Multiple Linear Regression

The phrase '**simple linear regression'** is used to indicate there is <u>one</u> predictor in the linear regression equation.

The phrase '**multiple linear regression'** is used to indicate there is <u>more than one</u> predictor in the linear regression equation.

Simple Linear Regression	**Multiple Linear Regression**
❑ One predictor in the equation.	❑ Two or more predictors in the equation.
❑ Unstandardized regression coefficient represents the marginal relationship between Y and X_1.	❑ Unstandardized regression coefficient represents the partial relationship between Y and X_1 after controlling for X_2.
❑ Standardized regression coefficient equals the zero-order correlation.	❑ Standardized regression coefficient <u>does not</u> equal the zero-order correlation unless all the predictors are uncorrelated.

Linear Regression Equation for a Two Predictor Model.

$$Y = b_0 + b_1 X_1 + b_2 X_2 + e$$
$$Y = \hat{Y} + e$$

where

Y = person's score on the dependent variable
$b_0 = Y$ intercept, the value of Y when $X_1 = 0$ and $X_2 = 0$
b_1 = partial regression coefficient, relating Y and X_1 after controlling for X_2
X_1 = person's score on the first predictor
b_2 = partial regression coefficient, relating Y and X_2 after controlling for X_1
X_2 = person's score on the second predictor
e = residual = prediction error for the i^{th} person
$\hat{Y}$ = predicted value for the i^{th} person

All factors, other than X_1 and X_2, that influence Y, are subsumed under the prediction error. Thus, e represents the variability in Y that is not explained or predicted by X_1 and X_2.

The predicted value of Y represents the variability in Y that can be explained or predicted by X_1 and X_2.

Least-squares Estimators of the Linear Regression Equation for a Two Predictor Model

Unstandardized Regression Coefficients (calculated from raw scores)	**Standardized Regression Coefficients** (calculated from Z scores)
$b_0 = \overline{Y} - b_1\overline{X}_1 - b_2\overline{X}_2$	$b_0 = \overline{Z}_Y - b_1\overline{Z}_1 - b_2\overline{Z}_2 = 0$
$b_1 = \frac{(r_{Y1} - r_{Y2}r_{12})S_Y}{(1-r_{12}^2)S_1}$	$b_1 = \frac{(r_{Y1} - r_{Y2}r_{12})}{(1-r_{12}^2)}$
$b_2 = \frac{(r_{Y2} - r_{Y1}r_{12})S_Y}{(1-r_{12}^2)S_2}$	$b_2 = \frac{(r_{Y2} - r_{Y1}r_{12})}{(1-r_{12}^2)}$

Characteristics of <u>Least-squares Estimators</u> include:

- The least-squares coefficients (b_0, b_1, b_2) cannot be calculated if one of the predictors has no variability.
- The least-squares coefficients (b_0, b_1, b_2) cannot be calculated if two or more predictors are perfectly correlated. If two predictors are perfectly correlated, they are said to be **collinear**.
- The least-squares residuals are uncorrelated with the predictors (independent variables).
- The least-squares residuals are uncorrelated with the predicted values, $\hat{Y}_i$.
- When the predictors are uncorrelated, the standardized regression coefficients equal the zero-order correlations of each predictor with the dependent variable.
- The standardized regression coefficients may have values greater than |1|.
- The unstandardized regression coefficients may have values greater than |1|.

Interpretation of the Least-squares Estimators:

- The unstandardized regression coefficient, b_1, represents the average change in the dependent variable, Y, for a one unit increase in X_1 when the other predictors are held constant.
- The standardized regression coefficient represents the average standard deviation change in Y for a one standard deviation increase in X_1 when other predictors are held constant.

How well does the multiple linear regression equation fit the data?

Variance of the Residuals = Mean Squared Error = Average "Squared" prediction error of the linear regression equation.

$$MSE = \frac{\sum e_i^2}{N - p - 1} = \frac{\left(1 - R^2\right) SS_Y}{N - p - 1}$$

- It is measured in squared units of the dependent variable.

Standard Deviation of the Residuals = Standard Error = Average prediction error of the linear regression equation.

$$S_E = \sqrt{MSE}$$

- It is measured in the units of the dependent variable.
- If the residuals are approximately normally distributed, then about 2/3 of the residuals are in the range ± 1(S_E), and about 95% are in the range ± 2(S_E).

Squared Multiple Correlation, R^2. The squared correlation involves a comparison of two models.

Reduced Model: $Y_i = b_0 + e_i$ where b_0 would be defined as $\overline{Y}$.

$$SS_Y = \sum e_i^2 = \sum \left(Y_i - \overline{Y}\right)^2$$

Full Model: $Y = b_0 + b_1 X_1 + b_2 X_2 + e$ where the regression coefficients are defined above.

$$SS_{\text{Re}\,sidual} = \sum e_i^2 = \sum \left(Y_i - \hat{Y}\right)^2$$

The difference between the total sum of squares for *Y* and the residual sum of squares for *Y* is called the **regression sum of squares**. The regression sum of squares gives the reduction in squared error due to the linear regression (i.e., using X_1 and X_2 to predict *Y*).

$$SS_{Regression} = SS_Y - SS_{Residual}$$

The $\boldsymbol{R^2}$ provides a relative measure of fit. It is the proportion of total variation in Y that can be predicted from the variability in X_1 and X_2. The value of $\boldsymbol{R^2}$ ranges between 0 and 1.

$$R^2 = \frac{\text{Variability in Y explained by } X_1 \text{ and } X_2}{\text{Total Variability in Y}} = \frac{\text{Regression SS}}{\text{Total SS for Y}} = \frac{SS_{\text{Regression}}}{SS_Y}$$

If the p predictors were uncorrelated with each other, the Coefficient of Determination, $\boldsymbol{R^2}$**,** could be calculated as:

$$R^2 = r_{Y1}^2 + r_{Y2}^2 + \ldots + r_{YP}^2$$

Since the p predictors are likely to be correlated at the sample level, the Coefficient of Determination, $\boldsymbol{R^2}$**,** must adjust for the correlation among the predictors:

$$R^2 = r_{Y1}b_1 + r_{Y2}b_2 + \ldots + r_{YP}b_P \text{ where } b = \text{standardized regression coefficients}$$

Multiple Correlation, *R*. It is the positive correlation between the actual Y scores and the predicted Y values, $\hat{Y} = b_0 + b_1X_1 + b_2X_2$. The value of $\boldsymbol{R}$ ranges between 0 and 1.

$$R = \sqrt{R^2}$$

Procedures for Estimating the Shrinkage of R^2

- Adjusted R^2
- Cross-Validation
- Data Splitting
- Formula-based Estimation of the Cross-Validity Coefficient

Adjusted R^2. The sample based R^2 is a positively biased estimator of the population coefficient. The Adjusted R^2 is one type of attempt to remove the positive bias.

$$Adj.R^2 = 1 - \left(1 - R^2\right)\left[\frac{N-1}{N-p-1}\right]$$

The adjusted R^2 modifies the R^2 to take the sample size and number of predictors into account because the degree of overestimation of R^2 is affected by the ratio of the number of predictors to the size of the sample. Other things equal, the larger this ratio, the greater the overestimation of R.

- Adding a predictor to the model does not hurt the $\mathbf{R^2}$; it can only increase it. Basically, you should always use & report the **adjusted $\mathbf{R^2}$**.
- When a selection procedure (forward, backward, or stepwise) is used, the R^2 is even more overestimated because of the capitalization on chance.

Cross-Validation

- Obtain two samples from the same population.
  Fall 1996 students, measure their Anxiety, Sleep, and Caffeine

  Fall 1997 students, measure their Anxiety, Sleep, and Caffeine
- The first sample (Fall 1996 students) is analyzed using multiple linear regression to obtain a prediction equation.
- The predictor scores (Sleep & Caffeine) from the second sample are used with the regression equation from the first sample to obtain predicted Anxiety of people in the second sample.

Fall 1996			Fall 1997			
X_1 Sleep	X_2 Caffeine	Y Anxiety	X_1 Sleep	X_2 Caffeine	Y Anxiety	$\hat{Y} = 19.17 - 0.34X_1 + 1.05X_2$
7	1	18	8	2	16	18.55
7	5	21	6	1	19	18.18
13	3	18	4	1	19	18.86
7	5	23	11	3	20	18.58
12	4	18	14	5	23	19.66
10	5	21	6	1	18	18.18
15	5	20	4	2	19	19.91
9	4	21	2	0	16	18.49
$\overline{X}_1 = 10$	$\overline{X}_2 = 4$	$\overline{Y} = 20$				Cross-Validity Coefficient $cor(Y, \hat{Y}) = .5395$

$\hat{Y} = 19.17 - 0.34X_1 + 1.05X_2$

- The Cross-Validity Coefficient is defined as the correlation between the actual Anxiety scores for the Fall 1997 students and the predicted Anxiety scores for the Fall 1997 students using the Fall 1996 regression equation.
- If the difference between the R^2 from the first sample (Fall 1996 students) and the **SQUARED cross-validity coefficient** from the second sample (Fall 1997 students) is small, the two samples can be combined and the regression equation for the combined samples can be used in future predictions.

