Math 455 Class 12 February 17

Senses in which linear regression is "best"

0. Linear regression is an orthogonal projection.
Othogonal projection is distance-minimizing.

1. Gauss-Markov Theorem:
the ~~best~~ least-squares estimators are BLUE
Best Linear Unbiased Estimators.

Remark: there are many possible sets of assumptions that ~~are~~ imply the conclusion of Gauss-Markov.
Search: "Gauss-Markov Assumptions".

Simple Example: Standard Hypotheses of Linear Regression include independence of errors.

To prove Gauss-Markov, need only that correlations are zero.

2. The least squares estimators are the maximum Likelihood Estimators.

PDF of Normal Dist.

$$f(\mu, \sigma; x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

What is the likelihood function?

It is the same function as the PDF.

BUT: We regard it as a function of the parameters, not x.

This means: the likelihood function is NOT a probability density function of the parameters.

Because a PDF needs to integrate to 1.

OK: What good is a likelihood function?

A: It is used in the method of maximum likelihood.

That is, if we wish to produce estimators for our parameters, given a bunch of data, we can write down the likelihood function (which has our data in it) and ask, what values of μ, σ. ~~maximi~~ (or whatever

parameters are in question), maximize the function?

These values $\hat{\mu}, \hat{\sigma}$ are called the maximum likelihood estimators of μ, σ.

In principle, with our Linear Regression Model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

and our data:

X	$x_1 \dots x_n$
Y	$Y_1 \dots Y_n$

We could write down a likelihood function and try to find $\hat{\beta}_0, \hat{\beta}_1$ maximizing the function.
(Note that our data are treated as known constants.)

This method yields exactly the least-squares estimators.

Q: Why is this true?

A: Because linear regression assumes:
- independence of errors ε_i
- normality of errors: $\varepsilon_i \sim N(0, \sigma^2)$

Q: How do we write down the likelihood function?

A: Use independence, which implies PDF is a product.

$\varepsilon_i \sim N(0, \sigma)$ $\qquad \varepsilon_i = Y_i - \beta_0 - \beta_1 X_i$

PDF of ε_i: $\dfrac{1}{\sqrt{2\pi}\,\sigma} \exp\left(- \dfrac{(Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2} \right)$

Joint PDF of $\varepsilon_1, \ldots, \varepsilon_n$: is a product.

of $\dfrac{1}{\sqrt{2\pi}\,\sigma} \exp\left(- \dfrac{(Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2} \right)$ (from $i=1$ to n)

How do we get from maximizing

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(- \frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2} \right)$$

to least squares?

2 standard reductions:

Suppose $f(\beta_0, \beta_1)$ is <u>any</u> function we want to maximize.

The $\hat{\beta}_0, \hat{\beta}_1$ that maximize f also work for $2f(\beta_0, \beta_1)$

In fact, if c is any positive number (not depending on β_0, β_1) then the $\hat{\beta}_0, \hat{\beta}_1$ work also for $cf(\beta_0, \beta_1)$.

Also: if $\hat{\beta}_0, \hat{\beta}_1$ are the values maximizing $\log f(\beta_0, \beta_1)$ then $\hat{\beta}_0, \hat{\beta}_1$ also work for $f(\beta_0, \beta_1)$.

Going back to our likelihood fcn.

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(- \frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2} \right)$$

We can ignore the constants $\frac{1}{\sqrt{2\pi}}, \sigma$.

in front.

Then take log: get new form

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

Again $\frac{1}{2\sigma^2}$ is a positive factor, as in $c\,f(\beta_0, \beta_1)$ above.

So now we have

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

maximizing this means minimizing the sum of squared residuals, because of the negative sign.

Q: Regression means "reversion to a previous or inferior state". Why is the subject called "Regression Analysis"?

Math 532 March 1 Segment 1

- Why is it called "Regression Analysis"?
- The paradox of "regression to the mean".

Regression literally means "reversion to a previous or inferior state."

What does this have to do with fitting a line to a bunch of points?

Statisticians studying data on heights of mothers (X) and daughters (Y) noticed:

Very tall mothers tended to have daughters taller than the mean but shorter than themselves.

Very short mothers tended to have daughters shorter than the mean, but taller than themselves.

It was as if there were a "mechanism" that "pulled" heights back to the mean value.

Noticing this trick does not truly resolve the paradox.

Something that will often help: writing down a model. We'll use a simple model that you already know: the bivariate normal distribution.

(X, Y) having the bivariate normal dist is specified by 5 parameters: $\mu_X, \sigma_X, \mu_Y, \sigma_Y, \rho$.

We'll assume for simplicity that

$$\mu_X = \mu_Y = \mu \qquad \sigma_X = \sigma_Y = \sigma \qquad 0 < \rho < 1.$$

The paradox purports to show that ~~$\sigma_X < \sigma_Y$~~ $\sigma_Y < \sigma_X$.

And we can compute $V[Y|X] = \sigma^2(1-\rho^2) < \sigma^2$

Now if it were true that $V[Y] = E[V[Y|X]]$

or that $V[Y] = V[Y|X]$

We would have $\sigma_Y = \sqrt{V[Y]} = \sigma\sqrt{1-\rho^2} < \sigma_X$.

But it is NOT true that

$$V[Y] = E[V[Y|X]]$$

They called this "regression to the mean".

They were using least-squares to predict Y from X.

The word "regression" stuck to least-squares.

Paradox of "regression to the mean":

If heights get closer to the mean with every generation, why isn't everyone very close to the average height?

Resolution of the paradox:

First notice a little trick in the statement:

We don't truly expect all heights to get closer to the mean: consider a woman who is (almost) exactly the average height.
We would expect her daughter's height to be further from the mean.

⚠ Many apparent paradoxes contain such tricks.

It is true that $E[Y] = E[E[Y|X]]$

But (Wackerly et. al. Thm 5.15)

$$V[Y] = E[V[Y|X]] + V[E[Y|X]]$$

$$= \sigma^2(1-\rho^2) + \rho^2\sigma^2 = \sigma^2.$$

So it is true that <u>given</u> the mother's height, the variance of the daughter's height is less than σ^2.

But the total variance σ_Y^2 has another term to account for variation in X.

⚠ It's worth writing down a model, because the model thinks of things that you don't!