Math 455 Class 11 February 14

New homework posted!

A practice midterm will be posted, which you will turn in as homework. This was an 80-minute midterm.

Today: properties of linear regression to know for the midterm.

Two optimality properties (from Ch. 2)

#1: Gauss-Markov theorem.

#2: The least squares estimators are the M.LE.

#1: Gauss-Markov: The least-squares estimators are BLUE.

B: Best L: Linear.

U: unbiased E: Estimators.

You do NOT have to know the proof.

You do have to know the meaning of the statement.

These a_i are allowed to depend on X_i and this dependence does NOT have to be linear.

To see that (*) is linear in the Y_i:

Note that $\sum_{i=1}^{n} (X_i - \bar{X}) = \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \bar{X}$

$= n\bar{X} - n\bar{X} = 0$

Also $\sum_{i=1}^{n} (X_i - \bar{X})\bar{Y} = \bar{Y} \sum_{i=1}^{n} (X_i - \bar{X})$

$= \bar{Y} \cdot 0 = 0.$

So (*) can be expressed:

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

call this "a_i": $\hat{\beta}_1$ is a linear fcn of the Y_i's.

What about $\hat{\beta}_0$? $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$

Why is this a linear function of $Y_1, \ldots, Y_n$?

Remarks on meaning of BLUE.

Recall the 1-explanatory variable regression:

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

Formulas for estimators $\hat{\beta}_1 = \dfrac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}$

that is $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$ (*1)

and $\hat{\beta}_0$ (conceptual formula)

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

[$(\bar{x}, \bar{y})$ is ON the regression line.]

Gauss Markov "L" says that $\hat{\beta}_0, \hat{\beta}_1$ are linear. What does this mean?

(*1) above doesn't look like a linear function.

"L" means it is a linear function of $y_1, \ldots, y_n$. That is, there are numbers $a_1, \ldots, a_n$ such that $\hat{\beta}_i = a_1 y_1 + \cdots + a_n y_n$.

They are estimators of the fixed (but unknown) parameters β_i. That is $E[\hat{\beta}_i] = \beta_i$

If $\tilde{\beta}_i$ is another linear estimator of β_i which is also unbiased, then $V[\hat{\beta}_i] \leq V[\tilde{\beta}_i]$

This is the meaning of "best".

Q: We said that $(\bar{X}, \bar{Y})$ is on the regression line. Suppose we run a regression without intercept. That is our model is $Y = 0 + \beta_1 X + \varepsilon$

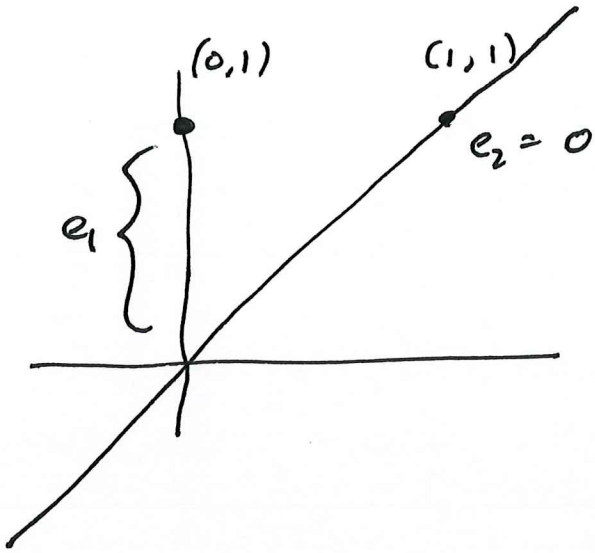
Does $(\bar{X}, \bar{Y})$ still have to be on the regression line? No!

Example: 2 data points: $(0,1)$, $(1,1)$

X	Y
0	1
1	1

What is $\hat{\beta}_1$?

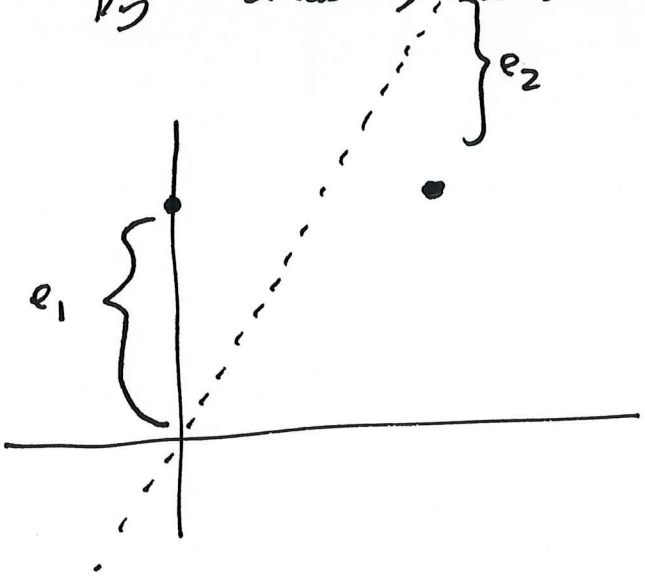
We minimize sum of squares!



$\hat{\beta}_1 = 1$

line: $Y = X$

Q: Why can't we reduce the sum of squares by drawing another line?



A: Because of the definition of the "residuals".

Sum of squares: $e_1^2 + e_2^2$

e_1 will always be 1 because 0 intercept.

with $\hat{\beta}_1 = 1$ $e_2^2 = 0$, it can't be smaller.

Q: Don't the residuals have to sum to 0?

A: In the model $Y = \beta_0 + \beta_1 X + \varepsilon$

the residuals e_i defined by

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + e_i$$

Satisfy $e_1 + \cdots + e_n = 0$.

In the model with 0 intercept, this doesn't have to be true. See example above.