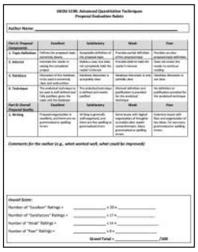
Class Reminder

- Your term project proposal is due on March 8
- Basic points to remember
  1. Define the general topic area for your project
  2. Explain why anyone would be interested in this topic
  3. Discuss the database you will use: source of data, any characteristics of the data you know now, etc.
  4. Give your best assessment of which technique from class will apply to your analysis
  5. Make sure the whole thing is highly readable and well-organized

Class Reminder

- Also review the proposal assessment document from the course website for details on the criteria to be used



GEOG 5190: Advanced Quantitative Techniques
Proposal Evaluation Rubric

Author Name: ______________________________

Part A: Proposal Components	Excellent	Satisfactory	Weak	Poor
1. Topic Definition	Defines the proposed topic extremely clearly	Acceptable definition of the proposed topic	Provides partial definition of the proposed topic	Provides no clear proposed topic definition
2. Interest	Interests the reader in seeing the completed project	Makes a case, but does not completely hold the reader's interest	Provides little to hold the reader's interest	Does not entice the reader to continue reading
3. Database	Discussion of the database to be used is extremely clear and well-written	Database discussion is acceptably clear	Database discussion is only partially clear	Database discussion is not clear
4. Technique	The analytical technique to be used is well defined and fully justified, given the topic and the database	The analytical technique is defined and mostly justified	Minimal definition and justification is provided for the analytical technique	No definition or justification provided for the analytical technique
Part B: Overall Proposal Quality	Excellent	Satisfactory	Weak	Poor
1. Writing	Proposal organization is excellent, and there are no grammatical or spelling errors	Writing is generally well-organized, and there are few spelling or grammatical errors	Some issues with logical organization of thoughts to enable clear reader comprehension. Many grammatical or spelling errors	Extensive issues with flow and organization of key ideas. Far too many grammatical or spelling errors

Comments for the author (e.g., what worked well, what could be improved):

Overall Score:

Number of "Excellent" Ratings = __________ x 20 = __________

Number of "Satisfactory" Ratings = __________ x 17 = __________

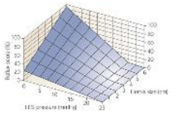
Number of "Weak" Ratings = __________ x 14 = __________

Number of "Poor" Ratings = __________ x 8 = __________

Grand Total = ______________ /100

Week 6

Multiple Correlation & Regression



Multiple Correlation & Regression

- This module is built on the foundation of simple correlation and regression that I'm assuming you have picked up in a previous statistics course
  - **Correlation:** how strong is the relationship between variables? (are the variables related?)
  - **Regression:** what is the precise form of the relationship between variables (prediction of one variable based on another)

Multiple Correlation & Regression

- We will not be exploring content that should have come up in any introduction to simple linear correlation and regression
  - Example: no discussion here of the *data requirements* for use of multiple corr./regression (such as normality, linearity, and homoscedascity)
  - However, much background from simple linear modeling carries over and needs to be considered in multiple correlation/regression applications

Multiple Correlation & Regression

- We will not be exploring content that should have come up in any introduction to simple linear correlation and regression
  - Example: no discussion here of the *data requirements* for use of multiple corr./regression (such as normality, linearity, and homoscedascity)

See my online handout "Dealing with Normality" for some practical tips on what to consider here

Multiple Correlation & Regression

- **Big thing to remember:** multiple correlation and regression involves the straightforward extension of simple correlation/regression concepts
- Multiple correlation/regression uses the same basic ideas as simple correlation/regression, just <u>a level up in complexity</u>

Multiple Correlation & Regression

- For example:

Simple $y = a + bx + \varepsilon$

Multiple $y = a + b_1x_1 + b_2x_2 + \varepsilon$

Where a = intercept/base constant

b = regression coefficient

ε = error term (recognizing the presence of residuals)

Multiple Correlation & Regression

- For example:

Simple $y = a + bx + \varepsilon$

Multiple $y = a + b_1x_1 + b_2x_2 + \varepsilon$

Where a = intercept/base constant

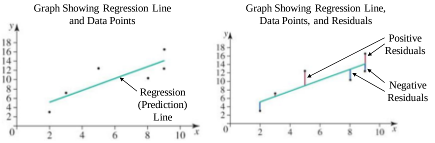
b = regression coefficient

We're not going to deal with this any further here

ε = error term (recognizing the presence of residuals)

