AAKASH GHOSH (19MS129)

APPLIED MICROECONOMETRICS

Contents

1 Basic Modelling and Linear Regression using OLS

1.1 Introduction

We shall mainly deal with micro data[1] on economic variables. The main objectives are:

1. **Causal analysis:** We try to differentiate between correlation and causation. For example: 2 rocks falling from a hill is a correlation as one doesn't cause the other. If a rock hits other, and they move, then it's causation: one causes the other.

2. **Economic modelling:** We shall see economic models are good for causal analysis but not so good for forecasts[2]

1.2 Linear models

Consider the equation:

$$y = \alpha + \beta x + \epsilon \tag{1.1}$$

Here, y is the dependent variable or outcome and x is the independent variable or the variable which causes change in y. It is also called the regressor or observed factor. ϵ is random error due to unobserved factors. Without ϵ the model becomes deterministic; with ϵ it is stochastic. We assume:

$$E[\epsilon|x] = 0$$

From this assumption we have - $E[\epsilon] = 0$ and $Cov(x, \epsilon) = 0$. This implies x and ϵ are uncorrelated and that the errors are random. α, β are the parameters of the model. We define $\alpha = E[y|x = 0]$ and $\beta = E\left[\frac{\partial y}{\partial x}\right]$. β is also defined as the marginal effect of x on y.

[1] Macro data can be in a sense considered to be aggregate of micro data

[2] On the other hand Machine learning models are very good in forecasts.

1.2.1 *Normal equations*

Note that From (1) we have $\epsilon = y - \beta x - \alpha$. Applying $E[\epsilon] = 0$ and $Cov(x, \epsilon) = 0$ we get:

$$\sum_{i=1}^{n}(y_i - \alpha - \beta x_i) = 0 \tag{1.2}$$

$$\sum_{i=1}^{n} x_i(y_i - \alpha - \beta x_i) = 0 \tag{1.3}$$

This two equations are known as normal equations. We solve to get:[3]

$$\widehat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{1.4}$$

$$\widehat{\alpha} = \bar{y} - \widehat{\beta}\bar{x} \tag{1.5}$$

Here, $\widehat{\beta}$ and $\widehat{\alpha}$ are estimates of β and α and not the quantities themselves.

Example 1.1: Linear modelling consumption with respect to income

In equation 1.1, if y is the consumption and x is the income, then we might get parameters like $\widehat{\alpha} = ₹500$ and $\widehat{\beta} = 0.75$. There are two key points to note:

1. Even when income or $x = 0$, $E[y] \neq 0$. This is because even if a person has no income, he has a minimum level of consumption to sustain his life. This money maybe acquired from:
   - Loans
   - Govt. Subsidiaries
   - Stealing
   - Begging (We assume begging is not a profession)
2. $\widehat{\beta}$ is the increase in consumption given an unit increase in the income. This is known as the marginal propensity of consume. It is generally less than 1 as a person generally sets aside a fraction of his/her income as savings/investments.

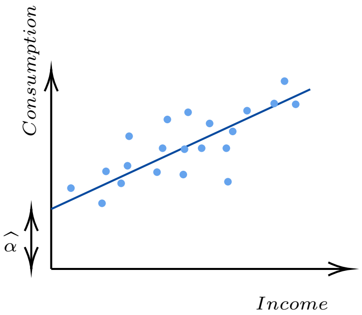


Figure 1.1: Income vs. Consumption modelled by a linear equation

1.2.2 *Linear models with multiple variables*

We consider the model given by:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots \tag{1.6}$$

Such a model is better because:

[3] **Estimates:** We believe that in nature there is some α and β which controls the parameters. We never know if the parameters we find are the actual parameters, we only know that they are estimates of the actual parameters.
Ref: Confidence in Parameters

1. We see the effects of other variables on y
2. Fitness of the model improves

For example if y is household consumption and x_1 is the income, then x_i's can be other factors like region of stay and number of mouths to feed.
In such models, a new question arises: Which variables do we need to consider. This is decided by two factors:

1. **Economic Theory:** In certain cases we have well established economic theory about the effect of one variable on others. **Example:** Effects of USD/INR exchange rates on exports.
2. **Economic Intuition:** Economists develop an intuition for economic variables and might take decisions based on them.

1.2.3 *Matrix representation of Linear models*

We define the following matrices:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_{11} & \dots & x_{k1} \\ 1 & x_{12} & \dots & x_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \dots & x_{kn} \end{bmatrix} \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \quad \beta = \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}$$

We can represent system of n linear models in k variables in the following way:

$$Y = X\beta + \epsilon \tag{1.7}$$

For quick reference, the final thing looks like this :

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \dots & x_{k1} \\ 1 & x_{12} & \dots & x_{k2} \\ 1 & x_{13} & \dots & x_{k3} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \dots & x_{kn} \end{bmatrix} \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Where the i^{th} equation reads:

$$y_i = \alpha + \sum_{r=1}^{k} x_{r,i}\beta_r + \epsilon_i$$

1.2.4 *Ordinary Least Squares(OLS) Estimation method*

We want to minimize the errors and one way to do this is by minimizing $\sum_{i=1}^{n} \epsilon_i^2 = \epsilon^t\epsilon$.[4]

$$\epsilon^t\epsilon = (Y - X\widehat{\beta})^t(Y - X\widehat{\beta}) \tag{1.8}$$

$$= Y^tY - \widehat{\beta}^tX^tY^t - Y^tX\widehat{\beta} + \widehat{\beta}^tX^tX\widehat{\beta} \tag{1.9}$$

