Selected Topics in Data Analysis

Raim, Ostrouchov, Neerchal

UMBC

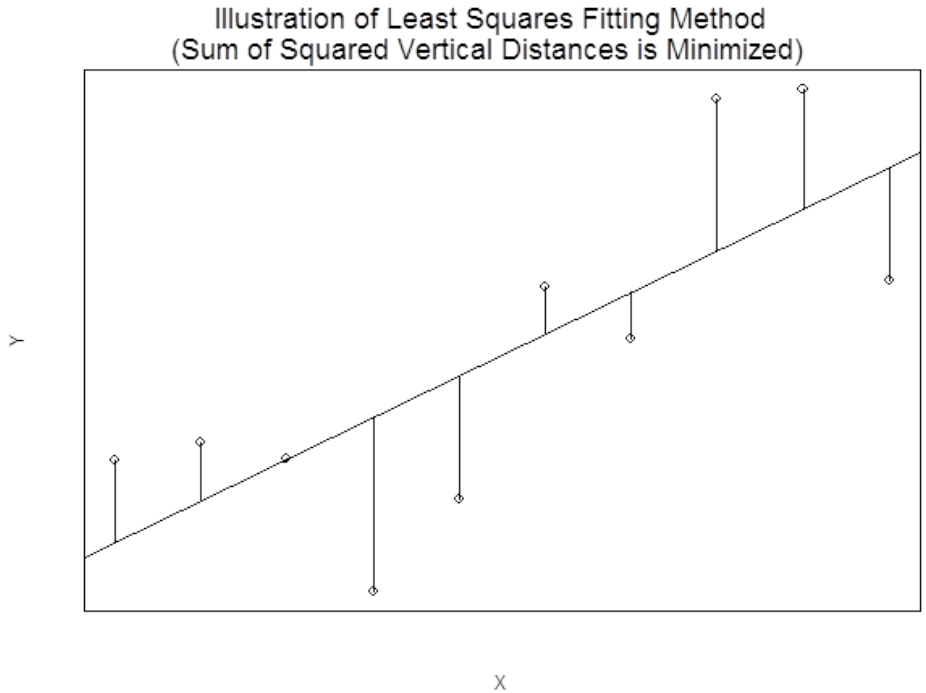
Introduction

1. Linear Regression
2. ANOVA
3. Logistic Regression

What is Regression?

- A way of predicting the value of one variable (dependent) from another variable (independent)
- It is a hypothetical relationship, an approximation at best
- A model for this relationship is assumed, usually linear
- That is, $Y \approx \beta_0 + \beta_1 X$

Method of Least Squares



$$\text{Min} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

$$\text{Slope } \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

$$\text{Intercept } \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Assumptions Underlying LR

- The model is correct: $E(Y \mid X) = \beta_0 + \beta_1 X$
- Variability is constant: $V(Y \mid X) = \sigma^2$
- Data are uncorrelated: $Cov(Y_i, Y_j) = 0$ for $i \neq j$
- Data are Gaussian: $Y_i \mid X \sim \text{Normal}(\beta_0 + \beta_1 X_i, \sigma^2)$

Under the LR Assumptions..

$$\begin{pmatrix} \hat{\beta}_0 \\ \\ \hat{\beta}_1 \end{pmatrix} \sim \begin{pmatrix} \beta_0 \\ \\ \beta_1 \end{pmatrix}, \quad \sigma^2 \begin{pmatrix} \frac{1}{n} + \frac{\bar{X}^2}{S_{XX}} & -\frac{\bar{X}}{S_{XX}} \\ \\ -\frac{\bar{X}}{S_{XX}} & \frac{1}{S_{XX}} \end{pmatrix}$$

where $S_{XX} = \sum_{i=1}^{n}(X_i - \bar{X})^2$.

- The usual inference procedures are carried out:
  $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$

  t-statistic: $t = \frac{\hat{\beta}_1}{SE_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{S_{XX}}}$ where , $\hat{\sigma^2} =$ Residual SS$= \sum_{i=1}^{n}(Y_i - \hat{Y})^2$

  Test: Reject H_0 if $\mid t \mid > t_{\alpha/2,n-2}$

- The ANOVA table, F-test, R-squared statistic, are obtained.

Album Sales Example

- From Field, Miles and Field (2012).
- Y: Album sales (CDs and downloads) in the week after release
- X: The amount (in units of £1000) spent promoting the record before release
- Data consists of 200 different music album releases.
- Objectives:
  - to assess the impact of promotion expediture on album sales
  - to predict album sales from promotion expenditure

Album Sales Example: Simple Linear Regression

. . . Demonstration . . .

(See `AlbumSalesLinearReg.Rmd`)

Prediction and Confidence Intervals

- Predicting Y corresponding to a given value of the predictor $X = x$

$$\text{Point Prediction:} \quad \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

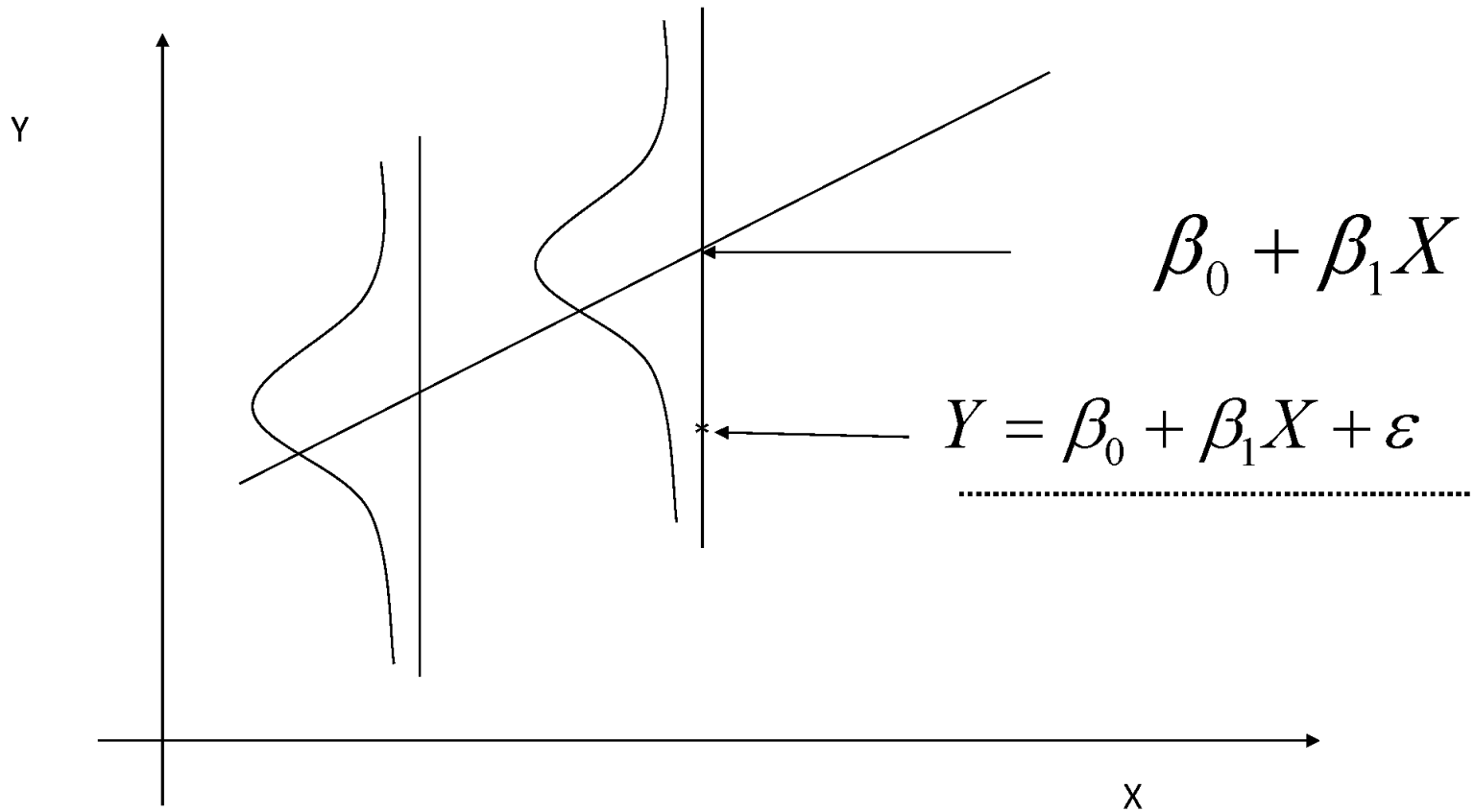
$$\text{Prediction Interval:} \quad \hat{\beta}_0 + \hat{\beta}_1 x \pm t_{\alpha/2,n-2}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{X})^2}{S_{xx}}}$$