For our example, R^2 = .8068 **and the** Squared Cross-Validity Coefficient = .2911**. The discrepancy is very large, so we would not use the prediction equation for future predictions.**

What is a large or small change? .10 or less

```
COMPUTE filter_$=(sample=0).
FILTER BY filter_$.
EXECUTE .
REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT response
  /METHOD=ENTER height weight positive  .
```

Regression (SAMPLE = 0)

Descriptive Statistics

	Mean	Std. Deviation	N
No. of Responses Received	9.0372	2.94580	188
Height of Person Given in Ad	66.5106	4.37725	188
Weight of Person Given in Ad	139.6170	7.10894	188
No. of Positive Adjectives in Ad	5.0798	1.46584	188

Correlations

		No. of Responses Received	Height of Person Given in Ad	Weight of Person Given in Ad	No. of Positive Adjectives in Ad
Pearson Correlation	No. of Responses Received	1.000	.444	-.301	.568
	Height of Person Given in Ad	.444	1.000	-.188	.252
	Weight of Person Given in Ad	-.301	-.188	1.000	-.133
	No. of Positive Adjectives in Ad	.568	.252	-.133	1.000
Sig. (1-tailed)	No. of Responses Received	.	.000	.000	.000
	Height of Person Given in Ad	.000	.	.005	.000
	Weight of Person Given in Ad	.000	.005	.	.035
	No. of Positive Adjectives in Ad	.000	.000	.035	.
N	No. of Responses Received	188	188	188	188
	Height of Person Given in Ad	188	188	188	188
	Weight of Person Given in Ad	188	188	188	188
	No. of Positive Adjectives in Ad	188	188	188	188

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	No. of Positive Adjectives in Ad, Weight of Person Given in Ad, Height of Person Given in Ad[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: No. of Responses Received

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.672[a]	.452	.443	2.19910

a. Predictors: (Constant), No. of Positive Adjectives in Ad, Weight of Person Given in Ad, Height of Person Given in Ad

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	732.904	3	244.301	50.517	.000[a]
	Residual	889.835	184	4.836		
	Total	1622.739	187			

a. Predictors: (Constant), No. of Positive Adjectives in Ad, Weight of Person Given in Ad, Height of Person Given in Ad

b. Dependent Variable: No. of Responses Received

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	1.8385	4.430		.415	.679	-6.902	10.579
	Height of Person Given in Ad	.1961	.038	.291	5.099	.000	.120	.272
	Weight of Person Given in Ad	-.0762	.023	-.184	-3.297	.001	-.122	-.031
	No. of Positive Adjectives in Ad	.9444	.114	.470	8.297	.000	.720	1.169

a. Dependent Variable: No. of Responses Received

```
IF (sample=1) predresp = 1.8385+.1961*height-.0762*weight+.9444*positive .
EXECUTE .

COMPUTE filter_$=(sample=1).
FILTER by filter_$.
EXECUTE.

CORRELATIONS
  /VARIABLES=predresp response
  /PRINT=TWOTAIL NOSIG
  /STATISTICS DESCRIPTIVES
  /MISSING=PAIRWISE .
```

Correlations

Descriptive Statistics

	Mean	Std. Deviation	N
Predicted # Responses	8.8897	2.03717	207
No. of Responses Received	8.8889	2.78209	207

Correlations

		Predicted # Responses	No. of Responses Received
Predicted # Responses	Pearson Correlation	1	.651**
	Sig. (2-tailed)	.	.000
	N	207	207
No. of Responses Received	Pearson Correlation	.651**	1
	Sig. (2-tailed)	.000	.
	N	207	207

**. Correlation is significant at the 0.01 level (2-tailed).

```
FILTER OFF.
USE ALL.
EXECUTE .
REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT response
  /METHOD=ENTER height weight positive  .
```

Regression (All DATA)

Descriptive Statistics

	Mean	Std. Deviation	N
No. of Responses Received	8.9595	2.85849	395
Height of Person Given in Ad	66.5190	4.29514	395
Weight of Person Given in Ad	139.9013	6.67923	395
No. of Positive Adjectives in Ad	5.0177	1.52715	395

Correlations

		No. of Responses Received	Height of Person Given in Ad	Weight of Person Given in Ad	No. of Positive Adjectives in Ad
Pearson Correlation	No. of Responses Received	1.000	.437	-.239	.582
	Height of Person Given in Ad	.437	1.000	-.139	.236
	Weight of Person Given in Ad	-.239	-.139	1.000	-.200
	No. of Positive Adjectives in Ad	.582	.236	-.200	1.000
Sig. (1-tailed)	No. of Responses Received	.	.000	.000	.000
	Height of Person Given in Ad	.000	.	.003	.000
	Weight of Person Given in Ad	.000	.003	.	.000
	No. of Positive Adjectives in Ad	.000	.000	.000	.
N	No. of Responses Received	395	395	395	395
	Height of Person Given in Ad	395	395	395	395
	Weight of Person Given in Ad	395	395	395	395
	No. of Positive Adjectives in Ad	395	395	395	395

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	No. of Positive Adjectives in Ad, Weight of Person Given in Ad, Height of Person Given in Ad[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: No. of Responses Received

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.665[a]	.442	.438	2.14301

a. Predictors: (Constant), No. of Positive Adjectives in Ad, Weight of Person Given in Ad, Height of Person Given in Ad

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1423.679	3	474.560	103.333	.000[a]
	Residual	1795.672	391	4.593		
	Total	3219.352	394			

a. Predictors: (Constant), No. of Positive Adjectives in Ad, Weight of Person Given in Ad, Height of Person Given in Ad

b. Dependent Variable: No. of Responses Received

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B	
Model		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	-3.3899	3.051		-1.111	.267	-9.388	2.608
	Height of Person Given in Ad	.2047	.026	.308	7.875	.000	.154	.256
	Weight of Person Given in Ad	-.0419	.017	-.098	-2.527	.012	-.074	-.009
	No. of Positive Adjectives in Ad	.9159	.074	.489	12.397	.000	.771	1.061

a. Dependent Variable: No. of Responses Received

Data Splitting (see copy of Pedhazur, p. 210)

If you have a large sample, e.g., N ≅ 500, then you can split the data into two random samples. Then, one sample may be used to find the prediction equation and the other may be used to calculate the cross-validity coefficient. If the difference between the R^2 and the squared cross-validity coefficient is small, the two samples are combined and prediction equation for the combined samples is used for future predictions.

Formula-based Estimation of the Cross-Validity Coefficient

…Assuming Predictors are Random Variables

$$R^2_{adjRP} = 1 - \left(\frac{N-1}{N-p-1}\right)\left(\frac{N-2}{N-p-2}\right)\left(\frac{N+1}{N}\right)\left(1-R^2\right)$$

…Assuming Predictors are Fixed Variables (unrealistic)

$$R^2_{adjRP} = 1 - \left(\frac{N-1}{N}\right)\left(\frac{N+p+1}{N-p-1}\right)\left(1-R^2\right)$$

Confidence Interval for R^2

$R^2 + 1.96\sqrt{\frac{4}{N}R^2\left(1-R^2\right)\left[1-\frac{(2p+1)}{N}\right]^2}$ is the upper 95% confidence limit

$R^2 - 1.96\sqrt{\frac{4}{N}R^2\left(1-R^2\right)\left[1-\frac{(2p+1)}{N}\right]^2}$ is the lower 95% confidence limit

Conducting a Test of the Null Hypothesis that All Regression Coefficients of a Multiple Regression Model are 0.

