Two variable linear regression analysis

1. Introduction

Econometrics has three major uses:

1. Describing economic reality.
2. Testing hypotheses about economic theory.
3. Forecasting future economic activity.

Econometricians attempt to quantify economic relationships that had previous been only theoretical. To undertake this requires 3 steps:

1. Specifying/identifying theoretical economic relationship between the variables.
2. Collecting the data on those variables identified by the theoretical model.
3. Obtaining estimates of the parameters in the theoretical relationship.

2. Correlation vs Regression analysis

2.1 Correlation

In Economics we are interested in the relation between 2 or more random variables, for example:

- Sales and advertising expenditure
- Personal consumption and disposable income
- Investment and interest rates
- Earnings and schooling

While there are many ways in which these pairs of variables might be related – a linear relationship is often a useful first approximation and this can be detected via a scatter plot, of one variable against the other.

A simple measure of linear association between two random variables *x* and *y* is the covariance, which for a sample of *n* pairs of observations $(x_1, y_1)\ldots(x_n, y_n)$ is calculated as[1]:

$$\text{cov}(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

The covariance measures the average cross product of deviations of *x*, around its mean, with *y*, around its mean. Consequently, if high (low) values of *x* - relative to its

[1] In EC124 we clearly distinguished between the random variable, *X*, and its realisation, *x*, here we are only going to use lower case letters for both the random variable and its realisations.

mean - are associated with high (low) values of *y* - relative to its mean – then we get a high positive covariance (see Figure 1). Conversely if high (low) values of *x* are associated with low (high) values of *y* we get a negative covariance (see Figure 2). A zero covariance occurs when there is no predominant association between the *x* and *y* values (see Figure 3). NOTE: The covariance is a linear association between *x* and *y* values and would be approximately zero for a quadratic association (see Figure 4).

The covariance measure is not scale free and multiplying the *x* variable by 100 multiplies the covariance by 100. A scale free measure is a correlation:

$$corr(x,y) \equiv \rho(x,y) = \frac{\text{cov}(x,y)}{\sqrt{V(x)V(y)}}$$

1. $-1 \le \rho(x,y) \le 1$
2. $\rho(x,y) = -1 \Rightarrow$ perfect negative association
3. $\rho(x,y) = 1 \Rightarrow$ perfect positive linear association
4. $\rho(x,y) = 0 \Rightarrow$ no linear association
5. As $|\rho(x,y)|$ increases $\Rightarrow$ stronger association.
6. $\rho(x,y) = \rho(y,x)$

(see handout 1 for some rules on expectations and variances).

2.2 Regression

Linear regression looks at the linear association between the random variables *x* and *y*, but in this case CAUSATION or DEPENDENCY is important. In particular, we talk about the variable, *x*, taking a specific value and we are interested in the response of *y* to a change in this value of *x*. So in our examples above we might be interested in

- Changes in sales caused by increased advertising expenditure
- Changes in personal consumption caused by increased disposable income
- Changes in investment caused by increased interest rates
- Changes in earnings caused by increased schooling

In the simplest type of linear regression analysis we model the relationship between 2 variables *y* and *x* and this is assumed to be a linear relationship. In particular, we are interested in the expected value of the random variable, *y*, given a specific value for *x*. Given linearity this is:

$$E(y \mid x) = \alpha + \beta x$$

when

$E(y \mid x=0)=\alpha$ = expected value of *y* when *x*=0 (invariably do not interpret this)

$E(y \mid x+1)=\alpha+\beta(x+1)$

therefore,

$\beta = E(y \mid x+1)-E(y \mid x)$ = change in the expected value of *y* for a unit increase in *x*.

y – is known as the **dependent variable** (**endogenous**)

x – is known as the **explanatory variable** (**exogenous**).

The actual values of the dependent variable, *y*, will invariably not be the same as the expected value. We denote the discrepency (error or disturbance) between the actual and expected value by ε_i, such that,

$\varepsilon_i = y_i - E(y_i \mid x_i) = y_i - \alpha - \beta x_i$

such that by rearranging we have

$$y_i = \alpha + \beta x_i + \varepsilon_i \qquad (1)$$

and this is the TRUE (but unknown) relationship between *y* and *x* and is made up of two components:

1. $\alpha + \beta x_i$ - the systematic part
2. ε_i - the random (non-systematic) component.

<u>3. Classical linear regression model (CLRM) assumptions</u>

To complete the model we need to specify the statistical properties of *x* and ε

1. $E(\varepsilon_i \mid x_i) = E(\varepsilon_i) = 0$ (so the error term is independent of x_i).
2. $V(\varepsilon_i \mid x_i) = \sigma^2$ (error variance is constant (homoscedastic) – points are distributed around the true regression line with a constant spread)
3. $\text{cov}(\varepsilon_i, \varepsilon_j \mid x_i) = 0 \quad i \neq j$ (the errors are serially uncorrelated over observations)
4. $\varepsilon_i \mid x_i \sim N(0, \sigma^2)$

Figure 5 for a plot of a TRUE regression and the distribution of points around this which are consistent with the CLRM assumptions. Figure 6 plots the actual TRUE disturbance terms from Figure 5, which are consistent with the CLRM assumptions.

4. Several important questions to be answered:

1. How do we estimate the population parameters (α, β, σ^2) of this statistical model?
2. What are the properties of the estimators?
3. How do we test hypotheses about these parameters?
4. How strong is the relationship?
5. Is the model adequate?

4.1 Estimation of the sample regression line

The statistical model is:

$y_i = \alpha + \beta x_i + \varepsilon_i \quad i = 1, \ldots n$

and we assume the CLRM assumptions above are satisfied. We are interested in estimating the population parameters, α, β, σ^2, we will call the estimates, *a*, *b* and s^2.

For *n* pairs of observations $(x_1, y_1) \ldots (x_n, y_n)$ we would like to find the straight line that best fits these points, denote:

$\hat{y}_i = a + bx_i$

as the predicted values from the regression line. In which case we can define

$y_i = \hat{y}_i + e_i = a + bx_i + e_i$

(where *e* are the residuals – difference between the actual value of *y* and its predicted value, $\hat{y}_i$), such that, $e_i = y_i - \hat{y}_i$, see Figure 7 for a diagrammatic illustration of these points).

For given values of *a* and *b* we can define a regression line (in Figure 8 we plot three alternative regression lines for a_i and b_i i=1,2,3). But we want *a* and *b* to have some desirable properties. The best estimates are those that make the residuals, e_i, as small as possible. However, as residuals can be both positive and negative, obtaining lines such that $\sum_{i=1}^{n} e_i = 0$ can yield a variety of equally good lines (in fact any line which cuts through $\bar{x}$ and $\bar{y}$ will have $\sum_{i=1}^{n} e_i = 0$. In fact in Figure 8 all lines have this property. The optimal solution is to minimise the *RSS* (Residual Sum of

Squares) $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$, with respect to the two unknown parameters *a* and *b*. Th achieve this we must differentiate the *RSS* expression with respect to *a* and *b* and set the resultant expressions equal to zero – this process of minimising the *RSS* to obtain the parameter estimates is known as **Ordinary Least Squares (OLS)**.