Multiple Correlation & Regression

Residual = Observed Value – Predicted Value



ε (error term) on previous slide recognizes that data do not fit the regression line perfectly

Multiple Correlation & Regression

<u>Analysis of residuals</u> can be a powerful means of better understanding the phenomenon you are studying and improving your regression model

We will see good spatial examples of this in our trend surface analysis discussion
(coming in two weeks)

Multiple Correlation & Regression

- <u>Key idea</u>: multiple corr./regression uses more than one independent variable to better account for complex, real-world situations
- <u>For example</u>: individual income level depends at least in part on (to name a few variables)
  - Education
  - Location
  - Age
  - Social connections …

Multiple Correlation & Regression

- Another example: evaporation depends on
  - Temperature
  - Pressure
  - Wind
  - Sunshine
  - Humidity …

Multiple Correlation & Regression

- Build all of these variables into a single, predictive equation (multiple regression)

$$y = a + b_1x_1 + b_2x_2$$

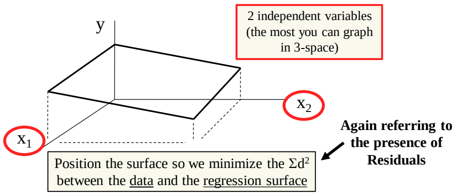
or $$x_1 = a + b_2x_2 + b_3x_3$$

or $$\boxed{x_1 = b_1 + b_2x_2 + b_3x_3}$$ (most common)

- The number of independent variables (right hand side variables) varies with how complex a situation you wish to model

Multiple Correlation & Regression

- With multiple regression, the essential idea is to fit the "surface" modeled by the regression equation to your dataset



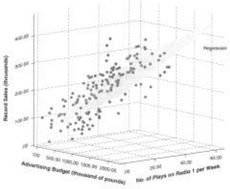
2 independent variables (the most you can graph in 3-space)

Again referring to the presence of Residuals

Position the surface so we minimize the Σd^2 between the data and the regression surface

Multiple Correlation & Regression

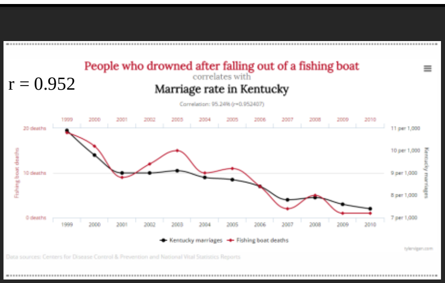
- Here's another representation of a similar situation, with the raw data included:



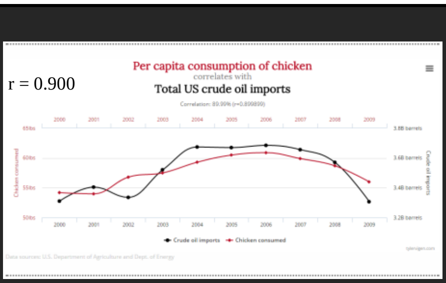
Multiple Correlation

- Let's specifically deal now with the multivariate extension of the correlation concept in particular
  - One reminder: correlation is not the same as causation (so please interpret correlations carefully)

r = 0.952

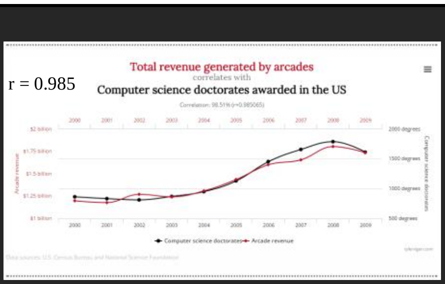


http://www.tylervigen.com/: 42 spurious correlations



r = 0.900

http://www.tylervigen.com/: 42 spurious correlations



r = 0.985

http://www.tylervigen.com/: 42 spurious correlations

Multiple Correlation

- Let’s specifically deal now with the multivariate extension of the correlation concept in particular
  - Because we now have multiple variables, we have more than one correlation value to consider (only have one corr. value with simple correlation)
  - New Concept: partial correlation coefficients

Partial Correlation Coefficients

- A measure of the correlation between the dependent variable and one independent variable
  - Holding constant the influence of all other independent variables
  - Only focusing on one pair of variables at a time: how strong is the interaction between the given two variables?

