ECON3389 Machine Learning in Economics

Module 1
Linear Regression III

Alberto Cappello

Department of Economics, Boston College

Fall 2024

Overview

Agenda:

- Qualitative predictors
- Interaction and non-linearities
- Issues in MLR

Readings:

- ISLR Ch.3, sections 3.4

Qualitative predictors

- OLS is a mathematical algorithm that find optimal solutions over a space of $p+1$ numerical variables $Y, X_1, X_2, \ldots$
- But not every observable feature/predictor has a natural numerical scale to be measured along.
- *Qualitative* or *factor* variables such as gender, race, city district or education major clearly are important in many statistical applications, but all of them lack numerical scale.
- The solution is to split the data for every such variable into non-overlapping groups and assign a binary 0/1 variable to identify every group.

Qualitative predictors

- The simplest case is when our qualitative predictor has only two possible values in the data, e.g. having a college degree:

$$college_i = \begin{cases} 0, \text{person i does not have a college degree} \\ 1, \text{person i has a college degree} \end{cases}$$

- Then a wage equation can take the form of

$$wage_i = \beta_0 + \beta_1 exper_i + \beta_2 college_i + \epsilon_i$$

or

$$wage_i = \begin{cases} \beta_0 + \beta_1 exper_i + & \epsilon_i, \quad \text{no college degree} \\ \beta_0 + \beta_1 exper_i + \beta_2 + & \epsilon_i, \quad \text{college degree} \end{cases}$$

- In this case β_2 gains a very special interpretation: it stands for a ceteris paribus fixed difference in average wage for college educated vs non-college educated workers.

Qualitative predictors

- If we have k possible values (groups), we need to create $k-1$ *dummy variables* that take values 0 and 1 to differentiate between $k-1$ groups and a baseline group.
- Each individual dummy variable will show the fixed difference between one of the groups and the baseline group. The difference between two dummies will show the fixed difference between those two groups only.

Qualitative predictors

- Suppose our qualitative predictor takes three possible values in the data, e.g. major $\in$ economics, maths, physics

$$economics_i = \begin{cases} 0, \text{person i does not have an economics degree} \\ 1, \text{person i has an economics degree} \end{cases}$$

$$math_i = \begin{cases} 0, \text{person i does not have a math degree} \\ 1, \text{person i has a math degree} \end{cases}$$

- Then a wage equation is

$$wage_i = \beta_0 + \beta_1 exper_i + \beta_2 economics_i + \beta_3 math_i + \epsilon_i$$

or

$$wage_i = \begin{cases} \beta_0 + \beta_1 exper_i + & \epsilon_i, & \text{neither econ nor math degree} \\ \beta_0 + \beta_1 exper_i + \beta_2 + & \epsilon_i, & \text{economics degree} \\ \beta_0 + \beta_1 exper_i + & \beta_3 + \epsilon_i, & \text{math degree} \end{cases}$$

Extensions of Linear Model

- The main issue with linear model is that all variables have fixed marginal effects:

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

- This is totally unrealistic in many cases — e.g. effect of years of education, standardized test scores, number of kids, etc.
- There are three most common ways to change that and yet retain the useful simplicity of a linear model: interactions, non-linear transformations and higher order polynomials.

Interactions

- In our previous analysis of the advertising data, we assumed that the effect on `sales` of increasing one advertising medium is independent of the amount spent on the other media.
- For example, in the model

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \epsilon$$

the average effect on `sales` of a one-unit increase in `TV` is always β_1, regardless of the amount spent on `radio`.

Interactions

- But suppose that spending money on radio advertising actually increases the effectiveness of TV advertising, so that the slope term for `TV` should increase as `radio` increases.
- In this situation, given a fixed budget of \$100 000, spending half on `radio` and half on `TV` may increase sales more than allocating the entire amount to either `TV` or to `radio`.
- In marketing, this is known as a *synergy effect*, and in statistics it is referred to as an *interaction effect*.
- We can capture this using the model

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{TV} \times \texttt{radio} + \epsilon$$

- the average effect on `sales` of a one-unit increase in `TV` is $(\beta_1 + \beta_3 \times \texttt{radio})$, which depends on `radio` spend

Interactions

- A more interesting way to add interactions is by interacting usual numerical variables with dummy (qualitative) variables:

$$wage_i = \beta_0 + \beta_1\, exper_i + \beta_2\, college_i + \beta_3\, exper_i \times college_i + \epsilon_i$$

- The model above allows college graduates to have not only fixed difference in wages, but also a different return to experience:

$$wage_i = \begin{cases} \beta_0 + & \beta_1 exper_i + & \epsilon_i, & \text{no college degree} \\ \beta_0 + \beta_2 + & (\beta_1 + \beta_3) exper_i + & \epsilon_i, & \text{college degree} \end{cases}$$

Non-Linear Transformations

- Another way to achieve non-constant marginal effects is to use non-linear transformation of regressors, e.g. natural logs:

$$\log(wage)_i = \beta_0 + \beta_1 \log(exper)_i + \beta_2 college_i + \epsilon_i$$

- The model still remains linear in parameters (betas), so OLS estimation proceeds as usual. However, marginal effect of years of experience on wage is no longer the same for every extra year of experience:

$$\beta_1 = \frac{\mathbb{E}[\Delta \log(wage)|\ldots]}{\Delta \log(exper)} \approx \frac{\mathbb{E}[\%\Delta wage|\ldots]}{\%\Delta exper}$$

- The latter fraction is known as *elasticity* and plays a very important role in Economics.

Polynomials

- The most straightforward way to make marginal effects vary with values of X is to add powers of corresponding regressors:

$$wage_i = \beta_0 + \beta_1\, exper_i + \beta_2\, exper_i^2 + \epsilon_i$$

- Marginal effects become variable:

$$\frac{\mathbb{E}[\Delta wage|\ldots]}{\Delta exper} = \beta_1 + 2\beta_2\, exper$$

  but individual β_j no longer have meaningful interpretation for most cases
- Higher order polynomial regression are very good at predicting, but mostly useless for inference (more on this in later chapters).

Issues in MLR

- There are lots of potential issue that may invalidate the results of any linear regression estimation.
- Most of those issues invalidate inference, as prediction power is much more robust. That is why many machine learning specialists do not care about these issues, but in Economics one must be very aware of potential pitfalls.
- Most common issues in cross-section data (i.e. not a time-series) are *multicollinearity*, *heteroscedasticity* and *outliers*.

Multicollinearity

- If ϵ is homoscedastic with no serial correlation, then the variance of OLS is

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

- That formula relies on matrix $\boldsymbol{X}'\boldsymbol{X}$ being invertible, i.e. having full rank. This is usually not a problem, unless some of your predictors are perfectly linearly related (e.g. including k dummy variables for k categories).
- However, if some variables in $\boldsymbol{X}$ are strongly, but not perfectly correlated, OLS will still work, but the variance of estimates will be very high. This is known as *multicollinearity* — a situation where OLS estimates become imprecise due to multiple predictors exhibiting notable linear co-movement.

Heteroscedasticity

- If $Var[\epsilon|\boldsymbol{X}] = \Sigma \neq \sigma^2\mathbb{I}$, then true OLS variance becomes

$$Var[\widehat{\beta}_{OLS}|\boldsymbol{X}] = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\Sigma\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

  and the usual regression standard errors become invalid.
- This is because OLS by default assumes that the amount of noise (variance of ϵ_i) in every observation i does not depend on X_{ij}, and thus "weights" every observation equally in terms of its noise/signal ratio, whatever the values of X_{ij}.
- But under HS some observations contain more noise and others contain less noise, depending on values of some or all of X, meaning that OLS no longer properly evaluates how precise $\widehat{\beta}_j$ estimates are.

Heteroscedasticity

- Two most common solutions for HS are *weighted least squares* and *robust standard errors*.
- The first option is about calculating weights w_i for each observation that are proportional to variance of ϵ_i in each observation. Then weighted least squares becomes

$$\widehat{\beta}_{WLS} = \left(\boldsymbol{X}'\boldsymbol{W}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{W}^{-1}\boldsymbol{Y}$$

- The second option involves estimating the matrix $\Sigma = Var\left[\epsilon|\boldsymbol{X}\right]$ and adjusting estimated standard errors without changing OLS estimates themselves:

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

Outliers

- Because OLS minimizes sum of squared distances between $\widehat{y}_i$ and y_i, outliers in terms of values of y_i gain significantly more weight on final position of estimated regression line.
- One must always perform visual analysis (e.g. scatter plots and boxplots) to identify the presence of outliers in the data.
- If such observations are found, they need to be treated differently — either as a separate sample of data or via the use of dummy variables.
- Alternatively, one may try using log-transformation of some or all variables to reduce the relative distance between data points.