Now note $M = Y^tX\widehat{\beta}$ is a scalar. Therefore, $M = M^t$ and $Y^tX\widehat{\beta} = \widehat{\beta}^tX^tY^t$. X^tX is a scalar, and we can separate $\widehat{\beta}^tX^tX\widehat{\beta}$ as $(X^tX)(\widehat{\beta}^t\widehat{\beta})$
Take derivative of 1.8 with respect to β to get:

[4] **Personal Note:** Things gets tricky here and none of the maths course I have taken has prepared me for this shit. Whatever is written here is found by looking up the following wiki pages: RSS,TSS,ESS and OLS. Additionally, explanation of formula is given HERE. Info about matrix calculus can be found HERE and recipes can be found HERE.

$$\frac{d}{d\beta}\epsilon^t\epsilon = -2X^tY + 2X^tX\widehat{\beta} = 0 \tag{1.10}$$

Simplify to get:

$$\widehat{\beta} = (X^tX)^{-1}X^tY \tag{1.11}$$

As long as X is full rank, X is positive and the β we obtain is a minimum. We define some terms:

List of formulas :

$$\frac{\partial x^t a}{\partial x} = \frac{\partial a^t x}{\partial x} = a$$

$$\frac{\partial a^t X b}{\partial X} = ab^t$$

$$\frac{\partial a^t X^t b}{\partial X} = ba^t$$

$$\frac{\partial a^t X a}{\partial X} = \frac{\partial a^t X^t a}{\partial X} = aa^t$$

$$\frac{\partial b^t A b}{\partial b} = 2Ab = 2b^tA$$

- **Explained sum of squares (ESS)**: A quantity used in describing how well a model, often a regression model, represents the data being modelled. In particular, the explained sum of squares measures how much variation there is in the modelled values.

$$ESS = \sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2 \tag{1.12}$$

- **Total sum of squares (TSS or SST)**: A quantity that appears as part of a standard way of presenting results of such analyses. For a set of observations y_i, $i \leq n$, it is defined as the sum over all squared differences between the observations and their overall mean.

$$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{1.13}$$

- **Residual sum of squares (RSS)**: The sum of the squares of residuals (deviations predicted from actual empirical values of data). It is a measure of the discrepancy between the data and an estimation model, such as a linear regression. A small RSS indicates a tight fit of the model to the data.

$$RSS = \sum_{i=1}^{n}(y_i - \widehat{y})^2 \tag{1.14}$$

We have the following result:

$$RSS = TSS - ESS \tag{1.15}$$

- **Coefficient of determination(R^2):** It is the proportion of the variation in the dependent variable that is predictable from the independent variable(s). It is a statistic used in the context of statistical models whose main purpose is either the prediction of future outcomes or the testing of hypotheses, on the basis of other related information. It provides a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model.

$$R^2 = \frac{ESS}{TSS} \tag{1.16}$$

High R^2 trap: Contrary to (my) Expectations, a very high R^2 probably means you have messed up your analysis/over-fitted your data. So take care

1.3 Non-Linear Models

Sometimes our data doesn't follow a linear tend. To take care of this we add an additional no-linear term to our model. Interpretations of some non-linear models are:

1. $y = \alpha + \beta_1 \ln x + \epsilon$
   It implies a 1% increase in x would increase y by 0.01β unit.

2. $\ln y = \alpha + \beta x + \epsilon$
   It implies a 1 unit change in x increases y by β times.

3. $\ln y = \alpha + \beta \ln x$
   It implies a 1% change in x produces a β% change in y

An example is given below.

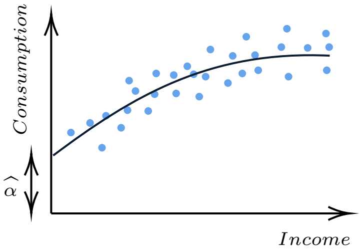


Figure 1.2: Non-Linear relation between Income and consumption. Note that the points don't have an exactly linear tendency

Example 1.2: Non-Linear modelling consumption with respect to income

For high income, the linear tendency of fig 1.1 is lost and is more like fig 1.2. To capture the diminishing slope we add a quadratic term to our model. The new model is therefore given by:

$$y = \alpha + \beta_1 x + \beta_2 x^2 + \epsilon \tag{1.17}$$

The marginal propensity to consume now is non-constant and is given by $\beta_1 + 2x\beta_2$. From the fact that the slop is decreasing with income, we conclude β_2 is negative.

1.4 Dealing with categorical parameter

It might so happen that a parameter we are dealing with is non-continuous and attains a discrete set of values. In that case, Eqn1.6 doesn't apply directly.

1.4.1 Parameters that take two values

We represent discrete parameters by d. Our model becomes:

$$y = \alpha + \beta_1 x + \beta_2 d \tag{1.18}$$

Where we assume that calculations are done with respect to a base category. For the category in question we set $d = 1$ otherwise(base case) $d = 0$. Therefore,

$$E[y|d=0] = \alpha + \beta_1 x \quad E[y|d=1] - E[y|d=0] = \beta_2 \tag{1.19}$$

Example 1.3: Variation of consumption with gender of head of family

It is found that consumption habits of household depends on gender of head of household. In female lead households, there is increased consumption of nutritious food. This has prompted governments of Latin American countries to pay cash subsidiaries in the hands of the woman of the house to make sure that the cash is used appropriately.
If we consider a male head of the house to be our base case then we have $\beta_2 > 0$ and if having a woman is the base case we have $\beta_2 < 0$

1.4.2 Parameters that take multiple values

If a parameter takes n values then we need $n-1$ variables. We fix a base parameter d_0. If the i^{th} value is present, we set $d_i = 1$ and $d_j(j \neq 0) = 0$

$$E[y|\text{Parameter takes } i^{th} \text{ value}] - E[y|\text{Parameter takes base value}] = \beta_i \tag{1.20}$$