- Estimating conditional mean $E(Y \mid X = x)$

$$\text{Point Estimate:} \quad \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$\text{Interval Estimate:} \quad \hat{\beta}_0 + \hat{\beta}_1 x \pm t_{\alpha/2,n-2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x-\bar{X})^2}{S_{xx}}}$$

Prediction and Confidence Intervals



Album Sales Example: Simple Linear Regression

. . . Demonstration . . .

(See `AlbumSalesLinearRegPred.Rmd`)

Multiple Linear Regression

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \epsilon$$

- The model is correct: $E(Y \mid X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$
- Variability is constant: $V(Y \mid X) = \sigma^2$
- Data are uncorrelated: $Cov(Y_i, Y_j) = 0$ for $i \neq j$
- Data are Gaussian: $Y_i \mid X \sim$ Normal

Multiple Linear Regression

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \epsilon$$

- Explaining or predicting a dependent variable as a quadratic or higher degree polynomial of an explanatory variable
  $E(Y \mid X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$

- Using other transformations of explanatory variables in the model
  $E(Y \mid X) = \beta_0 + \beta_1 X + \beta_2 log(X) + \beta_3 X^2$

- Capturing interaction between explanatory variables using cross-products in the regression model
  $E(Y \mid X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$

- Using categorical independent variables
  $E(Y \mid X) = \beta_0 + \beta_1 \text{ IND(Gender?)} + \beta_2 \text{ IND(Surgery?)} + \beta_3 \text{Age}$

MLR: First things first..

- Obtain summary of the data
  - Examine univariate graphics
  - Examine pairwise relationships
- Obtain multivariate summary of data
  - Correlations
  - Partial correlations
- Examine multivariate graphics
  - Pairwise scatterplots
  - Scatterplot matrix
  - Contour plots

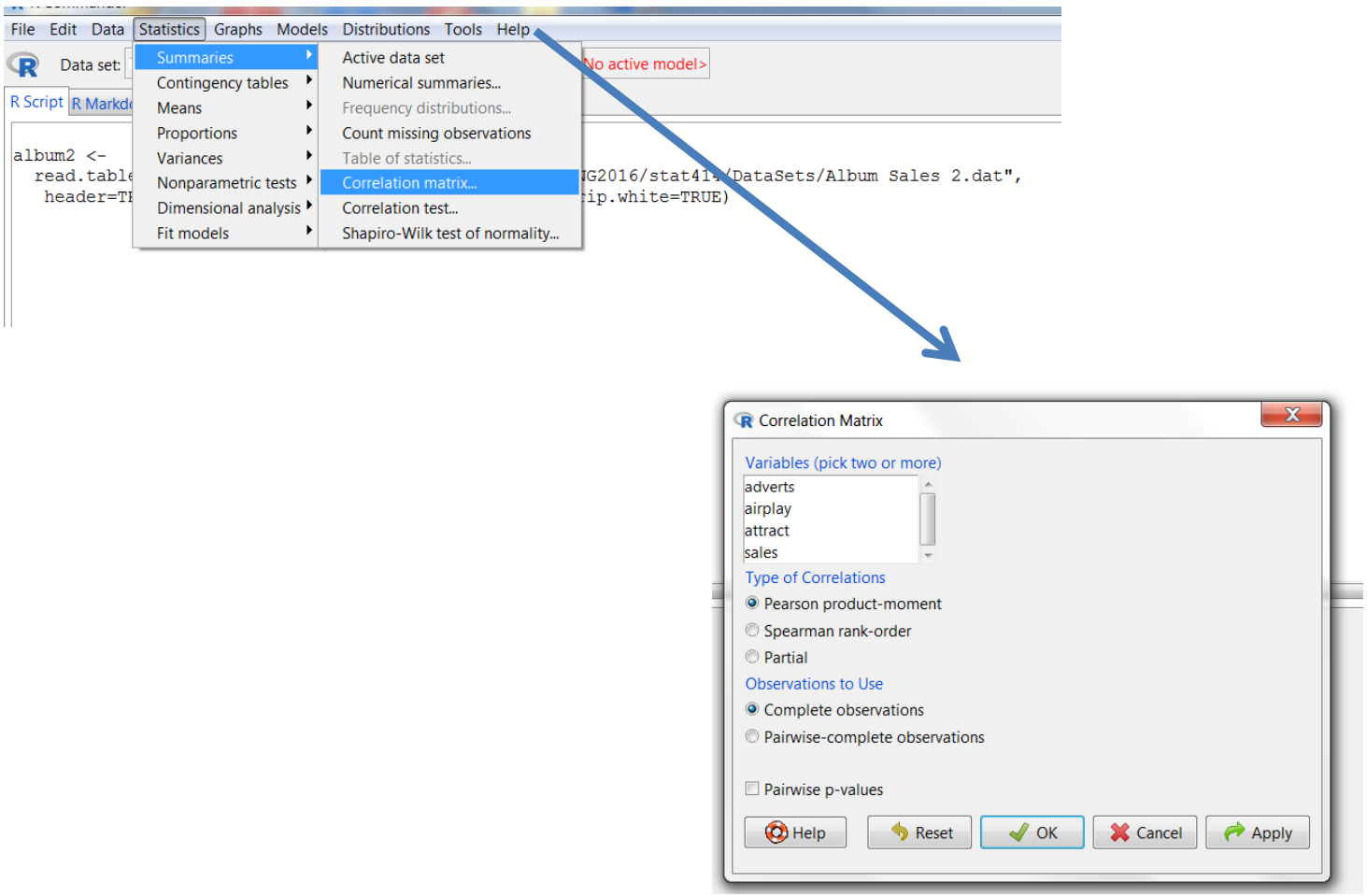
Album Sales Example

- From Field, Miles and Field (2012).
- Y: Album sales (CDs and downloads) in the week after release
- X1: Advert (same as before)
- X2: Airplay (amount of time the songs from the CD were played on air)
- X3: Attract (A score fore the attractiveness of the cover design)
- Data consists of 200 different music album releases.
- Objectives:
  - to assess the impact of promotion expediture on album sales
  - to predict album sales from promotion expenditure

