Math 455 Class 31 April 13

Practice midterm is posted to Gradescope.

Warning: not every possible topic is included.

One midterm topic we have not yet talked about

the "Box-Cox Transformation".

Suppose we think some other model is better than a linear model.

One thing we could do is introduce new predictors X_i^2, X_i^3, etc. or $\sqrt{X_i}$, etc.

Another thing we could do is change Y

e.g. replace Y by Y^2 or $\sqrt{Y}$ or Y^3.

Box-Cox is a method of considering all possible transformations Y^λ, where $\lambda = 0$ corresponds to $\log Y$.

Chapter 14: Categorical Variables

Variables with values like Red - Yellow - Blue that are not numbers.

R will produce regression output anyway.

This is done by translating the variables into numbers. There is a standard, default scheme "Reference Level and Dummy Codings."

for doing this, but many other choices are possible.

<u>Example:</u> We have a categorical var. with 3 possible values R, ~~B~~ Y, B

call our variable X.

X	D_1	D_2
R	1	0
B	0	0
R	1	0
Y	0	1
B	0	0

Replace X with D_1 and D_2.

Choose "Ref." level. B.

~~We~~ We <u>chose</u> ref level B.

D_1 & D_2 are 0 on B.

We chose D_1 to mean R. and

D_2 to mean Y.

this means : $D_1 = 1$ if $X = R$ and

$D_1 = 0$ otherwise.

$D_2 = 1$ if $X = Y$ and

$D_2 = 0$ otherwise.

Consequences: If X has 3 ~~levels~~

possible values, we need 2 dummy vars.

If X has d possible values, we need

(d-1) dummy variables.

Q: what if we have 8 categorical variables, each of which takes 10 possible values?

A: We need $8 \cdot (10-1) = 8 \cdot 9 = 72$ dummy vars.

Q: What if we don't have that many rows of data?

A: That is a problem!

Q: This "reference level" seems unsymmetric and arbitrary. Why not make it symmetric by including $D_3 = 1$ if $X = B$

$D_3 = 0$ otherwise

in the example?

A: Then $D_3 + D_2 + D_1$ is always 1.

This means that our design matrix X has linearly dependent rows, because we usually include a column of 1s.

This is a problem!

Q: I still think it is unsymmetric.

A: Other coding schemes are also possible.

See chapter 14.

The coding schemes don't change anything fundamental about the regression but they can change the interpretation of the coefficients.

Q: What is the interpretation of the coefficients for the standard "dummy-coding" scheme?

A: Imagine that we are trying to predict

Y = accident rate from

X = color of car.

remember we have $D_1 = 1$ if $X = R$

$D_2 = 1$ if $X = Y$.

the coeff $\hat{\beta}_1$ of D_1 is the <u>difference</u> in accident rate between red cars and blue cars.

Similarly for $\hat{\beta}_2$: it is the difference in accident rate between yellow cars and blue cars.