Bivariate Regression - Part II

Usually I present concepts and formulas first, and then work through examples. For variety, I will present the example first, and then give the rationale and procedures for working through it.

Data are collected from 20 individuals on their years of education (X) and annual income in thousands of dollars (Y). The data are as follows:

X	Y	XY	X^2	Y^2
2	5.0	10.0	4	25.00
4	9.7	38.8	16	94.09
8	28.4	227.2	64	806.56
8	8.8	70.4	64	77.44
8	21.0	168.0	64	441.00
10	26.6	266.0	100	707.56
12	25.4	304.8	144	645.16
12	23.1	277.2	144	533.61
12	22.5	270.0	144	506.25
12	19.5	234.0	144	380.25
12	21.7	260.4	144	470.89
13	24.8	322.4	169	615.04
14	30.1	421.4	196	906.01
14	24.8	347.2	196	615.04
15	28.5	427.5	225	812.25
15	26.0	390.0	225	676.00
16	38.9	622.4	256	1,513.21
16	22.1	353.6	256	488.41
17	33.1	562.7	289	1,095.61
21	48.3	1,014.3	441	2,332.89
$T_X = 241$	$T_Y = 488.3$	$T_{XY} = 6,588.3$	$T_{X^2} = 3,285$	$T_{Y^2} = 13,742.27$

Here is an SPSS analysis of the above:
Control cards:

```
DATA LIST FREE / Educ Income.
BEGIN DATA.
    2         5.0
    4         9.7
    8        28.4
    8         8.8
    8        21.0
   10        26.6
   12        25.4
   12        23.1
   12        22.5
   12        19.5
   12        21.7
   13        24.8
   14        30.1
   14        24.8
   15        28.5
   15        26.0
   16        38.9
   16        22.1
   17        33.1
   21        48.3
END DATA.
REGRESSION /DESCRIPTIVES ALL /STAT DEF CI
           /DEPENDENT INCOME /METHOD ENTER EDUC /
           /SCATTERPLOT (INCOME EDUC).
```

Selected output:

Regression

Descriptive Statistics

	Mean	Std. Deviation	Variance	N
INCOME	24.4150	9.78835	95.81187	20
EDUC	12.0500	4.47772	20.05000	20

Correlations

		INCOME	EDUC
Pearson Correlation	INCOME	1.000	.846
	EDUC	.846	1.000
Covariance	INCOME	95.812	37.068
	EDUC	37.068	20.050
Sig. (1-tailed)	INCOME	.	.000
	EDUC	.000	.
Sum of Squares and Cross-products	INCOME	1820.425	704.285
	EDUC	704.285	380.950
N	INCOME	20	20
	EDUC	20	20

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.846[a]	.715	.699	5.36642

a. Predictors: (Constant), EDUC

b. Dependent Variable: INCOME

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1302.054	1	1302.054	45.213	.000[a]
	Residual	518.372	18	28.798		
	Total	1820.425	19			

a. Predictors: (Constant), EDUC

b. Dependent Variable: INCOME

Coefficients[a]

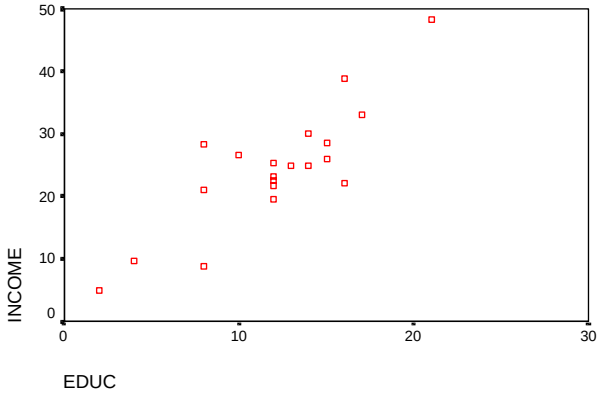
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	2.137	3.524		.607	.552	-5.266	9.541
	EDUC	1.849	.275	.846	6.724	.000	1.271	2.426

a. Dependent Variable: INCOME

Charts

Scatterplot

Dependent Variable: INCOME



Now we will consider what these numbers mean, and where they came from.

a. Determine $\hat{\mu}_x$, SST_X, s^2_X, s_X, $\hat{\mu}_y$, SST_Y, s^2_Y, SP_{XY}, s_{XY}.

Comment. These are generally familiar terms. The relatively new things are the subscripts on SST (which reflect the fact that there are 2 variables; if the subscript is missing, SST_y is assumed), SP_{xy} (which stands for the sum of products) and the sample covariance s_{xy}, which, as we will see, helps us to determine the nature of the relationship between X and Y. The SST and SP terms are also known as Cross-Products, and the notation XP_{xx}, XP_{yy}, and XP_{xy} is used.

The means and standard deviations are substantively interesting in and of themselves; the main value of the other terms is that they help us to compute other quantities that are of interest.

The formulas we need are

$$\hat{\mu}_X = \overline{X} = \frac{\sum X_i}{N},$$

$$SST_x = \sum(x_i - \overline{x})^2 = = \sum x_i^2 - (\sum x_i)^2 / N$$

$$s_X^2 = \frac{SST_x}{N - 1}$$

$$s_x = \sqrt{s_X^2},$$

$$\hat{\mu}_Y = \overline{Y} = \frac{\sum Y_i}{N},$$

$$SST_y = \sum(y_i - \overline{y})^2 = \sum y_i^2 - (\sum y_i)^2 / N$$

$$s_Y^2 = \frac{SST_y}{N - 1}$$

$$s_y = \sqrt{s_Y^2},$$

$$SP_{xy} = \sum(x_i - \overline{x})(y_i - \overline{y}) = \sum x_i y_i - \sum x_i \sum y_i / N,$$

$$s_{xy} = \frac{SP_{xy}}{N - 1}$$

Note: $SP_{xy} = SP_{yx}$, $s_{xy} = s_{yx}$.

Solution.

$\hat{\mu}_x$ = Σ X / N = 241/20 = 12.05,

$SST_x = \sum(x_i - \bar{x})^2 = \sum x_i^2 - (\sum x_i)^2 /N$ = 3,285 – 241^2/20 = (3,285 - 2904.05) = 380.95,

$s^2_X = SST_X/(N - 1)$ = 380.95/19 = 20.05,

s_X = 4.478

$\hat{\mu}_y$ = Σ Y / N = 488.3/20 = 24.415,

$SST_y = \sum(y_i - \bar{y})^2 = \sum y_i^2 - (\sum y_i)^2 /N$ = (13,742.27 –488.3^2) = (13,742.27 - 11,921.8445) = 1820.428,

$s^2_Y = SST_Y/(N - 1)$ = 1820.428 / 19 = 95.812,

s_Y = 9.788

$SP_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \sum x_i \sum y_i / N$ = (6,588.3 –241*488.3/20) = (6,588.3 - 5884.015) = 704.285,

$s_{xy} = SP_{xy}/(N - 1)$ = 704.285 / 19 = 37.068

---

b. Compute a and b. [VERY IMPORTANT AND INTERESTING]

Comment. As noted earlier, a and b are the sample estimates of the intercept and slope. The formulas are

$$b = \frac{\frac{1}{N-1}\sum(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{N-1}(x_i - \bar{x})^2} = \frac{SP_{xy}}{SST_x} = \frac{s_{xy}}{s_x^2},$$

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

Solution.

b = s_{XY}/s_X^2 = 37.068/ 20.05 = 1.8488; or

b = SP_{xy}/SST_x = 704.285/380.95 = 1.8488

$a = \bar{y} - b\bar{x}$ = 24.415 - 1.8488 * 12.05

= 24.415 - 22.27804 = 2.137

c. Compute SSR and SSE. [NECESSARY EVIL]

Comment. SST = Total sum of squares (default is y if no subscript is given), SSR = Regression sum of squares (also called SS Explained), SSE = error sum of squares (also called Residual sum of squares). Note that

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

e_i represents the amount of error in the regression prediction; that is, it is the extent to which the predicted score differs from the actual score.

Conversely, recall that, if b = 0, then $\hat{y} = \bar{y}$. Hence, $\hat{y}_i - \bar{y}$ represents the extent to which our prediction of Y is improved by our knowledge of X. Ergo, we get

$$SSR = \sum(\hat{y}_i - \bar{y})^2 = b^2 * SST_x = b^2 * s_X^2 * (N - 1)$$

$$= b * SP_{xy} = b * s_{xy} * (N - 1) = SST - SSE = SST_{\hat{y}}$$

$$SSE = \sum(y_i - \hat{y}_i)^2 = \sum e_i^2 = SST - SSR$$

Solution.

