Teoria e Pratica dei Modelli Statistici

Homework 3

Edoardo Ricca

1. *Write in a formal way the four standard assumptions on the model errors (zero mean, homoscedasticity, no correlation, normality).*

1. **zero mean**: given the linear model:

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

the following statement is assumed

$$\mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}$$

which means that the expected value of the errors is equal to zero.
Due to the linearity of expectations, this implies that:

$$\mathbb{E}[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta}$$

In the common case in which the covariates are stochastic as well, we should instead assume that:

$$\mathbb{E}[\boldsymbol{\varepsilon}|\mathbf{X}] = \mathbf{0}$$

which implies that

$$\mathbb{E}[\mathbf{y}|\mathbf{X}] = \mathbf{X}\boldsymbol{\beta}$$

In estimation, the estimate of the errors (the residuals) achieves a sum of zero through the orthogonality with respect to $\mathbf{X}$ which then leads to an estimated expectation (through the usage of the sample mean) of $\mathbf{0}$.

2. **homoscedasticity**: given the linear model presented above where n is the number of observations (and consequently the number of rows in $\mathbf{X}$), assuming that the variance of the errors is homoscedastic means:

$$\mathbb{V}ar[\varepsilon_i] = \sigma_\varepsilon^2 \quad \forall i \in [1, n]$$

indicating that the variance is equal for each error, meaning that it doesn't systematically change across the observations.

3. **no correlation**: given the aforementioned linear model, assuming no correlation means that:

$$\mathbb{C}ov(\varepsilon_i, \varepsilon_j) = 0 \iff \mathbb{C}or(\varepsilon_i, \varepsilon_j) = 0 \quad \forall i, j \in \mathbb{N}^+ \text{ with } i \neq j$$

which in matrix notation is written:

$$\mathbb{C}ov(\boldsymbol{\varepsilon}) = \mathbb{E}[\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}] = \sigma^2\mathbf{I}$$

since $\mathbb{C}ov(\varepsilon_i, \varepsilon_j) = \sigma_{\varepsilon_i}, \forall i = j$ and errors are homoscedastic by assumption.

4. **normality**: given the above linear model, assuming normally distributed errors means that:

$$\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2) \iff f_{\mathscr{E}}(\varepsilon_i) = \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} exp\left\{-\frac{\varepsilon_i^2}{2\sigma_\varepsilon^2}\right\}$$

where $f_{\mathscr{E}}(\varepsilon_i)$ is the probability density function of ε_i. In matrix notation is:

$$\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_\varepsilon^2\mathbf{I}) \iff f_{\mathscr{E}}(\boldsymbol{\varepsilon}) = \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2\mathbf{I}}} exp\left\{-\frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{2\sigma_\varepsilon^2\mathbf{I}}\right\}$$

Since $\mathbf{X} \perp\!\!\!\perp \boldsymbol{\varepsilon}$, stochastic covariates do not change the distribution of errors.

All the assumptions above leave to:

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma_{\boldsymbol{\varepsilon}}^2\mathbf{I})$$

2. *Prove that the OLS estimator is unbiased; list the standard assumptions of the model errors that are needed to get this result*

Given the OLS estimator:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

to prove that is unbiased, we simply compute its expected value and show that is equal to the true parameter β.

$$\begin{aligned}\mathbb{E}[\hat{\boldsymbol{\beta}}] &= \mathbb{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] = \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}[\mathbf{y}]\end{aligned}$$

due to the linearity of expectations. Using the above proved expected value of $\mathbf{y}$:

$$\mathbb{E}[\hat{\boldsymbol{\beta}}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

proving that the OLS estimator is unbiased.

To obtain this result, apart from the non-singularity and consequent invertibility of $\mathbf{X}$'$\mathbf{X}$, we need some of the assumptions aforementioned. In particular, the expected value of $\mathbf{y}$ is equal to $\mathbf{X}\boldsymbol{\beta}$ if and only if the expected value of the errors is equal to zero. We are therefore using the zero mean assumption. Since this estimator can be obtained without assuming a particular distribution of the errors (and consequently of $\mathbf{y}$) normality is not required.

Homoscedasticity and zero covariance too are not needed for the estimator to be unbiased although heteroscedastic errors can lead to mistakes in the estimation of the variance of $\hat{\boldsymbol{\beta}}$ and errors exhibiting ith-order autocorrelation are usually a symptom of a misspecified regression model (unless we are dealing with time-series analysis).

