Multiple Least Squares Regression

Christopher Ting

July 14, 2018

Course Web Site: http://cting.x10host.com/CUHKSZ/CUHKSZ.html

Broad Lesson Plan

Topics Covered

- Fama-French 3-factor model
- Classical assumptions of linear regression
- Ordinary least squares in matrix form
- Projection matrix and residual matrix
- Small sample properties of OLS estimator–unbiasedness, normality, efficiency, unbiased variance of residuals
- Statistical inference and confidence region
- Analysis of variance

Linear Models

- There are p explanatory variables X to explain the independent variable Y.
- Given n sets of observations of X and Y, the linear specification of their relationship is given by

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p} + \epsilon_i \,,$$

  where ϵ_i is the noise, disturbances, or errors.
- There are various ways to estimate the y intercept and the p parameters β_j, $j = 1, 2, \ldots, p$.
- The parameter β_j is also called the loading on X_j.

Example: Fama-French 3-Factor Model

$$r_{st} - r_{ft} = \beta_0 + \beta_1 \, \mathrm{MKT}_t + \beta_2 \, \mathrm{SMB}_t + \beta_3 \, \mathrm{HML}_t + \epsilon_t$$

❄ **Market Factor**

$$\mathrm{MKT}_t = r_{mt} - r_{ft} \, .$$

❄ **Size Factor**

$$\mathrm{SMB}_t = \frac{1}{3}\big(\text{Small Value} + \text{Small Neutral} + \text{Small Growth}\big) - \frac{1}{3}\big(\text{Big Value} + \text{Big Neutral} + \text{Big Growth}\big).$$

❄ **Value Factor**

$$\mathrm{HML}_t = \frac{1}{2}\big(\text{Small Value} + \text{Big Value}\big) - \frac{1}{2}\big(\text{Small Growth} + \text{Big Growth}\big) \, .$$

Classical Assumptions

- Linearity of the conditional expectation

$$\mathbb{E}\big(Y_i \big| X_{i,1}, X_{i,2}, \dots X_{i,p}\big) = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p}\,.$$

- Independent noise

$$\epsilon_1, \epsilon_2, \dots, \epsilon_n \text{ are i.i.d.}$$

with zero mean, $\mathbb{E}\big(\epsilon_i\big)$ for all i.

- Constant variance

$$\mathbb{V}(\epsilon_i) = \sigma_\epsilon^2\,, \qquad \text{for all } i.$$

- Gaussian noise

$$\epsilon_i \sim N\big(0, \sigma_\epsilon^2\big)\,, \qquad \text{for all } i.$$

Implications of Assumptions

- The first assumption is implied by

$$\mathbb{E}\left(\epsilon_i \middle| X_{i,1}, X_{i,2}, \dots X_{i,p}\right) = 0\,.$$

- From the first assumption, we also have

$$\beta_j = \frac{\partial \mathbb{E}\left(Y_i \middle| X_{i,1}, X_{i,2}, \dots X_{i,p}\right)}{\partial X_{i,j}}\,.$$

- Therefore, β_j is the change in the expected value of Y_i when $X_{i,j}$ changes one unit.

Conditional Variance of Independent Variable

- All n observations of p Xs are organized into the data matrix

$$\boldsymbol{X} = \begin{pmatrix} 1 & X_{1,1} & X_{1,2} & \dots & X_{1,p-1} & X_{1,p} \\ 1 & X_{2,1} & X_{2,2} & \dots & X_{2,p-1} & X_{2,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & X_{n,1} & \vdots & \dots & X_{n,p-1} & X_{n,p} \end{pmatrix}$$

- What is the conditional variance of Y_i given all the observations $\boldsymbol{X}$?

- Answer: It is simply the variance of the error.

$$\mathbb{V}\left(Y_i \middle| \boldsymbol{X}\right) = \sigma_\epsilon^2 \,.$$

In Matrix Form

- The independent variable is a vector of n rows, i.e.,

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

- The parameters $\beta_0, \beta_1, \ldots, \beta_p$ is also assembled as a vector $\boldsymbol{\beta}$ of $1 + p$ rows.

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

- Then

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}\,.$$

Error

- The i-th row of the $\boldsymbol{X}$ matrix corresponding to Y_i is transposed to form a column vector

$$\boldsymbol{x}_i = \begin{pmatrix} 1 \\ X_{i,1} \\ X_{i,2} \\ \vdots \\ X_{i,p} \end{pmatrix}$$

- Hence, $\boldsymbol{x}_i^\top \boldsymbol{\beta}$ is a scalar equal to $\beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p}$

- Accordingly the error term at i is given by

$$\epsilon_i = Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\,.$$

Ordinary Least Squares

- The ordinary least squares (OLS) estimator is defined as the value that minimizes the sum of the squared errors

$$\widehat{\boldsymbol{\beta}} = \min_{\boldsymbol{\beta}} s(\boldsymbol{\beta}) \,,$$

where

$$\begin{aligned} s(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \epsilon_i^2 \\ &= \sum_{i=1}^{n} \left(Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)^2 \\ &= \left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\right)^\top \left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\right) \\ &= \boldsymbol{Y}^\top \boldsymbol{Y} - 2\boldsymbol{Y}^\top \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} \\ &= \| \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} \|^2 \end{aligned}$$

Vector of Parameter Estimates

- The OLS minimizes the Euclidean distance between $\boldsymbol{Y}$ and $\boldsymbol{X}\boldsymbol{\beta}$.
- Differentiate $s(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ yields the first-order condition

$$-2\boldsymbol{X}^\top \boldsymbol{Y} + 2\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} = \mathbf{0}\,.$$

- The solution is

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{Y}\,.$$

- To verify that this is a minimum, the sufficient condition is that the second derivative, namely, $\boldsymbol{X}^\top \boldsymbol{X}$ is positive. Indeed, the rank of $\boldsymbol{X}^\top \boldsymbol{X}$ is $k := 1 + p$ and hence $\boldsymbol{X}^\top \boldsymbol{X}$ is a positive definite $k \times k$ matrix.

Insight!

- The fitted values are the vector $\widehat{Y} = X\widehat{\beta}$.
- The residuals are the vector $\widehat{\epsilon} = Y - X\widehat{\beta}$.
- Note that
$$Y = X\beta + \epsilon = X\widehat{\beta} + \widehat{\epsilon}\,.$$
- Also, the first order conditions can be written as
$$\begin{aligned} X^\top Y - X^\top X\widehat{\beta} &= 0 \\ X^\top\left(Y - X\widehat{\beta}\right) &= 0 \\ X^\top\widehat{\epsilon} &= 0 \end{aligned}$$
- The OLS residuals $\widehat{\epsilon}$ are orthogonal to $X^\top$

Projection Matrix

- $X\widehat{\beta}$ is the projection of Y onto the span of X, which is

$$X\widehat{\beta} = X\left(X^\top X\right)^{-1}X^\top Y$$

- Therefore, the matrix that projects Y onto the span of X is

$$P_X := X\left(X^\top X\right)^{-1}X^\top$$

Orthogonal Projection Matrix

- We have

$$\begin{aligned}\boldsymbol{\epsilon} &= \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} \\ &= \boldsymbol{Y} - \boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top \boldsymbol{Y} \\ &= \left(\boldsymbol{I}_n - \boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\right)\boldsymbol{Y}\end{aligned}$$

- So the matrix that projects $\boldsymbol{Y}$ onto the space orthogonal to the span of $\boldsymbol{X}$ is

$$\boldsymbol{M}_X = \boldsymbol{I}_n - \boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top = \boldsymbol{I}_n - \boldsymbol{P}_X$$

Insight!

- Since $\widehat{\boldsymbol{\epsilon}} = \boldsymbol{M}_X \boldsymbol{Y}$, we have

$$\boldsymbol{Y} = \boldsymbol{P}_X \boldsymbol{Y} + \boldsymbol{M}_X \boldsymbol{Y} \, .$$

- These two projection matrices decompose the n dimensional vector $\boldsymbol{Y}$ into two orthogonal components: the portion that lies in the k-dimensional space defined by $\boldsymbol{X}$; and the portion that lies in the orthogonal $(n-k)$-dimensional space.

- Note that both $\boldsymbol{P}_X$ and $\boldsymbol{M}_X$ are symmetric and idempotent.

Small Sample Properties: Unbiasedness

- For $\widehat{\beta}$ we have

$$\begin{aligned}\widehat{\beta} &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{X}\beta + \boldsymbol{\epsilon}) \\ &= \beta + (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{\epsilon},\end{aligned}$$

  then applying the strong exogeneity assumption and the law of iterated expectation, we have

$$\begin{aligned}\mathbb{E}\Big((\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{\epsilon}\Big) &= \mathbb{E}\Big(\mathbb{E}((\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{\epsilon}|\boldsymbol{X})\Big) \\ &= \mathbb{E}\Big((\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \mathbb{E}(\boldsymbol{\epsilon}|\boldsymbol{X})\Big) \\ &= 0.\end{aligned}$$

- So the OLS estimator is unbiased.

