Multiple Regression - Introduction

We will add a 2nd independent variable to our previous example. Data are collected from 20 individuals on their years of education (X_1), years of job experience (X_2), and annual income in thousands of dollars (Y). The data are as follows:

X_1	X_2	Y	X_1Y	X_2Y	X_1X_2	X_1^2	X_2^2	Y^2
2	9	5.0	10.0	45.0	18	4	81	25.00
4	18	9.7	38.8	174.6	72	16	324	94.09
8	21	28.4	227.2	596.4	168	64	441	806.56
8	12	8.8	70.4	105.6	96	64	144	77.44
8	14	21.0	168.0	294.0	112	64	196	441.00
10	16	26.6	266.0	425.6	160	100	256	707.56
12	16	25.4	304.8	406.4	192	144	256	645.16
12	9	23.1	277.2	207.9	108	144	81	533.61
12	18	22.5	270.0	405.0	216	144	324	506.25
12	5	19.5	234.0	97.5	60	144	25	380.25
12	7	21.7	260.4	151.9	84	144	49	470.89
13	9	24.8	322.4	223.2	117	169	81	615.04
14	12	30.1	421.4	361.2	168	196	144	906.01
14	17	24.8	347.2	421.6	238	196	289	615.04
15	19	28.5	427.5	541.5	285	225	361	812.25
15	6	26.0	390.0	156.0	90	225	36	676.00
16	17	38.9	622.4	661.3	272	256	289	1,513.21
16	1	22.1	353.6	22.1	16	256	1	488.41
17	10	33.1	562.7	331.0	170	289	100	1,095.61
21	17	48.3	1,014.3	821.1	357	441	289	2,332.89
T_{X1} = 241	T_{X2} = 253	T_Y = 488.3	T_{X1Y}= 6,588.3	T_{X2Y}= 6448.9	T_{X1X2} = 2999	T_{X1^2} = 3,285	T_{X2^2} = 3,767	T_{Y^2} = 13,742.27

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

```
DATA LIST FREE / Educ JobExp Income.
BEGIN DATA.
    2      9      5.0
    4     18      9.7
    8     21     28.4
    8     12      8.8
    8     14     21.0
   10     16     26.6
   12     16     25.4
   12      9     23.1
   12     18     22.5
   12      5     19.5
   12      7     21.7
   13      9     24.8
   14     12     30.1
   14     17     24.8
   15     19     28.5
   15      6     26.0
   16     17     38.9
   16      1     22.1
   17     10     33.1
   21     17     48.3
END DATA.
REGRESSION /DESCRIPTIVES ALL /STATISTICS ALL/DEPENDENT INCOME
           /METHOD ENTER EDUC JOBEXP/ SCATTERPLOT (EDUC JOBEXP) /
           /SCATTERPLOT (INCOME EDUC) / SCATTERPLOT (INCOME JOBEXP)/
           /SCATTERPLOT (INCOME *PRED) / .
```

Selected output:

Descriptive Statistics

	Mean	Std. Deviation	Variance	N
INCOME	24.4150	9.78835	95.81187	20
EDUC	12.0500	4.47772	20.05000	20
JOBEXP	12.6500	5.46062	29.81842	20

Correlations

		INCOME	EDUC	JOBEXP
Pearson Correlation	INCOME	1.000	.846	.268
	EDUC	.846	1.000	-.107
	JOBEXP	.268	-.107	1.000
Covariance	INCOME	95.812	37.068	14.311
	EDUC	37.068	20.050	-2.613
	JOBEXP	14.311	-2.613	29.818
Sig. (1-tailed)	INCOME	.	.000	.127
	EDUC	.000	.	.327
	JOBEXP	.127	.327	.
Sum of Squares and Cross-products	INCOME	1820.425	704.285	271.905
	EDUC	704.285	380.950	-49.650
	JOBEXP	271.905	-49.650	566.550
N	INCOME	20	20	20
	EDUC	20	20	20
	JOBEXP	20	20	20