3. *Derive the covariance matrix of the OLS estimator; list the standard assumptions of the model errors that are needed to get this result*

Given that

$$\mathbb{C}ov(\mathbf{y}) = \sigma_{\boldsymbol{\varepsilon}}^2 \mathbf{I}$$

and

$$\mathbb{C}ov(\mathbf{AX}) = \mathbf{A}\mathbb{C}ov(\mathbf{X})\mathbf{A}'$$

it is possible, while also using the exchangeability of transposition with respect to inversion, to prove that:

$$\begin{aligned}\mathbb{C}ov(\hat{\boldsymbol{\beta}}) &= \mathbb{C}ov\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\right) = \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{C}ov(\mathbf{y})((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')' \\ &= \sigma_{\boldsymbol{\varepsilon}}^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma_{\boldsymbol{\varepsilon}}^2(\mathbf{X}'\mathbf{X})^{-1}\end{aligned}$$

obtaining the covariance matrix of the OLS estimator.

For this proof to be true, the first given statement must be true as well. Due to the homoscedasticity assumption:

$$\mathbb{V}ar[y_i] = \mathbb{V}ar[\mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i] = \mathbb{V}ar[\varepsilon_i] = \sigma_\varepsilon^2$$

and due to the no-correlation assumption:

$$\mathbb{C}ov(y_i, y_j) = \mathbb{C}ov(\varepsilon_i, \varepsilon_j) = 0$$

which leads to the fact that $\mathbb{C}ov(\mathbf{y}) = \sigma_\varepsilon^2\mathbf{I}$ and that our proof is true. Therefore, in order to obtain this covariance matrix both the homoscedasticity and the no-correlation assumption are needed. No distribution is necessary to be specified.

4. *Write the optimal prediction of y for a future observation with covariates equal to x_0. In what sense this prediction is optimal?*

Given a vector of covariates x_0, the optimal prediction for a future observation is $\mathbf{X}\hat{\boldsymbol{\beta}}$ which is the sum of the individual estimated parameters (in the case of x^0). As shown in section 3.2.3 of FKLM, the prediction is optimal in the sense that minimizes the Euclidian norm between the observation vector $\mathbf{y}$ and the estimate $\hat{\mathbf{y}}$. To minimize this distance, $\hat{\boldsymbol{\beta}}$ is chosen such that the residual vector (which is the line connecting the observation and the prediction) is orthogonal to the prediction.

5. *Under the four standard assumptions on the model errors, write a statistic with chi-squared distribution which is based on the distance between the OLS estimates and the true parameter values.*

Using the assumptions presented, we concluded that $\mathbf{y}$ follows a normal distribution. Therefore, since if a matrix $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \mathbf{d} \in \mathbb{R}^q$ and $\mathbf{D}$ is a $(q \times p)$-matrix with $rk(D) = q \leq p$; then:

$$\mathbf{Y} = \mathbf{d} + \mathbf{D}\mathbf{X} \sim \mathcal{N}_q(\mathbf{d} + \mathbf{D}\boldsymbol{\mu}, \mathbf{D}\boldsymbol{\Sigma}\mathbf{D}')$$

we can conclude that $\hat{\boldsymbol{\beta}}$, being a linear combination of $\mathbf{y}$, follows:

$$\hat{\boldsymbol{\beta}} = \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$$

Given also that if $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} > 0$ it follows that:

$$\mathbf{Y} = (\mathbf{X} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}) \sim \chi^2_p$$

then we can conclude, since $\boldsymbol{\Sigma}^{-1}$ of $\hat{\boldsymbol{\beta}}$ is equal to $\frac{1}{\sigma^2}(\mathbf{X}'\mathbf{X})$ that:

$$\frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\mathbf{X}'\mathbf{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\sigma^2} \sim \chi^2_p$$

6. *Write a further assumption allowing to derive the asymptotic properties of the OLS estimator without normality and mention a typical situation when such assumption is true.*

If we cannot use the assumption of normally distributed errors, it is still possible to obtain valid results asymptotically. However, to obtain such results, we need the following assumption: given an indexed succession of design matrices $\mathbf{X}_n$ the following is true

$$\lim_{n \to \infty} \frac{1}{n} \mathbf{X}'_n \mathbf{X}_n = \mathbf{V}$$

where $\mathbf{V}$ is a positive definite matrix. Thanks to this assumption, we can obtain the following asymptotic properties:

1. $\hat{\boldsymbol{\beta}}_n$ is consistent for $\boldsymbol{\beta}$.
2. $\hat{\sigma}^2_n$ (obtained through ML or REML estimation) is consistent for σ^2.
3. The least square estimator has asymptotically the distribution:

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{V}^{-1})$$

Given these properties; not only can we conclude that, for a sufficiently large sample size n, the estimator has approximately a normal distribution even without assuming the normality of errors since, using property 3, we obtain:

$$\hat{\boldsymbol{\beta}}_n \overset{a}{\sim} \mathcal{N}(\boldsymbol{\beta}, \sigma^2 \mathbf{V}^{-1} n^{-1})$$

but also, if we replace σ^2 with its consistent estimator (property 2) and we consider the assumption presented above to be true then:

$$\hat{\boldsymbol{\beta}}_n \overset{a}{\sim} \mathcal{N}(\boldsymbol{\beta}, \hat{\sigma}^2_n (\mathbf{X}'\mathbf{X})^{-1})$$

indicating that the least square estimator has the same approximate (asymptotic) normal distribution with or without assuming the normality of errors.

It should be noticed, though, that usually the assumption introduced in this exercise is violated if the covariates follow a trend while is almost assured if the covariates are i.i.d stochastic covariates. An example of trend is $x_i = i^2$. In the former's case, we can substitute the assumption generating $\mathbf{V}$ with:

$$(\mathbf{X}_n'\mathbf{X}_n)^{-1} \to \mathbf{0}$$

which can be interpreted as the covariate information increasing to infinity as n goes to infinity. This is a necessary and sufficient condition for consistency of both $\hat{\boldsymbol{\beta}}_n$ and $\hat{\sigma}_n^2$. Moreover, to obtain asymptotic normality we need that:

$$\max_{i=1,\dots,n} \mathbf{x}_i'(\mathbf{X}_n'\mathbf{X})^{-1}\mathbf{x}_i \to 0 \quad \text{for} \quad n \to \infty$$

indicating that the influence of each observation is insignificant compared to the entire information matrix. Assuming the previous two statements to be true we obtain again, using the central limit theorem:

$$\hat{\boldsymbol{\beta}}_n \overset{a}{\sim} \mathscr{N}(\boldsymbol{\beta}, \hat{\sigma}_n^2(\mathbf{X}'\mathbf{X})^{-1})$$

7. *Derive the expectation and covariance matrix of the residuals.* In order to derive both the expectation and the covariance matrix of the residuals, we need to express them in terms of the prediction matrix $\mathbf{H}$. Since

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$$

by definition of residuals, we obtain:

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{H}\mathbf{y} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

Therefore, the expected value of the residuals is:

$$\begin{aligned}
\mathbb{E}[\hat{\boldsymbol{\varepsilon}}] &= \mathbb{E}[(\mathbf{I} - \mathbf{H})\mathbf{y}] = \\
&= \mathbb{E}[\mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] = \\
&= \mathbb{E}[\mathbf{y}] - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}[\mathbf{y}] = \\
&= \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{0}
\end{aligned}$$

which means that residuals, like errors, have zero mean.

Using the second property of covariance used in deriving the covariance of $\hat{\boldsymbol{\beta}}$ and the fact that $(\mathbf{I} - \mathbf{H})$ is both symmetric and idempotent, we can prove that:

$$\begin{aligned}
\mathbb{C}ov(\hat{\boldsymbol{\varepsilon}}) &= \mathbb{C}ov((\mathbf{I} - \mathbf{H})\mathbf{y}) = \\
&= (\mathbf{I} - \mathbf{H})\,\mathbb{C}ov(\mathbf{y})\,(\mathbf{I} - \mathbf{H})' = \\
&= (\mathbf{I} - \mathbf{H})\,\sigma_\varepsilon^2\mathbf{I}(\mathbf{I} - \mathbf{H})' = \sigma_\varepsilon^2(\mathbf{I} - \mathbf{H})
\end{aligned}$$

which implies that, differently from the errors, the residuals are neither homoscedastic nor uncorrelated.

