Misspecification and Data-Related Problems

Econometrics

J. Eduardo Vera-Valdés

eduardo@math.aau.dk

Department of Mathematical Sciences
Aalborg University
Denmark

AALBORG UNIVERSITY
DENMARK

Outline

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Introduction

Last lecture we study some properties of OLS.

They are product of a set of assumptions made on the model, particularly on the error terms.

Yet, there is no easy way to test if the assumptions are correct.

Today we will discuss the sort of issues that can arise if the assumptions are not valid.

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Outline

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Influential Observations and Leverage

OLS estimates are weighted averages of the elements of Y, where the weights are a function of the regressors X,

$$\hat{\beta}_j = [(X^t X)^{-1} X]_j Y$$

We can get a sense of the influence that each observation has on the estimates by comparing the results with and without the observation.

If the estimates change a lot between the two regressions, we may be suspicious of that observation.

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Influential Observations and Leverage

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Influential Observations and Leverage

Consider the regression given by

$$Y = X\beta^{(t)} + \alpha e_t + U,$$

where e_t is the unit basis vector having 1 in the t position and 0 otherwise.

Such a regression can help us to isolate the effect that observation t has on the regression.

By the FWL theorem, the estimator for $\beta^{(t)}$ is the same as the one from regression

$$M_t Y = M_t X\beta^{(t)} + V$$

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Influential Observations and Leverage

By the FWL theorem, $\hat{\alpha}$ is given by

$$\hat{\alpha} = \frac{e_t^T M_X Y}{e_t^T M_X e_t}.$$

So that the difference between the estimator with and without the t observation is given by

$$\hat{\beta} - \hat{\beta}^{(t)} = \frac{1}{(M_X)_{tt}}(X^T X)^{-1} X_t^T \hat{U}_t.$$

Which shows that the effect of an observation depends on both the residual (that depends on Y), and on the t observation of the matrix of regressors.

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Cook's Distance is a way to measure the effect that each data point has on the results.

For each observation, $t \in \{1, 2, \cdots, n\}$, we compute

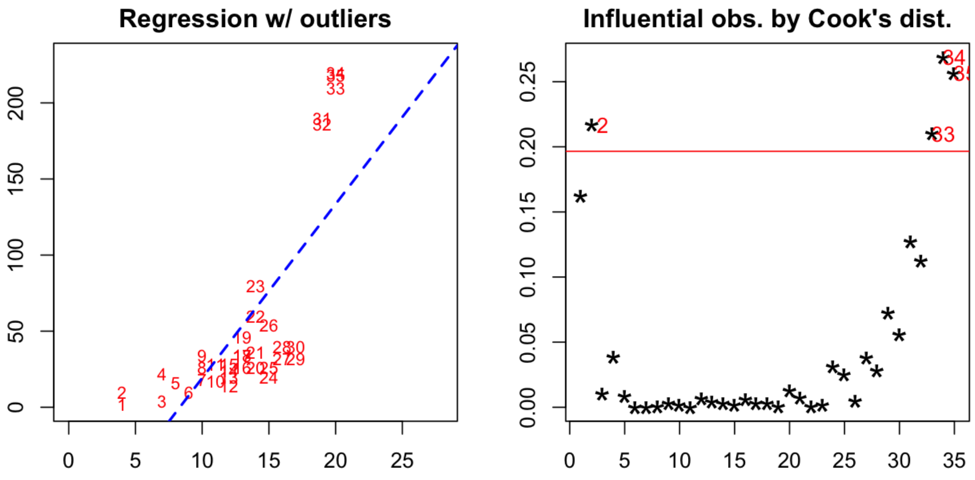
$$D_t = \frac{1}{ps^2} \sum_{j=1}^{n} \left(\hat{Y}_j - \hat{Y}_{j(-t)} \right)^2 = \frac{1}{ps^2} Y^T \left(P_{[X, e_t]} - P_X \right) Y,$$

where $\hat{Y}_j$ is the fit using all observations, $\hat{Y}_{j(-t)}$ is the fit removing observation t, s^2 is the estimator of the error variance, and p is the number of regressors.