Null Hypothesis:

Alternative Hypothesis:

By Hand	**Using SPSS**
Choose the Alpha Level: **Numerator df (*p*) =** **Denominator df (*N* – *p* – 1) =** **Critical Value:** **Calculate *F*** $F = \frac{\left(R^2 SS_Y\right)/p}{\left(1-R^2\right)SS_Y/\left(N-p-1\right)} = \frac{MS_{\text{Regression}}}{MSE}$ **or** $F = \frac{R^2/p}{\left(1-R^2\right)/\left(N-p-1\right)}$	**Choose the Alpha Level:** **<u>Identify…</u>** Numerator *df* Denominator *df* *F* Value *MSE* ***significance level***
Determine the significance by comparing *F* that was calculated to the *F* critical value. If the *F* that was calculated is equal or greater than the *F* critical value, then reject the null hypothesis.	**Determine the significance** by comparing the significance level to the alpha level. If the significance level is less than or equal to the alpha level, then reject the null hypothesis.

Interpretation:

Confidence Intervals & Hypothesis Tests for Individual Regression Coefficients of a Multiple Regression Model

Testing the Null Hypothesis: The regression coefficient is 0. The null also can be stated as "The variable is not a significant predictor of the dependent variable when controlling for the other predictors in the model."

By Hand	**Using SPSS**
Choose the Alpha Level: **Denominator df ($N - p - 1$) =** **Critical Values:** **Calculate the Standard Error for the Predictor:** $SE(b_j) = \left(\sqrt{\text{Variance Inflation Factor}}\right)\frac{\text{Standard Error of the Estimate}}{\sqrt{\text{SS}_\text{X}}}$ $= \left(\sqrt{\frac{1}{1-R_j^2}}\right)\frac{S_E}{\sqrt{SS_X}}$ **Calculate the *t* value for the Predictor:** $t = \frac{\text{Sample Regression Coefficient - Hypothesized Null Value}}{\text{Estimated Standard Error of the Regression Coefficient}} = \frac{b-0}{SE(b)}$	**Choose the Alpha Level:** **<u>Identify…</u>** Denominator *df* *t* Value ***significance level***
Determine the significance by comparing t that was calculated to the *t* critical values. If the |*t*| that was calculated is equal or greater than the *t* critical values, then reject the null hypothesis.	**Determine the significance** by comparing the significance level to the alpha level. If the significance level is less than or equal to the alpha level, then reject the null hypothesis.

Interpretation:

Calculating a Confidence Interval for the Regression Coefficient

$$\text{Regression Coefficient} = \text{Sample Regression Coefficient} \pm t_{crit}\left(\text{Estimated Standard Error of the Regression Coefficient}\right)$$
$$= b \pm t_{crit} SE(b)$$

Interpretation:

Correlations

Descriptive Statistics

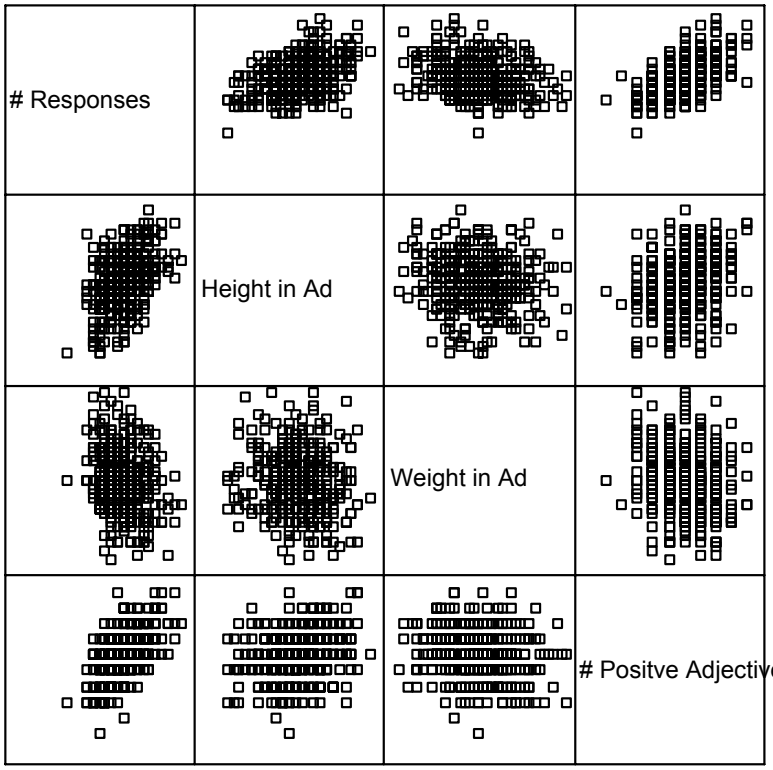
	Mean	Std. Deviation	N
# Responses	8.9595	2.85849	395
Height in Ad	66.5190	4.29514	395
Weight in Ad	139.9013	6.67923	395
# Positve Adjectives	5.0177	1.52715	395

Correlations

		# Responses	Height in Ad	Weight in Ad	# Positve Adjectives
# Responses	Pearson Correlation	1	.437**	-.239**	.582**
	Sig. (2-tailed)	.	.000	.000	.000
	N	395	395	395	395
Height in Ad	Pearson Correlation	.437**	1	-.139**	.236**
	Sig. (2-tailed)	.000	.	.006	.000
	N	395	395	395	395
Weight in Ad	Pearson Correlation	-.239**	-.139**	1	-.200**
	Sig. (2-tailed)	.000	.006	.	.000
	N	395	395	395	395
# Positve Adjectives	Pearson Correlation	.582**	.236**	-.200**	1
	Sig. (2-tailed)	.000	.000	.000	.
	N	395	395	395	395

**. Correlation is significant at the 0.01 level (2-tailed).

Graph



Regression

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	# Positve Adjectives, Weight in Ad, Height in Ad[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: # Responses

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.665[a]	.442	.438	2.14301	1.839

a. Predictors: (Constant), # Positve Adjectives, Weight in Ad, Height in Ad

b. Dependent Variable: # Responses

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1423.679	3	474.560	103.333	.000[a]
	Residual	1795.672	391	4.593		
	Total	3219.352	394			

a. Predictors: (Constant), # Positve Adjectives, Weight in Ad, Height in Ad

b. Dependent Variable: # Responses

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-3.390	3.051		-1.111	.267	-9.388	2.608
	Height in Ad	.205	.026	.308	7.875	.000	.154	.256
	Weight in Ad	-4.19E-02	.017	-.098	-2.527	.012	-.074	-.009
	# Positve Adjectives	.916	.074	.489	12.397	.000	.771	1.061

a. Dependent Variable: # Responses

Casewise Diagnostics[a]

Case Number	Std. Residual	# Responses
227	3.308	17.00

a. Dependent Variable: # Responses

Residuals Statistics[a]

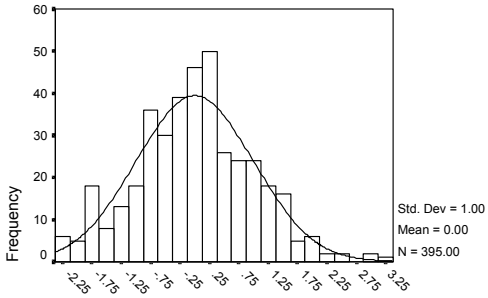
	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	3.8242	14.7527	8.9595	1.90089	395
Std. Predicted Value	-2.701	3.048	.000	1.000	395
Standard Error of Predicted Value	.10863	.40496	.20661	.06187	395
Adjusted Predicted Value	3.8201	14.7460	8.9604	1.90141	395
Residual	-4.8329	7.0884	.0000	2.13484	395
Std. Residual	-2.255	3.308	.000	.996	395
Stud. Residual	-2.258	3.314	.000	1.001	395
Deleted Residual	-4.8466	7.1155	-.0009	2.15615	395
Stud. Deleted Residual	-2.270	3.357	.000	1.004	395
Mahal. Distance	.015	13.072	2.992	2.463	395
Cook's Distance	.000	.044	.003	.005	395
Centered Leverage Value	.000	.033	.008	.006	395