8. *Explain why the residuals cannot be directly used to judge the homoscedasticity assumption and how this problem can be circumvented.*

For the reasons mentioned above, the residuals be cannot used to judge if errors are homoscedastic since homoscedastic errors do not imply homoscedastic residuals. Since errors are unknown, it becomes impossible to judge their homoscedasticity based only on the residuals and therefore it cannot be verified if one essential model assumption is met. To circumvent this problem, two solutions are possible:

1. **standardized residuals**: using standardization, which means dividing a variable by its standard deviation, we obtain the standardized residual r_i defined as:

$$r_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

   if the model assumptions are correct, standardized residuals $\mathbf{r}$ are homoscedastic. It is therefore possible to check whether the model assumptions are met by verifying if standardized residuals are homoscedastic.

2. **studentized residuals**: it is possible to obtain studentized residuals $r*_i$ by computing:

$$r*_i = \frac{\hat{\varepsilon}_{(i)}}{\hat{\sigma}_{(i)}\sqrt{1 + x'_{(i)}(\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1}x_i}} \sim t_{n-p-1}$$

   where the subscript (i) indicates that the variables lack the ith row (one observation is removed) and that estimators are computed without considering the ith observation. This is done to ensure that $\hat{\varepsilon}_{(i)} \perp\!\!\!\perp \hat{\sigma}_{(i)}$. As for the standardized residuals, studentized residuals can be used to check for model assumptions such as homoscedasticity of errors (and also to discover outliers).

9. *Suppose you fitted a model with an intercept and three covariates; considering the framework of general linear hypotheses, write the matrix C and the vector d to test the following hypotheses:*

(i) $5\beta_1 + 2 = 7\beta_2$
(ii)$\beta_1 = \beta_2 = \beta_3$

Assuming an intercept β_0 exists

1. $5\beta_1 + 2 = 7\beta_2 \implies \mathbf{d} = -2 \quad \wedge \quad \mathbf{C} = (0, 5, -7, 0, \ldots, 0)$

   such that:

$$\mathbf{C}\boldsymbol{\beta} = \mathbf{d} \iff (0, 5, -7, 0, \ldots, 0) \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} = \mathbf{d} \iff 5\beta - 7\beta_2 = -2$$

which is the initial hypothesis

2. $\beta_1 = \beta_2 = \beta_3 \implies \begin{cases} \beta_1 - \beta_2 = 0 \\ \beta_2 - \beta_3 = 0 \end{cases} \implies \mathbf{d} = \mathbf{0} \quad \wedge \quad \mathbf{C} = \begin{pmatrix} 0 & 1 & -1 & 0 & \dots & 0 \\ 0 & 0 & 1 & -1 & \dots & 0 \end{pmatrix}$

such that:

$$\mathbf{C}\boldsymbol{\beta} = \mathbf{d} \iff \begin{pmatrix} 0 & 1 & -1 & 0 & \dots & 0 \\ 0 & 0 & 1 & -1 & \dots & 0 \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \iff \begin{pmatrix} \beta_1 - \beta_2 \\ \beta_2 - \beta_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

which is, again, the initial hypothesis.

10. *Suppose you have to test the hypothesis $\beta = 1$. Draw a plot like figure 3.15 on page 129 to show graphically (i) the F test; (ii) the Wald test (only the numerator). Briefly explain the difference between the principles underlying the F test and the Wald test.*

We begin by assuming that errors are normally distributed.
The idea behind the F-test is very similar to the idea behind the Likelihood Ratio Test where we compute a test statistic based on the proportion between the restricted and unrestricted estimates. While the LRT deals with the likelihood function of estimated parameters, the F-test deals with the residual sum of squares. After computing the estimator $\hat{\boldsymbol{\beta}}^R$, which is the estimate of $\boldsymbol{\beta}$ under the restriction $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{d}$ obtained using the Lagrangian approach for restricted optimization, we can compute SSE_{H_0} which is the restricted residual sum of squares. By comparing SSE_{H_0} and SSE and by knowing their distributions (both χ^2 with different df) we obtain this test statistic following an F-distribution:

$$\frac{n-p}{r}\frac{\Delta SSE}{SSE} = \frac{n-p}{r}\frac{SSE_{H_0} - SSE}{SSE} \sim F_{r,n-p}$$

where r is the number of rows in $\mathbf{C}$ (the number of linearly independent restrictions), n is the number of observations and p the number of parameters.

This means that the closer $\hat{\boldsymbol{\beta}}^R$ is to $\hat{\boldsymbol{\beta}}$ the closer ΔSSE will be to 0, indicating that we should reject the null hypothesis for high values of F and fail to reject it for low values. More specifically, given a α level of significance, we will reject the null hypothesis if:

$$F > F_{r,n-p}(1-\alpha)$$

meaning that is larger than the (1 - α) quantile.

The similarities with the Wald test stem from the fact that, by expressing the F statistic in matrix notation, we obtain:

$$F = \frac{(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d})' \widehat{\mathbb{C}ov(\mathbf{C}\hat{\boldsymbol{\beta}})}^{-1} (\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d})}{r}$$

which means that:

$$W = rF$$

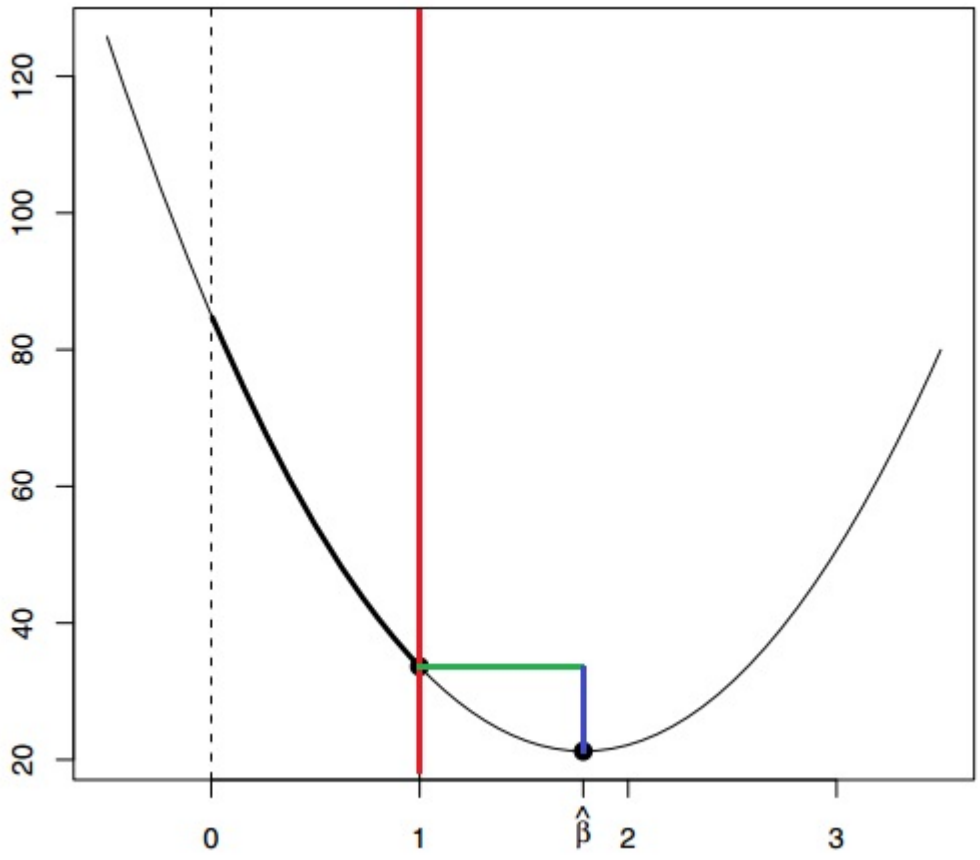


Figure 1: Restriction $\beta = 1$ (red), Wald test (green), F test (blue)

The idea behind the Wald test is to compute the difference between the estimate $\mathbf{C}\hat{\boldsymbol{\beta}}$ and its hypothesized value $\mathbf{d}$, all weighted by the inverse of the estimated covariance matrix.

Since the Wald test can be interpreted as r times the F-test, we can use the same interpretations for the F-test with the only difference of the constant r.

This connection between the two tests allows us to conclude that since, asymptotically $W \overset{a}{\sim} \chi^2_r$ then $F \overset{a}{\sim} \chi^2_r/r$ distribution which can be used even in the presence of non-normal errors.

Considering picture 3.15 (above) on page 143 with the hypothesis of $\beta = 1$ we can see the two tests graphically. Ignoring eventual constants and weights in both tests, we can visualize from the graph how the idea behind the Wald test is to measure the distance between the actual estimate and the hypothesized value by looking at the distance in green. Looking at the distance in blu, we can visualize the idea of the F-test involving the difference between the SSE of the restricted estimate and the SSE of the unrestricted estimate (being the vertical axis the residual sum of squares).

11. *Write the expression of the confidence interval for the mean of a new observation and the prediction interval for the value of a new observation; compute the difference between the length of the two intervals and comment.*

Due to the duality of the two-sided hypothesis test and confidence interval, it is possible to obtain the confidence intervals starting from the test statistics. Given a new observation with covariates x_0 we would obtain an optimal prediction given by $\hat{y}_0 = \mathbf{x}_0'\hat{\boldsymbol{\beta}}$ where $\hat{y}_0$ is the estimator for the conditional expected value of y: $\mathbb{E}[y_0]$ also indicated as μ_0. To obtain an interval estimation, however, we can use the test statistic:

$$\frac{\mathbf{x}_0'\hat{\boldsymbol{\beta}} - \mu_0}{\hat{\sigma}\sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}} \sim t_{n-p}$$

where the studentization occurs since $\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$ which implies that $\mathbf{x}_0'\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{x}_0'\boldsymbol{\beta}, \sigma^2\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)$, being a linear combination, and therefore we can standardize the test statistic and subsequently studentize by substituting the real standard deviation with its estimator.

Using the two quantiles of the test statistic we obtain:

$$\mathbb{P}\left(-t_{n-p}(1-\alpha/2) \leq \frac{\mathbf{x}_0'\hat{\boldsymbol{\beta}} - \mu_0}{\hat{\sigma}\sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}} \leq t_{n-p}(1-\alpha/2)\right) = 1-\alpha$$

$$\mathbb{P}\left(-t_{n-p}(1-\alpha/2)\hat{\sigma}\sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} \leq \mathbf{x}_0'\hat{\boldsymbol{\beta}} - \mu_0 \leq t_{n-p}(1-\alpha/2)\hat{\sigma}\sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}\right) = 1-\alpha$$

$$\mathbb{P}\left(-t_{n-p}(1-\alpha/2)\hat{\sigma}\sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} - \mathbf{x}_0'\hat{\boldsymbol{\beta}} \leq -\mu_0 \leq t_{n-p}(1-\alpha/2)\hat{\sigma}\sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} - \mathbf{x}_0'\hat{\boldsymbol{\beta}}\right) = 1-\alpha$$

$$\mathbb{P}\left(-t_{n-p}(1-\alpha/2)\hat{\sigma}\sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} + \mathbf{x}_0'\hat{\boldsymbol{\beta}} \leq \mu_0 \leq t_{n-p}(1-\alpha/2)\hat{\sigma}\sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} + \mathbf{x}_0'\hat{\boldsymbol{\beta}}\right) = 1-\alpha$$

from which we derive the $(1-\alpha)$ confidence interval for μ_0:

$$\left[\mathbf{x}_0'\hat{\boldsymbol{\beta}} - t_{n-p}(1-\alpha/2)\hat{\sigma}\sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}\,;\, \mathbf{x}_0'\hat{\boldsymbol{\beta}} + t_{n-p}(1-\alpha/2)\hat{\sigma}\sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}\right]$$

If we wanted instead to compute a prediction interval for a future observation y_0 rather than its mean, we would use the test statistic:

$$\frac{y_0 - \mathbf{x}_0'\hat{\boldsymbol{\beta}}}{\hat{\sigma}\sqrt{1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}} \sim t_{n-p}$$

obtained using the normal distribution of the prediction error $\hat{\varepsilon_0} = y_0 - \mathbf{x}_0'\hat{\boldsymbol{\beta}}$ as before:

$$\mathbb{P}\left(-t_{n-p}(1-\alpha/2) \leq \frac{y_0 - \mathbf{x}_0'\hat{\boldsymbol{\beta}}}{\hat{\sigma}\sqrt{1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}} \leq t_{n-p}(1-\alpha/2)\right) = 1-\alpha$$

$$\mathbb{P}\left(-t_{n-p}(1-\alpha/2)\hat{\sigma}\sqrt{1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} \leq y_0 - \mathbf{x}_0'\hat{\boldsymbol{\beta}} \leq t_{n-p}(1-\alpha/2)\hat{\sigma}\sqrt{1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}\right) = 1-\alpha$$

$$\mathbb{P}\left(-t_{n-p}(1-\alpha/2)\hat{\sigma}\sqrt{1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} + \mathbf{x}_0'\hat{\boldsymbol{\beta}} \leq y_0 \leq t_{n-p}(1-\alpha/2)\hat{\sigma}\sqrt{1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} + \mathbf{x}_0'\hat{\boldsymbol{\beta}}\right) = 1-\alpha$$

indicating the $(1-\alpha)$ confidence interval for y_0:

$$\left[\mathbf{x}_0'\hat{\boldsymbol{\beta}} - t_{n-p}(1-\alpha/2)\hat{\sigma}\sqrt{1+\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}\,;\, \mathbf{x}_0'\hat{\boldsymbol{\beta}} + t_{n-p}(1-\alpha/2)\hat{\sigma}\sqrt{1+\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}\right]$$

Computing the length of the confidence interval for μ_0 we obtain:

$$\begin{aligned}
&\mathbf{x}_0'\hat{\boldsymbol{\beta}} + t_{n-p}(1-\alpha/2)\hat{\sigma}\sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} - \mathbf{x}_0'\hat{\boldsymbol{\beta}} + t_{n-p}(1-\alpha/2)\hat{\sigma}\sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} = \\
=&t_{n-p}(1-\alpha/2)\hat{\sigma}\sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} + t_{n-p}(1-\alpha/2)\hat{\sigma}\sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} = \\
=&2t_{n-p}(1-\alpha/2)\hat{\sigma}\sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}
\end{aligned}$$

The length of the prediction interval for y_0 is instead:

$$\begin{aligned}
&\mathbf{x}_0'\hat{\boldsymbol{\beta}} + t_{n-p}(1-\alpha/2)\hat{\sigma}\sqrt{1+\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} - \mathbf{x}_0'\hat{\boldsymbol{\beta}} + t_{n-p}(1-\alpha/2)\hat{\sigma}\sqrt{1+\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} = \\
=&t_{n-p}(1-\alpha/2)\hat{\sigma}\sqrt{1+\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} + t_{n-p}(1-\alpha/2)\hat{\sigma}\sqrt{1+\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} = \\
=&2t_{n-p}(1-\alpha/2)\hat{\sigma}\sqrt{1+\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}
\end{aligned}$$

We can conclude that the difference between the length $CI_{\mu_0}(1-\alpha)$ and $PI_{y_0}(1-\alpha)$ is:

$$2t_{n-p}(1-\alpha/2)\hat{\sigma}\left[\sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} - \sqrt{1+\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}\right]$$

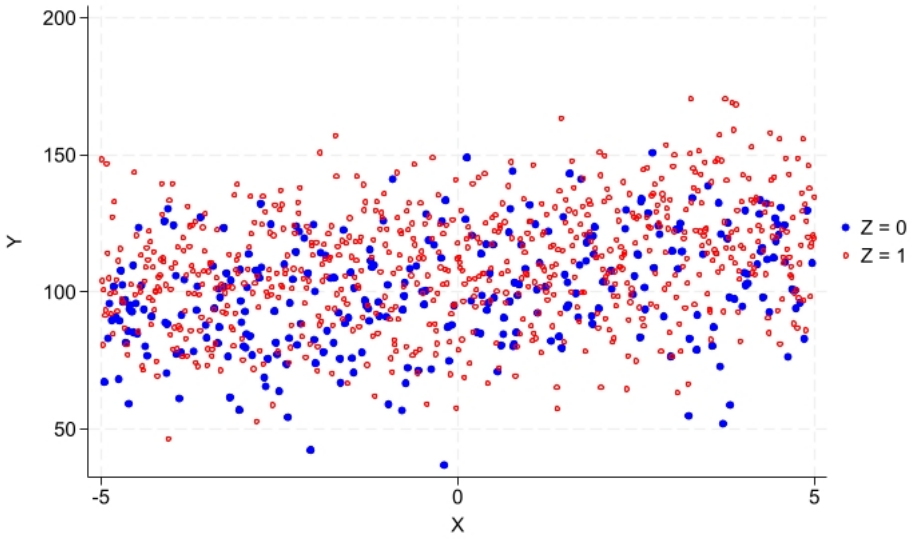
which is always negative and allows us to conclude that the length of the prediction interval is always greater than the confidence interval (as expected given the additional randomness involved in prediction).

Stata

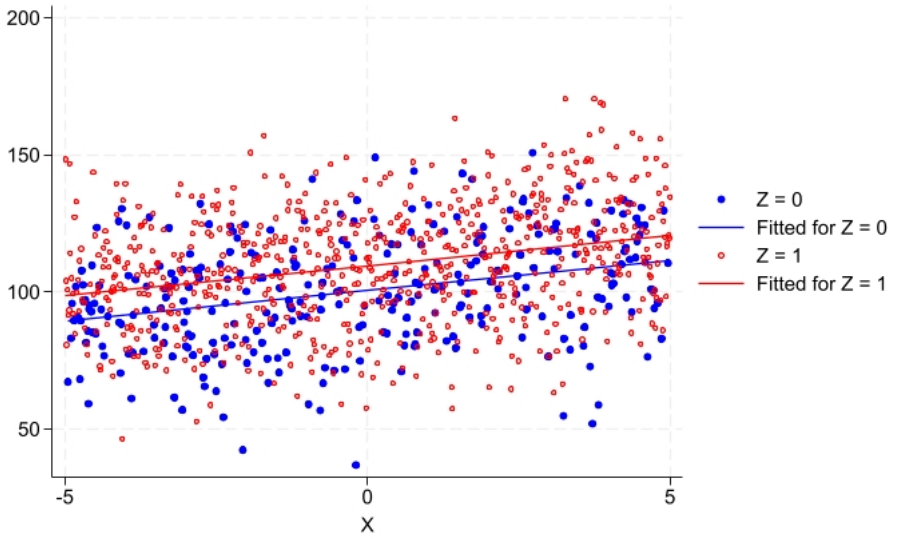
a. Make a plot, using different symbols for observations with Z=0 and Z=1.
b. Then fit the model with least squares and add the two regression lines (for Z=0 and Z=1) to the previous plot.
*c. Then make tests for the following hypotheses (you have to report the p-value and state if you reject or not at a significance level of 5%): betaX=0, betaX=1.5, betaX=1.9, betaZ=0, betaX=0 & betaZ=0, betaZ=4*betaX.*
d. Finally, predict the response Y for a new observation with X=3 and Z=0 and compute (i) the confidence interval for the mean of the response and (ii) the prediction interval for the value of the response.

The code can be found on the do file uploaded.

Point a.



Point b.



Point c.

Using the `test` function that implements the Wald test.

Hypothesis	$\beta_X = 0$	$\beta_X = 1.5$	$\beta_X = 1.9$	$\beta_Z = 0$	$\beta_X = 0 \wedge \beta_Z = 0$	$\beta_Z = 4 \cdot \beta_X$
F(1,997)	98.10	9.34	1.51	43.44	-	0.07
F(2,997)	-	-	-	-	74.41	-
p-value	0.0000	0.0023	0.2191	0.0000	0.0000	0.7879
5% level conclusion	rejected	rejected	failed to reject	rejected	rejected	failed to reject

Point d.

The computations lead to the following results, with an optimal prediction of 107.0461 and the following confidence and prediction intervals (at 95% level):

	Lower bound	Upper bound
CI	104.3819	109.7103
PI	67.81372	146.2785

As expected, the length of CI for μ_0 is significantly smaller than the PI.

12. *Using matrix operations, derive the formula of the prediction interval in the case of a model with an intercept and a single regressor x (so that the design matrix X has two columns); note that the derived formula expresses the length of the prediction interval as a function of x0, namely the value of the regressor for the new observation: which is the value of x0 such that the prediction interval has minimum length?*

Unfortunately, I was not able to derive the formula using matrix operations. I tried to prove it by computing $\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0$ but I only solved until this equations:

$$\frac{\sum_{i=1} x_i^2 - n x_0 \sum_{i=1} x_i + x_0 \sum_{i=1} x_i - x_0^2 \sum_{i=1} x_i^2}{n \sum_{i=1} x_i^2 - \left(\sum_{i=1} x_i\right)^2}$$

but I was unable to proceed further. I found another method by computing the variance of $\hat{y}$ but is not in matrix notation. Regarding the value of x_0 that minimizes the length, I assume it is the mean $\bar{x}$ since it would eliminate one of the terms inside the square root, indicating that the closer we are to the center of the data, the better our predictions are.