Classical Linear Regression: Extensions and Generalizations

Machine Learning & Computational Statistics

Ivailo Dimov

February 5, 2026

Readings

In addition to the lecture notes, the following are required readings in Hayashi (2000):

- Chapter 1.5[1], 1.6;
- Chapter 2.3-2.5, 2.9.[2]

[1] Time permitting, we will talk about this during class.

[2] You should learn the results from these sections in Chapter 2. You are not required to learn the proofs.

Review

Model Assumptions I

Recall the assumptions for the classical linear regression model

I. *Linearity:*

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\,. \tag{1}$$

II. *Strict Exogeneity:*

$$\mathbb{E}\left[\boldsymbol{\varepsilon}|\mathbf{X}\right] = \mathbf{0}\,. \tag{2}$$

III. *No Multicollinearity:*

$$\mathbb{P}\left[\mathrm{rank}(\mathbf{X}) = p\right] = 1\,. \tag{3}$$

IV. *Spherical Errors:*

$$\mathrm{var}\left[\boldsymbol{\varepsilon}|\mathbf{X}\right] = \sigma^2\mathbf{I}_n\,. \tag{4}$$

V. *Normality:*

$$\boldsymbol{\varepsilon}|\mathbf{X} \sim \mathcal{N}(0, \sigma^2\mathbf{I}_n)\,. \tag{5}$$

Predictions and Prediction Intervals for Classical Linear Regression

Predictions and Prediction Intervals for Classical Linear Regression I

Suppose we have estimated $\widehat{\beta}$ from observed data $\mathcal{D}$. We are interested in making predictions

$$\mathbf{X}_{\text{new}} \mapsto \mathbf{y}_{\text{new}} \, , \tag{6}$$

where $\mathbf{X}_{\text{new}} \in \mathbb{R}^{k \times p}$ is the new data matrix (consisting of observations that were not used in the estimation of $\widehat{\beta}$) and construct predictions intervals.

Typically we do not observe the new realized values of $\mathbf{y}_{\text{new}}$ so we estimate them from $\widehat{\beta}$ and the new data points $\mathbf{X}_{\text{new}}$.

Predictions and Prediction Intervals for Classical Linear Regression II

A natural estimate of $\mathbf{y}_{\text{new}}$ is

$$\widehat{\mathbf{y}}_{\text{new}} := \mathbf{X}_{\text{new}}\widehat{\boldsymbol{\beta}}\,, \tag{7}$$

with the property that

$$\begin{aligned}\mathbb{E}[\widehat{\mathbf{y}}_{\text{new}}|\mathbf{X}] &= \mathbb{E}[\mathbf{X}_{\text{new}}\widehat{\boldsymbol{\beta}}|\mathbf{X}] \\ &= \mathbf{X}_{\text{new}}\boldsymbol{\beta}\,.\end{aligned}$$

Predictions and Prediction Intervals for Classical Linear Regression III

How do we construct confidence intervals for the elements of $\widehat{\mathbf{y}}_{\text{new}}$?

Let us start with determining the variance of our estimator

$$\begin{aligned}\text{var}[\widehat{\mathbf{y}}_{\text{new}}|\mathbf{X}] &= \text{var}[\mathbf{X}_{\text{new}}\widehat{\boldsymbol{\beta}}|\mathbf{X}] \\ &= \mathbf{X}_{\text{new}}\text{var}[\widehat{\boldsymbol{\beta}}|\mathbf{X}]\mathbf{X}'_{\text{new}} \\ &= \sigma^2\mathbf{X}_{\text{new}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'_{\text{new}}\,.\end{aligned}$$

This suggests the following confidence interval for the mean of $\mathbf{y}_{\text{new}}$

$$\mathbb{E}[\widehat{\mathbf{y}}_{\text{new}}|\mathbf{X}] \in \mathbf{X}_{\text{new}}\widehat{\boldsymbol{\beta}} \pm t_{n-p,\alpha/2} \cdot \sqrt{s^2\,\text{diag}\left(\mathbf{X}_{\text{new}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'_{\text{new}}\right)}\,,$$

where $t_{n-p,\alpha/2}$ is the critical value and s^2 is our unbiased estimate of σ^2.

