2002.4 Regression - Introduction

the good stuff

Nonparametric Regression

A regression is a means of approximating the 'true' conditional expectation function, or the 'population regression function'.

Bivariate case

For a discrete X, can plot an average for every value of X. So years of ed. vs income - simple - avg income for each year.

For continuous X something different is needed. One approach is kernel regression - for each x, compute some weighted average of $Y|X$ in an interval of width h around x (the bandwidth).

The weights in this average come from a kernel function $K_h(x)$ such that $\int_{-\infty}^{\infty} K_h(x) = 1$ and $K_h(x) = K_h(-x)$.

Kernels

Uniform: Easiest kernel function.

$$K(x) = \begin{cases} \frac{1}{2} & \text{where } |x| \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

For my understanding I'm going to make up a "weighting function" W_h such that

$$W_h(x, x_0) = K\left(\frac{x - x_0}{h}\right)$$

Then a uniform kernel regression is:

$$\begin{aligned} \hat{E}]Y|X = x_0] &= \frac{\int_{-\infty}^{\infty} YK(\frac{x-x_0}{h})dx}{\int_{-\infty}^{\infty} K(\frac{x-x_0}{h})dx} \\ &= \frac{\int_{-\infty}^{\infty} YW_h(x, x_0)dx}{\int_{-\infty}^{\infty} W_h(x, x_0)dx} \end{aligned}$$

There are other kernel functions. The underlying logic is the same. Eg. Epanechikov kernel (a quadratic):

$$K(x) = \begin{cases} \frac{3}{4}(1-x)^2 & \text{where } |x| \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Choosing bandwidth - a bias-variance tradeoff. A smaller bandwidth will fit more closely but produce a high-variance estimator. As example: think of $h > \text{Range}(X)$; then the kernel regression will be constant, with 0 variance, but probably some bias...

Parametric (Linear) Regression

This is the probably more familiar case. Assume that the CEF takes the form $E(Y|X) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + u$. We create an estimator, $\hat{E}(Y|X) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots \hat{\beta}_k X_k + \hat{u}$ when we have k covariates. In matrix notation can write $X\hat{\beta} + \hat{u}$ where X is $n \times k$ and the first column of X is 1 by convention.

Why would we want to do this? We might have theoretical reason to believe linearity. Or it might be easier to work with or to interpret. Or maybe we just want to keep with longstanding tradition.

A common way to fit this regression line is to choose $\hat{\beta}$ to minimize the mean squared difference between Y and $X\hat{\beta}$.

Least Squares Formula, Bivariate Case

Ugh

We want to find $\hat{\beta}$ to minimize $S(\hat{\beta}) = \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$.

Expanding:

$$\begin{aligned} S(\hat{\beta}) &= \sum (\hat{\beta}_1 x_i + \hat{\beta}_0 - y_i)^2 \\ &= \sum \left[\hat{\beta}_1^2 x_i^2 + 2\hat{\beta}_0 \hat{\beta}_1 x_i - 2\hat{\beta}_1 x_i y_i + \ldots \right] \\ &= \hat{\beta}_1^2 \sum x_i^2 + 2\hat{\beta}_0 \hat{\beta}_1 \sum x_i - 2\hat{\beta}_1 \sum x_i y_i \ldots \\ &+ n\hat{\beta}_0^2 - \hat{\beta}_0 \sum y_i + \sum y_i^2 \end{aligned}$$

Now to get first-order conditions for $\hat{\beta}_1$:

$$\frac{\partial S}{\partial \hat{\beta}_1} = 2\hat{\beta}_1 \sum x_i^2 + 2\hat{\beta}_0 \sum x_i - 2\sum x_i y_i$$

$$0 = \frac{\partial S}{\partial \hat{\beta}_1} = \hat{\beta}_1 \sum x_i^2 + \hat{\beta}_0 \sum x_i - \sum x_i y_i$$

$$\hat{\beta}_1 = \frac{\sum x_i y_i - \hat{\beta}_0 \sum x_i}{\sum x_i^2}$$

and $\hat{\beta}_0$:

$$0 = \frac{\partial S}{\partial \hat{\beta}0} = 2\hat{\beta}_1 \sum x_i + n\hat{\beta}_0 - \sum y_i$$

$$\begin{aligned}\hat{\beta}_0 &= \frac{\sum y_i}{n} - \hat{\beta}_1 \frac{\sum x_i}{n} \\ &= \bar{y} - \hat{\beta}_1 \bar{x}\end{aligned}$$

Substitute to get:

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum x_i y_i}{\sum x_i^2} - \frac{\left[\bar{y} - \hat{\beta}_1 \bar{x}\right] \sum x_i}{\sum x_i^2} \\ &= \frac{\sum x_i y_i}{\sum x_i^2} + \frac{-n\bar{x}\bar{y} + \hat{\beta}_1 n \bar{x}^2}{\sum x_i^2} \\ &= \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2} + \frac{\hat{\beta}_1 n \bar{x}^2}{\sum x_i^2} \\ \hat{\beta}_1 \left[\sum x_i^2 - \sum \bar{x}^2\right] &= \sum x_i y_i - \sum \bar{x}\bar{y} \\ \hat{\beta}_1 &= \frac{\sum \left[x_i y_i - \bar{x}\bar{y}\right]}{\sum \left[x_i^2 - \bar{x}^2\right]} \\ &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}\end{aligned}$$

This is $\mathrm{Cov}(X, Y)/\mathrm{Var}(X)$ in the sample. It is also is the same idea as $(X'X)^{-1}X'Y$.

The residuals are 0 by construction.

Note that if $\hat{\beta}_1 = 0$ then $\hat{\beta}_0$ is the sample mean.

It's worth noting that u, called the error, represents the difference between $X\beta$ and Y, meaning the 'true' amount of variance in Y not explained by the covariates in X.

In contrast, $\hat{u}$ is the residual, which represents the degree to which $X\hat{\beta}$ explains Y. $E[\hat{u}] = 0$ by construction (were it not, you could shift the regression line up/down to reduce it further). We will come back to this.

Analysis of Variance for Lin. Reg

The variance of Y represents the deviations of y_i from $\bar{y}$. Let $\hat{y}_i$ be the predicted value at i, and $\hat{u}_i$ be the residual. Then, recalling that $\bar{y} = \bar{\hat{y}}$ and $\bar{\hat{u}} = 0$, we can break down the sample variance as follows:

$$\begin{aligned} y_i &= \hat{y}_i + \hat{u}_i \\ y_i - \bar{y} &= \hat{y}_i - \bar{y} + \hat{u}_i \\ (y_i - \bar{y})^2 &= (\hat{y}_i - \bar{y})^2 + (\hat{u}_i - \bar{\hat{u}})^2 \\ \mathrm{Var}(y) &= \mathrm{Var}(\hat{y}) + \mathrm{Var}(\hat{u}) \end{aligned}$$

Dividing both sides by $\mathrm{Var}(y)$ we get

$$1 = \mathrm{Var}(\hat{y})/\mathrm{Var}(y) + \mathrm{Var}(\hat{u})/\mathrm{Var}(y)$$

$$\frac{\mathrm{Var}(\hat{y})}{\mathrm{Var}(y)} = 1 - \frac{\mathrm{Var}(\hat{u})}{\mathrm{Var}(y_)} \text{ or } 1 - \frac{SSR}{SST}$$

This is the good old R^2, interpretable as "Share of observed variance in Y explained by the *linear* model."

Properties of Linear Regression under certain Assumptions

Under what circumstances can we use linear regression to estimate the population regression function?

The Gauss-Markov theorem (and some relatives) tells us what assumptions we need to make, and how far that gets us (pretty far, if you ask 60 years of social scientists)

There are several varying statements of these assumptions. Here's the one from class

Assumption I - Linearity in parameters

The 'true' relationship is linear in parameters with mean-0 errors. This isn't as restrictive as it seems. $\beta_0 + \beta_1 x^2$ is still perfectly fine. But something like $1/(1 + \exp(-X\beta))$ is trickier.

Assumption II - Random Sampling
Both Y and X are i.i.d. and follow the population model (ie are representative). This can be obtained by random sampling.

Assumption III - No (Perfect) Collinearity
$x_i \neq x_j$ for some i, j. The more collinearity, the worse the estimate I think.

Without this assumption, i.e. if $\mathrm{Var}(X)$ is 0, $\hat{\beta}$ is undefined (see denom. of the β_1 formula above.)

Assumption IV - Zero conditional mean
aka no confounders. $E[u|X] = 0$. If there's something influencing Y that's related to X, then no go. Implies the weaker condition $\mathrm{Cov}(X, u) = 0$.

Example from class: relationship between cigarette/pipe smoking choice and mortality. Found cigarette smokers have notably lower mortality. But that's because pipe smokers are older...

Now we can talk about the sampling distribution of our estimator, and demonstrate unbiasedness.

Assumption V - Homoskedasticity

$$\mathrm{Var}(u|X) = \sigma_u^2$$

In other words, the stochastic component is the same across observations - or the variance-covariance matrix is kI for some (non-negative) constant k.

Example of heteroskedasticity: imagine income vs. education. More education creates more dispersion in incomes (grad student vs mckinsey consultant...)

These assumptions **I-V** are the Gauss-Markov assumptions.

With these we can also talk about the variance of the sampling distribution.

Can show that

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma_u^2}{\sum (x_i - \bar{x})^2} = \frac{\sigma_u^2}{(n-1)S_x^2}$$

which means $\mathrm{Var}(\hat{\beta}_i)$ is low if $\mathrm{Var}(u)$ is low, and falls with n and $\mathrm{Var}(X)$.

We can estimate the stochastic component u with residuals $\hat{u}$, with a bias correction. We're assuming that the uncertainty in our estimate is representative of the fundamental uncertainty in the DGP.

The bias correction is to account for the fact that we've minimized $\hat{u}$ by construction. So the unbiased estimator for u is

$$\hat{\sigma}_u^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2$$

Then the variance of $\hat{\beta}_1$ is

$$\text{Var}(\hat{\beta}_1) = \frac{\hat{\sigma}_u^2}{\sum(x_i - \bar{x})^2} = \frac{SSR}{SSX}$$

Gauss-Markov Theorem

This theorem tells us that, given assumptions **I = IV**, an OLS estimate is BLUE: the *best unbiased linear estimator*. In other words, in the class of unbiased linear estimators, OLS has the lowest variance.

Variance

For some quantities of interest, e.g. confidence intervals, we might want to fully characterize the sampling distribution of the estimator. That's why we have:

Assumption VI - Normality

Assume the true errors are normally distributed and independent of X:

$$u \sim N(0, \sigma_u^2)$$

Equivalently,

$$Y|X \sim \mathcal{N}(X\beta, \sigma_u^2)$$

This is the systematic component/stochastic component disaggregation from Gov 2001.

With this assumption, we have the 'Classical Linear Model'. Given this, we can show that this is the BUE - best unbiased estimator (of any shape).

Asymptotics

We only need assumptions **I-IV** to show consistency of OLS estimator. In fact we can change **IV** to the weaker assumption $\mathrm{Cov}(X, u) = 0$.

GIven **I-V** we can show asymptotic normality.

Agnostic OLS

Honestly I don't really know what the point of this is. But if you want you can think of the OLS as estimating the *best linear approximation* to the true population regression function, rather than estimating the *linear* population regression function. But you still have to make the case that I should care about the linear approximation to the pop. reg function.

Quiz 2 ends here
