Warsaw Summer School 2023, OSU Study Abroad Program

Regression (multivariate)

Bivariate Correlation and Regression

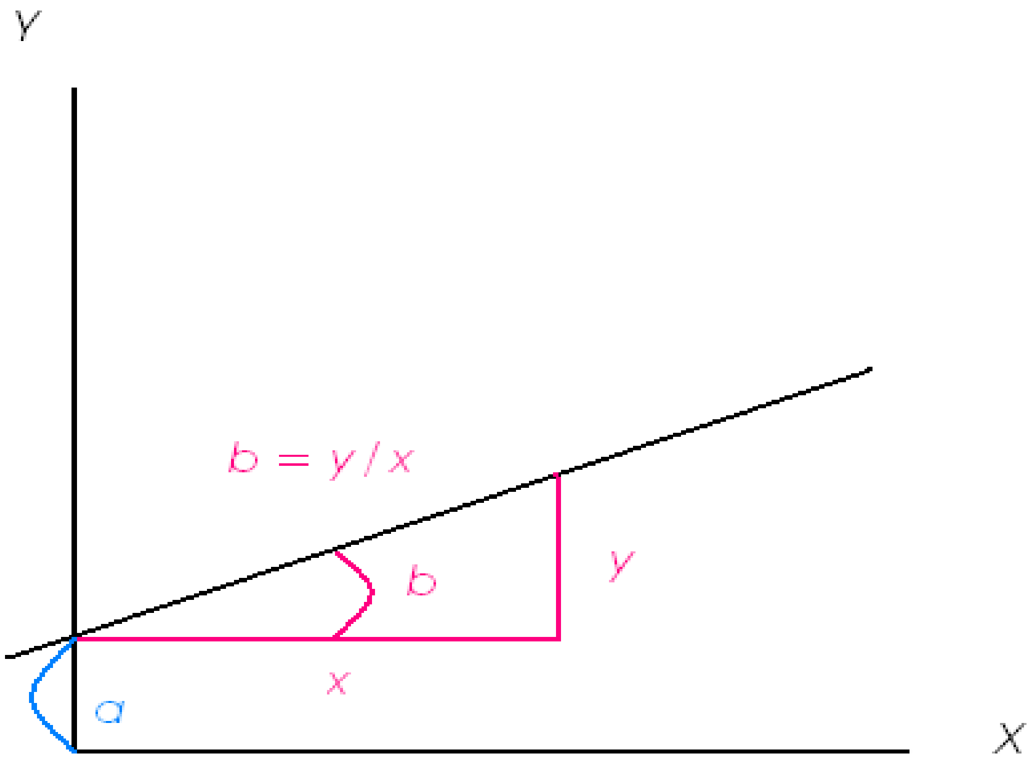
***Assumptions**:*

- Random sampling
- Both variables = continuous (or can be treated as such)
- Linear relationship between the two variables;
- Normally distributed characteristics of X and Y <u>in the population</u>

Linear Relationship

The line = mathematical function; can be expressed through the formula Y = a + b*X, where Y & X are our variables.

Y, the dependent variable, is expressed as a linear function of the independent (explanatory) variable X.

