Exploratory Data Analysis II

Therri Usher

Friday, February 27, 2015

Review: Data Processing

I looked up physical activity measures in the NHATS data. I found the following:

> *In the last month, did {you/SP} ever go walking for exercise? (pa1evrgowalk)*
>
> *In the last month, did {you/SP} ever spend time on vigorous activities that increased {your/his/her} heart rate and made {you/him/her} breathe harder? This includes things like working out, swimming, running or biking, or playing a sport. (pa1vigoractv)*

How can I use these measures to quantify physical activity?

Review: Data Processing

I can create a new indicator based on the two variables.

```r
# Load the necessary package for opening the data
library(foreign)

# Load the data set
dat <- read.dta("C:/Users/owner/Dropbox/Ph.D/Health Dispari

# Save the number of observations
n.obs <- dim(dat)[1]

# Create a new physical activity variable
phys.act <- matrix(nrow=n.obs, ncol=1)
```

Review: Data Processing

```
# Define the new physical activity variable
for(i in 1:n.obs)
{
  # Make sure that inapplicables, refused, etc. are NA
  if((dat$pa1evrgowalk[i]!=1 & dat$pa1evrgowalk[i]!=2)
     | (dat$pa1vigoractv[i]!=1 & dat$pa1vigoractv[i]!=2))
  {
    phys.act[i] <- NA
  } else if(dat$pa1evrgowalk[i]==1
            | dat$pa1vigoractv[i]==1)
  {
    phys.act[i] <- 1
  } else
  {
    phys.act[i] <- 0
  }
}
```

Review: Data Processing

Advantage: I have utilized the data that I have into an outcome variable.

Disadvantage: My outcome variable is not a comprehensive measure of physical activity but a self-report of whether the subject walked for exercise or spent time on vigorous activity in the past month.

Disclaimer

This lecture blurs the lines between exploratory data analysis and model fitting, requiring some knowledge and expertise about fitting regression models in particular.

Statistical Reasoning

Exploratory Data Analysis

One of the purposes of EDA is to help us decide which variables should be in a model.

We do this by assessing how they compare to the outcome variable and other predictor variables, either graphically or non-graphically.

Parsimony

Our goal is to find the best model that addresses our research question. But what if there are multiple best models?

We should aspire not to obtain just the "best" model, but the "best" model that is the simplest, i.e. the fewest parameters.

The principle of parsimony is that the simplest explanation that best explains is preferred[1]. Thereore, a parsimonious model is one that accomplishes the necessary goals, explanation or prediction, with the fewest predictor variables.

Keep in mind that if you include too many variables, your estimates are still unbiased and consistent but the estimated standard errors are inflated.

[1] Occam's razor. `http://en.wikipedia.org/wiki/Occam%27s_razor`

Model Specification

At the same time, the model must include all relevant variables, where relevant means that the variable is truly associated with the outcome variable. In other words, the model must be correctly specified.

If a relevant variable has been omitted, then the estimates of the model coefficients and other parameters are biased.

Assessing Model Misspecification

- Plots
  - Y versus $\hat{Y}$
  - $Y - \hat{Y}$ versus X for the different predictor variables
- Ramsey RESET Test
  1. Fit the planned model to obtain $\hat{Y}$
  2. Add $\hat{Y}^2$, $\hat{Y}^3$, $\ldots$, $\hat{Y}^k$ to the model
  3. Estimate the new model
  4. If at least one of the $\hat{Y}$ terms is significantly different from 0, there is evidence of model misspecification.

I do NOT recommend the Ramsey test. I would stick to looking at plots and relying on knowledge.

Collinearity

Collinearity (also called multicollinearity) occurs when two or more predictor variables are highly correlated.

This means that one variable can be written as a linear function of another with some error. If they are perfectly linear, then you **cannot** fit a regression model.

Remember that a regression model seeks to estimate independent effects of each predictor variable. If two variables are highly correlated, then the regression model cannot determine which variable is associated with the outcome and it cannot determine which variable is explaining the variation in the data.

As a result, collinearity can cause much higher standard errors and unexpected changes in coefficient estimates or signs.

Assessing Collinearity

There is no known threshold for collinearity and no one perfect method for handling it.

