# Bayesian Linear Regression Applied to Income Data

Senan Hogan-H.

MATH 153: Bayesian Statistics

May 2018

## 1 Wage Data and Frequentist Regression

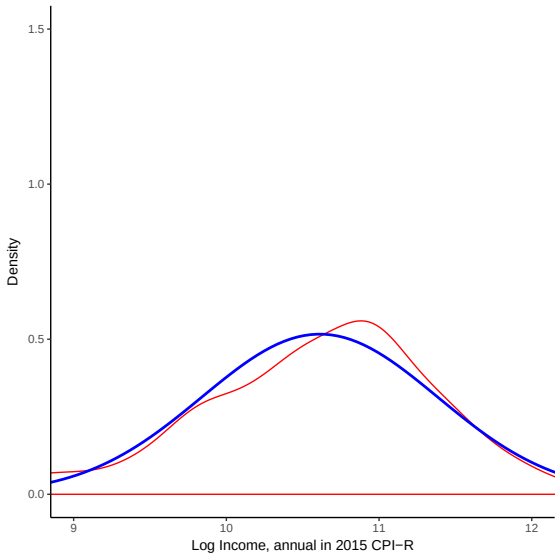
The distribution of income, here measured by annual income in 2015 CPI-R dollars, is well approximated for 2010-2016 by a $N(10.61,\ 0.77)$ distribution. The above shows the distribution of income (red) compared to the mentioned Normal distribution. Here, data is taken from the March CPS, a representative survey of annual data for individuals in the US [1]. The set is a 500 subset, to demonstrate Bayesian capabilities without washing out priors by the (available) thousands of observations.

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

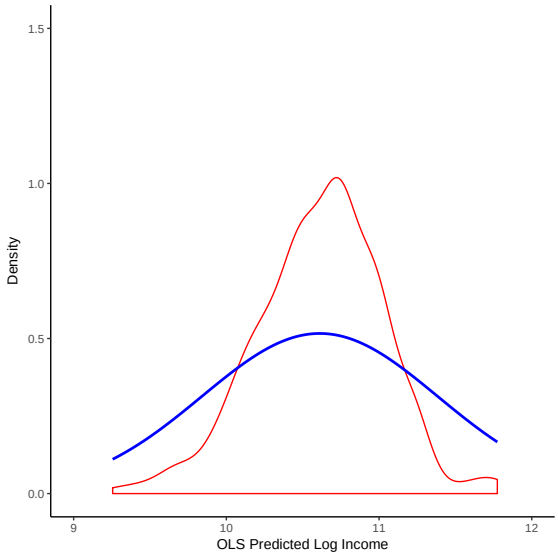
$$log(Y_i) = log(Y_0) + \rho s_i + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i \tag{1}$$

$Y_i$ represents a measure of income for an individual , $s_i$ years of education, and $Y_0$ the standard intercept. $\rho$, $\beta_1$, $\beta_2$ are coefficients to be estimated with residual $\varepsilon_i$. 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 $N(10.61,$




| | *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 |




0.77) distribution, showing how this approach fails to replicate the distribution, instead predicting 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 [5, 6]. 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 $\mu$ is the mean and $\sigma$

[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 $\sigma^2$. 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 [7].

## 3 Bayesian Approach

The Bayesian approach to estimating the distribution of wages has some notable differences. Let start with the priors.

Let $\rho(\beta, \sigma^2)$ be a prior on coefficients in the regression, so that $\beta = (\beta_1, \beta_2, \ldots, \beta_k)$ for an undetermined number, $k$, of regressors in matrix $X$. An uninformative prior in this case may be a uniform, or we could use a normal centred around some convenient means. In the standard Bayesian approach, we are looking for the posterior distribution, expressed as follows.

$$\rho(\beta, \sigma^2|Y, X) \propto \rho(Y|\beta, \sigma^2, X)\rho(\beta, \sigma^2)$$

Only in the case that the prior is conjugate to the posterior will the solution take an analytical form, which is not generally realistic so that more complicated MCMC sampling methods are required.

