Regression III
Critiques of Common Practice

Dave Armstrong

University of Wisconsin – Milwaukee
Department of Political Science

e: armstrod@uwm.edu
w: www.quantoid.net/teachicpsr/regression3

Model (Mis)Specification

Model (Mis)Specification and Omitted Variable Bias

- We'll talk in more depth about model selection (e.g., selecting which variables belong in the model) later on in the course.
- Finally, for today, we'll address the *potential* problem of omitted variable bias.
- First we must decide whether our model is one of empirical relationship or causal relationship. While we all probably want to make causal claims, there are some definite problems that come along with that.
  - A model of empirical relationship simply traces the empirical covariance of one variable with a set of other variables.
  - A causal model proposes that we have been able to ascertain and include all of the relevant causal determinants of $\boldsymbol{y}$ and as such, we can interpret coefficients as the *effect of* $\boldsymbol{x}$ on $\boldsymbol{y}$.

Misspecification

Suppose we're interested in the population model:

$$\boldsymbol{y}^* = \boldsymbol{X}^*\beta + \varepsilon = \boldsymbol{X}_1^*\beta_1 + \boldsymbol{X}_2^*\beta_2 + \varepsilon$$

where variables with a * superscript are mean deviated versions of the original variables. $\boldsymbol{X}_1^*$ and $\boldsymbol{X}_2^*$ are matrices of regressors.

Now, define $\boldsymbol{X}_2^* + \varepsilon \equiv \tilde{\varepsilon}$ such that

$$\boldsymbol{y}^* = \boldsymbol{X}_1^*\beta_1 + \tilde{\varepsilon}$$

Now we can see what happens to $\boldsymbol{b}_1$.

$\boldsymbol{b}_1$ and Mis-specification

$$\begin{aligned}
\boldsymbol{b}_1 &= \left(\boldsymbol{X}_1^{*\prime}\boldsymbol{X}_1^{*}\right)^{-1}\boldsymbol{X}_1^{*\prime}\boldsymbol{y}^{*} \\
&= \left(\frac{1}{n}\boldsymbol{X}_1^{*\prime}\boldsymbol{X}_1^{*}\right)^{-1}\frac{1}{n}\boldsymbol{X}_1^{*\prime}\boldsymbol{y}^{*} \\
&= \left(\frac{1}{n}\boldsymbol{X}_1^{*\prime}\boldsymbol{X}_1^{*}\right)^{-1}\frac{1}{n}\boldsymbol{X}_1^{*\prime}\left(\boldsymbol{X}_1^{*}\beta_1 + \boldsymbol{X}_2^{*}\beta_2 + \varepsilon\right) \\
&= \beta_1 + \left(\frac{1}{n}\boldsymbol{X}_1^{*\prime}\boldsymbol{X}_1^{*}\right)^{-1}\frac{1}{n}\boldsymbol{X}_1^{*\prime}\boldsymbol{X}_2^{*}\beta_2 + \left(\frac{1}{n}\boldsymbol{X}_1^{*\prime}\boldsymbol{X}_1^{*}\right)^{-1}\frac{1}{n}\boldsymbol{X}_1^{*\prime}\varepsilon
\end{aligned}$$

Taking probability limits produces:

$$\begin{aligned}
\text{plim } \boldsymbol{b}_1 &= \beta_1 + \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\beta_2 + \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\sigma}_{1\varepsilon} \\
&= \beta_1 + \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\beta_2
\end{aligned}$$

Where: $\boldsymbol{\Sigma}_{11} \equiv \text{plim } \left(\frac{1}{n}\right)\boldsymbol{X}_1^{*\prime}\boldsymbol{X}_1^{*}$, $\boldsymbol{\Sigma}_{12} \equiv \text{plim } \left(\frac{1}{n}\right)\boldsymbol{X}_1^{*\prime}\boldsymbol{X}_2^{*}$, and $\boldsymbol{\sigma}_{1\varepsilon} \equiv \text{plim } \left(\frac{1}{n}\right)\boldsymbol{X}_1^{*\prime}\varepsilon = 0$ (by assumption).

