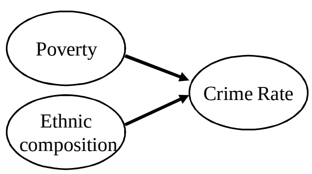
Multivariate Regression

Multivariate Regression

- Q: what if a dependent variable is affected by more than one independent variables?
- E.g. Crime rate can be predicted by
  - Poverty rate,
  - Ethnic composition
- Income can be determined by
  - Education
  - Occupation
  - Age
  - Gender

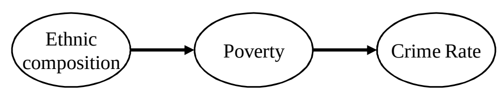
Multivariate Regression

- Possible relations among independent variables:
- 1. Each has a totally distinct (separate) effect on the dependent variable



Multivariate Regression

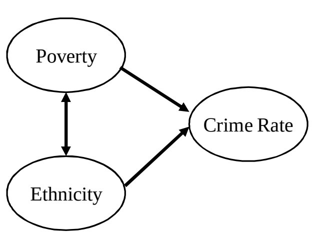
- 2. One variable "operates via" another (chain relationship)
  - i.e., poverty "mediates" the effect of ethnicity
  - Certain ethnicities are more likely to be in poverty
  - But has no additional impact beyond that



- Any observed bivariate relationship between ethnicity and crime rate is "**spurious**"
  - It disappears when poverty is controlled!

Multiple Regression

- 3. Both variables have partial independent effects
  - But may be partially mediated by each other



Multivariate Regression

- 4. There is no relation among variables
- 5. The relation is complex, conditional
  - Example: ethnicity matters "up to a point", but beyond that poverty matters more
  - Statistical interaction: the association between two variables changes as the value of a third variable change
- 6. Variables have a relationship that is nonlinear
- 7. The effect of both variables is actually mediated by some other unobserved variable
  - All observed effects are spurious

Multivariate Regression

- Q: What if a dependent variable is affected by more than one independent variable?
- Option #1: Do two separate bivariate regressions
  - One regression for each independent variable
  - This implicitly assume that neither independent variable mediates the other

Crime=a + b*Poverty

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-10.136	4.121		-2.460	.017
	Poverty	1.323	.275	.566	4.804	.000

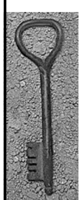
a. Dependent Variable: CrimeRate

Crime=a + b*White

Coefficients[a]

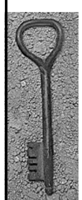
Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	56.766	6.962		8.153	.000
	White	-.571	.082	-.706	-6.983	.000

a. Dependent Variable: CrimeRate



Multivariate Regression

- Q: What if a dependent variable is affected by more than one independent variable?
- Option #1: Do several bivariate regressions
  - One regression for each independent variable
  - This assumes that neither independent variable mediates the other
  - But this assumption may be wrong: e.g. ethnic composition may affect poverty rate
  - What is the effect of poverty rate over and above the effect of ethnic composition?



Option #2: Multivariate Regression

- Examine "partial" relationships after the effects of other variables have been "controlled" (taken into account)
- Determine the effects of variables "over and above" other variables
  - And shows the **relative** impact of different factors on a dependent variable
- Several independent variables usually improve your predictions of the dependent variable

Multivariate Regression

Crime = a + b1*Poverty + b2*White

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	36.255	8.653		4.190	.000
	White	-.463	.080	-.573	-5.770	.000
	Poverty	.802	.232	.343	3.455	.001

a. Dependent Variable: CrimeRate

	Model 1	**Model 2**	**Model 3**
	Crime=a +b1* poverty	Crime=a +b2* white	Crime=a +b1*poverty + b2*white
a	-10.1364	56.76599	36.25453
b1	1.32296		0.80185
b2		-0.57115	-0.46322

- The signs for both variables are maintained
- Both coefficients decrease in absolute values – both poverty and white have smaller effects in model 3

