Statistics for CSAI II

7 — Multiple Regression

Travis J. Wiltshire, Ph.D. |

TILBURG UNIVERSITY

Modules

1. Introduction and Probability
2. Sampling Theory
3. Revisiting Hypothesis Testing & Intro to Correlation
4. Correlation
5. Intro to Regression
6. More Regression Centering and Checking Assumptions
7. *Multiple Regression and Assumptions*
8. Interactions
9. Multiple Regression with Categories
10. Multiple Regression with Polynomials
11. Mixed Models
12. Growth Curve Analysis

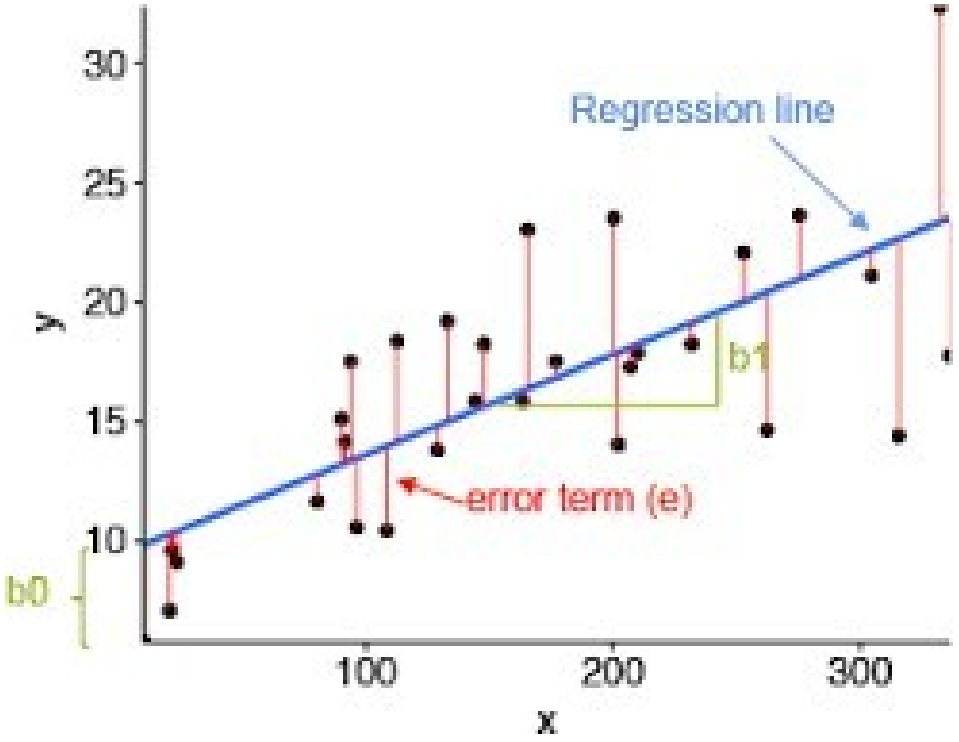
Outline

1. Simple vs Multiple Regression
2. Multiple Regression
3. Beta coefficients (standardized)
4. Comparing models
5. Entry methods

Simple Regression

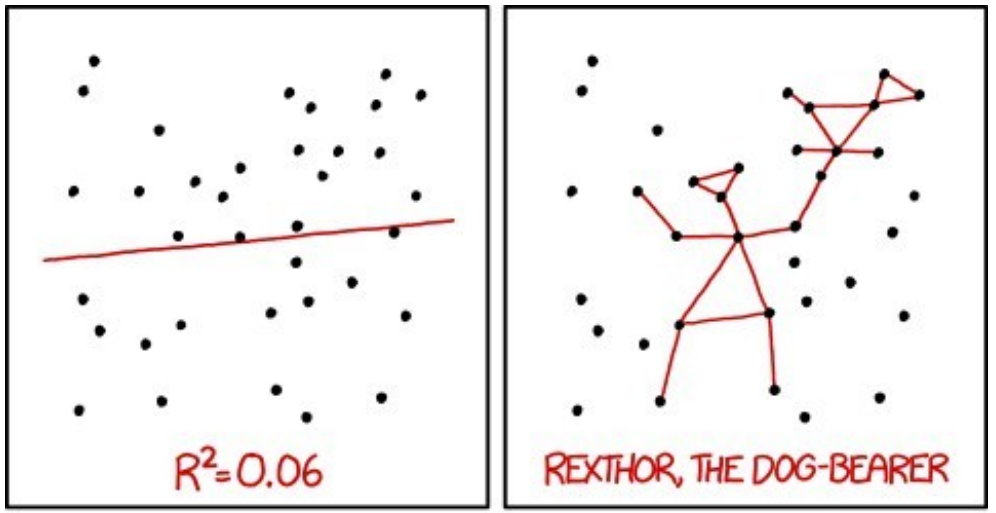
$$Y_i = b_0 + b_1 X_i + \varepsilon_i$$

- b_1
  - Regression coefficient for the predictor
  - Gradient (**slope**) of the regression line
  - Direction/strength of relationship
- b_0
  - **Intercept** (value of Y when $X = 0$)
  - Point at which the regression line crosses the Y-axis (ordinate)



Multiple Regression Analysis (MRA)

- Method for studying the relationship between a dependent variable and **two or more** independent variables.
- An equation-based model of the data: we seek to find the **linear combination of predictors** that correlate maximally with the outcome variable
- Purposes:
  - Prediction
  - Explanation
  - Theory building



Simple vs. Multiple Regression

$$Y_i = b_0 + b_1 X_i + \varepsilon_i$$

- One dependent variable Y predicted from one independent variable X
- One regression coefficient
- **R^2**: proportion of variation in dependent variable Y predictable from X

$$Y_i = b_0 + b_1 X1_i + b_2 X2_i + \ldots + b_k Xk_i + \varepsilon_i$$

- One dependent variable Y predicted from **a set of** independent variables (X1, X2 ….Xk)
- One regression coefficient for each independent variable
- **R^2**: proportion of variation in dependent variable Y predictable by **set of** independent variables (X's)

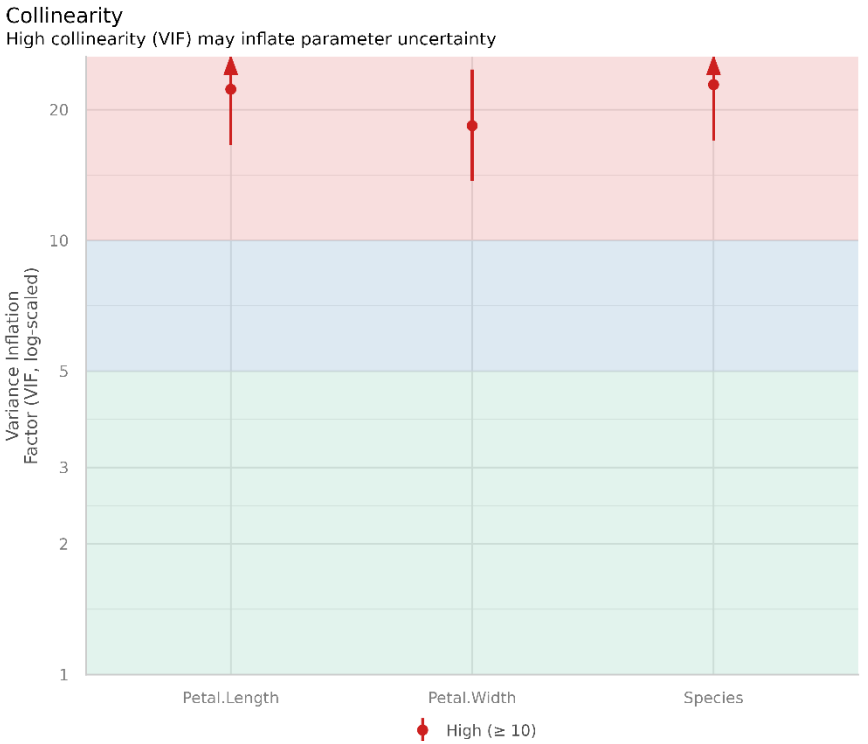
A more in depth comparison here

Assumptions

- **Linearity**: In the population, the relation between the dependent variable and each independent variable is linear when all the other independent variables are held constant.
- **Absence of multicollinearity:** Two predictors should not be too highly correlated (.8 or .9).
- **Homoscedasticity**: in the population, the variances of the dependent variable for each of the possible combinations of the levels of the X variables are equal.
- **Normality**: in the population, the scores on the dependent variable are normally distributed for each of the possible combinations of the level of the X variables; each of the variables is normally distributed
- **Independence**: the scores of any particular subject are independent of the scores of all other subjects