## 4 Model Selection

A frequentist setting does not factor any uncertainty in model selection, as in inclusion of variables. The labour economics literature regularly uses a Mincer wage equation with controls, barely even noting the exact functional form beyond a footnote. The Bayesian approach allows us to more robustly select model controls. The March CPS provides data for the standard regressors in a

Mincer equation, education and a measure of potential experience, as well as dummies for whether an individual is female, married, lives n rural area or city, and for whether they are black or Hispanic (so comparing to the white population).

Bayesian model selection involves considering the model evidence, for a given regression model $m$, $p(Y|m)$. This is defined as follows.

$$p(Y|m) = \int p(Y|\mathbf{X}, \beta, \sigma)\, p(\beta, \sigma)\, d\beta\, d\sigma$$

Below is some code for a Bayesian model selection of the Mincer Wage Equation, with a uniform prior on all coefficients (uninformative prior), showing the 5 most likely models with given controls. It is clear that the most likely model is that of the original Mincer equation with only controls included for the gender wage gap, and a wage penalty for the black population, with a posterior probability of 0.54.

| | P(B != 0 \| Y) | model 1 | model 2 | model 3 | model 4 | model 5 |
|---|---|---|---|---|---|---|
| Intercept | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| education | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| pot_exp | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| pot_exp_2 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| female | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| married | 0.14 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 |
| rural | 0.25 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 |
| centcity | 0.05 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Race_black | 0.92 | 1.00 | 1.00 | 1.00 | 0.00 | 1.00 |
| Race_hispanic | 0.05 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| BF | | 1.00 | 0.35 | 0.14 | 0.08 | 0.05 |
| PostProbs | | 0.54 | 0.19 | 0.08 | 0.04 | 0.03 |
| R2 | | 0.35 | 0.36 | 0.35 | 0.34 | 0.35 |
| dim | | 6.00 | 7.00 | 7.00 | 5.00 | 7.00 |
| logmarg | | -1334.99 | -1336.02 | -1336.92 | -1337.53 | -1337.89 |

It follows that a model of the following form will be estimated in this context, with variables as previously specified:

$$log(Y_i) = log(Y_0) + \beta_1 s_i + \beta_2 x_i + \beta_3 x_i^2 + \beta_4 female_i + \beta_5 Black_i + \varepsilon_i \tag{5}$$

# 5 Bayesian Regression Across Quantiles

Quantile regression in the frequentist paradigm involves adjusting the error function to produce an estimate for coefficient at given quantiles. Extension of this process to Bayesian regression allows for model uncertainty to be factored, as well as the use of priors on coefficients. This analysis is especially for wage data as we may expect returns to certain variables not be constant across the distribution of income. For example how do extra years of education affect income at the bottom 10% of the income distribution compared to the top, and what kind of distribution does Bayesian analysis estimate for either?

This section introduces the Bayesian Quantile regression, a relatively recent development in statistics, applied to the model developed previously. The equations follow that presented in [2], which builds the *BayesQR* package and is a great way to implement Bayesian quantile regression in R.

First, for a quantile of interest consider a three-parameter asymmetric Laplace distribution (ALD) density:

$$f(x|\mu,\sigma,\tau) = \frac{\tau(1-\tau)}{\sigma} e^{-\rho_\tau \frac{x-\mu}{\sigma}} \tag{6}$$

Here, $\rho_\tau(x) = x(\tau - I(x<0))$ and $I(.)$ is the indicator function. Bayesian implementation of quantile regression begins by forming a likelihood comprised of independent asymmetric Laplace densities for $\mu = x_i^T\beta$ for each $i = 1,\ldots,n$ (where $x_i$ is here defined as the vector of regressors, $\beta$ vector of coefficients). Before moving forward, a quantile of interest $\tau$ and priors for model parameters $\beta$ and $\sigma$, should be specified. The posterior distribution is then as follows, for a specified prior $\pi(.)$ and $ALD(.|\mu,\sigma,\tau)$ density of the ALD.

