Math 455 Class 7 February 4

R^2 "coefficient of determination" (Faraway 2.9)

"percentage of variance explained"

$$R^2 = 1 - \frac{\sum(\hat{Y}_i - Y_i)^2}{\sum(Y_i - \bar{Y})^2} = 1 - \frac{RSS}{\text{Total SS}}$$

↑ corrected for mean.

$$R^2 = Cor(Y, \hat{Y})^2$$

$$0 \le R^2 \le 1$$

R^2 is a measure of "how well the model fits the data" or "goodness of fit."

R^2 close to 1 — good

R^2 close to 0 — bad.

Warning: R^2 is only a measure of goodness.

How good is good enough? It depends on context.

Example: in a physics experiment, $R^2 = 0.9$ could be terrible - results not publishable.

If we are studying the effect of lead exposure on health problems, $R^2 = 0.06$ could be very significant.

R^2 is only one measure and not a substitute for looking at the data: see Faraway p. 25.

R^2 is connected to the F-statistic (Wackerly p. 627, see also Exercise 11.84 (a))

$$F = \frac{n-(k+1)}{k} \frac{R^2}{1-R^2}$$

(This is the F-statistic for the absolute F-test that shows up in the regression output.)

Remark on notation: in Wackerly k = # of "predictors" or "indep. variables" $x_1, \ldots, x_k$.

In Faraway, p = # of predictors $x_1, \ldots, x_p$.

In still other texts, p = # of parameters.

Notice that in the model with p predictors, we have $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$

$p+1$ parameters.

We said R^2 was a measure of goodness of fit. Another measure is $\hat{\sigma}^2$.

Remember that in the model $\varepsilon_i \sim N(0, \sigma^2)$ where we do NOT know σ^2.

$\hat{\sigma}^2$ is an estimator of σ^2

ICBST $E\left[\sum_{i=1}^{n} e_i^2\right] = \sigma^2(n-p)$

$=$

$E\left[\hat{\varepsilon}^T\hat{\varepsilon}\right]$

Estimator: $\hat{\sigma}^2 = \frac{\hat{\varepsilon}^T\hat{\varepsilon}}{n-p} = \frac{RSS}{n-p}$.

error for $\hat{\beta}_i$: $se(\hat{\beta}_i) = \sqrt{(X^TX)^{-1}_{ii}}\ \hat{\sigma}$.

Notation and Terminology: We assemble our predictor vars into a matrix X, the model matrix,

$$X = \begin{bmatrix} 1 & | & & | \\ 1 & & & \\ \vdots & x_1 & \cdots & x_p \\ & | & & | \\ 1 & | & & | \end{bmatrix}$$

n rows. of data.

p+1 cols.

parameters

"Variable space" vs "subject space".

Coordinates correspond to variables studied.

e.g. blood pressure

Coordinates correspond to subjects in experiment.

n vectors in $\mathbb{R}^{p+1}$

p+1 vectors in $\mathbb{R}^n$.

either way, it is the same matrix X

"model matrix" or "design matrix".

in R: model.matrix (lmod)

Q: Why is it called "regression" analysis?

Regression means "reversion to a previous or inferior state". We are fitting a line to a bunch of points. What sense does "regression" make here?

A: We admit that the terminology doesn't make a lot of sense. It developed for historical reasons.

Q: What were the historical reasons?

A: The phenomenon of "regression to the mean".

Example: Statisticians from hundreds of years ago studied the heights of mothers (X) and daughters (Y). We try to predict Y from X using least-squares. We notice:

- Very tall mothers tend to have daughters taller than the mean, but shorter than themselves.
- Very short mothers tend 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.

This was called "regression to the mean".

Since the statisticians were using least-squares, the word "regression" stuck to least-squares.

The Paradox of Regression to the mean:

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

Note that there are some simple reasons why this reasoning is deceptive. Consider a mother of exactly average height. Do we expect her daughter's height to be closer to the mean?

No: we expect the daughters height to be different from the mean.

A probability model for this:

X, Y have the bivariate normal dist.

(See Wackerly 5.10)

We assume $\mu_X = \mu_Y = \mu \qquad \sigma_X = \sigma_Y = \sigma.$

$0 < \rho < 1.$

The paradox argues that with each generation variance goes down: $\sigma_Y^2 < \sigma_X^2$.

We know in the model that $\sigma_Y = \sigma_X = \sigma$.

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

IF it <u>were</u> true that $E[V[Y|X]] = V[Y]$

<u>then</u> we <u>would</u> have $\sigma_Y = \sqrt{V[Y]}$

$= \sigma^2(1-\rho^2)^{1/2} < \sigma_X.$

$= \sigma$

But it is NOT true that

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

In fact (Wackerly 5.15):

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

In the context of the bivariate normal.

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

It is true that <u>given</u> the mothers height, the variance of the daughters height is $< \sigma^2$.

Total variance of daughters height has another term to account for variance in X.

In this model it is true that if the mother is tall, say $X = \mu + 3\sigma$.

then the expected value of Y is $\mu + 3\rho\sigma$.

Since $\rho < 1$ $\mu + 3\rho\sigma < \mu + 3\sigma$.

In the model (bivariate normal model)
ALL the statements of the paradox are true.
BUT the variance of daughters heights is
EXACTLY the variance of mothers' height.

This is part of why we write down models: "The model is smarter than you"

[This means that forcing yourself to write down a model provides a check on your reasoning.]

Note that this is true even if the model is wrong. Hence the saying:

"All models are wrong. Some models are useful."