Differentiating the *RSS* with respect to the parameter, *a*, and setting the expression to zero

$$\frac{\partial(\sum_{i=1}(y_i - a - bx_i)^2}{\partial a} = \frac{\partial\left[(y_1 - a - bx_1)^2 + ... + (y_n - a - bx_n)^2\right]}{\partial a} = 0 \qquad (2)$$

This entails differentiating each of the *n* terms in equation (2) with respect to *a*

$$\frac{\partial(\sum_{i=1}(y_i - a - bx_i)^2}{\partial a} = -2(y_1 - a - bx_1) - ... - 2(y_n - a - bx_n) = 0$$

$$= -2\sum_{i=1}^{n}(y_i - a - bx_i) = -2\sum_{i=1}^{n} e_i = 0 \Rightarrow \sum_{i=1}^{n} e_i = 0 \qquad (3)$$

Differentiating the *RSS* with respect to the parameter, *b*, and setting the expression to zero

$$\frac{\partial(\sum_{i=1}(y_i - a - bx_i)^2}{\partial b} = \frac{\partial\left[(y_1 - a - bx_1)^2 + ... + (y_n - a - bx_n)^2\right]}{\partial b} = 0 \qquad (4)$$

differentiating each of these *n* terms in equation (4) with respect to *b*

$$\frac{\partial(\sum_{i=1}(y_i - a - bx_i)^2}{\partial b} = -2x_1(y_1 - a - bx_1) - ... - 2x_n(y_n - a - bx_n) = 0$$

$$= -2\sum_{i=1}^{n} x_i(y_i - a - bx_i) = -2\sum_{i=1}^{n} x_i e_i = 0 \Rightarrow \sum_{i=1}^{n} x_i e_i = 0 \qquad (5)$$

NOTE

(i) Equation (3) implies that the residuals always sum to zero (providing there is an intercept in the model)

(ii) Equation (5) implies that the covariance (and hence correlation) between the residuals and the *x*'s is zero, that is, they are orthogonal.

The two equations (3) and (5) are referred to as the **NORMAL** equations. In Appendix 4 we estimate by OLS a simple two variable regression model in which we show that $\sum_{i=1}^{n} e_i = 0$ and $\sum_{i=1}^{n} x_i e_i = 0$.

Solving equation (3) for *a* we have

$$\sum_{i=1} e_i = \sum_{i=1} y_i - \sum_{i=1} a - \sum_{i=1} bx_i = \sum_{i=1} y_i - na - b\sum_{i=1} x_i = 0,$$

Solving for *a*, we get

$$a = \bar{y} - b\bar{x} \tag{6}$$

Substituting equation (6) in equation (5) we have

$$\Rightarrow \sum_{i=1}^{n} x_i(y_i - (\bar{y} - b\bar{x}) - bx_i) = \sum_{i=1} x_i(y_i - \bar{y}) - b\sum_{i=1} x_i(x_i - \bar{x}) = 0$$

that is,

$$b\sum_{i=1} x_i(x_i - \bar{x}) = \sum_{i=1} x_i(y_i - \bar{y})$$

$$b = \frac{\sum_{i=1} x_i(y_i - \bar{y})}{\sum_{i=1} x_i(x_i - \bar{x})} = \frac{\sum_{i=1} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1} (x_i - \bar{x})^2} \tag{7}$$

Finally we need to estimate σ^2 (the variance of the disturbance term, ε_i). This is estimated as:

$$s^2 = \frac{\sum_{i=1}^{n} e_i^2}{DoF} = \frac{RSS}{DoF} \tag{8}$$

where *DoF* are the degrees of freedom for the residuals. The *DoF* is the number of observations, *n*, less the number of restrictions we have on these residuals, that is, $\sum_{i=1}^{n} e_i = 0$ and $\sum_{i=1}^{n} x_i e_i = 0$ (which must be equal to the number of estimated parameters, *a* and *b* – in this case), i.e. *n*-2.

These **OLS** estimates have a number of desirable properties (known as the Gauss Markov Theorem) in that the estimators are **BLUE**:

(i) **Best** - have the minimum variance, such that $V(b) \leq V(b^*)$, where b^* is any alternative unbiased estimator.

(ii) **Linear** – linear function of the error term.

(iii) **Unbiased** - $E(a) = \alpha$ and $E(b) = \beta$

(iv) **Estimators**.

4.2 Properities of the OLS estimators

From equation (1) we have that $\bar{y} = \alpha + \beta\bar{x} + \bar{\varepsilon}$, in which case:

$$y_i - \bar{y} = \beta(x_i - \bar{x}) + \varepsilon_i - \bar{\varepsilon} \tag{9}$$

substituting equation (9) into equation (7) we have

$$b = \frac{\sum_{i=1}(x_i - \bar{x})\{\beta(x_i - \bar{x}) + (\varepsilon_i - \bar{\varepsilon})\}}{\sum_{i=1}(x_i - \bar{x})^2} = \beta\frac{\sum_{i=1}(x_i - \bar{x})^2}{\sum_{i=1}(x_i - \bar{x})^2} + \frac{\sum_{i=1}(x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon})}{\sum_{i=1}(x_i - \bar{x})^2}$$

$$b = \beta + \frac{\sum_{t}(x_i - \bar{x})\varepsilon_i}{\sum_{i=1}(x_i - \bar{x})^2} - \frac{\bar{\varepsilon}\overbrace{\sum_{i=1}(x_i - \bar{x})}^{0}}{\sum_{i=1}(x_i - \bar{x})^2} = \beta + \frac{\sum_{i=1}(x_i - \bar{x})\varepsilon_i}{\sum_{i=1}(x_i - \bar{x})^2} \tag{10}$$

From (10) we see that *b* as a **linear** function of the error term. Taking equation (10) we can show that our OLS estimator is unbiased:

$$b = \beta + \sum_{i=1}\omega_i\varepsilon_i, \quad \text{where } \omega_i = \frac{(x_i - \bar{x})}{\sum_{i=1}(x_i - \bar{x})^2} \tag{11}$$

Digression:

Now let us look at some of the properties of the variable ω_i .

(i) $\sum_{i=1}\omega_i = \frac{\sum_{i=1}(x_i - \bar{x})}{\sum_{i=1}(x_i - \bar{x})^2} = 0$

(ii) $\sum_{i=1}\omega_i^2 = \frac{\sum_{i=1}(x_i - \bar{x})^2}{\left(\sum_{i=1}(x_i - \bar{x})^2\right)^2} = \frac{1}{\sum_{i=1}(x_i - \bar{x})^2}$

(i) Unbiasedness

Consider first the slope coefficient, *b*,

$$E(b) = E(\beta) + E\left(\sum_{i=1}\omega_i\varepsilon_i\right)$$

as β is a constant, $E(\beta) = \beta$

$$E(b) = \beta + E\left(\sum_{i=1}\omega_i\varepsilon_i\right) = \beta + E(\omega_1\varepsilon_1 + \omega_2\varepsilon_2 + ... + \omega_n\varepsilon_n)$$

Now as ω_i is made up of x_i, which is by definition non-stochastic[2] (or for each *i* can be treated as a constant), consequently,

$$E(b) = \beta + \omega_1 E(\varepsilon_1) + \omega_2 E(\varepsilon_2) + ... + \omega_n E(\varepsilon_n) = \beta$$

as $E(\varepsilon_i) = 0 \; \forall \, i$.

Therefore the slope coefficient is **unbiased** estimator, that is, $E(b) = \beta$, that is, in repeated regressions of *y* on *x* the average of the coefficient estimates of *b*, will be equal to the true coefficient.