Multicollinearity

- VIF – Variance Inflation Factor
  - low = less than 5
  - moderate = between 5 and 10
  - high collinearity =larger than 10
- If high:
  - Remove or combine correlated predictors
  - Center or standardize variables (reduces numerical instability)
  - Stepwise/Regularization methods (e.g., Ridge regression)

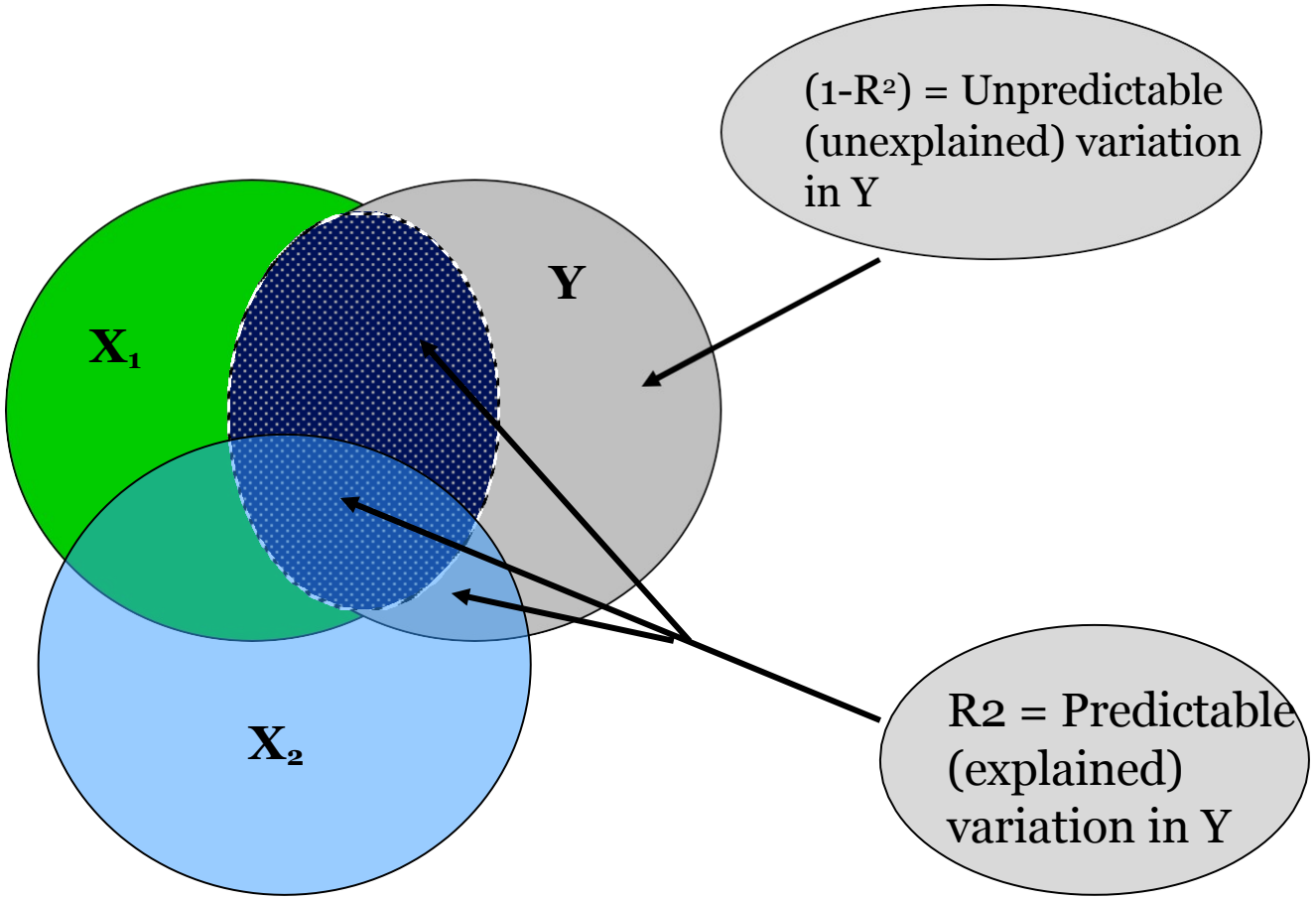


Example: Self Concept and Academic Achievement (N=103)

Statistic	Self-Concept: General (GSC)	Self-Concept: Academic (ASC)	Academic Achievement (AA)	Grade Point Average (GPA)
Correlation				
GSC	1.00			
ASC	.45	1.00		
AA	.15	.40	1.00	
GPA	.25	.50	.62	1.00
Mean	5.20	5.60	54.30	2.50
Standard Deviation	.92	1.26	10.65	.50

Proportion of Predictable and Unpredictable Variation

Where:
Y= AA
X_1 = ASC
X_2 =GSC



Example: The Model

$Y_i = b_0 + b_1X1_i + b_2X2_i + \ldots + b_nXn_i + \varepsilon_i$

- The b's are called **partial regression coefficients**

- **Our example - Predicting AA:**
  - $Y_i = b_0 + b_1 X_{ASC} + b_2 X_{GSC}$

    $= 49.72 + (1.48)X_{ASC} + (-.63)X_{GSC}$
- Predicted AA for person with GSC of 4 and ASC of 6
  - $Y_i = 49.72 + (1.48)(6) + (-.63)(4)$

    $= 56.05$

Various Significance Tests

- **Full model:** Testing R^2
  - Test R^2 through an F test [R output]
  - Test of competing models (difference between R^2) through an F test of difference of R^2s

- **Individual predictors:** Testing b coefficients
  - Test of each partial regression coefficient (b) by t-tests [R output]
  - Comparison of partial regression coefficients with each other: t-test of difference between **standardized** partial regression coefficients ( )

Comparing Partial Regression Coefficients