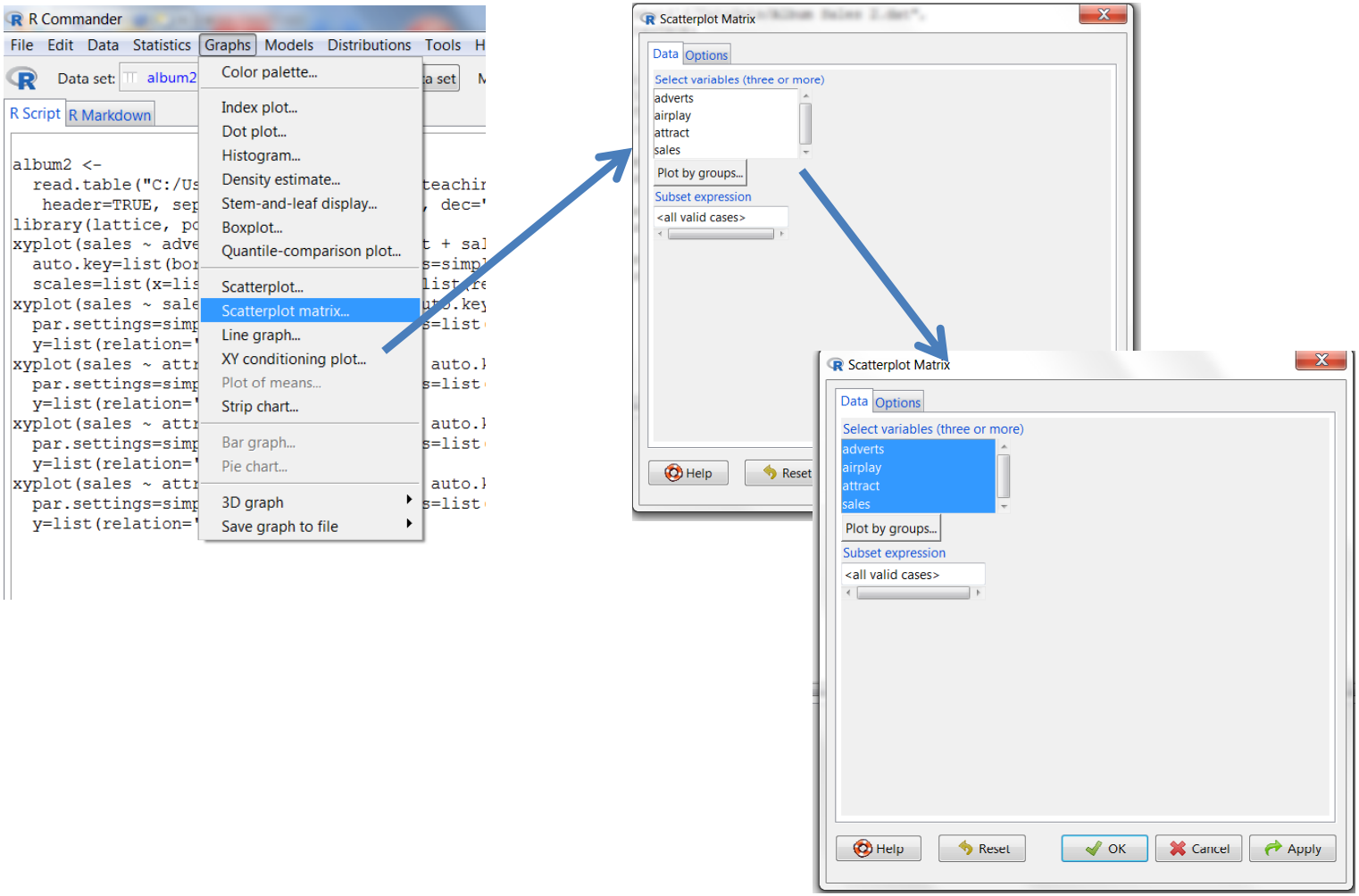
Using Rcmdr for quick (and ...) look

- A package called `Rcmdr` is useful for playing around with small size data set
- Author: John Fox, MacMaster University, Canada
- Textbook: Using the `R` Commander: A Point-and-Click Interface for `R`, Chapman and Hall-CRC Press, 2017.
- `Rcmdr` is installed like any other package
- Warning: `Rcmdr` will open its own window and will require some care when it is launched from within `RStudio`