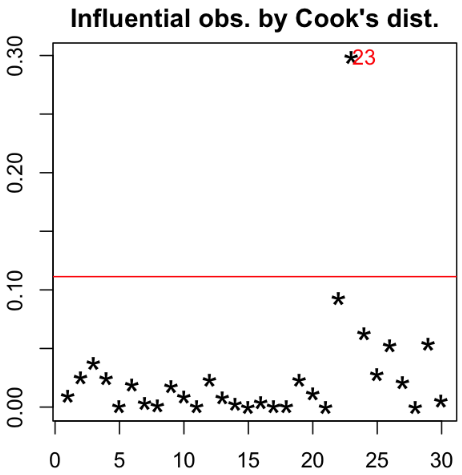
A Cook's distance greater than 3 or 4 times the average is called influential and may require further scrutiny.

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark


Regression w/ outliers
Influential obs. by Cook's dist.

Influential Observations and Leverage
Cook's Distance



J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Outline

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Collinearity

We saw last lecture that collinearity reduces the accuracy of the estimates, it causes the standard error for β_j to grow.

This means that the power of hypothesis tests are reduced by collinearity.

It is thus desirable to identify and address potential collinearity problems while fitting the model.

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Collinearity

We can compute the variance of the estimator associated to the j regressor in term of the rest of the regressors by

$$Var\left(\hat{\beta}_j\right) = \sigma^2 (X_j^T M_{X_{-j}} X_j)^{-1}.$$

Alternatively, we can write

$$Var\left(\hat{\beta}_j\right) = \frac{\sigma^2}{(n-1)Var(X_j)} \frac{1}{(1 - R^2_{X_j|X_{-j}})},$$

where $R^2_{X_j|X_{-j}}$ is the R^2 from a regression of X_j on all other regressors.

The last part in the product captures the relation between X_j and the rest of the regressors, it is called the variance inflation factor (VIF).

Collinearity

The *VIF* is given by

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}.$$

As a **rule of thumb**, a *VIF* value that exceeds 10 indicates a problematic amount of collinearity.

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Outline

Influential Observations and Leverage

Collinearity

Misspecification

Measurement Errors

Simultaneous Equations

Non-normality

Homoskedasticity and autocorrelation

Summing Up

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Overspecification

Sometimes we may have a set of possible regressors but we are uncertain about which ones to include.

A model is said to be overspecified if some variables are mistakenly included in the model.

In this sense, including irrelevant explanatory variables in a model makes the model larger than it need have been.

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Overspecification

Suppose we estimate the model

$$Y = X\beta + Z\gamma + U,$$

when the data are actually generated by

$$Y = X\beta + U.$$

Note that the estimated model is a special case of the correct model with $\gamma = 0$.

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Overspecification

By the FWL theorem, the estimator from the larger model is given by

$$\tilde{\beta} = (X^T M_Z X)^{-1} X^T M_Z Y.$$

We can show that it is unbiased

$$\tilde{\beta} = \beta + (X^T M_Z X)^{-1} X^T M_Z U.$$

And compute its variance

$$Var(\tilde{\beta}) = \sigma^2 (X^T M_Z X)^{-1}.$$

Overspecification

By the Gauss-Markov theorem, we know that the estimator from the correct (unaugmented) model, $\hat{\beta}$, is more efficient than $\tilde{\beta}$.

We prove this directly by showing that $Var(\tilde{\beta}) - Var(\hat{\beta})$ is a positive semidefinite matrix, which is equivalent to $Var^{-1}(\hat{\beta}) - Var^{-1}(\tilde{\beta})$ being a positive semidefinite matrix.

Note that

$$Var^{-1}(\hat{\beta}) - Var^{-1}(\tilde{\beta}) = \sigma^{-2}(P_Z X)^T (P_Z X),$$

which is positive semidefinite.

Thus, adding irrelevant variables make OLS inefficient but it remains unbiased.

Overspecification
Goodness of Fit

Another undesirable feature of adding irrelevant regressors is that the R^2 increases as we add more variables regardless of whether they actually are part of the model.

Note that the R^2 for both models have the same denominator, so that the difference relies purely on the numerator.

Now, the numerators for the augmented and unaugmented regression are given by $||P_{X,Z}Y||^2$ and $||P_X Y||^2$, respectively.

So the difference can be written as

$$Y^T(P_{X,Z} - P_X)Y,$$

which can be shown to be a quadratic form.

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

A solution to the problem is to modify the R^2 to account for the number of regressors in the model.