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Selection Criteria			
					R Square Change	F Change	df1	df2	Sig. F Change	Akaike Information Criterion	Amemiya Prediction Criterion	Mallows' Prediction Criterion	Schwarz Bayesian Criterion
1	.919[a]	.845	.827	4.07431	.845	46.332	2	17	.000	58.938	.210	3.000	61.925

a. Predictors: (Constant), JOBEXP, EDUC

b. Dependent Variable: INCOME

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1538.225	2	769.113	46.332	.000[a]
	Residual	282.200	17	16.600		
	Total	1820.425	19			

a. Predictors: (Constant), JOBEXP, EDUC

b. Dependent Variable: INCOME

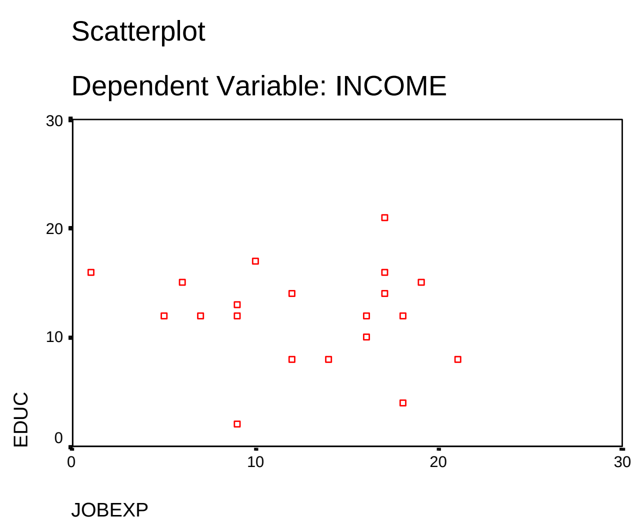
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		t	Sig.	95% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta	Std. Error			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	-7.097	3.626			-1.957	.067	-14.748	.554					
	EDUC	1.933	.210	.884	.096	9.209	.000	1.490	2.376	.846	.913	.879	.989	1.012
	JOBEXP	.649	.172	.362	.096	3.772	.002	.286	1.013	.268	.675	.360	.989	1.012

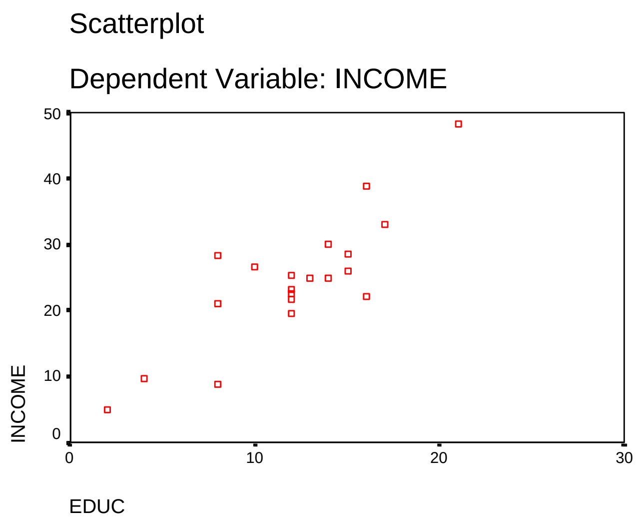
a. Dependent Variable: INCOME

Here are the <u>scatterplots</u> for the different variables we are examining. These will hopefully give you a better idea about how these variables are related to each other, and what r^2 and strength of association means.

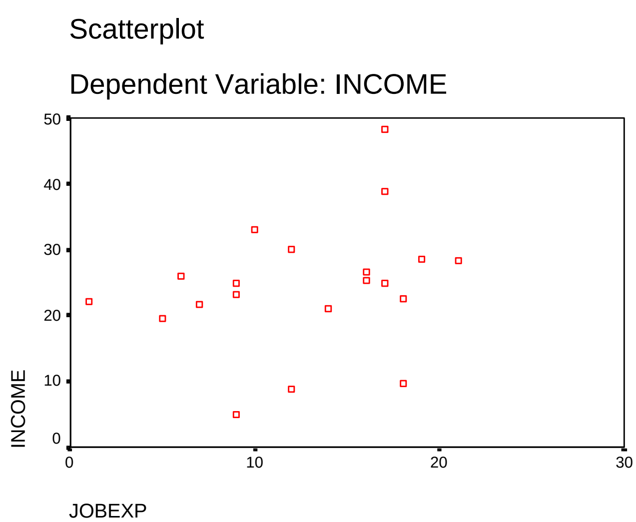
1. Education by job experience: (r = -.107). Note how there is almost no pattern to the dots, which is consistent with the very weak association between these variables.