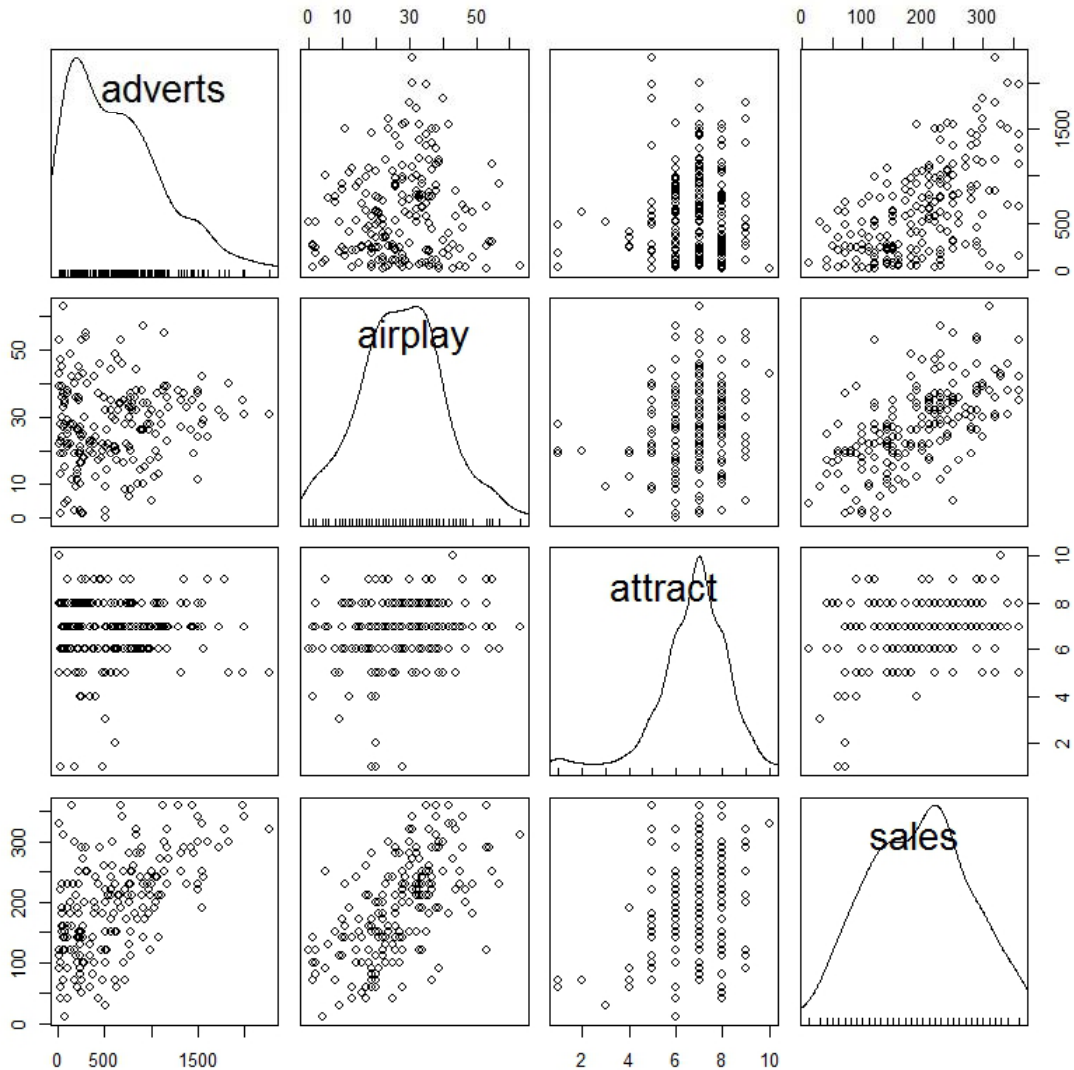
Using Rcmdr for quick (and ...) look



Using Rcmdr for quick (and ...) look



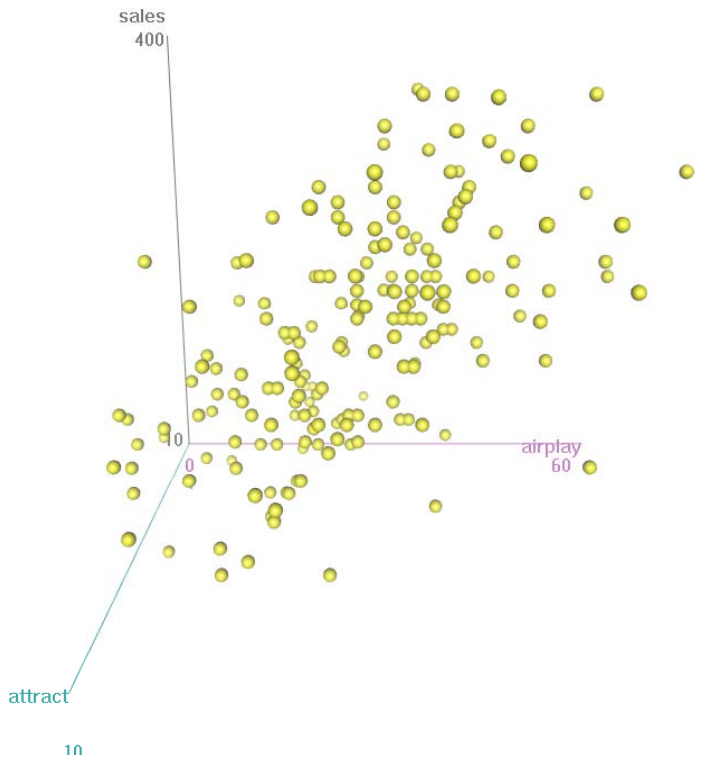
Using Rcmdr for quick (and ...) look



Using Rcmdr for quick (and ...) look

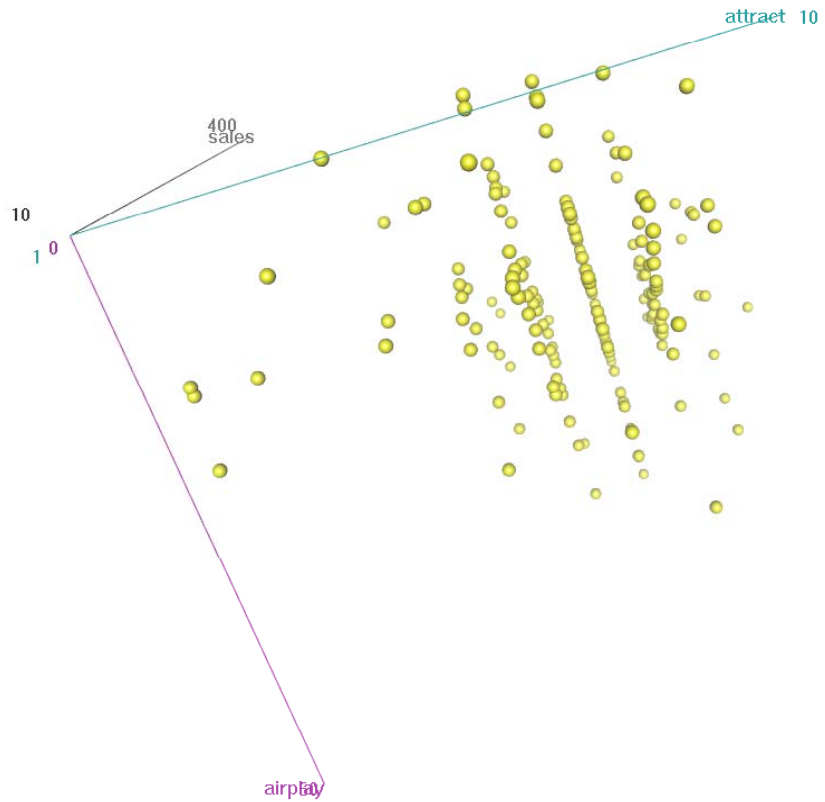
Graphs > 3D Scatterplot

Choose one response variable and two explanatory variables



Using Rcmdr for quick (and ...) look

You can rotate the graph to examine the scatter-cloud….



Method of Least Squares

$$\text{Minimze} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2} - \cdots - \beta_k X_{ik})^2$$

Solutions are obtained from the Normal Equations:

$$\begin{pmatrix} n & \sum X_1 & \dots & \sum X_k \\ \sum X_1 & \sum X_1^2 & \dots & \sum X_1 X_k \\ \dots & \dots & \dots & \dots \\ \sum X_k & \sum X_k X_1 & \dots & \sum X_k^2 \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_0 \\ \dots \\ \hat{\beta}_0 \end{pmatrix} = \begin{pmatrix} \sum Y \\ \sum X_1 Y \\ \dots \\ \sum X_k Y \end{pmatrix}$$

MLR for the Album Sales Data

```
album2 <- read_csv("AlbumSales2.csv")
album2.out <- lm(sales~adverts+airplay+attract,data=album2)
```

```
Call:
lm(formula = sales ~ adverts + airplay + attract, data = album2)

Residuals:
     Min       1Q   Median       3Q      Max
-121.324  -28.336   -0.451   28.967  144.132

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -26.612958  17.350001  -1.534    0.127
adverts       0.084885   0.006923  12.261  < 2e-16 ***
airplay       3.367425   0.277771  12.123  < 2e-16 ***
attract      11.086335   2.437849   4.548 9.49e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 47.09 on 196 degrees of freedom
Multiple R-squared:  0.6647,    Adjusted R-squared:  0.6595
F-statistic: 129.5 on 3 and 196 DF,  p-value: < 2.2e-16
```

MLR for the Album Sales Data

```
summary(album2.out)
anova(album2.out)
```

```
> summary(album2.out)

Call:
lm(formula = sales ~ adverts + airplay + attract, data = album2)

Residuals:
     Min       1Q   Median       3Q      Max
-121.324  -28.336   -0.451   28.967  144.132

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -26.612958  17.350001  -1.534    0.127
adverts       0.084885   0.006923  12.261   < 2e-16 ***
airplay       3.367425   0.277771  12.123   < 2e-16 ***
attract      11.086335   2.437849   4.548 9.49e-06 ***
---
Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  ' '  1

Residual standard error: 47.09 on 196 degrees of freedom
Multiple R-squared:  0.6647,    Adjusted R-squared:  0.6595
F-statistic: 129.5 on 3 and 196 DF,  p-value: < 2.2e-16
```

MLR for the Album Sales Data

```
> anova(album2.out)

Analysis of Variance Table

Response: sales
           Df  Sum Sq Mean Sq F value    Pr(>F)
adverts     1  433688  433688 195.600 < 2.2e-16 ***
airplay     1  381836  381836 172.214 < 2.2e-16 ***
attract     1   45853   45853  20.681 9.492e-06 ***
Residuals 196  434575    2217
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

Goodness-of-fit Measures

- Residual Sum of Squares
  - Predicted Value: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \cdots + \hat{\beta}_k X_{ik}$
  - Residual: $r_i = Y_i - \hat{Y}_i$
  - Residual Standard Error: $\sqrt{\frac{\sum (Y_i - \hat{Y}_i)_i^2}{n-(k+1)}}$

- Coefficient of Determination
  - $R^2 = \frac{SSTotal - RSS}{SSTotal}$
  - $0 \leq R^2 \leq 1$
  - R^2 is the square of the correlation coeffient between Y and its best predictor based on all X's
  - Stein's formula for adjusted $R^2 = 1 - \left(\frac{n-1}{n-k-1}\right)\left(\frac{n-2}{n-k-2}\right)\left(\frac{n+1}{n}\right)(1 - R^2)$
  - R uses $R^2 = 1 - \left(\frac{n-1}{n-k}\right)(1 - R^2)$

Testing for Significance of the Regression

Source	Degrees of freedom	Sum of Squares	Mean Squares	F, df,p-value
Model	k	SS_M	MS_M	
Residuals/Error	n-k-1	SS_R	MS_R	
Total	n-1			

- Testing for all coeffiencients
  - $H_0 : \beta_1 = 0, \beta_2 = 0, \cdots, \beta_k = 0$, H_1 : not H_0
  - Under H_0, $F \sim F_{k,n-k-1}$; Reject H_0 if $F > F_{\alpha,(k,n-k-1)}$
  - $F = 129.5$, on $(3, 196)$. $p - value = 2.2E - 16$
  - Conclusion: Regression is significant
  - Tests for individual coefficients can be read from the output also

Testing for Significance of Individual β's

