Summary of Results From Classical Linear Regression[1]

Algorithmic Trading and Quantitative Strategies

Petter Kolm

January 24, 2023

 If you find any typos, I would much appreciate you email them to me.

Readings

Most of this material has been covered in other statistics and data science courses you have taken. So you should be pretty familiar with it already.

In addition to the lecture notes, the following is required reading in Hayashi (2000):

- Chapter 1.1-1.6; and
- Chapter 2.3-2.5, 2.9.[2]

[2] You need to know the results from these sections in Chapter 2. You are not required to know the proofs. By the way, questions based on this material are very common in quant interviews.

Linear Models I

In this class we will build risk models based on a the linear *contemporaneous regression* specification

$$r_{i,t} = \alpha_i + \sum_{k=1}^{K} \beta_{ik} f_{k,t} + \varepsilon_{i,t} \,. \tag{1}$$

We will also discuss forecasting models based on linear the *time-series regression* (or *forecasting regression*) specification

$$r_{i,t+1} = \alpha_i + \sum_{k=1}^{K} \beta_{ik} f_{k,t} + \varepsilon_{i,t+1} \tag{2}$$

Of course, the choice of the factors $f_{k,t}$ is important. We could allow for time-varying parameters, i.e. $\alpha_{i,t}$, $\beta_{ik,t}$. Classical linear regression is a building block for many of these types of models.

The Classical Linear Regression Model

The Classical Linear Model I

We briefly review the classical approach to the linear regression problem

$$y = \mathbf{x}'\boldsymbol{\beta} + \varepsilon\,, \tag{3}$$

where $y \in \mathbb{R}$ and $\varepsilon \in \mathbb{R}$ are random variables, $\mathbf{x} \in \mathbb{R}^p$ is a random vector and $\boldsymbol{\beta} \in \mathbb{R}^p$ are parameters.

y is referred to as the dependent variable (or left-hand side variable),

$\mathbf{x}$ are the explanatory variables (right-hand side variables or regressors),

$\boldsymbol{\beta}$ are the regression coefficients, and

ε is referred to as the unobserved error (or residual).

The Classical Linear Model II

We assume that we have collected a dataset of the realizations of dependent and explanatory variables,

$$\mathcal{D} := (\mathbf{X}, \mathbf{y}) = \left\{(\mathbf{x}_i', y_i)\right\}_{i=1}^{n} \tag{4}$$

with $\mathbf{X} \in \mathbb{R}^{n \times p}$ and $\mathbf{y} \in \mathbb{R}^{n}$.

Our goal is to determine the parameters $\boldsymbol{\beta}$ that provide the "best fit" for the "model" given our data $\mathcal{D}$, so that

$$\mathbf{y} \approx \mathbf{X}\boldsymbol{\beta}\,. \tag{5}$$

Model Assumptions

Using the notation introduced above, for the classical linear regression model we make the following assumptions

I. *Linearity:*

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\,. \tag{6}$$

II. *Strict Exogeneity:*

$$\mathbb{E}\left[\boldsymbol{\varepsilon}|\mathbf{X}\right] = \mathbf{0}\,. \tag{7}$$

III. *No Multicollinearity:*

$$\mathbb{P}\left[\operatorname{rank}(\mathbf{X}) = p\right] = 1\,. \tag{8}$$

IV. *Spherical Errors:*

$$\operatorname{var}\left[\boldsymbol{\varepsilon}|\mathbf{X}\right] = \sigma^2\mathbf{I}_n\,. \tag{9}$$

V. *Normality:*

$$\boldsymbol{\varepsilon}|\mathbf{X} \sim \mathcal{N}(0, \sigma^2\mathbf{I}_n)\,. \tag{10}$$

Ordinary Least Squares I

Ordinary least squares (OLS) seeks the parameter estimate $\widehat{\boldsymbol{\beta}}$ that minimizes the *sum of squared residuals* (SSR)

$$\mathrm{SSR}(\boldsymbol{\beta}) := \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right)^2 . \tag{11}$$

As $\mathrm{SSR}(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$, we define our estimate via

$$\widehat{\boldsymbol{\beta}} := \mathrm{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 , \tag{12}$$

which can be shown to have the solution

$$\widehat{\boldsymbol{\beta}} := \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{y} . \tag{13}$$

Additional Concepts

We define the *residuals* as

$$\mathbf{r} := \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\,. \tag{14}$$

Note that

$$\mathrm{SSR} = \mathbf{r}'\mathbf{r} \tag{15}$$

Finally, in practice σ^2 is unknown and needs to be estimated. We know that

$$s^2 := \frac{\mathbf{r}'\mathbf{r}}{n-p} \tag{16}$$

is an unbiased estimator of σ^2.