2. Education by Income (r = .846). There is a much clearer and stronger linear association here.



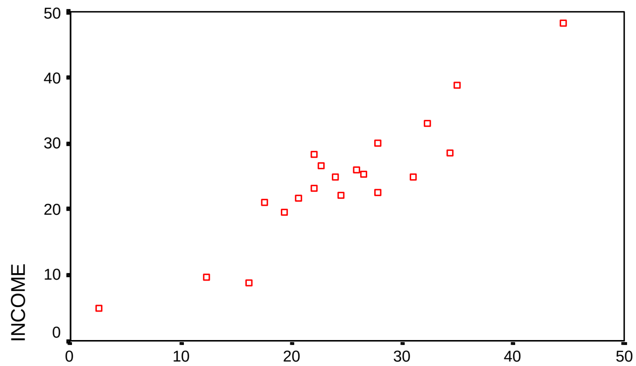
3. Job experience by income (r = .268). There appears to be linear association here, but, as the lower r would indicate, it does not seem to be as strong as was the case with education and income.



4. Predicted income by income (r = .919). This is a plot of $\hat{y}$ by y. As the r value suggests, the linear association is very clear (even more so than was the case with education and income), although not perfect.

Scatterplot

Dependent Variable: INCOME



a. Determine $\hat{\mu}_{X2}$, SST_{X2}, s^2_{X2}, s_{X2}, SP_{X1X2}, s_{X1X2}, SP_{X2Y}, s_{X2Y}.

Comment. The means, standard deviations, etc. for X_1 and Y are the same as before.

Solution.

$\hat{\mu}_{X2} = \Sigma X_2 / N = 253/20 = 12.65$,
$SST_{X2} = \Sigma X_2^2 - (\Sigma X_2)^2/N = (3{,}767 - 253^2) = (3{,}767 - 3200.45) = 566.55$,
$s^2_{X2} = SST_{X2}/(N - 1) = 566.55/19 = 29.82$,
$s_{X2} = 5.46$

$SP_{X1X2} = \Sigma X_1X_2 - \Sigma X1 \Sigma X2/N = (2{,}999 - 241 * 253 / 20) = (2{,}999 - 3048.65) = -49.65$,
$s_{X1X2} = SP_{X1X2}/(N - 1) = -49.65 / 19 = -2.613$

$SP_{X2Y} = \Sigma X_2Y - \Sigma X2 \Sigma Y/N = (6{,}448.9 - 253 * 488.3 / 20) = (6{,}448.9 - 6176.995) = 271.905$,
$s_{X2Y} = SP_{X2Y}/(N - 1) = 271.905 / 19 = 14.31$

---

b. Compute a, b_1, and b_2. [VERY IMPORTANT]

Comment. The formulas are

$$b_1 = \frac{(s_2^2 * s_{y1}) - (s_{12} * s_{y2})}{(s_1^2 * s_2^2) - s_{12}^2} = \frac{(SST_{X2} * SP_{X1Y}) - (SP_{X1X2} * SP_{X2Y})}{(SST_{X1} * SST_{X2}) - SP_{X1X2}^2}$$

$$b_2 = \frac{(s_1^2 * s_{y2})\ (s_{12} * s_{y1})}{(s_2^2 * s_1^2) - s_{12}^2} = \frac{(SST_{X1} * SP_{X2Y})\ (SP_{X1X2} * SP_{X1Y})}{(SST_{X2} * SST_{X1}) - SP_{X1X2}^2}$$

$$a = \hat{\mu}_y - (b_1 * \hat{\mu}_1) - (b_2 * \hat{\mu}_2)$$

Two Independent Variables - Proof [Optional].

$y_i = a + b_1x_{1i} + b_2x_{2i} + e_i$

==> $y_i - \underline{y} = a + b_1x_{1i} + b_2x_{2i} + e_i - \underline{y}$ (subtract $\underline{y}$ from both sides)

==> $\Sigma(y_i - \underline{y}) = \Sigma(b_1(x_{1i} - \underline{x}_1) + b_2(x_{2i} - \underline{x}_2))$ (substitute for a, sum all cases)

==> $SP_{y1} = b_1 * SST_{X1} + b_2 * SP_{12}$, (multiply by $x_{1i} - \underline{x}_1$)
$SP_{y2} = b_1 * SP_{12} + b_2 * SST_{X2}$ (multiply by $x_{2i} - \underline{x}_2$)

==> $b_1 = (SP_{y1} - b_2 * SP_{12})/SST_{X1}$, (from the last 2 equations)
$b_2 = (SP_{y2} - b_1 * SP_{12})/SST_{X2}$