```
Call:
lm(formula = sales ~ adverts + airplay + attract, data = album2)

Residuals:
     Min       1Q   Median       3Q      Max
-121.324  -28.336   -0.451   28.967  144.132

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -26.612958  17.350001  -1.534    0.127
adverts       0.084885   0.006923  12.261  < 2e-16 ***
airplay       3.367425   0.277771  12.123  < 2e-16 ***
attract      11.086335   2.437849   4.548 9.49e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 47.09 on 196 degrees of freedom
Multiple R-squared:  0.6647,    Adjusted R-squared:  0.6595
F-statistic: 129.5 on 3 and 196 DF,  p-value: < 2.2e-16
```

Comparing Nested Models

Full Model: $Y \approx \beta_0 + \beta_1 X_1 + \cdots + \beta_g X_g + \cdots + \beta_k X_k$

H_0 $\beta_{g+1} = \beta_{g+2} = \cdots = \beta_k = 0$

Reduced Model: $Y \approx \beta_0 + \beta_1 X_1 + \cdots + \beta_g X_g$

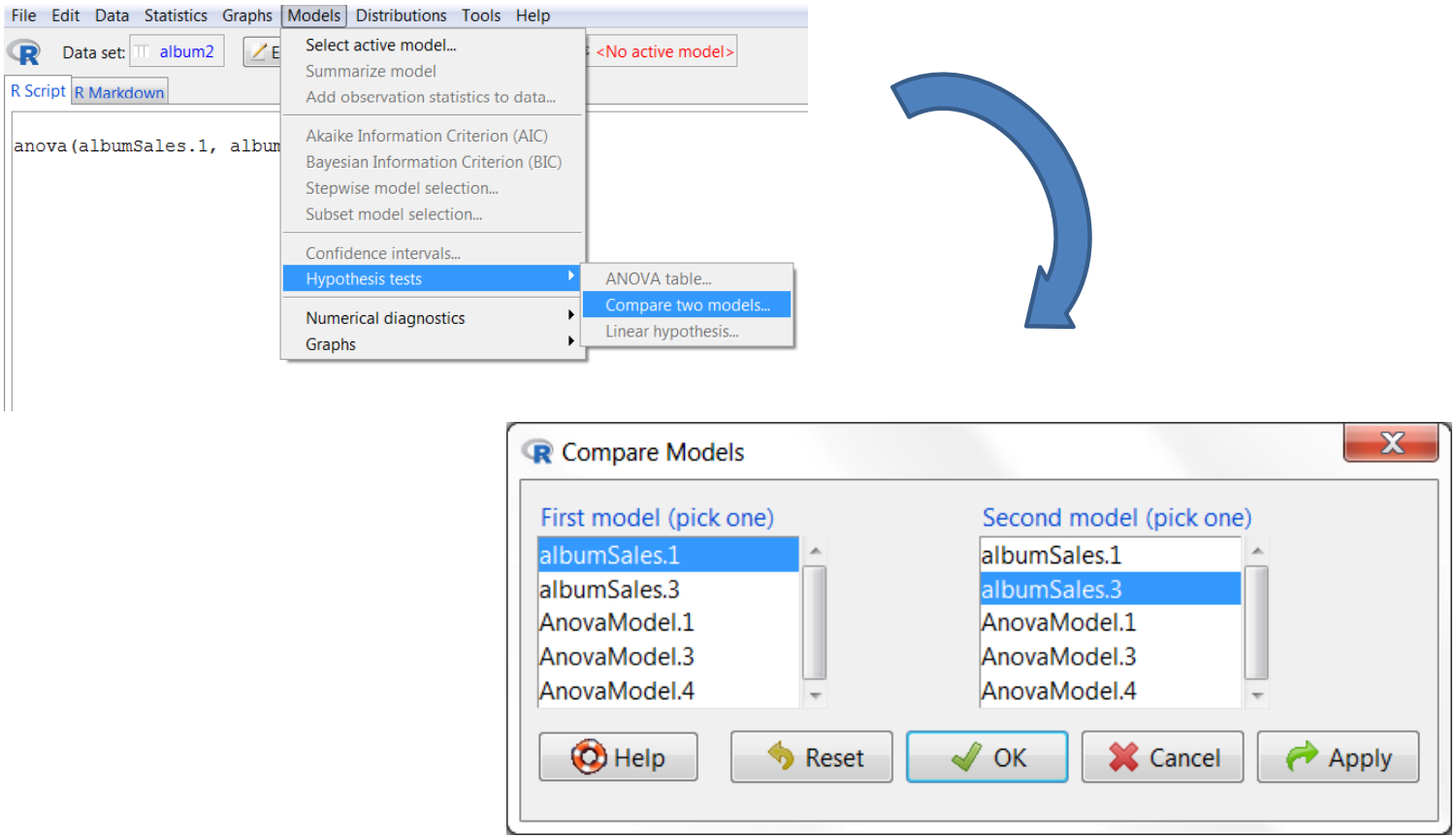
```
> albumSales.3 <- lm(sales ~adverts+airplay+attract,data=album2)
> albumSales.1 <- lm(sales ~adverts,data=album2)
> anova(albumSales.1,albumSales.3)
Analysis of Variance Table

Model 1: sales ~ adverts
Model 2: sales ~ adverts + airplay + attract
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    198 862264
2    196 434575  2    427690 96.447 < 2.2e-16 ***
---
Signif. codes:  0  '***' 0.001  '**' 0.01  '*' 0.05  '.' 0.1  ' ' 1
```

Comparing Nested Models in Rcmdr

Comparing nested models in Rcmdr

Testing for subsets of coefficients



Model Selection Criteria

- Akaike information criteria (AIC)
- Adjusted R-squared
- Cross-validation
- Leave-one-out (LOO)

Akaike information criteria (AIC)

$$AIC = n \ln\left(\frac{RSS}{n}\right) + 2k$$

where RSS is the sum of squared residuals in a model with k parameters.

- As we include more predictors (k), SSE decreases and therefore the first term decreases
- As we include more predictors (k), the second term increases
- AIC is trying to strike a trade off between goodness of fit (SSE) and parsimony (k)
- Related quanties are Schwarz information criteria, Bayesian information criteria etc
- Prefer models which give smaller AIC values

Cross-Validation

- Data Splitting
  - Training dataset: a subset of the original data set used for estimating the model to be evaluated
  - Test dataset: a subset of the original data set, set aside for evaluating the goodness of fit of the model

- Appropriate for examining how well does the model generalize from sample to population

- In principle, any reasonable goodness of fir criteria may be used. Typically, either SSE, or R2 , or adjusted R2 or similar prediction evaluation criteria is used.

- Typically, test data set may be based on a scientifically meaningful criteria (such as use data before a certain date for training, and data after that date for testing)

Cross-Validation

- Test data set may also be chosen at random. In this case, one may repeat the training-test exercise for a variety of choices of test data. [k-fold]

- Rules of thumb
  - At least 10 observations per parameter (Field: 50+8k)
  - Test dataset size is approximately 20% of the original dataset
  - In R,

```
      # K-fold cross-validation
            library(DAAG)
            cv.lm(df=mydata, fit, m=3) # 3 fold cross-
    validation
```

  - LOO is the extreme cas: Use $LOOR_i = Y_i - \hat{Y}_{i,-i}$, where $\hat{Y}_{i,-i}$ is the prediction of Y_i based on all other observations

