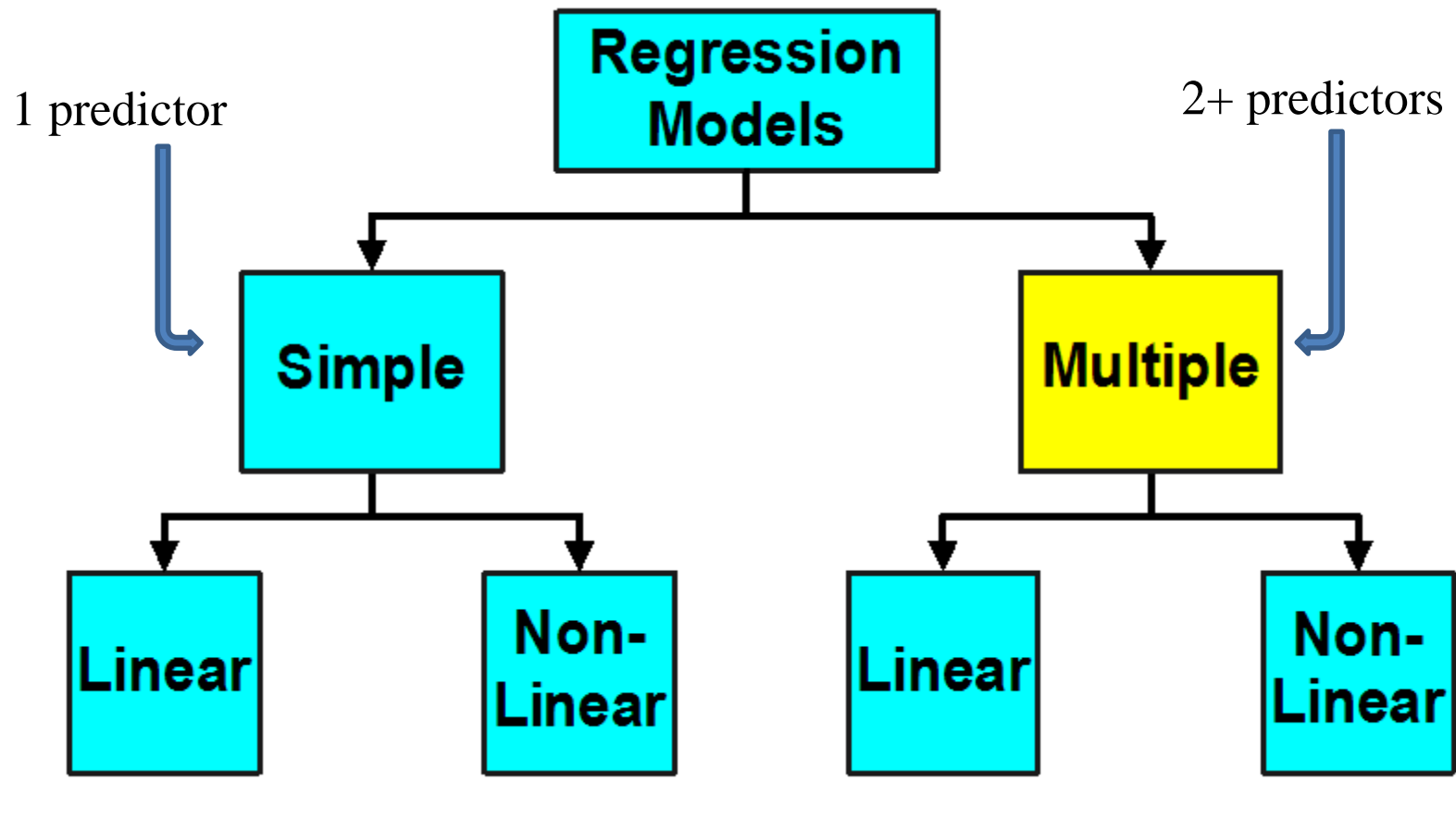


Linear model

Palm beach example

Multiple regression is not just one technique but a family of techniques that can be used to explore the relationship between one continuous dependent variable and a number of independent variables or predictors (usually continuous). Multiple regression is based on correlation but allows a more sophisticated exploration of the interrelationship among a set of variables.

