Multiple Linear Regression Example

Data: The data in this example were used in a study of teenage gambling.

Variables:

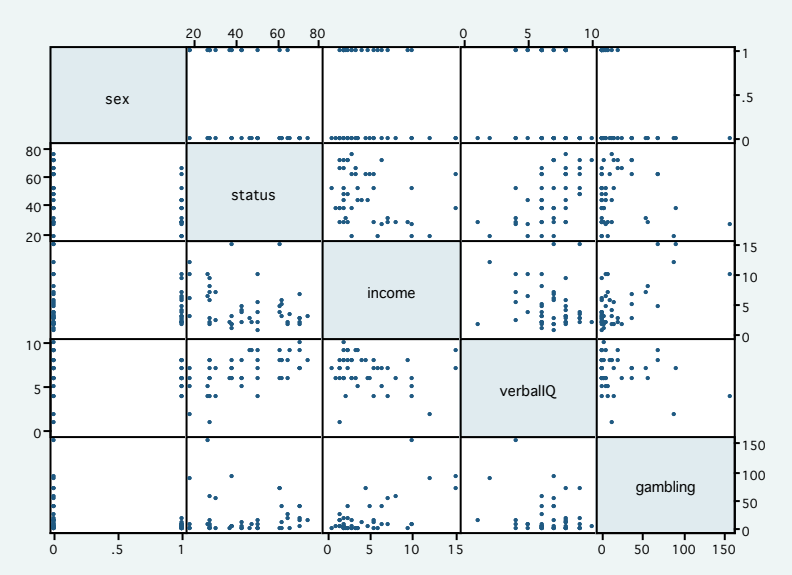
	ID	Individual
	Sex	Gender (0 = male; 1 = female)
	Status	Parents' occupation (higher numbers = higher status)
	Income	Pocket money and earnings (in pounds per week)
	VerbalIQ	Verbal intelligence (number of words out of 12 correctly defined)
	Gambling	Estimated expenditures on all forms of gambling in pounds per year (from survey)

Source: Ide-Smith & Lea (1988). *Journal of Gambling Behavior*, 4, 110-118

Here's a sample of the data from the 47 subjects:

ID	sex	status	income	verbalIQ	gambling
1	1	51	2	8	0
2	1	28	2.5	8	0
3	1	37	2	6	0
4	1	28	7	4	7.3
5	1	65	2	8	19.6
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
45	0	62	4.94	6	38
46	0	71	1.5	7	14.4
47	0	71	2.5	9	19.2
Mean	.4042553	45.23404	4.641915	6.659574	19.30106
Std. Dev	.4960529	17.26294	3.551371	1.856558	31.51587
Variance	.2460685	298.0091	12.61224	3.446808	993.2501

Scatterplot Matrix



1. **Before you look at the correlations, decide which variables you want to use to predict gambling. List these variables in the order in which you would put them into your prediction equation.**

Predictor #1: _____ **Predictor #2:** _____

Predictor #3: _____ **Predictor #4:** _____

Correlation Matrix

	sex	status	income	verbalIQ	gambling
sex	1.0000				
status	-0.4809	1.0000			
income	-0.1155	-0.2750	1.0000		
verbalIQ	-0.1070	0.5316	-0.1756	1.0000	
gambling	-0.4078	-0.0504	0.6221	-0.2201	1.0000

Squared Correlations

Let: X1 = sex X2 = status X3 = income X4 = VerbalIQ Y = Gambling

$R^2_{Y1} = 0.1663$

$R^2_{Y12} = 0.2454$ $R^2_{Y2} = 0.0025$

$R^2_{Y13} = 0.5014$ $R^2_{Y23} = 0.4027$ $R^2_{Y3} = 0.3870$

$R^2_{Y14} = 0.2366$ $R^2_{Y24} = 0.0546$ $R^2_{Y34} = 0.3997$ $R^2_{Y4} = 0.0484$

$R^2_{Y123} = 0.5058$ $R^2_{Y124} = 0.2629$ $R^2_{Y134} = 0.5263$ $R^2_{Y234} = 0.4450$ $R^2_{Y1234} = 0.5267$

We will now start building prediction models. Our most simple model would be one with no predictor variables.

Reduced Model A: $\hat{Y} = b_0$

2. What value would we use for b_0 ? How can we calculate SSY, the total variance in our dependent variable?

SSY = _____

3. Select your first predictor and write out your first **Full Model.** Then fill-in the ANOVA summary table and write your conclusions. Finally, verify your result with the omnibus F-ratio.

Full Model A: _____

Source	SS	df	MS	MSR

Omnibus F-ratio: _____

4. Write out the prediction equation (a computer will calculate this) and interpret the coefficients.

Equation: _____

5. Calculate $S_{Y|X} = S_Y \sqrt{1 - R^2} \sqrt{\frac{n-1}{n-2}}$ and explain what it represents.

6. Select your second predictor and write out your new Full and Reduced Models**.** Complete the ANOVA summary table and verify your results with the omnibus F-ratio.

Reduced Model B: _____

Full Model B: _____

Source	SS	df	MS	MSR
X_1 and X_2				
$X_2 \mid X_1$				
Error				
Total				

Omnibus F-ratio: _____

7. Write out the prediction equation (a computer will calculate this) and interpret the coefficients.

Equation: _____

8. What do you conclude about this prediction equation? What do you conclude about the predictive potency of your second predictor?

9. Calculate $R^2_{2|1}$?

10. Select your third predictor and write out your new Full and Reduced Models**.** Complete the ANOVA summary table and verify your results with the omnibus F-ratio.

Reduced Model C: _____

Full Model C: _____

Source	SS	df	MS	MSR
X_1 and X_2 and X_3				
$X_3 \mid X_1, X_2$				
Error				
Total				

Omnibus F-ratio: _____

11. Write out the prediction equation (a computer will calculate this) and interpret the coefficients.

Equation: _____

12. What do you conclude about this prediction equation? What do you conclude about the predictive potency of your third predictor?

13. Calculate $R^2_{3|1,2}$?

14. We're going to do something different this time. Suppose we group our predictor variables into two categories:

    Easy-to-get information: Sex, Income

    More costly-to-get information: Status, VerbalIQ

    We really want to accurately predict the amount of money a teenager spends on gambling in one year, but we don't have unlimited resources. First, we'll see how well the easy-to-get information predicts gambling expenditures. Then, we'll see how much more accurate our prediction would be if we used the more costly information. From this, we'll decide on which predictors to keep.

    Write out the full and reduced models in this situation:

    Reduced Model D: _____

    Full Model D: _____

15. Complete the ANOVA summary table and conduct the omnibus F-test:

Source	SS	df	MS	MSR

    Omnibus F-ratio: _____

16. Stata computes the following regression equation. Write out the equation and predict the gambling expenses for a male with: status = 30, income = 3, and VerbalIQ = 7.

gambling	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
sex	-22.11833	8.211115	-2.69	0.010	-38.68903	-5.54763
status	.0522338	.2811115	0.19	0.853	-.5150722	.6195399
income	4.961979	1.025392	4.84	0.000	2.892654	7.031305
verbalIQ	-2.959494	2.17215	-1.36	0.180	-7.34307	1.424083
_cons	22.55565	17.1968	1.31	0.197	-12.1489	57.26021

Equation: _____

Predicted Gambling Expenditures: _____

17. According to the output, which coefficients are significantly different from zero? What does this mean?

18. According to the output shown above, which predictor is the most potent?

19. According to the following output, which predictor is the most potent?

gambling	Coef.	Std. Err.	t	P>|t|	Beta
sex	-22.11833	8.211115	-2.69	0.010	-.3481377
status	.0522338	.2811115	0.19	0.853	.0286113
income	4.961979	1.025392	4.84	0.000	.5591415
verbalIQ	-2.959494	2.17215	-1.36	0.180	-.1743399
_cons	22.55565	17.1968	1.31	0.197	.