ECON3206/5206: Review of Linear regression model

Dr. Rachida Ouysse

School of Economics
UNSW

- Suppose we are interested in wage rates in the United states. Wages vary across workers and can be described using a probability distribution.
- Formally, we view the wage of an individual worker as a random variable $wage$ with **probability distribution**

$$F(w) = Pr(wage \leq w)$$

- A person wage is random: do not know the wage before it is measured. Observed wages are realizations from the distribution F
- We usually do not know F: we can learn about the distribution from many realizations of the *wage* variable.

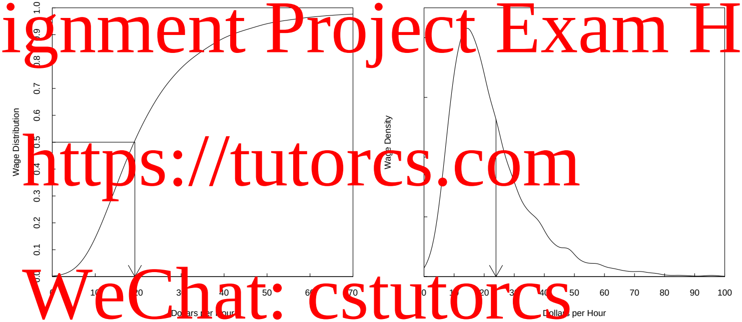


Figure 3.1: Wage Distribution and Density. All full-time U.S. workers

- Important measures of central tendency are the median and the mean. The median m of a continuous distribution F is the unique solution to

$$F(m) = \frac{1}{2}$$

The median U.S. wage in 2009 is $\$19.23$.

- A convenient measure (but not robust) of central tendency is the **mean** or **expectation**.
- The expectation of a random variable y with density f is

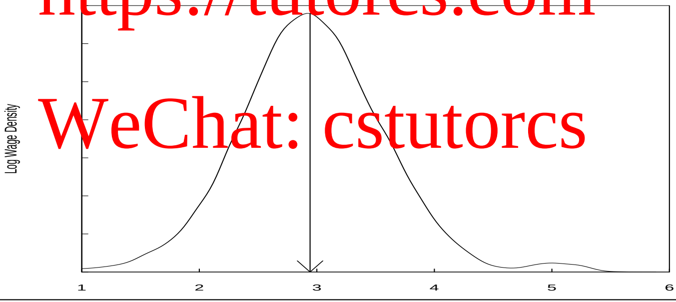
$$\mu = E(y) = \int_{-\infty}^{\infty} y f(y) dy$$

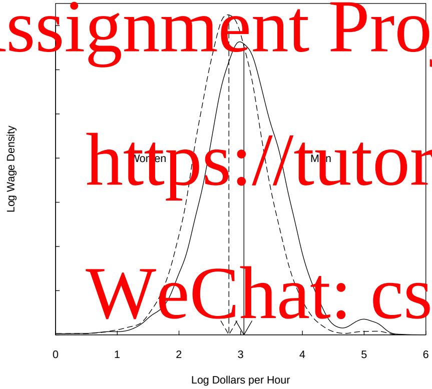
We use the single character y to denote the random variable, rather than the more cumbersome label $wage$

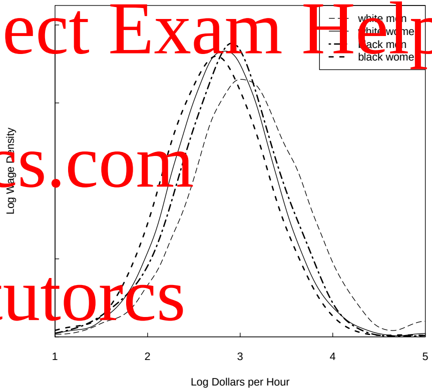
The mean wage in this example is $\$23.90$. The mean is not robust in the presence of substantial skewness or thick tails, which are both features of the wage distribution.

- In this context it is useful to transform the data by taking natural logarithm.