The whole model reads:

$$y = \alpha + \beta x + \sum_{i=1}^{n-1} \beta_i d_i \tag{1.21}$$

For quick reference In case the parameter takes the i^{th} value, the equation reads:

$$y = \alpha + \beta_i + \beta x \tag{1.22}$$

Example 1.4: Variation of income with caste

In India, one's caste is quite important in deciding one's income. We consider 4 categories:

1. **General:** Considered to be the base case.
2. **SC:** Represented by d_1
3. **ST:** Represented by d_2
4. **OBC:** Represented by d_3

Now if $d_1 > d_2$ then we can conclude that SC have a higher expected income than ST. Due to the linear nature of Eq1.20, changing which category is taken as base is quite easy. The relevant equations in this case are:

$$
\begin{aligned}
&y = \alpha + \beta x + \beta_1 d_1 + \beta_2 d_2 + \beta_3 d_3 && \\
&y = \alpha + \beta x && \text{For general caste} \\
&y = \alpha + \beta x + \beta_1 && \text{For SC} \\
&y = \alpha + \beta x + \beta_2 && \text{For ST} \\
&y = \alpha + \beta x + \beta_3 && \text{For OBC}
\end{aligned}
$$

1.5 *Confidence in parameters*

1.5.1 *Estimators*

Ref: Casella, Ch7: Point Estimation. [5]

Given a data-set we would like to determine the parameters of the relation they are following. But as we are working with a finite dataset we may always get an estimate of a parameter and never the parameter itself. All we claim is the expected value of the estimate is the actual parameter itself. For example consider two variables related by $y = \alpha + \beta x$. We have a dataset given by $\{x_i, y_i\}_{i\in\mathbb{N}}$. We estimate $\widehat{\beta}$ using some method(like OLS). Then what we claim is $E[\widehat{\beta}] = \beta$.

Definition [Point Estimator]: A point estimator is a function of the sample.

We use an appropriate estimator to estimate the parameters. The definition is purposefully vague to make sure we don't accidentally rule out any candidate.

Definition [Estimate]: The realized value of the estimator for some sample is known as the estimate.

1.5.2 *Variance in estimates*

Due to randomness, it is natural to have some variance in the estimates as a function of the sample. If variance of the estimator is low, then we have a higher confidence in the parameter it is estimating. We make some assumptions here:

1. $\epsilon \sim N(0, \sigma^2)$
2. $\widehat{\beta} \sim N(\beta, \sigma^2(X^tX)^{-1}$

We define:

$$\text{Standard Error } = \sqrt{\widehat{Var(\widehat{\beta})}} = \sqrt{\widehat{\sigma^2(X^tX)}}$$

The estimated error is:

$$\hat{\epsilon} = y - \hat{\alpha} - \hat{\beta}x$$

[5] George Casella and Roger L. Berger. *Statistical inference*. Duxbury advanced series. Brooks/Cole, Cengage Learning, Belmont, Calif., 2. ed., internat. student ed., [nachdr.] edition, 20. ISBN 9780495391876

The estimated variance in estimated error is:

$$\sigma^2 = \frac{1}{n-k}\sum \hat{\epsilon_i}^2$$

where

$$n = \text{Sample Size}$$

$$k = \text{Number of independent in model}$$

The $n-k$ in the denominator is to account for the degrees of freedom so that the estimate is unbiased. For an unbiased estimator of a parameter p we have:

$$E[\hat{p}] = p$$

For a sample size of degree n modelled with k parameters, the degree of freedom is given by $n-k$. We define the t statistic with $n-k$ degree of freedom as $\frac{\hat{\beta}}{S.E(\hat{\beta})}$. If the degrees of freedom is greater than 32, then significance of the t-statistic= significance of $\hat{\beta}$ statistic.

1.5.3 *Hypothesis testing and measure of significance*

Definition [Significance] : Measure of allowance of error. A significance at 5% level means that the error in the estimation of the parameter is less than 5%.
At a low level of significance we fail to reject the null hypothesis. This is distinctly different from accepting the alternate hypothesis.

1.6 *Omitted Variable Bias*

Ref: Cameron Trivedi Omitted Variable Bias(Sec 4.7.4) for a more mathematical treatment.[6]
Suppose y is dependent on x_1, x_2, x_3, but we model it based on x_1, x_2. This might be done because:

1. We lack data on x_3

2. We didn't know that the dependence existed, i.e. there was a mistake.

If $Cov(x_1, x_3) = Cov(x_2, x_3) = 0$ then there is no problem in estimates of marginal quantities. The problem arises when this is not the case. For example, let $y = \alpha + \beta x + \beta' x' + \epsilon$ but is modelled as $y = \alpha + \beta x + \epsilon$. Let there be an omitted variable x' such that $Cov(x, x') \neq 0$. Then in the modelling the model will capture the effect of x' on y through x and thus the effect of x alone on y is downplayed and the estimate we get will be incorrect.

This is a very back of the paper analysis but gives the feel for the thing:

$$\begin{aligned}\widehat{\beta} &= E\left[\frac{\partial y}{\partial x}\right] \\ &= E\left[\frac{\partial \beta x}{\partial x} + \frac{\partial \beta' x'}{\partial x} + \frac{\epsilon}{\partial x}\right] \\ &= E\left[\frac{\partial \beta x}{\partial x}\right] + E\left[\frac{\partial \beta' x'}{\partial x}\right] + E\left[\frac{\partial \epsilon}{\partial x}\right] \\ &= \beta + \beta' E\left[\frac{\partial x'}{\partial x}\right]\end{aligned}$$

Assume $\beta' > 0$. Then $\widehat{\beta} > \beta$ if $Cov(x, x') > 0$ and $\widehat{\beta} < \beta$ if $Cov(x, x') < 0$.

