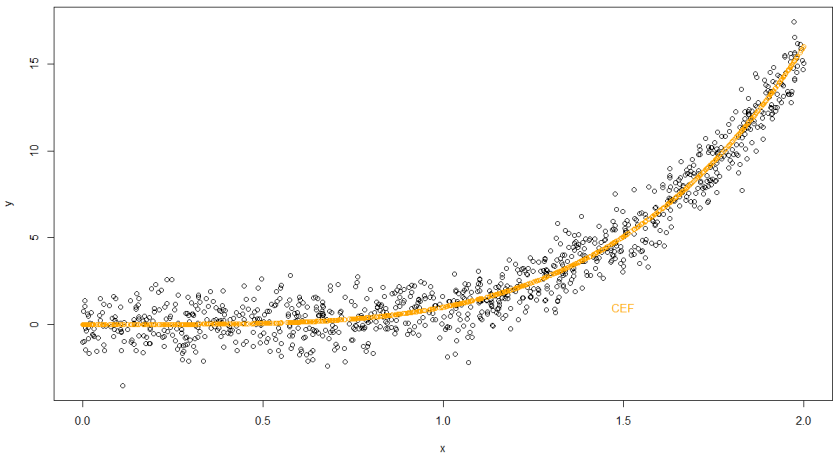
Example

Data generating process: $x = \text{runif}(1000, 0, 2)$; $y = x^4 + \text{rnorm}(1000, 0, 1)$



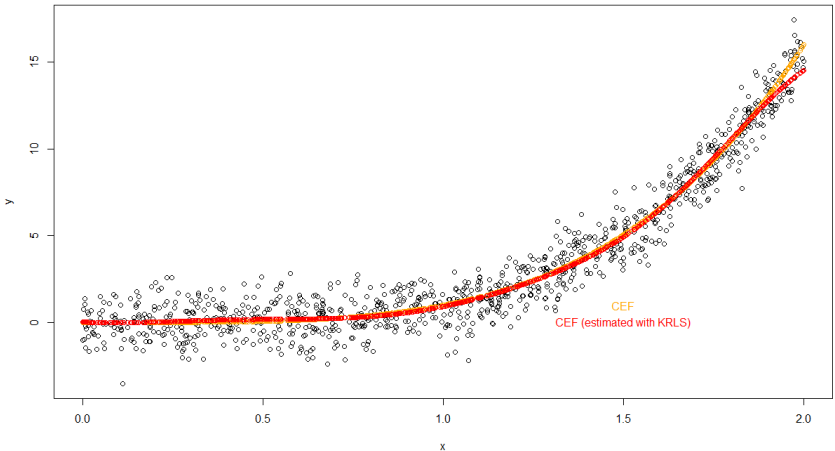
Example

Data generating process: $x = \text{runif}(1000, 0, 2)$; $y = x^4 + \text{rnorm}(1000, 0, 1)$



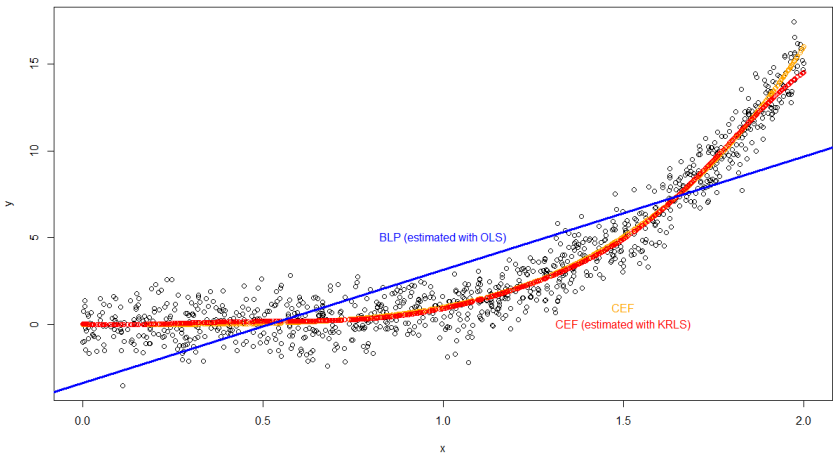
Example

Data generating process: $x = \text{runif}(1000, 0, 2)$; $y = x^4 + \text{rnorm}(1000, 0, 1)$



Example

Data generating process: $x = \text{runif}(1000, 0, 2)$; $y = x^4 + \text{rnorm}(1000, 0, 1)$



Estimating the BLP - bivariate case

The BLP of Y given X is $g(X) = \alpha + \beta X$, where

$\beta = \frac{\text{Cov}[X,Y]}{V[X]} = \frac{\mathbb{E}[XY]-\mathbb{E}[X]\mathbb{E}[Y]}{\mathbb{E}[X^2]-\mathbb{E}[X]^2}$
$\alpha = \mathbb{E}[Y] - \beta\mathbb{E}[X]$

We can estimate α, β from a sample of data as $\hat{\beta} = \frac{\overline{XY}-\overline{X}\times\overline{Y}}{\overline{X^2}-\overline{X}^2}$ and $\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X}$.

