Multiple Regression

Psychology 3256

Introduction

- Often we are interested in simple 1 to 1 variable relationships
- but let's say that $r = .50$ for some relationship
- $r = cov_{xy}/s_x s_y$
- How much variance is accounted for by one variable in the other?

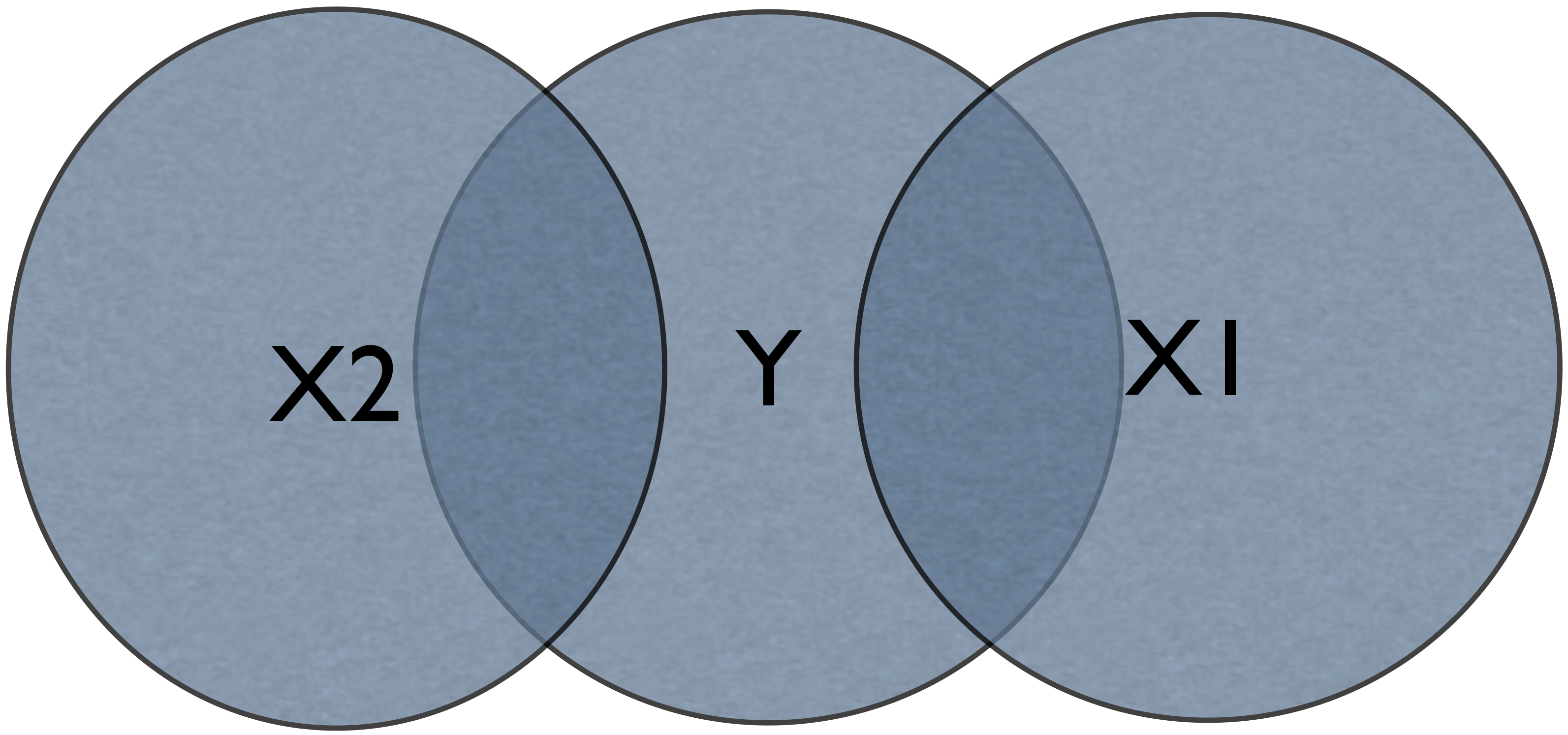
How much indeed

- well r deals with standard deviations, so square r
- $r^2 = .25$
- so we have accounted for 25 percent of the variance
- which means there is 75 percent left!

ergo..

- There must be other variables that account for the rest of the variance
- We deal with this by bringing them in to the model

Pretty pictures



In general, we have a model...

- $\hat{y} = b_0+b_1x_1+b_2x_2+...+b_{p-1}x_{p-1}+e$
- We have p-1 predictor variables
- This is for the data set itself, these are statistics, not parameters

In the population...