Partial Correlation Coefficients

- Notation
  - r = partial correlation (general symbol)
  - $r_{12.3}$ = partial correlation between the variables x_1 and x_2, with the variable x_3 held constant

$$x_1 = b_1 + b_2 x_2 + b_3 x_3$$

Partial Correlation Coefficients

- Zero Order Correlations
  - For the simple situation with only one dependent and one independent, we could write r_{12} (same as R or R_{12} with simple linear regression)
- First Order Correlations (2 independents)
  - $r_{12.3}$, $r_{13.2}$
- Second Order Correlations (3 independents)
  - $r_{12.34}$, $r_{13.24}$, $r_{14.23}$

Partial Correlation Coefficients

- Squared Partial Correlation Coefficients
  - For example, $r_{12.3}{}^{2}$
  - Meaning: the proportion of variance in the dependent variable explained by one independent variable when all other variables are held constant

Multiple Correlation Coefficient

- Multiple R
  - Gives the overall explanatory power of the multivariate model
  - Also called the “gross correlation”
    - $R_{1.23}$ (2 independent variables)
    - $R_{1.2345}$ (4 independent variables)
    - We use this notation so we might conceivably drop out of the model those variables that have little effect (such as going from $R_{1.2345}$ to $R_{1.235}$)

Multiple Correlation Coefficient

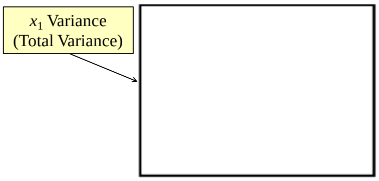
- Interpretation
  - R=1.0: all data points are on the multiple regression surface (perfect relationship)
  - R=0.0: no relationship whatsoever between dependent and independent variables (random)
  - Range of R is from 0 to 1 (note: no negative R)
  - Direction of relationship (slopes) are indicated by the b values only (not indicated by R)

Multiple Correlation Coefficient

- Relationship of partial correlations to the multiple R
  - Key point: $\Sigma r \neq R$
  - Partial correlations account for those proportions of the total variance that are not jointly accounted for
  - Need to add in the proportion of the total variance that is jointly accounted for by variable pairs

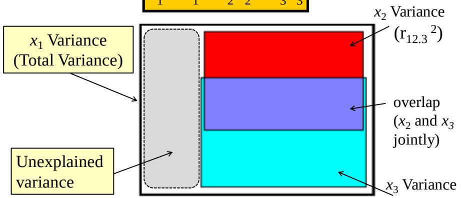
Multiple Correlation Coefficient

$$x_1 = b_1 + b_2x_2 + b_3x_3$$



Multiple Correlation Coefficient

$$x_1 = b_1 + b_2x_2 + b_3x_3$$



$R^2 = x_1$ Variance – Unexplained Variance

Multiple Regression

- Again, think of the basic form of the regression equation
  - $x_1 = b_1 + b_2x_2 + b_3x_3$ (add as many x-values as needed)
- Partial Regression Coefficients: “b values”
  - Also called “net regression coefficients”
  - Meaning: the unit change in x_1 associated with (for example) x_2, with all other x-values remaining constant

Multiple Regression

- **Partial Regression Coefficients: Notation**

$$x_1 = a + b_2x_2 + b_3x_3$$

$$x_1 = b_1 + b_2x_2 + b_3x_3$$

$$x_{1.23} = b_{1.23} + b_{12.3}x_2 + b_{13.2}x_3$$

Multiple Regression

- **β Coefficients**
  - Partial regression coefficients are measured in the units of the variables being used
  - β coefficients are standardized to the same scale
  - Useful when independent variables are in very different units/scales
    - Years of Education (1-20 years), House Size (500-10,000 square feet), Annual Income ($0 to $1 million)

