# 1. Review of Multiple Regression Models

## (1) Statistical relationships

- Aim: Characterise the stochastic relationship between a variable and a set of 'related' variable, e.g.:

| | |
|---|---|
| $P = \alpha + \beta Q + \gamma Y$ | $Y = income,\ \ Q = quantity,\ \ P = price$ |

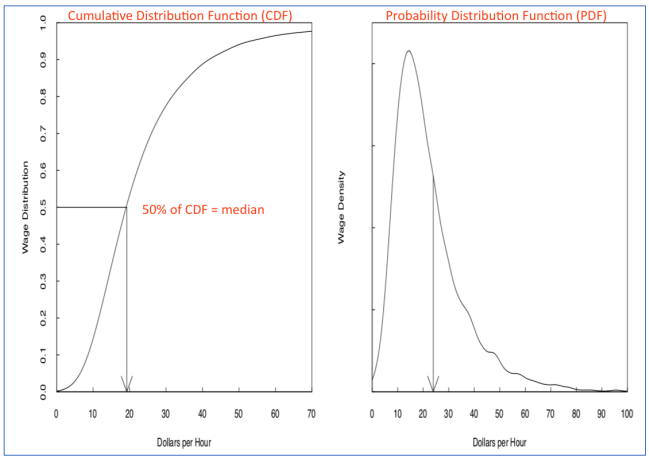
- Variation in $P$ arise with variation in $Q$ and random variation in its distribution
- There exists a conditional distribution $f(P|Q)$ and a conditional mean function $E[P|Q]$
- Variation in $P$ arise because of
  - Variation in the mean
  - Variation around the mean
  - (possibly) variation in a covariate, $Y$

## (2) Probability distribution

- E.g. wage rates in the US vary across workers – this can be described using a probability distribution by viewing wage as a random variable:

$$F(u) = \Pr\,(wage \le u)$$

- A person's wage is random, i.e. unknown before it is measured. Observed wages are realisations from the distribution $F$
- Generally, $F$ is unknown – we can learn about the distribution from many realisations of the $wage$ variable

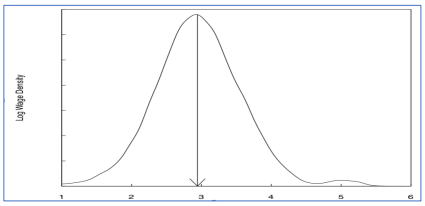



- Measures of central tendency:
  - **Median** – the median $m$ of a continuous distribution $F$ is the unique solution to $\mathrm{F(m)} = \frac{1}{2}$
  - **Mean** or **expectation** – a convenient, but not robust measure in the presence of substantial skewness or thick tails. The expectation of a RV $y$ with density $f$ is

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

## (3) Logarithmic transformation

- If there is substantial skewness and fat-tailed in the wage distribution → transform the data by taking the natural logarithmic $[\log\,(y)]$



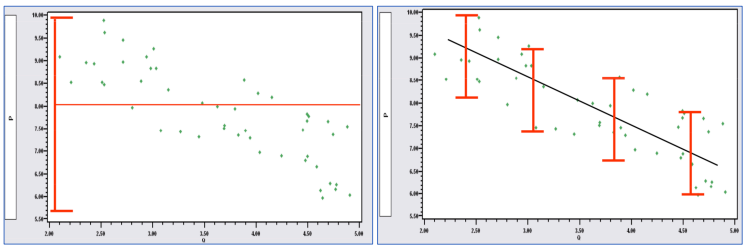

- The density of $\log(wage)$ is much less skewed and fat-tailed than the density of the level of wages, so its mean $E[\log(y)]$ is a better measure of central tendency of the distribution
- The geometric mean $\exp\{E[log(y)]\}$ is a robust measure of the central tendency of $y$

## (4) Conditional expectation

- Conditional expectation can be written with the generic notation – i.e. **conditional expectation function**:

$$E\left(\underbrace{y}_{dependent\ variable} \middle| \underbrace{x_1, x_2, \dots, x_k}_{multiple\ conditioning}\right) = m(x_1, x_2, \dots, x_k)$$

- Conditioning reduces variation → left plot below shows the variation of $P$ around $E(P)$. The right plot shows the variation around $E(P|Q)$



- For observational data $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$
  - If the data are cross sectional (CS) → reasonable to assume they are mutually independent
  - If the data are randomly gathered → reasonable to model each observation as a random draw from the same probability distribution (**independent and identically distributed** or **iid**)
- To study how the distribution of $y_i$ varies with $x_i$, 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:

| | |
|---|---|
| $y_i = E(y_i\|x_i) + (y_i - E[y_i\|x_i])$<br>$= E(y_i\|x_i) + u_i$ | $E(u_i\|x_i) = 0$<br>$u$ : conditional expectation functional error |

## (5) Linear regression model

- 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 → generally, we replace $m(\mathbf{x})$ with an approximation that is linear in $\mathbf{x}$
- It is convenient to augment the regressor vector $x$ by listing the number 1 as an element. This is called the constant or intercept term

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

  - $\mathbf{x} = (1, x_1, x_2, \dots, x_K)'$
  - $\beta = (\beta_0, \beta_1, \beta_2, \dots, \beta_K)'$
- Boldface letter indicates a column vector, i.e. one regressor $x$ and a constant term: $\beta = (\beta_0, \beta_1)$ and $\mathbf{x} = (1, x)'$ → $\mathbf{x}'\beta = \beta_0 + \beta_1 x$

## (6) Multiple Linear Regression (MLR) assumptions

**MLR1. Linearity**: The population model is linear in parameters (the $\beta$'s)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

- If MLR1 is violated, the model suffers from **functional form misspecification**. This occurs when:
  - The model does not account for some important nonlinearities (e.g. omitting important variables or squared terms)
- Generally, functional form misspecification causes biases in the remaining parameter estimators

**MLR2. Random Sampling**: We have random sampling of $n$ observations, $\{(x_{i1}, x_{i2}, \dots, x_{ik}, y_i) : i = 1,2, \dots, n\}$, following the population model in MLR1.

