Regression III: Advanced Methods

William G. Jacoby
Department of Political Science
Michigan State University

http://polisci.msu.edu/jacoby/icpsr/regress3
Determining Relative Importance

• If two independent variables are measured in exactly the same units, we can assess 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:

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

\[
Z_{iY} = B_{1*}Z_{i1} + \cdots + B_{k*}Z_{ik} + E_i^*
\]

- We interpret the effects of a standardized variable as the 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:

\[ b = (X'X)^{-1}X'y \]

Where $XX$ is the variance-covariance matrix.

• The matrix equation for the standardized coefficients is then:

\[ b^* = (R_{XX})^{-1}r_{XY} \]

• Here $R_{XX}$ is the correlation matrix for the X’s and $r_{XY}$ is the vector of correlations between the explanatory variables and Y
Standardized Variables: Cautions

• It makes little sense to standardize 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 difference in average level of $Y$ between two categories is lost

• We cannot standardize interaction effects (They are not dependent on the main effects)

• Standardized coefficients can never be compared across samples or populations
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.
  - Variables can be standardized individually:

  ```r
  > gini.std <- scale(gini)
  ```

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

  ```r
  > scaled.data <- data.frame(scale(Weakliem [, c('secpay', 'gini', 'gdp')]))
  ```
Relative Importance of Sets 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 variable versus the effects of another variable)
  - Standardized variables—where effects are measured in terms of their impact on the standard deviation units of the dependent variable—provides the generalization to comparing two sets rather than just a pair of predictors
Relative Importance of Sets of Predictors (2)

• Following from Silber et al. (1995) the ratio of variances of the contributions of two sets of variables, $\omega$, can be determined by:

$$
\omega = \frac{\beta^T X^T X \beta}{\gamma^T H^T H \gamma}
$$

Where $\beta$ is the coefficient vector and $X$ is the model matrix for the *set1* predictors; $\gamma$ is the coefficient vector and $H$ is the model matrix for the *set2* predictors

• If $\omega = 1$, then both sets of predictors contribute the same amount of variation in the dependent 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
The `relimp` package for R

- The `relimp` package for R implements the $\omega$ measure of relative importance of Silber et al.
- It can be used in two ways:
  1. The variables of interest can be specified in a command line, with each effect given the number corresponding to its column in the model matrix (or row in the regression output)

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

  2. Alternatively, sets of variables can be selected from a dialogue box by simply typing:

    ```
    >library(relimp)
    >relimp(model)
    ```
Relative Importance: An Example (1)

```r
> mod1<-lm(prestige~income+education+type)
> summary(mod1)
```

Coefficients:

|                  | Estimate | Std. Error | t value | Pr(>|t|) |
|------------------|----------|------------|---------|----------|
| (Intercept)      | 16.47249 | 9.53329    | 1.728   | 0.09172  |
| income           | 0.59755  | 0.08936    | 6.687   | 5.12e-08 *** |
| education        | 0.34532  | 0.11361    | 3.040   | 0.00416 ** |
| typebc           | -16.65751| 6.99301    | -2.382  | 0.02206 *  |
| typewc           | -31.31865| 5.07854    | -6.167  | 2.75e-07 *** |

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.744 on 40 degrees of freedom
Multiple R-Squared: 0.9131, Adjusted R-squared: 0.9044
F-statistic: 105 on 4 and 40 DF, p-value: < 2.2e-16
Relative Importance: An Example (2)

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

Relative importance summary for model

\[ \text{lm(formula = prestige ~ income + education + type)} \]

\[ \text{Numerator effects ("set1")} \]
\[ 1 \]
\[ \quad \text{income} \]
\[ 2 \]
\[ \text{Denominator effects ("set2")} \]
\[ \text{typebc} \]
\[ \text{typewc} \]

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)