

THE UNIVERSITY OF TEXAS
AT AUSTIN

WHAT STARTS HERE CHANGES THE WORLD

CHE384, From Data to Decisions: Measurement, Uncertainty, Analysis, and Modeling

Lecture 47

Multicollinearity

Chris A. Mack
Adjunct Associate Professor

<http://www.lithoguru.com/scientist/statistics/>

© Chris Mack, 2016 Data to Decisions 1

THE UNIVERSITY OF TEXAS
AT AUSTIN

WHAT STARTS HERE CHANGES THE WORLD

What's New in Multiple Regression

- **Multicollinearity**: often predictor variables are correlated with each other – they are not independent
 - Also called confounding
- Example: what body measures predict strength in a certain fitness test?
 - Height is correlated with strength
 - Weight is correlated with strength
 - But height and weight are correlated with each other! Are they really two different predictors?

© Chris Mack, 2016 Data to Decisions 2

THE UNIVERSITY OF TEXAS
AT AUSTIN

WHAT STARTS HERE CHANGES THE WORLD

Multicollinearity Example

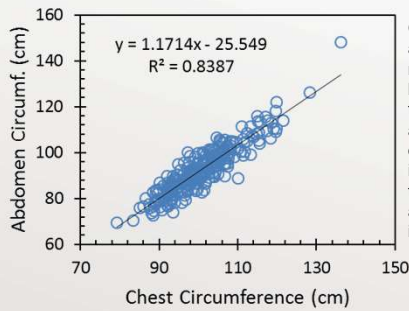
- Measuring % Body Fat is hard
 - Measure body density using full water immersion to calculate density (Siri's equation relates measured density to % Body Fat)
- Body Fat Model: The goal is to create a model of % Body Fat using easily obtainable measures
 - Height, Weight, Chest Circumference, Abdomen Circumference, Thigh Circumference, etc.

© Chris Mack, 2016 Data to Decisions 3

THE UNIVERSITY OF TEXAS
AT AUSTIN

WHAT STARTS HERE CHANGES THE WORLD

These Measures are Correlated



Abdomen Circumf. (cm)

$y = 1.1714x - 25.549$
 $R^2 = 0.8387$

Chest Circumference (cm)

Chest and Abdomen are both inputs to the model, but one can be used to predict the other with high accuracy (i.e., if one of these variables is in the model, adding the other does not add much new information)

© Chris Mack, 2016 Data to Decisions 4

THE UNIVERSITY OF TEXAS
AT AUSTIN

WHAT STARTS HERE CHANGES THE WORLD

The Correlation Matrix

- A common output of a multiple regression is the correlation matrix – the correlation coefficient for each pair of variables (response as well as predictors)
 - Only reveals pair-wise correlations

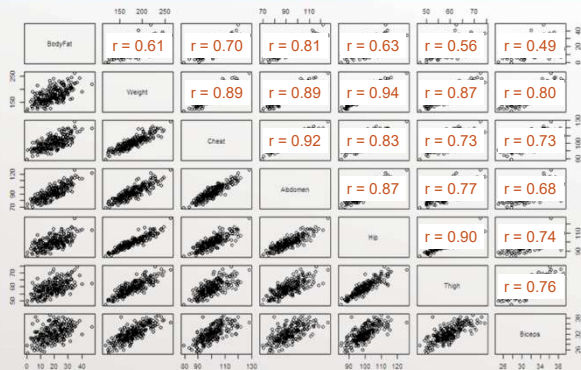
	Y	X ₁	X ₂	X ₃
Y	1.00	.945	.836	.701
X ₁		1.00	.781	.499
X ₂			1.00	.632
X ₃				1.00

We like predictor variables with high correlation to the response, but low correlations to other predictor variables

© Chris Mack, 2016 Data to Decisions 5

THE UNIVERSITY OF TEXAS
AT AUSTIN

WHAT STARTS HERE CHANGES THE WORLD



© Chris Mack, 2016 Data to Decisions 6

THE UNIVERSITY OF TEXAS AT AUSTIN WHAT STARTS HERE CHANGES THE WORLD

In the Extreme...

- For perfect multicollinearity (e.g., $r_{12} = \pm 1$) the design matrix is **singular** and therefore cannot be inverted
 - The OLS estimate does not exist
- To obtain a solution, one of the two perfectly correlated predictors must be removed from the model
- For correlation coefficients close to ± 1 , OLS math can produce large round-off errors

© Chris Mack, 2016 Data to Decisions 7

THE UNIVERSITY OF TEXAS AT AUSTIN WHAT STARTS HERE CHANGES THE WORLD

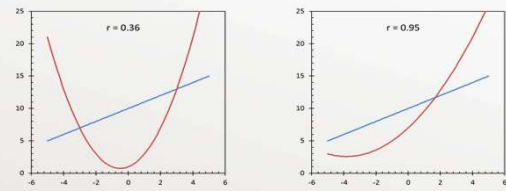
Orthogonal Parameters

- When two predictor variables are completely uncorrelated ($r_{12} = 0$), then we say those two are **orthogonal**
 - Adding or removing one of the two variables from the model does not affect the best-fit value of the other's coefficient (nor it's standard error)
 - This is ideally what we want, but we rarely come close
- We often try to design our experiment to minimize correlations between predictor variables

© Chris Mack, 2016 Data to Decisions 8

THE UNIVERSITY OF TEXAS AT AUSTIN WHAT STARTS HERE CHANGES THE WORLD

Are x and x^2 Correlated?



Correlation between x and x^2 over the range from a to $a + \Delta$:

$$r_{12} = \frac{(a + \frac{\Delta}{2})}{\sqrt{a^2 + a\Delta + \frac{4}{15}\Delta^2}}$$

If $a = 0$, $r_{12} = \sqrt{\frac{15}{16}} \approx 0.968$

Note: A standardized variable \tilde{x} and \tilde{x}^2 will always be uncorrelated.

© Chris Mack, 2016 Data to Decisions 9

THE UNIVERSITY OF TEXAS AT AUSTIN WHAT STARTS HERE CHANGES THE WORLD

Impact of Correlated Parameters

- When two or more predictor variables are highly correlated we can get:
 - Good fits, small $SE(Y)$, and good predictions (so long as the correlations remain constant)
 - Nonintuitive, biased values for the β_k
 - Large $SE(\beta_k)$ and large confidence intervals for the coefficients
- We can't interpret the β_k as a "marginal slope": holding all other predictor variables constant, what is the slope?
 - For highly correlated variables, we can't hold all but one variable constant!

© Chris Mack, 2016 Data to Decisions 10

THE UNIVERSITY OF TEXAS AT AUSTIN WHAT STARTS HERE CHANGES THE WORLD

Impact of Correlated Parameters

- When building a model with two or more highly correlated predictor variables:
 - Adding or removing a predictor causes large changes in the coefficients of correlated predictors
 - Different data sets produce models with very different coefficient values
 - A statistically significant model (passes the F-test) may have all statistically non-significant coefficients (each individually fails the t-test)

© Chris Mack, 2016 Data to Decisions 11

THE UNIVERSITY OF TEXAS AT AUSTIN WHAT STARTS HERE CHANGES THE WORLD

Lecture 47: What have we learned?

- What is multicollinearity?
- What happens to OLS if two predictor variables have perfect correlation?
- What is the opposite of perfect correlation between predictor variables?
- What is a correlation matrix and how is it used?
- What happens to models and predictions when multicollinearity exists?

© Chris Mack, 2016 Data to Decisions 12