Assessment tools include:

- Correlation matrices
- Running a regression model with one variable at a time then a model with both correlated variables

Assessing Collinearity

- Variance Inflation Factor
  1. For each variable, run a regression model with the variable as the outcome and all other predictor variables as the covariates: $X_j = X_1\beta_1 + ... + X_{j-1}\beta_{j-1} + X_{j+1}\beta_{j+1} + X_p\beta_p + \epsilon$.
  2. Obtain the R^2 value from the model
  3. Calculate $VIF = \frac{1}{1-R_j^2}$
  4. If $VIF_j > 1$, then there is evidence of collinearity.

  - Interpretation: If VIP=2, then the standard error of that coefficient is 2 times as large as it would be if it was uncorrelated with other predictor variables.
  - This only works for linear regression.
  - This is not a statistical test.

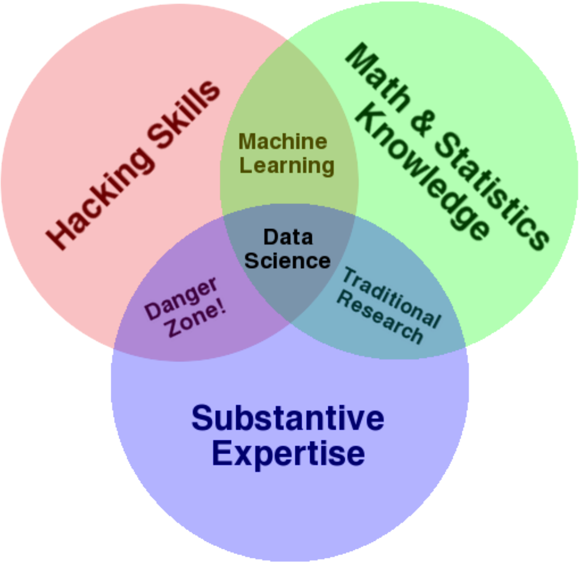
Addressing Collinearity

You can do the following things to try to fix collinearity:

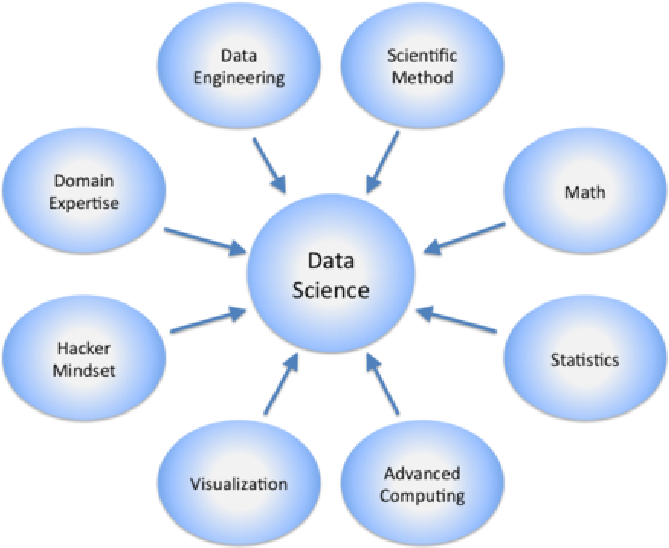
1. Keep all the variables.
   - You will lose statistical power but your estimates are still unbiased.
2. Throw out all but one of the correlated variables.
   - There's a chance that the model will become misspecified.
3. Create a composite score - a single measure that combines the correlated variables.
   - How do you create the composite score?

Scientific Knowledge

Data Science



Data Science

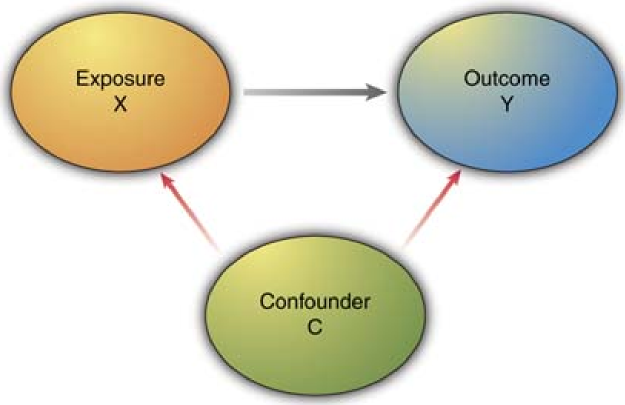


Confounding

One of the purposes of regression models is to find associations between predictors and the outcome independently of other predictors in the model. Therefore, it is a key tool for protecting against confounding.