- Which is the stronger predictor?
  - Predicting AA with a third predictor: $Y_{AA} = b_0 + b_1X_{ASC} + b_2X_{GSC}$ **$+ b_3X_{GPA}$**
  - Comparing partial regression coefficients b_1, b_2 and b_3 directly is not possible **-> why?**

- Solution: **standardize** partial regression coefficients (⇒ beta weights β_i)
  - Same as using standardized dependent and independent variables
  - On same scale so we can compare which is the strongest predictor

- Beta weights (β's) can also be tested for significance with t tests.

Standardized vs unstandardized regression coefficients

- Unstandardized coefficients
  - b coefficient is the amount by which our dependent variable changes if we change the independent variable by one unit *while keeping other independent variables constant*.
- Standardized coefficients
  - Measured in units of standard deviation. A beta value of 1.25 indicates that a change of one standard deviation in the independent variable results in a 1.25 standard deviations increase in the dependent variable *while keeping other independent variables constant*.
  - Allows for comparing the contribution of multiple independent variables, as they are expressed on the same scale.
  - Bonus: more meaningful interpretation of intercept: value of Y when all predictors are at their mean value (which is 0 in the case of scaled predictors)

Standardizing in R

Standardizing using a function:

```
mdl1 <- lm(aa ~ asc + gsc + gpa, data=selfconcept)
#install.packages("effectsize")
require("effectsize")
effectsize(mdl1)
```

Standardize in our model (not too common):

```
mdl2<-lm(scale(aa) ~ scale(asc) + scale(gsc) + scale(gpa), data=selfconcept)
summary(mdl2)
```