Mis-specification Bias

We assumed $\boldsymbol{\sigma}_{1\varepsilon} = 0$, but what about $\boldsymbol{\sigma}_{1\tilde{\varepsilon}}$?

$$\begin{aligned}
\text{plim } \frac{1}{n}\boldsymbol{X}_1^{*}\tilde{\varepsilon} &= \text{plim } \frac{1}{n}\boldsymbol{X}_1^{*}\left(\boldsymbol{X}_2^{*}\beta_2 + \varepsilon\right) \\
&= \boldsymbol{\Sigma}_{12}\beta_2 + \boldsymbol{\sigma}_{1\varepsilon}
\end{aligned}$$

So, $\boldsymbol{\sigma}_{1\tilde{\varepsilon}}$ can only be 0 if $\boldsymbol{\Sigma}_{12}$ (the correlation between $\boldsymbol{X}_1^{*}$ and $\boldsymbol{X}_2^{*}$ is 0 or β_2 the effect of $\boldsymbol{X}_2^{*}$ in the population is 0.

This is the classic *omitted variable bias*.

Omitted Variable Bias: The Phantom Menace

- We have the idea that the bias in our coefficients is monotonically decreasing (and approaching zero) as the proportion of relevant controls in our model increases.
- That is to say, if in the true data generating process (DGP), there are 100 variables - a model that includes 75 of them is *better* (coefficients have smaller bias) than a model that includes only 50 of them.
- Clarke (2005) shows that this is not necessarily the case.
  - Adding a subset of controls does not necessarily make the model *better*.
  - It could actually make the model *worse*.
- His recommendations:
  1. Focus on research design and look for natural experiments.
  2. Test theories on smaller, narrower domains (e.g., spatially or temporally narrower).

P-values

As Berk (2004) suggests - one of the fundamental question about statistical inference is - when p-values concern confidence intervals and statistical tests, to what does the probability refer? That is, "probability of what"?

- Assuming X is either fixed by design or considered fixed when the particular set of x values arise in the data, then random sampling results in inferences as one would expect.
- Sampling schemes other than randomness result in rather different properties with respect to inference.

Dealing with Apparent Populations

Below are some methods for dealing with non-random sample selection (particularly when we have non-randomly collected something more like a population). We will talk about each in turn.

1. Treat the data as the population.
2. Treat the data *as if* they were randomly sampled from a population
3. Redefine the population.
4. Invent a population.
5. Model-based sampling.

Treat Data Like a Population

- Description is the only game - the relationship you calculate is the relationship in the population.
- To the extent descriptions are not great (i.e., don't explain a lot of variation), the coefficients you calculate may still not provide insight into the DGP.
- Many problems do not require the frequentist thought experiment of infinite repeated sampling - treating the data as fixed is fine.
  - This might be particularly true in policy situations.

Treat the data *as if* they were randomly sampled from a population

Suggest that the data are approximately randomly sampled from a *real* population.

- Very good data and/or theory required to justify this.
- "Full disclosure" is insufficient - saying you assume your sample is a random sample without evaluation of those assumptions leaves readers not knowing which results to believe.
- Even if the population from which the sample is well-defined, sampling often does not happen in anything close to a random fashion.

Consequences:

- Regression parameters are bad estimates of population parameters (coefficients and variance explained can be attenuated).

Example

Sampling strategy: Collect water samples from a beach every Wednesday around noon over the Summer to measure levels of toxins from storm overflow. Want to infer to all days/times for that beach. Need to make (at least) the following assumptions:

- Toxin concentrations are independent of time of day and day of week.
- The 7-day time gap is sufficient to remove any "memory" (thus observations would be independent).

Data could (and should) be marshaled to provide evidence in favor of these assumptions.