- MLR2 can be violated due to:
  - Missing Data
  - Non-random samples → biased and inconsistent OLS estimator
  - Outliers

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

- MLR3 is violated when all seasonal dummies and the constant term are included in a regression

**MLR4. Zero Conditional Mean (ZCM)**: The error term $u$ has a conditional expected value of zero given any values of the independent variables

$$E(u|x_1, \dots, x_k) = 0$$

- MLR4 can be violated due to:
  - Misspecification of the functional form
  - Omitted variable bias (OVB): Omitting important factors correlated with any of the regressors
  - Measurement error in the explanatory variables
  - Endogeneity: Some explanatory variables are determined jointly with the dependent variable

**MLR1-4. Unbiasedness**: The OLS estimator $\hat{\beta}_j$, $j = 0, \dots, K$ is unbiased. i.e. its expected value is equal to the population parameter

$$E(\hat{\beta}_j) = \beta_j \quad \text{for} \quad j = 0, \dots, K$$

**MLR5. Homoskedasticity**: The error term has the same variance given any values of the explanatory variables. i.e. the variance of the error term does not depend on the explanatory variables

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

- This is a bad assumption if omitted variables are not correlated with the included variables, but have different order of magnitude for (group of) observations – e.g.
  - Units has different size (e.g. states, cities), omitted variable may be larger for more populous states or cities
  - Units at different point in time. Omitted variables may be more important at some points in time
  - Units that face different restrictions on their behaviour

**MLR1-5. Gauss Markov Theorem**: OLS estimators are

- Best (minimum variance and most efficient)
- Linear
- Unbiased
- Estimators

## (7) Deriving OLS estimators

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

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

For $y_i = \beta_0 + \beta_1 x_i + u_i$, the population regression coefficient $\beta_0$ and $\beta_1$ are defined by solving:

$$\beta_0, \beta_1 = argmin_{b_0, b_1} E[(y_i - b_0 - b_1 x_i)^2]$$

By solving the first order condition:

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

$$\therefore \beta_0 = E[y_i] - \beta_1 E[x_i]$$

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

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

- For multiple regressors

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_K x_{iK} + u_i$$

Let $\mathbf{x}_i = (1, x_{i1}, \dots, x_{ik})'$ be the $k \times 1$ vector of regressors (including the constant term) and $\beta = (\beta_0, \beta_1, \beta_2, \dots, \beta_K)$, then

$$y_i = \mathbf{x}_i' \beta + u_i$$

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

  - $\bar{x}_{ki}$ : 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

## (8) Omitting relevant variable

- Suppose the correct model has two sets of variables

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

If we compute the OLS estimator $\tilde{\beta}_1$ while omitting $X_2$, and observe that $V(\tilde{\beta}_1) < V(\hat{\beta}_1)$ 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, it reduces the variance
- (No free lunch)

$$E[\tilde{\beta}_1] = \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 \neq \beta_1 \quad \rightarrow \tilde{\beta}_1 \text{ is biased}$$

  - The bias can be huge and even reverse the sign of the coefficient
  - $\tilde{\beta}_1$ may be more 'precise' (smaller variance) but has positive bias. If bias is small, we may still favour the short regression
- (Free lunch?) Suppose $X_1'X_2 = 0$, then the bias goes away. Interpretation: the information is irrelevant ($\tilde{\beta}_1$ is the same as $\hat{\beta}_1$)

| | |
|---|---|
| $V(\hat{\beta}_1) = \frac{\sigma^2}{SST_1(1 - R_1^2)}$ | • $SST_1$: Total variation in $X_1$<br>• $R_1^2$: $R^2$ from the regressing $X_1$ on $X_2$ |
| $V(\tilde{\beta}_1) = \frac{\sigma^2}{SST_1}$ | • When $\beta_2 \neq 0$, $\tilde{\beta}_1$ is biased and $V(\tilde{\beta}_1) < V(\hat{\beta}_1)$<br>• When $\beta_2 = 0$, $\tilde{\beta}_1$ and $\hat{\beta}_1$ are unbiased and $V(\tilde{\beta}_1) < V(\hat{\beta}_1)$ |

## (9) What affects the variance of OLS?

- Variance of the OLS estimator of $\beta_j$, conditional on the sample values of the independent variables:

| | |
|---|---|
| $V(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}$ | • $SST_j = \sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2$ → total sample of variation in $X_j$<br>• $R_j^2$ : $R^2$ from the regression of $X_j$ on all other independent variables including constant term |

- The larger $\sigma^2$, 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 $\hat{\beta}_j$. To increase the sample variation of $X_j$, increase the sample size

## (10) Multicollinearity

- 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 section (9)
- Standard error (SE) of prediction will also tend to be larger if there are unnecessary $X$'s in the model

## (11) Stata

- `describe` and `summarize`: To describe and report summary statistics (detect categorical variables, any suspicious values, etc.)
- **Regression diagnostics**
  - After estimating a model, check the entire regression for: normality of the residuals, omitted and unnecessary variables, heteroskedasticity (hetero)
  - Test individual variables for: outliers, collinearity, functional form
  - `xi` indicates that 'i.foreign' is a dummy variable

```
. xi: reg log_p weight mpg forXmpg i.foreign
i.foreign         _Iforeign_0-1       (naturally coded; _Iforeign_0 omitted)
```

- `rvfplot`: Check the residuals