$$\psi(\beta,\sigma|x,y,\tau) \propto \pi(\beta,\sigma)\prod_{i=1}^{n} ALD(y_i|x_i^T\beta,\sigma,\tau) \tag{7}$$

## 5.1 MCMC Sampling for Regression on a Continuous Variable

This framework for implementing a quantile regression can take a few forms for the next steps, we'll focus on the format of a continuous response variable without adaptive LASSO for variable selection (as we've already looked at some skin-deep model selection).

A straightforward prior for $\beta$ is the multivariate normal distribution, Normal(mean = $\beta_0$, varcov

$= V_0$). For the prior on $\sigma$, consider the inverse gamma distribution invGamma(shape $= n_0$, scale $= s_0$), with density as follows.

$$f(x|n_0, s_0) = \frac{s_0^{n_0}}{\Gamma(n_0)} x^{-n_0-1} e^{-\frac{s_0}{x}} \tag{8}$$

In this Bayesian approach to quantile regression, the error term is assumed to follow the asymmetric Laplace distribution. [4] shows that the ALD can be represented as a location scale mixture of normal distributions, where the mixing distribution follows an exponential distribution, so that the regression can be written as:

$$log(Y_i) = y_i = x_i^T \beta + \varepsilon_i = x_i^T \beta + \theta v_i + \omega u_i \sqrt{\sigma v_i} \tag{9}$$

Here $v_i = \sigma z_i$, $\theta = \frac{1-2\tau}{\tau(1-\tau)}$, $\omega^2 = \frac{2}{\tau(1-\tau)}$, $z_i$ is a standard exponential and $u_i$ is a standard normal variate. So that the likelihood function takes the form as follows.

$$f(y_i|x_i^T\beta, \sigma, \tau) \propto e^{-\sum_{i=1}^n \frac{(y_i - x_i^T\beta - \theta v_i)^2}{2\omega^2 \sigma v_i}} \prod_{i=1}^n \frac{1}{\sqrt{\sigma v_i}} \tag{10}$$

Now a Gibbs sampling algorithm involves drawing $\beta$, $\sigma$ and $z$ from their full conditional distributions. Further computation by [2] show that the full conditional density of $\beta$ is given by:

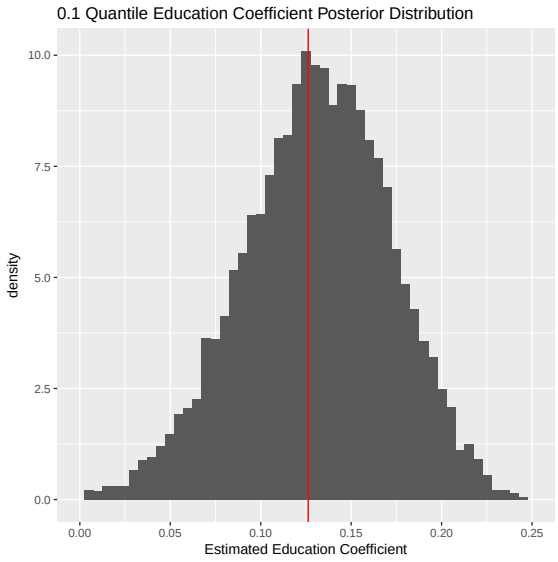
$$\psi(\beta|\sigma, v, y, x, \tau) \sim N(B, V) \tag{11}$$