Multiple Regression

- **β Coefficients**
  - Partial regression coefficients are measured in the units of the variables being used
  - β coefficients are standardized to the same scale

$$x_1 = a + \beta_2x_2 + \beta_3x_3$$

$$x_1 = \beta_1 + \beta_2x_2 + \beta_3x_3$$

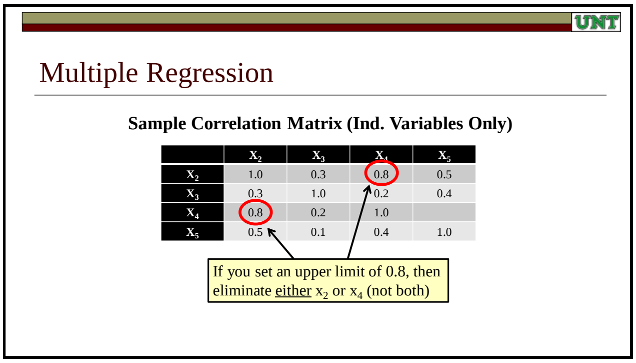
$$x_{1.23} = \beta_{1.23} + \beta_{12.3}x_2 + \beta_{13.2}x_3$$

Multiple Regression

- **Multicollinearity**
  - Definition: a high degree of relationship between independent variables
  - Multicollinearity is an issue with multiple regression: need to eliminate
  - No sense having an x_2 and an x_3 in the regression equation that are themselves highly related
  - Why have this complexity in your equation when it really adds very little to your explanation?

Multiple Regression

- **Approaches to Multicollinearity**
  - Arbitrary cut-offs: set a threshold (say 0.8 or 0.9) for the correlations between any two pairs of *independent* variables (if a correlation is greater than this, eliminate *one* of the independent variables)

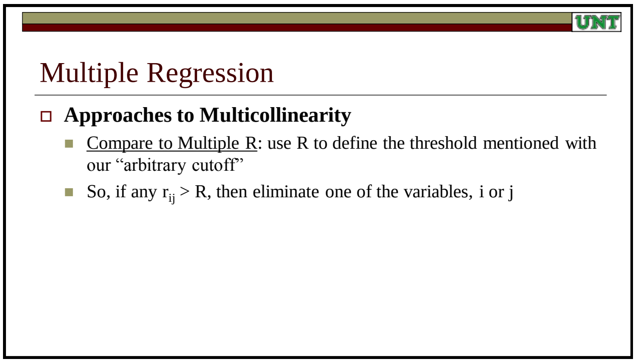


Multiple Regression

Sample Correlation Matrix (Ind. Variables Only)

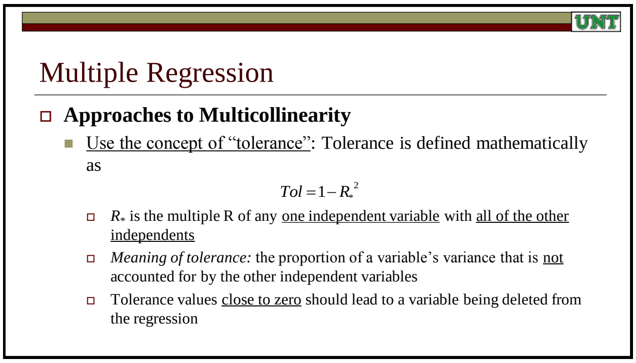
	X_2	X_3	X_4	X_5
X_2	1.0	0.3	0.8	0.5
X_3	0.3	1.0	0.2	0.4
X_4	0.8	0.2	1.0	
X_5	0.5	0.1	0.4	1.0

If you set an upper limit of 0.8, then eliminate either x_2 or x_4 (not both)



Multiple Regression

- **Approaches to Multicollinearity**
  - Compare to Multiple R: use R to define the threshold mentioned with our “arbitrary cutoff”
  - So, if any $r_{ij} > R$, then eliminate one of the variables, i or j

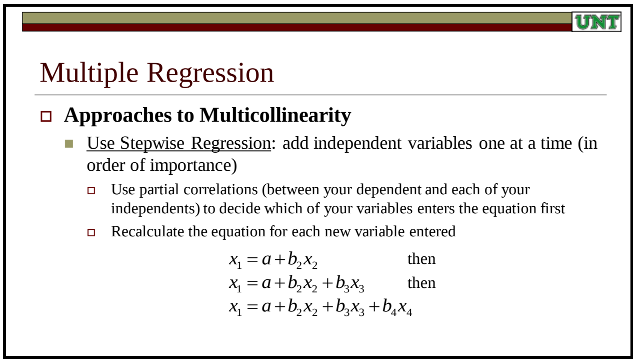


Multiple Regression

- **Approaches to Multicollinearity**
  - Use the concept of “tolerance”: Tolerance is defined mathematically as

