Quantitative Economics: Regression

Regression with the Population

Ian Crawford

Department of Economics

Nuffield College

Trinity Term, 2018

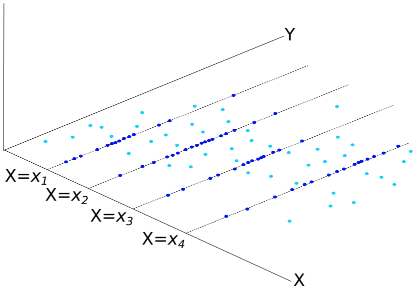
Weeks 3 & 4

Regression

1. Regression with the population
2. Regression with a sample
3. Regression and causal inference.

- The population CEF of a variable Y given a covariate X is the population expectation, or mean, of Y given X.
- The CEF is written
$$\mathbb{E}\left[Y|X\right]$$
and is regarded as a function of X.
- It's a description of how the mean of something (Y) varies as you change something else (X).
- Evaluated at a particular value of X, say x, the notation is
$$\mathbb{E}\left[Y|X=x\right]$$

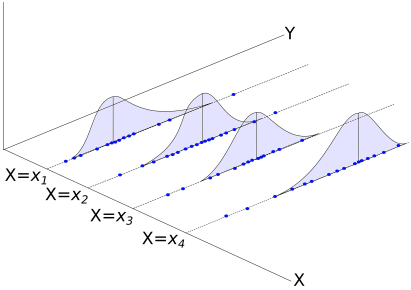
The Conditional Expectation Function
The Object of Interest



- This shows a scatter plot of $\{Y, X\}$.
- To simplify the figure we will focus on particular values of X denoted $x_1, \ldots, x_4$.

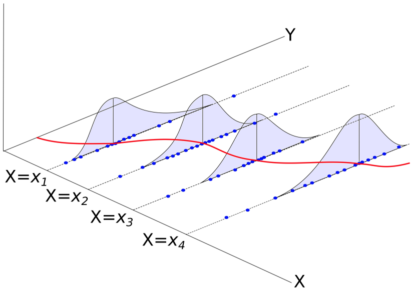
The Conditional Expectation Function

The Object of Interest



- We can think about the distribution of Y when X is equal to each of these values.
- The means of each of these conditional distributions are the conditional means

$$\mathbb{E}[Y|X = x_i]$$



- If we think of $\mathbb{E}[Y|X = x]$ and vary the value of x this traces out a function.
- The red curve shows the graph of the conditional mean of Y as we vary x.
- It's called the Conditional Expectation Function (CEF).

The CEF decomposition
Overview

- We can (almost) always write:

$$Y = \mathbb{E}[Y|X] + e$$

- The residual e enjoys certain properties:
  - it is mean zero $\mathbb{E}[e] = 0$
  - it is mean independent of X, that is $\mathbb{E}[e|X] = \mathbb{E}[e]$
  - it is uncorrelated with any function of X

Mean Zero

This is the property that $\mathbb{E}\left[e\right] = 0$. To show this take the CEF decomposition and take expectations

$$\begin{aligned} Y &= \mathbb{E}\left[Y|X\right] + e \\ \mathbb{E}\left[Y\right] &= \mathbb{E}\left[\mathbb{E}\left[Y|X\right] + e\right] \\ \mathbb{E}\left[Y\right] &= \mathbb{E}\left[\mathbb{E}\left[Y|X\right]\right] + \mathbb{E}\left[e\right] \end{aligned}$$

(the last line is due to the linearity of the $\mathbb{E}$ operator). The Law of Iterated Expectations says:

$$\mathbb{E}[Y] = \mathbb{E}\left[\mathbb{E}\left[Y|X\right]\right]$$

Hence $\mathbb{E}\left[e\right] = 0$.

Mean independence

This is the property that

$$\mathbb{E}\left[e|X\right] = \mathbb{E}\left[e\right]$$

To show this first rewrite the CEF decomposition

$$Y = \mathbb{E}\left[Y|X\right] + e$$

in terms of the residual:

$$e = Y - \mathbb{E}\left[Y|X\right]$$

Mean independence

Now sub for e in $\mathbb{E}\left[e|X\right]$:

$$\mathbb{E}\left[e|X\right] = \mathbb{E}\left(Y - \mathbb{E}\left[Y|X\right]|X\right)$$

Thanks to linearity of the $\mathbb{E}$ operator you can do the following:

$$\mathbb{E}\left[e|X\right] = \mathbb{E}\left[Y|X\right] - \mathbb{E}\left(\mathbb{E}\left[Y|X\right]|X\right)$$

Then to the Law of Iterated Expectations:[a]

$$\mathbb{E}\left[e|X\right] = \mathbb{E}\left[Y|X\right] - \mathbb{E}\left[Y|X\right] = 0$$

which is the mean independence property.

[a] Which works for conditional means too: $\mathbb{E}\left(\mathbb{E}\left[Y|X\right]|X\right) = \mathbb{E}\left[Y|X\right]$

Uncorrelatedness

Consider any function of the variable X. Call this function $h(X)$. Multiply it by the residual:

$$h(X)\, e$$

Then take the expectation:

$$\mathbb{E}\left[h(X)\, e\right]$$

We need to show that this is zero.

The CEF decomposition

Uncorrelatedness

By the Law of Iterated Expectations again[a]

$$\mathbb{E}\left[h\left(X\right)e\right] = \mathbb{E}\left[h\left(X\right)\mathbb{E}\left[e|X\right]\right]$$

Finally, by mean independence, $\mathbb{E}\left[e|X\right] = 0$ so

$$\mathbb{E}\left[h\left(X\right)e\right] = 0$$

[a] Here's the LIE: $\mathbb{E}\left[Y\right] = \mathbb{E}\left(\mathbb{E}\left[Y|X\right]\right)$. Now write $h(X)e = Y$ and sub in: $\mathbb{E}\left[h(X)e\right] = \mathbb{E}\left(\mathbb{E}\left[h(X)e|X\right]\right)$. Taking expectation of $h(X)e$ holding X fixed means $h(X)$ is fixed too, so $\mathbb{E}\left(\mathbb{E}\left[h(X)e|X\right]\right) = \mathbb{E}\left(h(X)\mathbb{E}\left[e|X\right]\right)$.

The CEF decomposition

$$Y = \mathbb{E}\left[Y|X\right] + e$$

- This means that we can take any variable Y and decompose it into two bits
  - a description of how, on average, it varies with respect to some other variable X,
  - a residual which is mean independent of X.
- More intuitively it says that we can break up Y into
  - a bit which is “explained" by X,
  - a bit which isn't.

The CEF as a Predictor

- The CEF can be used to predict the expected value of Y for different values of X.
- Predictors can be good, or bad. We use a *loss function* to help us determine whether a predictor is any good.
- A *best predictor* is one that minimises *expected loss*.
- There are many interesting loss functions but the leading example is the *squared loss*, which you have already seen in the context of establishing the mean as measure of location.

The CEF as a Predictor

- Let $m(X)$ denote a predictor of Y which uses X to predict Y. Let the error/deviation between Y and the prediction be

$$e = Y - m(X)$$

- The squared loss

$$e^2$$

penalises bigger prediction errors by more than smaller prediction errors; assigns no loss to predictions which are correct: $0^2 = 0$; penalises bad predictions symmetrically: $e^2 = (-e)^2$

- Its expected value

$$\mathbb{E}\left[e^2\right]$$

is proportional to the squared "distance" between the predictor and the data.

The CEF as a Predictor

- Suppose we want to select the best predictor in the least-squares sense.
- The CEF is the solution.
- Formally the problem is

$$\min_{m(X)} \mathbb{E}\left[e^2\right]$$

i.e. "choose a function $m(X)$ to minimise the expected loss when $m(X)$ is used to predict Y."

The CEF as a Predictor

Best predictor

Firstly take the residual $e = Y - m(X)$ and square it

$$e^2 = [Y - m(X)]^2$$

Now add and subtract $\mathbb{E}[Y|X]$

$$e^2 = [(Y - \mathbb{E}[Y|X]) + (\mathbb{E}[Y|X] - m(X))]^2$$

Expand it out

$$\begin{aligned} e^2 \quad = \quad & (Y - \mathbb{E}[Y|X])^2 \\ & + 2(Y - \mathbb{E}[Y|X])(\mathbb{E}[Y|X] - m(X)) \\ & + (\mathbb{E}[Y|X] - m(X))^2 \end{aligned}$$

The CEF as a Predictor

Best predictor

$$
\begin{aligned}
e^2 &= \left(Y - \mathbb{E}\left[Y|X\right]\right)^2 \\
&\quad + 2\left(Y - \mathbb{E}\left[Y|X\right]\right)\left(\mathbb{E}\left[Y|X\right] - m\left(X\right)\right) \\
&\quad + \left(\mathbb{E}\left[Y|X\right] - m\left(X\right)\right)^2
\end{aligned}
$$

This describes the loss.
We need to minimise its expectation by choosing $m\left(X\right)$.
Remain calm: the first term doesn't involve $m\left(X\right)$ so we can disregard it.

The CEF as a Predictor

Best predictor

The second term is

$$2\left(Y - \mathbb{E}\left[Y|X\right]\right)\left(\mathbb{E}\left[Y|X\right] - m\left(X\right)\right)$$

We can write this as

$$2e\left(\mathbb{E}\left[Y|X\right] - m\left(X\right)\right) = h\left(X\right)e$$

where $h\left(X\right) = 2\left(\mathbb{E}\left[Y|X\right] - m\left(X\right)\right)$.
Remain calm: thanks to the CEF decomposition we can rely on the fact that $\mathbb{E}\left[h\left(X\right)e\right] = 0$ so we can disregard this term too.

The CEF as a Predictor

Best predictor

The final term is $\left(\mathbb{E}\left[Y|X\right] - m\left(X\right)\right)^2$. This is the only one which matters.

Remain calm: this is simple quadratic so it is minimised (at zero) when $\mathbb{E}\left[Y|X\right] - m\left(X\right) = 0$ or

$$m\left(X\right) = \mathbb{E}\left[Y|X\right]$$

...and we've just shown that

$$\mathbb{E}\left[Y|X\right] = \underset{m(X)}{\arg\min} \mathbb{E}\left[\left(Y - m\left(X\right)\right)^2\right]$$

CEF - summary

Things to be able to demonstrate

The CEF decomposition and variance decomposition.
The CEF residual is mean zero.
The CEF residual is mean independent of X.
The CEF residual is uncorrelated with any function of X.
The CEF is a best predictor of Y.

The Linear Regression Model

The Linear Regression Model

The LRM parameters
The LRM as best linear predictor of the CEF
The LRM as best linear approximation to the CEF
Errors in the CEF and LRM

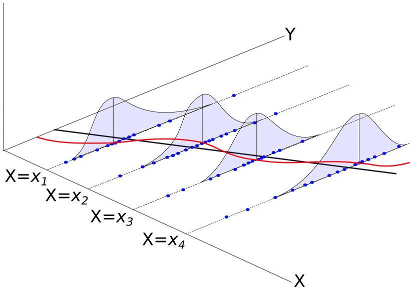
Linear Regression and the CEF

- The CEF is a function.
- The linear regression model (LRM) is .... err ... a model of the CEF which is .... err.....linear:

$$\mathbb{E}[Y|X] = b_0 + b_1 X$$

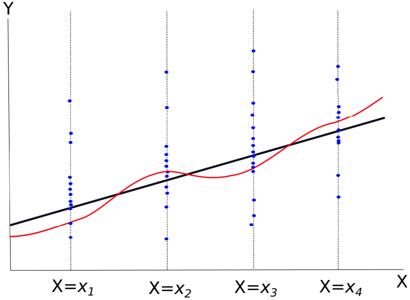
  where $\{b_0, b_1\}$ are unknown "parameters" in the jargon.
- Like *all* models, the linear regression model is a simplification/idealisation of the object of study.

Linear Regression and the CEF



- The red curve is the CEF.
- The straight line in the our linear model of it: the Linear Regression Model.

Linear Regression and the CEF



- Looking at it from "above" we see the not-necessarily-linear CEF and the Linear Regression Model of it.
- Remember that the LRM is a "model", not the real thing.

The Linear Regression Model
The coefficients

$$Y = b_0 + b_1 X + u$$

- We are after the values of $\{b_0, b_1\}$ which minimise the expected loss (expected sum of squares).
- These optimal values we will denote by $\{\beta_0, \beta_1\}$

$$\{\beta_0, \beta_1\} = \underset{b_0, b_1}{\arg\min} \mathbb{E}\left[\left(Y - [b_0 + b_1 X]\right)^2\right]$$

- Finding them is just a bog-standard optimisation problem:

$$\min_{b_0, b_1} \mathbb{E}\left[\left(Y - [b_0 + b_1 X]\right)^2\right]$$

The LRM/OLS coefficients

Differentiate with respect to b_0 :

$$\frac{\partial \mathbb{E}\left[\left(Y-\left[b_0+b_1X\right]\right)^2\right]}{\partial b_0} = \mathbb{E}\left[2\left(Y-\left[b_0+b_1X\right]\right)(-1)\right]$$
$$= -2\mathbb{E}\left[Y-b_0-b_1X\right]$$

Then with respect to b_1 :

$$\frac{\partial \mathbb{E}\left[\left(Y-\left[b_0+b_1X\right]\right)^2\right]}{\partial b_1} = \mathbb{E}\left[2\left(Y-\left[b_0+b_1X\right]\right)(-X)\right]$$
$$= -2\mathbb{E}\left[\left(Y-b_0-b_1X\right)X\right]$$

The LRM/OLS coefficients

Then set them equal to zero for a min

$$\mathbb{E}\left[Y - b_0 - b_1 X\right] = 0$$

This condition says $\mathbb{E}\left[u\right] = 0$: the optimal value of the $\{b_0, b_1\}$ must be such that the residual is on average zero.

$$\mathbb{E}\left[\left(Y - b_0 - b_1 X\right) X\right] = 0$$

This last condition says $\mathbb{E}\left[uX\right] = 0$: the optimal value of the $\{b_0, b_1\}$ must be such that the residual is uncorrelated with X.

The LRM/OLS coefficients

Since the foc's are linear we can rewrite them as:

$$\mathbb{E}\left[Y\right] - b_0 - b_1\mathbb{E}\left[X\right] = 0$$

$$\mathbb{E}\left[YX\right] - b_0\mathbb{E}\left[X\right] - b_1\mathbb{E}\left[XX\right] = 0$$

Thus we have two simultaneous linear equations with two unknowns to solve for. We will denote the optimal values of $\{b_0, b_1\}$ which solve this by $\{\beta_0, \beta_1\}$.

The Linear Regression Model
The coefficients

The LRM/OLS coefficients

Substituting the first into the second gives:

$$\begin{aligned}
\beta_1 \mathbb{E}[XX] &= \mathbb{E}[YX] - (\mathbb{E}[Y] - \beta_1 \mathbb{E}[X])\,\mathbb{E}[X] \\
&= \mathbb{E}[YX] - \mathbb{E}[Y]\,\mathbb{E}[X] + \beta_1 \mathbb{E}[X]^2 \\
\beta_1 \mathbb{E}[XX] - \beta_1 \mathbb{E}[X]^2 &= \mathbb{E}[YX] - \mathbb{E}[Y]\,\mathbb{E}[X]
\end{aligned}$$

So the β_1 coefficient is

$$\beta_1 = \frac{\mathbb{E}[YX] - \mathbb{E}[Y]\,\mathbb{E}[X]}{\mathbb{E}[XX] - \mathbb{E}[X]^2}$$

The LRM/OLS coefficients

Or (using the definition of the population variance and covariance)

$$\begin{aligned} Cov\left(Y,X\right) &= \mathbb{E}\left[YX\right]-\mathbb{E}\left[Y\right]\mathbb{E}\left[X\right] \\ Var\left(X\right) &= \mathbb{E}\left[XX\right]-\mathbb{E}\left[X\right]^2 \end{aligned}$$

gives

$$\beta_1=\frac{Cov\left(Y,X\right)}{Var\left(X\right)}$$

The LRM/OLS coefficients

With β_1 in hand we can then get β_0 :

$$\beta_0 = \mathbb{E}\left[Y\right] - \beta_1 \mathbb{E}\left[X\right]$$

The Linear Regression Model

The population linear least squares regression is defined therefore as:

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

where

$$\beta_0 = \mathbb{E}\left[Y\right] - \beta_1 \mathbb{E}\left[X\right]$$

and

$$\beta_1 = \frac{Cov\left(Y, X\right)}{Var\left(X\right)}$$

The residuals are mean zero and uncorrelated with X by construction.

The Linear Regression Model
An approximation to the CEF

- If the population CEF is itself linear then the linear regression function recovers it *exactly*.
- But even when this is not the case you can still use linear regression as an approximation.
- In fact the LRM is the best linear approximation to the CEF.
- This justification for the LRM goes back (at least) to Goldberger and has also been picked up by more recent authors such as Angrist and Pischke.

The Linear Regression Model

An approximation to the CEF

- The linear regression function provides *the best linear approximation* to the CEF (even though the CEF may not be linear).
- That is, the linear regression function solves

$$\{\beta_0, \beta_1\} = \underset{b_0, b_1}{\arg\min} \mathbb{E}\left[\left(\mathbb{E}\left[Y|X\right] - \left[b_0 + b_1 X\right]\right)^2\right]$$

- That is the LRM coeffcents $\{\beta_0, \beta_1\}$ minimise the distance between the (maybe wiggly) CEF and a linear function of X.

- The best way to see this is to show that the values of $\{b_0, b_1\}$ which solve the CEF approximation problem ...

$$\min_{b_0, b_1} \mathbb{E}\left[\left(\mathbb{E}\left[Y|X\right] - \left[b_0 + b_1 X\right]\right)^2\right]$$

... are the same as those which solve the linear least squares problem:

$$\min_{b_0, b_1} \mathbb{E}\left[\left(Y - \left[b_0 + b_1 X\right]\right)^2\right]$$

The LRM as an approximation to the CEF

Consider the linear regression problem

$$\min_{b_0, b_1} \mathbb{E}\left[(Y - [b_0 + b_1 X])^2\right]$$

Now focus on the minimand:

$$(Y - [b_0 + b_1 X])^2$$

now add and subtract $\mathbb{E}[Y|X]$

$$(Y - [b_0 + b_1 X])^2 = [(Y - \mathbb{E}[Y|X]) + (\mathbb{E}[Y|X] - [b_0 + b_1 X])]^2$$

The Linear Regression Model

An approximation to the CEF

The LRM as an approximation to the CEF

Then expand this quadratic:

$$
\begin{aligned}
(Y - [b_0 + b_1 X])^2 \quad = \quad & (Y - \mathbb{E}[Y|X])^2 \\
& + 2(Y - \mathbb{E}[Y|X])(\mathbb{E}[Y|X] - [b_0 + b_1 X]) \\
& + (\mathbb{E}[Y|X] - [b_0 + b_1 X])^2
\end{aligned}
$$

As before the first term is irrelevant and the second term will have an expected value of zero (because it's the residual times a function of X and we've already seen that object has zero expectation). This just leaves us with the third term.

The LRM as an approximation to the CEF

So minimising

$$\left(Y - [b_0 + b_1 X]\right)^2$$

is equivalent to minimising

$$\left(\mathbb{E}\left[Y|X\right] - [b_0 + b_1 X]\right)^2$$

Best Linear Approximation

An approximation to the CEF

The LRM as an approximation to the CEF

Since minimising $(Y - [b_0 + b_1 X])^2$ is equivalent to minimising $(\mathbb{E}[Y|X] - [b_0 + b_1 X])^2$, the linear least squares problem (which we have already solved)

$$\min_{b_0, b_1} \mathbb{E}\left[(Y - [b_0 + b_1 X])^2\right]$$

is the same as the CEF approximation problem:

$$\min_{b_0, b_1} \mathbb{E}\left[(\mathbb{E}[Y|X] - [b_0 + b_1 X])^2\right]$$

Since the problems are the same so are the solutions (i.e. $\{\beta_0, \beta_1\}$)

Linear Regression and the CEF

- The meaning and properties of the residual in the LRM are important matters.
- The residual in the LRM represents the variability in the dependent variable which was not captured by your linear combination of the explanatory variables.
- The residuals are errors between the observation and your ability to predict it stemming from
  - measurement errors
  - specification errors (a non-linear CEF and/or a CEF which depends on variables other than X).
- The properties of the LRM residual are often diagnostic of problems with your LRM.

Linear Regression and the CEF

- Note that, unless the CEF really is linear, the CEF and the linear regression are not the same.

$$\begin{aligned} Y &= \mathbb{E}\left[Y|X\right] + e \\ Y &= b_0 + b_1 X + u \end{aligned}$$

- In particular this means that the LRM residual u has properties which are somewhat alike those of the the CEF residual e, but they are not the same.

The CEF and the LRM

The CEF

$$Y = \mathbb{E}[Y|X] + e$$

$$\begin{aligned} \mathbb{E}[e] &= 0 \\ \mathbb{E}[eh(X)] &= 0 \\ \mathbb{E}[e|X] &= 0 \end{aligned}$$

The LRM

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

$$\begin{aligned} \mathbb{E}[u] &= 0 \\ \mathbb{E}[uX] &= 0 \\ \mathbb{E}[u|X] &= ? \end{aligned}$$

The CEF residual e is mean independent of X, but the LRM residual u may not be. This is a very important distinction.

CEF & LRM

Things to be able to demonstrate

Definition and derivation of the LRM
The LRM residual is mean zero
The LRM residual is uncorrelated with X.
The LRM is the best linear predictor of Y.
The LRM is the best linear approximation to the CEF.

The Multiple Regression Model

The Multiple Regression Model

The Multivariate LRM parameters
The Frisch-Waugh-Lovell Theorem
Co-linear regressors

Multiple Regression

- The notion of a CEF expands in an obvious way from the bivariate case

$$\mathbb{E}\left[Y|X\right]$$

to the multivariate case

$$\mathbb{E}\left[Y|X_1, X_2, ..., X_K\right]$$

- Similarly the linear model expands in an obvious way from

$$\mathbb{E}\left[Y|X\right] = \beta_0 + \beta_1 X$$

to accommodate other regressors:

$$\mathbb{E}\left[Y|X_1, X_2, ..., X_K\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_K X_K$$

Multiple Regression

- You can derive the OLS formulae for $\{\beta_0, \beta_1, ..., \beta_K\}$ by solving the least squares optimisation problem, but it's somewhat ugly (and it would take all the fun out of the Econometrics option next year).
- But the end result is really just a simple extension of the bivariate case.
- For the β_0 term in the bivariate case we had

$$\beta_0 = \mathbb{E}[Y] - \beta_1 \mathbb{E}[X]$$

- Now we have

$$\beta_0 = \mathbb{E}[Y] - \beta_1 \mathbb{E}[X_1] - \beta_2 \mathbb{E}[X_2] - ... - \beta_K \mathbb{E}[X_K]$$

Multiple Regression

- For the β_1 term in the bivariate case we had

$$\beta_1 = \frac{Cov\,(Y, X)}{Var\,(X)}$$

- Now we have for the multivariate case

$$\beta_k = \frac{Cov\left(Y, \widetilde{X}_k\right)}{Var\left(\widetilde{X}_k\right)}$$

- $\widetilde{X}_k$ is the residual from a regression of X_k on all of the other regressors (i.e. X_1 to X_K excluding X_k).

$$X_k = \alpha_0 + \sum_{i=1, i \neq k}^{i=K} \alpha_i X_i + \tilde{X}_k$$

Multiple Regression

- The result is known as the *Frisch-Waugh-Lovell Theorem* and is fiddly to prove but the intuition is reasonably clear:
- $\widetilde{X}_k$ is the variation in X_k which cannot be explained as a linear combination of the other regressors: it is particular to X_k.
- The bivariate regression of Y on $\widetilde{X}_k$ therefore focuses on the problem of predicting Y using *just* the independent variation in X_k with the effects of the other regressors removed.

Multiple Regression

- If X_k is *perfectly* explained by a linear combination of the other regressors[1] (i.e. $\widetilde{X}_k = 0$) it has no independent contribution to the prediction problem and $\beta_k = 0$.
- Extending this idea to the rest of the regressors and the coefficients all reflect the individual predictive contribution of *each individual regressor alone*.

[1] Known as "(multi)collinearity".

Population Regression - summary

Things you should now know

What the CEF is
The properties of the CEF and its residual
The definition and derivation of the LRM
How the LRM relates to the CEF
The properties of the LRM and its residual
FWL and the multivariate LRM