Regression

PSYC 300B - Lecture 5
Dr. J. Nicol

Regression

- Regression is a statistical procedure for determining the equation for the straight line that best fits a set of data
- The equation for the best-fitting straight line is called the regression equation
- The regression equation makes it possible to predict the value of the outcome variable ($\hat{Y}$) for any given value of a predictor variable or variables (*X*)
- The regression equation explains how differences in one variable relate to differences in another and that allows us to predict a person's scores on one variables from knowledge of that person's score on another variable(s)

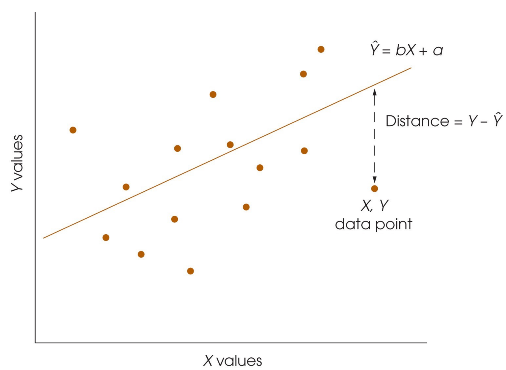
The Linear Equation

- ***Slope (b):*** determines the direction and the degree to which the regression line is tilted (i.e., shape)
- The slope indicates how many units of change you expect in ***Y*** for a one-unit change in ***X***
- ***Y-intercept (a):*** determines the point where the regression line crosses the *Y*-axis (i.e., location)
- The intercept is the predicted value of ***Y*** when ***X*** = 0
- Task is to solve for those values of ***b*** and ***a*** that will produce the best-fitting linear function (i.e., the one with the smallest deviation scores—the least squared error)

The Least Squared Error

- The regression equation provides the best prediction for a value of *Y* (i.e., $\hat{Y}$) for a given value of *X* and results in the ***least squared error (residual)*** between the observed data and the regression line (i.e., line of best fit)
- ***Residuals (errors)*** are the differences between the predicted values for the outcome variable based on the regression equation and the actual value of the outcome variable at each value of X ($\hat{Y}$ - Y)

The regression equation for the line of best-fit produces the smallest (least) sum of squared errors between the regression line and the actual data



The regression equation for Y is the linear equation:

$$\hat{Y} = bX + a$$

where

$$b = \frac{SP}{SS_X}$$

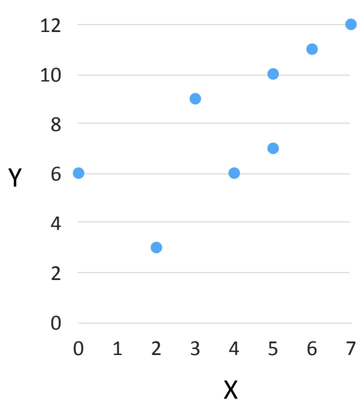
and

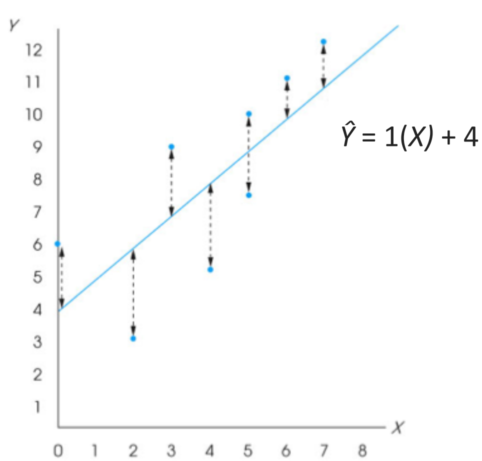
$$a = M_Y - bM_X$$

$$Covariance = \frac{\sum(X - M_x)(Y - M_Y)}{N - 1} = \frac{SP}{N - 1}$$

$$r = \frac{Covariance}{(s_X)(s_Y)} = \frac{\sum(X - M_x)(Y - M_Y)}{(N - 1)(s_X)(s_Y)}$$

X	Y
2	3
6	11
0	6
4	6
5	7
7	12
5	10
3	9



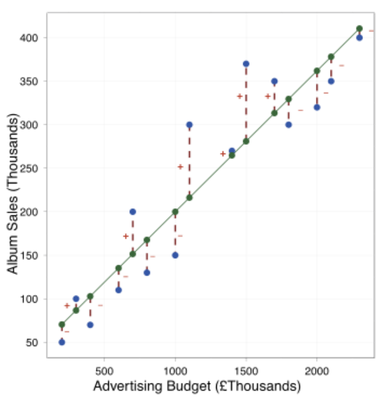


Standard Error of the Estimate

- The regression equation allows us to make predictions, but it does not tell us how good those predictions are
- Indicates the average distance between the regression line and the actual data (i.e., average error when using the regression equation to make predictions)
- It is the standard deviation of the errors that we make when using the regression equation to make predictions about *Y*

$$s_{Y-\hat{Y}} = \sqrt{\frac{SS_{ERROR}}{df_{ERROR}}} = \sqrt{MS_{ERROR}}$$

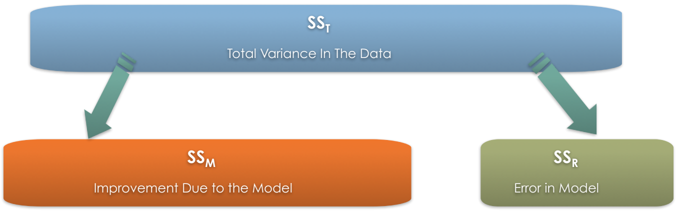
How many albums do you predict would be sold if $1,000,000 was spent on advertising?



Assessing the Goodness of Fit

- The ***mean*** is a model of no relationship (i.e., $R^2 = 0$) between the outcome and predictor variable(s)
- When we use the mean as the model, we can calculate the difference between the observed values and the values predicted by the mean
- If the best-fitting model is any good then it should have significantly less error associated with it compared to the baseline model (i.e., the mean)

When the model results in better predictions than using the mean, then $SS_{REGRESSION}$ is much greater than $SS_{RESIDUAL}$



R^2 and Predictable Variability

- R^2 measures the proportion of variability in the outcome variable that can be explained by the predictor variable (i.e., variability shared by the predictor(s) and outcome)
- So (1 - R^2) measures variability in the outcome variable that cannot be accounted for by the predictor variable(s)

$$SS_{REGRESSION} = R^2 SS_Y$$

$$SS_{ERROR} = (1 - R^2) SS_Y$$

Degrees of Freedom

$$df_{TOTAL} = N - 1$$

$$df_{REGRESSION} = \#\ of\ predictors$$

$$df_{ERROR} = df_{TOTAL} - df_{REGRESSION}$$

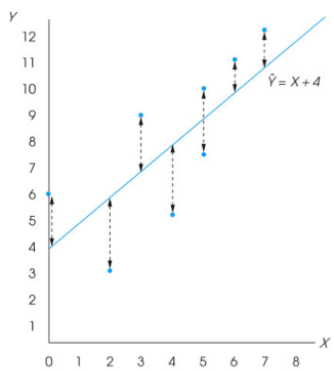
Mean Square and F

$$MS_{REGRESSION} = \frac{SS_{REGRESSION}}{df_{REGRESSION}}$$

$$MS_{ERROR} = \frac{SS_{ERROR}}{df_{ERROR}}$$

$$F = \frac{MS_{REGRESSION}}{MS_{ERROR}}$$

X	Y
2	3
6	11
0	6
4	6
5	7
7	12
5	10
3	9



- Compute R^2
- Calculate the line-of-best-fit and determine if the regression equation, with X as the predictor, accounts for a significant proportion (α = 0.05) of the variance in *Y* scores (the outcome)
- Compute the standard error of the estimate ($s_{Y-\hat{Y}}$)

Descriptive Statistics

	Mean	Std. Deviation	N
Y	8.0000	3.02372	8
X	4.0000	2.26779	8

Correlations

		Y	X
Pearson Correlation	Y	1.000	.750
	X	.750	1.000
Sig. (1-tailed)	Y	.	.016
	X	.016	.
N	Y	8	8
	X	8	8

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.750[a]	.563	.490	2.16025

a. Predictors: (Constant), X

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	36.000	1	36.000	7.714	.032[b]
	Residual	28.000	6	4.667		
	Total	64.000	7			

a. Dependent Variable: Y

b. Predictors: (Constant), X

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	4.000	1.630		2.454	.050
	X	1.000	.360	.750	2.777	.032

a. Dependent Variable: Y

- A professor obtains SAT scores and first-year GPAs for a sample of N = 15 students
- SAT scores have a M_{SAT} = 580 with SS_{SAT} = 22,400 and GPAs have a M_{GPA} = 3.10 with SS_{GPA} = 1.26, and SP = 84
- Find the regression equation for ***predicting GPA from SAT scores***
- Compute R^2
- Determine if the regression equation accounts for a significant proportion (α = 0.05) of the variance in GPA (i.e., if SAT scores are a significant predictor of GPA)
- Compute the standard error of the estimate ($s_{Y\text{-}\hat{Y}}$)

- Compute R^2
- Calculate the line-of-best-fit and determine if the regression equation, with X as the predictor, accounts for a significant proportion ($\alpha = 0.05$) of the variance in *Y* scores
- Compute the standard error of the estimate ($s_{Y-\hat{Y}}$)

X	Y
5	10
1	4
4	5
7	11
6	15
4	6
3	5
2	0