[6] A. Colin Cameron and P. K. Trivedi. *Microeconometrics: methods and applications*. Cambridge University Press, New York, NY, 2005. ISBN 9780511125812

Example 1.5: Variation of income with age and education

It is obvious that income increases with both age and education. Suppose one is interested in the effect of education on income (and this interest is important as this is used to formulate the compulsory education plan: we need to have an accurate grasp of how income increases with increase in education, so we know if government funding of education is beneficial). As no. of years of education depends on age, if we don't include the effect of age in our calculations then our estimate of the effect will be biased(**Ref: Examples of regression in ch3**). By the side-note above we see that the estimate will be higher than the actual value.

One naive way to emit omitted variable bias is to consider every variable in the model. But this is not efficient as:

1. It is important o simplify the model as much as we can to reduce cost of computation.

2. Often we don't have enough data.

Ref: Instrumental Variable for Omitted bias correction.

1.7 *Interaction variables*

Suppose there is a categorical variable d which takes two values: 0 and 1. We are interested in knowing variation of the dependent variable y with respect to the independent variable x as v takes different values. We do this by defining an interaction variable $\tilde{x} = x \times d$(i.e. $\tilde{x_i} = x_i d_i$ for all pairs (x_i, d_i) in our dataset). Let our initial model was $y = \alpha + \beta x + \gamma d + \sum_r \beta_r x_r$ where x_r's are the other independent variables in consideration. We define our new model as $y = \alpha + \beta x + \gamma d + \tilde{\beta}\tilde{x} + \sum_r \beta_r x_r$. Then we have:

$$\frac{\partial y}{\partial x} = \beta + \tilde{\beta} d$$

as $\tilde{\beta}\tilde{x} = \tilde{\beta} x d$. Therefore, the marginal change in y with x β if $d = 0$ and $\beta + \tilde{\beta}$ if $d = 1$.

2 Instrumental Variables

2.1 Introduction

As mentioned before, direct regression of one variable over other might not give the proper marginal coefficient. In those cases we use instrumental variables.

2.2 Endogenity errors

In econometrics, endogeneity broadly refers to situations in which an explanatory variable is correlated with the error term. Endogenity errors are errors where the independent variable is not uncorrelated with the error terms. Examples of such errors are:

1. Omitted variable bias

2. Measurement errors

There are two ways to correct endogenity errors. They are:

1. Using an instrumental variable

2. Using a structured equation model.

2.3 Correction using Instrumental variable

We look for a variable z which is correlated to the dependent variable but is exogenous to the error term. Such a variable is called an instrumental variable. Such a variable satisfies two main conditions:

1. The IV is causally related to x(This is generally found by intuition)

2. The IV is uncorrelated with ϵ(This is verified statistically)

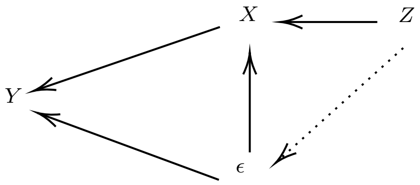


Figure 2.1: The solid line implies correlation. The dashed line implies no correlation.

In particular, the new estimate $\hat{\beta}$ is calculated as:

$$\hat{\beta}_{IV} \equiv \frac{\Delta y/\Delta z}{\Delta x \Delta z} = \frac{(Z^t Z)^{-1}(Z^t Y)}{(Z^t Z)^{-1}(Z^t X)} = (Z^t X)^{-1}(Z^t Y)$$

We note that this is different from $\hat{\beta}_{OLS} = (X^t X)^{-1}(X^t Y)$

2.4 2 step least square(2SLS)

Given an instrumental variable x, this method is used to find β.

1. Regress z on x
2. Predict $\tilde{x}$ using the above regression
3. Regress y on $\tilde{x}$

Intuitively, this method works because $\tilde{x}$ is basically x with the effect of ϵ stripped off (Due to no correlation between z and ϵ, z treats the effect of ϵ on x as random errors).

2.5 Example of 2SLS using simulated data in R

2.5.1 Creation of the simulated dataset

This requires the $MASS$ package. We simulate $x = 0.75 + z + v$. We correlate v and ϵ by $(v, \epsilon) \sim N(0, 0, \sum)$ where $\sum$ is the covariance matrix given by $\sum = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$ We therefore have $Cov(x, \epsilon) = Cov(z, \epsilon)$. We model $z \sim N(2, 1)$. Set $y = 0.25 + 0.5x + \epsilon$.

```
> library("MASS")
> set.seed(567)
> n<-10000
> mu<-c(0,0)
> cov<-matrix(c(1,0.8,0.8,1),ncol=2)
> er<-mvrnorm(n,mu,cov)
> dim(er)
[1] 10000    2
> e<-er[ ,1]
> v<-er[,2]
> z<-rnorm(n,2,1)
> x<-0.75+z+v
> y<-0.25+(0.5*x)+e
```

2.5.2 2SLS analysis on x, z, y

We shall assume that we only know about x, y and z and do a 2SLS analysis. First we shall do a normal regression of y on x and check the bias.

```
> reg<-lm(y~x)
> summary(reg)

Call:
lm(formula = y ~ x)
```

```
Residuals:
    Min     1Q Median     3Q    Max
-3.3054 -0.5500 0.0093 0.5397 3.5087

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.855263 0.017967 -47.6  <2e-16 ***
x           0.895699  0.005799  154.4  <2e-16 ***
---
Signif. codes: 0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

Residual standard error: 0.82 on 9998 degrees of freedom
Multiple R-squared: 0.7047, Adjusted R-squared: 0.7046
F-statistic: 2.385e+04 on 1 and 9998 DF, p-value: < 2.2e-16
```

