Math 455 Class 4 January 28

Syllabus is posted.. with TA office Hours

Notes from first 3 classes are posted.

Tentative date for 1st midterm: early March (March 2? 4?)

Homework 1 (No math content) due today.

Homework 2 (Review of t- and F-dist) due Friday Jan. 30.

Q: How do we show in Homework 2, that <statistic> has <distribution>?

A: Refer to the definition of <statistic> and <distribution>. You can also do this indirectly, by referring to theorems that quote the definitions. You learned such theorems in Math 447/8.

For example: Suppose $Y_1, \dots, Y_n$ are indep.

normal RVs with mean μ and variance σ^2.

Then $\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$ is normal with mean μ

and variance $\frac{\sigma^2}{n}$. (Theorem 7.2)

Further if $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$

then $\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$

has the χ^2 dist. with $(n-1)$ d.o.f.

and $\bar{Y}$ and S^2 are independent. (Thm. 7.3)

So when you do problems 7.37, 7.38, etc. your answer to "Why?" should involve stuff like this: ~~the~~ the definitions, or theorems referencing the definitions of <distribution>.

Recall the axioms of linear regression.

(which must be memorized for the test, or your 1st ML interview.)

(1) Linear Relationship (Approximate)

$$Y = \beta_0 + \beta_1 X + \varepsilon \leftarrow \text{error.}$$

Y is a linear function of X plus an error.

where β_0, β_1 are fixed but unknown parameters.

Our job is to estimate β_0, β_1 from a data set $(x_1, y_1) \cdots (x_n, y_n)$ generated by this process.

---

(2) in $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ (Indep.)

the errors ε_i are indep. RVs.

---

(3) (Normality) $\varepsilon_i \sim N(0, \sigma^2)$.

ε_i is a normal RV with mean 0.

and variance σ^2

---

(4) (Homoskedasticity) All the errors have the same variance σ^2.

We will see that all these axioms can be weakened in various ways.

We defined the OLS (Ordinary Least-Squares) estimators $\hat{\beta}_0, \hat{\beta}_1$ by defining

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

and saying $\hat{\beta}_0, \hat{\beta}_1$ are the values of β_0, β_1 that minimize $Q(\beta_0, \beta_1)$

And we obtained formulas (calculus)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

$$\left[\text{Think: } \hat{\beta}_1 = \frac{Cov(X, Y)}{Var(X)} \right.$$

$\hat{\beta}_0$ satisfies : $(\bar{X}, \bar{Y})$ is on the regression line.

$$\hat{\beta}_0 + \bar{X} \hat{\beta}_1 = \bar{Y} \quad \text{or} \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Remarks about the OLS estimators:

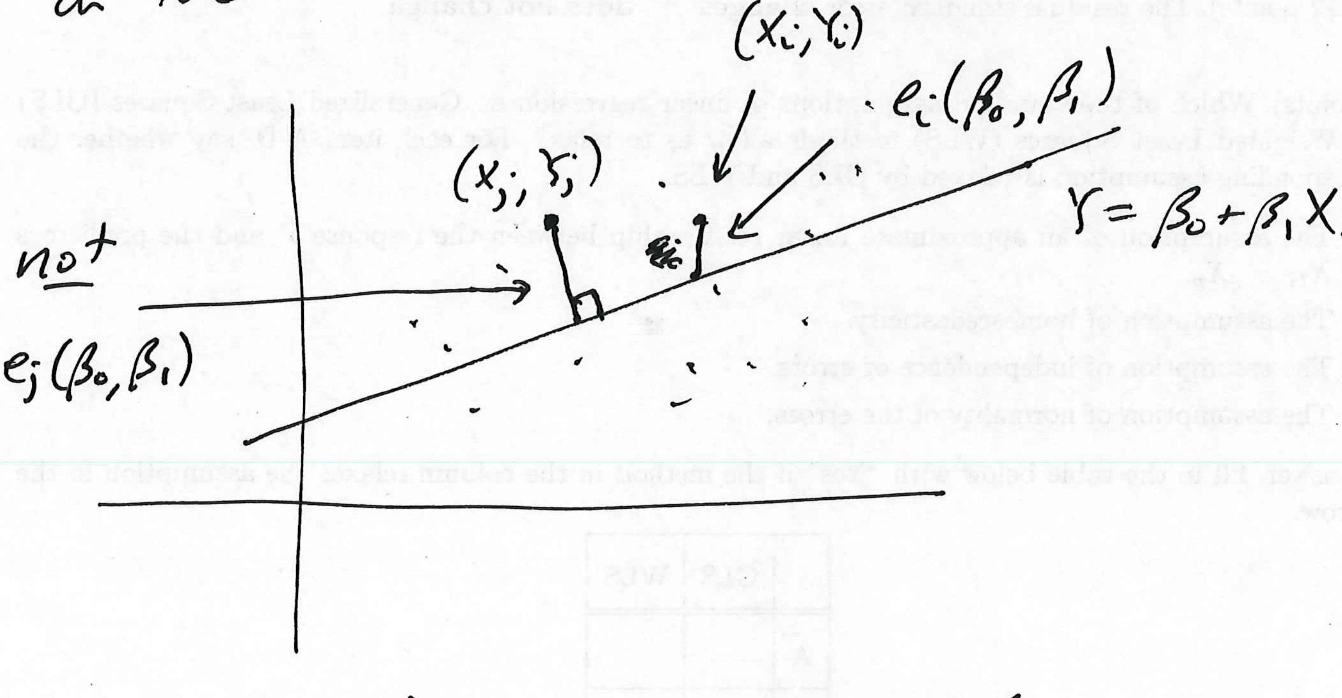
(1) They are <u>estimators</u> in the sense of Math 448.

