SAN JOSÉ STATE UNIVERSITY

School of Social Work
ScWk 242
Spring 2009
Peter Allen Lee

Linear Regression

Terms and concepts you should know:

- Simple linear regression
- Prediction
- Equation of a Line
- Slope, coefficient, constant
- F test and its relationship to regression
- R-square
- Adjusted R-square
- Beta
- Scatterplot and its relationship to regression
- Directional Hypothesis within Regression
- Multiple Regression
- Control Variables
- Dummy Coding
- Dummy Variable
- Baseline
- Rotating Baselines
- Logistic Regression

Simple Linear Regression

Conceptually, simple linear regression is used to predict the value of the dependent variable when the value of the independent variable and the constant is known. You can also use simple linear regression to examine the relationship between the independent variable and the dependent variable.

General notation: (equation of a line)

$$Y = c + bX$$

Y = dependent variable
X = independent variable
b = slope of the line or the coefficient of the X variable
c = constant (or the point of intercept at the Y axis)

To begin our understanding of linear regression, let's re-visit the relationship between exercise and stress:

2. H^{o}: There is no relationship between exercise time and stress.

H^{a}: There is a relationship between exercise time and stress.

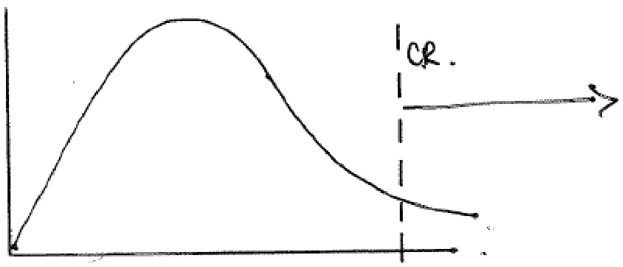
3. Alpha level = .05

4. Statistical test and sampling distribution = Simple Linear Regression and normal distribution

5. Critical region and degrees of freedom

The critical value for F to test the significance of the overall model at alpha .05, df = 1, 23 = 4.28

*** The df can be found by looking at the regression and residual df portions of the printout under the ANOVA box.



CV =
α= .05 and df = 1,28

***** Running linear regression**

Enter data for exercise time and stress. For this example, the relevant variable names are **exercise** and **stress**. Click on **Analyze**, **Regression**, **Linear**... In the **Dependent** box put **stress** and in the **Independent(s)** box put **exercise**. Leave the other settings alone for now. Click **OK**.

----

6. Table of results (use the SPSS print-out)

Regression

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	exercise[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: stress

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.788[a]	.622	.608	5.19343

a. Predictors: (Constant), exercise

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1240.659	1	1240.659	45.999	.000[a]
	Residual	755.208	28	26.972		
	Total	1995.867	29			

a. Predictors: (Constant), exercise

b. Dependent Variable: stress

Coefficients[a]