As expected, the effect of x is overestimated and there is a positive bias. Now we shall regress z on x and predict $\tilde{x} = x1$.

```
> reg1<-lm(x~z)
> x1<-predict(reg1,data=z)
```

Now we shall regress y on X to get $\hat{\beta}_{IV}$.

```
> reg2<-lm(y~x1)
> summary(reg2)

Call:
lm(formula = y ~ x1)

Residuals:
    Min     1Q Median     3Q    Max
-4.9194 -0.9456 0.0069 0.9678 5.1786

Coefficients:
           Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.23087   0.04166  5.542 3.07e-08 ***
x1          0.50169   0.01420 35.319 < 2e-16 ***
---
Signif. codes: 0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

Residual standard error: 1.423 on 9998 degrees of freedom
Multiple R-squared: 0.1109, Adjusted R-squared: 0.1108
F-statistic: 1247 on 1 and 9998 DF, p-value: < 2.2e-16
```

We note that $\hat{\beta}_{IV}$ gives a much better estimate than β_{OLS}

3 Probit models

3.1 Introduction

A model which has categorical output can be divided in three types:

1. **Single choice model:** The dependent variable has two possible outcomes. Examples are: Predicting if a loan/credit card is to be approved, if a student will take a STEM major, etc. Such a variable can be assumed to take two values: 0 or 1

2. **Multiple choice model with unordered outcomes:** The dependent variable can have multiple outcomes, but they are unordered. Examples are: Picking one of the multiple streams with equivalent values in college.

3. **Multiple choice model with ordered outcomes:** The dependent variable can have multiple outcomes, but they are unordered. One example is employment status: there are three states: unemployed, part-time employed and full time employed.

Here, we will mainly focus on the models of the first type. If we do an OLS, then there is a possibility that we will get values other than 0 or 1(and in some cases, we might get a negative value as well). So we go for a non-linear model.

3.2 Dummy variables

A dummy variable is one that takes only the value 0 or 1 to indicate a choice of a single choice model. They can be thought of as numeric stand-ins for qualitative facts in a regression model, sorting data into mutually exclusive categories (such as smoker and non-smoker). We define a cutoff value c_0. If $y*$ is the dummy variable corresponding to y and if estimated value of $y* > c_0$ then we assume y takes the value corresponding to $y* = 1$ otherwise it takes the values corresponding to $y* = 0$.

3.3 Probit model

Some commonly used models are probit and logit. In particular, $\hat{\beta}_{probit} = \hat{\beta}_{logit} \times 0.87$. We assume the following:

1. $\epsilon \sim N(0, 1)$
2. Parameters β and σ^2 can't be identified separately: one depends on the other. What we instead try to do is identify β/σ^2. In particular, we scale the data so that $\sigma^2 = 1$

If our model is of the form $y = \alpha + x\beta + \epsilon$ then $y > c \Rightarrow (\alpha - c) + \beta x + \epsilon > 0$ Now we can choose $y*$ appropriately to normalize everything to fit our assumption.

3.4 Goodness of fit for a probit model

R^2 is no longer a good measure as y takes only two values. A better measurement is the following index:

$$\text{Percent correctly predicted} = \frac{\text{Total no. of correct prediction}}{\text{Total no. of predictions}}$$

If $\hat{y}$ is the predicted value of y then correct prediction occurs when $(\hat{y}, y) = (0, 0)$ or $(1, 1)$. An incorrect prediction occurs when $(\hat{y}, y) = (0, 1)$ or $(1, 0)$. We generally assume $c_0 = 0.5$

3.5 Receiver Operating Characteristic(ROC) curve

This is a plot of % correct prediction of $y = 1$ vs % incorrect prediction of $y = 0$ for varying values of c_0. A higher area under the curve implies a better model. For $c = 0$ we note that there is no correct prediction of $y = 1$ and all prediction of $y = 0$ is incorrect. Conversely, for $c = 0$ we note that there is no incorrect prediction of $y = 1$ and all prediction of $y = 0$ is correct.

3.6 Use of probit models

1. Used to check if a credit card/loan application is fit. A probit model based on previous customer records is useful in this case.
2. Used in consulates for application of visa.

3.7 Insample and Outsample Prediction

Insample Prediction is predicting the result a of the data from the data itself. Outsample prediction is predicting data different from the data used in sample. % correctly predicted of outsample prediction is a better fitness of test.

4 *Example of data analysis on R using simwagem.txt*

4.1 *Introduction*

Here the data is given in text format. Data in this format has two main benefits:

1. It is easier to read
2. It has a smaller file size

The data contains wage analysis of labor market for males. The analysis is done separately for males and females because of two reasons:

1. Factors like having a family/kids (social factors) affect wage
2. There is a pay gap between males and females for the same work. This is discriminatory in nature.

Discrimination: When there is no clear reason for a difference in treatment of one human with respect to other, we say the cause of the difference is discrimination.

4.2 *Preliminary Analysis*

In the preliminary analysis we look at

1. Number of observations
2. Number of variables
3. Idea about the parameters and their values

In this case such an analysis may include things like looking at person's id, age, martial status, religion, caste etc. For categorical variables, the majority group is taken as the base. This helps in better presentation. In *R* such an analysis can be made with table and summary commands.