in particular they are RVs, and we can

talk about E, V, and unbiasedness etc.

(2) They are defined by considering only errors

in the Y-direction:



$$e_i(\beta_0, \beta_1) = Y_i - \beta_0 - \beta_1 X_i$$

We are NOT measuring error by perp. distance. (We could, but this is not part of the def'n of linear regression.)

(3) The variables Y and X are not on an equal footing: Y is called the dependent variable and X the independent variable. (Other terminology is also common.)

- We measure errors in ~~X~~ Y, not X.
- Y is is RV but X is NOT.
- X might be deliberately chosen by the experimenter. (See: course in Exp. Design)
- The model is a "conditional model": it is a model for the dist of $Y|X$. It makes no statement about "the distribution of X".

(4) The estimators $\hat{\beta}_0, \hat{\beta}_1$ (OLS estimators) are linear in the Y_i: We saw that we could rewrite

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

This is a linear function of the Y_i (not X_i)

$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$ is also linear in the Y_i

(5) The OLS estimators are not arbitrary.

there are various optimality properties.

One such is this: linear regression is orthogonal projection.

We change our point of view: instead of n points $(x_1, y_1) \cdots (x_n, y_n)$ in $\mathbb{R}^2$,

think $y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ $x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ 2 vectors in $\mathbb{R}^n$.

Our problem is now to write:

$$y = \beta_0 \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} + \beta_1 x = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_1 \\ \\ \beta_0 + \beta_1 x_n \end{bmatrix}$$

This is in general impossible: the best we can do is look for ~~the~~ a point $\vec{y}$ in the 2-dim'l subspace $\{\beta_0 \vec{1} + \beta_1 x\}$ generated by $\vec{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$ and x which is closest to y.

What is the distance from $\beta_0 \vec{1} + \beta_1 \vec{x}$ to y?

It is $\sqrt{\sum (y_i - \beta_0 - \beta_1 x_i)^2} = \sqrt{Q}$.

minimizing distance or distance^2 gives the same $\hat{y}$ and the same values $\hat{\beta}_0, \hat{\beta}_1$.

The solution is $\hat{y}$ is the orthogonal projection of y on the plane generated by $\vec{1}$, x

---

More optimality properties to be covered later.

(6) The Gauss - Markov Theorem.

the OLS are BLUE

B - Best (min variance)

L - linear (in the Y_i)

U - unbiased $E[\hat{\beta}_0] = \beta_0$ $E[\hat{\beta}_1] = \beta_1$

E - estimators $\hat{\beta}_0, \hat{\beta}_1$ est. of β_0, β_1

(7) The OLS estimators are the same as what we would get from MLE: Maximum Likelihood Estimation gives us the OLS estimators. (This depends on ε_i being normal.)

Remark: The GM Theorem does not require normality of the ε_i.