Now looking at the intercept, *a*,

$$E(a) = E(\bar{y}) - E(b\bar{x}) = E(\alpha + \beta\bar{x} + \bar{\varepsilon}) - \bar{x}E(b) = \alpha + \beta\bar{x} + E(\bar{\varepsilon}) - \beta\bar{x} = \alpha$$

as $E(\bar{\varepsilon}) = 0$.

Therefore the intercept is an **unbiased** estimator, that is, $E(a) = \alpha$ (see Figures 9 and 10).

(ii) Variances

Firstly looking at the variance of the slope coefficient, *b*,

$$V(b) = E[(b - E(b))^2] = E[(b - \beta)^2] = E\left[\left(\sum_{i=1} \omega_i \varepsilon_i \right)^2 \right] \qquad (12)$$

$$\begin{aligned} V(b) = E[(\omega_1\varepsilon_1 + \omega_2\varepsilon_2 + ... + \omega_n\varepsilon_n)^2] = E(&\omega_1^2\varepsilon_1^2 + \omega_2^2\varepsilon_2^2 + ... + \omega_n^2\varepsilon_n^2 + \\ &2\omega_1\omega_2\varepsilon_1\varepsilon_2 + 2\omega_1\omega_3\varepsilon_1\varepsilon_3 + ... + 2\omega_1\omega_n\varepsilon_1\varepsilon_n + \\ &2\omega_2\omega_3\varepsilon_2\varepsilon_3 + ... + 2\omega_2\omega_n\varepsilon_2\varepsilon_n + \\ &+ .+.2\omega_{n-1}\omega_n\varepsilon_{n-1}\varepsilon_n) \end{aligned}$$

As the ω_i terms are exogeneous these will be brought outside the expectation operator, that is

$$\begin{aligned} V(b) = &\omega_1^2 E(\varepsilon_1^2) + \omega_2^2 E(\varepsilon_2^2) + \ldots + \omega_n^2 E(\varepsilon_n^2) + 2\omega_1\omega_2 E(\varepsilon_1\varepsilon_2) + 2\omega_1\omega_3 E(\varepsilon_1\varepsilon_3) + \ldots + 2\omega_1\omega_n E(\varepsilon_1\varepsilon_n) + \\ &2\omega_2\omega_3 E(\varepsilon_2\varepsilon_3) + \ldots + 2\omega_2\omega_n E(\varepsilon_2\varepsilon_n) + \ldots + 2\omega_{n-1}\omega_n E(\varepsilon_{n-1}\varepsilon_n) \end{aligned}$$

As $V(\varepsilon_i) = E(\varepsilon_i^2) = \sigma^2$ and $\text{cov}(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i\varepsilon_j) = 0, i \neq j$, all of the cross-product terms are zero and we have

$$V(b) = \omega_1^2\sigma^2 + \omega_2^2\sigma^2 + \ldots + \omega_n^2\sigma^2 = \sigma^2 \sum_{i=1} \omega_i^2 = \frac{\sigma^2}{\sum_{i=1} (x_i - \bar{x})^2} \qquad (13)$$

[2] If you do not have non-stochastic x_i you require independence between x_i and ε_i.

As *b* is a linear function of the error term, ε, see equation (10), and as ε is normally distributed (by assumption), then *b* (*a*) will follow a normal distribution, $b \sim N\left(\beta, \frac{\sigma^2}{\sum_{i=1}^{n}(x_t - \bar{x})^2}\right)$. Additionally as *s* is the sum of squared normally distributed residuals, then $\frac{(n-2)s^2}{\sigma^2} \sim \chi^2_{n-2}$ - as can be seen in Figure 10 the distribution of *b* is Normal.

4.3 Hypothesis testing

All hypothesis testing follows a 5-step procedure:

1. $H_0 : \beta = \beta_0$
2. **2.** $H_0 : \beta \neq \beta_0$
3. Choose some appropriate significance level of α, and find the corresponding value from the t-distribution, denoted $-t_{\alpha/2,DoF}$ and $t_{\alpha/2,DoF}$, where *DoF* is the degrees of freedom of the model, that is, the number of observation, *n*, minus the number of restrictions on the residuals, 2, (or number of estimated parameters in the model, *a* and *b*).
4. $t = \frac{b-\beta_0}{s_b} \sim t_{\alpha/2,DoF}$, where $s_b = \sqrt{\frac{s^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$ = standard error of *b* (in replacing σ^2 by s^2 in the standard error of *b*, we are essentially scaling a N(0,1) by a $\sqrt{\chi^2_{DoF}/DoF}$ and this yields a t_{DoF} distribution.
5. If t is either less than $-t_{\alpha/2,DoF}$, or greater than $t_{\alpha/2,DoF}$, then we have observed an event which occurs with a probability of less than α and should therefore reject *H0*. The decision rule is: Reject *H0* if $t = \left|\frac{b-\beta_0}{s_b}\right| > t_{\alpha/2,DoF}$; Do not reject *H0* if $t = \left|\frac{b-\beta_0}{s_b}\right| < t_{\alpha/2,DoF}$.

4.4 Measure of goodness of fit

Having used the Ordinary Least Squares (OLS) methods to estimate the sample regression line, we now ask whether the estimated model fits the data well. As OLS minimises the RSS or $\sum_{i=1}^{n} e_i^2$, this means that the estimates, *a* and *b* produce a smaller RSS than an alternative pair of estimators. But is the model "good" at explaining movements in y_i?

Our OLS regression has:

$$\underbrace{y_i}_{\text{Actual values}} = \underbrace{a + bx_i}_{\hat{y}_i} + \underbrace{e_i}_{\text{Residuals}}$$

taking averages

$$\bar{y} = \bar{\hat{y}} + \underbrace{\bar{e}}_{0} \Rightarrow \bar{y} = \bar{\hat{y}}$$

and subtracting these two equations we have:

$$y_i - \bar{y} = (\hat{y}_i - \bar{\hat{y}}) + e_i$$

squaring both sides we get:

$$(y_i - \bar{y})^2 = (\hat{y}_i - \bar{\hat{y}})^2 + e_i^2 + 2(\hat{y}_i - \bar{\hat{y}})e_i$$

Now taking sums we have

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{TSS} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}_{ESS} + \underbrace{\sum_{i=1}^{n} e_i^2}_{RSS} + 2\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})e_i$$

But the last term is zero, as from equation (5) we have $\sum_{i=1}^{n} x_i e_i = 0$ in which case:

$\sum_{i=1}^{n} \hat{y}_i e_i$. Consequently, we have: *TSS=ESS+RSS*

where TSS = Total sum of squares (the total amount of variation in the dependent variable, *y*), ESS = Explained sum of squares (the amount of variation the model explained) and RSS= Residual sum of squares (amount of variation the model did NOT explain). Our measure of goodness of fit, is the coefficient of determination or R-square and this is calculated as:

$$R^2 = 1 - \frac{RSS}{TSS} = \frac{ESS}{TSS}$$

and so it can be interpreted as the proportion of the total variation in the dependent variable that the model (x_t) can explain.

Use of R^2

(a) Goodness of fit measure for simple regression model

(b) It can be used to compare (choose) between different linear models (as long as the dependent variable is identical in all models).

(c) It is NOT correct to say that a model with a "high" R^2 is a good model.

(see Appendix 3 to see a Stata output, which is labelled and Appendix 4 for an example calculation of R^2.