SSR = b² * s_x^2 * (N - 1) = b² * SST_x = 1.8488² * 380.95 = 1302.11; or
SSR = b * s_{xy} * (N - 1) = b * SP_{xy} = 1.8488 * 704.285 = 1302.08
SSE = SST - SSR = 1820.43 - 1302.11 = 518.32

Q: Show that, if b = 0, SSR = 0. Conversely, show that, if X and Y are perfectly associated, SSR = SST and SSE = 0

A: If b = 0, then

$$\hat{y}_i = \bar{y}, hence\ SSR = \sum(\bar{y} - \bar{y})^2 = 0$$

If X and Y are perfectly associated,

$$y_i = \hat{y}_i, hence\ SSE = \sum(\hat{y}_i - \hat{y}_i)^2 = 0, SSR = SST - SSE = SST - 0 = SST$$

d. Compute the (sample) standard error of the estimate (SEE or s_e). [SOMEWHAT INTERESTING]

Comment. The formula for the SEE (also sometimes just called the standard error) is

$$s_e = \sqrt{\frac{SSE}{N - K - 1}}$$

The value of s_e can be interpreted in a manner similar to the sample standard deviation of the values of x about $\bar{x}$. Given that $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, then approximately 68.3% of the observations will fall within $\pm 1s_e$ units of the regression line, 95.4% will fall within $\pm 2s_e$ units, and 99.7% will fall within $\pm 3s_e$ unit. Using this gives one a good indication of the fit of the regression line to the sample data. Note that increases in sample size increase both the numerator and the denomination, so s_e is relatively unaffected by the size of the sample.

Solution.

$s_e = \sqrt{(SSE/(n-2))} = \sqrt{(518.32/18)} = 5.366$

---

e. Compute s_b, the standard error of the regression coefficient b. [INTERESTING]

Comment. s_b is a measure of the amount of sampling error in the regression coefficient b, just as $s_{\underline{x}}$ is a measure of the sampling variability in $\bar{x}$. The formula is

$$s_b = \frac{s_e}{\sqrt{SST_x}} = \sqrt{\frac{MSE}{SST_x}}$$

(MSE is defined shortly). Once we have s_b, we will be able to proceed much the same as we do when we conduct tests concerning the population mean. A t-test (with N - 2 d.f., since a and b have been estimated) can be used to test the null hypothesis H_0: ß = ß$_0$. This test is very similar to the t-test about a population mean, as we are again testing a mean (ß), the population is assumed to be normal (the ε_i's) and the population standard deviation is unknown. In the present case, the sample statistic is b (rather than $\underline{x}$) and the sample standard error is s_b. Incidentally, note that increases in the sample size cause the standard error to go down.

Solution.

$s_b = s_e / \sqrt{(SST_x)} = 5.366 / \sqrt{(380.95)} = .2749$

f. Compute the 95% confidence interval for ß. [INTERESTING]

Comment. Do this the same way you would a c.i. for a population mean, i.e. proceed much as you would for single sample tests, case III, σ unknown. d.f. = N - 2. The c.i. is

$$b \pm t_{\alpha/2} * s_b \text{, i.e.}$$

$$b - t_{\alpha/2} * s_b \leq \ \leq b + t_{\alpha/2} * s_b$$

Solution.

95% c.i. = $b \pm t_{\alpha/2,n-2} * s_b = b \pm t_{.025,18} * s_b =$
$1.8488 \pm 2.101 * .2749 \Rightarrow$
$1.27 \leq ß \leq 2.43$

Note that 0 does NOT fall in the confidence interval, suggesting b significantly differs from 0.

---

g. Use a T-test to test whether b significantly differs from 0. [INTERESTING]

Comment. Again, this is very similar to single sample tests, case III.

Solution.

Step 1. H_0: ß = 0
H_A: ß <> 0

Step 2. An appropriate test stat is

$$T_{N-2} = \frac{b - \beta_0}{s_b} = \frac{b}{s_b} = \frac{b}{.2749}$$

Step 3. For α = .05, accept H_0 if $-2.101 \leq T \leq 2.101$

Step 4. The computed value of the test statistic is

$$T_{18} = \frac{b - \beta_0}{s_b} = \frac{b}{s_b} = \frac{1.8488}{.2749} = 6.73$$

Step 5. Reject H_0.