Types of Linear Models



Linear Model of k predictors

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \xi$$

$$Y_i = \beta_0 + \sum_{i=1}^k \beta_i X_i + \xi$$

Linear Model (2 predictors)

The population simple linear regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

where

- ✓ $Y \rightarrow$ **dependent variable**, we wish to explain or predict
- ✓ $X_i \rightarrow$ **independent variables**, also called the **predictors**
- ✓ ε is the **error term**, the only random component in the model
- ✓ $\beta_0, \beta_1, \beta_2$ are the **model parameters which can be estimated using method of least square**

$$\varepsilon \sim n(0, \sigma^2)$$

The Method of Least Squares (Estimating parameters)

The difference between observed and estimated values of dependent variable should be minimum

SSE → Sum of squares of errors should be minimum

Estimating the intercept and slope: least squares estimation

Sum of deviation between observed and estimated values of dependent variable should be minimum

Or

Error sum of square should be minimum

$$SSE = \sum (Y_i - \hat{Y}_i)^2$$

$$SSE = \sum \varepsilon^2$$

$$SSE = \sum (Y_i - \beta_0 - \beta_1 X_i - \beta_2)^2$$

Standard Error of the Estimate

The **standard error of the estimate**, s_e , is defined by

$$s_e = \sqrt{\frac{SSE}{n-2}},$$

where SSE is the error sum of squares.

$$s_e = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{\sum e_i^2}{n-2}}.$$

Multiple linear model in R

Assumptions of Multiple Regression

- Sample size
- Multicollinearity and singularity
- Outliers
- Normality, linearity, homoscedasticity, independence of residuals

Sample size

How many cases or participants do you need?

Different authors tend to give different guidelines concerning the number of cases required for multiple regression. Stevens (1996, p. 72) recommends that 'for social science research, about 15 participants per predictor are needed for a reliable equation'.

Sample size

Tabachnick and Fidell (2007, p. 123) give a formula for calculating sample size requirements, taking into account the number of independent variables that you wish to use:

$N > 50 + 8m$ (where m = number of independent variables). If you have five independent variables, you will need 90 cases. More cases are needed if the dependent variable is skewed.

Multicollinearity and singularity

This refers to the relationship among the independent variables. Multicollinearity exists when the independent variables are highly correlated ($r = .9$ and above). Singularity occurs when one independent variable is actually a combination of other independent variables (e.g. when both subscale scores and the total score of a scale are included).

Multicollinearity and singularity

Multiple regression doesn't like multicollinearity or singularity and these certainly don't contribute to a good regression model, so always check for these problems before you start.

Outliers

Multiple regression is very sensitive to outliers (very high or very low scores). Checking for extreme scores should be part of the initial data screening process. You should do this for all the variables, both dependent and independent, that you will be using in your regression analysis. Outliers can either be deleted from the data set or, alternatively, given a score for that variable that is high but not too different from the remaining cluster of scores.

Outliers

Outliers on dependent variable can be identified from the standardized residual plot that can be requested. Tabachnick and Fidell (2007, p. 128) define outliers as those with standardized residual values above 3.3 (or less than -3.3).

Normality, linearity, homoscedasticity and independence of residuals

These assumptions can be checked from the residuals scatterplots which are generated as part of the multiple regression procedure. Residuals are the differences between the obtained and the predicted dependent variable (DV) scores.

Normality, linearity, homoscedasticity & independence of residuals

The residuals scatterplots allow you to check:

- normality: the residuals should be normally distributed about the predicted DV scores
- linearity: the residuals should have a straight-line relationship with predicted DV scores
- homoscedasticity: the variance of the residuals about predicted DV scores should be the same for all predicted scores

Normality, Linearity, and Homoscedasticity of Residuals

Examination of residuals scatterplots provides a test of assumptions of normality, linearity, and homoscedasticity between predicted DV scores and errors of prediction. Assumptions of analysis are that the residuals are normally distributed about the predicted DV scores, that residuals have a horizontal- line relationship with

