

Data Analysis: The OLS Model

Mark Hendricks

August Review

UChicago Financial Mathematics

Outline

The Classic Model

Classic Inference

Large Sample Properties

Assumption: Full-rank

Assumption 1: $\mathbf{X}'\mathbf{X}$ is full rank.

Equivalently, assume that there is no exact linear relationship among any of the regressors.

- ▶ Clearly, the existence of OLS estimator requires that this assumption be satisfied.
- ▶ Multicollinearity refers to the case where this assumption fails.

Assumption: Exogeneity

Assumption 2: ϵ is exogenous to the regressors, \mathbf{x} .

$$\mathbb{E}[\epsilon | \mathbf{x}] = 0$$

The exogeneity assumption,

- ▶ implies that ϵ is uncorrelated with \mathbf{x} .
- ▶ implies that ϵ is uncorrelated with any function of \mathbf{x} .
- ▶ does NOT imply that ϵ is independent of \mathbf{x} .

Statistics as variables

To judge the OLS forecast, remember that

- ▶ A statistic is a function of random variables.
- ▶ Thus, a statistic is itself a random variable with a mean, variation, and distribution.
- ▶ A good statistic/forecast will be centered tightly around the true population value.

Unbiased statistics

An estimate is **unbiased** if its expectation equals the population value.

- ▶ Consider the sample estimator, \bar{x} , for a sample of n .
- ▶ Suppose we have a variable x with population mean μ_x ;
- ▶ Verify that \bar{x} is unbiased:

$$\begin{aligned}\mathbb{E}[\bar{x}] &= \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n x_i\right] \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}[x_i] \\ &= \mu_x\end{aligned}$$

Is OLS estimate unbiased?

Check if the OLS estimator is unbiased:

$$\begin{aligned}\mathbb{E}[\mathbf{b}] &= \mathbb{E}\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\right] \\ &= \mathbb{E}\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})\right] \\ &= \boldsymbol{\beta} + \mathbb{E}\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}\right]\end{aligned}$$

But $\mathbb{E}[\mathbf{b}] = \boldsymbol{\beta}$ if, and only if,

$$\mathbb{E}\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}\right] = \mathbf{0}$$

which is guaranteed by the exogeneity assumption but not simply by orthogonality.

Variance of the mean estimator

Continue with the example of the sample mean estimator, \bar{x} for sample size n .

- ▶ Suppose $\text{var}[x] = \sigma^2$.
- ▶ What is the variance of the sample mean estimator, \bar{x} ?

$$\text{var}[\bar{x}] = \frac{\sigma^2}{n}$$

Variance of OLS estimator

The variance of the OLS estimator is

$$\text{var}[\mathbf{b} | \mathbf{x}] = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\Sigma\mathbf{X} (\mathbf{X}'\mathbf{X})^{-1}$$

where $\Sigma = \mathbb{E}[\epsilon\epsilon' | \mathbf{x}]$.

- ▶ From above we have that $\mathbb{E}[\epsilon] = 0$.
- ▶ Thus, Σ is the variance of ϵ .
- ▶ So far we have used the first two moments of ϵ , along with its relation to \mathbf{x} , but no assumption on the distribution of ϵ has been made.

Heteroscedasticity and autocorrelation of residuals

- ▶ **Heteroscedasticity** refers to the case where the residuals but have distinct variances,

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

- ▶ **Autocorrelation** refers to the case where residuals are correlated. In this case, Σ is not diagonal.

Assumption: Homoscedastic and orthogonal residuals

Assumption 3:

The residuals are uncorrelated across observations, with identical variances,

$$\Sigma = \mathbb{E} [\epsilon \epsilon' | \mathbf{x}] = \sigma^2 \mathcal{I}_n$$

Gauss-Markov Theorem

With these assumptions, the OLS estimator, \mathbf{b} , is the minimum variance linear unbiased estimator of β .

- ▶ The assumption on Σ simplifies the variance of the OLS estimator,

$$\begin{aligned}\text{var}[\mathbf{b} | \mathbf{x}] &= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\sigma^2\mathcal{I}_n\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}\end{aligned}$$

- ▶ Any other linear, unbiased estimator of β will have larger variance.
- ▶ This is known as the Gauss-Markov theorem. It depends on the above assumptions regarding linearity, exogeneity, full-rank, and residual covariance structure.

Outline

The Classic Model

Classic Inference

Large Sample Properties

OLS Inference

The distribution of the OLS estimates is required in order to assess statistical significance.

- ▶ Above the mean and variance of \mathbf{b} were derived without making any distributional assumptions.

Assumption: Normality of residuals

Assumption 4: The residuals, ϵ are normally distributed.

$$\epsilon | \mathbf{x} \sim \mathcal{N}(\mathbf{0}, \Sigma)$$

Distribution of OLS estimator

Assumptions 1, 2, 3, 4 imply

$$\mathbf{b} | \mathbf{x} \sim \mathcal{N}(\boldsymbol{\beta}, \Omega)$$

where

$$\Omega = \sigma^2 (X'X)^{-1}$$

Often, these 4 assumptions are referred to as the **classical regression model**.