$$Tol = 1 - R_*^2$$

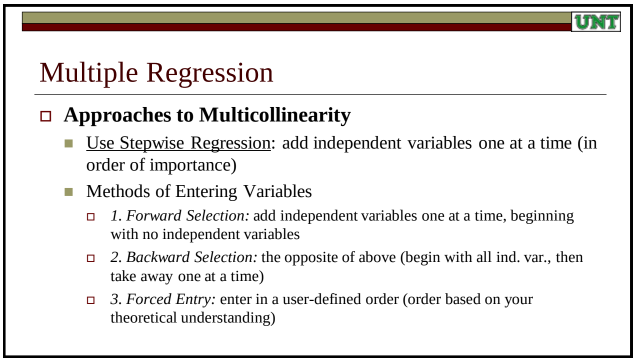
    - R_* is the multiple R of any one independent variable with all of the other independents
    - *Meaning of tolerance:* the proportion of a variable’s variance that is not accounted for by the other independent variables
    - Tolerance values close to zero should lead to a variable being deleted from the regression



Multiple Regression

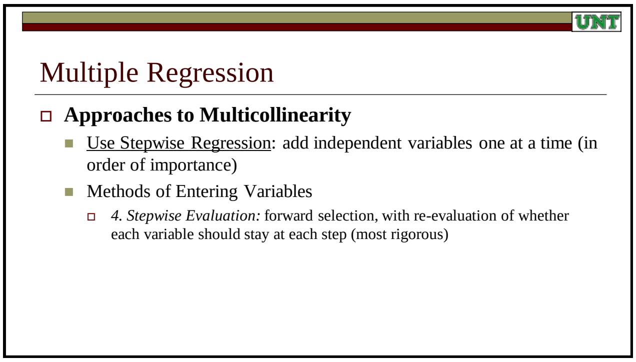
- **Approaches to Multicollinearity**
  - Use Stepwise Regression: add independent variables one at a time (in order of importance)
    - Use partial correlations (between your dependent and each of your independents) to decide which of your variables enters the equation first
    - Recalculate the equation for each new variable entered

$$x_1 = a + b_2x_2 \quad \text{then}$$
$$x_1 = a + b_2x_2 + b_3x_3 \quad \text{then}$$
$$x_1 = a + b_2x_2 + b_3x_3 + b_4x_4$$



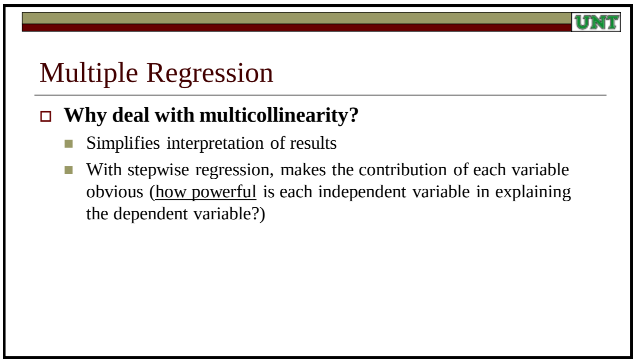
Multiple Regression

- **Approaches to Multicollinearity**
  - Use Stepwise Regression: add independent variables one at a time (in order of importance)
  - Methods of Entering Variables
    - *1. Forward Selection:* add independent variables one at a time, beginning with no independent variables
    - *2. Backward Selection:* the opposite of above (begin with all ind. var., then take away one at a time)
    - *3. Forced Entry:* enter in a user-defined order (order based on your theoretical understanding)



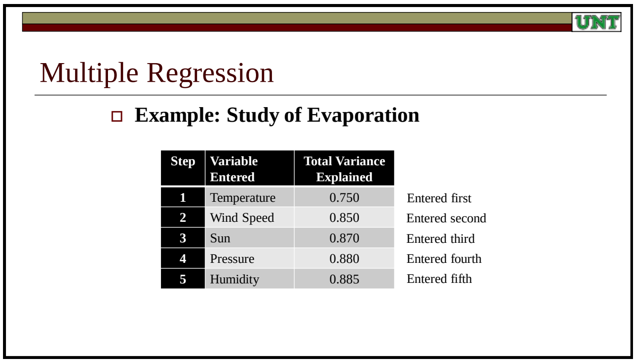
Multiple Regression

- **Approaches to Multicollinearity**
  - Use Stepwise Regression: add independent variables one at a time (in order of importance)
  - Methods of Entering Variables
    - *4. Stepwise Evaluation:* forward selection, with re-evaluation of whether each variable should stay at each step (most rigorous)