  The mean of the random variable $log(wage)$ also denoted $log(y)$ is \$2.95.
- The density of log wages is much less skewed and fat-tailed than the density of the level of wages, so its mean $E(log(y)) = 2.95$ is a much better measure of central tendency of the distribution.
- In fact, the geometric mean $exp(E(log(y))) = \$19.11$ is a robust measure of central tendency of y!!




Log Wage Density
0
1
2
3
4
5
6
Log Dollars per Hour


1
2
3
4
5
Log Dollars per Hour

- Is the wage distribution the same for all workers, or does the wage distribution vary across subpopulations?

- the plots above displays the densities of log wages in the U.S. men and women with their means (3.05 and 2.81).
- the means displayed are called the **conditional means** (or **conditional expectations**) of log wages given gender:

$$E(log(wage)|gender = man) = 3.05$$

$$E(log(wage)|gender = woman) = 2.81$$

- Here the conditioning variable $gender$ is a random variable from the viewpoint of econometric analysis.
- We can use more than one variable in the conditioning of the expectation:

$$E(log(wage)|gender = man, race = white) = 3.07$$

- In many cases it is convenient to simplify notation by writing variables using single characters, typically y, x, and/or z.

- Typically in econometrics it is conventional to denote the dependent variable by the letter y and the conditioning variables by the letter x, and multiple conditioning by the subscripted letters $x_1, x_2, \cdots, x_k$.
- Conditional expectation can be written with the generic notation

$$E(y|x_1, x_2, \cdots, x_k) = m(x_1, x_2, \cdots, x_k)$$

- This is called the **conditional expectation function**. For example, the conditional expectation of $y = log(wage)$ given $(x_1, x_2) = (gender, race)$ is given by

	men	woman
white	3.07	2.82
black	2.86	2.73
other	3.03	2.86

- An econometrician has observational data

$$\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$$

- If the data are cross-sectional, it is reasonable to assume they are mutually independent
- If the data are randomly gathered, it is reasonable to model each observation as a random draw from the same probability distribution. In this case the data are **independent and identically distributed**, or **iid**.

- To study how the distribution of y_i varies with x_i, we can focus on the conditional density of y_i given x_i and its conditional mean $m(\mathbf{x}_i)$.
- The conditional mean function is the regression function.

$$y_i = E[y_i|x_i] + (y_i - E[y_i|x_i]) = E[y_i|x_i] + \mu_i$$

- $E[\mu_i|x_i] = 0$.
- μ is called the conditional expectation function error.

- While the conditional mean $m(\mathbf{x})$ is the best predictor of y among all functions of $\mathbf{x}$, its functional form is typically unknown.
- For empirical implementation and estimation, it is typical to replace $m(\mathbf{x})$ with an approximation.
- Most commonly, this approximation is linear in $\mathbf{x}$.
- It is convenient to augment the regressor vector x by listing the number 1 as an element. We call this the **constant** or **intercept** term.

$$m(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k = \mathbf{x}'\beta$$

where

$$\mathbf{x} = (1, x_1, x_2, \cdots, x_K)'$$

$$\beta = (\beta_0, \beta_1, \beta_2, \cdots, \beta_K)'$$

- Boldface letter indicates a column vector. In the case of one regressor x and a constant term: $\beta = (\beta_0, \beta_1)$ and $\mathbf{x} = (1, x)'$, and $\mathbf{x}'\beta = \beta_0 + \beta_1 x$.

  (Wisdom: Models should have a constant term unless the theory says they should not.)

Assumption

MLR.1 Linearity : The population model is linear in the parameters:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u, \quad (1)$$

where β_i, $i = 0, \cdots, k$ are the unknown (constants) parameters of interest, x_i's are the regressors which can be assumed to be either fixed or random, and μ the random error.

If the linearity assumption is violated then the regression model is misspecified. This is known as *functional form misspecification* (although this is still linear in β's).

- The model does not account for some important nonlinearities;

- Omitting important variables is also model misspecification;

- Generally functional form misspecification causes biases in the remaining parameter estimators.