Figure 5.1

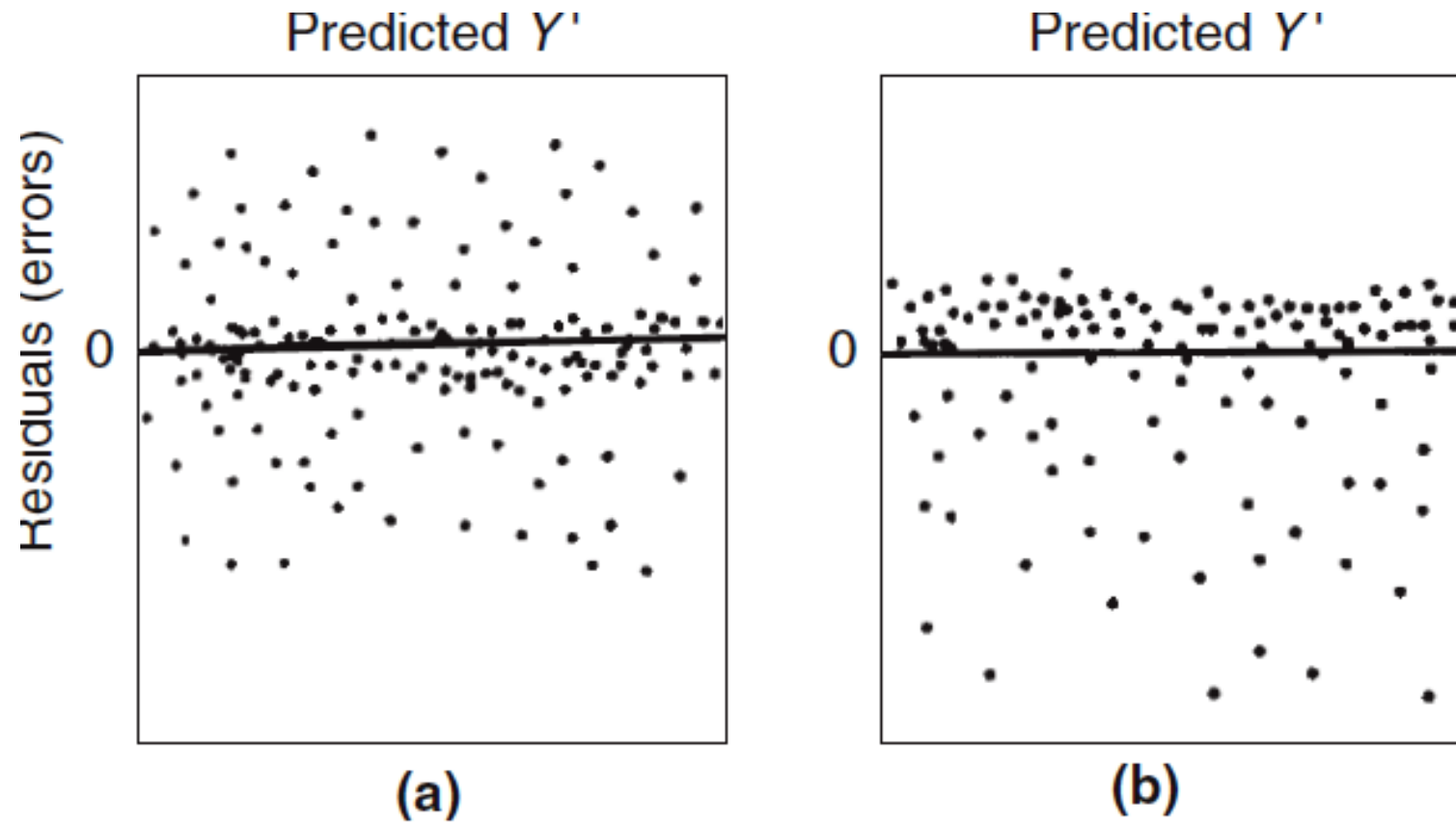
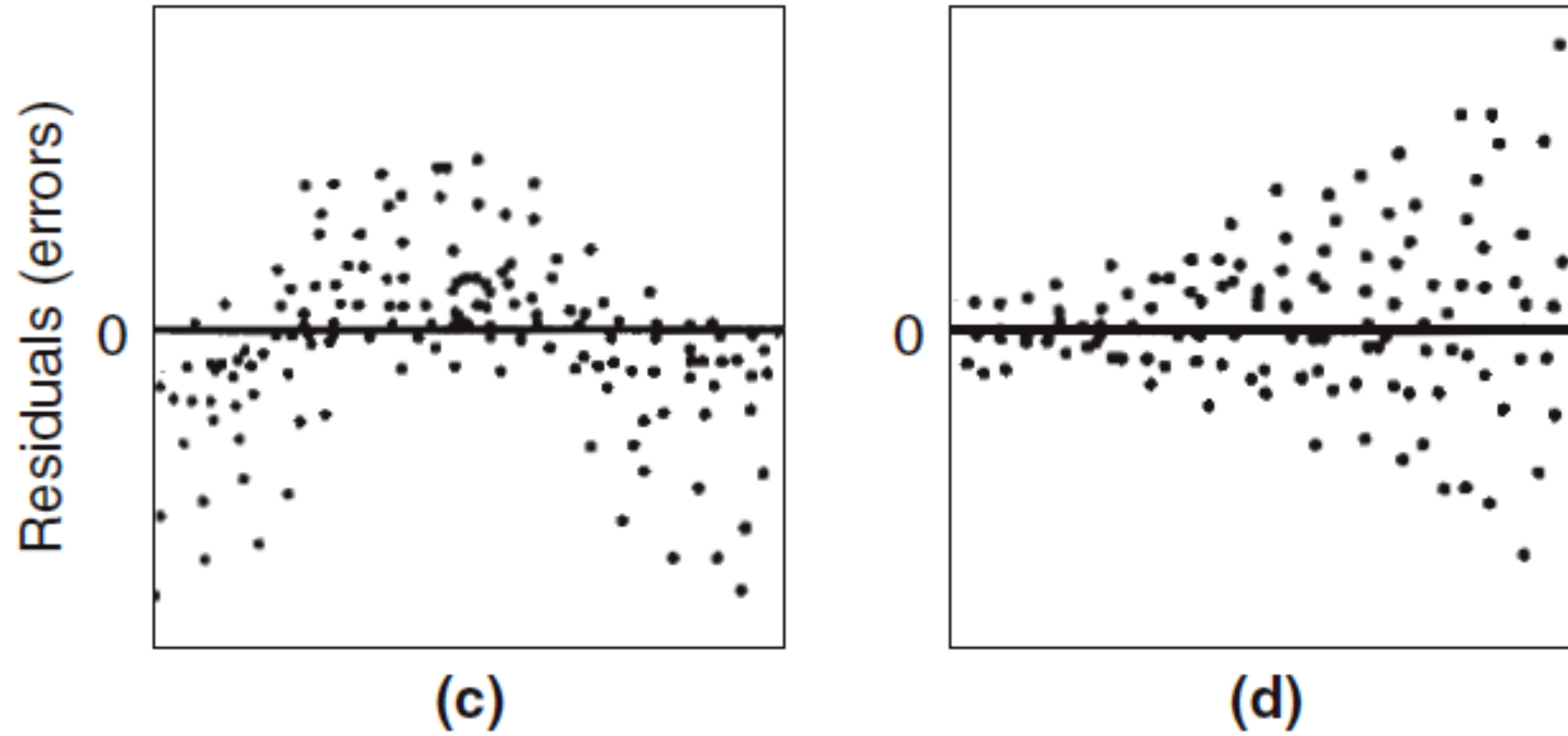