1. **Summary:** This is good for continuous data. In our dataset, such variables include age and earning.

```
> summary(df$age)
   Min. 1st Qu. Median   Mean 3rd Qu.   Max.
  15.00   30.00  38.00  38.36  47.00  60.00
> summary(df$earnings)
   Min. 1st Qu. Median   Mean 3rd Qu.   Max.
  194.9   431.1  552.4  578.2  697.8 1164.0
```

2. **Table:** This is good for discrete data. In this case, they include things like martial status, religion being muslim, being urban etc.

```
> table(df$single)
    0     1
88275 15658
> table(df$muslim)
    0     1
89221 14712
> table(df$urban)
    0     1
61520 42413
```

We would in general consider years of education to be continuous. But this is not so. We take a look at the following output:

```
> table(df$eduyrs)

    0     5     8    10    12    14    15    18
26266 13534 20088 16183 10351 2203 11369 3939
```

We note that there are actually only a few values that the column takes. This corresponds to each major step in education: 5 years is primary education, 8 years is middle school, 10 years is matriculation, 12 years is higher secondary education and so on. Note that most people have 10 years of education. Generally, at this stage one is 15 years of age and is able to work legally. Apart from that, there is also government incentives which serves as secondary motivation.

4.3 *Analysis: Effect of Education on Earning*

We define earning as the money we get for our work. This is different from income and mostly consist of our salary. Income which don't contribute to earning are stuff like money from tenants, interests etc. Take a look at the following code and figure 3.1.

```
> summary(df$earnings)
Min. 1st Qu. Median   Mean 3rd Qu.   Max.
194.9   431.1  552.4  578.2  697.8  1164.0
```

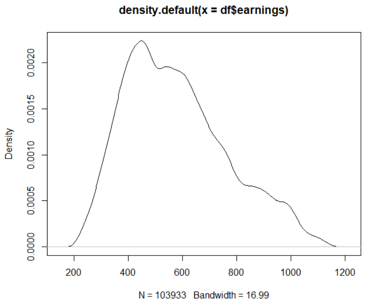


Figure 4.1: Density plot of monthly earning. It is easy to conclude that people are generally poor.

We note the following about employment in India

1. The minimum wage is close to the Govt. minimum wage. This is to prevent exploitation.
2. In rural areas, you are eligible to a job card. The card guarantees 100 days of work per year
3. In case, the government is unable to procure suitable work for you, then you still get paid the Govt. mandated minimum wage.

We do a naive regression of earnings on years of education.

```
> reg<-lm(earnings~eduyrs,data = df)
> summary(reg)

Call:
lm(formula = earnings ~ eduyrs, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-265.999 -79.272   3.259  77.503 275.537

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 353.15207  0.54125  652.5  <2e-16 ***
eduyrs      29.73722   0.05817  511.2  <2e-16 ***
---
Signif. codes: 0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

Residual standard error: 101.5 on 103931 degrees of freedom
Multiple R-squared: 0.7155, Adjusted R-squared: 0.7155
F-statistic: 2.614e+05 on 1 and 103931 DF, p-value: < 2.2e-16
```

We get $R^2 \sim 71.5\%$. But this analysis is not exactly correct. There are other factors in play. For example, a higher age corresponds to more experience and thus a higher income.

```
> reg<-lm(earnings~eduyrs+age,data = df)
> summary(reg)

Call:
lm(formula = earnings ~ eduyrs + age, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-230.519 -71.704   2.092  74.729 173.149

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 135.39391  1.00186  135.1  <2e-16 ***
eduyrs      30.62081   0.04673  655.3  <2e-16 ***
```

```
age          5.50303   0.02283  241.1   <2e-16 ***
---
Signif. codes: 0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

Residual standard error: 81.28 on 103930 degrees of freedom
Multiple R-squared: 0.8175, Adjusted R-squared: 0.8175
F-statistic: 2.328e+05 on 2 and 103930 DF, p-value: < 2.2e-16
```

We note by also considering age to be an important independent variable controlling education, we have a higher R^2 of $\sim 81.7\%$.[1] We now do one more regression analysis: one which takes all the variable into account.

```
> reg<-lm(earnings~age+single+eduyrs+selfemp
+regularemp+urban+st+sc+obc+muslim+otherr,data=df)
> summary(reg)

Call:
lm(formula = earnings ~ age + single + eduyrs + selfemp +
    regularemp +
   urban + st + sc + obc + muslim + otherr, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-121.181 -25.954   0.291  33.178 112.668

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 101.32540  0.68272 148.415 < 2e-16 ***
age          5.29716   0.01355 390.972 < 2e-16 ***
single      -3.19974   0.41293 -7.749 9.36e-15 ***
eduyrs      26.38504   0.02648 996.250 < 2e-16 ***
selfemp     23.90837   0.33478 71.415  < 2e-16 ***
regularemp  45.48712   0.38584 117.892 < 2e-16 ***
urban      138.32079   0.26696 518.132 < 2e-16 ***
st          -7.68017   0.45671 -16.816 < 2e-16 ***
sc          -9.20168   0.40521 -22.708 < 2e-16 ***
obc         -7.04464   0.30392 -23.179 < 2e-16 ***
muslim      -7.45145   0.37556 -19.841 < 2e-16 ***
otherr      -3.00097   0.45105 -6.653 2.88e-11 ***
---
Signif. codes: 0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

Residual standard error: 40.06 on 103921 degrees of freedom
Multiple R-squared: 0.9557, Adjusted R-squared: 0.9557
F-statistic: 2.037e+05 on 11 and 103921 DF, p-value: < 2.2e-16
```

We note that taking more variables improves our estimates R^2. This regression is important because of the following fact: of all the factors that contribute to education, no. of years of education is the only