Multiple Regression in R

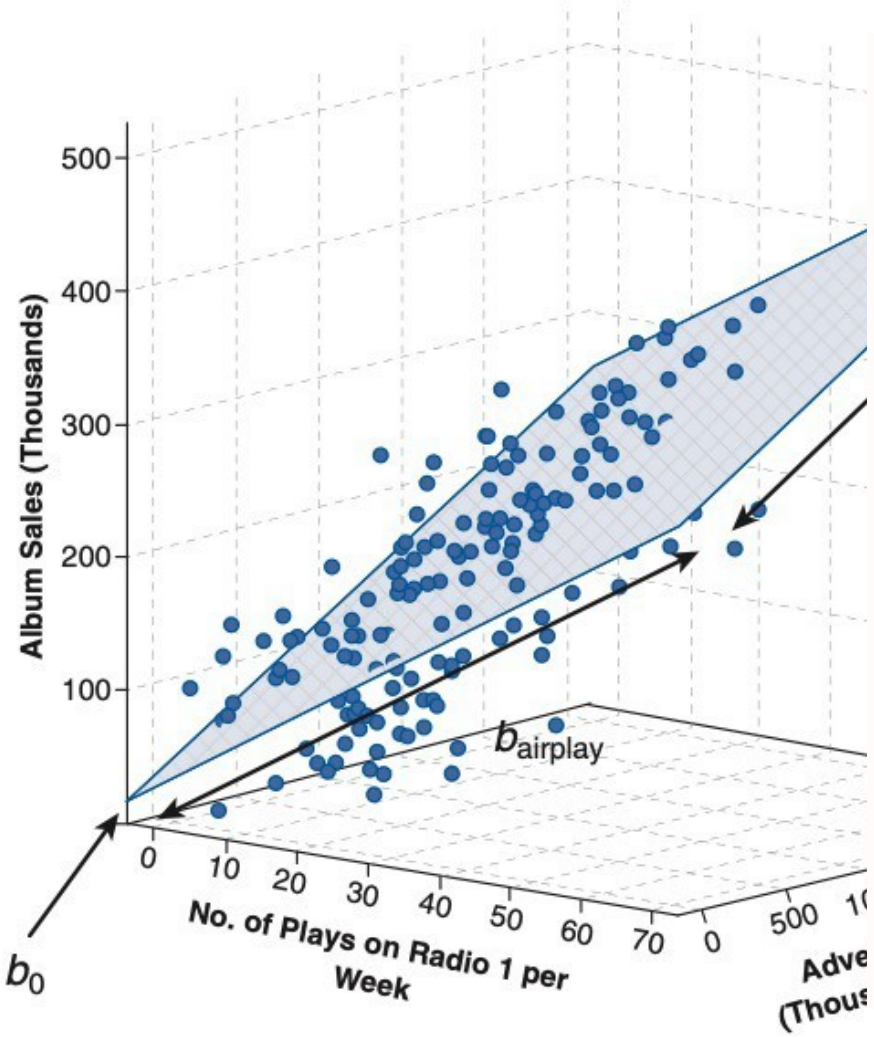
- Multiple linear regression:
  - Still use lm() but we just add more variables into the equation (using +)
- Confidence intervals on estimates
  - confint()
- Standardized Beta coefficients
  - effectsize() from effectsize package (include standardizes CIs)
  - lm.beta() from QuantPsyc package