Example

Suppose that the correct specification of the wage equation is:

$$log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 (exper)^2 + \mu, \quad (2)$$

then the return for an extra year of experience is

$$\frac{\partial wage}{\partial exper} = wage \times [\beta_2 + 2\beta_3 exper].$$

If the estimated model is instead:

$$log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \mu, \quad (3)$$

then use of the biased (upward) OLS estimator of β_2 can be misleading.
If the estimated model is instead:

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 (exper)^2 + \mu, \quad (4)$$

$$\partial wage / \partial exper = \beta_2 + 2\beta_3 exper \quad (5)$$

Assumption

MLR2. Random Sampling:
We have a random sample of n observations, $\{(x_{i1}, x_{i2}, \cdots, x_{ik}, y_i) : i = 1, 2, \cdots, n\}$, following the population model in Assumption 1.
Nonrandom sampling causes OLS estimator to be biased and inconsistent.

Scenarios where Assumption 2 does not hold include:

- Missing Data
- Nonrandom Samples
- Outliers

Assumption

MLR3. No Perfect Collinearity:
In the sample and in the population, none of the independent variables is constant, and there are no exact linear relationships among the independent variables.

Scenarios where Assumption 3 is violated include:

- One independent variable is a **linear combination** of one or more other regressors. It is not a problem to include nonlinear functions of the same variables.
  - For example include consumption, investment and income on the right hand side of the regression equation. In national accounts, national income is the sum of consumption and investment
  - Including all seasonal dummies and the constant term in the regression

Assumption

MLR4. Zero Conditional Mean:
The error term μ has a conditional expected value of zero given any values of the independent variables,

$$E(u|x_1, \cdots, x_K) = 0$$

This assumption fails for many reasons, these include:

- Misspecification of the functional form
- Omitting important factors correlated with any of the regressors: omitted variables bias.
- Measurement error in the explanatory variables (more later, W. Ch. 15).
- Endogeneity and Simultaneity: some explanatory variables are determined jointly with the dependent variable

Theorem

Unbiasedness
Under Assumptions MLR1-MLR4, the ordinary least squares (OLS) estimator, $\widehat{\beta}_j$, $j = 0, \cdots, K$ is unbiased. That is its expected value is equal to the population parameter,

$$E\left(\widehat{\beta}_j\right) = \beta_j,, \text{ for } j = 0, \cdots, K$$

- OLS estimator minimizes the sum of squared residuals. For the simple case of one regressor x_1, $\widehat{\beta} = (\widehat{\beta}_0, \widehat{\beta}_1)$ minimizes,

$$SSR(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_1)^2$$

Consider the case of multiple regressors:

$$y_i \quad = \quad \beta_0 + \beta_1 x_{i1} + \mu_i \tag{6}$$

The population regression coefficients β_0 and β_1 are defined by solving:

$$\beta_0, \beta_1 = \text{argmin}_{b_0, b_1} E\left[(y_i - b_0 - b_1 x_i)^2\right]$$

The first order conditions,

$$\frac{\partial E\left[(y_i - \beta_0 - \beta_1 x_i)^2\right]}{\partial \beta_0} = E\left[-2(y_i - \beta_0 - \beta_1 x_i)\right] \quad = \quad 0 \tag{7}$$

$$\frac{\partial E\left[(y_i - \beta_0 - \beta_1 x_i)^2\right]}{\partial \beta_1} = E\left[-2x_i(y_i - \beta_0 - \beta_1 x_i)\right] \quad = \quad 0 \tag{8}$$

Solving for β_0 and β_1:

$$\beta_1 = \frac{Cov(y_i, x_i}{V(x_i)} \tag{9}$$

$$\beta_0 = E\left[y_i\right] - \beta_1 E\left[x_i\right] \tag{10}$$

Consider the case of multiple regressors:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{iK} + \mu_i \tag{11}$$

Matrix notation:
Let $\mathbf{x}_i = (1, x_{i1}, \cdots, x_{iK})'$ be the $k \times 1$ vector of regressors (including the constant term) and $\beta = (\beta_0, \beta_1, \cdots, \beta_K)$, then:

$$y_i = \mathbf{x}_i' \beta + \mu_i \tag{12}$$

- Useful representation! The population regression coefficients are defined by:

$$\beta_k = \frac{Cov(y_i, \overline{x}_{ki})}{V(\overline{x}_{ki})}, \tag{13}$$

  where $\overline{x}_{ki}$ is the residual from **a regression of x_{ki} on all other variables**.
- Each coefficient in a multivariate regression is the bivariate slope coefficient for the corresponding regressor, after "**partialling out**" all the other variables in the model.