Redefine the Population

Another seemingly reasonable strategy might be to redefine the population post-sampling such that the sample is a random draw from the population.

- Since the missing data occurred after the sampling procedure happened, this is also not appropriate.
- The process that made the data unavailable may be confounded with the relationship of interest.

This makes the justification for the inferential process circular. Convenience samples are, simply, not good fodder for (frequentist) inference.

> *To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of. (Fisher, 1938)*

Invent a Population: Superpopulation

The superpopulation argument is one of the mot prevalent in political science.

- Definition is often circular - the population is the set of other possible circumstances from which this sample could be a random sample.
- Superpopulations are not real and thus not well-defined, so inferences to superpopulations are tenuous at best.
- Some superpopulations *could* exist, but to make inferences to those sorts of populations, the conditions under which the superpopulation exists should be both well-defined and theoretically/empirically justifiable.

Even more problematic is the much of our "population" data suffers from non-random, perhaps non-ignorable missingness.

Superpopulation Example (Berk)

For example, monthly economic indicators from the year 2000 might be treated as "random" realizations of indicators that *could* have existed. Things that need to be specified for this to make sense:

- The data generating process (i.e., a strong and convincing theory, particularly about the random processes leading to this rather than another set of values),
- The conditioning factors pertaining to the observation - what were the particularities of the year 2000 (e.g., exchange rates with other countries, patterns of tariffs and constraints on international trade). These serve as the foundation for the superpopulation.
- Technically, even then you would need to show empirically that there are other years like the year 2000 and that the year 2000 can be meaningfully treated as a sample.

Model-Based Sampling

Propose a model by which nature produces data inferences are made to the model said to be responsible for the data generating process.

- The outcome variables are thought to be random realizations from the model (that is, the model could have generated different values of y).
- Since the outcome variables are random variables, regardless of how many of them we have (even if we have all N of them), each one was one realization of a random process.
- Even if we have the population, uncertainty remains regarding the parameters of the model that generated this (and could generate another) realization.
- The sampling scheme is irrelevant because all observations are assumed to be caused by the same natural process.

Justification of Model-based Sampling

Distributional assumption is key - why would we expect deviations from expectation to be normally distributed?

- CLT says that the sum of a bunch of independent, normal random variables approaches normality.
- The disturbance term in the model can be thought of in such a way as to justify such a distribution.
- A story is still needed to justify this - what sorts of things comprise the disturbance term? Can they really be thought to be independent?

Problems in Model-based Sampling

At best in the literature a hypothesis is done with the null hypothesis being the assumed distribution - failure to reject is taken as evidence the null is true.

- Treats failure to reject "accepting the null".
- These tests often have low power (leading to fewer rejections of falls null hypotheses than we would like).
- Many different random processes (i.e., distributions) can produce functionally equivalent-looking data.

For inference to make sense, the method that nature uses to make data has to be well-understood and explicated.

Sampling Conclusions

- Remember, we are data analysts/social scientists, not magicians.
- Assumptions, conditions, models required for the frequentist thought experiment of one kind or another to make sense must be reasonable, theoretically justified and empirically evaluated.

What Does "Holding Constant" Mean?

Mathematically:

- A covariance adjustment - removing variance in both x and y that can be explained by z (e.g., partial out the effect of z).

Substantively:

- Independent variables must be *independently manipulable*.
- What would it mean, when predicting income, to hold occupation constant while "manipulating" education?
  - Education is theoretically manipulable - people can gain more education and interventions aimed at such can be undertaken.
  - What would it mean to hold an eventual occupation constant while intervening on education?
  - Post hoc considerations are rather more helpful, but still unsatisfying.

Covariance Adjustment $\not\Rightarrow$ Independent Manipulability

Some Cautions about Statistical Inference

- Models have to be right and sampling procedures well-justified for inference to make sense. Failure on either count means inferences are "right" to a greater or lesser degree.
- p-values will almost always be anti-conservative. Since we rarely (never) take into account model selection uncertainty, total variability is grossly underestimated. Even formal procedures to correct for multiple testing will be insufficient here.
- Inferences should be done on a validation dataset or cross-validation should be used to prevent overfitting and capitalizing on chance through exploration.