Correlations

		CrimeRate	Poverty	White
CrimeRate	Pearson Correlation	1	.566**	-.706**
	Sig. (2-tailed)		.000	.000
	N	51	51	51
Poverty	Pearson Correlation	.566**	1	-.389**
	Sig. (2-tailed)	.000		.005
	N	51	51	51
White	Pearson Correlation	-.706**	-.389**	1
	Sig. (2-tailed)	.000	.005	
	N	51	51	51

**. Correlation is significant at the 0.01 level (2-tailed).

Multivariate Regression

- Poverty rate has a positive effect on crime rate **<u>over and above</u>** (i.e. "controlling for") ethnicity composition.
- Ethnicity has a negative effect on crime rate, over and above poverty rate.
- Both variables have smaller effects than in bivariate regressions
- Possible interpretation: correlation between poverty and ethnicity

Multivariate Regression Model

- A two-independent variable regression model:

$$Y = a + b_1 X_1 + b_2 X_2 + e$$

- The full multiple regression model is:

$$Y = a + b_1 X_1 + b_2 X_2 \ldots + b_k X_k + e$$

Multivariate Regression: Slopes

- b_1, b_2,… "partial slopes", "partial regression coefficients"
- b_i: the average change in Y associated with one unit change in X_i, <u>when the other independent variables are held constant</u>
- Example: with %white held constant, what is the average change in crime rate with one unit change in poverty rate?

Multiple Regression: Slopes

- Regression slope for the two variable case:

$$b_1 = \left(\frac{s_Y}{s_{X_1}}\right)\frac{r_{YX_1} - r_{YX_2} r_{X_1 X_2}}{1 - r^2_{X_1 X_2}}$$

- b_1 = slope for X_1 – controlling for the other independent variable X_2
- b_2 is computed symmetrically. Swap X_1s, X_2s
- Compare to bivariate slope: $b_{YX} = \frac{s_Y}{s_X} r_{YX}$

Multivariate Regression: Slopes

$$b_1 = \left(\frac{s_Y}{s_{X_1}}\right)\frac{r_{YX_1} - r_{YX_2} r_{X_1 X_2}}{1 - r^2_{X_1 X_2}} \qquad versus - b_{YX} = \frac{s_Y}{s_X} r_{YX}$$

- What happens to b_1 if X_1 and X_2 are totally uncorrelated?
- Answer: The formula reduces to the bivariate
- If two variables (X1, X2) are correlated, the X variable that is more correlated with Y will have a higher slope in multivariate regression (compared to bivariate regression); the slope of the less-correlated variable will shrink
- Thus, slopes for each variable are adjusted to how well the other variable predicts Y; it is the slope "controlling" for other variables

Multivariate Regression Slopes

- One last thing to keep in mind…

$$b_1 = \left(\frac{s_Y}{s_{X_1}}\right)\frac{r_{YX_1} - r_{YX_2} r_{X_1 X_2}}{1 - r^2_{X_1 X_2}} \qquad versus - b_{YX} = \frac{s_Y}{s_X} r_{YX}$$

- What happens to b_1 if X_1 and X_2 are almost perfectly correlated?
- Answer: The denominator approaches zero
  - The slope "blows up", approaching infinity
- Highly correlated independent variables can cause trouble for regression models… watch out

Multivariate Regression Estimation

- Calculating b's involves solving a set of equations to minimize squared error
  - – Analogous to bivariate
- The optimal estimator has minimum variance and is referred to as "BLUE":
  - – Best Linear, Unbiased Estimate
- The BLUE Multiple Regression has more assumptions than bivariate
  - – Will discuss assumptions later

Interpreting Multivariate Results

- Multivariate regression with two variables: A, B
- If slopes of A, B are the same as bivariate, then **each has an independent effect**
- If **A** remains large, **B** shrinks to zero we typically conclude that **effect of B was spurious, or operates through A**
- If both A and B shrink a little, each has an effect, then some overlap or mediation is occurring

Standardized Regression Coefficients

- Problem: Regression slopes reflect the units of the independent variables. How do you compare the size of coefficients of two variables?
  - – Poverty and ethnicity are measured in different units. How do you know which slope is larger?
- Answer: Create "standardized" coefficients

Standardized Regression Coefficients

- Standardized Coefficients
  - – Also called "Betas" or "Beta Weights"
  - – Symbol: Greek b with asterisk: β^*
  - – Equivalent to Z-scoring (standardizing) all independent variables before doing the regression
- Formula of coefficient for X_j:

$$\beta_j^* = \left(\frac{s_{X_j}}{s_Y}\right) b_j$$

- Result: The unit is standard deviation
  - Betas: Indicates the effect of 1 standard deviation change in X_j on Y

Standardized Regression Coefficients

Coefficients[a]

Model		Unstandardized Coefficients B	Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	36.255	8.653		4.190	.000
	White	-.463	.080	-.573	-5.770	.000
	Poverty	.802	.232	.343	3.455	.001

a. Dependent Variable: CrimeRate

- Betas give you a sense of which variables "matter more": ethnicity is in fact more important than poverty
- An increase of 1 standard deviation in Poverty results in a 0.34 standard deviation increase in Crime rate

R-Square in Multiple Regression

- Recall the R-square formula: $R^2 = \frac{SS_{REGRESSION}}{SS_{TOTAL}}$
- The same concept applies in the multivariate case. But, $SS_{regression}$ and SS_{total} are based on the multivariate regression
- R^2: Share of variations in Y explained by **ALL** independent variables in the model
- Example: Crime rate=a + b1*poverty +b2*white

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.774[a]	.599	.582	6.930

a. Predictors: (Constant), Poverty, White

Poverty White

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.566[a]	.320	.306	8.926

a. Predictors: (Constant), Poverty

Poverty

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.706[a]	.499	.489	7.665

a. Predictors: (Constant), White

White

R-Square in Multiple Regressions

- Typically, the addition of new variables results in better prediction of Y, less error (e), and thus a larger R-square
- Caution: Do not overload variables, correlation between variables can bring you trouble – multicolliniality.
- Theory determines what variables to be added

Dummy Variables

- Question: How can we incorporate nominal variables (e.g., race, gender) into regression?
- Answer: Dummy variables
  - Dichotomous variables coded to indicate the presence (1) or absence (0) of something
- Example: Gender, coded 1=Female, 2=Male
  - Make two new variables
    - one for "female" =1, "others"=0
    - another for "male"=1, "others"=0
  - Include **only one** of the variables in the regression, the other is the baseline

Dummy Variables

- Example: Race
  - Coded 1=white, 2=black, 3=other (like GSS)
- Make 3 dummy variables:
  - "DWHITE" is 1 for whites, 0 for everyone else
  - "DBLACK" is 1 for blacks, 0 for everyone else
  - "DOTHER" is 1 for "others", 0 for everyone else
- Then, include only two of the three variables in the multivariate regression model
- <u>Always leave one category out as the baseline for comparison</u>

Dummy Variables: Interpretation

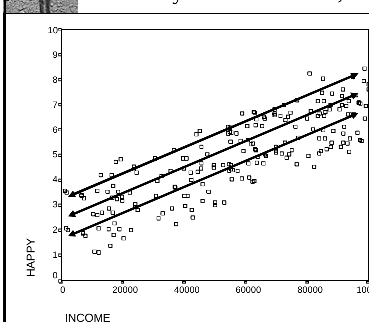
- Consider the following regression equation:

$$HAPPY_i = a + b_1 INCOME_i + b_2 DFEMALE_i + e_i$$

- Q: What if the case is a male
- Answer: DFEMALE is 0, so the entire term becomes zero.
- If the case is female, the b_2 coefficient remains in the equation and adds to the constant
  - Thus, the coefficient for dummy variables reflects "differences in the constant" between groups
  - Tells which gender is higher/lower overall– not slope!

Dummy Variables: Interpretation

- Visually: Women = blue, Men = red



Overall slope for all data points

Note: Line for men, women have same slope… but one is high other is lower.

The constant differs!