Small Sample Properties: Conditional Normality

* $\widehat{\boldsymbol{\beta}}|\boldsymbol{X} \sim N\left(\boldsymbol{\beta}, (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\sigma_\epsilon^2\right)$, where $\sigma_\epsilon^2 = \mathbb{E}\left(\epsilon_i^2\right)$.
* Proof: Note that $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{\epsilon}$, then $\widehat{\boldsymbol{\beta}}$ is a linear function of $\boldsymbol{\epsilon}$.
* It is sufficient to obtain the conditional variance.

$$\begin{aligned}
\mathbb{V}(\widehat{\boldsymbol{\beta}}|\boldsymbol{X}) &= \mathbb{E}\left(\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^\top \middle| \boldsymbol{X}\right) \\
&= \mathbb{E}\left((\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top \boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \middle| \boldsymbol{X}\right) \\
&= (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \mathbb{E}\left(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top \middle| \boldsymbol{X}\right) \boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \\
&= (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\sigma_\epsilon^2.
\end{aligned}$$

* Notice that the above results is true irrespective of the distribution of $\boldsymbol{\epsilon}$.

Small Sample Properties: Efficiency

- (Gauss-Markov theorem): In the classical linear regression model, the OLS estimator, $\widehat{\beta}$, is the minimum variance linear unbiased estimator of β.
- Proof: Consider a new linear estimator $\widetilde{\beta} = \boldsymbol{C}\boldsymbol{Y}$, where $\boldsymbol{C}$ is given by

$$\boldsymbol{C} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top + \boldsymbol{D}\,,$$

  and $\boldsymbol{D}$ is a nonzero matrix.
- If this estimator is unbiased, then we must have $\boldsymbol{C}\boldsymbol{X} = \boldsymbol{I}$ since

$$\begin{aligned}\mathbb{E}(\boldsymbol{C}\boldsymbol{Y}) &= \mathbb{E}(\boldsymbol{C}\boldsymbol{X}\beta + \boldsymbol{C}\boldsymbol{\epsilon}) \\ &= \boldsymbol{C}\boldsymbol{X}\beta\end{aligned}$$

- Therefore, $\boldsymbol{C}\boldsymbol{X} = \boldsymbol{I}$.

Small Sample Properties: Efficiency (cont)

- The variance of an unbiased $\widetilde{\boldsymbol{\beta}}$ is

$$\mathbb{V}\left(\widetilde{\boldsymbol{\beta}}\right) = \boldsymbol{C}\boldsymbol{C}^{\top}\sigma_{\epsilon}^{2}.$$

- Since $\boldsymbol{D} = \boldsymbol{C} - (\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}$ and $\boldsymbol{C}\boldsymbol{X} = \boldsymbol{I}$, then $\boldsymbol{D}\boldsymbol{X} = \boldsymbol{0}$.
- Consequently,

$$\begin{aligned}
\mathbb{V}\left(\widetilde{\boldsymbol{\beta}}\middle|\boldsymbol{X}\right) &= \left(\boldsymbol{D} + (\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}\right)\left(\boldsymbol{D} + (\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}\right)^{\top}\sigma_{\epsilon}^{2} \\
&= \left(\boldsymbol{D}\boldsymbol{D}^{\top} + \left(\boldsymbol{X}^{\top}\boldsymbol{X}\right)^{-1}\right)\sigma_{\epsilon}^{2} \\
&= \boldsymbol{D}\boldsymbol{D}^{\top}\sigma_{\epsilon}^{2} + \mathbb{V}\left(\widehat{\boldsymbol{\beta}}\middle|\boldsymbol{X}\right)
\end{aligned}$$

- Since $\boldsymbol{D}\boldsymbol{D}^{\top}$ is positive definite, it follows that

$$\mathbb{V}\left(\widetilde{\boldsymbol{\beta}}\right) \geq \mathbb{V}\left(\widehat{\boldsymbol{\beta}}\right).$$

Small Sample Properties: Unbiased $\widehat{\sigma}_\epsilon^2$

- The estimate for σ_ϵ^2 is

$$\widehat{\sigma}_\epsilon^2 = \frac{1}{n-k}\widehat{\boldsymbol{\epsilon}}^\top\widehat{\boldsymbol{\epsilon}} = \frac{1}{n-k}\boldsymbol{\epsilon}^\top \boldsymbol{M}_X \boldsymbol{\epsilon}\,,$$

  where $\boldsymbol{M}_X$ is idempotent matrix that projects onto the space orthogonal to the span of $\boldsymbol{X}$ and $\widehat{\boldsymbol{\epsilon}} = \boldsymbol{M}_X \boldsymbol{Y}$.

