Regression III

Variable Importance

Dave Armstrong

Goals for this Lecture

1. Understand how to compare variable effects.
2. Presentation of comparable effects.

Comparing Effect Sizes

Two things to keep in mind.

1. Variable's variability
2. Coefficient size

In both cases, you to be cognizant of multi-term variables - interactions, polynomials, factors.

Notes

Type notes here...

Determining Relative Importance

If two explanatory variables are measured in exactly the same units, we can (kind of) asses their relative importance in their effect on y quite simply

- The larger the coefficient, the stronger the effect
- This does not, however, take into account the variable's variance.

A better rule would be:

- For variables measured in the same units with roughly the same variance, bigger coefficients mean larger effects.

Consider the example below.

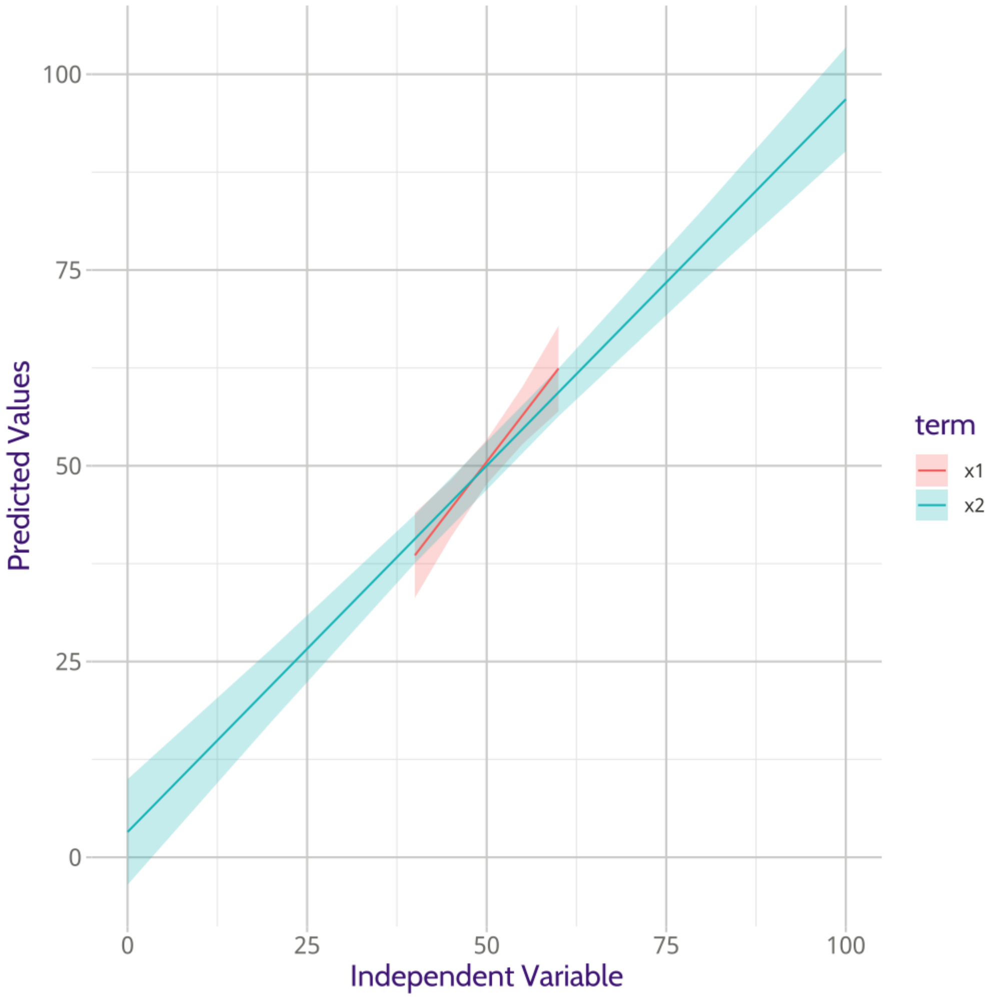
Notes

Type notes here...