```
            cv.lm(df=mydata, fit, m=n) # Leave-One-Out (LOO)
```

Variable Selection

- Forward
  - Start with a model containing one predictor with the corresponding the highest
  - Sequentially add predictors providing the largest increase in Variable once entered stays in R^2
- Backward
  - Start with the model including all predictors
  - Sequentially drop predictors providing the least decrease in R^2
  - Dropped variables stay out
- Stepwise
  - Start the forward method
  - Sequentially add predictors providing the largest increase in R^2
  - At each step evaluate all variables in the model for dropping as in the backward method
- stepAIC() in the `MASS` package; `leaps` package for all-subsets
- These are traditional methods, more sophisticated methods will be discussed later

Residual Diagnostics

- Outliers
  - An observation with large residual
  - An observation whose dependent-variable value is unusual given its values on the predictor variables
  - An outlier may indicate a sample peculiarity or may indicate a data entry error or other problem
- Leverage
  - An observation with an extreme value on a predictor variable
  - Leverage is a measure of how far an independent variable deviates from its mean
  - These leverage points can have an effect on the estimate of regression coefficients
- Influence
  - Influence can be thought of as the confluence of leverage and outlier property
  - Removing the observation substantially changes the estimate of coefficients

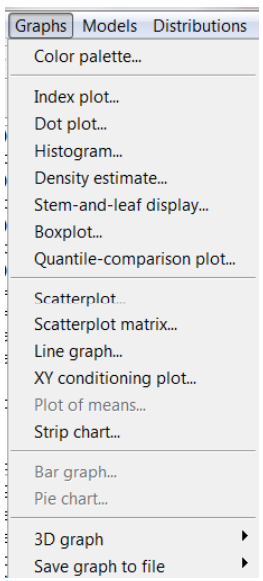
Checking Assumptions

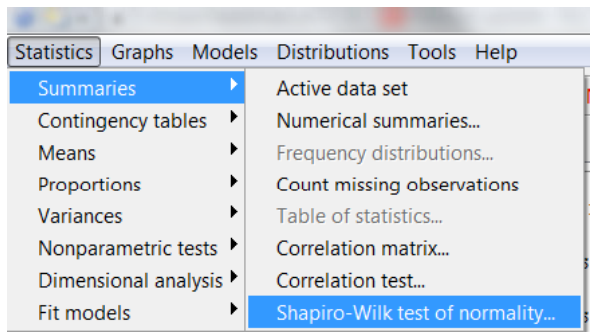
- Linearity of the model
- Distributional assumptions on the errors
  - mean zero?
  - constant variance?
  - Gaussian?
- Independence of errors

Residual Diagnostics

Checking normality

- First save the residuals to a dataframe:
  - album2$resid <- resid(albumSales.3)
- Now use the Rcmdr graphics and statistics





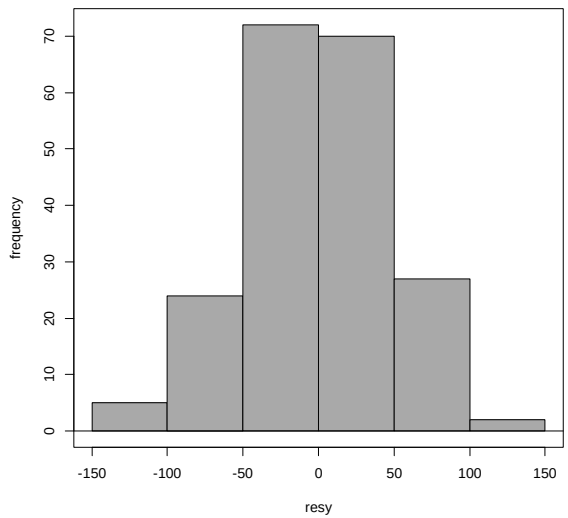
Residual Diagnostics

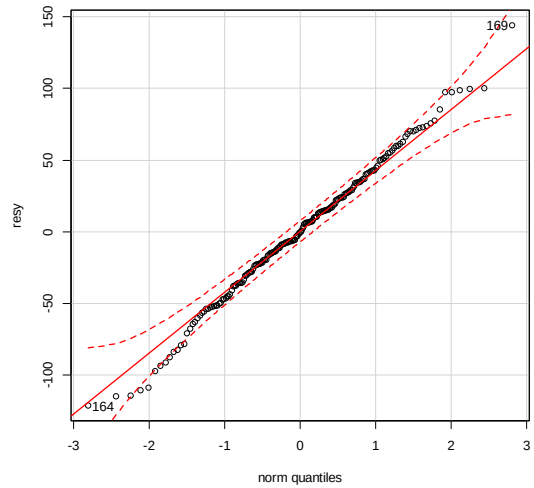
- Histogram
- Q-Q plot
- Shapiro-Wilk statistic

```
Rcmdr>  with(album2,
shapiro.test(resy))