Assumption

MLR5. Homoskedasticity

The error term has the same variance given any values of the explanatory variables:

$$Var(\mu_i|x_{i1},\cdots,x_{iK})=\sigma^2$$

- Homoskedasticity : the variance of the error term does not depend on the explanatory variables.
- When is this a bad assumption?

  If omitted variables are not correlated with the included variables, but have a different order of magnitude for (groups of) observations.

Theorem

Gauss Markov
Under Assumptions MLR1-MLR5, OLS estimator is **BLUE**.

- What happens to OLS estimator if one/all of these assumptions does not hold?

- Suppose the correct model has two sets of variables,

$$y = X_1\beta_1 + X_2\beta_2 + \epsilon$$

- Compute least squares omitting X_2. Denote this estimator by $\widetilde{\beta_1}$ Some easily proved results:
  - $V(\widetilde{\beta_1})$ is smaller than $V(\widehat{\beta_1})$. i.e., you get a smaller variance when you omit X_2. (One interpretation: Omitting X_2 amounts to using extra information ($\beta_2 = 0$). Even if the information is wrong (see the next result), it reduces the variance. (This is an important result.)

- **(No free lunch)**

$$E[\widetilde{\beta_1}] = \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 \neq \beta_1.$$

So, $\widetilde{\beta_1}$ is **biased**.

The bias can be huge. Can reverse the sign of a price coefficient in a "demand equation."
$\widetilde{\beta_1}$ may be more "**precise**." Smaller variance but positive bias. If bias is small, may still favor the short regression.

- (Free lunch?) Suppose $X_1'X_2 = 0$. Then the bias goes away. Interpretation, the information is not "right," it is irrelevant. $\widetilde{\beta}_1$ is the same as $\widehat{\beta}_1$.
- It can be shown that

$$V(\widehat{\beta}_1) = \frac{\sigma^2}{SST_1(1 - R_1^2)}$$

where SST_1 is the total variation in X_1 and R_1 is the $R-squared$ from the regression of X_1 on X_2. Furthermore,

$$V(\widetilde{\beta}_1) = \frac{\sigma^2}{SST_1}$$

- when $\beta_2 \neq 0$, $\widetilde{\beta}_1$ is biased and $V(\widetilde{\beta}_1) < V(\widehat{\beta}_1)$;
- when $\beta_2 = 0$, both $\widetilde{\beta}_1$ and $\widehat{\beta}_1$ are unbiased, and $V(\widetilde{\beta}_1) < V(\widehat{\beta}_1)$;

- The variance of the OLS estimator of β_j, conditional on the sample values of the independent variables is

$$V(\widehat{\beta}_j) = \frac{\sigma^2}{SST_j(1-R_j^2)} \tag{14}$$

where $SST_j = \sum_{i=1}^{n}(X_{ij} - \overline{X}_j)^2$ is the total sample variation in X_j and R_j^2 is the R-squared from the regression of X_j on all other independent variables including constant term.

- The larger σ^2, the larger is the variance of OLS estimator. More noise means difficult to estimate the partial effect of any variable.
- The larger the total variation in X_j, the smaller is the variance of $\widehat{\beta}_j$. To increase the in sample variation of X_j, one can increase the sample size!

- The variance of an estimated coefficient will tend to be larger if there are other X's in the model that can predict X_j. This is reflected by a high R_j^2 in equation 14;
- The standard error of prediction will also tend to be larger if there are unnecessary or redundant X's in the model.

This is a variant on linear regression that downplays the influence of outliers

- First performs the original OLS regression
- Drops observations with Cook's distance > 1
- Calculates weights for each observation based on their residuals
- Performs weighted least squares regression using these weights.