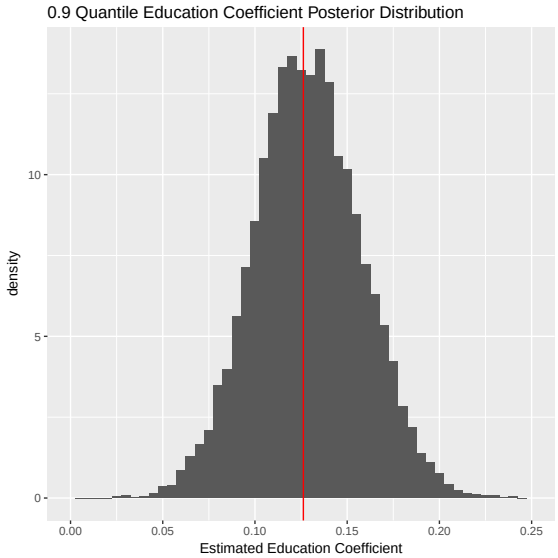
where

$$V^{-1} = \sum_{i=1}^n \frac{x_i^T x_i}{\omega^2 \sigma v_i} + V_0^{-1} \tag{12}$$

$$B = V(\sum_{i=1}^n \frac{x_i(y_i - \theta v_i)}{\omega^2 \sigma v_i}) \tag{13}$$

With more complicated, yet tractable, conditional posterior distributions for $\sigma$ and $v$ provided in [2].




## 5.2 Application to Returns to Education

Below I apply this Bayesian quantile regression to the March CPS wage data. Quantiles $\tau = 0.1, 0.9$ are considered, as the top and bottom 10% of the income distribution are most widely considered as comparison groups in the economic literature [3].

Below are the MC chain samples of the posterior distribution for estimated coefficient for years of education at the two quantiles. The frequentist estimate is layed over the desnsity plots, showing how returns to education are estimated to be more variable for 0.1 quantile than for the 0.9 quantile, though both are not far from the frequentist point estimate.

# 6 Appendix: R code

```
> set.seed(47)
> library(tidyverse)
> library(data.table)
> library(stargazer)
> library(BAS)
> library(bayesQR)
> library(xtable)
> CPS.data <- fread('CPS_data.csv', header = T, sep = ',', showProgress = FALSE)
> CPS.data$pot_exp <- CPS.data$age - CPS.data$education - 6
> CPS.data$pot_exp_2 <- CPS.data$pot_exp^2
> CPS.data <- CPS.data %>% subset(year>=2010 & year<=2016)
> CPS.data <- CPS.data %>% subset(select = c(
+   'rincp_ern', # real annual earnings (no unearned income), for person
+   'education', #education variable
+   'pot_exp',
+   'pot_exp_2',
+   'female', # whether female
+   'married', # whether married
+   'rural', # whether live in rural area
+   'suburb', # whether live in suburbs
+   'centcity', # whether they live in a central city
+   'selfemp', # self-employed
+   'Race_white', 'Race_black', 'Race_hispanic' # race vairables
+ ))
> n_sample <- 500
> CPS.data <- CPS.data %>% sample_n(n_sample)
> wage_mean <- mean(log(CPS.data$rincp_ern))
> wage_sd <- sd((log(CPS.data$rincp_ern)))
```

```
> CPS.data %>% ggplot() +
+   geom_density(aes(x = log(rincp_ern)), colour = 'red') +
+   stat_function(fun = function(x) dnorm(x,
+                                         mean = wage_mean,
+                                         sd = wage_sd),
+                 colour = 'blue', size = 1) +
+   labs(x= 'Log Income, annual in 2015 CPI-R', y='Density') +
+   coord_cartesian(xlim=c(9,12), ylim=c(0,1.5)) +
+   theme_classic()
> Mincer.reg <- CPS.data %>%
+   lm( log(rincp_ern) ~ education + pot_exp + pot_exp_2,
+       #female + Race_black + Race_hispanic,
+       data=.)
> stargazer(Mincer.reg,
+           title = 'Mincer Equation Results',
+           covariate.labels = c('Years education', 'Potential experience',
+                                '(Potential experience)$^2$'),
+           #dep.var.caption  = 'Income Measure',
+           #dep.var.labels   = c('Log hourly wage', 'Log annual income'),
+           omit.stat=c("LL","ser","f"),
+           header = FALSE, float = FALSE, no.space = TRUE)
> predicted.data %>% ggplot() +
+   geom_density(aes(x = OLS_predicted_Log_income), colour = 'red') +
+   stat_function(fun = function(x) dnorm(x,
+                                         mean = wage_mean,
+                                         sd = wage_sd),
+                 colour = 'blue', size = 1) +
+   labs(x= 'OLS Predicted Log Income', y='Density') +
+   coord_cartesian(xlim=c(9,12), ylim=c(0,1.5)) +
+   theme_classic()
```

```
> wage_equation <- log(rincp_ern) ~ education + pot_exp + pot_exp_2 +
+   female + married + rural + centcity + Race_black + Race_hispanic
> Bayes_av.reg <- bas.lm( wage_equation ,
+                         data = CPS.data,
+                         prior = "BIC",
+                         modelprior = uniform(),
+                         na.action = "na.omit")
> xtable(summary(Bayes_av.reg))
> # https://rpubs.com/mfondoum/bayesian_linear_regression
>
> # Next remove the other variables, for coding convenience
> CPS.data <- CPS.data %>% subset(select = c(
+   'rincp_ern', # real annual earnings (no unearned income), for person
+   'education', #education variable
+   'pot_exp',
+   'pot_exp_2',
+   'female', # whether female
+   'Race_black' # race variable
+ ))
> # Equation chosen by above model selection
> wage_equation <- log(rincp_ern) ~ education + pot_exp + pot_exp_2 +
+   female  + Race_black
> n <- 10000
> Bayes_q10.reg <- bayesQR(wage_equation ,
+                        data = CPS.data,
+                         quantile = 0.1, alasso = F,
+                         ndraw = n,
+                         seed = 47)
> Bayes_q10.data <- as.data.frame(Bayes_q10.reg[[1]]$betadraw)
> colnames(Bayes_q10.data) <- Bayes_q10.reg[[1]]$names
```

```
> Bayes_q90.reg <- bayesQR(wage_equation ,
+                        data = CPS.data,
+                         quantile = 0.9, alasso = F,
+                         ndraw = n,
+                         seed = 47)
> Bayes_q90.data <- as.data.frame(Bayes_q90.reg[[1]]$betadraw)
> colnames(Bayes_q90.data) <- Bayes_q90.reg[[1]]$names
> # summary(Bayes_q10.reg, burnin = n/5)
> # summary(Bayes_q90.reg, burnin = n/5)
> freq.coeff <- as.numeric(Mincer.reg$coefficients[2])
> Bayes_q10.data %>% ggplot() +
+   geom_histogram(aes(x=education, y = ..density..), binwidth = .005) +
+   labs(title = "0.1 Quantile Education Coefficient Posterior Distribution",
+        x = "Estimated Education Coefficient") +
+   xlim(0,0.25) + geom_vline(xintercept = freq.coeff, colour = 'red')
> Bayes_q90.data %>% ggplot() +
+   geom_histogram(aes(x=education, y = ..density..), binwidth = .005) +
+   labs(title = "0.9 Quantile Education Coefficient Posterior Distribution",
+        x = "Estimated Education Coefficient") +
+   xlim(0,0.25) + geom_vline(xintercept = freq.coeff, colour = 'red')
>
```

## References


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

[2] Dries F Benoit, Dirk Van den Poel, et al. bayesqr: A bayesian approach to quantile regression. *Journal of Statistical Software*, 76(i07), 2017.

[3] Chinhui Juhn, Kevin M Murphy, and Brooks Pierce. Wage inequality and the rise in returns to skill. *Journal of political Economy*, 101(3):410–442, 1993.

[4] Hideo Kozumi and Genya Kobayashi. Gibbs sampling methods for bayesian quantile regression. *Journal of statistical computation and simulation*, 81(11):1565–1578, 2011.

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

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

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