Multiple Regression

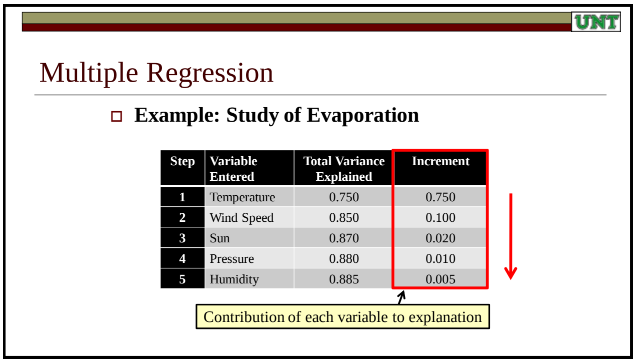
- **Why deal with multicollinearity?**
  - Simplifies interpretation of results
  - With stepwise regression, makes the contribution of each variable obvious (how powerful is each independent variable in explaining the dependent variable?)



Multiple Regression

- **Example: Study of Evaporation**

Step	Variable Entered	Total Variance Explained	
1	Temperature	0.750	Entered first
2	Wind Speed	0.850	Entered second
3	Sun	0.870	Entered third
4	Pressure	0.880	Entered fourth
5	Humidity	0.885	Entered fifth

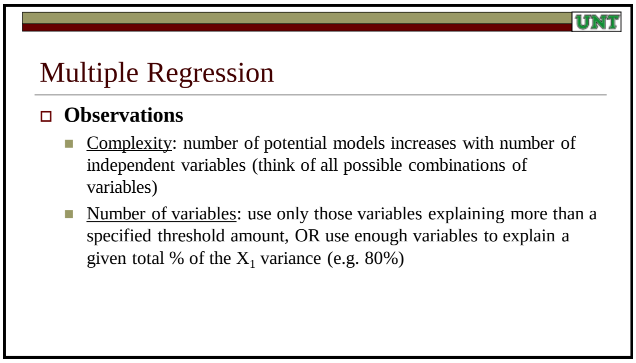


Multiple Regression

- **Example: Study of Evaporation**

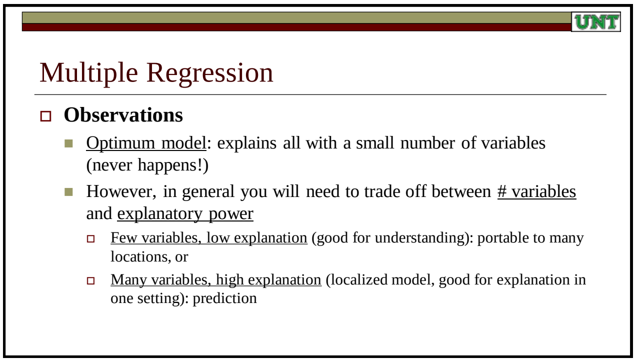
Step	Variable Entered	Total Variance Explained	Increment
1	Temperature	0.750	0.750
2	Wind Speed	0.850	0.100
3	Sun	0.870	0.020
4	Pressure	0.880	0.010
5	Humidity	0.885	0.005

Contribution of each variable to explanation



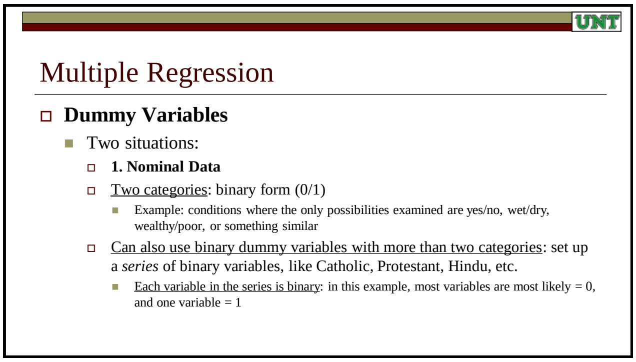
Multiple Regression

- **Observations**
  - Complexity: number of potential models increases with number of independent variables (think of all possible combinations of variables)
  - Number of variables: use only those variables explaining more than a specified threshold amount, OR use enough variables to explain a given total % of the X_1 variance (e.g. 80%)