Confounding

There is actually debte about what a confounder truly is. For this purpose, a confounder can be thought of as a variable that is correlated with both the independent variable and the outcome. (Wikipedia)



Confounding

Commonly, confounders are either associated with a predictor variable or a cause of the predictor variable and are commonly a cause of the outcome.

Confounders can lead to spurious relationships, where the relationship between a predictor and outcome is incorrectly estimated due to not accounting for the confounder. An example would be alcohol confounding the relationship between smoking and lung cancer.

The problem is confounding relies on causality, which is hard to prove. Therefore, it is hard to prove confounding. However, we are often aware of variables that could be *potential* confounders.

Addressing Confounding

These are the classic ways of addressing confounding:

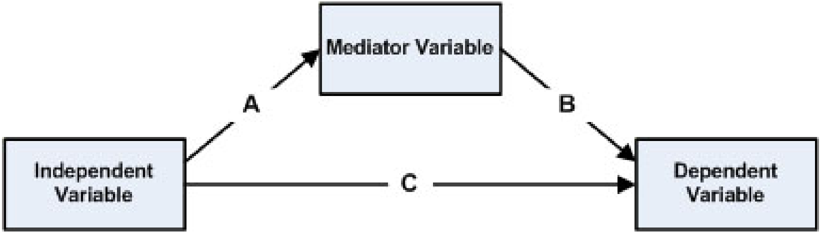
1. Stratification
   - Run the same model on different levels of the confounder.
   - Not good for continuous confounders.
2. Randomization
   - If you are using an observational study or survey, then this is not feasible.
3. Regression
   - Does not always take care of ALL the confounding.
4. Causal inference tools
   - Good for observational data but requires a lot of assumptions.

Mediation

A mediator is a variable that is caused by the predictor variable, which then causes the outcome. (Wikipedia)

An example of a mediator is diet between income and health statuses.

As a result, mediators create indirect effects between a predictor variable and the outcome. The indirect effects plus the direct effect equals the total effect.



The Problem

Regression cannot tell the difference between confounding and regression.

If you adjust for a confounder, then you help eliminate some of the confounding. If you adjust for a mediator, you lose some of the association between the predictor and the outcome. Therefore, the estimate of the association is smaller (or larger) than it should be.

However, it is not so simple to tell a confounder from a mediator. Oftentimes, you have to rely on a hypothesized framework.

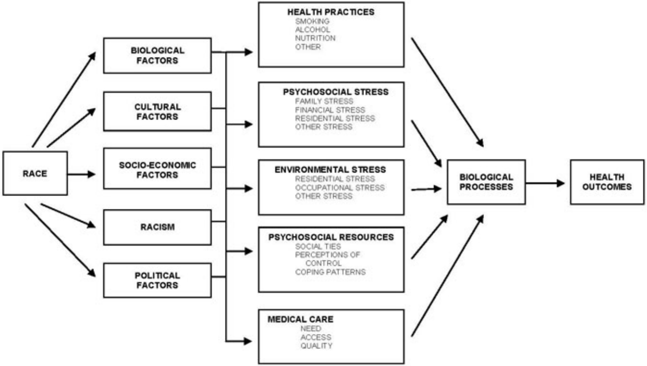
Conceptual Framework

A conceptual framework is an organized background of ideas regarding how variables of interest relate to each other.

Conceptual frameworks are based off of known knowledge but can also contain hypotheses that you believe are true. It is important that you indicate which relations have been proven and which are hypothesized.

Conceptual frameworks helps to visualize the background knowledge of a research problem and to decide what variables to adjust for.