[1] **Mincerian Wage Equation:** The Mincerian wage equation models wage by years of schooling and experience in the labour market.(Copied from Wikipedia)

variable that can be improved by government policy. In particular, we note one year of education improved income by ₹26.39. This analysis is also done with *log* i.e how many times does the earnings increase with education(**Ref: Non-Linear models**).

```
> df$lnearnings<-log(df$earnings)
> reg<-lm(lnearnings~age+single+eduyrs+selfemp
+regularemp+urban+st+sc+obc+muslim+otherr,data=df)
> summary(reg)

Call:
lm(formula = lnearnings ~ age + single + eduyrs + selfemp +
    regularemp +
   urban + st + sc + obc + muslim + otherr, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.43261 -0.02919 0.01062 0.04444 0.16697

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.433e+00 1.125e-03 4829.045 < 2e-16 ***
age         9.741e-03 2.233e-05 436.292 < 2e-16 ***
single     -5.662e-03 6.805e-04  -8.320 < 2e-16 ***
eduyrs      4.741e-02 4.364e-05 1086.157 < 2e-16 ***
selfemp     7.126e-02 5.517e-04 129.162 < 2e-16 ***
regularemp  9.372e-02 6.358e-04 147.396 < 2e-16 ***
urban       2.088e-01 4.399e-04 474.592 < 2e-16 ***
st         -1.375e-02 7.526e-04 -18.272 < 2e-16 ***
sc         -8.478e-03 6.678e-04 -12.696 < 2e-16 ***
obc        -2.349e-03 5.008e-04  -4.690 2.74e-06 ***
muslim     -8.332e-03 6.189e-04 -13.462 < 2e-16 ***
otherr      1.585e-03 7.433e-04   2.132   0.033 *
---
Signif. codes: 0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

Residual standard error: 0.06601 on 103921 degrees of freedom
Multiple R-squared: 0.9606, Adjusted R-squared: 0.9606
F-statistic: 2.305e+05 on 11 and 103921 DF, p-value: < 2.2e-16
```

Therefore, 1 year of education increases income by 4.7%

We note that being single decreases income. This is because, generally, being married implies that you have enough income to support your family

4.4 *Analysis: Is return to education higher in urban area?*

Ref: Interaction Variables

This is a direct application of interaction variables. We show the required code and output below.

```
> df$interact<-df$eduyrs*df$urban
> reg<-lm(earnings~age+single+eduyrs
+selfemp+regularemp+urban+st+sc+obc+muslim+otherr+interact,data=df)
> summary(reg)

Call:
lm(formula = earnings ~ age + single + eduyrs + selfemp +
    regularemp +
    urban + st + sc + obc + muslim + otherr + interact, data
        = df)

Residuals:
    Min      1Q  Median      3Q     Max
-45.744 -6.806 -0.004  6.808  50.282

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 148.831707 0.176647 842.536 < 2e-16 ***
age          5.000869  0.003430 1458.019 < 2e-16 ***
single      -3.196287  0.104279 -30.651  < 2e-16 ***
eduyrs      20.058044  0.008424 2380.935 < 2e-16 ***
selfemp     27.115272  0.084584 320.574  < 2e-16 ***
regularemp  48.728864  0.097472 499.926  < 2e-16 ***
urban       18.452035  0.118166 156.154  < 2e-16 ***
st          -8.126321  0.115336 -70.458  < 2e-16 ***
sc          -5.025905  0.102386 -49.088  < 2e-16 ***
obc         -3.021164  0.076819 -39.328  < 2e-16 ***
muslim      -0.668853  0.095001  -7.040 1.93e-12 ***
otherr      -0.485820  0.113924  -4.264 2.01e-05 ***
interact    14.952930  0.012106 1235.158 < 2e-16 ***
---
Signif. codes: 0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

Residual standard error: 10.12 on 103920 degrees of freedom
Multiple R-squared: 0.9972, Adjusted R-squared: 0.9972
F-statistic: 3.055e+06 on 12 and 103920 DF, p-value: < 2.2e-16
```

Therefore, in urban areas, one year of education gives an additional income of ₹14.95

4.5 Prediction: Being self-employed

In most first world countries, being self-employed is a sign of entrepreneur mindset i.e. someone who has big ideas and is willing to work towards it. On the other hand, in Third World countries, this is often an indentation that a person has failed to secure a stable job. In our data set, if a person is self-employed, it is marked as 0. We try to model it based on age, education, location(urban or rural) and martial status.

```
> probit1<-glm(selfemp~age+eduyrs+urban+single,
family=binomial(link="probit"),data=df)
> summary(probit1)

Call:
glm(formula = selfemp ~ age + eduyrs + urban + single, family
    = binomial(link = "probit"),
   data = df)

Deviance Residuals:
   Min       1Q   Median       3Q      Max
-1.5117 -1.1062 -0.6935  1.1619  1.9103

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.5237845 0.0189361 -27.66 <2e-16 ***
age          0.0165732 0.0004227  39.20  <2e-16 ***
eduyrs      -0.0099715 0.0007519 -13.26  <2e-16 ***
urban       -0.2425356 0.0082958 -29.24  <2e-16 ***
single      -0.3917176 0.0136754 -28.64  <2e-16 ***
---
Signif. codes: 0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

(Dispersion parameter for binomial family taken to be 1)

   Null deviance: 143251 on 103932 degrees of freedom
Residual deviance: 136823 on 103928 degrees of freedom
AIC: 136833

Number of Fisher Scoring iterations: 4

> pr<-predict(probit1,type="response",data=df)
> yhat<-as.numeric(pr>0.5)
```

Now once we have the predicted values of y(which is yhat) we can find % correctly predicted.

```
> c_pr<-as.numeric(df$selfemp==yhat)
> table(c_pr)
c_pr
    0     1
41855 62078
```

The index can be easily calculated after this.