Figure 5.1



Missing Value Analysis

Figure 5.1(a) of residual plots given on previous two slides shows that the assumptions are fulfilled and everything is fine. (cigar shaped with horizontal line)

Figure (b) illustrates a failure of normality, with a skewed distribution of residuals.

Linearity of relationship between predicted DV scores and errors of prediction is also assumed. If nonlinearity is present, the overall shape of the scatterplot is curved instead of rectangular, as shown in Figure (c).

Typical heteroscedasticity is a case in which the band becomes wider at larger predicted values, as illustrated in Fig(d). (cone shaped)

Steps in Multiple Regression Model

- **Step 1** Hypothesize the deterministic component of the model. This component relates the mean, $E(y)$, to the independent variables x_1, x_2, \dots, x_k . This involves the choice of the independent variables to be included in the model.
- **Step 2** Use the sample data to estimate the unknown model parameters $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ in the model.
- **Step 3** Specify the probability distribution of the random error term, ε , and estimate the standard deviation of this distribution, σ .

Steps in Multiple Regression Model

- **Step 4** Check that the assumptions on ε are satisfied, and make model modifications if necessary.
- **Step 5** Statistically evaluate the usefulness of the model.
- **Step 6** When satisfied that the model is useful, use it for prediction, estimation, and other purposes.

Palm Beach Atlantic University

University students often complain that universities reward professors for research but not for teaching, and argue that professors react to this situation by devoting more time and energy to the publication of their findings and less time and energy to classroom activities. Professors counter that research and teaching go hand in hand; more research makes better teachers. A student organization at Palm Beach Atlantic University decided to investigate the issue.