- Note that many results were derived without any distributional assumptions.

OLS Z-test

Testing the significance of an element of \mathbf{b} , would simply be a z-test:

$$\frac{b_j - \beta_j}{\sigma \omega_{jj}} \sim Z$$

where ω_{jj}^2 is the (j, j) element of $(\mathbf{X}'\mathbf{X})^{-1}$. Thus, $\sigma^2 \omega_{ij}^2$ is the (i, j) element of Ω .

- ▶ However, this statistic depends on σ , which is unknown.
- ▶ Instead one must use the sample estimate of the variance,

$$s^2 = \frac{\mathbf{e}'\mathbf{e}}{n - k - 1}$$

OLS t-test

The t-test assesses statistical significance of an element of \mathbf{b} ,

$$\frac{b_j - \beta_j}{s \omega_{jj}} \sim t(n - k - 1)$$

which depends only on observable data as well as the hypothesized value, β_j .

OLS F-test

An F-test will determine the joint significance of the linear regression:

$$\frac{R_{y,x}^2}{1 - R_{y,x}^2} \left(\frac{n - k - 1}{k} \right) \sim F(k, n - k - 1)$$

Namely, this tests whether all coefficients are jointly equal to zero. (We are assuming the regression includes a constant.)

- Note the use of the R-squared stat.
- It is simple to generalize this to test whether **b** is jointly equal to a non-zero hypothesis vector, **β^*** .

Problems with multicollinearity

If regressor j is highly correlated with the other regressors, then the variance of the coefficient estimate can be written as

$$\text{var}[b_j | \mathbf{x}] = \left(\frac{1}{1 - R_{x_j, \mathbf{x}_{[-j]}}^2} \right) \frac{\sigma^2}{\sum_{i=1}^n (\mathbf{x}_{i,j} - \bar{\mathbf{x}}_j)^2}$$

where

$$R_{x_j, \mathbf{x}_{[-j]}}^2$$

denotes the R-squared from regressing x_j on the remaining regressors, (all columns except column j of \mathbf{X} .)

Variance inflation factor

The term

$$\frac{1}{1 - R_{x_j, \mathbf{x}_{[-j]}}^2}$$

is known as the **variance inflation factor**.

- ▶ The VIF is closely related to the condition number of $\mathbf{X}'\mathbf{X}$.
- ▶ The condition number in linear algebra measures the sensitivity of inverting a matrix.
- ▶ It compares the largest and smallest eigenvalues.
- ▶ Most software packages will warn the user if the condition number (VIF) is too big.

Example of multicollinearity

Consider a regression examining how the unemployment rate responds to the short and long ends of the U.S. Treasury yield curve.

- ▶ Such a regression would be at least a crude attempt to think about the dual mandate of the Fed.
- ▶ They are supposed to balance a stable money supply against stimulating full employment.
- ▶ Monetary stimulation could effect the yield curve and unemployment.

Regressing with multicollinearity

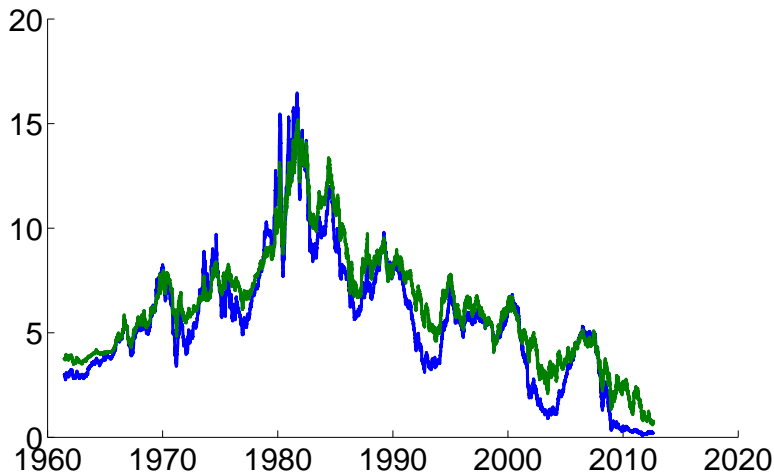


Figure: $VIF=7.6$ Condition of $X'X = 13.7$

Warning: Correlation of interest rates.

This is an unsophisticated model, but it serves as a basic warning.

- ▶ Many applications use interest rates as explanatory variables.
- ▶ However, many different rates are highly correlated.
- ▶ Recall that multicollinearity decreases confidence in the OLS estimate.

Variation in regressors

The sample variance of x_j reduces the variance of the OLS estimate b_j .

- ▶ Again write

$$\text{var}[b_j | \mathbf{x}] = \left(\frac{1}{1 - R_{x_j, \mathbf{x}_{[-k]}}^2} \right) \frac{\sigma^2}{\sum_{i=1}^n (\mathbf{x}_{i,j} - \bar{\mathbf{x}}_j)^2}$$

- ▶ The denominator of the second term is the scaled sample variance.
- ▶ If there is not enough variation in the regressor data, then the OLS estimation can not precisely estimate β_j .