Significant vs Not Significant

Gelman and Stern (2006) argue that the distinction between significant and not significant is less interesting than we think.

- Comparing levels of statistical significance is not appropriate.
- When comparing models, more matters than sign and significance.
  - Need to understand whether two estimates are statistically and substantively different from *each other*.

Determining Relative Importance

- If two explanatory variables are measured in exactly the same units, we can asses their relative importance in their effect on y quite simply
  - The larger the coefficient, the stronger the effect
- Often, however, our explanatory variables are not all measured in the same units, making it difficult to assess relative importance
- This problem can be overcome for quantitative variables by using standardized variables
- We can generalize standardization to include sets of variables, thus incorporating factors, interactions, and multiple effects

Standardized Regression Coefficients

- Standardized coefficients enable us to compare the relative effects of two or more explanatory variables that have different units of measurement
- Standardized coefficients convert all the variables into standard deviation units:

$$\frac{Y_i - \bar{Y}}{S_y} = \left(B_1 \frac{S_1}{S_y}\right) \frac{X_{i1} - \bar{X}_1}{S_1} + \cdots + \left(B_k \frac{S_k}{S_y}\right) \frac{X_{ik} - \bar{X}_k}{S_k} + \frac{E_i}{S_y}$$

- We interpret the effects of a standardized variable as the average number of standard deviation units Y changes with an increase in one standard deviation in X
- Since they don't have a standard deviation, standardized coefficients for factors are meaningless

Standardized Coefficients using Matrices

- Recall that the matrix equation for the least-squares slopes is:

$$\boldsymbol{B} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

  where $\boldsymbol{X}'\boldsymbol{X}$ is the variance-covariance matrix
- The matrix equation for the standardized coefficients is then:

$$\boldsymbol{b*} = \boldsymbol{R}_{XX}^{-1}\boldsymbol{r}_{Xy}$$

  where $\boldsymbol{R}_{XX}$ is the correlation matrix between the explanatory variables and $\boldsymbol{r}_{Xy}$ is the vector of correlations between the explanatory variables and the outcome variable.

Standardized Variables: Cautions

- It makes little sense to standardized dummy variables:
  - It cannot be increased by a standard deviation so the regular interpretation for standardized coefficients does not apply
  - Moreover, the standard interpretation of the dummy variable showing differences in level between two categories is lost
- We cannot standardize interaction effects
  - They are dependent on the main effects
  - We can, however, standardize quantitative variables beforehand and construct interaction terms afterwards.

Standardized Variables in **R**

- Unlike some statistical packages, **R** does not automatically return standardized coefficients
- A separate model must be fitted to a dataset for which all quantitative variables have been standardized
  - This is done using the scale function.
  - You can do this in the model:

```
> Weakliem <- read.table("http://www.quantoid.net/files/reg3/weakliem.txt
> Weakliem$democrat <- factor(Weakliem$democrat, levels=0:1,
+    labels = c("Non Democracy", "Democracy"))
> mod <- lm(scale(secpay) ~ scale(gini)*democrat, data=Weakliem)
```

  - Alternatively, all the quantitative variables can be standardized at the same time by creating a new scaled dataset:

```
> library(DAMisc)
> scaled.data <- scaleDataFrame(Weakliem)
> mod <- lm(secpay ~ gini*democrat, data=scaled.data)
```

Relative Importance of a Set of Predictors (1)

- In the standardized variables case, we can easily determine relative importance by the ratio of the two standardized coefficients
  - In other words, we assess the ratio of the standard deviations of the two contributions to the linear predictor
- Imagine now that we are interested in the relative effects of two sets of variables (e.g., a set of dummy regressors for a single variables versus the effects of another variable)
  - Instead of individual standardize variables, we explore the relative contributions that the set of variables make to the dispersion of the fitted values