If women=1, men=0: The constant (a) reflects men only. Dummy coefficient (b) reflects increase for women (relative to men)

Dummy Variables: Interpretation

- A higher constant generates a different line, either higher or lower
  - Variable: DFEMALE (women = 1, men = 0)
  - A positive coefficient (b) indicates that women are consistently higher compared to men (on dep. var.)
  - A negative coefficient indicated women are lower
- Example: If DFEMALE coeff = 1.2:
  - "Women are on average 1.2 points higher than men"
  - Comparison is always with group "left out" of model
- Thus, dummy variables are like t-test, ANOVA

Dummy Variables: example

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-10.537	4.089		-2.577	.013
	Poverty	1.237	.279	.529	4.429	.000
	Coast	3.602	2.548	.169	1.414	.164

a. Dependent Variable: CrimeRate

Dummy variable: coast (1, 0)

Crime=a + b1*poverty + b2*coast

Coastal states: crime=-10.537+1.237*poverty + 3.602

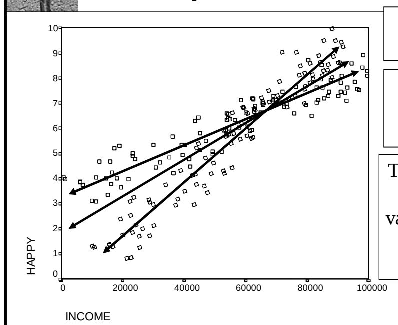
Non-coastal states: crime=-10.537+1.237*poverty

Interaction Terms

- Question: What if you suspect that a variable has a totally different slope for two different sub-groups in your data?
- Example: Income and Happiness
  - Perhaps men are more materialistic -- an extra dollar increases their happiness a lot
  - If women are less materialistic, each dollar has a smaller effect on income (compared to men)
- Issue isn't men = "more" or "less" than women
  - Rather, the slope of a variable coefficient (for income) differs across groups

Interaction Terms

- Visually: Women = blue, Men = red



Overall slope for all data points

Note: Here, the slope for men and women differs.

The effect of income on happiness (X1 on Y) varies with gender (X2). This is called an "interaction effect"

Interaction Terms

- Interaction effects: Differences in the relationship between two variables within categories of a third variable
- Strategy #1: Analyze each group separately
  - For female: $HAPPY_i = a + b_1 Income_i + e_i$
  - For male: $HAPPY_i = a + b_1 Income_i + e_i$
  - Two different coefficients, two intercepts
  - No statistical test for $\beta_f = \beta_m$

Interaction Terms

- Strategy #2: Multiply the two variables of interest: (DFEMALE, INCOME) to create a new variable
  - Called: DFEMALE*INCOME
  - Add that variable to the multiple regression model

$$HAPPY_i = a + b_1 Income_i + b_2 Dfemale_i + b_3 Income_i * Dfemale_i + e_i$$

Interaction Terms

$$HAPPY_i = a + b_1 Income_i + b_2 Dfemale_i + b_3 Income_i * Dfemale_i + e_i$$

- A positive b_3 indicates the slope for income is higher for women vs. men
  - A negative effect indicates the slope is lower
  - Size of coefficient indicates change in slope
- If b3 is significant, $\beta_f \neq \beta_m$
- E.g.
  - Dfemale: b = 2
  - Income: b =0.5
  - Income*Dfemale: b = -0.2
- Interpretation: Slope for income is 0.5 for men, 0.3 for women

Interaction Terms: example

Coefficients[a]

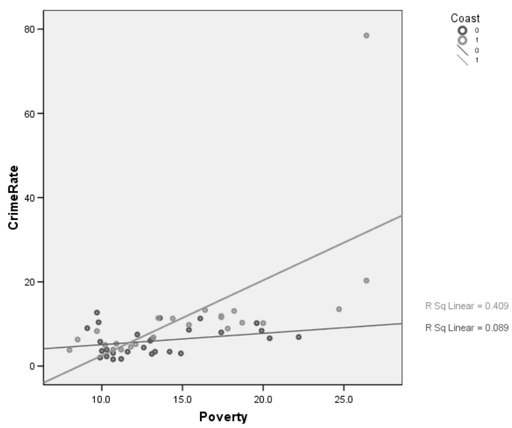
Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	2.406	6.009		.400	.691
	Poverty	.269	.434	.115	.619	.539
	Coast	-17.790	8.020	-.834	-2.218	.031
	PovertyCoast	1.519	.544	1.205	2.794	.008

a. Dependent Variable: CrimeRate

Crime=a+b1*Poverty+b2*Coast+b3*PovertyCoast

Coastal states: crime=2.4+0.27*Poverty-17.8+1.52*Poverty

Non-coastal states: crime=2.4+0.27*Poverty



Interaction Terms

- Continuous variable can also interact with each other
- Example: Perhaps poverty has a smaller effect on crime rate in states with more whites
  - As %white increases, the slope for poverty would decrease
- Simply multiply poverty and white create the interaction term “poverty*white”
  - And add it to the model

Example: Interaction

Crime=a + b1*White+ b2*Poverty + B3*PovertyWhite

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-34.352	9.330		-3.682	.001
	White	.463	.112	.572	4.121	.000
	Poverty	5.152	.496	2.204	10.391	.000
	PovertyWhite	-.059	.006	-1.836	-9.150	.000

a. Dependent Variable: CrimeRate

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	36.255	8.653		4.190	.000
	White	-.463	.080	-.573	-5.770	.000
	Poverty	.802	.232	.343	3.455	.001

Example: Interaction

- What are the changes?
  - Coefficient for White is positive now
  - Negative effect for interaction
- Interpretation: the effect of poverty depends on the level of %white; the effect of %white depends on poverty rate

Hypothesis Tests: the model

- The significance of the overall model
  - For $X_1, X_2 \ldots X_k$,
- Null/Alternative hypotheses :
  - H0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$
  - H1: at least one $\beta_i \neq 0$
- Use a F-test: $F = \frac{MSR}{MSE}$

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4914.461	3	1638.154	92.890	.000[a]
	Residual	828.860	47	17.635		
	Total	5743.322	50			

a. Predictors: (Constant), PovertyWhite, White, Poverty

b. Dependent Variable: CrimeRate

Hypothesis Tests: slopes

- Independent test for all slopes (b) of X_i
- Null/Alternative hypotheses are the same for variables:
  - H0: $\beta_k = 0$
  - H1: $\beta_k \neq 0$
  - Or, one-tailed tests: H1: $\beta_k >= 0$, H1: $\beta_k < 0$
- Hypothesis tests are about the slope controlling for other variables in the model
  - They may disagree with bivariate hypothesis tests

Hypothesis Tests: slopes

- Calculate the test statistic:

$$t_{N-K-1} = \frac{b_i}{s_{b_i}}$$

- Where b is a slope, s_b is a standard error
- i represents the ith independent variable
- K = total number of independent variables
- T-test degrees of freedom depends on how many variables are in the entire equation
- Compare observed t-value to critical t

Hypothesis Tests: slopes

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-34.352	9.330		-3.682	.001
	White	.463	.112	.572	4.121	.000
	Poverty	5.152	.496	2.204	10.391	.000
	PovertyWhite	-.059	.006	-1.836	-9.150	.000

a. Dependent Variable: CrimeRate

Hypothesis Tests

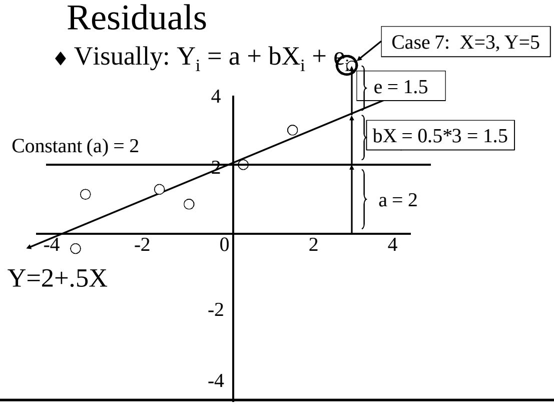
- Why are some variables not significant?
  - Obvious reason: X is not a cause of Y
  - Inadequate sample size
    - Easier to detect relationship in the population with a large sample; collect more data
  - Specification error
    - Nonlinear relationship → scatterplot
    - Nonlinear regression, or transform the variable
  - Restricted variance in X
    - Too little variation in X: descriptive statistics will reveal
    - Collect data with more extreme values
    - E.g. age and income, data on college students only