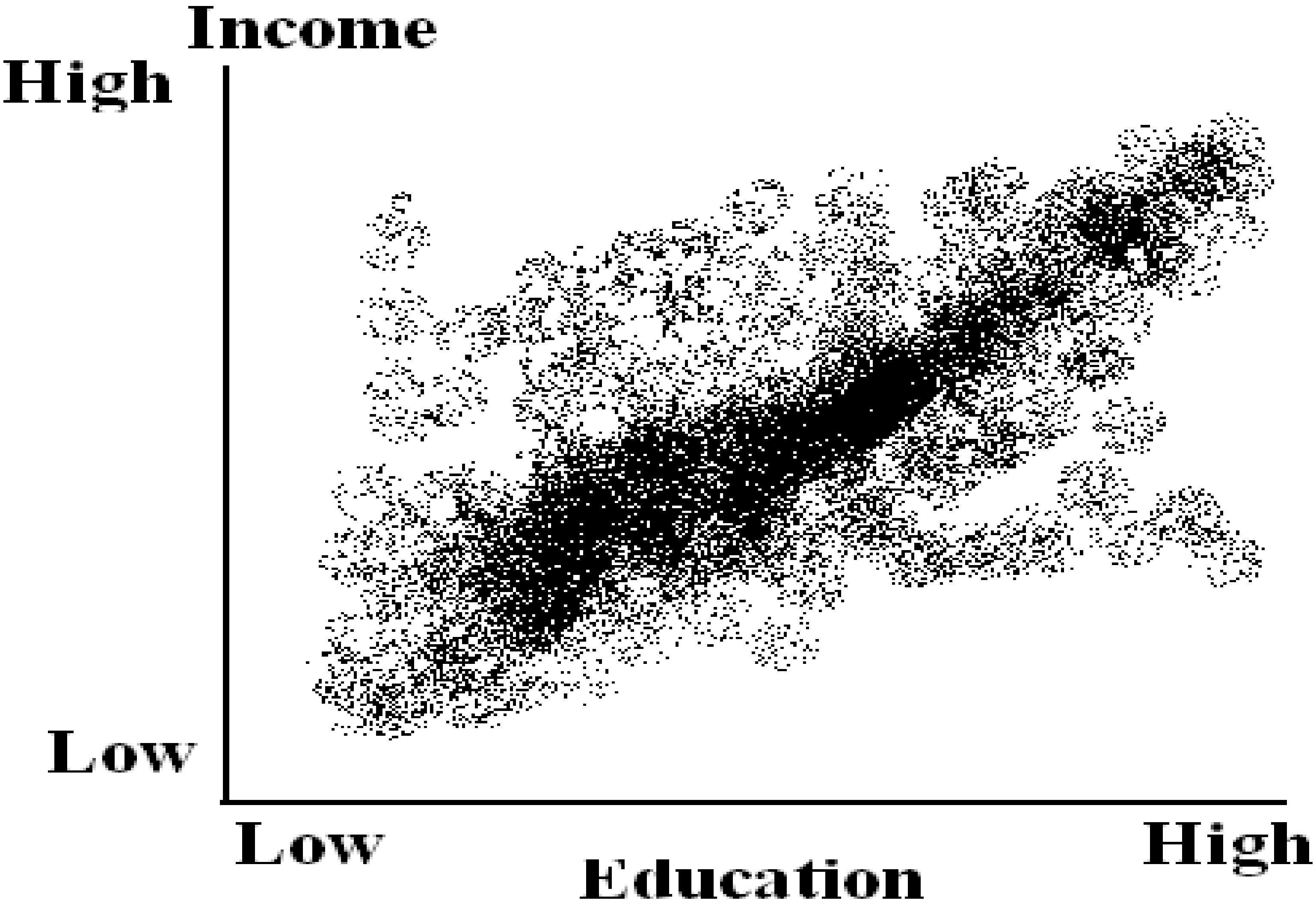


Model vs Reality

The function $Y = a + b*X$ is a model

In reality we do not have one line

In reality data are scattered



Regression

Finds the best fitting straight line between X & Y:

the line that goes through the means of X and of Y, and

minimizes the sum of squared errors (distances) btw. the data points (observed values of Y) and the line (predicted values of Y)

□ least square regression

Z-scores

.

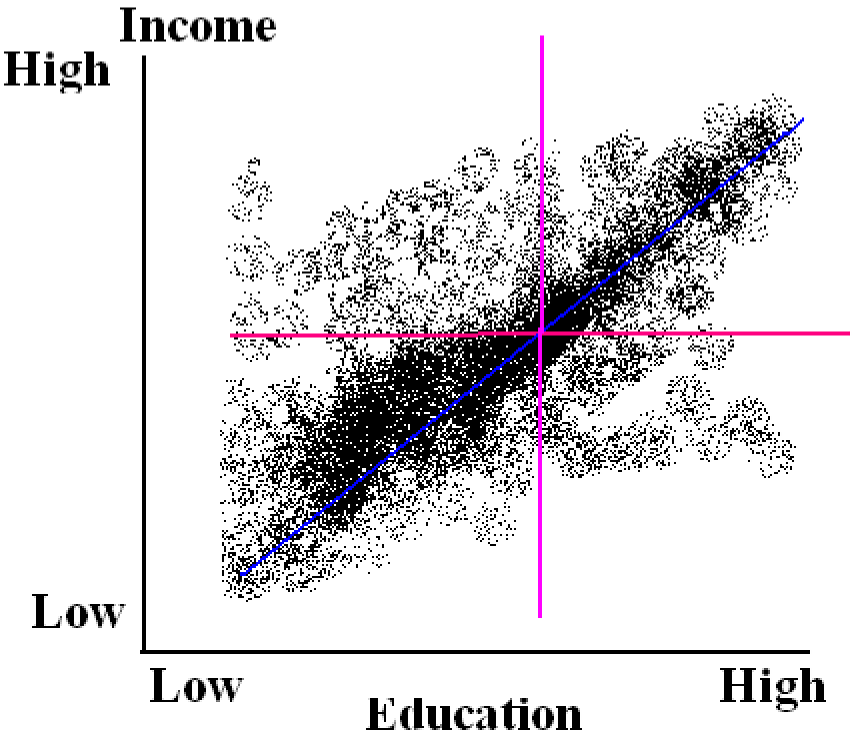
$$\mathbf{Z\ score\ (X_i) = X_i - Mean\ for\ X\ /\ \sqrt{s^2}}$$

The numerical value of the z-score specifies the distance from the mean expressed in terms of the proportion of the standard deviation.

For z-score distribution: mean = 0; st. dev = 1.

Hence, in regression, when “raw” scores □ z-scores:

For z scores



Bivariate Regression

= used to predict the dependent variable (DV) from the independent variable (IV);

Observed distribution (what the data show):

Y = a + b*X **+ e**.

Y = **observed values of DV**

a = the intercept (the value of Y when X = 0)

b = the regression coefficient (the slope), indicating the amount of change in Y given a unit change in X

X = the independent variable

e = error term

..

(

$$\hat{Y} = a + b * X$$

(**Prediction model**)

$\hat{Y} =$ **predicted values** for the dependent variable, Y

a = the intercept (the value of Y when X = 0)

b = the regression coefficient (the slope), indicating the amount of change in Y given a unit change in X

X = the independent variable

Residuals

Prediction model gives the points on the line.

But, in reality, not all points fall on the line.

Predicted minus *Observed* values of Y at each value of X = errors of prediction (i.e. error terms, **residuals**!!!)

Y = a + bX **+ e**; □ **e** = Y- (a +b*X) □ $e = Y - \hat{Y}$

e > 0: under-prediction (Y > Y-hat)

e < 0: over-prediction (Y < Y-hat)

Ex: We checked data; Jill was paid $7.50/week, not 16. What’s the prediction error?

Method of Least Squares

The regression line minimizes the sum of error terms:

$$SSE = \sum (Y - \hat{Y})^2$$

The methods of least square provides the prediction equation
$\hat{Y} = a + bX$ having the minimal value of SSE.

a, b = least square estimates

Goal:
arrive at a set of regression coefficients (bs) for the IVs that bring
$\hat{Y}$s as close as possible to Ys values

b and onstant a

▪

$$\hat{Y} = a + b * X$$

$$b = \frac{\Sigma\left(X - \overline{X}\right)\left(Y - \overline{Y}\right)}{\Sigma\left(X - \overline{X}\right)^2}$$

▪ $$a = \hat{Y} - b * X$$

Ex: if **a** = 4, X = years of education Y = dollars earned/week, interpret the intercept of the mode

If someone had no education (if X=0), then we would expect their weekly wage to be $4

Covariance

In regression analysis we ask: to what extent could we predict Y knowing our variable X?

Prediction means that values X and Y go together or co-vary.

Covariance is sum of products, or SP,

$$SP = \Sigma\left(X - \overline{X}\right)\left(Y - \overline{Y}\right)$$

Sums of squares for X:

$$SSx = \Sigma\left(X - \overline{X}\right)^2$$

Note that in the regression equation of Y on X, $\hat{Y} = a + b * X$

b = SP / SSx

$$b = \frac{\Sigma\left(X - \overline{X}\right)\left(Y - \overline{Y}\right)}{\Sigma\left(X - \overline{X}\right)^2}$$

Interpretation of b (unstandardized coefficients)

$b > 0$, positive relationship	(X has a positive effect on Y)
$b < 0$, negative relationship	(X has a negative effect on Y)
$b = 0$, no relationship	(X has no effect on Y)

<u>Generic</u>: a one unit (original measurement of X) increase in X produces a ***b*** unit (original measurement of Y) increase (if $b>0$) or decrease (if $b<0$) in Y

<u>Ex:</u> If $b = 1.5$, X =years of education, Y =dollars earned/week, interpret the effect of education on weekly wages.

For a one year increase in X, we would expect weekly wages to go up by $1.50

Regression for Z-scores of X and Y

Transform all the responses into Z-scores

Calculate the coefficients to create the line.

For Z-scores in the linear relationship of the type: $\hat{Y} = a + b*X$, we have:

Predicted $Z_y = \beta * Z_x$ because **a = 0**

$$a = \overline{Y} - b\overline{X}$$ but $$Z\overline{Y} = 0, and \quad Z\overline{X} = 0$$

Why bother with standardized (beta) coefficients?