Predictions and Prediction Intervals for Classical Linear Regression IV

In practice, we are often interested in an interval for the prediction error

$$\mathbf{e} := \mathbf{y}_{\text{new}} - \widehat{\mathbf{y}}_{\text{new}}\,.$$

We compute the variance of the prediction error

$$\begin{aligned}\text{var}[\mathbf{y}_{\text{new}} - \widehat{\mathbf{y}}_{\text{new}}|\mathbf{X}] &= \text{var}[\mathbf{y}_{\text{new}}|\mathbf{X}] + \text{var}[\widehat{\mathbf{y}}_{\text{new}}|\mathbf{X}] - 2\text{cov}[\mathbf{y}_{\text{new}}, \widehat{\mathbf{y}}_{\text{new}}|\mathbf{X}] \\ &= \sigma^2\mathbf{I}_k + \sigma^2\mathbf{X}_{\text{new}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'_{\text{new}} \\ &= \sigma^2\left(\mathbf{I}_k + \mathbf{X}_{\text{new}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'_{\text{new}}\right)\,,\end{aligned}$$

where k is the number of rows of $\mathbf{X}_{\text{new}}$.

Predictions and Prediction Intervals for Classical Linear Regression V

Why is $\operatorname{cov}[\mathbf{y}_{\mathrm{new}}, \widehat{\mathbf{y}}_{\mathrm{new}} | \mathbf{X}] = 0$?To see this, let $\mathbf{y}_{\mathrm{new}} := \mathbf{X}_{\mathrm{new}} \boldsymbol{\beta} + \widetilde{\boldsymbol{\varepsilon}}$.

Then

$$\begin{aligned} \operatorname{cov}[\mathbf{y}_{\mathrm{new}}, \widehat{\mathbf{y}}_{\mathrm{new}} | \mathbf{X}] &= \operatorname{cov}[\widetilde{\boldsymbol{\varepsilon}}, \mathbf{X}_{\mathrm{new}} \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{y} | \mathbf{X}] \\ &= 0\,, \end{aligned}$$

since $\widetilde{\boldsymbol{\varepsilon}}$ is independent of $\mathbf{y}$.

Predictions and Prediction Intervals for Classical Linear Regression VI

This gives us the *prediction intervals* for $\mathbf{y}_{\text{new}}$

$$\mathbf{y}_{\text{new}}|\mathbf{X} \in \mathbf{X}_{\text{new}}\widehat{\boldsymbol{\beta}} \pm t_{n-p,\alpha/2} \cdot \sqrt{s^2 \,\text{diag}\left(\mathbf{I}_k + \mathbf{X}_{\text{new}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_{\text{new}}'\right)},$$

where $t_{n-p,\alpha/2}$ and s^2 are defined as above. Observe that this prediction interval is exactly a factor of s wider than the confidence interval determined above.

Predictions and Prediction Intervals for Classical Linear Regression VII

Remark

The prediction interval is affected not only by the variance of the true $\mathbf{y}_{\text{new}}$ (due to random shocks), but also by the variance of its mean since the parameters $\widehat{\boldsymbol{\beta}}$ in general are imprecise and hence have non-zero variance. In other words, the prediction interval combines the uncertainty in parameter estimates *and* the uncertainty coming from the randomness in a new observation. Therefore, it is natural that a prediction interval will be wider than a confidence interval.

For you: What does the prediction region look like when homoscedasticity is no longer valid?

The Generalized Linear Regression Model and Generalized Least Squares

The Heteroscedastic Error Assumption

Up to this point we have assume errors are homoscedastic (spherical). We will now discuss some of the consequences that follow when this assumption no longer holds.

Specifically, we assume the errors are heteroscedastic

$$\mathrm{var}\left[\boldsymbol{\varepsilon}|\mathbf{X}\right] = \sigma^2\mathbf{V}(\mathbf{X}), \tag{8}$$

where $\mathbf{V}(\mathbf{X}) \in \mathbb{R}^{n \times n}$ is invertable and symmetric. To simplify notation we will abbreviate $\mathbf{V} \equiv \mathbf{V}(\mathbf{X})$. The resulting regression model is referred to as the *generalized linear regression model* (GLR).

The Generalized Linear Regression Model I

We summarize the main results of GLR:

- The Gauss-Markov Theorem no longer holds for the OLS estimator. Thus, OLS is no longer BLUE.
- The t-ratio is no longer distributed as the t-distribution. Thus, the t-test is no longer valid.
- The F-statistic is no longer distributed as the F- distribution. Thus, the F-test is no longer valid.
- However, the OLS estimator is still unbiased.[3]

[3] Recall that unbiasedness does not require homoscedasticity.

Estimation With Known $\mathbf{V}$: Generalized Least-Squares I

Using Cholesky we decompose

$$\mathbf{V}^{-1} = \mathbf{C}'\mathbf{C}\,. \tag{9}$$

Left multiplying the regression problem

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\,, \tag{10}$$

by $\mathbf{C}$, we obtain

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

We can write this as

$$\widetilde{\mathbf{y}} = \widetilde{\mathbf{X}}\boldsymbol{\beta} + \widetilde{\boldsymbol{\varepsilon}} \tag{11}$$

where $\widetilde{\mathbf{y}} := \mathbf{C}\mathbf{y}$, $\widetilde{\mathbf{X}} := \mathbf{C}\mathbf{X}$ and $\widetilde{\boldsymbol{\varepsilon}} := \mathbf{C}\boldsymbol{\varepsilon}$.

Estimation With Known $\mathbf{V}$: Generalized Least-Squares II

We see that the covariance matrix of the transformed error terms are spherical, that is

$$\begin{aligned}\operatorname{var}[\widetilde{\varepsilon}|\mathbf{X}] &= \mathbb{E}(\widetilde{\varepsilon}\widetilde{\varepsilon}'|\mathbf{X}) \\ &= \mathbb{E}[\mathbf{C}\varepsilon\varepsilon'\mathbf{C}'|\mathbf{X}] \\ &= \mathbf{C}\,\mathbb{E}[\varepsilon\varepsilon'|\mathbf{X}]\mathbf{C}' \\ &= \sigma^2\mathbf{C}\mathbf{V}\mathbf{C}' \\ &= \sigma^2\mathbf{I}_n\end{aligned}$$

For you: Verify that the assumptions (I)-(IV) of the CLR are satisfied.

Estimation With Known $\mathbf{V}$: Generalized Least-Squares III

Computing the estimator of the transformed model, we obtain

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \left(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}\right)^{-1}\widetilde{\mathbf{X}}'\widetilde{\mathbf{y}} \\
&= \left((\mathbf{CX})'(\mathbf{CX})\right)^{-1}(\mathbf{CX})'\mathbf{Cy} \\
&= \left(\mathbf{X}'\mathbf{C}'\mathbf{CX}\right)^{-1}\left(\mathbf{X}'\mathbf{C}'\mathbf{Cy}\right) \\
&= \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}\,.
\end{aligned}$$

This is referred to as the *generalized least squares* (GLS) estimator.

Let us also compute the conditional variance of the GLS estimator

$$\begin{aligned}
\mathrm{var}[\widehat{\boldsymbol{\beta}} \mid \mathbf{X}] &= \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\cdot\mathrm{var}[\mathbf{y}\mid\mathbf{X}]\cdot\mathbf{V}^{-1}\mathbf{X}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1} \\
&= \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\cdot\left(\sigma^2\mathbf{V}\right)\cdot\mathbf{V}^{-1}\mathbf{X}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1} \\
&= \sigma^2\cdot\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\,.
\end{aligned}$$

Estimation With Known $\mathbf{V}$: Generalized Least-Squares IV

The direct application of the Gauss-Markov Theorem to our transformed model leads to the following result:

Proposition

Generalized least squares with known covariance matrix is BLUE.

Estimation With Known $\mathbf{V}$: Generalized Least-Squares V

Note that the transformation we used above does not effect $\boldsymbol{\beta}$. The parameter vector is exactly the same. We have only transformed the data for the purpose of applying OLS. Therefore $\widehat{\boldsymbol{\beta}}$ and s^2 can be taken directly from the model fit of the transformed versions of the variables.

For example, we can compute prediction intervals as follows

$$\mathbf{y}_{\mathrm{new}}|\mathbf{X} \in \mathbf{X}_{\mathrm{new}}\widehat{\boldsymbol{\beta}} \pm t \cdot \sqrt{s^2 \,\mathrm{diag}\left(\mathbf{V}_{\mathrm{new}} + \mathbf{X}_{\mathrm{new}}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'_{\mathrm{new}}\right)}\,,$$

where t is the appropriate critical value.

Observe that we only need the diagonal of $\mathbf{V}_{\mathrm{new}}$ here. For prediction, we do not care about the covariance between predictions. Only the raw variances matter, and they can be completely different than the variance of the data used for fitting the data.

Weighted Least Squares

If the matrix $\mathbf{V}$ is diagonal, so only homoskedasticity is “broken,” we can solve the problem with weighted least squares (WLS).
If the variance of is known to follow the equation

$$\mathbb{E}[\varepsilon\varepsilon'|\mathbf{X}] = \sigma^2 \operatorname{diag}(v_1, \ldots, v_n)\,.$$

Then $\mathbf{C}$ is the diagonal matrix

$$\mathbf{C} = \operatorname{diag}(\frac{1}{\sqrt{v_1}}, \ldots, \frac{1}{\sqrt{v_n}})\,, \tag{12}$$

and the transformed model is just a “weighted form” of the original

$$\begin{aligned} \widetilde{y}_i &= \frac{y_i}{\sqrt{v_i}} \\ \widetilde{\mathbf{X}}_{ij} &= \frac{\mathbf{X}_{ij}}{\sqrt{v_i}}\,, \end{aligned}$$

for all i, j.

Generalizing Classical Linear Regression: Going Beyond Normality and Strict Exogeneity

Revisiting Some of the Assumptions of CLR I

We discussed above how to go beyond homoscedasticity, leading to GLR and GLS.

Two other (very) restrictive assumptions of CLR are:

- strict exogeneity,
- normality.

These are not satisfied in many time series and panel data regression models, where frequently errors are (a) not strictly exogenous[4] and (b) serially correlated. In practice, these assumptions may also not be satisfied in practice (especially for financial data) for the cross-sectional regression models that we are discussing in this course.

Question: Can we generalize these assumptions somehow?

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.

While the mathematics is not hard, it requires "machinery" from probability theory in order to properly derive the results (i.e. various limit theorems) that go beyond the scope of this course.

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.

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

Convergence in Probability

The sequence of random vectors $\{\mathbf{z}_n\} \in \mathbb{R}^k$ *converges in probability* to a constant vector $\boldsymbol{\alpha} \in \mathbb{R}^k$ if, for any $\varepsilon > 0$,

$$\lim_{n \to \infty} \Pr\left(|z_{nk} - \alpha_k| > \varepsilon\right) = 0 \text{ for all } k\,.$$

We denote this by

$$\mathbf{z}_n \underset{p}{\to} \boldsymbol{\alpha}\,, \tag{13}$$

and

$$\text{plim}_{n \to \infty} \mathbf{z}_n = \boldsymbol{\alpha}\,. \tag{14}$$

Consistent Estimator

An estimator $\widehat{\boldsymbol{\beta}}_n$ is a *consistent estimator* of $\boldsymbol{\beta}$ if

$$\widehat{\boldsymbol{\beta}}_n \underset{p}{\longrightarrow} \boldsymbol{\beta}\,. \tag{15}$$

Convergence in Distribution

The sequence of random vectors $\{\mathbf{z}_n\} \in \mathbb{R}^k$ *converges in distribution* to a random vector $\mathbf{z} \in \mathbb{R}^k$ if the joint cumulative distribution function (CDF) $F_n(\mathbf{x})$ of $\mathbf{z}_n$ converges to the joint CDF $F(\mathbf{x})$ of $\mathbf{z}$ for all $\mathbf{x} \in \mathbb{R}^k$ where $F(\mathbf{x})$ is continuous. We denote this

by

$$\mathbf{z}_n \underset{d}{\rightarrow} \mathbf{z}\,. \tag{16}$$

Intuition I

Let us start developing some intuition as to how this is going to work.

We can rewrite the OLS estimator as follows

$$\begin{aligned}\widehat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}\left(\frac{1}{n}\mathbf{X}'\mathbf{y}\right) \\ &= \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i \cdot y_i\right) .\end{aligned}$$

Notice how we have made the estimator's dependence on n explicit.

Intuition II

Using the above, we can write the sampling error

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\varepsilon} \\
&= \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i\right) \\
&= \mathbf{S}_{\mathbf{xx}}^{-1}\overline{\mathbf{g}}
\end{aligned}
$$

where

$$
\begin{aligned}
\mathbf{S}_{\mathbf{xx}} &:= \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i' \,, \\
\overline{\mathbf{g}} &:= \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i \,.
\end{aligned}
$$

Important: While our notation does not make it explicit, the sample means $\mathbf{S}_{\mathbf{xx}}$ and $\overline{\mathbf{g}}$ depend on the sample size n.

Intuition III

When the number of rows of $\mathbf{X}$ becomes large, i.e. $n \to \infty$, then under "some reasonable" assumptions

$$\mathbf{S}_{\mathbf{xx}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' \underset{p}{\to} \mathbb{E}[\mathbf{x}_i \mathbf{x}_i'] \,,$$

$$\mathbf{S}_{\mathbf{xx}}^{-1} = \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \underset{p}{\to} \left(\mathbb{E}[\mathbf{x}_i \mathbf{x}_i'] \right)^{-1} \,,$$

$$\overline{\mathbf{g}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \varepsilon_i \underset{p}{\to} \mathbb{E}[\mathbf{x}_i \varepsilon_i] \equiv \mathbf{0} \,.$$

Therefore $\widehat{\boldsymbol{\beta}} \underset{p}{\to} \boldsymbol{\beta}$, showing that the estimator $\widehat{\boldsymbol{\beta}}$ is consistent.

Large Sample Assumptions for Linear Regression I

We will make the following assumptions:

I'. *Linearity:*

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i\,. \tag{17}$$

II'. *Stationarity and IID:* The $(p+1)$-dimensional vector stochastic process $\{y_i, \mathbf{x}_i\}$ is jointly stationary, independent and identically distributed (IID).[5]

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

$$\mathbb{E}[\mathbf{x}_i\varepsilon_i] = \mathbf{0}\,. \tag{18}$$

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:*

$$\mathbb{E}[\varepsilon_i|\varepsilon_{i-1},\ldots,\varepsilon_1,\mathbf{x}_i,\ldots,\mathbf{x}_1,] = 0\,. \tag{19}$$

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

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

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

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

[5] To be more general, one can replace IID with ergodic and all results we will discuss 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

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

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

$$\mathbf{S}_{\mathbf{xx}} := \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i' = \frac{1}{n}\mathbf{X}'\mathbf{X} \,.$$

Hypothesis Testing in the Large Sample Case

The Robust t-test I

We want to test the null hypothesis

$$H_0 : \beta_j = b_j \,, \tag{20}$$

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

$$H_a : \beta_j \neq b_j \,. \tag{21}$$

The Robust t-test II

From the Proposition above we have that under the null

$$\sqrt{n}\big(\widehat{\beta}_j - b_j\big) \mid H_0 \underset{d}{\rightarrow} \mathcal{N}\left(0, \mathrm{avar}[\widehat{\beta}_j]\right), \tag{22}$$

$$\widehat{\mathrm{avar}[\widehat{\beta}_j]} \mid H_0 \underset{p}{\rightarrow} \mathrm{avar}[\widehat{\beta}_j]. \tag{23}$$

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

It follows that

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

where

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

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*[6]. 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 tell it apart 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.

[6] 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

$$H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{c}\,, \tag{24}$$

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*

$$W := n \cdot (\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{c})' \left(\mathbf{R} \cdot \widehat{\mathrm{avar}[\widehat{\boldsymbol{\beta}}]} \cdot \mathbf{R}' \right)^{-1} (\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{c})\,. \tag{25}$$

One can show that

$$W \underset{d}{\rightarrow} \chi^2_c\,, \tag{26}$$

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

$$F = \frac{(\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{c})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{c})/k}{s^2}\,, \tag{27}$$

$$W = n \cdot (\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{c})' \left(\mathbf{R} \cdot \widehat{\mathrm{avar}[\widehat{\boldsymbol{\beta}}]} \cdot \mathbf{R}'\right)^{-1} (\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{c})\,. \tag{28}$$

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

Time Series and Panel Data Models

A Comment on Time Series and Panel Data Models

We will not cover time series models in this class. It is a whole course on its own. If you are curious about it, see for example Hayashi (2000) and Tsay (2005).

Similarly, if you are interested in panel data models, see for example Wooldridge (2010) and Hsiao (2014).

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.

Hsiao, Cheng (2014). *Analysis of Panel Data*. 54. Cambridge 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.

Tsay, Ruey S. (2005). *Analysis of Financial Time Series*. John Wiley & Sons.

White, Halbert (1980). "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for