Multiple Regression

- **Observations**
  - Optimum model: explains all with a small number of variables (never happens!)
  - However, in general you will need to trade off between # variables and explanatory power
    - Few variables, low explanation (good for understanding): portable to many locations, or
    - Many variables, high explanation (localized model, good for explanation in one setting): prediction



Multiple Regression

- **Dummy Variables**
  - Two situations:
    - **1. Nominal Data**
    - Two categories: binary form (0/1)
      - Example: conditions where the only possibilities examined are yes/no, wet/dry, wealthy/poor, or something similar
    - Can also use binary dummy variables with more than two categories: set up a *series* of binary variables, like Catholic, Protestant, Hindu, etc.
      - Each variable in the series is binary: in this example, most variables are most likely = 0, and one variable = 1

Multiple Regression

- **Dummy Variables**
  - Two situations:
    - **2. Ordinal**
    - Variable with values 1, 2, 3, 4, … (some kind of ranking)
    - Example: one study in the journal *Environment and Planning A* used this road classification:
      - 0 = non-motorable road (no vehicle use)
      - 1 = dead-end road
      - 2 = through road
      - 3 = state highway

Multiple Regression

- **Dummy Variables**
  - Two situations:
    - **2. Ordinal**
    - Negative view: Shouldn’t use in a multiple regression since intervals aren’t meaningful; can replace single ordinal variable with a series of nominal variables (think of road example, what would we do?)
    - Positive view: Go ahead and use in multiple regression, as often intervals are close to equal, and results are the same as with the negative view

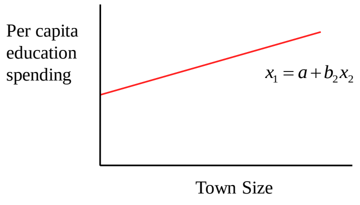
Multiple Regression

- **Other data issues**
  - **Interval/ratio data:** in general, no problem
    - However, some problems with use of percentages
    - Issue: limit the range of the regression using a percentage variable – how meaningful is a regression prediction for a relative humidity of 200%?
  - **Mixing data types:** can mix all four variable types (nominal, ordinal, interval, ratio) in a single equation

Multiple Regression

- **What if you do mix data types in a single regression?**
  - What would that mean? Let’s have a look at an example to clarify the interpretation
  - Example: the basic idea from one classic study of the relationship between *per capita education expenditures* and *town size* in the United Kingdom

Multiple Regression



Multiple Regression

- **Now, introduce a dummy variable into the simple regression equation we started with**
  - New Variable: x_3 = “political orientation”
    - x_3 = 0 if the town tends to vote for one political party
    - x_3 = 1 if the town tends to vote for another party

$$x_1 = \underbrace{a + b_2x_2}_{\text{Original equation}} + \underbrace{b_3x_3}_{\text{New dummy variable}}$$

Multiple Regression

- **What does this mean?**
  - **If $x_3 = 1$:** we are effectively adding an increment on the old equation, with the increment equal to the value of b_3

$x_1 = a + b_2x_2 + b_3x_3$ becomes

$x_1 = a + b_2x_2 + b_3$ since $x_3 = 1$

Multiple Regression

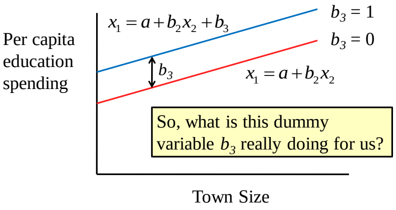
- **What does this mean?**
  - **If $x_3 = 0$:** we are effectively back to the original, old equation, so the effect of the b_3 term simply disappears

$x_1 = a + b_2x_2 + b_3x_3$ but $x_3 = 0$, so

$x_1 = a + b_2x_2$ the original equation

Multiple Regression

- **We see the result of this on our graph**



So, what is this dummy variable b_3 really doing for us?