Palm Beach Atlantic University

They randomly selected 50 psychology professors employed by PBA. The students recorded the salaries of professors, their average teaching evaluations (on a 10-point scale), and the total number of journal articles published in their careers. These data are stored in columns 1 to 3, respectively. Perform a complete analysis (produce the regression equation, assess it, and diagnose it) and report your findings

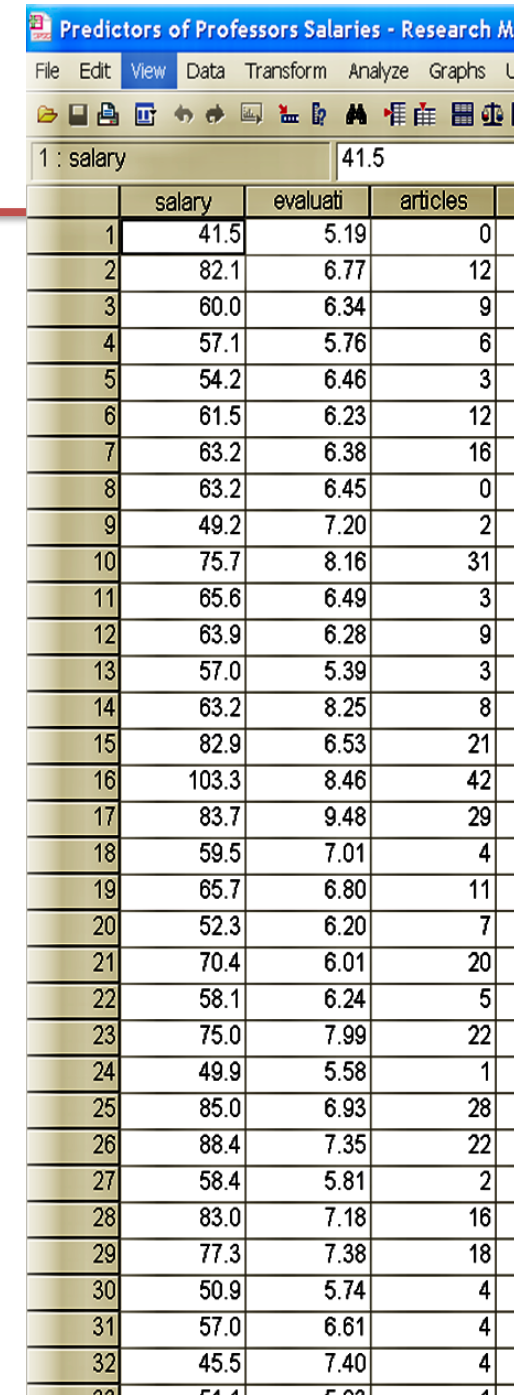
Example

$Y \rightarrow$ Salaries of Professors

The variables that we've isolated
as influencing the salaries:

$X_1 \rightarrow$ Course Evaluations

$X_2 \rightarrow$ Number of Publications



	salary	evaluati	articles
1	41.5	5.19	0
2	82.1	6.77	12
3	60.0	6.34	9
4	57.1	5.76	6
5	54.2	6.46	3
6	61.5	6.23	12
7	63.2	6.38	16
8	63.2	6.45	0
9	49.2	7.20	2
10	75.7	8.16	31
11	65.6	6.49	3
12	63.9	6.28	9
13	57.0	5.39	3
14	63.2	8.25	8
15	82.9	6.53	21
16	103.3	8.46	42
17	83.7	9.48	29
18	59.5	7.01	4
19	65.7	6.80	11
20	52.3	6.20	7
21	70.4	6.01	20
22	58.1	6.24	5
23	75.0	7.99	22
24	49.9	5.58	1
25	85.0	6.93	28
26	88.4	7.35	22
27	58.4	5.81	2
28	83.0	7.18	16
29	77.3	7.38	18
30	50.9	5.74	4
31	57.0	6.61	4
32	45.5	7.40	4

R code

```
>myDataFr=data.frame(read.csv("f://courses/data  
mining/palmbeachatlantic.csv"))  
  
>View(myDataFr)  
  
>plot(myDataFr)  
  
>palmBeachLM <-  
      lm(myDataFr$Salary~myDataFr$Evalua  
tion+myDataFr$Articles)  
  
>plot(fitted(palmBeachLM),resid(palmBeachLM))  
  
>qqnorm(resid(palmBeachLM))
```

Correlation matrix

```
> cor(data)
      salary      art      Eval
salary 1.0000000 0.8478048 0.5774398
art     0.8478048 1.0000000 0.6393230
Eval    0.5774398 0.6393230 1.0000000
```