The **adjusted** R^2 is given by

$$\bar{R}^2 = 1 - \frac{(n-1)Y^T M_X Y}{(n-k)Y^T M_\iota Y},$$

where k is the number of regressors, and n is the sample size.

The adjusted R^2 can be motivated as a ratio of unbiased variance estimators.

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Underspecification

Now we analyse the opposite situation, assume we fail to include variables that should be in the regression.

Thus, now the correct model is

$$Y = X\beta + Z\gamma + U,$$

while we estimate

$$Y = X\beta + U.$$

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Underspecification

The estimator from the smaller model is

$$\tilde{\beta} = (X^T X)^{-1} X^T Y,$$

so that

$$E[\tilde{\beta}] = \beta + (X^T X)^{-1} X^T Z \gamma + (X^T X)^{-1} X^T U.$$

The second term does not disappear in general making the estimator biased.

Moreover, even asymptotically, the second term does not disappear so that the estimator is inconsistent.

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Outline

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Measurement Errors

Many economic variables are measured with error.

For example, macroeconomic time series are often based on surveys, and they suffer from sampling variability.

Whenever there are measurement errors, the values economists observe inevitably differ, to a greater or lesser extent, from the true values that economic agents presumably act upon.

The effect that measurement errors have on the estimator depend on whether they happen in the dependent or independent variables.

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Measurement Errors

Assume the correct model is

$$Y = X\beta + U,$$

but we do not observe the regressand Y directly. Instead, we observe a noisy, badly measured, version of it.

Thus, we observe

$$Y^* = Y + V,$$

where V is the measurement error, that we assume is an IID term with variance σ_V^2.

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Measurement Errors

We estimate the model

$$Y^* = X\beta + U.$$

Substituting for Y^*

$$Y = X\beta + U - V = X\beta + W,$$

where $W = U - V$.

If we assume that the measurement error on the regressand is independent from the regressors, then $E[W|X] = E[U - V|X] = 0$.

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Measurement Errors

On the other hand, assume the correct model is

$$Y = X\beta + U,$$

but we do not observe the regressors X directly. Instead, we observe a noisy, badly measured, version of it.

Thus, we observe

$$X^* = X + V,$$

where V is the measurement error, that we assume is an IID term with variance σ_V^2.

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Measurement Errors

Thus, we estimate the model

$$Y = X^*\beta + U.$$

Substituting for X^*

$$Y = X\beta + V\beta + U = X\beta + W,$$

where $W = V\beta + U$.

In this case, note that $E[W|X] = -\beta V \neq 0$ [unless $\beta = 0$].

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Outline

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Simultaneous Equations

Economic theory often suggests that two or more endogenous variables are determined simultaneously.

For example, consider the price-quantity determined by the supply and demand equations given by

$$q_t = \gamma_d p_t + X_t^d \beta_d + u_t^d,$$

$$q_t = \gamma_s p_t + X_t^s \beta_s + u_t^s.$$

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Simultaneous Equations

The solution for this system is given by

$$\begin{bmatrix} q_t \\ p_t \end{bmatrix} = \begin{bmatrix} 1 & -\gamma_d \\ 1 & -\gamma_s \end{bmatrix}^{-1} \left(\begin{bmatrix} X_t^d \beta_d \\ X_t^s \beta_s \end{bmatrix} + \begin{bmatrix} u_t^d \\ u_t^s \end{bmatrix} \right).$$

Thus, $E[u_t^d | p_t] \neq 0$, $E[u_t^s | p_t] \neq 0$, $E[u_t^d | q_t] \neq 0, E[u_t^s | q_t] \neq 0$.

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Outline

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Non-normality

Asymptotic Theory

For tests for linear restrictions and we obtained their distributions, we typically assume that the errors follow a Normal distribution.

