ECON3389 Machine Learning in Economics

Module 1
Linear Regression II

Alberto Cappello

Department of Economics, Boston College

Fall 2024

Overview

Agenda:

- Motivation and interpretation
- Estimation/inference
- Variable selection

Readings:

- ISLR Ch.3, sections 3.2

Motivation

- Main issue with simple linear regression: it always violates our key assumption of $f(X)$ being a regression function:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

$$f(X) = \mathbb{E}[Y|X] \quad \Rightarrow \quad \mathbb{E}[\epsilon|X] = 0$$

- This is because in SLR error term ϵ contains all other factors that affects Y besides X, and among those factors some are bound to be related to X.

Motivation

- Main issue with simple linear regression: it always violates our key assumption of $f(X)$ being a regression function:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

$$f(X) = \mathbb{E}[Y|X] \quad \Rightarrow \quad \mathbb{E}[\epsilon|X] = 0$$

- This is because in SLR error term ϵ contains all other factors that affects Y besides X, and among those factors some are bound to be related to X.
- Prediction is also much more precise when one has access to a range of feature/predictors, as opposed to just a single one.

Motivation

- Main issue with simple linear regression: it always violates our key assumption of $f(X)$ being a regression function:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

$$f(X) = \mathbb{E}[Y|X] \quad \Rightarrow \quad \mathbb{E}[\epsilon|X] = 0$$

- This is because in SLR error term ϵ contains all other factors that affects Y besides X, and among those factors some are bound to be related to X.
- Prediction is also much more precise when one has access to a range of feature/predictors, as opposed to just a single one.
- MLR also allows to perform inference that is not possible in SLR, such as joint significance tests or relative importance of factors.

Multiple Linear Regression

We now have p explanatory variables $X_1, X_2, \ldots, X_p$:

$$Y = \beta_0 \cdot 1 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$$

- Given zero conditional mean assumption, β_j is the *average marginal effect* of X_j on Y:

$$\mathbb{E}[\epsilon|X] = 0 \quad \Rightarrow \quad \beta_j = \frac{\partial \mathbb{E}[Y|X]}{\partial X_j} = \frac{\mathbb{E}[\Delta Y|X]}{\Delta X_j}$$

  i.e. average effect on Y of a one unit increase in X_j holding all other factors fixed.

Interpreting Regression Coefficients

- Is the effect of X on Y *causal*? Generally NO.

Interpreting Regression Coefficients

- Is the effect of X on Y *causal*? Generally NO.
- In real life data regressors are rarely uncorrelated, which leads to two major issues:
  - Precision of estimation for individual coefficients decreases, sometimes dramatically.
  - Claims of causality are harder to make — when X_j changes, everything else may change as well.

Interpreting Regression Coefficients

- Is the effect of X on Y *causal*? Generally NO.
- In real life data regressors are rarely uncorrelated, which leads to two major issues:
  - Precision of estimation for individual coefficients decreases, sometimes dramatically.
  - Claims of causality are harder to make — when X_j changes, everything else may change as well.
- There maybe omitted variables within the ϵ which maybe correlatred with X

Interpreting Regression Coefficients

- Is the effect of X on Y *causal*? Generally NO.
- In real life data regressors are rarely uncorrelated, which leads to two major issues:
  - Precision of estimation for individual coefficients decreases, sometimes dramatically.
  - Claims of causality are harder to make — when X_j changes, everything else may change as well.
- There maybe omitted variables within the ϵ which maybe correlatred with X
- It is possible to design an experimental study in a way that guarantees that predictors are uncorrelated or even completely independent.
  - Each coefficient can then be precisely estimated and tested separately from all others.
  - One can make causal interpretations.

Estimation problems in MLR

- **Multicollinearity**

Estimation problems in MLR

- **Multicollinearity**
  - High correlation coefficient

Estimation problems in MLR

- **Multicollinearity**
  - High correlation coefficient
  - High R^2 with low t-stats

Estimation problems in MLR

- **Multicollinearity**
  - High correlation coefficient
  - High R^2 with low t-stats
  - High Variance Inflation Factor (VIF)

- **Omitted variables**
  Example 1: Y is hourly wage, X_1 is years of working experience, X_2 is gender (male/female). We can expect $\widehat{\beta}_1 > 0$, but what about $\widehat{\beta}_2$? If it is statistically different from zero, does this imply gender discrimination?

Estimation in MLR

- Our sample is $\{y_i, x_{1i}, x_{2i}, \ldots, x_{pi}\}_{i=1}^{n}$. Given estimates $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p$, our predicted (estimated) outcome is

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{1i} + \widehat{\beta}_2 x_{2i} + \ldots + \widehat{\beta}_p x_{pi}$$

- The same logic as in SLR leads us to OLS estimates as the ones that minimize RSS:

$$RSS(\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p) = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \rightarrow \min_{\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p}$$

- Closed form solutions still exist, but they become cumbersome when usual scalar notation is used.

Matrix Notation

$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix} \quad \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{21} & \dots & x_{p1} \\ 1 & x_{12} & x_{22} & \dots & x_{p2} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{1n} & x_{2n} & \dots & x_{pn} \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_p \end{pmatrix} \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_0 \\ \epsilon_1 \\ \dots \\ \epsilon_n \end{pmatrix}$$

Matrix Notation

$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix} \quad \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{21} & \dots & x_{p1} \\ 1 & x_{12} & x_{22} & \dots & x_{p2} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{1n} & x_{2n} & \dots & x_{pn} \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_p \end{pmatrix} \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_0 \\ \epsilon_1 \\ \dots \\ \epsilon_n \end{pmatrix}$$

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \qquad \widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} \qquad \boldsymbol{e} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$$

Matrix Notation

$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix} \quad \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{21} & \dots & x_{p1} \\ 1 & x_{12} & x_{22} & \dots & x_{p2} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{1n} & x_{2n} & \dots & x_{pn} \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_p \end{pmatrix} \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_0 \\ \epsilon_1 \\ \dots \\ \epsilon_n \end{pmatrix}$$

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \qquad \widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} \qquad \boldsymbol{e} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$$

Then $RSS = \boldsymbol{e}'\boldsymbol{e}$ and our OLS estimates are defined as

$$\widehat{\boldsymbol{\beta}}_{OLS} = \underset{\widehat{\boldsymbol{\beta}}}{\text{argmin}}\, RSS = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

Small sample properties

- ZCM assumption ensures that OLS is unbiased:

$$\mathbb{E}[\widehat{\boldsymbol{\beta}}_{OLS}|\boldsymbol{X}] = \mathbb{E}\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}|\boldsymbol{X}\right] = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}] =$$

$$= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbb{E}[\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}|\boldsymbol{X}] = \underbrace{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}}_{=\mathbb{I}}\boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\underbrace{\mathbb{E}[\boldsymbol{\epsilon}|\boldsymbol{X}]}_{=0} =$$

$$= \boldsymbol{\beta}$$

Small sample properties

- ZCM assumption ensures that OLS is unbiased:

$$\mathbb{E}[\widehat{\boldsymbol{\beta}}_{OLS}|\boldsymbol{X}] = \mathbb{E}\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}|\boldsymbol{X}\right] = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbb{E}\left[\boldsymbol{Y}|\boldsymbol{X}\right] =$$
$$= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbb{E}\left[\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}|\boldsymbol{X}\right] = \underbrace{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}}_{=\mathbb{I}}\boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\underbrace{\mathbb{E}\left[\boldsymbol{\epsilon}|\boldsymbol{X}\right]}_{=0} =$$
$$= \boldsymbol{\beta}$$

- If $\boldsymbol{\epsilon}$ is homoscedastic with no serial correlation, then the variance of OLS is

$$Var[\widehat{\boldsymbol{\beta}}_{OLS}|\boldsymbol{X}] = Var\left[\boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\epsilon}|\boldsymbol{X}\right] = Var\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\epsilon}|\boldsymbol{X}\right] =$$
$$= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'Var\left[\boldsymbol{\epsilon}|\boldsymbol{X}\right]\left((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right)' = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\sigma^2\mathbb{I})\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} =$$
$$= \sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} = \sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$$

Small Sample Properties

- Just like in SLR, under assumptions of $\mathbb{E}[\boldsymbol{\epsilon}|\boldsymbol{X}] = 0$ and $Var[\boldsymbol{\epsilon}|\boldsymbol{X}] = \sigma^2\mathbb{I}$ OLS is BLUE — has least variance among all linear unbiased estimators.

Small Sample Properties

- Just like in SLR, under assumptions of $\mathbb{E}[\boldsymbol{\epsilon}|\boldsymbol{X}] = 0$ and $Var[\boldsymbol{\epsilon}|\boldsymbol{X}] = \sigma^2\mathbb{I}$ OLS is BLUE — has least variance among all linear unbiased estimators.
- If one assumes normality of the error term ϵ, OLS estimates have exact normal sampling distributions:

$$\widehat{\boldsymbol{\beta}}_{OLS} \sim \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}\right)$$

$$\frac{\widehat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}_{OLS}}{se(\widehat{\boldsymbol{\beta}}_{OLS})} \sim t_{n-p-1}$$

- However, assuming normality of the error term is often just as unrealistic in MLR as in SLR. In addition, unbiasedness only works with repeated samples, while we usually have access only to a single sample.

Large Sample Properties (Asymptotics)

Good news — just like in SLR, OLS estimates in MLR are consistent

$$\text{plim}\,\widehat{\boldsymbol{\beta}}_{OLS} = \text{plim}\left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y}\right) = \text{plim}\,\left(\boldsymbol{\beta} + \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{\epsilon}\right) =$$

$$= \boldsymbol{\beta} + \text{plim}\left(\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\frac{1}{n}\boldsymbol{X}'\boldsymbol{\epsilon}\right) = \boldsymbol{\beta} + \text{plim}\left(\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right)\cdot\text{plim}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{\epsilon}\right) =$$

$$= \boldsymbol{\beta} + \mathbb{E}\left[X'X\right]\cdot\mathbb{E}\left[X\boldsymbol{\epsilon}\right] = \boldsymbol{\beta}$$

and asymptotically normal

Large Sample Properties (Asymptotics)

Good news — just like in SLR, OLS estimates in MLR are consistent

$$\operatorname{plim} \widehat{\boldsymbol{\beta}}_{OLS} = \operatorname{plim} \left(\left(\boldsymbol{X}'\boldsymbol{X} \right)^{-1} \boldsymbol{X}'\boldsymbol{Y} \right) = \operatorname{plim} \left(\boldsymbol{\beta} + \left(\boldsymbol{X}'\boldsymbol{X} \right)^{-1} \boldsymbol{X}'\boldsymbol{\epsilon} \right) =$$

$$= \boldsymbol{\beta} + \operatorname{plim} \left(\left(\frac{1}{n} \boldsymbol{X}'\boldsymbol{X} \right)^{-1} \frac{1}{n} \boldsymbol{X}'\boldsymbol{\epsilon} \right) = \boldsymbol{\beta} + \operatorname{plim} \left(\left(\frac{1}{n} \boldsymbol{X}'\boldsymbol{X} \right)^{-1} \right) \cdot \operatorname{plim} \left(\frac{1}{n} \boldsymbol{X}'\boldsymbol{\epsilon} \right) =$$

$$= \boldsymbol{\beta} + \mathbb{E}\left[X'X \right] \cdot \mathbb{E}\left[X\boldsymbol{\epsilon} \right] = \boldsymbol{\beta}$$

and asymptotically normal

$$\widehat{\boldsymbol{\beta}}_{OLS} \underset{n \to \infty}{\overset{d}{\sim}} \mathcal{N}\left(\boldsymbol{\beta}, \frac{\sigma^2}{n} \left(\boldsymbol{X}'\boldsymbol{X} \right)^{-1} \right)$$

$$\sqrt{n} \left(\widehat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}_{OLS} \right) \underset{n \to \infty}{\overset{d}{\sim}} \mathcal{N}\left(0, \sigma^2 \left(\boldsymbol{X}'\boldsymbol{X} \right)^{-1} \right)$$

Statistical Inference

- Standard t-test for significance can interpreted in the same way as in SLR
- Results from example with advertising data, using all three variables:

Variable	Coefficient	SE	t	p-value
Intercept	2.939	0.3119	9.42	<0.0001
TV	0.046	0.0014	32.81	<0.0001
radio	0.189	0.0086	21.89	<0.0001
newspaper	-0.001	0.0059	-0.18	<0.8599

- Correlation matrix:

Variable	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.00000

Statistical Inference

MLR offers much wider range of possible analysis avenues:

- Is at least one of the predictors $X_1, X_2, \ldots, X_p$ useful in predicting the response?
- Do all the predictors help to explain Y, or is only a subset of the predictors useful?
- How well does the model fit the data?
- Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

F-test for linear restrictions

- Standard significance tests only look at one variable at a time. What if we want to asses the *joint significance* of several variables at once?
- This means imposing multiple restrictions at once, e.g. $k = 3$ restrictions:

$$H_0 : \beta_1 = \beta_4 = \beta_5 = 0$$

F-test for linear restrictions

- Standard significance tests only look at one variable at a time. What if we want to asses the *joint significance* of several variables at once?
- This means imposing multiple restrictions at once, e.g. $k = 3$ restrictions:

$$H_0 : \beta_1 = \beta_4 = \beta_5 = 0$$

- The idea of an *F-test* is to compare how much worse our model's fit gets once we impose those restrictions and run OLS with them:

$$F = \frac{(RSS_r - RSS_{ur})/k}{RSS_{ur}/(n-k-1)} = \frac{(R^2_{ur} - R^2_r)/k}{(1-R^2_{ur})/(n-p-1)} \sim \mathcal{F}_{k,n-p-1}$$

F-test for linear restrictions

- Standard significance tests only look at one variable at a time. What if we want to asses the *joint significance* of several variables at once?
- This means imposing multiple restrictions at once, e.g. $k = 3$ restrictions:

$$H_0 : \beta_1 = \beta_4 = \beta_5 = 0$$

- The idea of an *F-test* is to compare how much worse our model's fit gets once we impose those restrictions and run OLS with them:

$$F = \frac{(RSS_r - RSS_{ur})/k}{RSS_{ur}/(n-k-1)} = \frac{(R^2_{ur} - R^2_r)/k}{(1-R^2_{ur})/(n-p-1)} \sim \mathcal{F}_{k,n-p-1}$$

- If $F > \mathcal{F}^{\alpha}_{k,n-p-1}$, we reject H_0 on significance level α.

Is at least one predictor useful?

- This questions corresponds to special case of

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$$

- In Econometrics this is known as regression significance test, and it is automatically performed for every linear regression.
- For our advertising data the result of this test is

$$R^2 = 0.897, \quad F = 570, \quad \text{p-value} < 0.00001 \Rightarrow H_0 \text{ is rejected}$$

Which variables are important?

- Generally we don't have any knowledge as to which variables we should test for joint significance

Which variables are important?

- Generally we don't have any knowledge as to which variables we should test for joint significance
- We have lots of independent variables. We need to decide which is the best model? For our current purposes best = highest Adj R squared

Which variables are important?

- Generally we don't have any knowledge as to which variables we should test for joint significance
- We have lots of independent variables. We need to decide which is the best model? For our current purposes best = highest Adj R squared
- We have to decide two things: The number of variables to include in our model and which variables are those?

Which variables are important?

- Generally we don't have any knowledge as to which variables we should test for joint significance
- We have lots of independent variables. We need to decide which is the best model? For our current purposes best = highest Adj R squared
- We have to decide two things: The number of variables to include in our model and which variables are those?
- The most direct approach is called *best subsets* regression: we compute OLS fit for all possible subsets and then choose between them based on some criterion that balances training error with model size.
  - But this is often not feasible, since they are 2^p possible subsets of p regressors, e.g. with $p = 40$ there are over a billion models!
- Instead, the two most commonly used approaches are *forward selection* and *backward selection*.
- We can also use statistics like Mallow's C_p, Akaike information criterion (AIC), Bayesian information criterion (BIC), Cross-validation (CV)