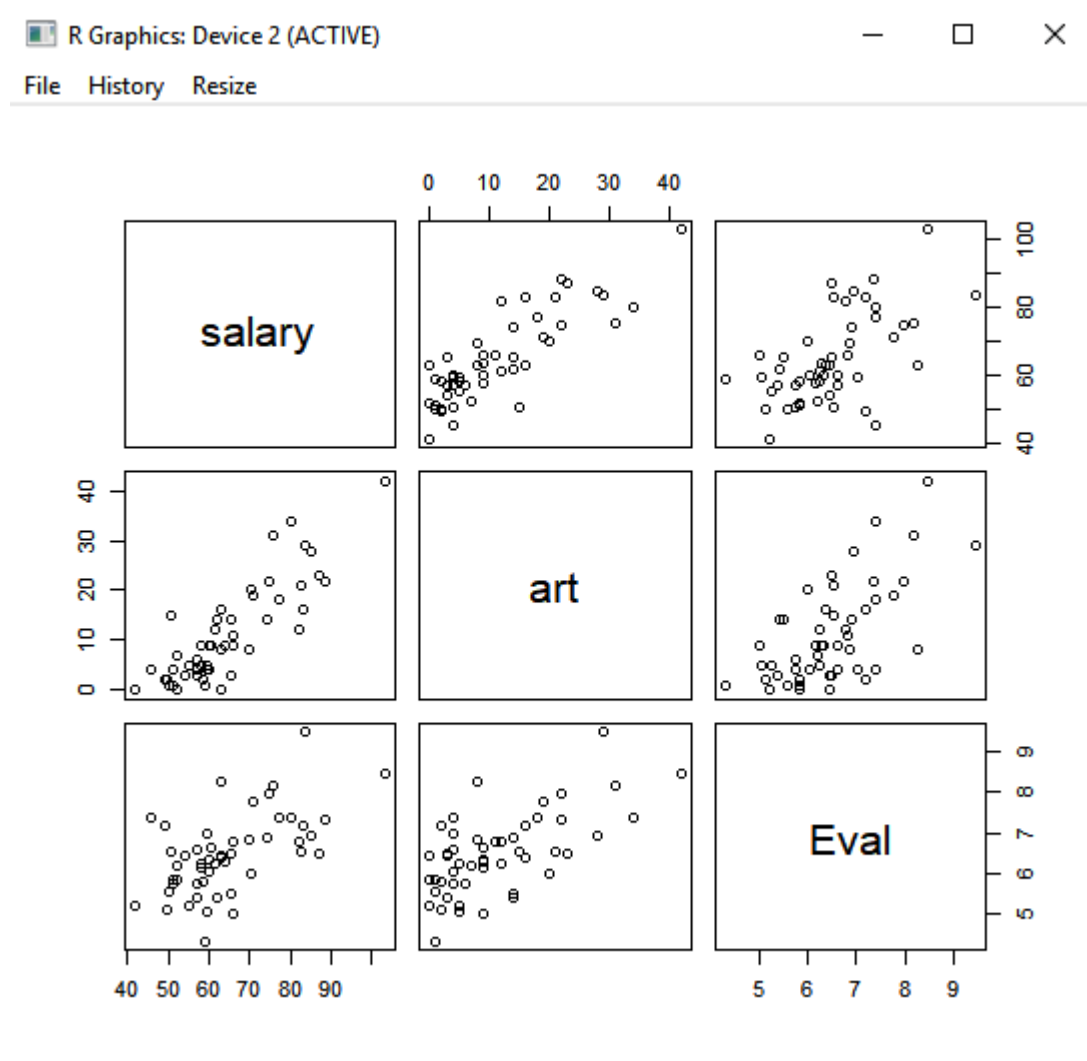
Descriptive Statistics

	Mean	Std. Deviation	N
Salary (\$1,000)	64.694	12.9929	50
Average Teaching Evaluation (10-point scale)	6.4814	1.00267	50
Number of Published Articles	11.20	9.904	50

Correlations

		Salary (\$1,000)	Average Teaching Evaluation (10-point scale)	Number of Published Articles
Pearson Correlation	Salary (\$1,000)	1.000	.577	.848
	Average Teaching Evaluation (10-point scale)	.577	1.000	.639
	Number of Published Articles	.848	.639	1.000
Sig. (1-tailed)	Salary (\$1,000)	.	.000	.000
	Average Teaching Evaluation (10-point scale)	.000	.	.000
	Number of Published Articles	.000	.000	.
N	Salary (\$1,000)	50	50	50
	Average Teaching Evaluation (10-point scale)	50	50	50
	Number of Published Articles	50	50	50