- Proof:

$$\begin{aligned}
\mathbb{E}\big(\widehat{\sigma}_\epsilon^2\big) &= \frac{1}{n-k}\mathbb{E}\Big(\mathrm{Trace}\big(\boldsymbol{\epsilon}^\top \boldsymbol{M}_X \boldsymbol{\epsilon}\big)\Big) \\
&= \frac{1}{n-k}\mathbb{E}\Big(\mathrm{Trace}\big(\boldsymbol{M}_X \boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top\big)\Big) \\
&= \frac{1}{n-k}\mathrm{Trace}\big(\mathbb{E}\big(\boldsymbol{M}_X \boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top\big)\big) \\
&= \frac{1}{n-k}\sigma_\epsilon^2\,\mathrm{Trace}(\boldsymbol{M}_X) \\
&= \sigma_\epsilon^2.
\end{aligned}$$

Summary

- Parameter estimates for the linear model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is a k-vector:

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{Y}$$

- The variance-covariance of the estimate $\widehat{\boldsymbol{\beta}}$ is

$$\mathbb{V}\left(\widehat{\boldsymbol{\beta}}\right) = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \sigma_\epsilon^2$$

Statistical Inference

- The marginal distribution of any regression coefficient $\widehat{\beta}_j$ is normal with mean β_j and variance $\sigma_\epsilon^2 c_{jj}$, where c_{jj} is the j-th diagonal element of the matrix $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$.
- Since $\widehat{\boldsymbol{\beta}} \big| \boldsymbol{X} \sim N\big(\boldsymbol{\beta}, (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\sigma_\epsilon^2\big)$, we have, for all $j = 0, 1, \ldots, p$,

$$t \text{ statistic} = \frac{\widehat{\beta}_j - \beta_j}{\widehat{\sigma}_\epsilon \sqrt{c_{jj}}} \sim t_{n-k}$$

- The $\alpha\%$ significance level for β_j is

$$\widehat{\beta}_j - q\widehat{\sigma}_\epsilon \sqrt{c_{jj}} \leq \beta_j \leq \widehat{\beta}_j + q\widehat{\sigma}_\epsilon \sqrt{c_{jj}}\,.$$

  where q is the $\big(1 - \alpha/2\big)$-th quantile of the t_{n-k} distribution.
- The estimate $\widehat{\beta}_j$ is said to be statistically significant at the $\alpha\%$ significance level if the absolute value of the computed t statistic is greater than q.

Confidence Interval for Mean Response

- What is the confidence interval for the mean response $\boldsymbol{\beta}^\top \boldsymbol{x}$ for a given observation $\boldsymbol{x}$?
- Given the unbiased estimate $\widehat{\boldsymbol{\beta}}$, the variance of $\widehat{\boldsymbol{\beta}}^\top \boldsymbol{x}$ is

$$\mathbb{V}\left(\widehat{\boldsymbol{\beta}}^\top \boldsymbol{x}\right) = \mathbb{V}\left(\boldsymbol{x}^\top \widehat{\boldsymbol{\beta}}\right) = \boldsymbol{x}^\top \mathbb{C}(\widehat{\boldsymbol{\beta}})\boldsymbol{x} = \sigma_\epsilon^2 \boldsymbol{x}^\top \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{x}$$

- Hence a $100 \times (1-\alpha)\%$ confidence interval for the mean response $\boldsymbol{\beta}^\top \boldsymbol{x}$ is

$$\widehat{\boldsymbol{\beta}}^\top \boldsymbol{x} \pm q\sigma_\epsilon \sqrt{\boldsymbol{x}^\top \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{x}}$$

Prediction Interval for a New Observation

- Suppose a future observation of $\boldsymbol{x}$ is obtained. Then, since $\epsilon \sim N(0, \sigma_\epsilon^2)$, we have

$$\mathbb{V}\left(y - \widehat{\boldsymbol{\beta}}^\top \boldsymbol{x}\right) = \mathbb{V}(\epsilon) + \mathbb{V}\left(\widehat{\boldsymbol{\beta}}^\top \boldsymbol{x}\right)$$

- Hence a $100 \times (1-\alpha)\%$ prediction interval for y is

$$\widehat{\boldsymbol{\beta}}^\top \boldsymbol{x} \pm q\sigma_\epsilon \sqrt{1 + \boldsymbol{x}^\top \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{x}}$$

Confidence Region for All Regression Coefficients

- The statistic

$$\frac{\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)^{\top}\left(\boldsymbol{X}^{\top}\boldsymbol{X}\right)\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)/k}{\sigma_{\epsilon}^{2}} \sim F_{k,n-k}$$