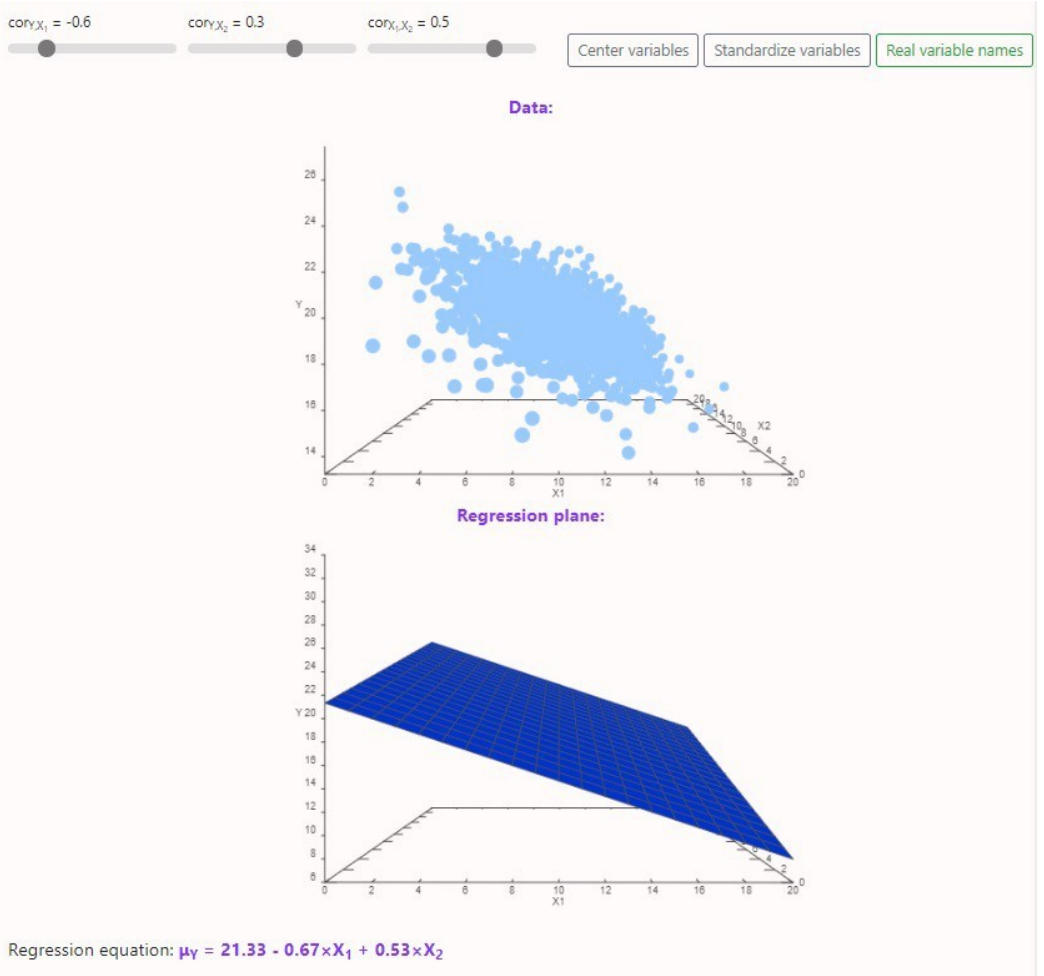
Run a regression on Album sales 2 data

- Load the Album Sales 2.dat file into R.
- Generate descriptive statistics for the data and interpret them
- Make (and save) a simple linear model predicting albums sales from advertising
  - Interpret the output
- Make (and save) a multiple regression model predicting album sales from advertising and airplay
- Generate beta estimates and confidence intervals for your b/beta estimates
  - Interpret the results
  - Which of the predictors has a stronger relationship with the outcome?

Visualizing the Regression Plane

FIGURE 7.8
Scatterplot of the relationship between album sales, advertising budget and radio play





- Check out this link

Comparing model fit

We often want to know if the model with more variables is better, because parsimonious modeling is preferred

- Using ANOVA (compare model 1 to model 2)
- Comparing AIC the **Akaike Information Criterion** using `extractAIC()`
  - Smaller AIC values are better
- Root mean square error using `performance()` from performance package
  - Square root of the variance of the residuals (an
  - Useful for models with aim of prediction
  - Smaller is better

$$\mathrm{AIC} = n \cdot \ln\left(\frac{\mathrm{RSS}}{n}\right) + 2k \cdot$$

- n = number of observations,
- RSS = residual sum of squares,
- k = number of estimated parameters (including intercept).