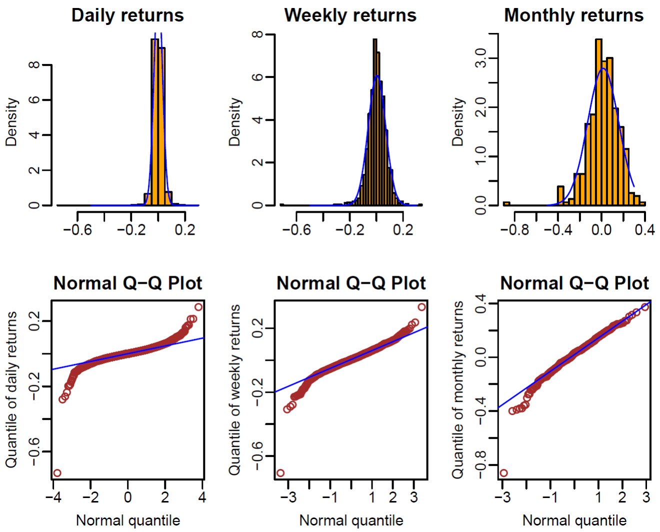
Hence, it can be shown that the tests follow the t or F distributions.

Nonetheless, the normality assumption may be a strong assumption to make for some economic data, particularly financial data.

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Non-normality
Asymptotic Theory



Histogram and Q-Q plots for returns on Apple stock (1985/1 to 2011/2).

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Luckily, the t and F tests that we discussed in the previous lecture are asymptotically valid under (reasonably well-behaved) non-normal errors.

Suppose we are interested in estimating the model

$$Y = X\beta + U,$$

where $U \sim IID(0, \sigma^2)$, $E[U_t|X_t] = 0$, and $E[U_t^2|X_t] = \sigma^2$.

Furthermore, assume that

$$\operatorname{plim} \frac{1}{n} X^T X = S_{XX}.$$

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

The key is to write the statistic as a function of quantities to which we can apply either a Law of Large Numbers or a Central Limit Theorem.
Recall that the t statistic for the test $\beta_2 = 0$ is given by

$$t_{\beta_2} = \left(\frac{Y^T M_X Y}{n-k}\right)^{-1/2} \frac{x_2^T M_{X_1} Y}{(x_2^T M_{X_1} x_2)^{1/2}}.$$

The second factor can be written as

$$\frac{x_2^T M_{X_1} Y}{(x_2^T M_{X_1} x_2)^{1/2}} = \frac{n^{-1/2} x_2^T M_{X_1} Y}{(n^{-1} x_2^T M_{X_1} x_2)^{1/2}},$$

and it can be shown using the LLN and CLT that this factor converge to a normal distribution.
Thus,

$$t_{\beta_2} \sim^a N(0,1).$$

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Similar arguments can be used to show that the numerator and denominator of the F statistic converge to the square of a Normal distribution, i.e., a chi-square distribution.

Thus, the F statistic indeed follows a F distribution asymptotically.

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Asymptotic tests rely on having an infinite sample, which of course is unfeasible with real data.

Thus, the finite sample distributions of the statistics just discussed differ in general from their asymptotic distributions.

We can use the bootstrap to approximate the finite sample distribution of the test statistics.

The errors committed by both asymptotic and bootstrap tests diminish as n increases, but those committed by bootstrap tests diminish more rapidly.

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Non-normality
Bootstrap Tests

The idea of the bootstrap is to generate more observations by resampling the errors. We then generate more data to construct more test statistics, and compare our original test statistic against the bootstrapped ones.

The bootstrap can be **parametric** if we assume a distribution for the errors. Then we simply draw new errors from that distribution.

Otherwise, they can be **semiparametric** if we do not assume a distribution and we simply generate the new errors by resampling with replacement from the original ones.

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Outline

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Homoskedasticity and autocorrelation

In the general case when the error terms are not assumed to be heteroskedastic and no autocorrelated, the variance matrix is given by

$$Var(U) = \Omega.$$

Thus, the variance of the estimator can be computed to be

$$Var(\hat{\beta}) = (X^T X)^{-1} X^T \Omega X (X^T X)^{-1}.$$

There is usually no way to know whether $s^2(X^T X)^{-1}$ is larger or smaller than the true variance of $\hat{\beta}$ above. Nonetheless, tests based on the former will be misleading.

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Outline

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Summing Up

- We have analysed some regression misspecifications and data-related problems.
- Outliers can wrongly influence the estimators.
- Collinearity can inflate their variance.
- Adding irrelevant regressors makes OLS inefficient, while failing to add relevant regressors makes them biased and inconsistent.
- Measurement errors in the regressand increases the variance, while in the regressors make the estimators biased.
- Simultaneity biases the estimators.
- In case of non-normal errors, asymptotic or bootstrapped tests can be used.
- Heteroskedasticity and autocorrelation produces misleading test statistics.

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark