Bayesian Linear Regression Applied to Income Data

Senan Hogan-H.

MATH 153: Bayesian Statistics

May 2018

1 Wage Data and Frequentist Regression

The distribution of income, here measured by annual income in 2015 CPI-R dollars, is well approximated for 2010-2016 by a $N(10.61, 0.77)$ distribution. The above shows the distribution of income (red) compared to the mentioned Normal distribution. Here, data is taken from the March CPS, a representative survey of annual data for individuals in the US [1].

The distribution of income, and thus inequality of income, is regularly explained in labour economics by a frequentist linear regression of the following form:

$$log(Y_i) = log(Y_0) + \rho s_i + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i \tag{1}$$

Y_i represents a measure of income for an individual , s_i years of education, and Y_0 the standard intercept. ρ, β_1, β_2 are coefficients to be estimated with residual ε_i. Potential experience, x_{it}, is defined as age minus years of education minus 6, i.e. $x_{it} = Age_{it} - s_{it} - 6$. The equation may also include a dummy variable for race, gender, and possibly other variables to control for these differences.

An extra year of educations is associated with a rise of around 10% in income, and the measure of potential experience with around 6% but deteriorating for higher levels of experience (shown by the negative estimate on the quadratic term). The predictions are compared below to the $N(10.61, 0.77)$ distribution, showing how this approach fails to replicate the distribution, instead predicting

	Dependent variable:
	log(rincp_ern)
Years education	0.126^{***}
	(0.011)
Potential experience	0.064^{***}
	(0.009)
(Potential experience)^2	-0.001^{***}
	(0.0002)
Constant	8.147^{***}
	(0.183)
Observations	500
R^2	0.283
Adjusted R^2	0.278
Note:	*p<0.1; **p<0.05; ***p<0.01

a more equal distribution than the one observed.

This regression approach has been names the Mincer wage equation or earnings function, which dates back to some of the first studies that focus on wage inequality [2, 3]. This model is extremely influential in labour economics to describe and predict inequality in wages in the US population. Its influence comes in part from its theoretical foundations and simplicity in interpretation, yet is documented as being only accurate in predicting wages[1] for the 1950s, and less so after. The approach is classic in the frequentist, econometrics paradigm and is still (often egregiously) used in economic research today for predictive purposes.

2 Generalisation of the Mincer Wage Equation

Suppose that the income distribution, Y_i, follows a standard Mincer wage equation, as follows (and as in equation 1):

$$log(Y_i) = log(Y_0) + \rho s_i + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i \tag{2}$$

Call this distribution the Log-Normal (LN) distribution. It has the following probabilistic density function (pdf) and cumulative distribution function (cdf), where μ is the mean and σ

[1] Where the equation may estimated independently for different years.

standard deviation:

$$f_{LN}(y|\mu,\sigma) = \frac{1}{\sigma y\sqrt{2}\pi} e^{-\frac{(log(y)-\mu)^2}{2\sigma^2}} \tag{3}$$

$$F_{LN}(y|\mu,\sigma) = \frac{1}{\sqrt{2}\pi} \int_{-\infty}^{\frac{log(y)-\mu}{\sigma}} e^{\frac{-x^2}{2}} dx \tag{4}$$

If income follows this specific distribution the error term follows a normal distribution with expectation zero and variance σ^2. However, this model is only useful when economists are considering the effect of conditional means (and their change) on the distribution of income. For example, the framework is useful for considering an effect on mean income of a uniform increasing in education, but not useful for raising income for those at the bottom of the income distribution only [4].

3 Bayesian Approach

Explain differences, building framework for Bayesian inference.

Uncertainty in model selection, build model by model averaging. Some equations to explain what that is.

4 Regression Across Quantiles

Some equations to minimise functions and explain what quantile regression is.

Apply to wage data, showing different returns to education, wage gap etc.

********** * Start estimating quantile 1 of 2 in total * ********** Current iteration : [1] 500 Current iteration : [1] 1000 ********** * Start estimating quantile 2 of 2 in total * ********** Current iteration : [1] 500 Current iteration : [1] 1000

5 Appendix: R code

References

[1] Center for economic and policy research. march cps uniform extracts, version 1.0. washington, dc., 2016.

[2] Jacob Mincer. Investment in human capital and personal income distribution. *Journal of political economy*, 66(4):281–302, 1958.

[3] Jacob A Mincer. Schooling and earnings. In *Schooling, experience, and earnings*, pages 41–63. NBER, 1974.

[4] Masato Okamoto et al. Mincer earnings regression in the form of the double pareto-lognormal model. Technical report, 2016.