Same-unit comparisons

```r
set.seed(519)
dat <- tibble(
  x1 = runif(250, .4,.6)*100,
  x2 = runif(250, .1, .9)*100,
  yhat = -50 + x1 + x2,
  y = yhat + rnorm(250, 0, sd(yhat))
)
mod <- lm(y ~ x1 + x2, data=dat)
summary(mod)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -60.886 -15.191  -2.236  15.011  67.736
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -56.37585   12.50197  -4.509 1.01e-05 ***
## x1            1.19258    0.23687   5.035 9.23e-07 ***
## x2            0.93575    0.06178  15.147  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.86 on 247 degrees of freedom
## Multiple R-squared:  0.4997,    Adjusted R-squared:  0.4956
## F-statistic: 123.3 on 2 and 247 DF,  p-value: < 2.2e-16
```



Notes

Type notes here...

Other Methods of Comparison

If explanatory variables are not all measured in the same units, it is difficult to assess relative importance

- This problem can be overcome for quantitative variables by using standardized variables.
- For other types of variables, we need a different method.

Notes

Type notes here...

Standardized Regression Coefficients

- Standardized coefficients enable us to compare the relative effects of two or more explanatory variables that have different units of measurement

- If this is the un-standardized model

  $$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + e_i$$

  The standardized coefficients are:

- Fully standardized: $b_j^* = b_j \frac{s_{x_j}}{s_y}$
  - For every one-standard-deviation change in x_j, we expect a b_j^* *standard deviation* change in y holding all other model covariates constant.
- x-standardized: $b_j^* = b_j s_{x_j}$
  - For every one-standard-deviation change in x_j, we expect a b_j^* *unit* change in y holding constant all other model covariates.

Notes

Type notes here...

How Many Standard Deviations?

Gelman (2008, *Statistics in Medicine*) suggests dividing quantitative predictors by two rather than one standard deviations.

- Binary variables have a standard deviation of $\sqrt{p(1-p)}$.
- For symmetric binary variables, this is: $\sqrt{.5 \times .5} = 0.5$
    - In which case a change from $0 \rightarrow 1$ would be two standard deviations.
- If we divide by 1 SD, quantitative variables will have a standard deviation of 1, twice the size of that of binary variables.
    - Dividing by two standard deviations gives the resulting quantitative variable a standard deviation of 0.5, the same as a symmetric binary variable.

Notes

Type notes here...

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.

```r
mod <- lm(scale(prestige) ~ scale(income) + scale(education) +
            type, data=Duncan)
```

- Alternatively, all the quantitative variables can be standardized at the same time by creating a new scaled dataset (from the `{DAMisc}` package):

```r
scaled.data <- scaleDataFrame(Duncan,
                              numsd = 2)
mod <- lm(prestige ~ income + education + type,
          data = scaled.data)
```

Notes

Type notes here…

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 multi-term variables: interaction effects or polynomials

- Interactions are dependent on the main effects
- We can, however, standardize quantitative variables beforehand and construct higher-order terms afterwards.
- Regardless of this, we cannot determine importance of multi-term variables by looking at any single coefficient.

Notes

Type notes here...

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

Notes

Type notes here...

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:

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

  Where β is the coefficient vector and $\mathbf{X}$ is the model matrix for the *set1 predictors*; γ is the coefficient vector and $\mathbf{H}$ 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

Notes

Type notes here...

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). For example:

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

Notes

Type notes here...