At this point, we substitute the value for b_2 into the b_1 equation:

$$b_1 = \frac{SP_{Y1}\ \ \dfrac{SP_{Y2}\ \ b_1 * SP_{12}}{SST_{X2}} * SP_{12}}{SST_{X_1}}$$

$$= \frac{(SST_{X_2} * SP_{Y1})\ (SP_{Y2} * SP_{12}) + (b_1 * SP_{12}^2)}{SST_{X_1} * SST_{X_2}}$$

(In the latter step, we multiply both numerator and denominator by SST_{X2}).
We now need to isolate b_1 on the left-hand side. First, we multiply both sides by the right-hand denominator:

$$b_1 * SST_{X_1} * SST_{X_2} = (SST_{X_2} * SP_{Y1})\ (SP_{Y2} * SP_{12}) + (b_1 * SP_{12}^2)$$

We now subtract $b_1 * SP_{12}^2$ from both sides:

$$b_1 * (SST_{X_1} * SST_{X_2}\ \ SP_{12}^2) = (SST_{X_2} * SP_{Y1})\ (SP_{Y2} * SP_{12})$$

Now we simply divide both sides by ($SST_{X1} * SST_{X2} - SP_{12}^2$), yielding the formula for b_1 given originally. Through a similar process, we can prove the formula for b_2.

Though the change in formulas between the bivariate and multivariate case may seem inexplicable, there is a logic and consistent pattern behind the formulas. Fully understanding this logic, however, requires knowledge of matrix algebra. In case you happen to know matrix algebra: if $X_{0i} = 1$ for all cases, then it is very easy to show that $b = (X'X)^{-1}X'Y$. See Hayes, Appendix D, if you want more information on this.

Any number of IVs - Proof that $b = (X'X)^{-1}X'Y$ [Optional]. Let X be an N x K matrix (i.e. N cases, each of which has K X variables, including X_0.) Y is an N x 1 matrix. e is an N x 1 matrix. Then, if the assumptions of OLS regression are met,

$Y = Xb + e$	
$Y – e = Xb$	Subtract e from both sides
$X'(Y – e) = X'Xb$	Premultiply both sides by X’
$X'Y = X'Xb$	If the assumptions of OLS regression are met, $X'e = 0$ because the Xs are uncorrelated with the residuals of Y
$(X'X)^{-1}X'Y = (X'X)^{-1}X'Xb$	Premultiply both sides by $(X'X)^{-1}$
$(X'X)^{-1}X'Y = b$	$(X'X)^{-1}X'X = I$ and $Ib = b$

Solution.

$b_1 = (s_2^2 * s_{y1} - s_{12} * s_{y2}) / (s_1^2 * s_2^2 - s_{12}^2) =$
$(29.82 * 37.07 - -2.61 * 14.31) / (20.05 * 29.82 - -2.61^2) =$
$1142.78 / 591.08 = 1.933$; or,
$b_1 = (SST_{X2} * SP_{X1Y} - SP_{X1X2} * SP_{X2Y}) / (SST_{X1} * SST_{X2} - SP_{X1X2}^2) =$
$(566.55 * 704.285 - -49.65 * 271.905) / (380.95 * 566.55 - -49.65^2) =$
$412512.75/213362.1 = 1.933$

$b_2 = (s_1^2 * s_{y2} - s_{12} * s_{y1}) / (s_2^2 * s_1^2 - s_{12}^2) =$

(20.05 * 14.31 - -2.61 * 37.07) / (29.82 * 20.05 - -2.61²) =

383.67 / 591.08 = .649; or,

$b_2 = (SST_{X1} * SP_{X2Y} - SP_{X1X2} * SP_{X1Y}) / (SST_{X2} * SST_{X1} - SP_{X1X2}^2) =$

(380.95 * 271.905 - -49.65 * 704.285) / (566.55 * 380.95 - -49.65²) =

138549.96/213362.1 = .649

$a = \hat{\mu}_Y - b_1 * \hat{\mu}_{X1} - b_2 * \hat{\mu}_{X2}$ = 24.415 - 1.933 * 12.05 - .649 * 12.65 =

24.415 - 23.29265 - 8.20985 = -7.0875

---

c. Compute SSR and SSE. [NECESSARY EVIL]