Finite Sample Properties of OLS I

Proposition (Finite-sample properties of OLS)

1. Under assumptions (I)-(III), we have that $\mathbb{E}[\widehat{\boldsymbol{\beta}}|\mathbf{X}] = \boldsymbol{\beta}$.
2. Under assumptions, (I)-(IV), we have that
   (a) $\mathrm{var}[\widehat{\boldsymbol{\beta}}|\mathbf{X}] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$,
   (b) $\widehat{\boldsymbol{\beta}}$ is the best linear unbiased estimator (the Gauss-Markov Theorem),
   (c) $\mathrm{cov}[\widehat{\boldsymbol{\beta}}, \mathbf{r}|\mathbf{X}] = 0$, and
   (d) $\mathbb{E}[s^2|\mathbf{X}] = \sigma^2$.
3. If σ^2 is unknown, then an estimate of $\mathrm{var}[\widehat{\boldsymbol{\beta}}|\mathbf{X}]$ is given by

$$\widehat{\mathrm{var}[\widehat{\boldsymbol{\beta}}|\mathbf{X}]} = s^2 \cdot \left(\mathbf{X}'\mathbf{X}\right)^{-1} .$$

Hypothesis Testing under Normality

The t-test I

We want to test the null hypothesis

$$H_0 : \beta_j = b_j\,, \tag{17}$$

where $b_j \in \mathbb{R}$ is some known value versus the alternative hypothesis

$$H_a : \beta_j \neq b_j\,. \tag{18}$$

The t-test II

We define the t-ratio (or t-statistic)

$$
\begin{aligned}
t_j &:= \frac{\widehat{\beta}_j - b_j}{\sqrt{s^2\left((\mathbf{X}'\mathbf{X})^{-1}_{jj}\right)}} \\
&= \frac{\widehat{\beta}_j - b_j}{\mathrm{SE}(\widehat{\boldsymbol{\beta}}_j)}\,,
\end{aligned}
$$

where $\mathrm{SE} := \sqrt{s^2\left((\mathbf{X}'\mathbf{X})^{-1}_{jj}\right)}$ is referred to as the *standard error* of the OLS estimate $\widehat{\beta}_j$, and s^2 is an estimate of σ^2 (that is typically unknown)

The t-test III

Recall that under the classical linear regression assumptions (I)-(V) the t-ratio follows a t-distribution with $(n - p)$ degrees of freedom under the null hypothesis, that is

$$t_j | \mathbf{X}, H_0 \sim t_{n-p} \, .$$

The t-test IV

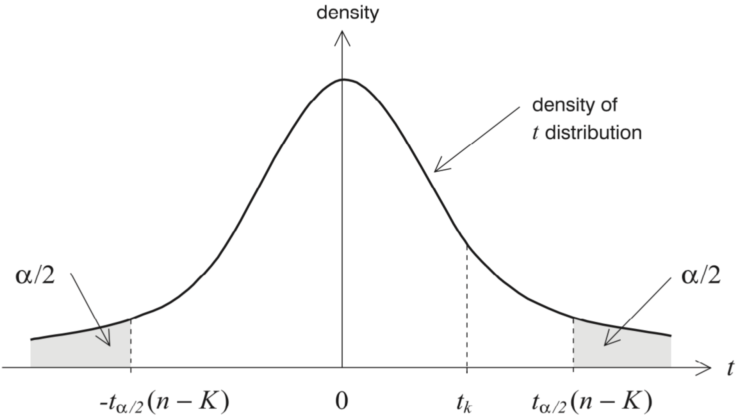

Figure 1: Illustration of how to perform a t-test at significance level α. (Source: Hayashi (2000)).

The F-test

We want to test the null hypothesis

$$H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{d}\,, \tag{19}$$

where $\mathbf{R} \in \mathbb{R}^{d \times p}$ with rank d, and $\mathbf{d} \in \mathbb{R}^d$ are known. We define

the F-statistic

$$F := \frac{(\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{d})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{d})/d}{s^2}\,. \tag{20}$$

Under the classical linear regression assumptions (I)-(V) the F-statistic follows an F-distribution with $(d, n-p)$ degrees of freedom under the null hypothesis, that is

$$F|\mathbf{X}, H_0 \sim F_{(d,n-p)}\,.$$

Large Sample Theory for Linear Regression

Revisiting Some of the Assumptions of CLR I

Two (very) restrictive assumptions of CLR are:

- strict exogeneity,
- normality.

In financial applications we frequently encounter time series and panel data regression models. Here CLR assumptions are not satisfied as frequently errors are (a) not strictly exogenous[3] and (b) serially correlated.

Revisiting Some of the Assumptions of CLR II

First, let us recall why we need these assumptions:

- Strict exogeneity is needed for unbiasedness.
- Normality is used for constructing hypothesis tests, confidence intervals and regions, etc.

One way it to let the number of observations become large (i.e. $n \to \infty$ in the limit). This is referred to as the *large sample theory* of linear regression.

As the results are important for empirical work and modeling of financial and economic data, we will summarize the results and discuss them from an intuitive perspective. If you are interested, Hayashi (2000) does a good job describing the details.

[3] For them to be strictly exogenous they need to be orthogonal to *past, current and future regressors*.

Large Sample Assumptions for Linear Regression I

We will make the following assumptions:

I'. *Linearity:*

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i\,. \tag{21}$$

II'. *Stationarity and IID:* The $(p+1)$-dimensional vector stochastic process $\{y_i, \mathbf{x}_i\}$ is jointly stationary, independent and identically distributed (IID).[4]

III'. *Predetermined Regressors*: All regressors are are orthogonal to the contemporaneous error term

$$\mathbb{E}[\mathbf{x}_i \varepsilon_i] = \mathbf{0}\,. \tag{22}$$

IV'. *Rank Condition:* The matrix $\boldsymbol{\Sigma}_{\mathbf{xx}} := \mathbb{E}[\mathbf{x}_i \mathbf{x}_i'] \in \mathbb{R}^{p \times p}$ is invertable.

V'. *Conditional Expectation of Errors:*

$$\mathbb{E}[\varepsilon_i | \varepsilon_{i-1}, \ldots, \varepsilon_1, \mathbf{x}_i, \ldots, \mathbf{x}_1,] = 0\,. \tag{23}$$

VI'. *Finite 4th Moment for Regressors:* $\mathbb{E}[(x_{ik}x_{ij})^2] < \infty$ for all k, j.

Large Sample Assumptions for Linear Regression II

Comments:

- Note that

$$\mathbf{\Sigma_{xx}} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' .$$

- We define the finite sample approximation of $\mathbf{\Sigma_{xx}}$ as

$$\mathbf{S_{xx}} := \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' .$$

[4] To be more general, we can replace IID with ergodic and all results shown here will still hold.

Large Sample Properties I

Proposition (Large Sample Properties)

1. Under assumptions (I')-(IV'), $\widehat{\boldsymbol{\beta}}$ is a consistent estimator for $\boldsymbol{\beta}$.
2. Under assumptions, (III') and (V'), we have that

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \underset{d}{\rightarrow} \mathcal{N}(\mathbf{0}, \mathrm{avar}(\widehat{\boldsymbol{\beta}})) \text{ as } n \rightarrow \infty$$

where

$$\mathrm{avar}(\widehat{\boldsymbol{\beta}}) := \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1} \mathbf{S} \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}$$

with $\boldsymbol{\Sigma}_{\mathbf{xx}} := \mathbb{E}[\mathbf{x}_i \mathbf{x}_i']$ and $\mathbf{S} := \mathbb{E}[\varepsilon_i^2 \mathbf{x}_i \mathbf{x}_i']$.

Large Sample Properties II

3. Under assumptions (I'), (II') and (VI'), a consistent estimator of $\mathbf{S}$ is given by
$$\widehat{\mathbf{S}} := \frac{1}{n}\sum_{i=1}^{n} \widehat{\varepsilon}_i^{\,2}\mathbf{x}_i\mathbf{x}_i' ,$$
where $\widehat{\varepsilon}_i := y_i - \mathbf{x}_i'\widehat{\beta}$.
4. Under assumptions (I')-(VI'), a consistent estimator of $\mathrm{avar}[\widehat{\boldsymbol{\beta}}]$ is given by
$$\widehat{\mathrm{avar}[\widehat{\boldsymbol{\beta}}]} = \mathbf{S}_{\mathbf{xx}}^{-1}\widehat{\mathbf{S}}\mathbf{S}_{\mathbf{xx}}^{-1}$$
where
$$\mathbf{S}_{\mathbf{xx}} := \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i' = \frac{1}{n}\mathbf{X}'\mathbf{X} .$$