4.5 Model adequacy

A better way to detect model adequacy than the use of R^2 is to ensure that the residuals, e_t, are consistent with the assumptions that you make about the disturbance term, ε_t. That is, we would require that:

1. The residuals are serially uncorrelated
2. The residuals have a constant error variance
3. The residuals are normally distributed

The assumption that the errors average to zero, is not a test as by construction the residuals sum to zero, see equation (5). Testing for the exogeneity of x_t is non-trivial and will not be considered here. However, one can look for non-constant variance in the residuals and serial correlation in the residuals (and we will look at these issues later in the module).

5. Interpreting coefficients

OLS is appropriate for all model that are linear in parameters and each of the following models is linear in parameters (although the model is often non-linear in the variables *Y* and *x*):

1. $y_i = \alpha + \beta x_i + \varepsilon_i$ $\qquad \dfrac{\partial y_i}{\partial x_i} = \beta = \dfrac{\text{Expected change in y}_i}{1\uparrow \text{ in } x_i}$

2. $y_i = \alpha + \dfrac{\beta}{x_i} + \varepsilon_i$ $\qquad \dfrac{\partial y_i}{\partial (1/x_i)} = \beta = \dfrac{\text{Expected change in y}_i}{1\uparrow \text{ in } (1/x_i)}$ (meaningless?)

$$\frac{\partial y_i}{\partial x_i} = \frac{-\beta}{x_i^2} = \frac{\text{Expected change in y}_i}{1\uparrow \text{ in } x_i}$$

3. $y_i = \alpha + \beta \ln(x_i) + \varepsilon_i$

$E(y_i) = \alpha + \beta \ln(x_i)$ $\qquad E(y_i^+) = \alpha + \beta \ln(x_i^+)$ in which case:

$E(y_i^+) - E(y_i) = \beta[\ln(x_i^+) - \ln(x_i)] = \beta \ln(x_i^+ / x_i)$ and β is the change in the expected value of y_i when $\ln(x_i^+ / x_i) = 1$. So when does $\ln(x_i^+ / x_i) = 1$?

Digression

Define the growth rate of the variable *x* as *g*,

$$g = \frac{(x^+ - x)}{x} = \frac{x^+}{x} - 1 \Rightarrow 1 + g = \frac{x^+}{x}$$

Now taking nature logs of both sides we get:

$\ln(1+g) = \ln(x^+ / x) = \ln(x^+) - \ln(x) \equiv \Delta \ln(x) \approx g$ (if *g* is SMALL) as:

g	ln(1+*g*)	*g*	ln(1+*g*)	*g*	ln(1+*g*)
0.01	0.0099	0.10	0.0953	-0.03	-0.0305
0.02	0.0198	0.20	0.1823	-0.06	-0.0619
0.03	0.0296	-0.01	-0.0101	-0.10	-0.1054
0.06	0.0583	-0.02	-0.0202	-0.20	-0.2231

$\ln(x_i^+ / x_i) = 1 \Rightarrow \exp(\ln(x_i^+ / x_i)) = \exp(1) \Rightarrow x_i^+ / x_i \Rightarrow 1 + g = 2.718$ or $g = 1.718$

(a) If we consider g=0.01 then $\ln(x_i^+ / x_i) = \ln(1.01) \approx 0.01$ then

$$0.01\beta = \frac{\text{Expected change in y}_i}{1\% \uparrow \text{ in } x_i}$$

(b) If we consider g=0.10 then $\ln(x_i^+ / x_i) = \ln(1.10) \approx 0.095$ then

$$0.095\beta = \frac{\text{Expected change in y}_i}{10\% \uparrow \text{ in } x_i}$$

(c) If we consider g=1.0 then $\ln(x_i^+ / x_i) = \ln(2) \approx 0.693$ then

$$0.693\beta = \frac{\text{Expected change in y}_i}{100\% \uparrow \text{ in } x_i}$$

4. $\ln(y_i) = \alpha + \beta x_i + \varepsilon_i$

$E(\ln y_i) = \alpha + \beta x_i$ $\quad$ $E(\ln y_i^+) = \alpha + \beta x_i^+$ in which case:

$E(\ln y_i^+) - E(\ln y_i) = \beta(x_i^+ - x_i)$ and β is the change in the expected value of $\ln y_i$ when $(x_i^+ - x_i) = 1$. So $100 * [\exp(\beta) - 1] = \dfrac{\text{Expected \% change in y}_i}{1 \uparrow \text{ in } x_i}$

Note: If β is small $100 * \beta = \dfrac{\text{Expected \% change in y}_i}{1 \uparrow \text{ in } x_i}$.

5. $\ln(Y_i) = \alpha + \beta \ln(x_i) + \varepsilon_i$

$E(\ln y_i) = \alpha + \beta \ln(x_i)$ $\quad$ $E(\ln y_i^+) = \alpha + \beta \ln(x_i^+)$ in which case:

$E(\ln y_i^+) - E(\ln y_i) = \beta(\ln(x_i^+) - \ln(x_i))$ and β is the change in the expected value of $\ln y_i$ when $(\ln x_i^+ - \ln x_i) = 1$.

So $0.01\beta = \dfrac{\text{Expected proportionate change in y}_i}{1\% \uparrow \text{ in } x_i}$ or $\beta = \dfrac{\text{\% change in y}_i}{1\% \uparrow \text{ in } x_i}$.