a. Dependent Variable: # Responses

Charts

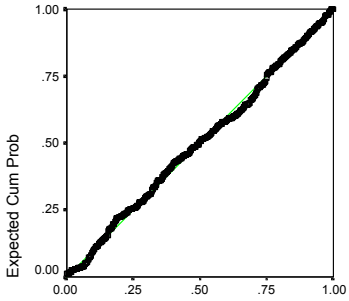
Histogram

Dependent Variable: # Responses



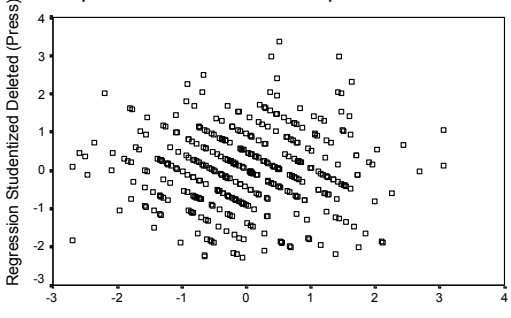
Normal P-P Plot of Regression Stand

Dependent Variable: # Responses



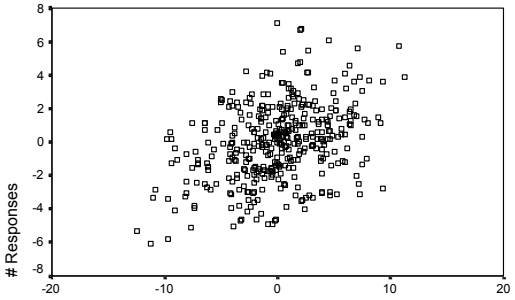
Scatterplot

Dependent Variable: # Responses



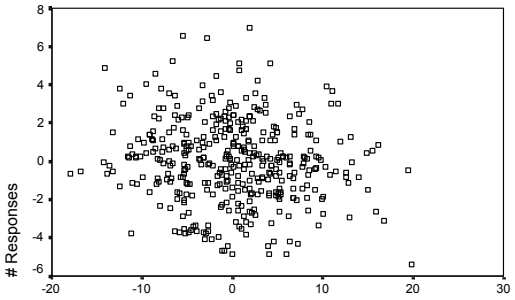
Partial Regression Plot

Dependent Variable: # Responses



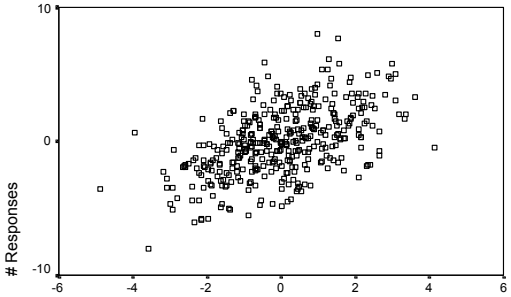
Partial Regression Plot

Dependent Variable: # Responses



Partial Regression Plot

Dependent Variable: # Responses



Correlations

Descriptive Statistics

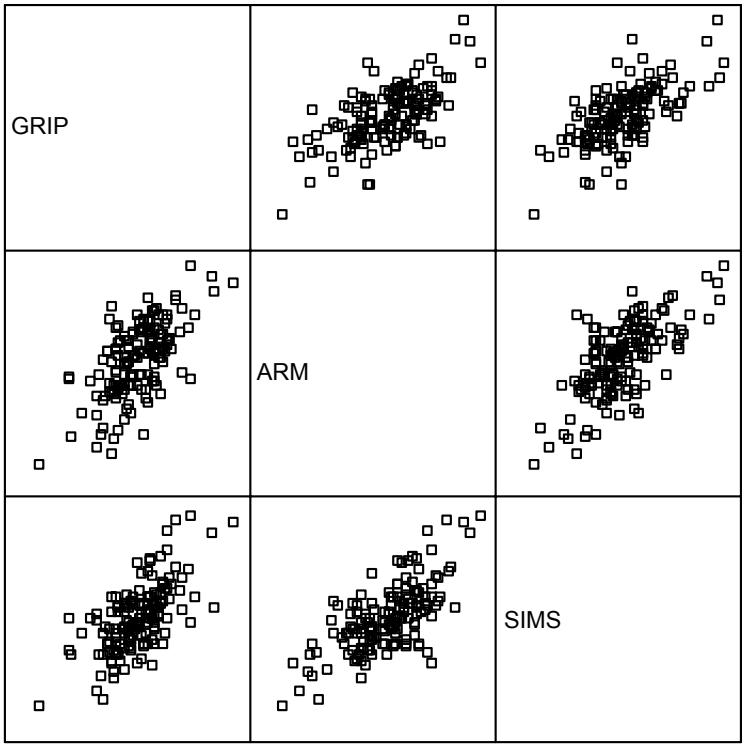
	Mean	Std. Deviation	N
GRIP	110.231	23.6299	147
ARM	78.752	21.1093	147
SIMS	.2018	1.67897	147

Correlations

		GRIP	ARM	SIMS
GRIP	Pearson Correlation	1	.630**	.640**
	Sig. (2-tailed)	.	.000	.000
	N	147	147	147
ARM	Pearson Correlation	.630**	1	.686**
	Sig. (2-tailed)	.000	.	.000
	N	147	147	147
SIMS	Pearson Correlation	.640**	.686**	1
	Sig. (2-tailed)	.000	.000	.
	N	147	147	147

**. Correlation is significant at the 0.01 level (2-tailed).

Graph



Regression

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	ARM, GRIP[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: SIMS

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.736[a]	.542	.536	1.14392	1.938

a. Predictors: (Constant), ARM, GRIP

b. Dependent Variable: SIMS

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	223.136	2	111.568	85.261	.000[a]
	Residual	188.431	144	1.309		
	Total	411.567	146			

a. Predictors: (Constant), ARM, GRIP

b. Dependent Variable: SIMS

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-5.434	.462		-11.77	.000	-6.347	-4.521
	GRIP	2.447E-02	.005	.344	4.744	.000	.014	.035
	ARM	3.731E-02	.006	.469	6.462	.000	.026	.049

a. Dependent Variable: SIMS

Residuals Statistics[a]

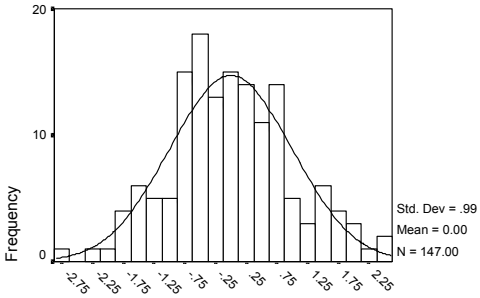
	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	-4.0153	3.7429	.2018	1.23626	147
Std. Predicted Value	-3.411	2.864	.000	1.000	147
Standard Error of Predicted Value	.09449	.34842	.15524	.05123	147
Adjusted Predicted Value	-3.9995	3.6409	.2008	1.23630	147
Residual	-3.1846	2.8634	.0000	1.13606	147
Std. Residual	-2.784	2.503	.000	.993	147
Stud. Residual	-2.818	2.524	.000	1.004	147
Deleted Residual	-3.2637	2.9109	.0010	1.16133	147
Stud. Deleted Residual	-2.889	2.573	.001	1.011	147
Mahal. Distance	.003	12.551	1.986	2.223	147
Cook's Distance	.000	.100	.007	.014	147
Centered Leverage Value	.000	.086	.014	.015	147

a. Dependent Variable: SIMS

Charts

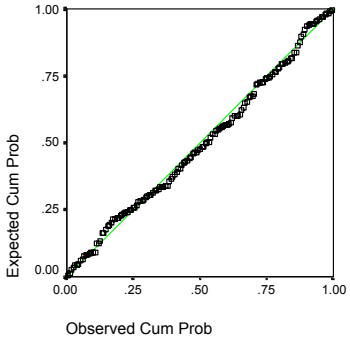
Histogram

Dependent Variable: SIMS



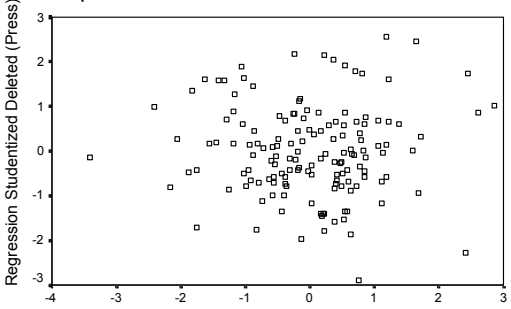
Normal P-P Plot of Regression Stand

Dependent Variable: SIMS



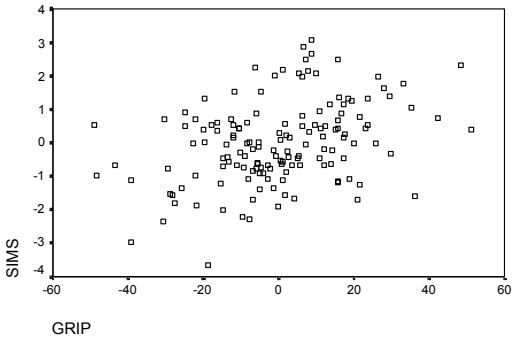
Scatterplot

Dependent Variable: SIMS



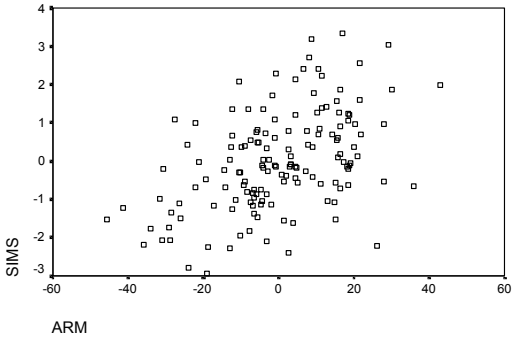
Partial Regression Plot

Dependent Variable: SIMS



Partial Regression Plot

Dependent Variable: SIMS



Selecting the Best Regression Equation for Prediction

- In predictive research the main emphasis is on practical applications, whereas in explanatory research the main emphasis is on understanding phenomena.
- Behavioral researchers frequently taken the results of purely predictive studies and use them for the purpose of explaining phenomena. This is inappropriate!
- When appropriately used, regression analysis in predictive research poses few difficulties in interpretation. It is the use and interpretation of regression analysis in explanatory research that is fraught with ambiguities and potential misinterpretations (Pedhazur, 1997, p. 198).