        Shapiro-Wilk normality test

data:  resy
W = 0.99483, p-value = 0.7253
```

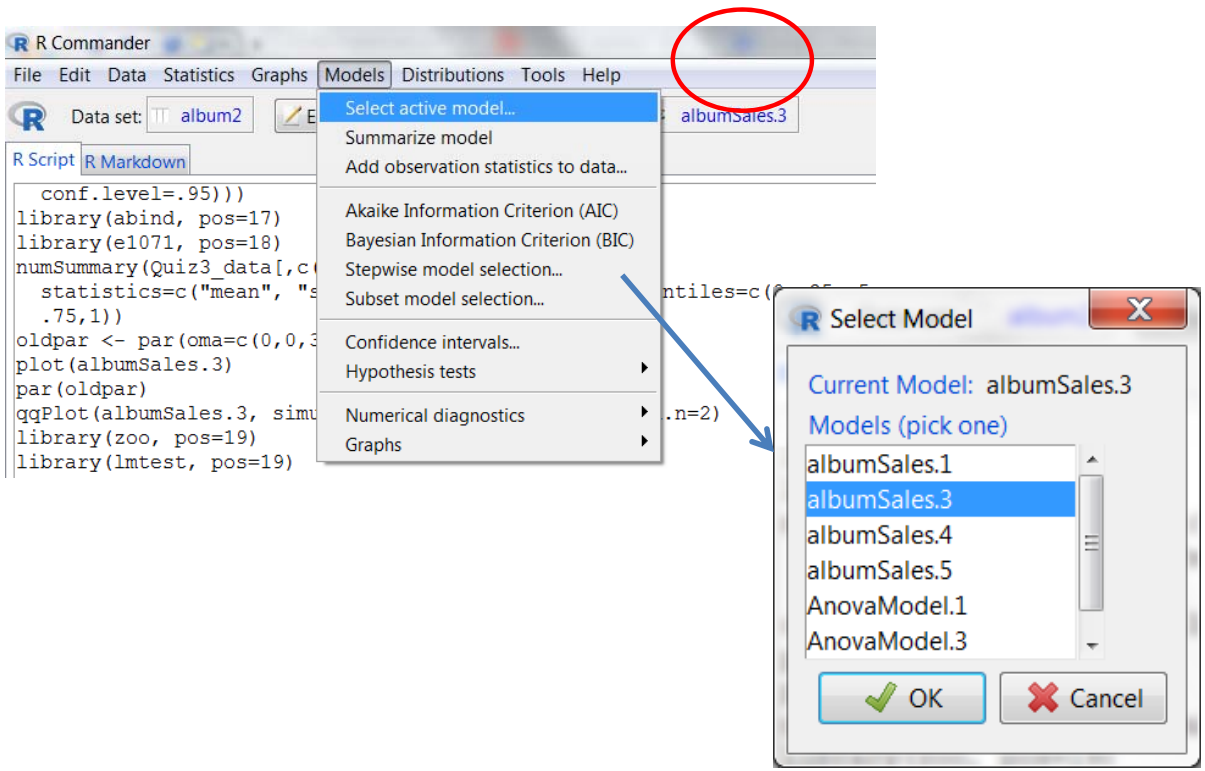




Residual Diagnostics

Further diagnostics

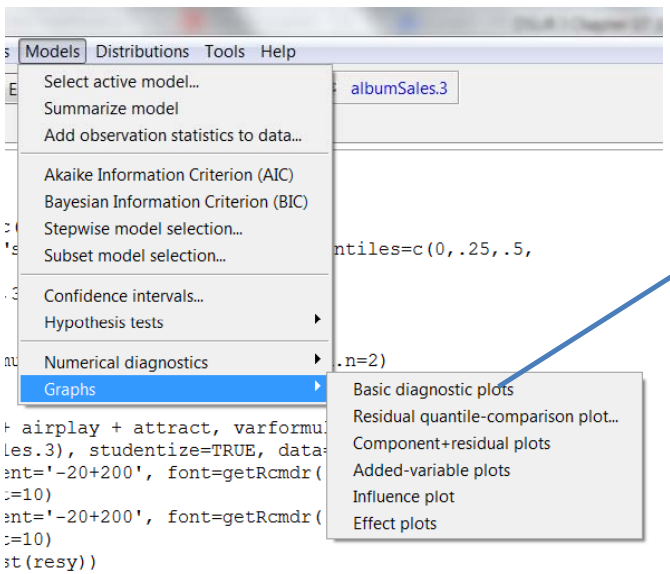
- Load the “model object” to Rcmdr

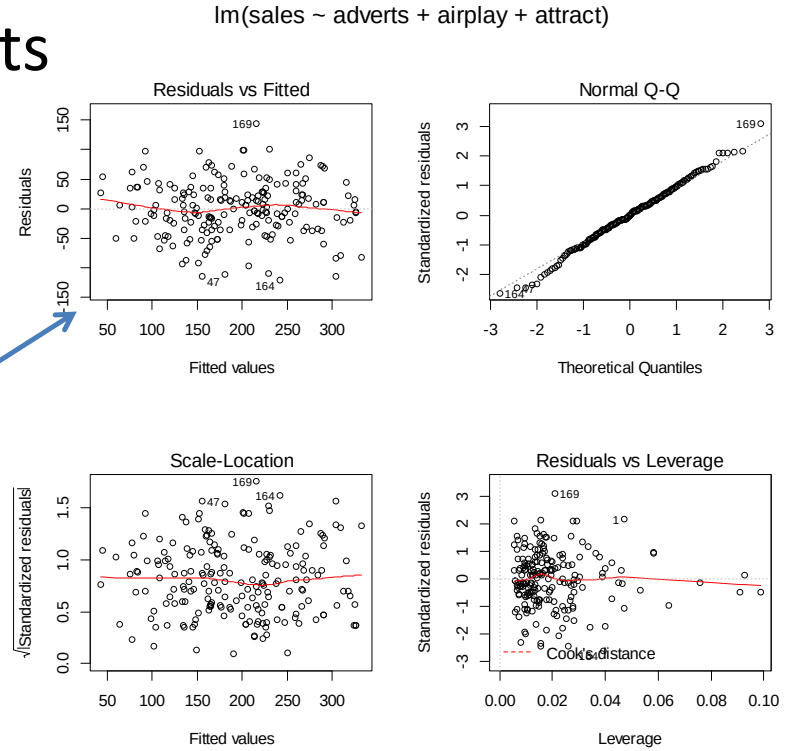


Residual Diagnostics

Checking constancy of variance

- Basic diagnostic plots



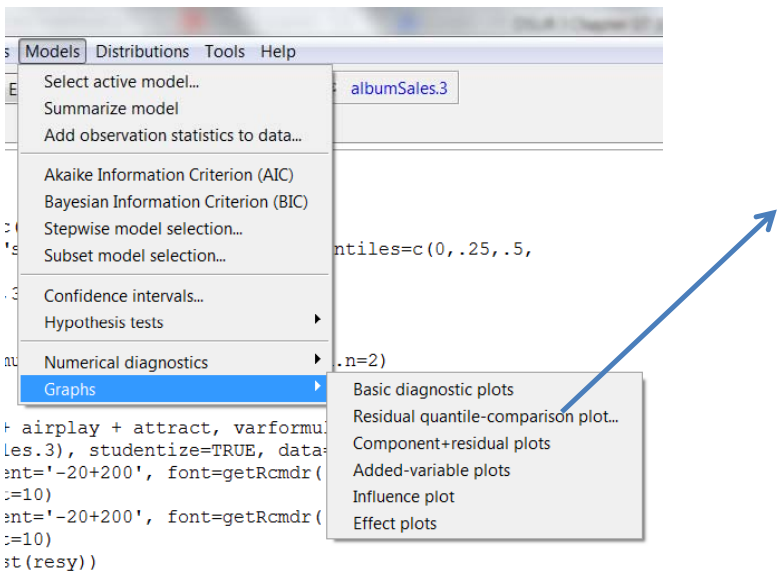


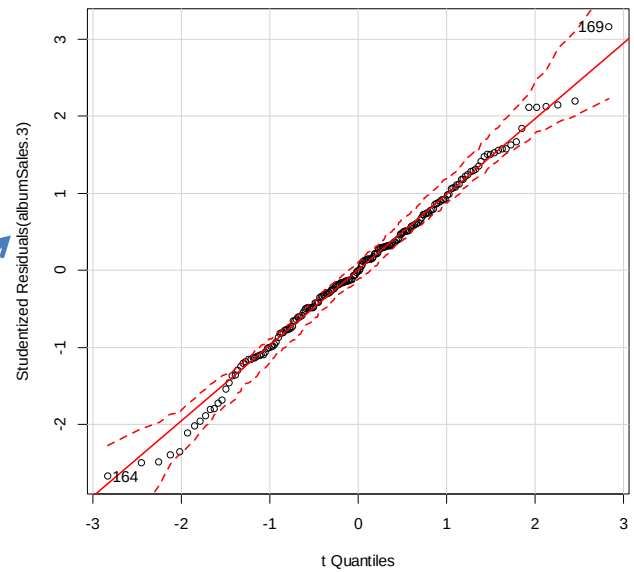
- Cone-shaped scatter plots would indicate violations of homoscadasticity

Residual Diagnostics

Checking constancy of variance

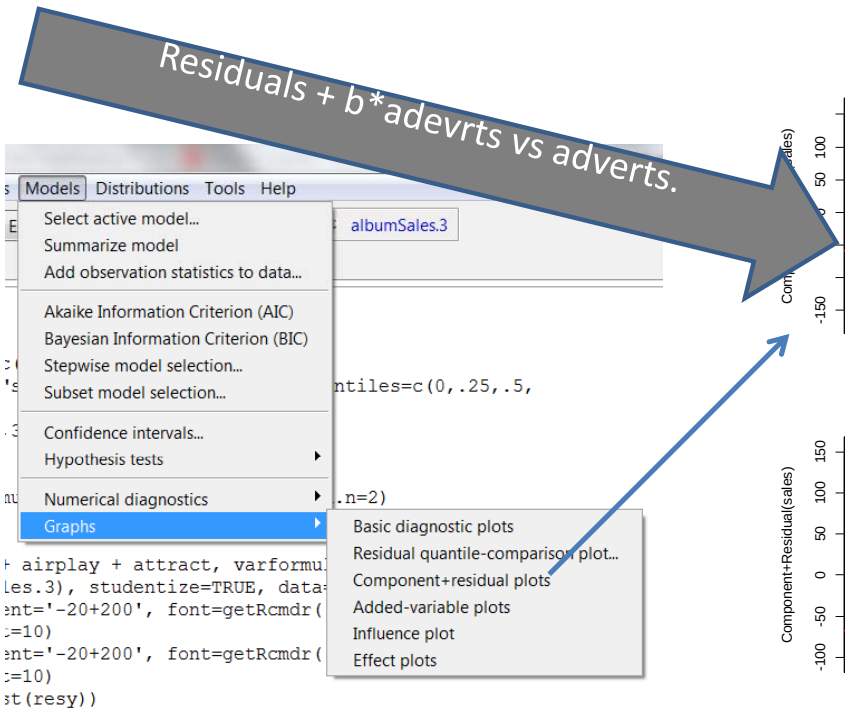
- QQ plot