- Let F_{α} be the $(1-\alpha)$-th quantitle of the $F_{k,n-k}$ distribution. Then the region that simultaneously contains all regression coefficients with probability $1-\alpha$ is

$$\left\{\boldsymbol{\beta}: \frac{\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)^{\top}\left(\boldsymbol{X}^{\top}\boldsymbol{X}\right)\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)/k}{\sigma_{\epsilon}^{2}} \leq F_{\alpha}\right\}.$$

Analysis of Variance

- The total variation in Y is the variation of Y with respect to $\overline{Y}$ in terms of the sum of squares.

$$\text{TSS} = \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2$$

- It is partitioned into
  - the variation that can be explained by $X_1, \ldots, X_p$. The explained sum of squares is

$$\text{ESS} = \sum_{i=1}^{n} \left(\widehat{Y}_i - \overline{Y}\right)^2$$

  - the variation that cannot be explained, which is the residual error sum of squares

$$\text{RSS} = \sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i\right)^2$$

Results

- Hence

$$\text{TSS} = \text{RSS} + \text{ESS}\,.$$

- The coefficient of determination for the regression, or R^2, is

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}\,.$$

- Note that TSS is purely a quantity obtained purely from the dependent variable Y.

Degrees of Freedom

- The number of degrees of freedom of a regression with an intercept and p explanatory variables is k.

- To estimate variance of the residuals, we use the residual mean sum of squares

$$\widehat{\sigma}_{\epsilon}^{2} = \frac{\sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i\right)^2}{n-k}$$

F Tests

- Suppose there are two models I and II. Model I is a subset of Model II. The excess regression sum of squares of Model II relative to Model I is

$$\begin{aligned} \text{RSS}\left(\text{II} \,\middle|\, \text{I}\right) &= \text{RSS for Model II} - \text{RSS for Model I} \\ &= \text{ESS for Model II} - \text{ESS for Model I} \end{aligned}$$

- Correspondingly, the relative degrees of freedom of Model II with k_{II} degrees of freedom to to Model I with k_{I} degrees of freedom is

$$\text{df}_{\text{I}I|I} = k_{\text{II}} - k_{\text{I}} \,.$$

- The F statistic is

$$\frac{\frac{\text{RSS}\left(\text{I}I|I\right)}{k_{\text{II}}-k_{\text{I}}}}{\widehat{\sigma}^2_\epsilon} \sim F_{\text{df}_{k_{\text{II}}-k_{\text{I}},n-k_{\text{II}}}}$$

Adjusted R^2

- The coefficient of determination R^2 is always increased by adding more explanatory variables to the linear regression model, even if they do not really explain Y.
- Recall that

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\frac{1}{n}\text{RSS}}{\frac{1}{n}\text{TSS}}$$

  where RSS is the residual sum of squares.
- The bias in R^2 can be removed, resulting in the adjusted R^2 denoted by $\overline{R}^2$

$$\overline{R}^2 = 1 - \frac{\frac{1}{n-k}\text{RSS}}{\frac{1}{n-1}\text{TSS}} = 1 - \frac{\widehat{\sigma}_\epsilon^2}{s^2}$$

  where s^2 is the sample variance of Y_i.

Model Selection

- Find a subset of explanatory variables that provides a parsimonious regression model by minimizing one of the following criteria
  - AIC
$$\text{AIC} = \ln\left(\widehat{\sigma}_\epsilon^2\right) + \frac{2k}{n}\,.$$
  - BIC
$$\text{BIC} = \ln\left(\widehat{\sigma}_\epsilon^2\right) + \frac{k}{n}\ln(n)\,.$$
  - C_p statistic [Mallows 1973]:
Suppose there are M explanatory variables, and $p \leq M$. Then C_p is defined as
$$C_p = \frac{\text{RSS}(p)}{\widehat{\sigma}_{\epsilon,M}^2} - n + 2k\,.$$

Newey-West Standard Errors

- The standard errors that are "robust" against heteroscedasticity and serial correlation were proposed by Newey and West in 1987.
- Define

$$\boldsymbol{Q} := \sum_{j=-k}^{k} \frac{k-|j|}{k} \frac{1}{T} \sum_{t=1}^{T} \widehat{\epsilon}_t \boldsymbol{x}_t \boldsymbol{x}_{t-j}^\top \widehat{\epsilon}_{t-j}$$

and

$$\boldsymbol{G} := \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{x}_t \boldsymbol{x}_t^\top$$

- Asymptotic variance-covariance matrix of the regression coefficients

$$\mathbb{V}\left(\widehat{\boldsymbol{\beta}}\right) = \frac{1}{T} \boldsymbol{G}^{-1} \boldsymbol{Q} \boldsymbol{G}^{-1}$$