Wednesday PM

- Presentation of AM results
- Multiple linear regression
  - Simultaneous
  - Stepwise
  - Hierarchical
- Logistic regression

Multiple regression

- Multiple regression extends simple linear regression to consider the effects of multiple independent variables (controlling for each other) on the dependent variable.
- The line fit is:
  $Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots$
- The coefficients (b_i) tell you the independent effect of a change in one dependent variable on the independent variable, in natural units.

Multiple regression in SPSS

- Same as simple linear regression, but put more than one variable into the independent box.
- Equation output has a line for each variable:

Coefficients: Predicting Q2 from Q3, Q4, Q5

	Unstandardized		Standardized		
	B	**SE**	**Beta**	**t**	**Sig.**
(Constant)	.407	.582		.700	.485
Q3	.679	.060	.604	11.345	.000
Q4	-.028	.095	-.017	-.295	.768
Q5	.112	.066	.095	1.695	.091

- Unstandardized coefficients are the average effect of each independent variable, controlling for all other variables, on the dependent variable.

Standardized coefficients

- Standardized coefficients can be used to compare effect sizes of the independent variables *within the regression analysis.*
- In the preceding analysis, a change of 1 standard deviation in Q3 has over 6 times the effect of a change of 1 sd in Q5 and over 30 times the effect of a change of 1 sd in Q4.
- However, βs are not stable across analyses and can't be compared.

Stepwise regression

- In simultaneous regression, all independent variables are entered in the regression equation.
- In stepwise regression, an algorithm decides which variables to include.
- The goal of stepwise regression is to develop the model that does the best prediction with the fewest variables.
- Ideal for creating scoring rules, but atheoretical and can capitalize on chance (post-hoc modeling)

Stepwise algorithms

- In *forward* stepwise regression, the equation starts with no variables, and the variable that accounts for the most variance is added first. Then the next variable that can add new variance is added, if it adds a significant amount of variance, etc.
- In *backward* stepwise regression, the equation starts with all variables; variables that don't add significant variance are removed.
- There are also hybrid algorithms that both add and remove.

Stepwise regression in SPSS

- Analyze…Regression…Linear
- Enter dependent variable and independent variables in the independents box, as before
- Change “Method” in the independents box from “Enter” to:
  - Forward
  - Backward
  - Stepwise

Hierarchical regression

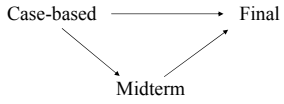
- In hierarchical regression, we fit a hierarchy of regression models, adding variables according to theory and checking to see if they contribute additional variance.
- You control the order in which variables are added
- Used for analyzing the effect of dependent variables on independent variables in the presence of moderating variables.
- Also called *path analysis*, and equivalent to *analysis of covariance (ANCOVA)*.

Hierarchical regression in SPSS

- Analyze…Regression…Linear
- Enter dependent variable, and the independent variables you want added for the smallest model
- Click “Next” in the independents box
- Enter additional independent variables
- …repeat as required…

Hierarchical regression example

- In the hyp data, there is a correlation of -0.7 between case-based course and final exam.
- Is the relationship between final exam score and course format moderated by midterm exam score?



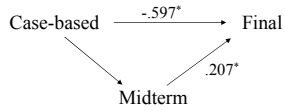
Hierarchical regression example

- To answer the question, we:
  - Predict final exam from midterm and format (gives us the effect of format, controlling for midterm,
    and the effect of midterm, controlling for format)
  - Predict midterm from format
    (gives us the effect of format on midterm)
- After running each regression, write the βs on the path diagram:

Predict final from midterm, format

Coefficients

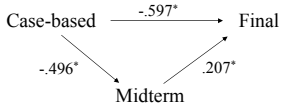
	B	SE	Beta	t	Sig.
(Constant)	50.68	4.415		11.479	.000
Case-based course	-26.3	3.563	**-.597**	-7.380	.000
midterm exam score	.156	.061	**.207**	2.566	.012



Predict midterm from format

Coefficients

	B	SE	Beta	t	Sig.
(Constant)	63.43	3.606		17.59	.000
Case-based course	-29.2	5.152	**-.496**	-5.662	.000



- Conclusions: The course format affects the final exam both directly and through an effect on the midterm exam. In both cases, lecture courses yielded higher scores.

Logistic regression

- Linear regression fits a line.
- Logistic regression fits a cumulative logistic function
  - S-shaped
  - Bounded by [0,1]

- This function provides a better fit to binomial dependent variables (e.g. pass/fail)
- Predicted dependent variable represents the probability of one category (e.g. pass) based on the values of the independent variables.

Logistic regression in SPSS

- Analyze…Regression…Binary logistic (or multinomial logistic)
- Enter dependent variable and independent variables
- Output will include:
  - Goodness of model fit (tests of misfit)
  - Classification table
  - Estimates for effects of independent variables
- Example: Voting for Clinton vs. Bush in 1992 US election, based on sex, age, college graduate

Logistic regression output

- Goodness of fit measures:

-2 Log Likelihood	2116.474	*(lower is better)*
Goodness of Fit	1568.282	*(lower is better)*
Cox & Snell - R^2	.012	*(higher is better)*
Nagelkerke - R^2	.016	*(higher is better)*

	Chi-Square	df	Significance
Model	18.482	3	.0003

(A significant chi-square indicates poor fit (significant difference between predicted and observed data), but most models on large data sets will have significant chi-square)

Logistic regression output

Classification Table
The Cut Value is .50

		Predicted Bush B	Predicted Clinton C	Percent Correct
Observed				
Bush	B	0	661	.00%
Clinton	C	0	907	100.00%
			Overall	57.84%

Logistic regression output

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
FEMALE	.4312	.1041	17.2	1	.0000	.0843	1.5391
OVER65	.1227	.1329	.85	1	.3557	.0000	1.1306
COLLGRAD	.0818	.1115	.53	1	.4631	.0000	1.0852
Constant	-.4153	.1791	5.4	1	.0204		

- B is the coefficient in log-odds; Exp(B) = e^B gives the effect size as an odds ratio.
- Your odds of voting for Clinton are 1.54 times greater if you're a woman than a man.

Wednesday PM assignment

- Using the semantic data set:
  - Perform a regression to predict total score from semantic classification. Interpret the results.
  - Perform a one-way ANOVA to predict total score from semantic classification. Are the results different?
  - Perform a stepwise regression to predict total score. Include semantic classification, number of distinct semantic qualifiers, reasoning, and knowledge.
  - Perform a logistic regression to predict correct diagnosis from total score and number of distinct semantic qualifiers. Interpret the results.