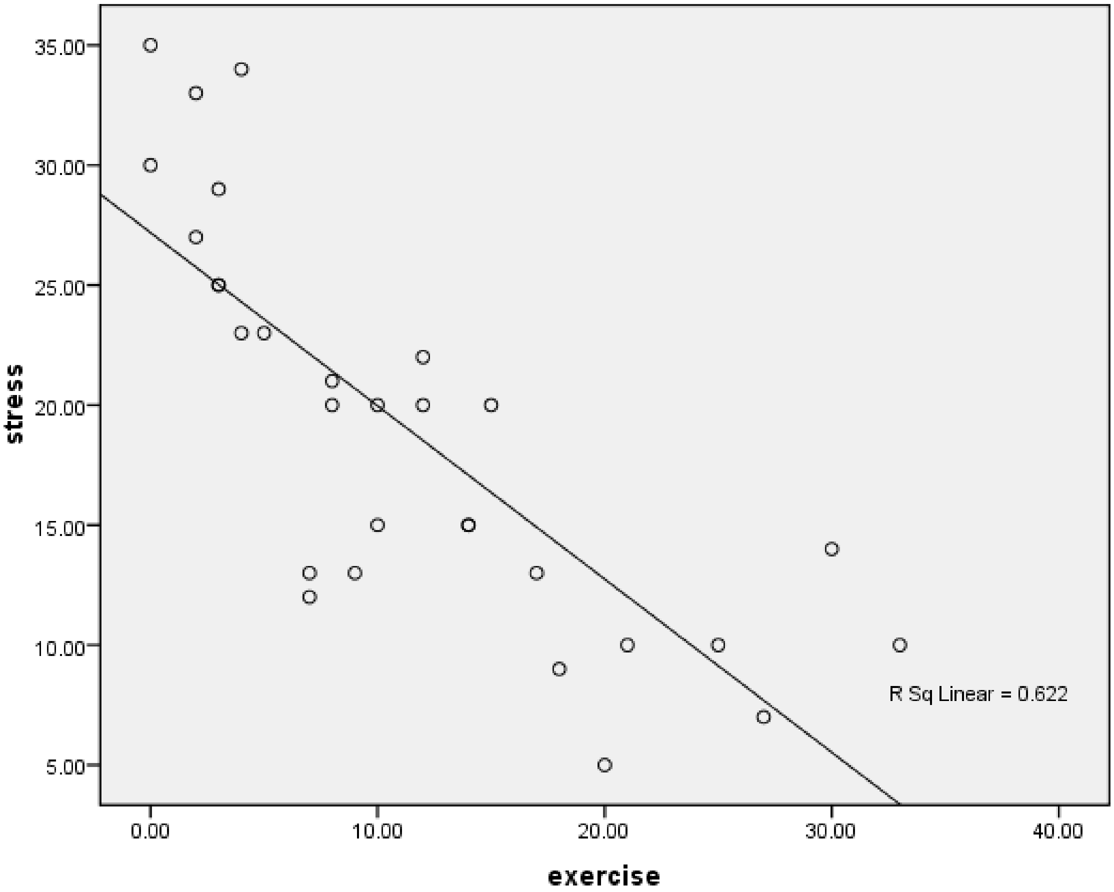
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	27.189	1.543		17.621	.000
	exercise	-.722	.106	-.788	-6.782	.000

a. Dependent Variable: stress

7. Results:

   From a simple linear regression analysis, the overall model with exercise time as a predictor of stress was significant (adjusted R-square = .608, F (1, 28) = 45.999, *p* = .001). Those who spent more time exercising also had lower stress levels (b coefficient = -.788, p = .001).

8. Interpretation: See correlation notes



***** Reviewing the Scatterplot (only with one independent and one dependent variable)**

Click on **Graphs**, **Scatter**, and then select **Simple** and click on **Define**. Then in the Simple Scatterplot box, move the **exercise** variable into the X-axis box and the **stress** variable into the Y-axis box. Independent variables are always put on the X-axis and dependent variables on the Y-axis. Click **OK**.

***** Adding the Regression Line**

In the Output view, double click on the scatterplot picture you have just created. The Chart Editor will appear. Click on **Chart** and then **Options**. In the **Scatterplot Options** box, click the **Total** box under the part that says **Fit Line**. Click **OK**.

What is the relationship between what you see in the graph and the output in your linear regression?

***** Note on equation of a line:**

Y = c + bX

Y = 27.189 + (-.722) (X) or Stress = 27.189 + (-.722) (Exercise Time)

***** Try plugging in a number for X (exercise time) and calculate the corresponding predicted value for Y (Stress). Does it match your line in the graph?**

***** Remember though that we report the Beta (or the standardized coefficient called beta). Why? Beta is the value expressed in <u>standardized units of measurement</u> which makes your variables comparable (similar to the concept of z-scores).**

***** Note that if you have a directional hypothesis (one-tailed test) for the independent variable, you may divide the p-value of specific independent variables by 2, just as we have always done with directional hypotheses and resulting 2-tailed p-values. REMEMBER, you would do this ONLY for the variables in which you have a directional hypothesis.**

***** Why do we report the Adjusted R-square by convention instead of the regular R value?**

***** What is the conceptual meaning of "error" in the linear equation Y = c + bX + error**

Multiple Linear Regression

Conceptually, multiple regression is used to predict the value of the dependent variable when the value of many independent variables and the constant are known. You can also use multiple regression to examine the relationship of the independent variables with the dependent variable.

Let's revisit the scenario where you are the medical social worker conducting a preliminary investigation about the correlates of exercise. You are given access to the hospital database and obtain a random sample of 30 adults who had recently completed a hospital outreach program about improving exercise habits. This time however, you are examining average hours of aerobic exercise per week, self-esteem, life satisfaction, and IQ as predictors of stress. [Try doing this in SPSS following the instructions above or those in your book]

$$\mathbf{Y = c + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4}$$

***** What are the b's and X's? What is the Y? What is c ?**

1.

2. H^o: Average hours of aerobic exercise per week, self-esteem, life satisfaction, and IQ are not predictors of stress.

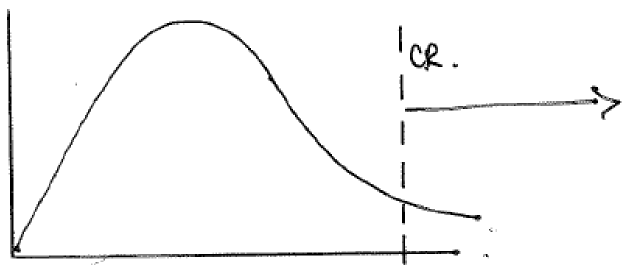
   H^a: Average hours of aerobic exercise per week, self-esteem, life satisfaction, and IQ are predictors of stress.

3. Alpha level = .05

4. Statistical test and sampling distribution = Multiple Linear Regression and normal distribution

5. Critical region and degrees of freedom

   The critical value for F to test the significance of the overall model at alpha .05, df = 3, 26 = 2.89.
   *** The df can be found by looking at the regression and residual df portions of the printout under the ANOVA box.



CV =
α= .05 and df =

6. Results

Regression

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	iq, satisfaction, self-esteem, exercise[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: stress

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.836[a]	.698	.650	4.90780

a. Predictors: (Constant), iq, satisfaction, self-esteem, exercise

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1393.705	4	348.426	14.466	.000[a]
	Residual	602.162	25	24.086		
	Total	1995.867	29			

a. Predictors: (Constant), iq, satisfaction, self-esteem, exercise

b. Dependent Variable: stress

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	8.914	15.651		.570	.574
	exercise	-.568	.201	-.620	-2.825	.009
	self-esteem	-.195	.195	-.216	-1.000	.327
	satisfaction	-.186	.177	-.165	-1.046	.305
	iq	.264	.143	.239	1.843	.077

a. Dependent Variable: stress

7. Results:

   From a multiple linear regression analysis, the overall model predicting stress by the average hours of aerobic exercise per week, self-esteem, life satisfaction, and IQ was significant (adjusted R-square = .650, F (4, 25) = 14.466, *p* = .001). Specifically, within the model, those who exercised more had significantly lower stress (b coefficient = -.620, p = .009). IQ approached significance; those with higher IQ had higher stress (b coefficient = .239, p = .077). The relationship between self-esteem and stress (b coefficient = -.216, p = .327), and life satisfaction with stress (b coefficient = -.165, p = .305) were not significant.

8. Interpretation:

   More exercise predicts reduced stress. Blah blah.... IQ should be investigated blah blah... [Further explanation would be added depending on additional evidence contained in your research and from your literature review...].

***** What are control variables?**

***** What would you recommend for future research given these results?**

***** What if you had directional hypotheses for some or all of your independent variables?**

***** What is the advantage of multivariate analysis over bivariate and univariate analyses?**

***** What would be the regression equation given these results?**

Stress = 8.914 + (-.568)(exercise) + (-.195)(self-esteem) + (-.186)(life satisfaction) + (.264)(IQ)