Linear Regression

PSY 5102: Advanced Statistics for Psychological and Behavioral Research 2

Goals

- Understand linear regression with a single predictor
- Understand how we assess the fit of a regression model
  - Total Sum of Squares
  - Model Sum of Squares
  - Residual Sum of Squares
  - F
  - R^2
- Understand how to conduct a simple linear regression using SPSS
- Understand how to interpret a regression model

What is Regression?

- A way of predicting the value of one variable from another
  - It is a hypothetical model of the relationship between two variables
  - The model used is a linear one
  - Therefore, we describe the relationship using the equation of a straight line

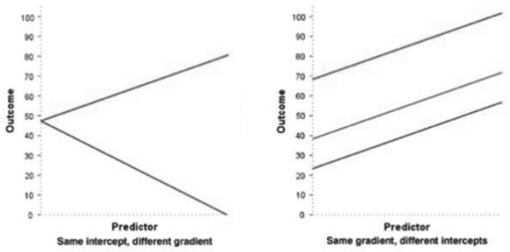
Describing a Straight Line

- Y = A + BX + e
- A
  - Y-Intercept (value of Y when X = 0)
  - Point at which the regression line crosses the Y-axis (ordinate)
- B
  - Regression coefficient for the predictor
  - Gradient (slope) of the regression line
  - Direction/strength of relationship
- e
  - Error term for the model

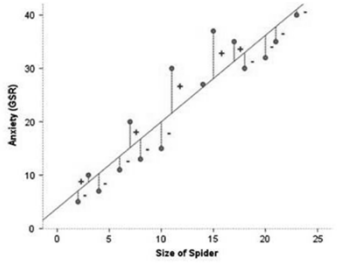
Additional Information

- Y = A + BX + e vs. Y' = A + BX
  - The "Y" formula is for the "actual score" and requires an error term. The "Y'" formula is for the estimated value of Y and does not require an error term.
- How to calculate the slope (B)
  - $B = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma X^2 - (\Sigma X)^2} = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\Sigma(X-\bar{X})^2} = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{SS_x}$
- How to calculate the Y-intercept (A)
  - $A = \bar{Y} - B\bar{X}$

Intercepts and Gradients



The Method of Least Squares



This scatterplot shows the association between the size of a spider and the anxiety response it elicited in participants

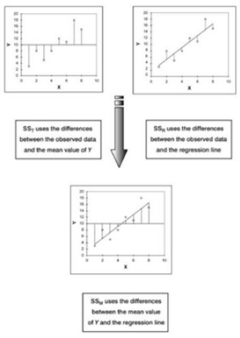
The regression line shows the best fitting line (i.e., the line that minimizes the distance between itself and the data points)

The vertical lines show the differences (or residuals) between the actual data and the predicted values

How Good is the Model?

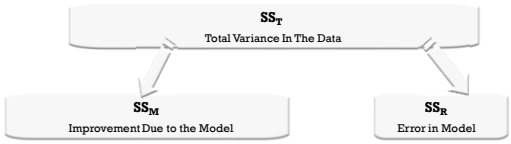
- The regression line is only a model based on the data
- This model may not reflect reality
  - We need some way of testing how well the model fits the observed data
  - How?

Sums of Squares



- SS_{Total}
  - Total variability: variability between scores and the mean
- $SS_{Residual}$
  - Residual/Error variability: variability between the regression model and the actual data
- SS_{Model}
  - Model variability: difference in variability between the model and the mean

Testing the Model: ANOVA



- If the model results in better prediction than using the mean, then we expect SS_M to be much greater than SS_R

Testing the Model: ANOVA

- Mean Squared Error
  - Sums of Squares are total values
  - They can be expressed as averages called Mean Squares (MS)
    - To get MS, the SS is divided by its corresponding degrees of freedom
    - $MS_M = \frac{SS_M}{df_M}$
      - df_M is the number of predictor variables in the model
    - $MS_R = \frac{SS_R}{df_R}$
      - df_R is the number of observations minus the number of parameters being estimated (i.e., the number of regression coefficients including the constant)

$$F = \frac{MS_M}{MS_R}$$

Testing the Model: R^2

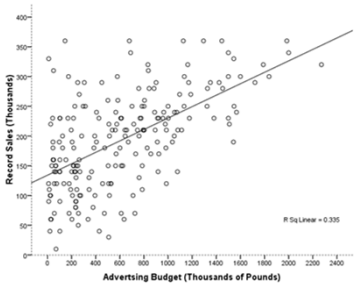
- R^2
  - The proportion of variance accounted for by the regression model

$$R^2 = \frac{SS_M}{SS_T}$$

Regression: An Example

- A record company boss was interested in predicting record sales from advertising
- Data
  - 200 different album releases
- Outcome variable:
  - Sales (CDs and Downloads) in the week after release
- Predictor variable:
  - The amount (in £s) spent promoting the record before release

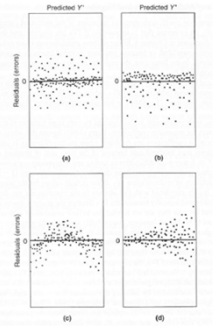
Step One: Graph the Data



Assumptions

- Simple linear regression has a number of distributional assumptions that can be examined by looking at a residual plot (i.e., plot showing the differences between the obtained and predicted values of the criterion variable)
  1. The residuals should be normally distributed around the predicted criterion scores (normality assumption)
  2. The residuals should have a horizontal-line relationship with predicted criterion scores (linearity assumption)
  3. The variance of the residuals should be the same for all predicted criterion scores (homoscedasticity assumption)

Examine Residual Plots



Plots of predicted values of the criterion variable (Y) against residuals

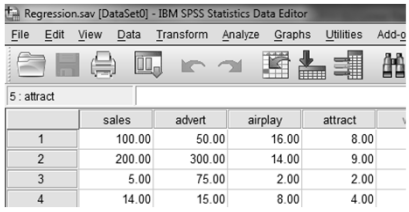
A. Shows that the assumptions of regression are met

B. Shows that the data fails to meet the normality assumption

C. Shows that the data fails to meet the linearity assumption

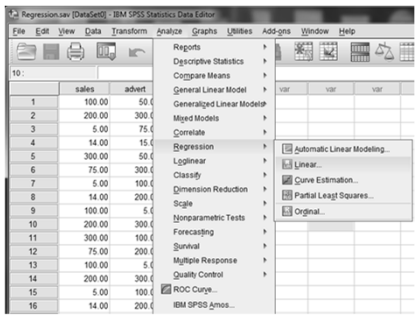
D. Shows that the data fails to meet the homoscedasticity assumption

Regression Using SPSS

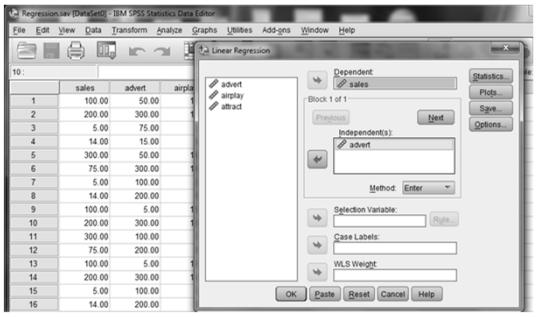


	sales	advert	airplay	attract
1	100.00	50.00	16.00	8.00
2	200.00	300.00	14.00	9.00
3	5.00	75.00	2.00	2.00
4	14.00	15.00	8.00	4.00

Regression Using SPSS



Regression Using SPSS



Output: Model Summary

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.578[a]	.335	.331	65.9914

a. Predictors: (Constant), Advertising Budget (thousands of pounds)

This is the R which reflects the degree of association between our regression model and the criterion variable

Output: Model Summary

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.578[a]	.335	.331	65.9914

a. Predictors: (Constant), Advertising Budget (thousands of pounds)

This is R^2 which represents the amount of variance in the criterion variable that is explained by the model (SS_M) relative to how much variance there was to explain in the first place (SS_T). To express this value as a percentage, you should multiply this value by 100 (which will tell you the percentage of variation in the criterion variable that can be explained by the model).

Output: Model Summary

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.578[a]	.335	.331	65.9914

a. Predictors: (Constant), Advertising Budget (thousands of pounds)

This is Adjusted R^2 which corrects R^2 for the number of predictor variables included in the model. It will always be less than or equal to R^2. It is a more conservative estimate of model fit because it penalizes researchers for including predictor variables that are not strongly associated with the criterion variable

Output: Model Summary

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.578[a]	.335	.331	65.9914

a. Predictors: (Constant), Advertising Budget (thousands of pounds)

This is the standard error of the estimate which is a measure of the accuracy of predictions. The larger the standard error of the estimate, the more error in our regression model

Output: ANOVA

SS_M

SS_R

SS_T

MS_M

MS_R

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	433687.833	1	433687.833	99.587	.000[a]
	Residual	852264.167	198	4354.870		
	Total	1295952.000	199			

a. Predictors: (Constant), Advertising Budget (thousands of pounds)

b. Dependent Variable: Record Sales (thousands)

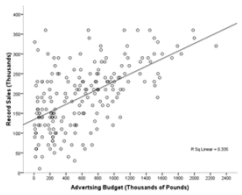
The ANOVA tells us whether the model, overall, results in a significantly good degree of prediction of the outcome variable. However, the ANOVA does not tell us about the individual contribution of variables in the model (although in this simple case there is only one variable in the model which allows us to infer that this variable is a good predictor)

SPSS Output: Model Parameters

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	134.140	7.537		17.799	.000
	Advertsing Budget (thousands of pounds)	.096	.010	.578	9.979	.000

a. Dependent Variable: Record Sales (thousands)



This is the Y-intercept (i.e., the place where the regression line intercepts the Y-axis). This means that when £0 in advertising is spent, the model predicts that 134,140 records will be sold (unit of measurement is "thousands of records")

SPSS Output: Model Parameters

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	134.140	7.537		17.799	.000
	Advertsing Budget (thousands of pounds)	.096	.010	.578	9.979	.000

a. Dependent Variable: Record Sales (thousands)

This t-test compares the value of the Y-intercept with 0. If it is significant, then it means that the value of the Y-intercept (134.14 in this example) is significantly different from 0

SPSS Output: Model Parameters

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	134.140	7.537		17.799	.000
	Advertsing Budget (thousands of pounds)	.096	.010	.578	9.979	.000

a. Dependent Variable: Record Sales (thousands)

This is the unstandardized slope or gradient coefficient. It is the change in the outcome associated with a one unit change in the predictor variable. This means that for every 1 point the advertising budget is increased (unit of measurement is thousands of pounds) the model predicts an increase of 96 records being sold (unit of measurement for record sales was thousands of records)

SPSS Output: Model Parameters

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	134.140	7.537		17.799	.000
	Advertsing Budget (thousands of pounds)	.096	.010	.578	9.979	.000

a. Dependent Variable: Record Sales (thousands)

This is the standardized regression coefficient (β). It represents the strength of the association between the predictor and the criterion variable. If there is only one predictor, then β is equal to the Pearson product-moment correlation coefficient (the closer β is to +1 or -1, then the better the prediction of Y from X [or X from Y])

SPSS Output: Model Parameters

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	134.140	7.537		17.799	.000
	Advertsing Budget (thousands of pounds)	.096	.010	.578	9.979	.000

a. Dependent Variable: Record Sales (thousands)

This t-test compares the magnitude of the standardized regression coefficient (β) with 0. If it is significant, then it means that the value of β (0.578 in this example) is significantly different from 0 (i.e., the predictor variable is significantly associated with the criterion variable)

Using The Model

- Y' = A + BX
- Record Sales = A + B(Advertising Budget)
- Record Sales = 134.14 + (0.09612 x Advertising Budget)
- Expected record sales with £0 advertising budget
  - 134.14 + (0.09612 x 0) = 134.14 + 0 = 134.14 = 134,140 records
- Expected record sales with £100,000 advertising budget
  - 134.14 + (0.09612 x 100) = 134.14 + 9.612 = 143.752 = 143,752 records
- Expected record sales with £500,000 advertising budget
  - 134.14 + (0.09612 x 500) = 134.14 + 48.06 = 182.2 = 182,200 records