Multiple Regression

- **Q: then what would a series of dummy variables do for us?**

$$x_1 = a + b_2x_2 + b_3x_3 + b_4x_4$$

x_2 = a usual, ratio variable
x_3 and x_4 = binary dummy variables

Multiple Regression

- **Another analytical issue: dummy variables and perfect multicollinearity**
  - Say you are interested in adding to your multiple regression model of highway accidents the impact of traffic light status: three possibilities to account for
  - You could use three, mutually-exclusive dummy variables to model this situation:
    - x_1 = 0 or 1 (red light)
    - x_2 = 0 or 1 (yellow light)
    - x_3 = 0 or 1 (green light)

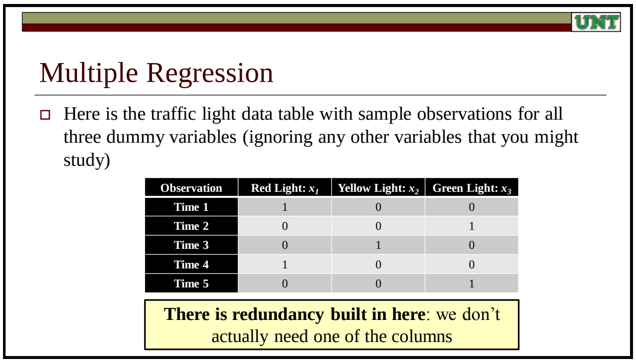
Impossible to have more than one light on at the same time

Multiple Regression

- Here is the traffic light data table with sample observations for all three dummy variables (ignoring any other variables that you might study)

Observation	Red Light: x_1	Yellow Light: x_2	Green Light: x_3
Time 1	1	0	0
Time 2	0	0	1
Time 3	0	1	0
Time 4	1	0	0
Time 5	0	0	1

Q: do we need this entire table to give us all the information we need?

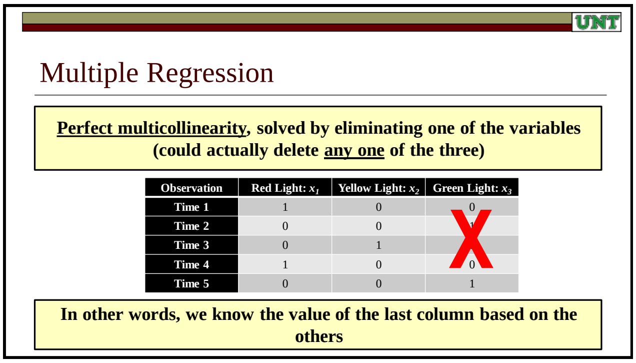


Multiple Regression

- Here is the traffic light data table with sample observations for all three dummy variables (ignoring any other variables that you might study)

Observation	Red Light: x_1	Yellow Light: x_2	Green Light: x_3
Time 1	1	0	0
Time 2	0	0	1
Time 3	0	1	0
Time 4	1	0	0
Time 5	0	0	1

There is redundancy built in here: we don't actually need one of the columns

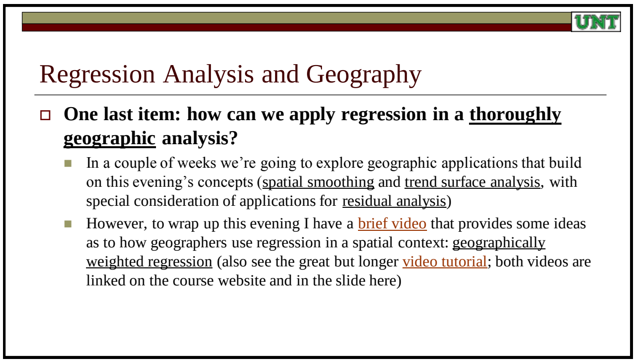


Multiple Regression

<u>Perfect multicollinearity</u>, solved by eliminating one of the variables (could actually delete <u>any one</u> of the three)

Observation	Red Light: x_1	Yellow Light: x_2	Green Light: x_3
Time 1	1	0	0
Time 2	0	0	1
Time 3	0	1	[illegible]
Time 4	1	0	0
Time 5	0	0	1

In other words, we know the value of the last column based on the others



Regression Analysis and Geography

- **One last item: how can we apply regression in a <u>thoroughly geographic</u> analysis?**
  - In a couple of weeks we're going to explore geographic applications that build on this evening's concepts (<u>spatial smoothing</u> and <u>trend surface analysis</u>, with special consideration of applications for <u>residual analysis</u>)
  - However, to wrap up this evening I have a <u>brief video</u> that provides some ideas as to how geographers use regression in a spatial context: <u>geographically weighted regression</u> (also see the great but longer <u>video tutorial</u>; both videos are linked on the course website and in the slide here)