Removing Bias – the SIMEX Procedure

Tom Struppeck, FCAS

3/28/2017

What is linear regression?

- One variable, called the response, is expressed as a linear combination of the other variables, called the predictors, plus some noise (the errors).
- Different fields of study use other terms for response and predictors.
- We will assume that the noise has expected value 0. This is not a constraint if the model has an intercept term, which ours will.
- We will also assume that our errors are uncorrelated with each other and all have constant variance (homoscedastic).

The Gauss-Markov Theorem

- Under the assumptions on the previous slide, the Best Linear Unbiased Estimator (BLUE) of the coefficients is given by the Ordinary Least Squares (OLS) estimator.
  - Estimator: a function of the data
  - Unbiased: the expected value is the true value
  - Linear: the estimator is a linear function of the data
  - Best: among such estimators, this one has minimal variance

  So, an estimator that is BLUE is unbiased, by definition, and he Gauss-Markov Theorem guarantees that OLS estimators are BLUE, so this talk should be very short.

Meanwhile, in the real world …

- The coefficient estimates are often biased towards zero.
- p-values are conditional probabilities of true coefficient values being equal to zero, so bias towards zero will overstate p-values.
- Overstated p-values may make significant predictors appear to not have statistical significance, when in fact they would without the bias.

- Of course, Gauss-Markov says that none of this can happen.

Is the Gauss-Markov Theorem wrong?

- Of course not, but the conclusion of the theorem does not hold. This can only mean that the hypotheses of the theorem are not satisfied.
  - There don't seem to be many hypotheses …
    - We use OLS to obtain our estimates.
    - ?? What other hypothesis is there?

It must be the errors

- We need them to have expected value zero.
  - OK
- We need them to have constant variance.
  - OK
- We need them to be uncorrelated.
  - OK

- ?? Now what?

What are the errors?

- From the second slide:
  - One variable, called the response, is expressed as a linear combination of the other variables, called the predictors, plus some noise (the errors).
- To explore this further, let's assume just one predictor. This case is sometimes called Simple Linear Regression.

Simple Linear Regression

- Call the response variable Y and the predictor X.
- Model: $$Y = \beta_1 X + \beta_0 + \varepsilon$$
- OLS estimates: $$\hat{\beta}_1 = r_{XY} \frac{s_Y}{s_X}$$ and $$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$
- Where: r_{XY} is the sample correlation between X and Y and

  s_X and s_Y are the sample variances of X and Y, respectively.

Simple Linear Regression

- Call the response variable Y and the predictor X.
- Model: $Y = \beta_1 X + \beta_0 + \varepsilon$
- OLS estimates: $\hat{\beta}_1 = r_{XY} \frac{s_Y}{s_X}$ and $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$
- We now have a linear equation relating X and Y found by OLS.
- Is it unique?

What if …

- We call the response variable X and the predictor Y?
- Model: $X = \gamma_1 Y + \gamma_0 + \delta$
- OLS estimates: $\hat{\gamma}_1 = r_{YX} \frac{s_X}{s_Y}$ and $\hat{\gamma}_0 = \bar{X} - \hat{\gamma}_1 \bar{Y}$
- Is this the same equation? If so, the slope coefficients should be reciprocals.

$$\hat{\beta}_1 = r_{XY} \frac{s_Y}{s_X} \qquad \text{vs} \qquad \hat{\gamma}_1 = r_{YX} \frac{s_X}{s_Y}$$

They almost look like reciprocals

- Except that the correlation of X and Y and the correlation of Y and X are equal, so unless they are ± 1 (perfect correlation) there are (at least) two OLS lines.
- The difference is that the errors are fully attributed to the response variable; the predictor is assumed to be know with perfect accuracy.
- When there are measurement errors in the predictors, the hypotheses of the Gauss-Markov Theorem are not met.

A latent variable model

$$Y_{observed} = \beta X_{actual} + \alpha + \varepsilon_i$$

$$X_{observed} = X_{actual} + \delta_i$$

Two types of measurement noise

- The errors in the predictors (the measurement noise) might be independent of the actual value.
  - We take a measurement and expect to get it right on average, but we might be either over or under

- Or the errors might be highly correlated with the actual value.
  - This is the case when we round. If we round to the nearest integer, 0.9 is always rounded to 1.0, so the measurement error is completely determined by the actual value. (Perfect negative correlation in this case.)

A formula for the bias

$$E(\hat{\beta}) = \beta \frac{\sigma_X^2 + \sigma_{X\delta}}{\sigma_X^2 + 2\sigma_{X\delta} + \sigma_\delta^2}$$

Which can be written as:

- A credibility-weighted sum of the true value and zero:

$$E(\hat{\beta}) = \beta\left(\frac{\sigma_X^2 + \sigma_{X\delta}}{\sigma_X^2 + 2\sigma_{X\delta} + \sigma_\delta^2}\right) + 0\left(\frac{\sigma_{X\delta} + \sigma_\delta^2}{\sigma_X^2 + 2\sigma_{X\delta} + \sigma_\delta^2}\right)$$

- This is what causes the bias towards zero (attenuation).

What can be done about it?

- In the case of simple linear regression, we have a formula for the amount of attenuation, so we can simply adjust for it.
- But in other settings, the attenuation still occurs and a tool for adjusting for it is needed.
- Problem statement:
  - We have some observed predictor values (the predictor) that are the sum of the true predictor values and one unit of random noise. From these, estimate what the slope coefficients would be if the noise were not present.

Idea:

- There is no situation so bad that it can't be made worse by adding more noise.

- We'd love to just subtract the noise, but of course that is impossible.
- Instead we add more!
- We know that adding more noise will attenuate the coefficient estimates even more than they are. If we can quantify this, we may have a way to infer what would happen if we had no noise.

This is SIMEX

- SIMEX has two parts, the SIM and the EX
- SIMulation:
  - Generate additional random noise and add it to the observed data
  - Fit a regression using OLS, record the slope coefficient
  - Repeat many times to obtain the sampling distribution of the new slope
  - Add even more noise and repeat the above steps
- EXtrapolation:
  - Fit an appropriate curve through the average slope coefficient estimates as a function of the amount of noise (for original data, noise level = 1)
  - Evaluate the implied slope at noise level = 0.

An example showing the SIMEX extrapolation: