Columbia University
Statistical Analysis and Time Series

IEOR-4709

A. Capponi | **Problem Set #4** | Issued: March 24, 2025
Spring 2025 | Due: **BEFORE CLASS** April 2, 2025

Note: Please put the number of hours that you spent on this homework set on top of the first page of your homework. The TA in charge of this homework's review session is Jose Sidaoui Gali.

The notation in the problem below is the same as the notation adopted in class.

Ex. 1.
Consider the simple linear regression model

$$y_i = \alpha + \beta x_i + \epsilon_i,$$

where ϵ_i is a zero-mean random variable. Moreover, ϵ_i's have equal variance σ^2, ϵ_i's are independent, ϵ_i's are Gaussian with mean zero and variance σ^2. Do the following:

- Find the MLE estimators for α and β given observed samples $x_{1:n} = (x_1, \ldots, x_n)$ and $y_{1:n} = (y_1, \ldots, y_n)$.
- Show that the MLE estimator of σ^2 is $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i^2$.

Solution. The likelihood function of $x_{1:n} = (x_1, \ldots, x_n)$ and $y_{1:n} = (y_1, \ldots, y_n)$ is

$$L(\alpha, \beta, \sigma^2; x, y) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\alpha-\beta x_i)^2}{2\sigma^2}} = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i-\alpha-\beta x_i)^2}$$

and the corresponding log-likelihood function is

$$LL(\alpha, \beta, \sigma^2; x, y) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \alpha - \beta x_i)^2.$$

We take the derivative of the log-likelihood function with respect to α, β and σ^2 respectively to get

$$\begin{aligned}
\frac{1}{2\sigma^2}\sum_{i=1}^n 2(y_i - \alpha - \beta x_i) &= 0 \\
\frac{1}{2\sigma^2}\sum_{i=1}^n 2x_i(y_i - \alpha - \beta x_i) &= 0 \\
-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^n (y_i - \alpha - \beta x_i)^2 &= 0.
\end{aligned}$$

We solve the above three equations for α, β and σ^2 and get the MLE estimators as below

$$\begin{aligned}
\hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x} \\
\hat{\beta} &= \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} := S_{xy}/S_{xx} \\
\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2 = \frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i^2.
\end{aligned}$$

Ex. 2.
Show that the variance and covariance estimators in the simple linear regression model are

$$\begin{aligned}
Var(\hat{\beta}) &= S_{xx}^{-1}\sigma^2 \\
Cov(\hat{\alpha}, \hat{\beta}) &= -\bar{x}S_{xx}^{-1}\sigma^2 \\
Var(\hat{\alpha}) &= \frac{\sigma^2}{n} + \bar{x}^2 S_{xx}^{-1}\sigma^2
\end{aligned}$$

Solution.

$$\begin{aligned}
Var(\hat{\beta}) &= Var(\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{S_{xx}}) \\
&= Var(\sum_{i=1}^n \frac{(x_i - \bar{x})}{S_{xx}} y_i) \\
&= \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{S_{xx}^2} Var(y_i) \\
&= \frac{1}{S_{xx}}\sigma^2,
\end{aligned}$$

where the second line follows because $\sum(x_i - \bar{x})\bar{y} = 0$, the third line follows because y_i's are independent and the last line follows since $\sum(x_i - \bar{x})^2 = S_{xx}$.

$$\begin{aligned}
Cov(\hat{\alpha}, \hat{\beta}) &= Cov(\bar{y} - \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{S_{xx}}\bar{x}, \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{S_{xx}}) \\
&= Cov(\sum_{i=1}^n (\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}}) y_i, \sum_{i=1}^n \frac{(x_i - \bar{x})}{S_{xx}} y_i) \\
&= \sum_{i=1}^n (\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}})(\frac{(x_i - \bar{x})}{S_{xx}})\sigma^2 \\
&= -\sum_{i=1}^n \frac{(x_i - \bar{x})^2\bar{x}}{S_{xx}^2}\sigma^2 \\
&= -\frac{\bar{x}}{S_{xx}}\sigma^2,
\end{aligned}$$

where the third line follows since y_i's are independent with constant variance σ^2, the fourth line follows because $\sum(x_i - \bar{x}) = 0$.

$$
\begin{aligned}
Var(\hat{\alpha}) &= Var(\bar{y} - \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{S_{xx}}\bar{x}) \\
&= Var(\sum_{i=1}^{n}(\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}})y_i) \\
&= \sum_{i=1}^{n}(\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}})^2\sigma^2 \\
&= (\frac{n}{n^2} + \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2\bar{x}^2}{S_{xx}^2})\sigma^2 \\
&= \frac{\sigma^2}{n} + \frac{\bar{x}^2}{S_{xx}}\sigma^2,
\end{aligned}
$$

where the fourth line follows since $\sum(x_i - \bar{x}) = 0$.

Ex. 3.

Consider the following linear regression model:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

where x_1, x_2 and the constant are three explanatory variables, y is the response variable, and ε is normally distributed noise. Suppose that 20 data points of $(x_{i,1}, x_{i,2}, y_i), i = 1, \ldots, 20$ are observed. The following are some simple statistics:

$$\bar{y} := \frac{1}{20}\sum_{i=1}^{20} y_i = 5, \quad \sum_{i=1}^{20}(y_i - \bar{y})^2 = 20, \quad \sum_{i=1}^{20}\hat{\varepsilon}_i^2 = 10$$

where $\hat{\varepsilon}_i$'s are the residuals of y_i's in the regression model.

1. Find an unbiased estimate of the variance of ε.
2. Test whether $\beta_1 = \beta_2 = 0$. Use 1% as the significance level.
3. Calculate the R-squared and the adjusted R-squared of the model.

Solution:

1. The unbiased estimate of the variance of ε is $\hat{\sigma}^2_{smpl} = \frac{S_{rsdl}}{n-p} = \frac{\sum_{i=1}^{20}(y_i - \hat{y}_i)^2}{n-p} = \frac{10}{20-3} = \frac{10}{17}$.
2. This is a F test with the reduced model being: $y = \alpha + \epsilon$, then we know $\hat{\alpha} = \bar{y}$, and $S_{rdcd,rsdl} = \sum_{i=1}^{20}(y_i - \bar{y})^2 = 20$, thus,

$$F = \frac{(S_{rdcd,rsdl} - S_{full,rsdl})/(p-q)}{S_{full,rsdl}/(n-p)} = \frac{(20-10)/(3-1)}{10/(20-3)} = \frac{17}{2} = 8.5$$

Where the threshold value for 1% significance level and d.f.s $p - q = 2$ and $n - p = 17$ of F distribution is ≈ 6.112. As $8.5 > 6.112$, we reject the null hypothesis $\beta_1 = \beta_2 = 0$.

3. Since one of the explanatory variables is the constant, then $R^2 = 1 - \frac{S_{rsdl}}{S_{ttl}} = 1 - \frac{\sum_{i=1}^{20}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{20}(y_i - \bar{y})^2} = 1 - \frac{10}{20} = 0.5$, and $R^2_{adj} = 1 - \frac{S_{rsdl}/(n-p)}{S_{ttl}/(n-1)} = 1 - \frac{10/17}{20/19} \approx 0.441$

Ex. 4.
Consider a multiple linear regression model

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon.$$

It is desired to test the null hypothesis $H_0 : c_2\beta_2 + c_3\beta_3 = 0$ versus the alternative hypothesis $H_1 : c_2\beta_2 + c_3\beta_3 \neq 0$, where c_2, c_3 are real numbers. Provide the exact rule for rejecting the null hypothesis if the desired significance level of the test is ρ.

Solution. We have that $Var(\hat{\beta}) = (\mathbf{X}^\top\mathbf{X})^{-1}\sigma^2$ and thus we have the standard error of $c_2\hat{\beta}_2 + c_3\hat{\beta}_3$ equal to

$$s.e.(c_2\hat{\beta}_2 + c_3\hat{\beta}_3) = \hat{\sigma}_{rsdl}\sqrt{c_2^2(\mathbf{X}^\top\mathbf{X})^{-1}_{22} + c_3^2(\mathbf{X}^\top\mathbf{X})^{-1}_{33} + 2c_2c_3(\mathbf{X}^\top\mathbf{X})^{-1}_{23}},$$

where $\hat{\sigma}^2_{rsdl} := \frac{1}{n-p}S_{rsdl}$ and $(\mathbf{X}^\top\mathbf{X})^{-1}_{ij}$ is the element of matrix $(\mathbf{X}^\top\mathbf{X})^{-1}$ on the ith row and jth column.

Since $c_2\hat{\beta}_2 + c_3\hat{\beta}_3$ is a normal random variable with mean $c_2\beta_2 + c_3\beta_3$, we have that

$$T = \frac{c_2\hat{\beta}_2 + c_3\hat{\beta}_3}{s.e.(c_2\hat{\beta}_2 + c_3\hat{\beta}_3)}$$

follows a student-t distribution with $n - p$ degrees of freedom under H_0. Therefore we reject the null hypothesis if $|T| > t_{n-p,1-\frac{\rho}{2}}$.

Ex. 5.
Consider the monthly return data of IBM and S&P500 index in period January 2010 through December 2023. You can download the data set from Yahoo finance or any other sources (please, report the source). Do the following:

(a) Test December effect, i.e., test whether the returns of IBM in December has different alpha and beta from the returns in the other months.

(b) Provide a 95% confidence interval for the difference between the beta of IBM returns in December and the beta of IBM returns in the other months.

Solution:
(The data sets for the S&P 500 Index and IBM can be downloaded from the yfinance package in Python)

Firstly, we download the close prices for IBM and S&P 500 index in period Jnuary 2010-December 2023, and compute their monthly returns in this period.

Let x denote the returns of S&P 500 index, and y denote the returns of IBM, assume the risk-free rate is 0 for simplicity. Then, we run a simple linear regression, $y = \alpha + \beta x + \epsilon$, the results of $\hat{\alpha}$ and $\hat{\beta}$ are the desired alpha and beta. Now, in order to test the December effect, we define a new

vector, Dec, where $Dec_i = 1$ when the i^{th} return data point is the return in December, and $Dec_i = 0$ otherwise.

Now, consider a new model,

$$y = \alpha_1 + \beta_1 x + \alpha_2 Dec + \beta_2 (x * Dec) + \epsilon$$

In this new model, if the data point is not from December, it will reduce to our original model $y = \alpha + \beta x + \epsilon$. However, if the data point is from December, we will have two new coefficients α_2, β_2 to measure December's impact on alpha and beta, i.e. α_2, β_2 are the differences of alpha and beta of the returns in December from the returns in the other months.

1. To test whether the returns of IBM in December has different alpha and beta from the returns in the other months is equivalent to test $H_0 : \alpha_2 = 0$ and $H_0 : \beta_2 = 0$. Using the results of linear regression in the Lecture Notes, we have: $\hat{\alpha_2} = -.0102$, $\hat{\beta_2} = .2141$, and the t-test statistics for $H_0 : \alpha_2 = 0$ and $H_0 : \beta_2 = 0$ are $-.983$ and 1.12. Since this is a student t-test with degree of freedom $n - p = 163$, the threshold value for significance level 1% and 5% are approximately 2.358 and 1.658, because $-.983 < 2.358$ for $H_0 : \alpha_2 = 0$ and $1.12 < 1.658$ for $H_0 : \beta_2 = 0$, we fail to reject $H_0 : \alpha_2 = 0$ and also cannot reject $H_0 : \beta_2 = 0$, i.e. we cannot assure that the alpha of the returns in December is different from that in the other months.

2. Based on our analysis above, the confidence interval desired is actually the 95 % confidence interval for β_2. Since $\hat{\beta_2} = .2141$, $SE(\hat{\beta_2}) = 0.191$, the threshold value for significance level 5% is approximately 1.658, the confidence interval for β_2 is [.19841 , 0.2297].

Refer to the .ipynb file for the code and full solution with the regression summary.