Relative Importance of a Set of Predictors (2)

- Following from Silber et al. (1995) the ratio of variances of the contributions of two sets of variables, ω, can be determined by:

$$\omega = \frac{\boldsymbol{\beta' X' X \beta}}{\boldsymbol{\gamma' H' H \gamma}}$$

  Where β is the coefficient vector and $\boldsymbol{X}$ is the model matrix for the *set1 predictors*; γ is the coefficient vector and $\boldsymbol{H}$ is the model matrix for the *set2 predictors*
- If $\omega = 1$, then both sets of predictors contribute the same amount of variation in the outcome variable
- MLE also provides an approximate test of $H_0 : \omega = 1$ which refers to the standard normal distribution, yielding the standard confidence intervals, thus making the test generalizable to GLMs

Standardization vs Relative Importance

- If each set contains a single quantitative variable, then the relative importance is just the ratio of the absolute value of standardized coefficients.
- You can think of relative importance as a generalization of standardization to sets of predictors (not so much for interpretation, but for evaluating effect size).

The `relimp` Package in **R**

- The `relimp` package for **R** implements the ω measure of relative importance of Silber et al.
- The variables of interest can be specified in a command line, with each effect given the number corresponding to its column(s) in the model matrix (or row in the regression output)

```
> library(relimp)
> relimp(model, set1=1:3, set2=4:5)
```

Relative Importance: An Example (1)

```
> mod1<-lm(prestige~income+education+type, data=Duncan)
> summary(mod1)

Call:
lm(formula = prestige ~ income + education + type, data = Duncan)

Residuals:
    Min      1Q  Median      3Q     Max
-14.890  -5.740  -1.754   5.442  28.972

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.18503    3.71377  -0.050  0.96051
income        0.59755    0.08936   6.687 5.12e-08 ***
education     0.34532    0.11361   3.040  0.00416 **
typeprof     16.65751    6.99301   2.382  0.02206 *
typewc      -14.66113    6.10877  -2.400  0.02114 *
---
Signif. codes:  0
```


Relative Importance: An Example (2)

```
> library(relimp)
> relimp(mod1, set1=2, set2=4:5)

Relative importance summary for model
    lm(formula = prestige ~ income + education + type, data = Duncan)

      Numerator effects ("set1")      Denominator effects ("set2")
1                         income                          typeprof
2                                                           typewc

Ratio of effect standard deviations: 1.332
Log(sd ratio):                0.287   (se 0.276)

Approximate 95% confidence interval for log(sd ratio): (-0.255,0.829)
Approximate 95% confidence interval for sd ratio:      (0.775,2.29)
```

In this instance, the ω measure suggests these two sets of predictors contribute equally to the variation in the dependent variable.

Readings

Today: Critiques, Cautions and Miscellany

* Clarke (2005)
* Berk, Western and Weiss (1994)
– Berk (2004) * Gelman and Stern (2006)
* Silber, Rosenbaum and Ross (1995)

Berk, Richard A. 2004. *Regression Analysis: A Constructive Critique*. Thousand Oaks, CA: Sage.

Berk, Richard A., Bruce Western and Robert E. Weiss. 1994. "Statistical Inference for Apparent Populations with discussion." *Sociological Methodology* 25:421–485.

Clarke, Kevin. 2005. "The Phantom Menace: Omitted Variable Bias in Econometric Research." *Conflict Management and Peace Science* 22(4):341–352.

Gelman, Andrew and Hal Stern. 2006. "The Difference Between 'Significant' and 'Not Significant' is not Itself Statistically Significant." *The American Statistician* 60(4):328–331.

Silber, Jeffrey H., Paul R. Rosenbaum and Richard N. Ross. 1995. "Comparing the Contributions of Groups of Predictors: Which Outcomes Vary with Hospital Rather Than Patient Characteristics." *Journal of the American Statistical Association* 90(429):7–18.