Example of Building and Using a Bivariate Regression Model

In most prediction situations, we want to know the value of a variable that we don't have, either because that variable hasn't yet occurred (as in this example), because we can not afford to measure the variable, or because it is unethical to obtain the data.

However, we think we have a variable, called the predictor, that we can substitute for that difficult-to-obtain criterion. Using the predictor we can create a regression formula that allows us to predict the difficult-to-obtain criterion. Once we have the formula, we can estimate any person's criterion score from their predictor score.

Prediction is a *two step* process:

First, we must obtain one sample (called the **modeling sample)** that includes **both** the **criterion variable** and the **predictor variable(s)**. This will require patience, cost, or an ethical trade-off, depending upon what makes the criterion variable difficult to obtain. We will use this **modeling sample** to assess the utility of and build the linear regression model.

Second, we will have one or more samples (called the **application sample(s)**) for which we have only the predictor variable. We will use the linear regression model to estimate the criterion variable score for the people in this/these sample(s).

Example of Selecting, Building and Using a Simple Linear Regression

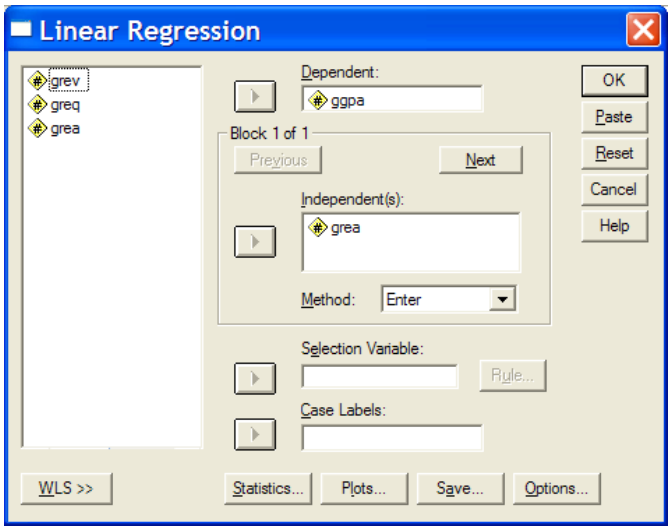
The graduate faculty decided that they only wanted to admit students who were likely to have 1st year GPA of at least 3.500. Naturally, one can't know a student's 1st year GPA before they are admitted, so ... "This is a job for linear regression!!" Looking over the records for the last couple of years, a statistician was able to compile a data base of 30 students for which their GRE and 1st year GPA data were available. This comprised the "modeling sample". The decision was to use the GREA (Analytic) score as the predictor and to construct the regression model to predict 1st year GPA.

Step #1 -- Building the Regression model using the Modeling Sample

SPSS

Analyze → Regression → Linear

- Move criterion variable into "Dependent" window
- Move predictor variable into "Independent(s)" window



Syntax

```
REGRESSION
  /STATISTICS COEFF OUTS R ANOVA
  /DEPENDENT ggpa
  /METHOD=ENTER grea.
```

SPSS Output: SPSS Output:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.611[a]	.374	.351	.4832

a. Predictors: (Constant), Analytic subscore of GRE

We can see that there is a significant and substantial correlation between the predictor and the criterion -- looks like a good basis for linear regression

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3.897	1	3.897	16.693	.000[a]
	Residual	6.537	28	.233		
	Total	10.435	29			

a. Predictors: (Constant), Analytic subscore of GRE

b. Dependent Variable: 1st year graduate gpa -- criterion variable

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-.949	1.047		-.907	.372
	Analytic subscore of GRE	7.540E-03	.002	.611	4.086	.000

a. Dependent Variable: 1st year graduate gpa -- criterion variable

Reminder -- 7.540E-.03 is exponential notation for .007450 (R-03) means move the decimal three places to the left)

From this we would construct the regression formula y' = bx + a = gpa' = (.007540 * grea) - .949485

Now the faculty has a model. The next year, they build a data base that contains the GRE scores for each of the 15 applicants.

Example Write-up:

Students with higher GREA scores tended to have a higher first year graduate GPA, r(30) = .661, p < .001. The resulting regression model was : gpa' = (.007540 * grea) - .949485, indicating that GGPA is expected to increase by .0075 for each 1-point increase in GREA, or to increase by .75 for each 100-pint increase.

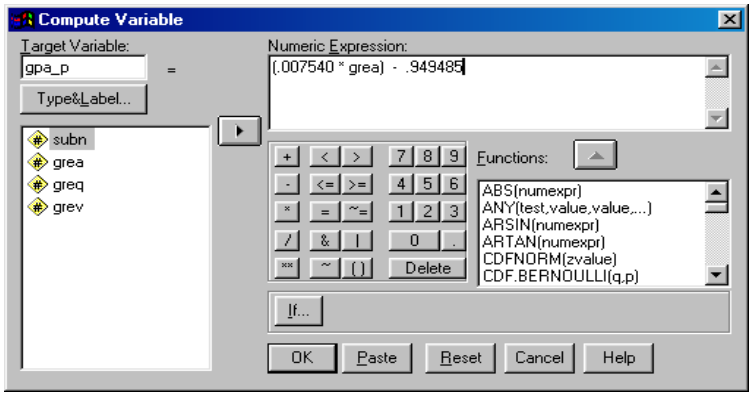
Table 1 Univariate statistcs and correlations (N = 30).

Variable	Mean	Std Dev
GPA	3.3133	.5998
GREA	565.3333	48.6177

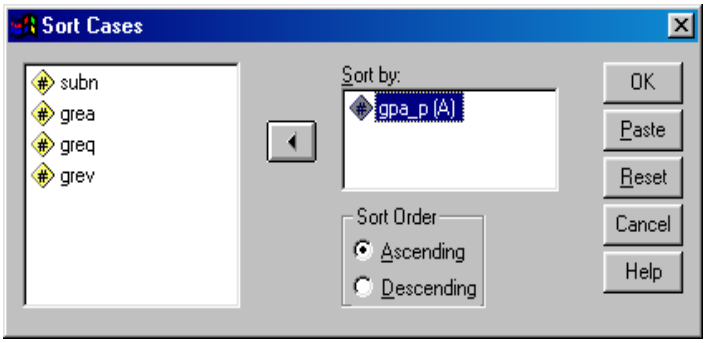
Step #2 -- Applying the regression formula to a group of applicants

* **Note: This is a new dataset including only applicant's GRE scores (no GPA, since they're applicants)**

Transform → Compute

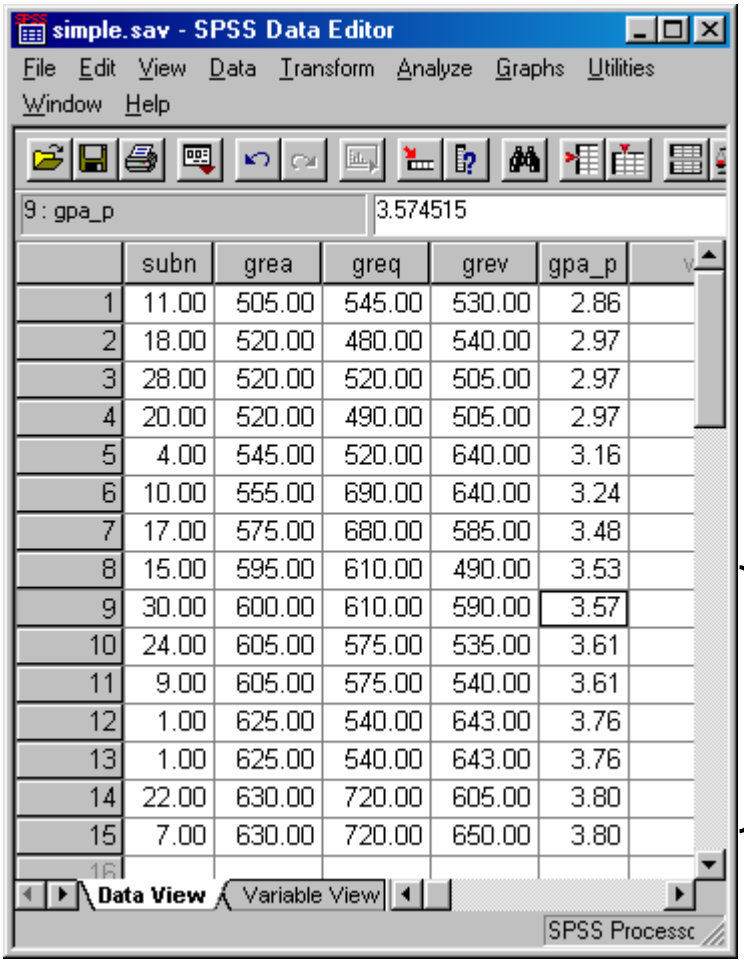


Data → Sort Cases



COMPUTE gpa_p = (007540 * grea) - .949485.
EXE.

SORT CASES BY ggpa.



	subn	grea	greq	grev	gpa_p
1	11.00	505.00	545.00	530.00	2.86
2	18.00	520.00	480.00	540.00	2.97
3	28.00	520.00	520.00	505.00	2.97
4	20.00	520.00	490.00	505.00	2.97
5	4.00	545.00	520.00	640.00	3.16
6	10.00	555.00	690.00	640.00	3.24
7	17.00	575.00	680.00	585.00	3.48
8	15.00	595.00	610.00	490.00	3.53
9	30.00	600.00	610.00	590.00	3.57
10	24.00	605.00	575.00	535.00	3.61
11	9.00	605.00	575.00	540.00	3.61
12	1.00	625.00	540.00	643.00	3.76
13	1.00	625.00	540.00	643.00	3.76
14	22.00	630.00	720.00	605.00	3.80
15	7.00	630.00	720.00	650.00	3.80

Notice there is no "GPA" column, because these are applicants who have not completed any graduate classes -- if they had we wouldn't need "predicted gpa" scores for them!

Notice that the rank order for GREA and gpa_p are the same! This will always happen, because the regression formula involves only linear transformations, which will not change the rank order of a set of scores.

If we wanted to select students who were likely to have a 1st Year GPA of at least 3.5, we would select these folks.

But consider the slim difference between the predicted values of folks # 17 and #15 – especially considering the SEE (.4832).

Keep in mind that while correlation analyses are about hypotheses, theories, and other ***ideas***, regression analyses used for selection purposes are about ***people*** and decisions that influence their lives!!