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{p-1} X_{p-1} + \varepsilon$$

$$\varepsilon NID(0, \sigma_{\varepsilon}^2)$$

- ε is not a prediction error, it is individual variation
- Note it is Y not a predicted y hat

What we get

- if p-1 =1 we get a line
- if p-1 =2 we get a surface or plane
- if p-1 >2 we get a hyperplane in hyperspace!
- Best not to try and visualize a hyperplane..

And you thought you were done with ANOVA....

- We can find out if our regression model is significant with ANOVA
- Variance due to regression (the model)
- Variance due to residual

Yes, ANOVA

SV	df	SS	MS	F
regression	p-1		SSREG/	MSREG/MSRES
residual	n-p		SSRES/p-1	
TOTAL	n-1			

- This analysis is about the whole model
- Not about individual variables
- One thing that is the sum of it’s parts

Finer grained analysis

- So, of course the model is significant, or it bloody well better be
- We are much more concerned with how much extra variation is accounted for by adding another variable to in to the model
- R^2 = SSREG/SSTOTAL

Adding variables

- If you have a model with 5 x variables and you add a 6th, does R^2 go up?
- It has to
- by how much?
- is it enough to deal with the loss of df and the increase in complexity?

So look at something other than R^2

$$R_a^2 = (\frac{n-1}{n-p})\frac{SSE}{SST}$$

- adjusted R^2
- This is weighted by the number of variables in the model, it can go down when more variables are added

Which is the best model?

sv	df	SS
Reg	1	30 / 50
Res	n-2	100 / 80

X1 / X2

sv	df	SS
Reg	2	80
Res	n-3	50

X1 X2

Sums of Squares

- There are Type I and Type II SS
- Type I SS depend on the order variables go in to the model, Type IIs don't
- Let's say we have a three variable model, X1 X2 and X3

Comparison

	Type I	Type II
X1	SSR(X1)	SSR(X1 | X2 X3)
X2	SSR(X1 X2)	SSR(X2 | X1 X3)
X3	SSR(X1 X2 X3)	SSR(X3 | X1 X3)

Why should you care?

- If there is no correlation between variables, then Type 1 = Type 2
- If there is a correlation Type 1 $\neq$ Type 2
- a bit more on this later..

What can TII give you?

- So, Type IIs give you the extra variation accounted for by having a variable in the model, given the others are already there
- This can give us the coefficient of partial determination
- Sort of the opposite of R^2 which is the coefficient of multiple determination

Extra variation

- So it (the coefficient of partial determination) gives us the extra variation accounted for by adding in another variable
- You can square it and get the partial correlation, which is pretty useful

Why does this matter?

- think about the model
- $\hat{y} = b_0+b_1x_1+b_2x_2+...+b_{p-1}x_{p-1}+e$
- nothing there about two variables together
- this is a problem called multicolinearity

so you are violating an assumption…

- the bs will change
- how do we detect it?
- look at correlation between x variables
- you might have to chuck something

another assumption

- we assume a linear model
- what if it is not linear?

$$Y = \lambda_0^{\lambda_1^{xe}}$$

Aaaahhhh!

$$\log y = \log \lambda_0 + \log x_1 \lambda_1 + \log e$$

$$y = b_0 + b_1 x_1 + e$$

- intrinsically linear
- careful, not everything is

we assume an additive model

- There is no mention of interactions
- but you could put something in $x_1 x_2$
- Tough to know what the term should be, EDA is the key

Selection of predictors

- qualitative
- ok if binary
- 0 and 1, not 1 and 2
- watch Likert scales
- experimental variables can be good, no colinearity etc

Model Building

- So how do you choose what variables to use?
- Much different from ANOVA, we are making a prediction with Multiple Regression
- you usually start out with a lot of variables

you could do all of them

- 3 variables, there are 7 models
- 4 there are 15
- for 10 there are like a zillion

Residual plots

- Can be very useful
- Can find anomalies
- Can find non linear relationships

Forward Selection

- An automatic method
- start with the x that has the highest R^2
- add in the next variable that gives the biggest jump in R^2
- Keep going until the jump in R^2 is not big enough

How big is big enough?

$$F^* = \frac{MSR(X_1 | X_2)}{MSR(X_1 X_2)}$$

Backwards elimination

- The opposite
- Start with all of the variables in the model
- Delete variables that contribute the least
- Smallest F*

Stepwise Regression

- Combine the two
- Go forward
- Check F* for each variable
- Drop or add if necessary
- Set criteria for adding and dropping
- F* to enter >= F* to leave

The thing is...

- The automatic methods only look at Fs
- Not residual plots
- don’t care about multicolinearity
- don’t worry about non linear stuff

An approach

- Start with a correlation matrix
- pick a subset, if small enough, do all models
- try all 3 automatic methods
- check for outliers residual plots
- do it again!