Example: Considering variation in the regressor

The standard error of the OLS estimator depends on the variation in the regressors.

- ▶ The standard error of b decreases as the variation in X increases, holding other things equal.
- ▶ Recall that net interest margin refers to the spread in the lending and borrowing of banks.
- ▶ Consider using this as a regressor, for some financial or economic data.

Outline

The Classic Model

Classic Inference

Large Sample Properties

Is OLS robust?

How good is OLS if the assumptions do not hold?

- ▶ Financial data is usually non-normal—violating Assumption 4.
- ▶ Time-series models will almost always violate exogeneity—Assumption 2.
- ▶ Macro-economic data typically has correlated residuals, while asset prices show time-varying volatility—violations of Assumption 3.

OLS corrections

Two main ways to address these problems:

- ▶ Large sample properties. (Relax assumptions 2, 4.)
- ▶ Robust standard errors. (Relax assumption 3.)

Instrumental Variable Regression (IV) is also very important in dealing with assumption 2, but will not be discussed here.

Non-normality

Applications often do not satisfy **Assumption 4**, upon which the inference results relied.

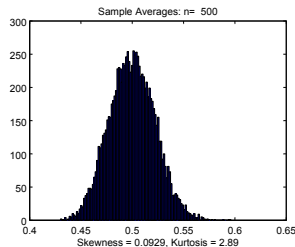
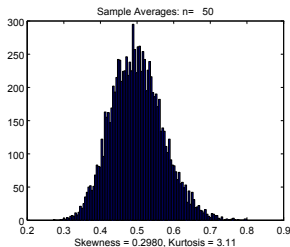
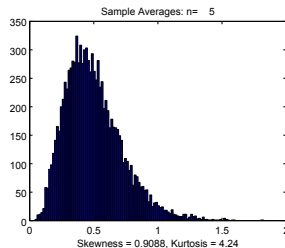
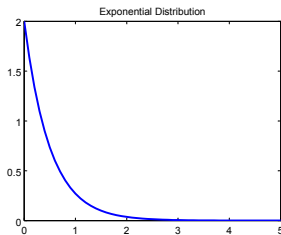
- ▶ However, the asymptotic distribution of the OLS estimate is an application of the Central Limit Theorem.
- ▶ In practice, inference often relies on having large data sets and appealing to the asymptotic results.

Central Limit Theorem

A reminder:

- ▶ As sample size increases, the sample average statistic converge to a normal distribution.
- ▶ Slightly more complicated for non-iid data, but weaker versions hold.
- ▶ Note that the OLS estimator can be rewritten as a sample average of ϵ , so we can apply the CLT!

Example - Central Limit Theorem



Assumption: Orthogonality of population residuals

Assumption 5: The population residuals are uncorrelated with the regressors.

$$\mathbb{E} [\mathbf{x}'\epsilon] = \mathbf{0}$$

- ▶ This assumption is much weaker than **Assumption 2**.
- ▶ This is a restriction on the population variables, not the fitted estimates, which have zero correlation by construction.

Consistency

A sample statistic is **consistent** if it converges to the true population value in probability.

- ▶ Suppose that **Assumptions 1, 5** hold.
- ▶ Then the OLS estimator, **b** is consistent,

$$\text{plim } \mathbf{b} = \boldsymbol{\beta}$$

- ▶ In practice, more attention is paid to having a consistent estimator than an unbiased estimator, due to the weaker assumption.

Asymptotic distribution of OLS

Under **Assumptions 1,3, 5**, the OLS estimate is asymptotically normal,

$$\mathbf{b} | \mathbf{x} \sim^{\text{asym}} \mathcal{N}(\boldsymbol{\beta}, \Omega)$$

where

$$\Omega = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$$

Heteroscedastic and autocorrelated inference

For many applications, particularly in time-series, **Assumption 3** is clearly false.

- For practical purposes, this is not a big problem for inference.

Under **Assumptions 1, 5**, the OLS estimate is asymptotically normal,

$$\mathbf{b} | \mathbf{x} \sim^{\text{asym}} \mathcal{N}(\boldsymbol{\beta}, \Omega)$$

where

$$\Omega = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\Sigma\mathbf{X} (\mathbf{X}'\mathbf{X})^{-1}$$

OLS without iid errors

With non-iid errors, OLS is still unbiased (or consistent).

- ▶ Thus, it is appropriate to estimate with OLS, but one must use the larger variance given by

$$\text{var}[\mathbf{b} | \mathbf{x}] = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\Sigma\mathbf{X} (\mathbf{X}'\mathbf{X})^{-1}$$

- ▶ Non-OLS estimators, such as GLS, may have lower variances which allow for more confident inference.

References

- ▶ Cochrane, John. *Asset Pricing*. 2001.
- ▶ Greene, William. *Econometric Analysis*. 2011.
- ▶ Hamilton, James. *Time Series Analysis*. 1994.
- ▶ Wooldridge, Jeffrey. *Econometric Analysis of Cross Section and Panel Data*. 2011.