Figure 1: Positive correlation between y and x

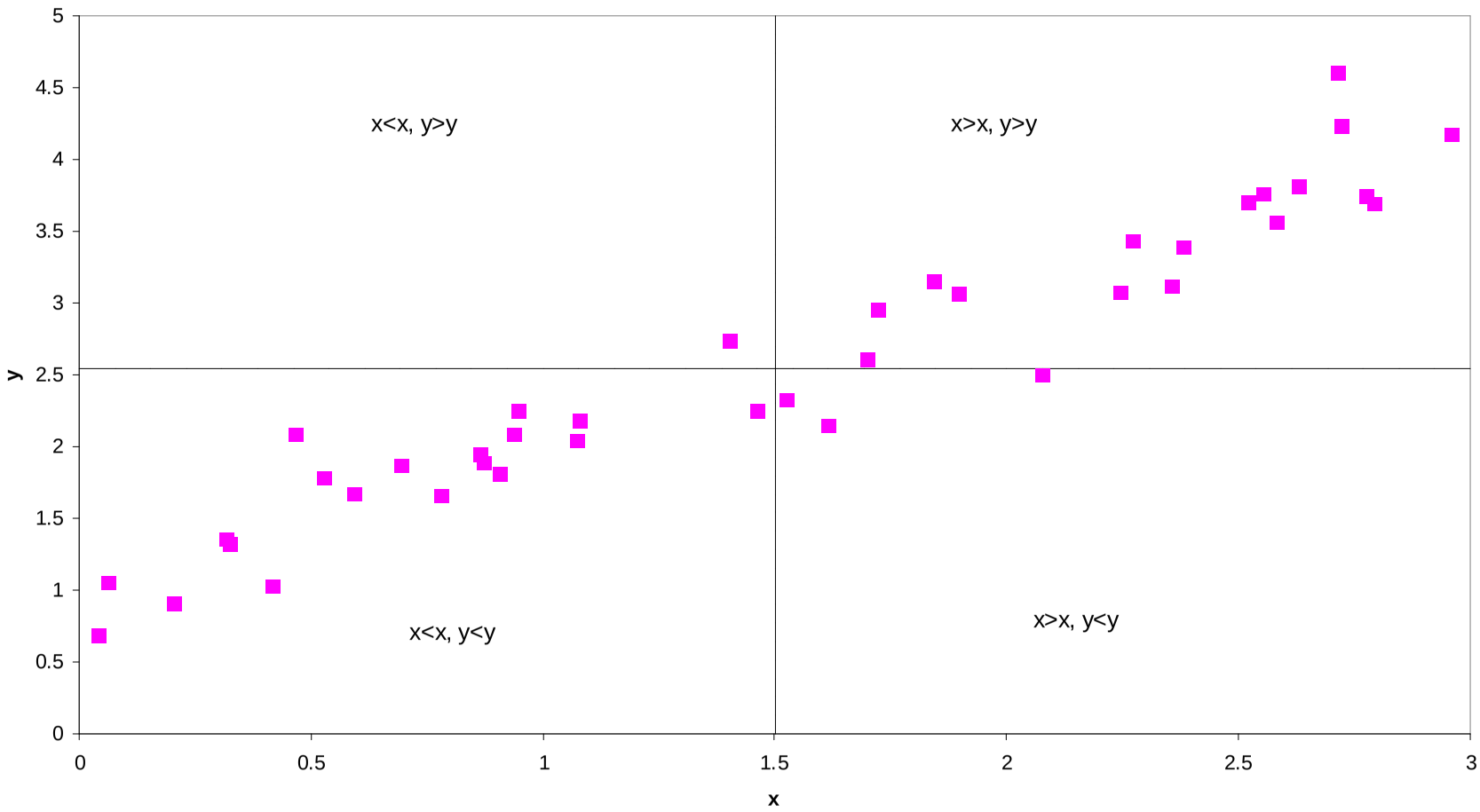


Figure 2: Negative correlation between y and x

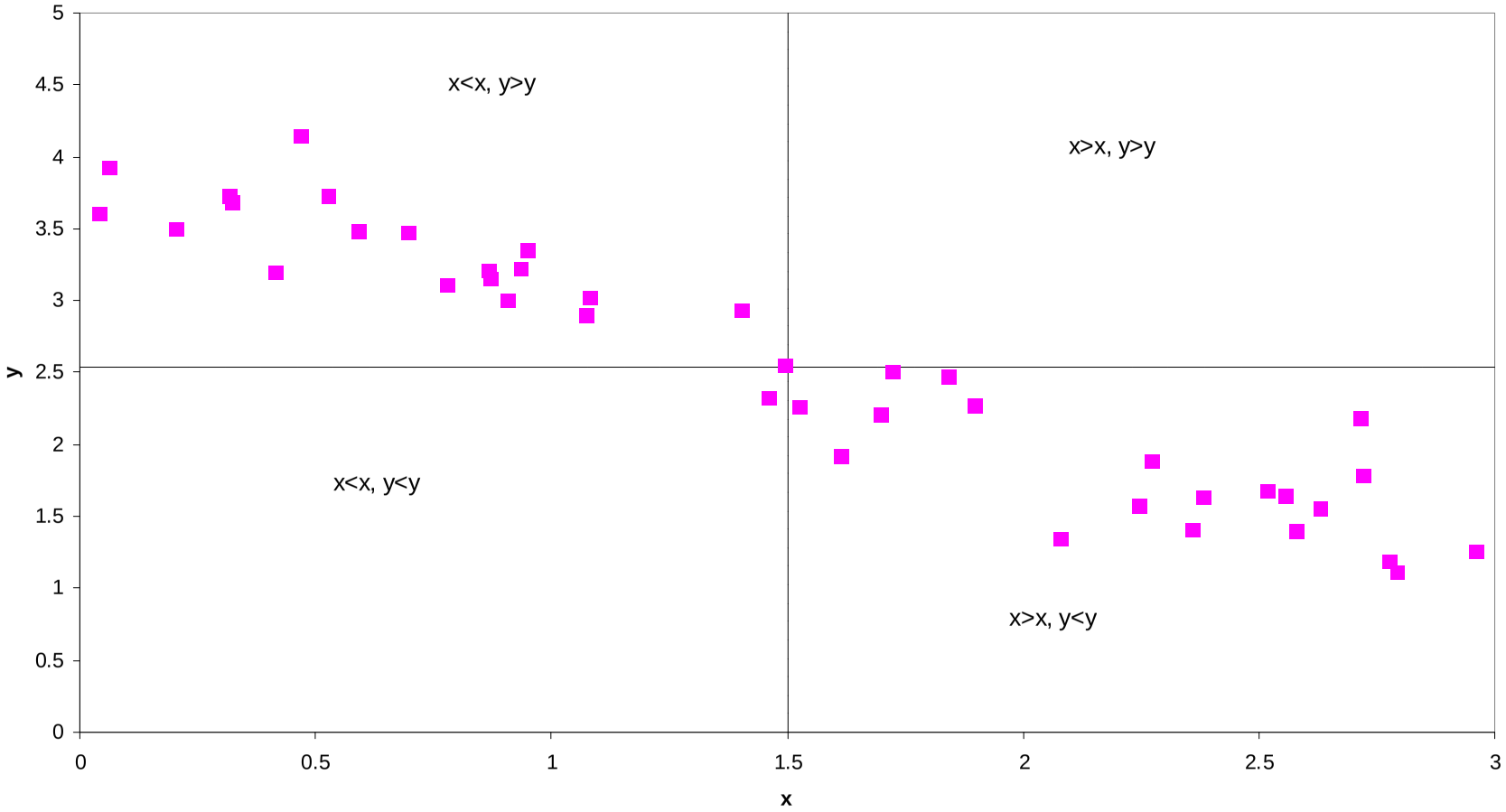


Figure 3: Zero correlation between y and x