The Robust t-test I

We want to test the null hypothesis

$$H_0 : \beta_j = b_j \, , \tag{24}$$

where $b_j \in \mathbb{R}$ is some known value versus the alternative hypothesis

$$H_a : \beta_j \neq b_j \, . \tag{25}$$

The Robust t-test II

From the Proposition above we have that under the null

$$\sqrt{n}\left(\widehat{\beta}_j - b_j\right) \mid H_0 \underset{d}{\rightarrow} \mathcal{N}\left(0, \mathrm{avar}[\widehat{\beta}_j]\right), \tag{26}$$

$$\widehat{\mathrm{avar}[\widehat{\beta}_j]} \mid H_0 \underset{p}{\rightarrow} \mathrm{avar}[\widehat{\beta}_j]. \tag{27}$$

From here on, we will drop the conditioning on H_0 to keep the notation more compact.

It follows that

$$t_j := \frac{\sqrt{n}\left(\widehat{\beta}_j - b_j\right)}{\sqrt{\widehat{\mathrm{avar}[\widehat{\beta}_j]}}} = \frac{\widehat{\beta}_j - b_j}{\mathrm{SE}^*(\widehat{\beta}_j)} \underset{d}{\rightarrow} \mathcal{N}(0,1)$$

where

$$\mathrm{SE}^*(\widehat{\beta}_j) := \sqrt{\frac{1}{n}\widehat{\mathrm{avar}[\widehat{\beta}_j]}} = \sqrt{\frac{1}{n}\left(\mathbf{S}_{\mathbf{xx}}^{-1}\widehat{\mathbf{S}}\mathbf{S}_{\mathbf{xx}}^{-1}\right)_{jj}}.$$

The Robust t-test III

$\mathrm{SE}^*(\widehat{\beta}_j)$ is referred to as the *heteroskedasticity consistent standard error*, *robust standard error*, or *Eicker-Huber-White standard error*[5]. Of course, the reason for this terminology is that the error term can be conditionally heteroskedastic (remember we did not assume conditional homoskedasticity).

We will refer to this t-ratio as the *robust t-ratio* in order to distinguish it from the t-ratio in the finite sample case.

The Robust t-test IV

Key differences between the robust t-ratio and t-ratio:

- The formulas for the standard errors are different. (For you: If errors are conditionally homoskedastic, how do the standard errors compare?)
- To perform hypothesis tests and construct confidence intervals, we use the normal distribution rather than a t-distribution.
- Note that for finite n the robust t-ratio is approximately normally distributed.

[5] After the original authors Eicker (1967), Huber (1967), and White (1980).

The Wald Statistic ("A Robust Version of the F-Test") I

We want to test the null hypothesis

$$H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{c}\,, \tag{28}$$

where $\mathbf{R} \in \mathbb{R}^{k \times p}$ of full rank k, and $\mathbf{c} \in \mathbb{R}^d$ are known. We define

the *Wald statistic*

$$W := n \cdot (\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{c})' \left(\mathbf{R} \cdot \widehat{\mathrm{avar}[\widehat{\boldsymbol{\beta}}]} \cdot \mathbf{R}' \right)^{-1} (\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{c})\,. \tag{29}$$

One can show that

$$W \underset{d}{\rightarrow} \chi^2_c\,, \tag{30}$$

from which hypothesis testing and construction of confidence regions immediately follow.

The Wald Statistic ("A Robust Version of the F-Test") II

Compare the formulas for the F- and Wald statistics

$$F = \frac{(\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{c})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{c})/k}{s^2}, \tag{31}$$

$$W = n \cdot (\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{c})' \left(\mathbf{R} \cdot \widehat{\mathrm{avar}[\widehat{\boldsymbol{\beta}}]} \cdot \mathbf{R}' \right)^{-1} (\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{c}). \tag{32}$$

The differences between the two tests are similar to our discussion of the differences between the t-ratios.

References

Eicker, Friedhelm (1967). “Limit Theorems for Regressions with Unequal and Dependent Errors”. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Vol. 1. 1. Berkeley, CA: University of California Press, pp. 59–82.

Hayashi, Fumio (2000). *Econometrics*. Princeton University Press.

Huber, Peter J. (1967). “The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions”. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Vol. 1. 1. University of California Press, pp. 221–233.

White, Halbert (1980). “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity”. In: *Econometrica: journal of the Econometric Society*, pp. 817–838.