Details of AIC here

Compare the models

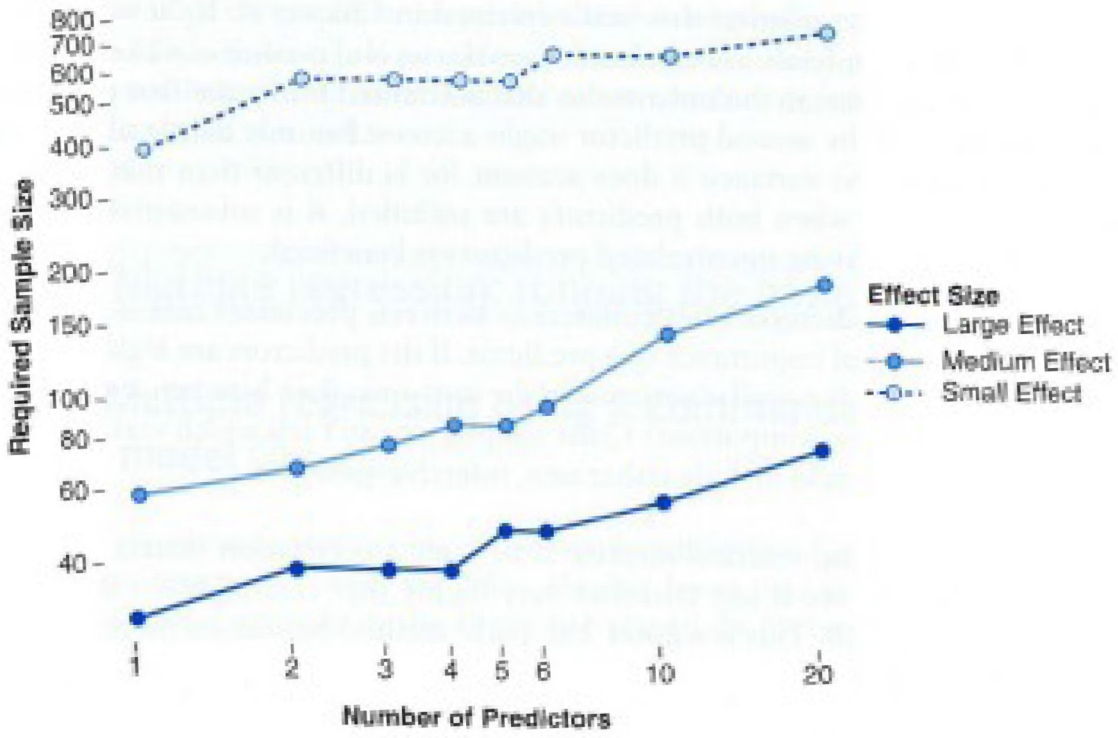
- Compare your two models and see if the second one improved the fit using the anova() function
  - F test here tells you if there is a difference in fit of the data by the two models
- Check out the AIC using extractAIC() or performance()
- Or try plot(compare_performance(mod1, mod2)) from performance package
- Now do a quick global assessment of model assumptions using gvlma() and check_model()

Different Ways of Building Regression Models

- **Hierarchical**: independent variables entered in stages (based on theory)
- **Simultaneous Forced Entry**: all independent variables entered together
- **Stepwise**: independent variables entered according to some order
  - By size or correlation with dependent variable
  - In order of significance
- **All-subsets methods**: try every combination of variables to see which one gives the best fit (number of models will grow exponentially with number of predictors!)

Different (data mining) Entry Methods in R

- We've been using hierarchical method
- Can do stepwise with stepAIC() function from MASS package
- 'backward' – puts all variables in model and sees if AIC is lowered by removing variables
- 'forward' – enters variables one at a time and tests if the model is improved



Issues with Stepwise Entry Methods

- Useful for
  - Exploration of large datasets without existing theory
  - Quick and computationally cheap model selection
- Problems such as
  - Overfitting
  - Lacking theory
  - Biased estimates
  - Inflated error
- Some better alternatives:
  - Regularization such as
    - LASSO/RIDGE/ELASTIC NET REGRESSSION

Why there are problems with this method and some alternatives here

Try out the stepwise model

- First create (and save) a new model that has all predictors in the data
- Then save a new model using the stepAIC() from the MASS package on that model and specify the direction
- Use your-model-name-here$anova to get the results

Evaluate and compare the final model to the last model

- Look at the summary of this stepwise model with all variables and interpret them
- Generate standardized beta estimates and CIs
- Compare this last model that includes all the variables to the one with only two of the predictors
- Do check of the assumptions of the model both visually and with gvlma

Preparing for Module 8: Interactions

1. Read A primer on Interaction Effects in Multiple Linear Regression https://quantpsy.org/interact/interactions.htm

Thanks! See you next time! Questions?

Material from Fields et al Discovering Statistics with R.