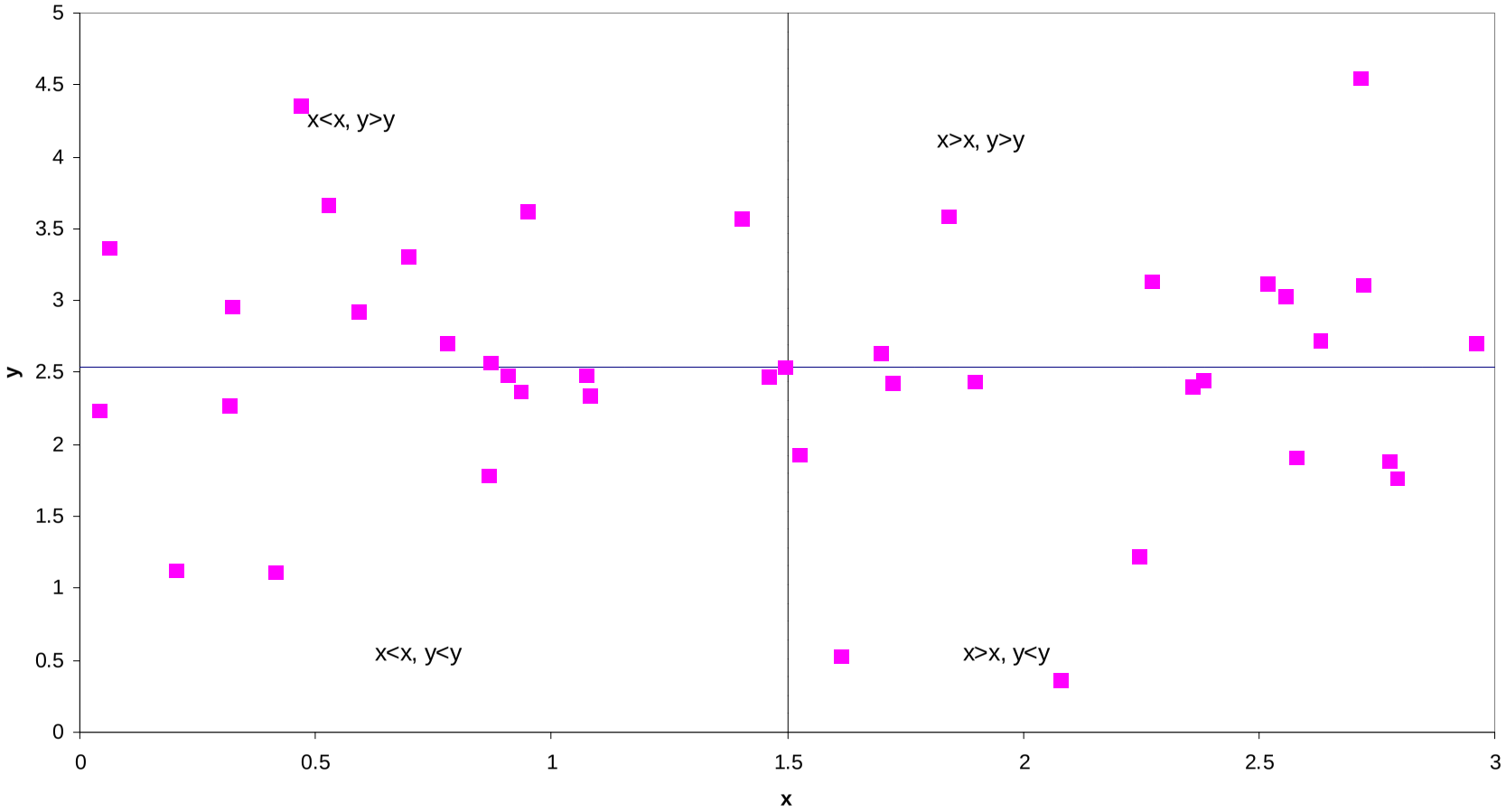


Figure 4: Zero correlation between y and x

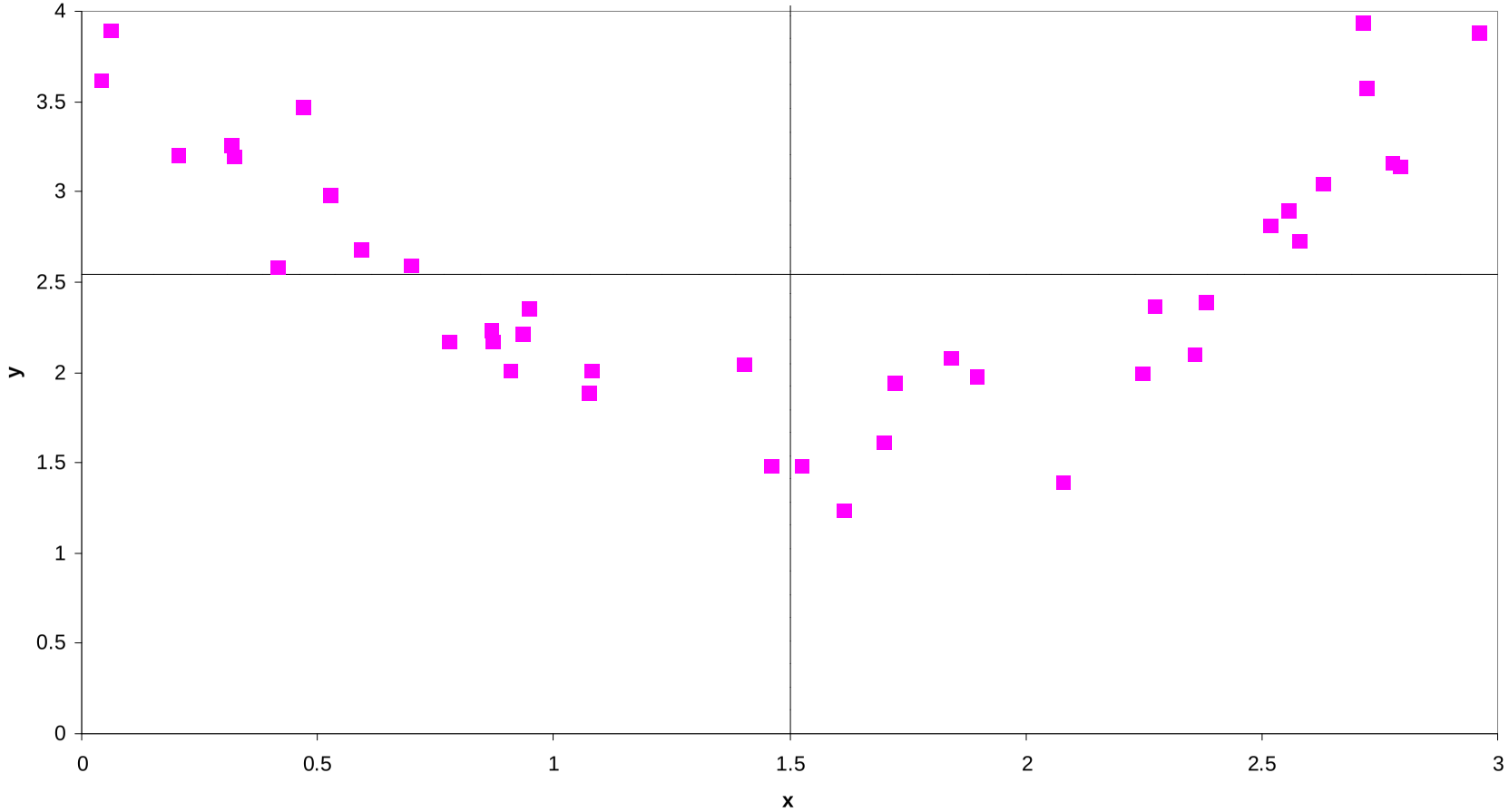


Figure 5: The two variable regression model

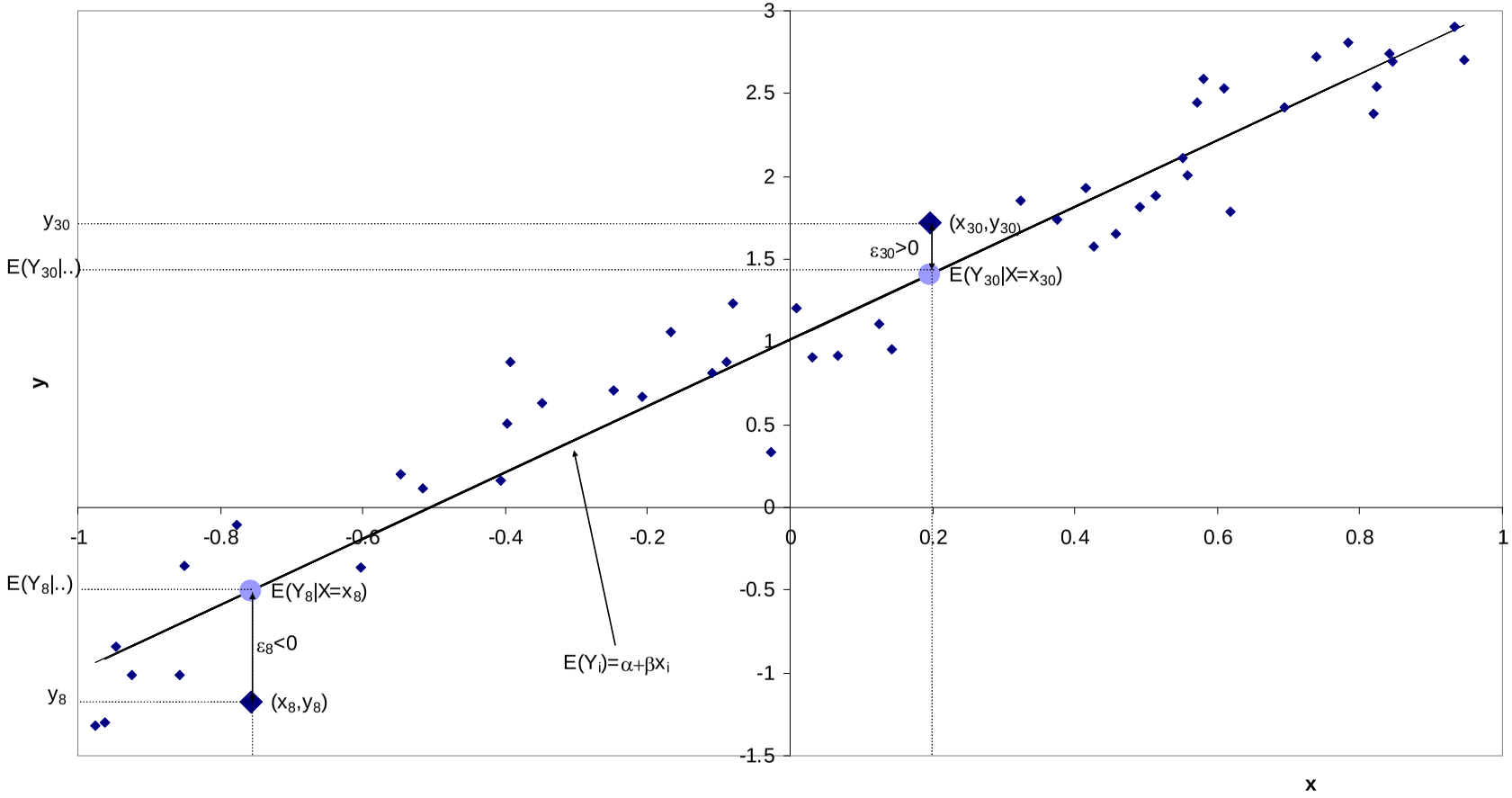


Figure 6: Disturbances which satisfy the CLRM assumptions

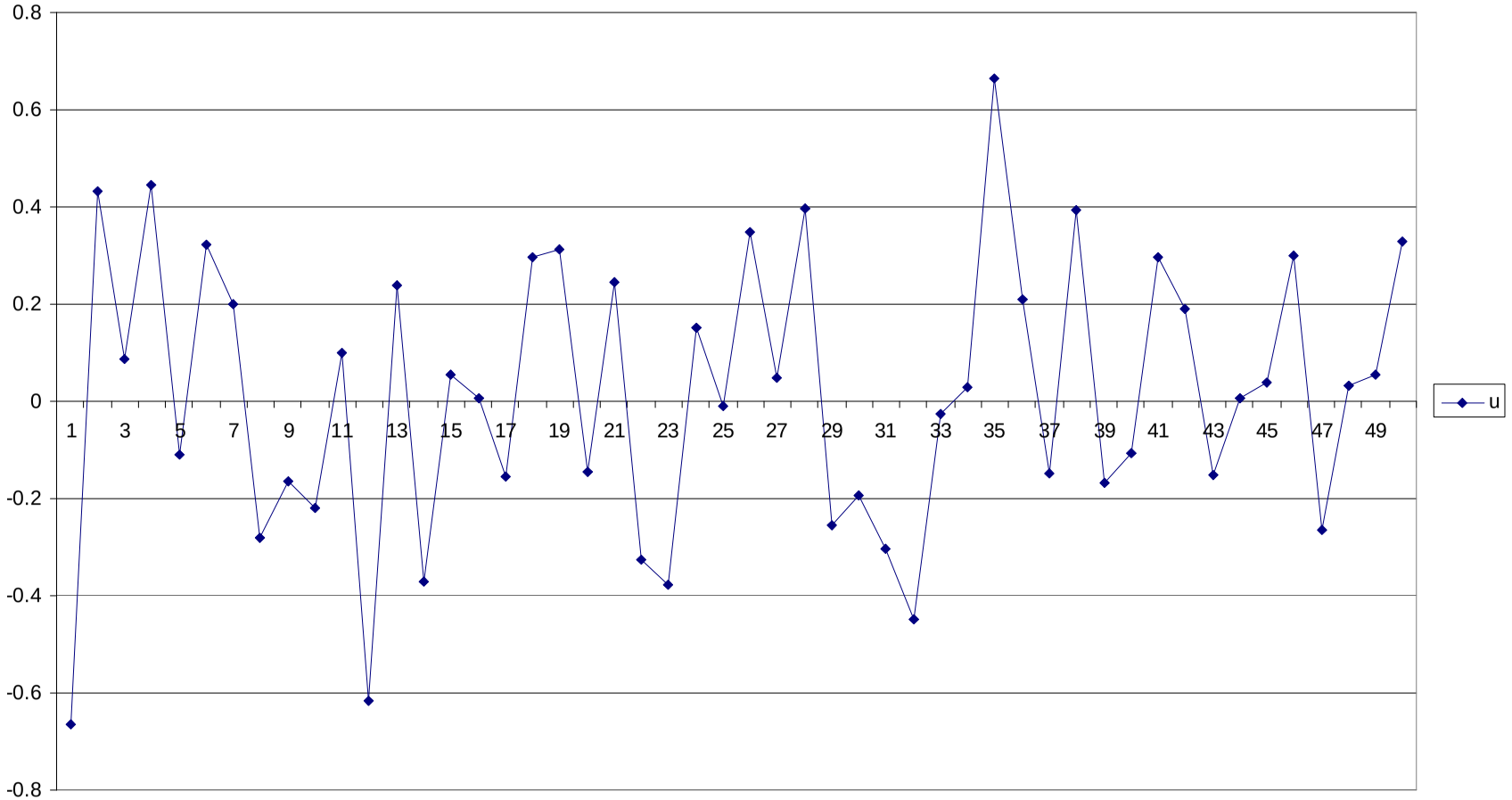


Figure 7: The estimated two variable regression model

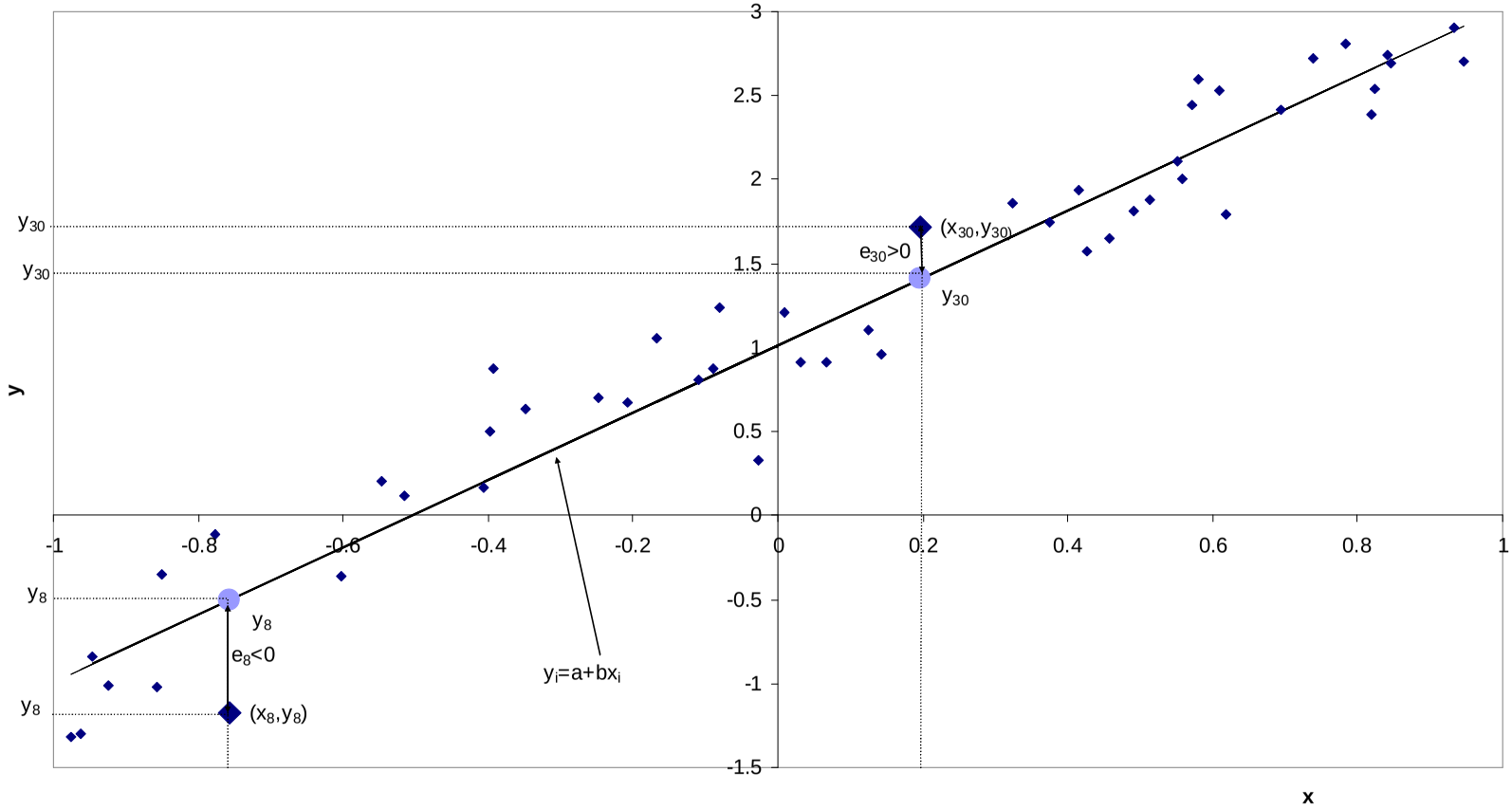