Methods for Selecting the Best Predictors

- Theoretical Basis
- All Subsets Regression
- Backward Elimination
- Forward Selection
- Stepwise Selection (hybrid of backward elimination & forward selection)
- Chunkwise or Blockwise Selection

If you have 10 or fewer predictors and are using SAS, I recommend inspecting the all subsets regression output to find a good place to start for selecting the best predictors.

Most people would recommend using the stepwise selection procedure with a liberal alpha. Then, you would consider revising the model based on common sense, etc. The stepwise procedure might be used with groups of predictors (i.e., chunkwise).

SAS Regression Notes *recommends*

⇒ *Simpler regression models are preferred*. Therefore, consider models with fewer variables unless a more complex model contains important variables.

⇒ *Practical considerations are important*. For example, certain variables may be expensive or impractical to measure.

⇒ *For each candidate model, you should examine the data for unusual observations and verify inference assumptions, which can help eliminate certain candidate models*. If a particular model of a certain size is not appropriate, you can consider other models of the same size.

⇒ *With model-selection methods, there is a tendency to overspecify the model.* Recall that a model is overspecified if the model, on average, fits the data better than the true regression curve.

Predicting Anger Expression Scores

All Possible Subsets Procedure (RSQUARE on SAS)

- **Used when you have a *small number* of predictors in your maximum model.**
- **Calculates all possible models for your predictors, 2^k.**

k = # Predictors	# Models Compared (including intercept only model)
3	8
5	32
7	128

The REG Procedure
Model: MODEL1
Dependent Variable: EXPRESS

R-Square Selection Method

Number in Model	R-Square	Variables in Model
1	0.1502	FRUSTRAT
1	0.0659	LIFE
1	0.0552	AGESELF
1	0.0517	FEEL
1	0.0315	STRANGE
1	0.0036	AGEOTHER
1	0.0000	FRIENDS
2	0.2081	FEEL FRUSTRAT
2	0.1986	AGESELF FRUSTRAT
2	0.1593	FRUSTRAT LIFE
2	0.1564	STRANGE FRUSTRAT
2	0.1502	AGEOTHER FRUSTRAT
2	0.1502	FRIENDS FRUSTRAT
2	0.1312	AGESELF LIFE
2	0.1308	FEEL LIFE
2	0.1115	FEEL AGESELF
2	0.1026	STRANGE LIFE
2	0.0859	STRANGE FEEL
2	0.0701	AGEOTHER LIFE
2	0.0659	FRIENDS LIFE
2	0.0627	STRANGE AGESELF
2	0.0623	AGEOTHER FEEL
2	0.0554	AGEOTHER AGESELF
2	0.0552	FRIENDS AGESELF
2	0.0534	FRIENDS FEEL
2	0.0331	STRANGE FRIENDS
2	0.0323	STRANGE AGEOTHER
2	0.0036	FRIENDS AGEOTHER
3	0.2609	FEEL AGESELF FRUSTRAT
3	0.2224	FEEL FRUSTRAT LIFE
3	0.2153	STRANGE FEEL FRUSTRAT
3	0.2129	AGESELF FRUSTRAT LIFE
3	0.2109	AGEOTHER FEEL FRUSTRAT
3	0.2097	FRIENDS FEEL FRUSTRAT
3	0.2030	FEEL AGESELF LIFE
3	0.1997	AGEOTHER AGESELF FRUSTRAT
3	0.1988	STRANGE AGESELF FRUSTRAT
3	0.1986	FRIENDS AGESELF FRUSTRAT
3	0.1714	STRANGE FEEL LIFE
3	0.1693	STRANGE FRUSTRAT LIFE
3	0.1596	AGEOTHER FRUSTRAT LIFE
3	0.1593	FRIENDS FRUSTRAT LIFE
3	0.1568	STRANGE FRIENDS FRUSTRAT
3	0.1565	STRANGE AGEOTHER FRUSTRAT
3	0.1502	FRIENDS AGEOTHER FRUSTRAT
3	0.1438	AGEOTHER FEEL LIFE
3	0.1400	STRANGE AGESELF LIFE
3	0.1332	FRIENDS FEEL LIFE
3	0.1315	AGEOTHER AGESELF LIFE
3	0.1313	FRIENDS AGESELF LIFE
3	0.1197	STRANGE FEEL AGESELF

```
       3       0.1149    AGEOTHER FEEL AGESELF
       3       0.1131    FRIENDS FEEL AGESELF
       3       0.1042    STRANGE FRIENDS LIFE
       3       0.1035    STRANGE AGEOTHER LIFE
       3       0.0909    STRANGE AGEOTHER FEEL
       3       0.0859    STRANGE FRIENDS FEEL
       3       0.0701    FRIENDS AGEOTHER LIFE
       3       0.0635    FRIENDS AGEOTHER FEEL
       3       0.0632    STRANGE FRIENDS AGESELF
       3       0.0628    STRANGE AGEOTHER AGESELF
       3       0.0554    FRIENDS AGEOTHER AGESELF
       3       0.0340    STRANGE FRIENDS AGEOTHER
-------------------------------------------------------------------
       4       0.2822    FEEL AGESELF FRUSTRAT LIFE
       4       0.2625    FRIENDS FEEL AGESELF FRUSTRAT
       4       0.2611    STRANGE FEEL AGESELF FRUSTRAT
       4       0.2611    AGEOTHER FEEL AGESELF FRUSTRAT
       4       0.2347    STRANGE FEEL FRUSTRAT LIFE
       4       0.2267    AGEOTHER FEEL FRUSTRAT LIFE
       4       0.2244    FRIENDS FEEL FRUSTRAT LIFE
       4       0.2169    STRANGE AGEOTHER FEEL FRUSTRAT
       4       0.2157    STRANGE FRIENDS FEEL FRUSTRAT
       4       0.2135    AGEOTHER AGESELF FRUSTRAT LIFE
       4       0.2129    STRANGE AGESELF FRUSTRAT LIFE
       4       0.2129    FRIENDS AGESELF FRUSTRAT LIFE
       4       0.2127    STRANGE FEEL AGESELF LIFE
       4       0.2122    FRIENDS AGEOTHER FEEL FRUSTRAT
       4       0.2074    AGEOTHER FEEL AGESELF LIFE
       4       0.2053    FRIENDS FEEL AGESELF LIFE
       4       0.1998    STRANGE AGEOTHER AGESELF FRUSTRAT
       4       0.1997    FRIENDS AGEOTHER AGESELF FRUSTRAT
       4       0.1988    STRANGE FRIENDS AGESELF FRUSTRAT
       4       0.1778    STRANGE AGEOTHER FEEL LIFE
       4       0.1714    STRANGE FRIENDS FEEL LIFE
       4       0.1698    STRANGE FRIENDS FRUSTRAT LIFE
       4       0.1693    STRANGE AGEOTHER FRUSTRAT LIFE
       4       0.1596    FRIENDS AGEOTHER FRUSTRAT LIFE
       4       0.1568    STRANGE FRIENDS AGEOTHER FRUSTRAT
       4       0.1456    FRIENDS AGEOTHER FEEL LIFE
       4       0.1404    STRANGE FRIENDS AGESELF LIFE
       4       0.1400    STRANGE AGEOTHER AGESELF LIFE
       4       0.1315    FRIENDS AGEOTHER AGESELF LIFE
       4       0.1221    STRANGE AGEOTHER FEEL AGESELF
       4       0.1200    STRANGE FRIENDS FEEL AGESELF
       4       0.1162    FRIENDS AGEOTHER FEEL AGESELF
       4       0.1053    STRANGE FRIENDS AGEOTHER LIFE
       4       0.0909    STRANGE FRIENDS AGEOTHER FEEL
       4       0.0633    STRANGE FRIENDS AGEOTHER AGESELF
-------------------------------------------------------------------
       5       0.2841    FRIENDS FEEL AGESELF FRUSTRAT LIFE
       5       0.2828    AGEOTHER FEEL AGESELF FRUSTRAT LIFE
       5       0.2824    STRANGE FEEL AGESELF FRUSTRAT LIFE
       5       0.2631    STRANGE FRIENDS FEEL AGESELF FRUSTRAT
       5       0.2625    FRIENDS AGEOTHER FEEL AGESELF FRUSTRAT
       5       0.2613    STRANGE AGEOTHER FEEL AGESELF FRUSTRAT
       5       0.2374    STRANGE AGEOTHER FEEL FRUSTRAT LIFE
       5       0.2350    STRANGE FRIENDS FEEL FRUSTRAT LIFE
       5       0.2283    FRIENDS AGEOTHER FEEL FRUSTRAT LIFE
       5       0.2173    STRANGE FRIENDS AGEOTHER FEEL FRUSTRAT
       5       0.2158    STRANGE AGEOTHER FEEL AGESELF LIFE
       5       0.2136    STRANGE AGEOTHER AGESELF FRUSTRAT LIFE
       5       0.2135    FRIENDS AGEOTHER AGESELF FRUSTRAT LIFE
```

```
       5       0.2133    STRANGE FRIENDS FEEL AGESELF LIFE
       5       0.2129    STRANGE FRIENDS AGESELF FRUSTRAT LIFE
       5       0.2093    FRIENDS AGEOTHER FEEL AGESELF LIFE
       5       0.1998    STRANGE FRIENDS AGEOTHER AGESELF FRUSTRAT
       5       0.1778    STRANGE FRIENDS AGEOTHER FEEL LIFE
       5       0.1698    STRANGE FRIENDS AGEOTHER FRUSTRAT LIFE
       5       0.1405    STRANGE FRIENDS AGEOTHER AGESELF LIFE
       5       0.1223    STRANGE FRIENDS AGEOTHER FEEL AGESELF
       6       0.2846    FRIENDS AGEOTHER FEEL AGESELF FRUSTRAT LIFE
       6       0.2841    STRANGE FRIENDS FEEL AGESELF FRUSTRAT LIFE
       6       0.2829    STRANGE AGEOTHER FEEL AGESELF FRUSTRAT LIFE
       6       0.2632    STRANGE FRIENDS AGEOTHER FEEL AGESELF FRUSTRAT
       6       0.2376    STRANGE FRIENDS AGEOTHER FEEL FRUSTRAT LIFE
       6       0.2163    STRANGE FRIENDS AGEOTHER FEEL AGESELF LIFE
       6       0.2136    STRANGE FRIENDS AGEOTHER AGESELF FRUSTRAT LIFE
-------------------------------------------------------------------------
  7       0.2846    STRANGE FRIENDS AGEOTHER FEEL AGESELF FRUSTRAT LIFE
```

Backward Elimination Procedure

- Begins by calculating statistics for the maximum model.
- Then the variables are deleted from the model one by one until all the variables remaining in the model produce significant Partial F tests according to the significance level for removal.
- At each step, the variable showing the smallest contribution to the model is deleted.
- Once a variable is removed from the model it remains out of the model.

Regression

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	life satisf, #friends, age-self, feel anger, age-other, frustration, #strangers[a]	.	Enter
2	.	#strangers	Backward (criterion: Probability of F-to-remove >= .100).
3	.	age-other	Backward (criterion: Probability of F-to-remove >= .100).
4	.	#friends	Backward (criterion: Probability of F-to-remove >= .100).

a. All requested variables entered.

b. Dependent Variable: express anger

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.533[a]	.285	.231	4.4670
2	.533[b]	.285	.239	4.4434
3	.533[c]	.284	.247	4.4216
4	.531[d]	.282	.253	4.4046

a. Predictors: (Constant), life satisf, #friends, age-self, feel anger, age-other, frustration, #strangers

b. Predictors: (Constant), life satisf, #friends, age-self, feel anger, age-other, frustration

c. Predictors: (Constant), life satisf, #friends, age-self, feel anger, frustration

d. Predictors: (Constant), life satisf, age-self, feel anger, frustration

ANOVA[e]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	746.127	7	106.590	5.342	.000[a]
	Residual	1875.686	94	19.954		
	Total	2621.814	101			
2	Regression	746.119	6	124.353	6.298	.000[b]
	Residual	1875.695	95	19.744		
	Total	2621.814	101			
3	Regression	744.928	5	148.986	7.620	.000[c]
	Residual	1876.885	96	19.551		
	Total	2621.814	101			
4	Regression	739.965	4	184.991	9.535	.000[d]
	Residual	1881.849	97	19.401		
	Total	2621.814	101			

a. Predictors: (Constant), life satisf, #friends, age-self, feel anger, age-other, frustration, #strangers

b. Predictors: (Constant), life satisf, #friends, age-self, feel anger, age-other, frustration

c. Predictors: (Constant), life satisf, #friends, age-self, feel anger, frustration

d. Predictors: (Constant), life satisf, age-self, feel anger, frustration

e. Dependent Variable: express anger

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-15.296	10.625		-1.440	.153	-36.393	5.801
	#strangers	1.080E-02	.510	.002	.021	.983	-1.001	1.023
	#friends	-.227	.483	-.043	-.469	.640	-1.186	.733
	age-other	3.359E-02	.137	.022	.245	.807	-.239	.306
	feel anger	.386	.126	.276	3.055	.003	.135	.637
	age-self	.251	.101	.243	2.484	.015	.050	.451
	frustration	.693	.231	.311	2.996	.003	.234	1.152
	life satisf	.421	.251	.168	1.677	.097	-.078	.920
2	(Constant)	-15.170	8.767		-1.730	.087	-32.575	2.234
	#friends	-.224	.463	-.043	-.484	.630	-1.144	.696
	age-other	3.344E-02	.136	.022	.246	.807	-.237	.304
	feel anger	.386	.126	.276	3.072	.003	.136	.635
	age-self	.250	.091	.243	2.733	.007	.068	.431
	frustration	.691	.219	.310	3.161	.002	.257	1.125
	life satisf	.422	.247	.168	1.712	.090	-.067	.912
3	(Constant)	-14.708	8.520		-1.726	.088	-31.620	2.204
	#friends	-.232	.460	-.044	-.504	.616	-1.145	.681
	feel anger	.381	.123	.272	3.090	.003	.136	.625
	age-self	.254	.090	.246	2.831	.006	.076	.431
	frustration	.699	.215	.314	3.251	.002	.272	1.126
	life satisf	.417	.244	.166	1.705	.091	-.068	.902
4	(Constant)	-15.194	8.433		-1.802	.075	-31.930	1.543
	feel anger	.370	.121	.265	3.061	.003	.130	.611
	age-self	.254	.089	.247	2.844	.005	.077	.431
	frustration	.701	.214	.315	3.272	.001	.276	1.126
	life satisf	.413	.243	.164	1.696	.093	-.070	.896

a. Dependent Variable: express anger

Excluded Variables[d]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics: Tolerance
2	#strangers	.002[a]	.021	.983	.002	.687
3	#strangers	.001[b]	.008	.994	.001	.688
	age-other	.022[b]	.246	.807	.025	.908
4	#strangers	-.013[c]	-.128	.898	-.013	.743
	age-other	.025[c]	.280	.780	.029	.912
	#friends	-.044[c]	-.504	.616	-.051	.973

a. Predictors in the Model: (Constant), life satisf, #friends, age-self, feel anger, age-other, frustration

b. Predictors in the Model: (Constant), life satisf, #friends, age-self, feel anger, frustration

c. Predictors in the Model: (Constant), life satisf, age-self, feel anger, frustration

d. Dependent Variable: express anger

Forward Elimination Procedure

- Begins with no predictors in the model.
- For each of the predictors, FORWARD calculates F statistics that reflect the variable's contribution to the model if it is included.
- Variables are added to the model one by one until no remaining variables produce a significant partial F based on the entry significance level.
    - If several predictors are significant, FORWARD adds the predictor that has the largest Partial F value.
    - If no predictors are significant, FORWARD stops.
- Once a variable is in the model, it stays.

Regression

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	frustration	.	Forward (Criterion: Probability-of-F-to-enter <= .050)
2	feel anger	.	Forward (Criterion: Probability-of-F-to-enter <= .050)
3	age-self	.	Forward (Criterion: Probability-of-F-to-enter <= .050)

a. Dependent Variable: express anger

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.388[a]	.150	.142	4.7203
2	.456[b]	.208	.192	4.5795
3	.511[c]	.261	.238	4.4466

a. Predictors: (Constant), frustration

b. Predictors: (Constant), frustration, feel anger

c. Predictors: (Constant), frustration, feel anger, age-self

ANOVA[d]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	393.715	1	393.715	17.670	.000[a]
	Residual	2228.098	100	22.281		
	Total	2621.814	101			
2	Regression	545.585	2	272.792	13.007	.000[b]
	Residual	2076.229	99	20.972		
	Total	2621.814	101			
3	Regression	684.132	3	228.044	11.534	.000[c]
	Residual	1937.681	98	19.772		
	Total	2621.814	101			

a. Predictors: (Constant), frustration

b. Predictors: (Constant), frustration, feel anger

c. Predictors: (Constant), frustration, feel anger, age-self

d. Dependent Variable: express anger

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	9.401	6.161		1.526	.130	-2.823	21.625
	frustration	.863	.205	.388	4.204	.000	.456	1.271
2	(Constant)	1.661	6.633		.250	.803	-11.502	14.823
	frustration	.882	.199	.396	4.422	.000	.486	1.277
	feel anger	.337	.125	.241	2.691	.008	.088	.585
3	(Constant)	-12.403	8.349		-1.486	.141	-28.972	4.166
	frustration	.862	.194	.387	4.452	.000	.478	1.247
	feel anger	.350	.122	.250	2.876	.005	.108	.591
	age-self	.237	.090	.230	2.647	.009	.059	.415

a. Dependent Variable: express anger

Excluded Variables[d]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics
						Tolerance
1	#strangers	-.082[a]	-.858	.393	-.086	.932
	#friends	-.001[a]	-.012	.990	-.001	1.000
	age-other	.009[a]	.091	.927	.009	.982
	feel anger	.241[a]	2.691	.008	.261	.999
	age-self	.220[a]	2.445	.016	.239	.998
	life satisf	.107[a]	1.039	.301	.104	.806
2	#strangers	-.088[b]	-.950	.344	-.096	.932
	#friends	-.041[b]	-.451	.653	-.045	.974
	age-other	.054[b]	.586	.559	.059	.951
	age-self	.230[b]	2.647	.009	.258	.997
	life satisf	.134[b]	1.343	.183	.134	.798
3	#strangers	.016[c]	.156	.876	.016	.764
	#friends	-.039[c]	-.446	.656	-.045	.974
	age-other	.011[c]	.125	.901	.013	.919
	life satisf	.164[c]	1.696	.093	.170	.789

a. Predictors in the Model: (Constant), frustration

b. Predictors in the Model: (Constant), frustration, feel anger

c. Predictors in the Model: (Constant), frustration, feel anger, age-self

d. Dependent Variable: express anger

Stepwise Selection Procedure

- The stepwise method is a modification of the forward-selection technique and differs in that predictors already in the model do not necessarily stay there.

- Predictors are added one by one to the model, and the F statistics for a predictor to be added must be significant as defined by the ENTRY significance level.

- After a variable is added, the stepwise method looks at all the predictors already included in the model and deletes any predictor that does not produce a significant F at the REMOVAL significance level.

- Only after this check is made and the necessary deletions completed can another variable be added to the model.

- This process ends when none of the variables **outside** the model are significant according to the significance level for entry and all variables **inside** the model are significant according to the significance level for removal…**or when** the variable to be added to the model is the one just deleted from it.

Regression

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	frustration	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
2	feel anger	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
3	age-self	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).

a. Dependent Variable: express anger

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.388[a]	.150	.142	4.7203
2	.456[b]	.208	.192	4.5795
3	.511[c]	.261	.238	4.4466

a. Predictors: (Constant), frustration

b. Predictors: (Constant), frustration, feel anger

c. Predictors: (Constant), frustration, feel anger, age-self

ANOVA[d]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	393.715	1	393.715	17.670	.000[a]
	Residual	2228.098	100	22.281		
	Total	2621.814	101			
2	Regression	545.585	2	272.792	13.007	.000[b]
	Residual	2076.229	99	20.972		
	Total	2621.814	101			
3	Regression	684.132	3	228.044	11.534	.000[c]
	Residual	1937.681	98	19.772		
	Total	2621.814	101			

a. Predictors: (Constant), frustration

b. Predictors: (Constant), frustration, feel anger

c. Predictors: (Constant), frustration, feel anger, age-self

d. Dependent Variable: express anger

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	9.401	6.161		1.526	.130	-2.823	21.625
	frustration	.863	.205	.388	4.204	.000	.456	1.271
2	(Constant)	1.661	6.633		.250	.803	-11.502	14.823
	frustration	.882	.199	.396	4.422	.000	.486	1.277
	feel anger	.337	.125	.241	2.691	.008	.088	.585
3	(Constant)	-12.403	8.349		-1.486	.141	-28.972	4.166
	frustration	.862	.194	.387	4.452	.000	.478	1.247
	feel anger	.350	.122	.250	2.876	.005	.108	.591
	age-self	.237	.090	.230	2.647	.009	.059	.415

a. Dependent Variable: express anger

Excluded Variables[d]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics
						Tolerance
1	#strangers	-.082[a]	-.858	.393	-.086	.932
	#friends	-.001[a]	-.012	.990	-.001	1.000
	age-other	.009[a]	.091	.927	.009	.982
	feel anger	.241[a]	2.691	.008	.261	.999
	age-self	.220[a]	2.445	.016	.239	.998
	life satisf	.107[a]	1.039	.301	.104	.806
2	#strangers	-.088[b]	-.950	.344	-.096	.932
	#friends	-.041[b]	-.451	.653	-.045	.974
	age-other	.054[b]	.586	.559	.059	.951
	age-self	.230[b]	2.647	.009	.258	.997
	life satisf	.134[b]	1.343	.183	.134	.798
3	#strangers	.016[c]	.156	.876	.016	.764
	#friends	-.039[c]	-.446	.656	-.045	.974
	age-other	.011[c]	.125	.901	.013	.919
	life satisf	.164[c]	1.696	.093	.170	.789

a. Predictors in the Model: (Constant), frustration

b. Predictors in the Model: (Constant), frustration, feel anger

c. Predictors in the Model: (Constant), frustration, feel anger, age-self

d. Dependent Variable: express anger

Chunkwise or Blockwise Selection Procedures

Technically, forward selection is applied to blocks (sets) of predictors while using any of the predictor-selection methods to select predictors from each block.

Demographic: Physical Health, Age, Gender, Income,

Psychological: # Life Events in last year, Prior Depression, Anxiety, Anger

For example, I would use stepwise selection to determine which of the demographic variables were significant predictors of depression.

- Model depression = health age gender income/selection = stepwise;
- Selected Model ⇒ depression = gender health;

Then, I would use stepwise selection to determine which of the psychological variables were significant predictors of depression given the key demographic variables are already in the model.

- Model depression = gender health events prior anxiety anger /
  selection = stepwise include=2;
- Selected Model ⇒ depression = gender health events prior

Variant on the Blockwise Selection Procedure

1. Use Stepwise Selection to determine the best demographic predictors of depression.
   - Model depression = health age gender income/selection = stepwise;
   - Selected Model ⇒ depression = gender health;
2. Use Stepwise Selection to determine the best psychological predictors of depression.
   - Model depression = events prior anxiety anger/selection = stepwise;
   - Selected Model ⇒ depression = events prior anxiety
3. Use Stepwise Selection to determine the best combination of demographic and psychological predictors of depression.
   - Model depression = health gender events prior anxiety/ selection = stepwise;
   - Selected Model ⇒ depression = events gender anxiety

Including Categorical Predictors within the Linear Regression Model

- We have used linear regression to predict a continuous dependent variable using information from continuous independent (predictor) variables.

- With some modification, we can include categorical variables as predictors in linear regression.

Modification of Categorical Variables for Use in Linear Regression

Ways to Modify a Categorical Variable:
- Reference Group Coding
- Effect Coding
- Orthogonal Coding (not discussed)

Suppose we wanted to predict a person's **Income** based on their **Race** (0 =White, 1 = Black, 2 = Hispanic, 3 = Other).
- Income, the dependent variable, is continuous.
- Race, the predictor, is categorical and has 4 categories.