Comment. The formulas are

$$SSR = \sum(\hat{y}_i - \overline{y})^2$$

$$= b_1^2 * SST_{X_1} + b_2^2 * SST_{X_2} + 2 b_1 b_2 * SP_{X_1 X_2}$$

$$= (b_1^2 * s_{X_1}^2 + b_2^2 * s_{X_2}^2 + 2 b_1 b_2 * s_{X_1 X_2}) * (N - 1)$$

$$= b_1 * SP_{X_1 Y} + b_2 * SP_{X_2 Y}$$

$$= (b_1 * s_{X_1 Y} + b_2 * s_{X_2 Y}) * (N - 1) = SST \quad SSE = SST_{\hat{y}}$$

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

For a proof of one of the above, note that, according to the rules for expectations,
$\hat{y} = a + b_1 x_1 + b_2 x_2 \implies v(\hat{y}) = b_1^2 s_{X1}^2 + b_2^2 s_{X2}^2 + 2 b_1 b_2 s_{X1X2}$.

Solution.

$SSR = b_1^2 * SST_{X1} + b_2^2 * SST_{X2} + 2b_1b_2SP_{X1X2}$

= 1.933² * 380.95 + .649² * 566.55 + 2 * 1.933 * .649 * -49.65 = 1537.47; or,

$SSR = (b_1^2 s_{X1}^2 + b_2^2 s_{X2}^2 + 2b_1b_2 s_{X1X2}) * (N - 1)$

= 1.933² * 20.05 + .649² * 29.82 + 2 * 1.933 * .649 * -2.613 * 19 = 1537.49; or,

$SSR = b_1 * SP_{X1Y} + b_2 * SP_{X2Y}$ = 1.933 * 704.285 + .649 * 271.905 = 1537.85; or,

$SSR = (b_1 * s_{X1Y} + b_2 * s_{X2Y}) * (N - 1)$ = (1.933 * 37.068 + .649 * 14.31) * 19 = 1537.85

SSE = SST - SSR = 1820.425 - 1537.47 = 282.96

---

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

Comment. The formula for the SEE is the same as in the bivariate case; however, K = 2 in this example, since there are two independent variables.

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

As before, s_e is the standard deviation of the residuals. 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 $\pm 1 s_e$ units of the regression line, 95.4% will fall within $\pm 2 s_e$ units, and 99.7% will fall within $\pm 3 s_e$ unit. Using this gives one a good indication of the fit of the regression line to the sample data.

Solution.

$s_e = \sqrt{(SSE/(N - K - 1))} = \sqrt{(282.96/17)} = 4.08$

---

e. Compute s_{bk}, k = 1, K, the standard errors of the regression coefficients b_k. [IMPORTANT]

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

$$s_{b_k} = \frac{s_e}{\sqrt{(1 - r_{12}^2) * SST_{x_k}}} = \sqrt{\frac{SSE}{(1 - r_{12}^2) * SST_{x_k} * (N - K - 1)}}$$

$$= \frac{s_e}{\sqrt{(1 - r_{12}^2) * s_{x_k}^2 * (N - 1)}} = \sqrt{\frac{SSE}{(1 - r_{12}^2) * s_{x_k}^2 * (N - 1) * (N - K - 1)}}$$

$$= \sqrt{\frac{1 - r_{y12}^2}{(1 - r_{12}^2) * (N - K - 1)}} * \frac{s_y}{s_{x_k}} = s_{b_{k'}} * \frac{s_y}{s_{x_k}}$$

Once we have s_{bk}, 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 - 3 d.f., since a, b_1 and b_2 have been estimated) can be used to test the null hypothesis H_0: $ß_k = ß_0$. This test is very similar to the t-test about a population mean, as we are again testing a mean ($ß_k$), 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 $\bar{x}$) and the sample standard error is s_{bk}.

Solution.

$$s_{b_1} = \frac{s_e}{\sqrt{(1 - r_{12}^2) * SST_{x_1}}} = \frac{4.08}{\sqrt{(1 - .107^2) * 380.95}} = .210$$

$$= \sqrt{\frac{SSE}{(1 - r_{12}^2) * SST_{x_1} * (N - K - 1)}} = \sqrt{\frac{282.96}{(1 - .107^2) * 380.95 * 17}} = .210$$

$$= \sqrt{\frac{1 - r_{y12}^2}{(1 - r_{12}^2) * (N - K - 1)}} * \frac{s_y}{s_{x_1}} = \sqrt{\frac{1 - .845}{(1 - .107^2) * 17}} * \frac{9.788}{4.478}$$

$$= .096 * \frac{9.788}{4.478} = .210$$

$$s_{b_2} = \frac{s_e}{\sqrt{(1 \;\; r_{12}^2) * SST_{x_2}}} = \frac{4.08}{\sqrt{(1 \;\; .107^2) * 566.55}} = .172$$

$$= \sqrt{\frac{SSE}{(1 \;\; r_{12}^2) * SST_{x_2} * (N \;\; K \;\; 1)}} = \sqrt{\frac{282.96}{(1 \;\; .107^2) * 566.55 * 17}} = .172$$

$$= \sqrt{\frac{1 \;\; r_{y12}^2}{(1 \;\; r_{12}^2) * (N \;\; K \;\; 1)}} * \frac{s_y}{s_{x_2}} = \sqrt{\frac{1 \;\; .845}{(1 \;\; .107^2) * 17}} * \frac{9.788}{5.461}$$

$$= .096 * \frac{9.788}{5.461}$$

We will discuss standard errors in greater detail later.

f. Compute the 95% confidence intervals for $ß_k$. [IMPORTANT]

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 - K - 1 = N - 3. The c.i. is

$$b_k \pm t_{\alpha/2} * s_{b_k}, i.e.$$

$$b_k - t_{\alpha/2} * s_{b_k} \leq \beta_k \leq b_k + t_{\alpha/2} * s_{b_k}$$

Solution.

95% c.i. for $b_1 = b_1 \pm t_{\alpha/2,n-3} * s_{b1} = b_1 \pm t_{.025,17} * s_{b1} =$
$1.933 \pm 2.110 * .210 \implies$
$1.490 \le ß_1 \le 2.376$

95% c.i. for $b_2 = b_2 \pm t_{\alpha/2,n-3} * s_{b2} = b_2 \pm t_{.025,17} * s_{b2} =$
$.649 \pm 2.110 * .172 \implies$
$.286 \le ß_2 \le 1.012$

Note that 0 does NOT fall in either confidence interval, suggesting the b's significantly differ from 0.

---

g. Do a t-test to determine whether b_1 significantly differs from 0. [IMPORTANT]

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

Solution.

Step 1. $H_0: ß_1 = 0$
$H_A: ß_1 <> 0$

Step 2. An appropriate test stat is

$$T_{N-K-1} = \frac{b_k - \beta_{k0}}{s_{b_k}} = \frac{b_k}{s_{b_k}}$$

In this case, k = 1, DF = 17, and $s_{bk} = s_{b1} = .210$.

Step 3. For $\alpha = .05$, accept H_0 if $-2.11 \le T_{17} \le 2.11$

Step 4. For b_1, the computed value of the test statistic is

$$T_{N-K-1} = \frac{b_k - \beta_{k0}}{s_{b_k}} = \frac{b_k}{s_{b_k}} = \frac{1.933}{.209} = 9.205$$

Step 5. Reject H_0.

If we repeat the process for b_2, we get $T_{17} = .649/.172 = 3.773$. Again, we would 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, 2), for MSE d.f. = N - K - 1 = N - 3 in this case.

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

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

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

MST = SST/(N-1) = 1820.428/19 = s^2_Y = 95.81
MSR = SSR/K = SSR/2 = 1537.47/2 = 768.74,
MSE = SSE/(N - K - 1) = 282.96/17 = 16.64. Or,
MSE = s^2_e = 4.08^2 = 16.65

---

i. Construct the ANOVA table. [IMPORTANT FOR THE F VALUE -- AND FOR NESTED COMPARISONS]

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 = 1537.47	K = 2	SSR / K = 768.74	MSR/MSE = 46.20*
Error (or residual)	SSE = 282.96	N - K - 1 = 17	SSE / (N-K-1) = 16.64	
Total	SST = 1820.43	N - 1 = 19	SST / (N - 1) = 95.81	

NOTE: For an F with d.f. = 2,17 and $\alpha = .05$, accept H_0 if $F \leq 3.59$. Also, note that, unlike the bivariate regression case, the T-tests and the F-test are not equivalent to each other. The F-test is a test of the hypothesis

H_0: $ß_1 = ß_2 = ... = ß_K = 0$
H_A: At least one $ß_k$ does not equal 0.

The F-test can also be thought of as a test of
H_0: $\rho = 0$
H_A: ρ <> 0

---

j. Compute R_{yx1x2} and R^2_{yx1x2}, i.e. Multiple R and Multiple R^2 [IMPORTANT, ALBEIT OVER-RATED]

Comment. R^2_{yx1x2} is the proportion of variance in y that is accounted for, or explained, by X_1 and X_2. r^2 is also called the coefficient of determination. R^2_{yx1x2} represents the strength of the linear relationship that is present in the data. The closer y is to $\hat{y}$, the bigger R^2_{yx1x2} will be. In a multiple regression, R and R^2 range from 0 to 1. R_{yx1x2} is our estimate of the population parameter rho (ρ). Formulas:

$$R^2_{yx_1x_2} = SSR/SST,$$

$$R_{yx_1x_2} = \sqrt{SSR/SST} = \sqrt{R_{yx_1x_2}{}^2}$$

Solution.

R^2_{yx1x2} = SSR/SST = 1537.47/1820.43 = .845,

$R_{yx1x2} = \sqrt{.845} = .919$

---

k. Test whether R_{yx1x2} significantly differs from 0. [IMPORTANT]

Comment. As noted above, the F-test is a test of

H_0: $\rho = 0$
H_A: $\rho <> 0$

The t-test procedure we also used in the bivariate regression case is not appropriate when there is more than 1 independent variable.

---

l. Alternative formulas for F and R^2: [SOMETIMES VERY USEFUL]

$$F = \frac{R^2 * (N - K - 1)}{(1 - R^2) * K},$$

$$R^2 = \frac{F * K}{(N - K - 1) + (F * K)}$$

[OPTIONAL] Proof:

$R^2 = SSR/SST$,	[as defined above]
$SST/SST = (SSR + SSE)/SST = 1$	[substitute for SST]
$= SSR/SST + SSE/SST = R^2 + SSE/SST$	[rearrange terms, substitute in r^2]
$\Rightarrow 1 = R^2 + SSE/SST$	[from last two lines]
$\Rightarrow SSE/SST = 1 - R^2$	[substract r^2 from both sides]

Further,

$F = MSR/MSE = (SSR/K)/(SSE/[N - K - 1])$	[defn of F, MSR, MSE]
$= \dfrac{SSR * (N - K - 1)}{SSE * K}$	[rearrange terms]
$= \dfrac{SSR/SST * (N - K - 1)}{SSE/SST * K}$	[divide top and bottom by SST]
$\Rightarrow F = \dfrac{R^2 * (N - K - 1)}{(1 - R^2) * K}$	[substitute for SSR/SST and SSE/SST]

And, for R^2,

$R^2/(1 - R^2) = (F * K)/(N - K - 1)$ [Multiply both sides by K/(N - K - 1)]

$\Longrightarrow (1 - R^2)/R^2 = (N - K - 1)/(F * K)$ [take reciprocals]

$\Longrightarrow (1 - R^2 + R^2)/r^2 = ((N - K - 1) + (F * K))/(F * K)$ [add 1 to both sides]

$\Longrightarrow 1/R^2 = (F * K)/ ((N - K - 1) + (F * K))/(F * K)$ [since $1 - r^2 + r^2 = 1$]

$\Longrightarrow R^2 = (F * K)/ ([N - K - 1] + [F * K])$ [take reciprocals]

Note that, as R^2 gets bigger, F will increase; F also increases as the sample size increases. Hence, the value of F is dependent on both the strength of association and on the sample size. Conversely, changes in sample size have no necessary effect on R^2.

These alternative formulas can be <u>very</u> useful, since it is not unusual for either F or R^2 to not be reported, while the other necessary information is.

In the present case,

$$F = \frac{R^2 * (N - K - 1)}{(1 - R^2) * K} = \frac{.845 * 17}{.155 * 2} = 46.34,$$

$$R^2 = \frac{F * K}{(N - K - 1) + (F * K)} = \frac{46.20 * 2}{17 + (46.20 * 2)} = \frac{92.4}{109.4} = .845$$

---

m. Compute Adjusted R^2.

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

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

Note that, unlike regular R^2, Adjusted R^2 can actually get smaller as additional variables are added to the model. As N gets bigger, the difference between R^2 and Adjusted R^2 gets smaller and smaller.

$$\text{Adjusted R}^2 = 1 - \left(\frac{(N-1)(1-R^2)}{(N-K-1)} \right) = 1 - \left(\frac{(20-1)(1-.845)}{(20-2-1)} \right) = .827$$