We are not assuming anything about the distribution of X, Y. This is always a valid procedure for estimating the BLP (in that these estimates are consistent for the BLP). Though the BLP can be a better or worse approximation to the CEF depending on the distribution of X, Y and the linearity of their relationship.

Estimating the BLP - multivariate case

Let $(Y_1, \mathbf{X}_1), (Y_2, \mathbf{X}_2), \ldots, (Y_n, \mathbf{X}_n)$ be n IID random vectors, where $\mathbf{X}_i = (X_{i,1}, X_{i,2}, \ldots, X_{i,p})$ is a random vector of p "explanatory" variables for each unit i.

The BLP of $\mathbf{Y}$ given $\mathbb{X}$ is $g(\mathbb{X}) = \mathbb{X}\beta$,

where $\beta^\top = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p) \in \mathbb{R}^{p+1}$, and $\mathbb{X}_{n\times(p+1)}$ is a matrix with n rows for the n observations or units and p columns for the p explanatory variables as well as a column of 1's for the intercept.

$$\mathbb{X}_{n\times(p+1)} = \begin{pmatrix} 1 & X_{1,1} & X_{1,2} & \cdots & X_{1,p} \\ 1 & X_{2,1} & X_{2,2} & \cdots & X_{2,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & X_{n,2} & \cdots & X_{n,p} \end{pmatrix}$$

β can be estimated using OLS regression.

Ordinary least squares

OLS regression minimizes the sum of squared differences $e_i = Y_i - \mathbf{X}_i^\top \mathbf{b}$ across all units i using the same set of coefficients $\mathbf{b}$ to estimate β:

$$\begin{aligned}\hat{\beta} &= \arg\min_{\mathbf{b}\in\mathbb{R}^{p+1}} \sum_i^n e_i^2 \\ &= \arg\min_{\mathbf{b}\in\mathbb{R}^{p+1}} \|\mathbf{e}\|_2^2 \\ &= \arg\min_{\mathbf{b}\in\mathbb{R}^{p+1}} \|\mathbf{Y} - \mathbb{X}\mathbf{b}\|_2^2 \\ &= \arg\min_{\mathbf{b}\in\mathbb{R}^{p+1}} (\mathbf{Y} - \mathbb{X}\mathbf{b})^\top (\mathbf{Y} - \mathbb{X}\mathbf{b})\end{aligned}$$

This can be solved by setting the derivative wrt $\mathbf{b}$ of $(\mathbf{Y} - \mathbb{X}\mathbf{b})^\top (\mathbf{Y} - \mathbb{X}\mathbf{b})$ equal to zero.

$$\hat{\beta} = (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top \mathbf{Y};\ \hat{\mathbf{Y}} = \hat{g}(\mathbb{X}) = \mathbb{X}\hat{\beta}$$

In the PSET you will derive this and show a few properties.
The suggested resources and these slides should get you started.

Bivariate and multivariate connection

Recall that the sample covariance between two vectors x, y can be written as $\hat{\text{Cov}}[x, y] = \frac{x_c^\top y_c}{n-1}$ and that the sample variance of a vector can be written as $\hat{V}[x] = \frac{x_c^\top x_c}{n-1}$.

We can see a connection between the bivariate and multivariate estimates of the BLP:

bivariate: $\hat{\beta} = \frac{\hat{\text{Cov}}[X,Y]}{\hat{V}[X]} = \frac{\frac{x_c^\top y_c}{n-1}}{\frac{x_c^\top x_c}{n-1}} = \frac{x_c^\top y_c}{x_c^\top x_c} = (x_c^\top x_c)^{-1} x_c^\top y_c$

multivariate: $\hat{\beta} = (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top \mathbf{Y}$

It turns out that the centering doesn't change the estimate of $\hat{\beta}$ so, as we might expect, these can be written in a similar form.