Initially, we would code Race a categorical variable in SAS. To analyze the data via PROC REG, we would have to recode RACE.

If we have 4 categories for race, we need 3 Dummy Variables in our regression equation.
If we have 8 categories for race, we need 7 Dummy Variables in our regression equation.

If you have *a* categories, you need *a -1* dummy variables.

	Reference Group Coding			Effect Coding		
Race	**D1**	**D2**	**D3**	**D1**	**D2**	**D3**
0 = White	0	0	0	-1	-1	-1
1 = Black	1	0	0	1	0	0
2 = Hispanic	0	1	0	0	1	0
3 = Other	0	0	1	0	0	1

Interpretation of the Regression Output with Reference Group Coding	Interpretation of the Regression Output with Effect Coding
constant = mean of the reference group (assigned zeros throughout)	**constant** = grand mean of the dependent variable (all people)
regression coefficient = deviation of the mean of the group identified in the dummy variable from the mean of the reference group.	**regression coefficient** = deviation of the mean of the group identified in the dummy variable from the grand mean. It is the treatment effect.
The *t* tests for ***b*** are equivalent to pairwise comparisons for comparing the groups to the reference group. (treatment vs. control)	**The *t* tests** for ***b*** are meaningless (unless you want to test whether the group differs from the grand mean).
The R^2 and *F* values are not influenced by the choice of coding (i.e., Reference Group, Effects, or Orthogonal).	**The R^2 and *F* values** are not influenced by the choice of coding (i.e., Reference Group, Effects, or Orthogonal).
The Null Hypothesis for *F* is that all the group means are equal…one-way anova. (test of equal intercepts)	The Null Hypothesis for *F* is that all the group means are equal…one-way anova. (test of equal intercepts)
Pairwise Comparisons: The *t* tests do some of them, you would have to recode the dummy variables to obtain all possible pairwise comparisons.	**Pairwise Comparisons:** Test of the difference between two regression coefficients. $$t = \frac{b_j - b_{j'}}{\sqrt{\frac{MSE}{n}(2)}}$$

Regression

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	D3, D1, D2[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: INCOME

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.368[a]	.135	.117	6.5720

a. Predictors: (Constant), D3, D1, D2

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	986.062	3	328.687	7.610	.000[a]
	Residual	6305.831	146	43.191		
	Total	7291.893	149			

a. Predictors: (Constant), D3, D1, D2

b. Dependent Variable: INCOME

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	20.941	1.594		13.138	.000
	D1	1.213	2.421	.049	.501	.617
	D2	.269	2.194	.013	.123	.902
	D3	5.861	1.723	.394	3.402	.001

a. Dependent Variable: INCOME

Frequencies: For your Information...not part of Regression Output

RACE

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Black	17	11.3	11.3	11.3
	Hispanic	13	8.7	8.7	20.0
	Other	19	12.7	12.7	32.7
	White	101	67.3	67.3	100.0
	Total	150	100.0	100.0	

Two predictors (one categorical and one continuous)

Suppose we wanted to predict a person's **Income** based on their **Years Education** and **Race** (0 = White, 1 = Black, 2 = Hispanic, 3 = Other).

- Income, the dependent variable, is continuous.
- Years Education (12 = HS degree, 16 = Bachelors degree, etc.) is continuous.
- Race is categorical and has 4 categories.

First, we need to recode RACE into dummy variables.

	Reference Group Coding			Effect Coding		
Race	**D1**	**D2**	**D3**	**D1**	**D2**	**D3**
0 = White	0	0	0	-1	-1	-1
1 = Black	1	0	0	1	0	0
2 = Hispanic	0	1	0	0	1	0
3 = Other	0	0	1	0	0	1

Second, we need to check whether there is an interaction between Race and Years Education.
Equivalently,

- Test that the relationship between income and education is the same for each racial group.
- Test that the Slopes are the Same across the Races
- Test of Parallelism

ED1 = D1*EDUC
ED2 = D2*EDUC
ED3 = D3*EDUC

If the interaction is <u>significant</u>, we cannot discuss the effect of education on Salary without taking Race into account. Likewise, we cannot discuss the effect of Race on Salary without taking Years Education into account.

^^
Testing for an Interaction Between RACE and EDUC
^^

Regression

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	D3, EDUC, D1, D2[a]	.	Enter
2	ED2, ED1, ED3[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: INCOME

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.470[a]	.221	.200	6.2582	.221	10.295	4	145	.000
2	.501[b]	.251	.214	6.2030	.030	1.865	3	142	.138

a. Predictors: (Constant), D3, EDUC, D1, D2

b. Predictors: (Constant), D3, EDUC, D1, D2, ED2, ED1, ED3

ANOVA[c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1612.876	4	403.219	10.295	.000[a]
	Residual	5679.018	145	39.166		
	Total	7291.893	149			
2	Regression	1828.197	7	261.171	6.788	.000[b]
	Residual	5463.696	142	38.477		
	Total	7291.893	149			

a. Predictors: (Constant), D3, EDUC, D1, D2

b. Predictors: (Constant), D3, EDUC, D1, D2, ED2, ED1, ED3

c. Dependent Variable: INCOME

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	5.797	4.078		1.421	.157
	EDUC	1.051	.263	.301	4.001	.000
	D1	.999	2.306	.040	.433	.666
	D2	.591	2.091	.028	.283	.778
	D3	4.941	1.657	.332	2.982	.003
2	(Constant)	-22.023	12.717		-1.732	.085
	EDUC	2.981	.876	.854	3.402	.001
	D1	24.396	20.326	.984	1.200	.232
	D2	33.371	15.664	1.592	2.130	.035
	D3	35.649	13.606	2.398	2.620	.010
	ED1	-1.628	1.390	-.966	-1.171	.243
	ED2	-2.282	1.085	-1.557	-2.103	.037
	ED3	-2.119	.931	-2.234	-2.277	.024

a. Dependent Variable: INCOME

- We may test whether race is a predictor of salary after adjusting for years education (Test for equal intercepts among the races).

  Full: $Income = b_0 + b_1 Educ + b_2 D1 + b_3 D2 + b_4 D3 + \varepsilon$

  Reduced: $Income = b_0 + b_1 Educ + \varepsilon$

  If the RACE F test is significant, then we would conclude that the average salary differs among the races after adjusting for years of education (intercepts differ).

  If the RACE F test is not significant, we would conclude RACE is not a significant predictor of the salary one earns after adjusting for years of education.

- We may test whether years education is a predictor of salary after adjusting for race.

  Full: $Income = b_0 + b_1 Educ + b_2 D1 + b_3 D2 + b_4 D3 + \varepsilon$

  Reduced: $Income = b_0 + b_2 D1 + b_3 D2 + b_4 D3 + \varepsilon$

  If the *t* test for educ is significant, then we would conclude that education is related to one's salary after adjusting for race.

  If the *t* test for educ is not significant, then we would conclude that education is not a significant predictor after adjusting for race.

If **Race** is significant, but **Educ** is not, then we have a one-way anova:

$$Income = b_0 + b_2 D1 + b_3 D2 + b_4 D3 + \varepsilon$$

If **Educ** is significant, but **Race** is not, then we have simple linear regression:

$$Income = b_0 + b_1 Educ + \varepsilon$$

^^

Testing for Race, Adjusting for EDUC
Testing for EDUC, Adjusting for RACE

^^

Regression

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	EDUC[a]	.	Enter
2	D1, D2, D3[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: INCOME

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics: R Square Change	F Change	df1	df2	Sig. F Change
1	.368[a]	.135	.129	6.5280	.135	23.114	1	148	.000
2	.470[b]	.221	.200	6.2582	.086	5.344	3	145	.002

a. Predictors: (Constant), EDUC

b. Predictors: (Constant), EDUC, D1, D2, D3

ANOVA[c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	984.985	1	984.985	23.114	.000[a]
	Residual	6306.908	148	42.614		
	Total	7291.893	149			
2	Regression	1612.876	4	403.219	10.295	.000[b]
	Residual	5679.018	145	39.166		
	Total	7291.893	149			

a. Predictors: (Constant), EDUC

b. Predictors: (Constant), EDUC, D1, D2, D3

c. Dependent Variable: INCOME

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	5.816	4.031		1.443	.151
	EDUC	1.282	.267	.368	4.808	.000
2	(Constant)	5.797	4.078		1.421	.157
	EDUC	1.051	.263	.301	4.001	.000
	D1	.999	2.306	.040	.433	.666
	D2	.591	2.091	.028	.283	.778
	D3	4.941	1.657	.332	2.982	.003

a. Dependent Variable: INCOME