Relative Importance: An Example (1)

```r
mod1<-lm(interlocks ~ log(assets) + sector + nation, data=Ornste
summary(mod1)$coefficients
```

```
##                 Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) -28.4429670  4.9271875 -5.7726578 2.465024e-08
## log(assets)   5.9907825  0.6813797  8.7921354 3.235865e-16
## sectorBNK    17.3227304  5.1846800  3.3411378 9.710771e-04
## sectorCON    -2.7126874  5.4241073 -0.5001168 6.174628e-01
## sectorFIN    -1.2744881  3.4121039 -0.3735197 7.090998e-01
## sectorHLD    -2.2916036  4.6132359 -0.4967454 6.198350e-01
## sectorMAN     1.2440168  2.3665722  0.5256619 5.996209e-01
## sectorMER    -0.8801086  3.0346472 -0.2900201 7.720577e-01
## sectorMIN     1.7566138  2.4447619  0.7185214 4.731527e-01
## sectorTRN     1.8888418  3.3023169  0.5719747 5.678882e-01
## sectorWOD     5.1056070  3.0990366  1.6474820 1.008012e-01
## nationOTH    -3.0533129  3.0872167 -0.9890180 3.236759e-01
## nationUK     -5.3294006  3.0714272 -1.7351544 8.403005e-02
## nationUS     -8.4912938  1.7174063 -4.9442544 1.458432e-06
```

```r
library(relimp)
relimp(mod1, set1=3:11, set2=12:14)
```

```
##
## Relative importance summary for model
##     lm(formula = interlocks ~ log(assets) + sector + nation, data = Orn
##
##        Numerator effects ("set1")      Denominator effects ("set2")
## 1                        sectorBNK                         nationOTH
## 2                        sectorCON                          nationUK
## 3                        sectorFIN                          nationUS
## 4                        sectorHLD
## 5                        sectorMAN
## 6                        sectorMER
## 7                        sectorMIN
## 8                        sectorTRN
## 9                        sectorWOD
##
## Ratio of effect standard deviations: 0.858
## Log(sd ratio):                       -0.153   (se 0.314)
##
## Approximate 95% confidence interval for log(sd ratio): (-0.768,0.461)
## Approximate 95% confidence interval for sd ratio:      (0.464,1.586)
```

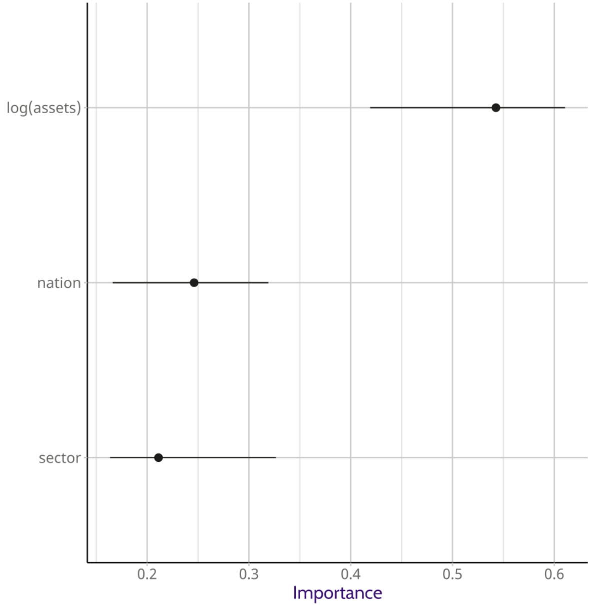
Notes

Type notes here...

Plotting "Importance"

Using Silber et. al.'s definition of importance, we could make a plot of not the relative importance, but the absolute importance of variables.

```
library(psre)
s <- srr_imp(mod1, data=Ornstein, R=1500, pct=TR
ggplot(s, aes(y=reorder(var, importance, mean),
              x=importance,
              xmin=lwr, xmax=upr)) +
  geom_pointrange() +
  theme_classic() +
  mytheme() +
  labs(x="Importance", y="")
```



Notes

Type notes here...

Importance of Interactions

We can also use the function above with interactions:

```r
data(Prestige, package="carData")
mod <- lm(prestige ~ income*education + women + type,
          data=Prestige)
srr_imp(mod, Prestige, pct=TRUE,
        combine_terms = list("Interaction" = c("income",
                                               "education",
                                               "income:education")))
```

```
##           var importance         lwr       upr
## 1       women 0.06654899 0.005393675 0.1361451
## 2        type 0.18670933 0.111036343 0.3254353
## 3 Interaction 0.74674168 0.596982745 0.8498367
```

Notes

Type notes here...