Matrix Scatterplots



```
call:
lm(formula = myDataFr$Salary ~ myDataFr$Evaluation + myDataFr$Articles)

Coefficients:
      (Intercept)  myDataFr$Evaluation  myDataFr$Articles
          47.7682             0.7762             1.0620
```

The unstandardized regression equation is shown here:
 $Y' = 47.768 + 0.776(X_1) + 1.062(X_2)$

summary(palmBeachLM)

Console ~/ ↗

call:

```
lm(formula = salary ~ Eval + art)
```

Residuals:

Min	1Q	Median	3Q	Max
-18.2597	-3.7048	-0.9026	4.2693	16.3324

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	47.7682	7.6256	6.264	1.07e-07	***
Eval	0.7762	1.2987	0.598	0.553	
art	1.0620	0.1315	8.078	1.95e-10	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.009 on 47 degrees of freedom

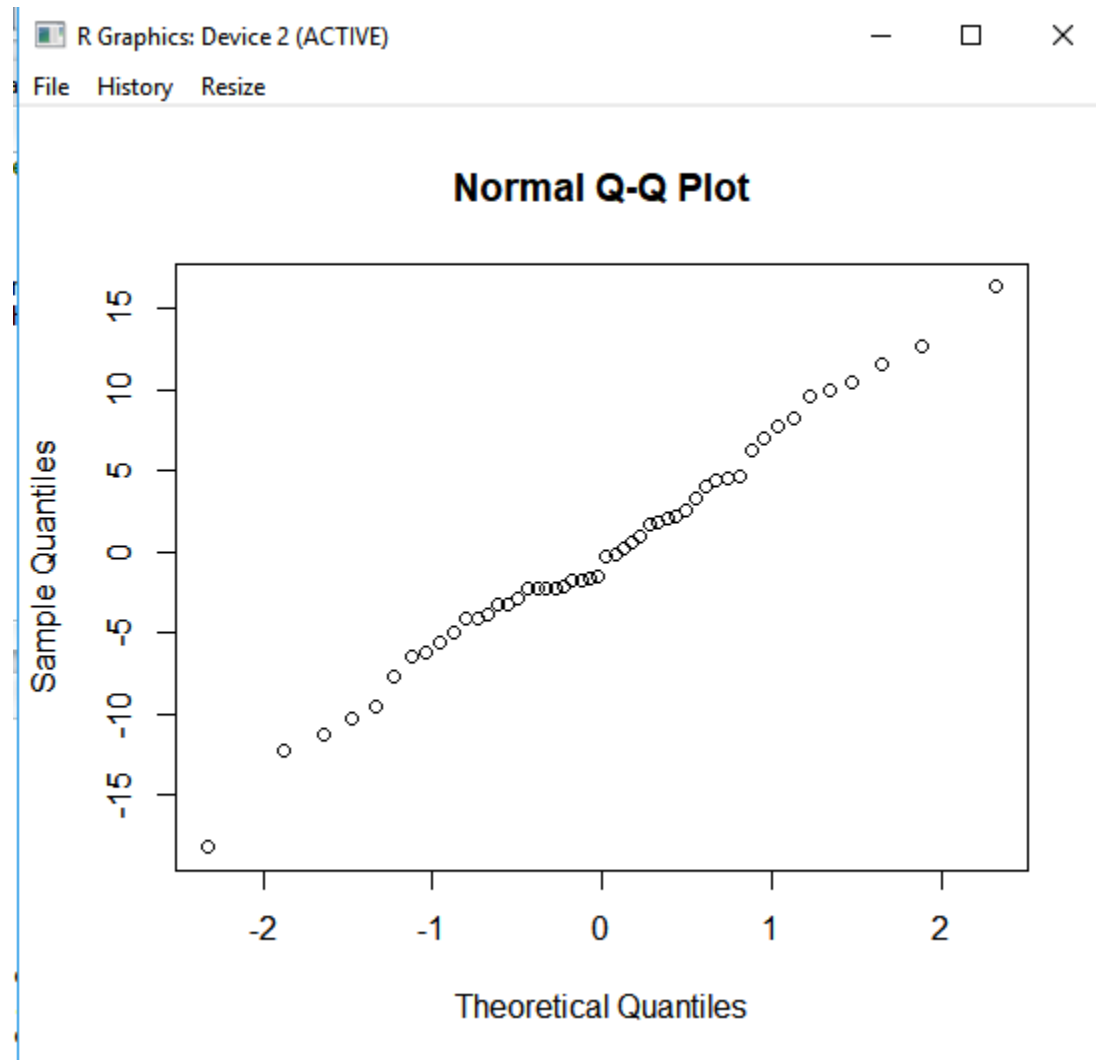
Multiple R-squared: 0.7209, Adjusted R-squared: 0.709

F-statistic: 60.7 on 2 and 47 DF, p-value: 9.453e-14

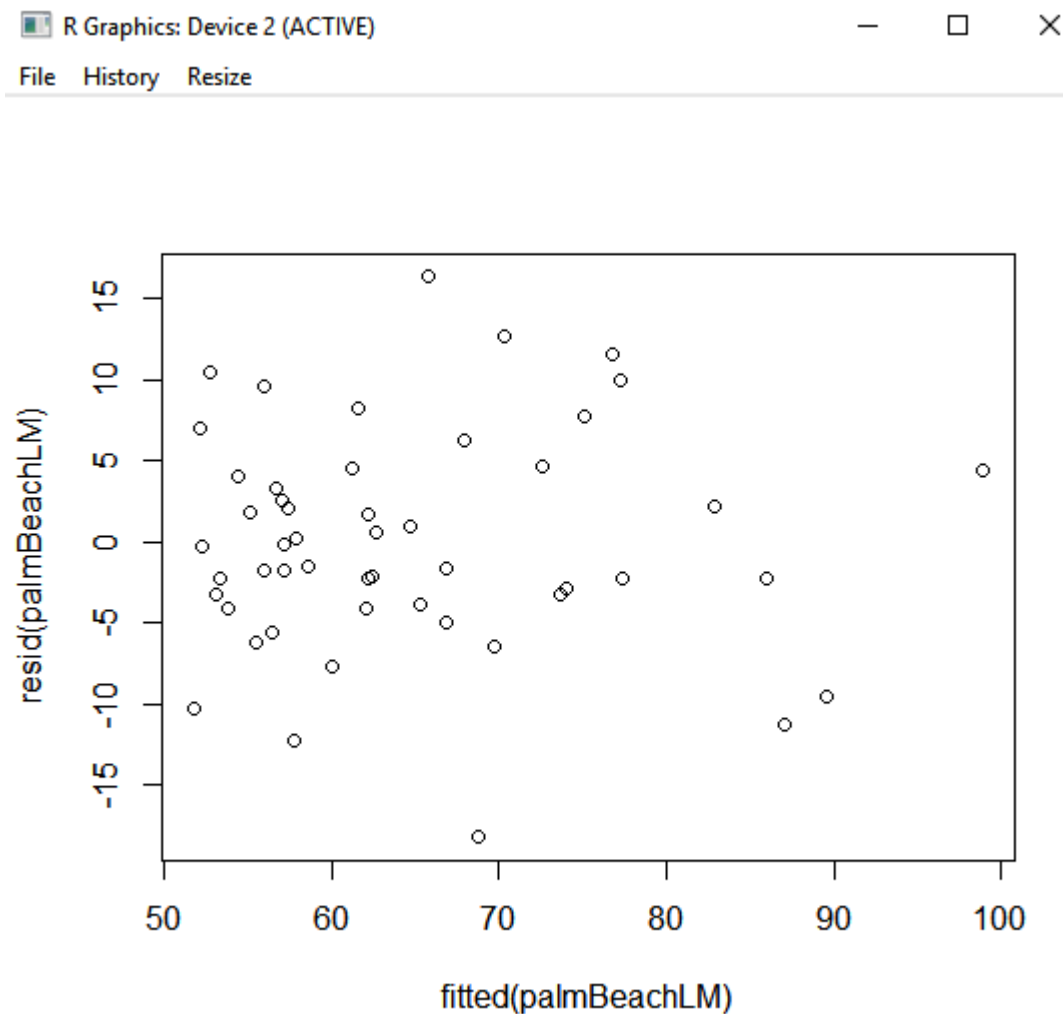
Now we must determine the fit of the model by estimating R^2 and statistically testing all the estimates. The squared multiple correlation coefficient is analogous to r^2 the term used in SLR. This value tells us how much of the variance accounted for in Y is attributable to the independent variables. In our example, R^2 is 0.721. This indicates that a meaningful 72.1% of the variance in the model is explained. 72.1% of the variance in salaries can be attributed to teaching performance and publication record.

The remaining step in the evaluation of the regression equation is to estimate the contribution of each variable in the study. If all the b values are significant

Normal prob. plot



Residual plot



Questions?