Predicting Ys

- Regression can be used for explanation and prediction
- Crime=36.255 + 0.801*Poverty – 0.463*White
  - NY: poverty rate 16.4, %white:77.2
  - Predicted: 13.65 vs. observed: 13.3
- Crime=-34.352 + 5.15*Poverty + 0.496*White –0.059*Poverty*White
- Happy=2.5 + 0.05*Income +1.2*Dfemale
  - income in 1000s
- Happy=2 + 0.05*Income + 2*Dfemale - 0.2*Income*Dfemale

Residuals

- ♦ Visually: $Y_i = a + bX_i + e_i$



Residuals

$$R_i = Y_i - \hat{Y}_i$$

- ♦ Prediction is rarely perfect. But residual should follow a certain pattern
- ♦ Analysis of residuals can help us to detect the violation of certain assumptions.

Model Selection

- ♦ Use theory to guide your model building
- ♦ It is common to conduct a series of multiple regressions. Q: Do the new variables substantially improve the model?
- ♦ Idea #1: Look for increase in Adjusted R-Square
- ♦ Idea #2: Conduct a F-test
  - – Recall that F-tests allow comparisons of variance (e.g., SSbetween to SSwithin)

Model Selection

- ♦ F-tests require "nested models"
  - – Models are the same, except for addition of new variables
  - – You can't compare totally different models this way

$$F_{(K_2-K_1)(N-K_2-1)} = \frac{(R_2^2 - R_1^2)/(K_2 - K_1)}{(1-R_2^2)/(N-K_2-1)}$$

- Tests following Hypotheses:
  - H0: Two models have the same R-square
  - H1: Two models have different R-square
- ♦ A significant F-test indicates the second model (with additional variables) is a significant improvement (in R-square) compared to the first.

Model Selection

- ♦ Idea #3: use model selection methods in statistical software
  - – Backward elimination
    - Start with all variables in the model
    - Delete variable one by one till all variables in the model produce significant F statistics at a defined significance level
    - At each step, the variable showing the smallest contribution to the model is deleted
    - You can define significance level (e.g. 0.5)

Model Selection

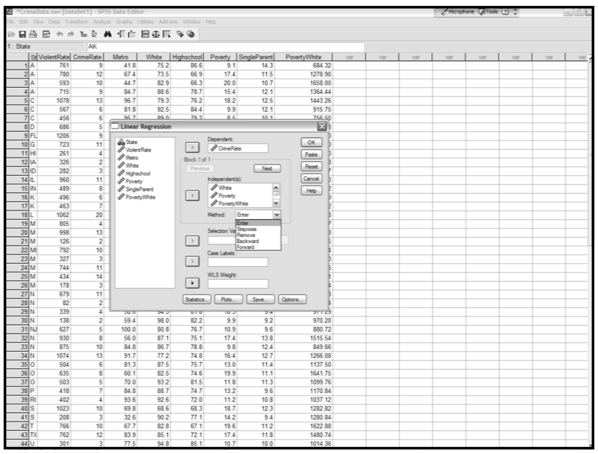
- ♦ Idea #3: use model selection methods in statistical software
  - – Backward elimination
  - – Forward selection
    - Start with no variables in the model
    - Add the variable with the largest F-statistics
    - Recalculate the F-statistics for the rest variables
    - Add the one with the largest F
    - Once a variable is in the model, it stays
    - You can define significance level (e.g. 0.5)

Model Selection

- Idea #3: use model selection methods in statistical software
  - Backward elimination
  - Forward selection
  - Stepwise
    - Combination of forward and backward
    - Variables are added one by one to the model
    - But previous variables can be deleted
    - Selection stops when no variables outside the model is significant at the pre-defined significance level, and every variable in the model is significant

Model Selection

- Idea #3: use model selection methods in statistical software
  - Backward elimination
  - Forward selection
  - Stepwise
  - Limitations:
    - The final model may not be sensible
      - Interaction stays but main effect is out (**Main effect has to be in the model if interaction is used**)
      - Sometimes an important variable may be excluded
- Use theories to guild the model building



Summary

- Multivariate vs. bivariate regression
- Standardized regression coefficients
- Dummy variables
- Interaction terms
- Hypothesis tests
- Predictions, and residuals
- Model selection