Conceptual Framework



RACE

BIOLOGICAL FACTORS

CULTURAL FACTORS

SOCIO-ECONOMIC FACTORS

RACISM

POLITICAL FACTORS

HEALTH PRACTICES
SMOKING
ALCOHOL
NUTRITION
OTHER

PSYCHOSOCIAL STRESS
FAMILY STRESS
FINANCIAL STRESS
RESIDENTIAL STRESS
OTHER STRESS

ENVIRONMENTAL STRESS
RESIDENTIAL STRESS
OCCUPATIONAL STRESS
OTHER STRESS

PSYCHOSOCIAL RESOURCES
SOCIAL TIES
PERCEPTIONS OF CONTROL
COPING PATTERNS

MEDICAL CARE
NEED
ACCESS
QUALITY

BIOLOGICAL PROCESSES

HEALTH OUTCOMES

Goals of the Model

How a model is built depends on the goal for the model.

1. Prediction
   - All that matters is how well the model predicts the outcome.
2. Inference
   - Not only do you care about the estimates but also their standard errors.
   - Confounding becomes a major concern.
   - Statistical power is also a major concern.

Variable Selection

There are various methods for selecting variables to put into a model. Some of the most common are:

- Prior knowledge
- Change in estimates
- Stepwise selection
  - Backward elimination
  - Forward selection

NOTE: I DO NOT recommend leaving in only significant coefficients.

Prior Knowledge

Variable selection is entirely determined by scientific knowledge and hypotheses.

Think about:

- What variables confound the relationship of interest?
- Are there any variables that mediate the relationship?
- What is commonly known in the scientific community?

A drawback is that you might not have such a parsimonious model.

Change in Estimates

This is meant to be a method of detecting confounding.

Fit a model without adjusting for the potential confounder. Then, fit a second model with adjusting for the potential confounder. If the estimate of the primary variable changes, there is evidence of confounding.

The rule of thumb is a change of 10% or more.

Stepwise Selection

Stepwise regression is an automated process of building a model by either adding in or removing predictors based on their p-values.

Backward Elimination

1. Start with all predictors in the model.
2. Remove the predictor with the highest p-value over the level of significance, usually 0.05, 0.10, or 0.15.
3. Refit the model.
4. Repeat step 2.
5. Stop when all the p-values are less than the level of significance.

Forward Selection

1. Start with no variables in the model.
2. For the variables not in the model, calculate their p-value if they are added to the model.
3. Keep the variable with the lowest p-value under the level of significance.
4. Repeat step 2.
5. Stop when no new predictors have p-values under the level of significance.

Stepwise Regression

Stepwise regression combines forward selection and bckward elimination. It has the same steps as forward selection but if any of the earlier predictors become non-significant, they can be removed.

A Word of Caution

Stepwise regression can be simple to understand and easy to compute. However,

- The model is prone to overfitting.
- It is possible to miss the “optimal” model.
- The process does not adjust for multiple testing.
- The model tends to be oversimplified.
- It does not account for the total number of parameters in the model.
- It does not incorporate any scientific knowledge into the selection process.

Advanced Variable Selection

Advanced methods include:

- Penalized regression
- Principal components
- Propensity scores

Wrap-Up

In my opinion, a model should have the following properties:

- Parsimony
- Identifiability
  - Ability to obtain unique estimates for the parameters of the model, given the data
- Goodness of fit
- Scientific relevance and consistency
- Predictive power
  - Including being able to predict on observations outside of the sample

Advice

- Use statistical findings as well as scientific knowledge to determine which variables you want to include.
- Utilize EDA tools to examine relationships between variables and incorporate that knowledge into the building of the model.
- If you know a variable is a confounder of the relationship of interest, include the variable no matter what.
- If the variable is a strong, categorical confounder, you can always stratify your analysis.
- It's better to err on the side of caution when it comes to adding variables. However, do not create "kitchen sink" models.

Active Learning Exercise

This exercise involves creating a conceptual framework. You have two options for creating the framework:

- Create a framework for your final project, i.e. a framework for the relationship between your variable of interest and your outcome.
- Create a framework for the relationship between education and a health status of choice.

Draw it out, similar to the diagram presented during the lecture. Be sure to indicate what is known in the scientific literature and what you hypothesize to be true. Discuss ideas with your classmates.