---

h. Compute MST, MSR, and MSE. [NECESSARY EVIL]

Comment. The only trick is figuring out the d.f. For MST, d.f. = N - 1, for MSR, d.f. = K where K is the number of b's that have been estimated (in this case, 1), for MSE d.f. = N - K - 1 = N - 2 in this case.

$$MST = \frac{SST}{N - 1} = s_y^2,$$

$$MSR = \frac{SSR}{K} = \frac{SSR}{1},$$

$$MSE = \frac{SSE}{N - K - 1} = \frac{SSE}{N - 2} = s_e^2$$

MST = SST/(N-1) = 1820.428/19 = s^2_Y = 95.812

MSR = SSR/K = SSR/1 = 1302.08

MSE = SSE/(N-2) = 518.32/18 = 28.79. Or,
MSE = $s^2_e = 5.366^2$ = 28.79

---

i. Construct the ANOVA table. [F TEST IS OF PRIMARY INTEREST]

General format:

Source	SS	d.f.	MS	F
Regression (or explained)	SSR	K	SSR / K	MSR/MSE
Error (or residual)	SSE	N - K - 1	SSE / (N-K-1)	
Total	SST	N - 1	SST / (N - 1)	

For this problem:

Source	SS	d.f.	MS	F
Regression (or explained)	SSR = 1302.08	K = 1	SSR / K = 1302.08	MSR/MSE = 45.23
Error (or residual)	SSE = 518.32	N - K - 1 = 18	SSE / (N-K-1) = 28.79	
Total	SST = 1820.40	N - 1 = 19	SST / (N - 1) = 95.81	

NOTE: The degrees of freedom are K, N-K-1. For an F with d.f. = 1,18 and $\alpha = .05$, accept H_0 if $F \leq 4.41$. Also, note that 45.23 = approximately 6.73^2, the value we got for the T statistic squared. The F test is a test of whether any betas significantly differ from zero; or equivalently, it is a test of whether R^2 differs from 0.

---

j. Compute r_{yx} and r^2_{yx} [INTERESTING]

Comment. r^2_{yx} is the proportion of variance in y that is accounted for, or explained, by X. r^2 is also called the coefficient of determination. r^2_{yx} represents the strength of the linear relationship that is present in the data. The closer y is to $\hat{y}$, the bigger r^2_{yx} will be. In a bivariate regression, r can range from -1 to 1, while r^2 ranges from 0 to 1. r_{yx} is the bivariate correlation between y and x; it is our estimate of the population parameter rho (ρ). Formulas:

$$r_{yx} = \frac{s_{xy}}{s_x s_y},$$

$$r^2_{yx} = SSR/SST$$

Solution.

$r_{yx} = s_{XY}/s_X s_Y = 37.068 / (4.478 * 9.788) = .846$
$r^2_{yx} = .846^2 = .716$. Or,
$r^2_{yx} = SSR/SST = 1302.08 / 1820.428 = .715$

NOTE: We also use R_{yx} and R_{yx}^2, which are referred to as Multiple R and Multiple R^2 (terms that make more sense when there is more than one IV being used).

$$R_{yx} = R_{y\hat{y}} = \sqrt{SSR/SST}\,, R^2_{yx} = SSR/SST$$

That is R_{yx} is the correlation of the observed value of Y with the predicted value of Y. This number must always be positive or zero, hence $R_{yx} = ABS(r_{yx})$

---

k. Test whether r_{yx} significantly differs from 0.

Comment. The appropriate test statistic is given in step 2.

Solution.

Step 1. H_0: $\rho = 0$
H_A: $\rho <> 0$

Step 2. The appropriate test statistic is

$$t_{N-2} = \frac{r_{yx}\sqrt{N-2}}{\sqrt{1-r_{yx}^2}}$$

Step 3. For $\alpha = .05$ and d.f. = n - 2 = 18, accept H_0 if $-2.101 \leq T \leq 2.101$

Step 4. The computed value of the test statistic is

$$t_{N-2} = \frac{r_{yx}\sqrt{N-2}}{\sqrt{1-r_{yx}^2}} = \frac{.846*\sqrt{18}}{\sqrt{1-.715}} = 6.72$$

Step 5. Reject H_0

Incidentally, note that, in the bivariate regression case, a test of H_0: $\rho = 0$ yields the same results as a test of H_0: ß = 0. This will generally *not* be true as more predictor variables are added.

Alternatively, the F test could also be used, and would yield the same result. Indeed, the above approach only works in a bivariate regression. Once more variables are added you need to do an F test. The main advantage of the above approach is that you could test a one-tailed alternative hypothesis.

---

l. Compute Adjusted R^2.

Comment. R^2 is biased upward, particularly in small samples. Therefore, *adjusted* R^2 is sometimes used. The formula is

$$\text{Adjusted } R^2 = 1-\left(\frac{(N-1)(1-R^2)}{(N-K-1)}\right)$$

In this case,

$$\text{Adjusted } R^2 = 1-\left(\frac{(N-1)(1-R^2)}{(N-K-1)}\right) = 1-\frac{(20-1)(1-.715)}{20-1-1} = 1-\frac{19*.285}{18} = 1-\frac{5.415}{18} = .699$$