Enable comparing relative magnitude of IVs effects (in multivariate regression):

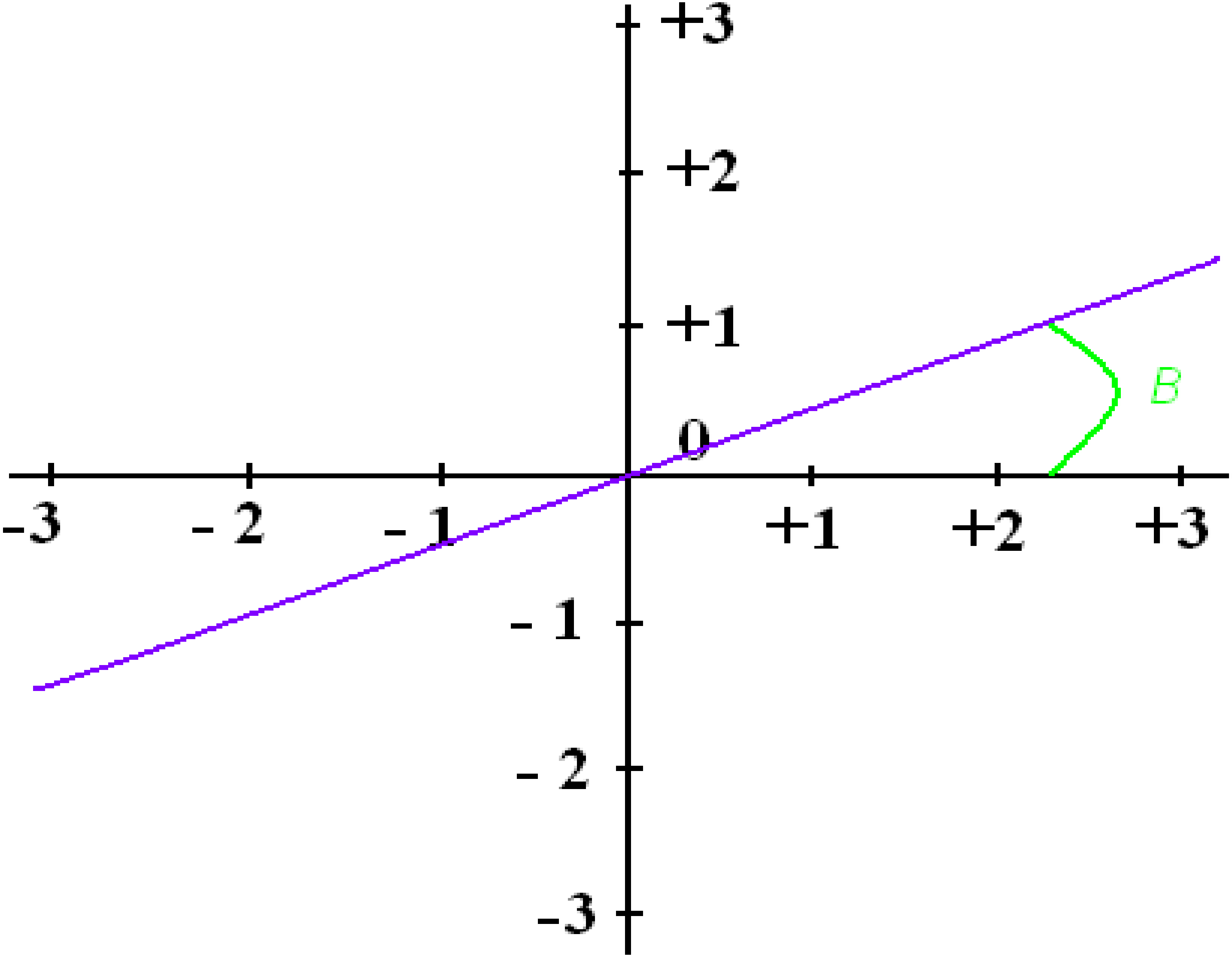
Beta coefficients are expressed in units of standard deviations

For bivariate regressions:

β (beta) $= b = r$,

where r = Pearson’s correl. Coefficient

□ Values of beta coeff. will fall within +/-1 range (same as r)


+3
+2
+1
0
B
-3
- 2
- 1
+1
+2
+3
- 1
- 2
-3

Interpretation of Beta (standardized coeff)

Generic:

A one standard deviation increase in X produces a ***Beta*** standard deviation increase (if beta>0) or decrease (if beta <0) in Y

Ex: If β = 0.2 for the effect of education on weekly wages

For a 1 standard deviation increase in education (X), we expect a 0.2 standard deviation (i.e. about a fifth of a standard deviation) increase in weekly wages.

Hypothesis Testing

1) ANOVA and regression (F statistic for whole model)

2) T-tests for the effect of each of the independent variables (H0: b = 0)

$$Y = a + b_1X_1 + b_2X_2, \ldots, b_nX_n$$

$$Y_i = a + b_1X_{i1} + b_2X_{i2}, \ldots , b_nX_{in} + e_i$$

$_i$

• •

MultivaRIATE