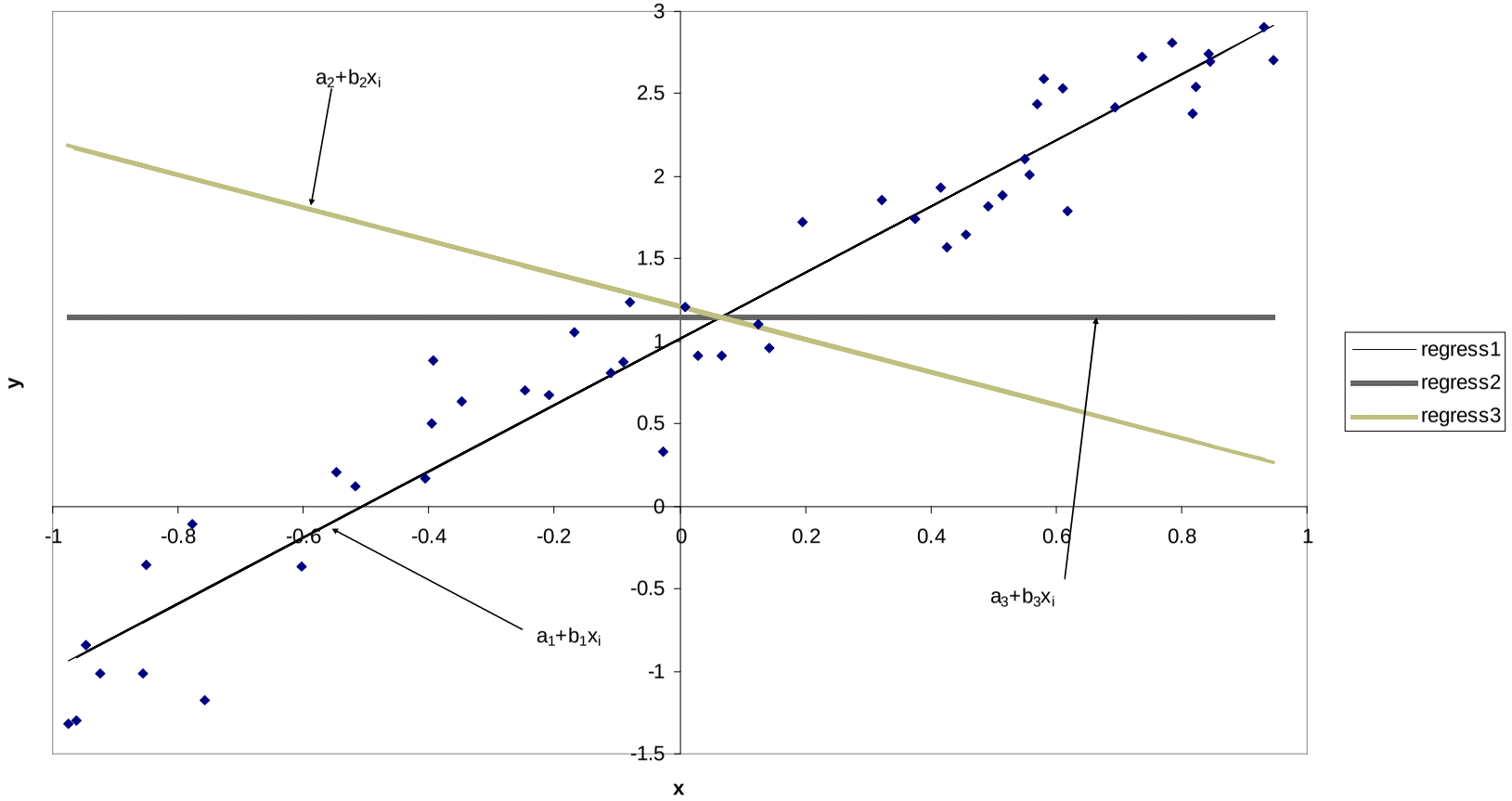


Figure 8: Two-variable regression - line of best fit



Appendix 1: Stata output for two-variable regression

regress lhourpay edage if(edage>=0 & edage<=95)