4.6 *Analysis: Fitness of probit model*

Ref: Insample vs. Outsample prediction
We first divide the data in two parts: one for training and one for testing

```
> nrow(df)
[1] 103933
> df$n<-seq(1,nrow(df),1)
> train<-df[df$n<51966,]
> test<-df[df$n>51966,]
```

Now we model the train data and test it on test data.

```
>
    probit<-glm(selfemp~age+eduyrs+urban+single,family=binomial(link="probit"),data=train)
> pr<-predict(probit1,type="response",newdata=test)
> yhat<-as.numeric(pr>0.5)
> c_pr<-as.numeric(test$selfemp==yhat)
> x<-table(c_pr)
> x
c_pr
    0     1
21308 30659
> x[2]/(x[1]+x[2])
        1
0.5899706
```

The % correctly predicted in outsample prediction is ≈ 59%

5 R cookbook

Note: Due to randomness, there might be small deviations from the data shown here. This is OK

5.1 Introduction to R environment

1. To get current working directory

```
> getwd()
[1] "/MS/ag19ms129"
```

2. To change working directory

```
> setwd("~/Documents/HU codes")
> getwd()
[1] "/MS/ag19ms129/Documents/HU codes"
```

3. Simulate a normal distribution of x with $n = 10000, \mu = 50, \sigma = 5$ and e with $n = 10000, \mu = 0, \sigma = 5$

```
> x<-rnorm(10000,50,5)
> e<-rnorm(10000,0,5)
```

Here $\leftarrow$ is the assignment operator. $rnorm(n, \mu, \sigma)$ generates a n large dataset with mean μ and variance σ. Thus, the first line means the variable is assigned the value x

4. Check mean of x, variance of x and how distribution of e looks like

```
> mean(x)
[1] 49.88043
> var(x)
[1] 24.90584
> de<-density(e)
> plot(de)
```

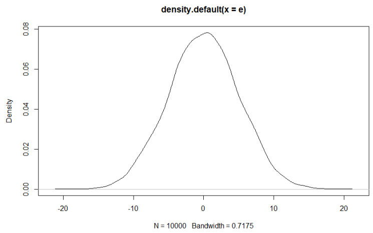
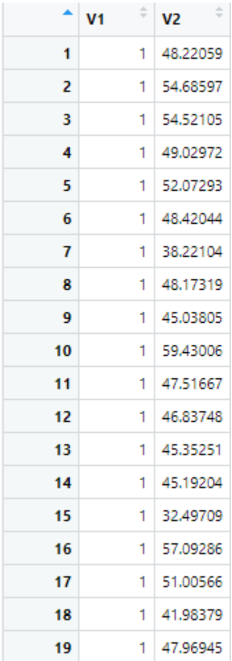


Figure 5.1: Density plot of e. The main thing to note is the rough bell shaped figure

5. Simulate $y = 2.5 + 0.75x + e$

```
> y<-2.5+(0.75*x)+e
> mean(y)
[1] 39.94653
> var(y)
[1] 39.66292
```



	V1	V2
1	1	48.22059
2	1	54.68597
3	1	54.52105
4	1	49.02972
5	1	52.07293
6	1	48.42044
7	1	38.22104
8	1	48.17319
9	1	45.03805
10	1	59.43006
11	1	47.51667
12	1	46.83748
13	1	45.35251
14	1	45.19204
15	1	32.49709
16	1	57.09286
17	1	51.00566
18	1	41.98379
19	1	47.96945

Figure 5.2: How X looks like. Snippet from $R - Studio$

6. Make required X matrix

```
> X<-matrix(data=NA,10000,2)
> X[,1]<-1
> X[,2]<-x
```

Leaving the first entry blank implies change is to be applied to all the columns

7. We now make the required Y matrix. We do like we did for X

```
> Y<-matrix(data=NA,10000,1)
> Y[,1]<-y
```

5.2 Analysis

Once we have a simulated X and Y, we shall pretend to not know about the relation between them, and try to find estimates parameters α and β (as per eqn1.1) which might have generated this.
There are commands to automate most of the things needed. But we will work through all the steps once to get a feel for things.

5.2.1 Manual Walk-through

8. We will find $\widehat{\beta}$ using eqn1.11 First we shall calculate $(X^tX)^{-1}$ and call it $XtXi$. Then we shall calculate X^tY and call it XtY. Then we multiply them to get B (our estimate of $\widehat{\beta}$). This, of course, can be done in a single step. Note that the $solve(M)$ command is used to find inverse of matrix M. To get product of two matrices A, B we use $A\% * \%B$[1]

```
> XtXi<-solve(t(X)%*%X)
> XtY<-t(X)%*%Y
> B<-(XtXi)%*%XtY
> B
          [,1]
[1,] 2.0679723
[2,] 0.7593873
```

[1] **Inaccurate $\widehat{\alpha}$ fix:**
One might notice that their $\widehat{\alpha}$ is rather inaccurate. This is OK: the most important thing is accuracy of $\widehat{\beta}$

Therefore, we get:

$$\widehat{\alpha} = 2.06\ldots \quad \widehat{\beta} = 0.759\ldots \tag{5.1}$$

9. Finding the standard error manually

5.2.2 *Analysis of simwagem*

Bibliography

A. Colin Cameron and P. K. Trivedi. *Microeconometrics: methods and applications.* Cambridge University Press, New York, NY, 2005. ISBN 9780511125812.

George Casella and Roger L. Berger. *Statistical inference.* Duxbury advanced series. Brooks/Cole, Cengage Learning, Belmont, Calif., 2. ed., internat. student ed., [nachdr.] edition, 20. ISBN 9780495391876.