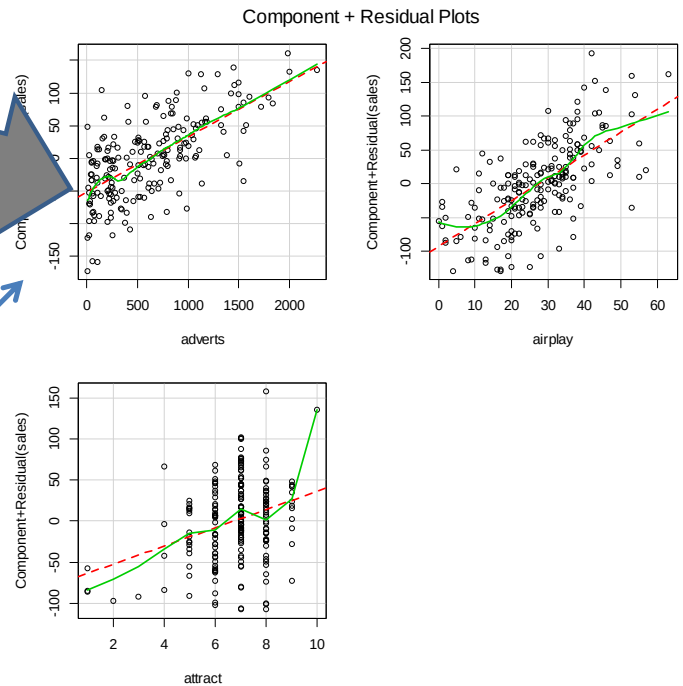


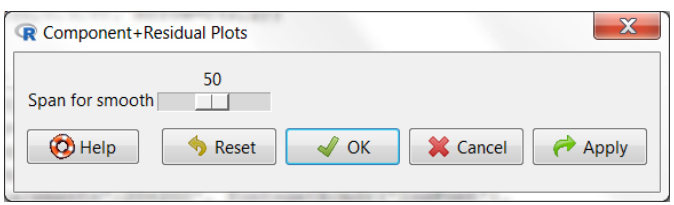


Residual Diagnostics

Checking for nonlinearity



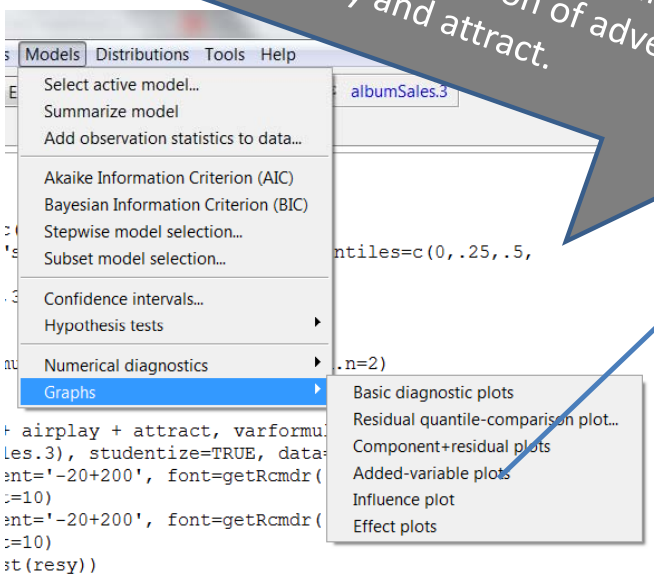


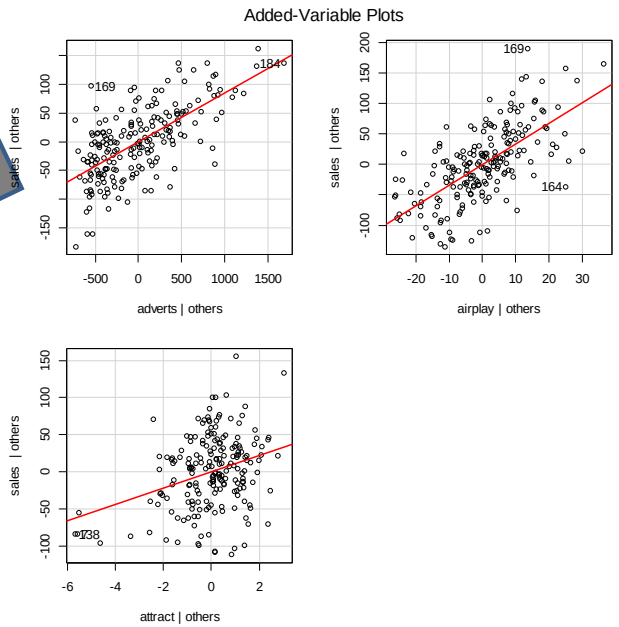


Residual Diagnostics

Checking for leverage

Residuals from the regression of sales on airplay and attract are plotted against residuals from the regression of adverts on airplay and attract.



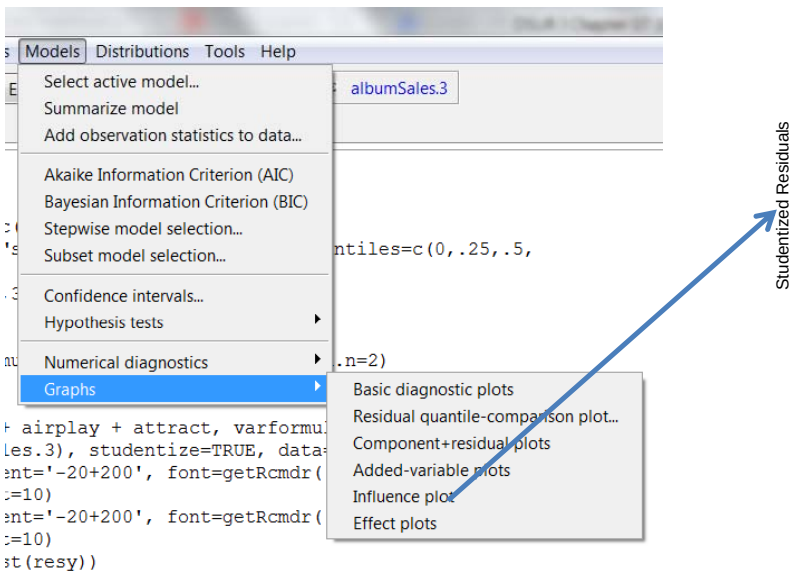


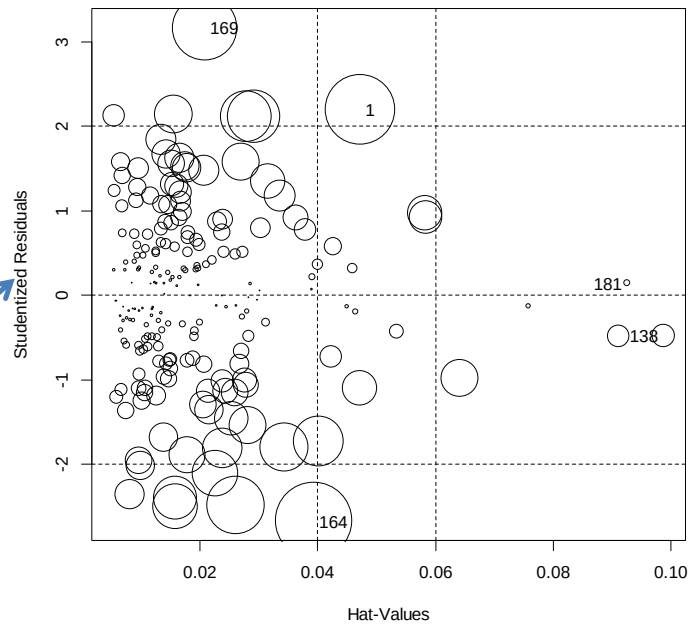
- High leverage observations show in added variable plots as points horizontally distant from the rest of the data.

Residual Diagnostics

Influence plot

- Graphical methods



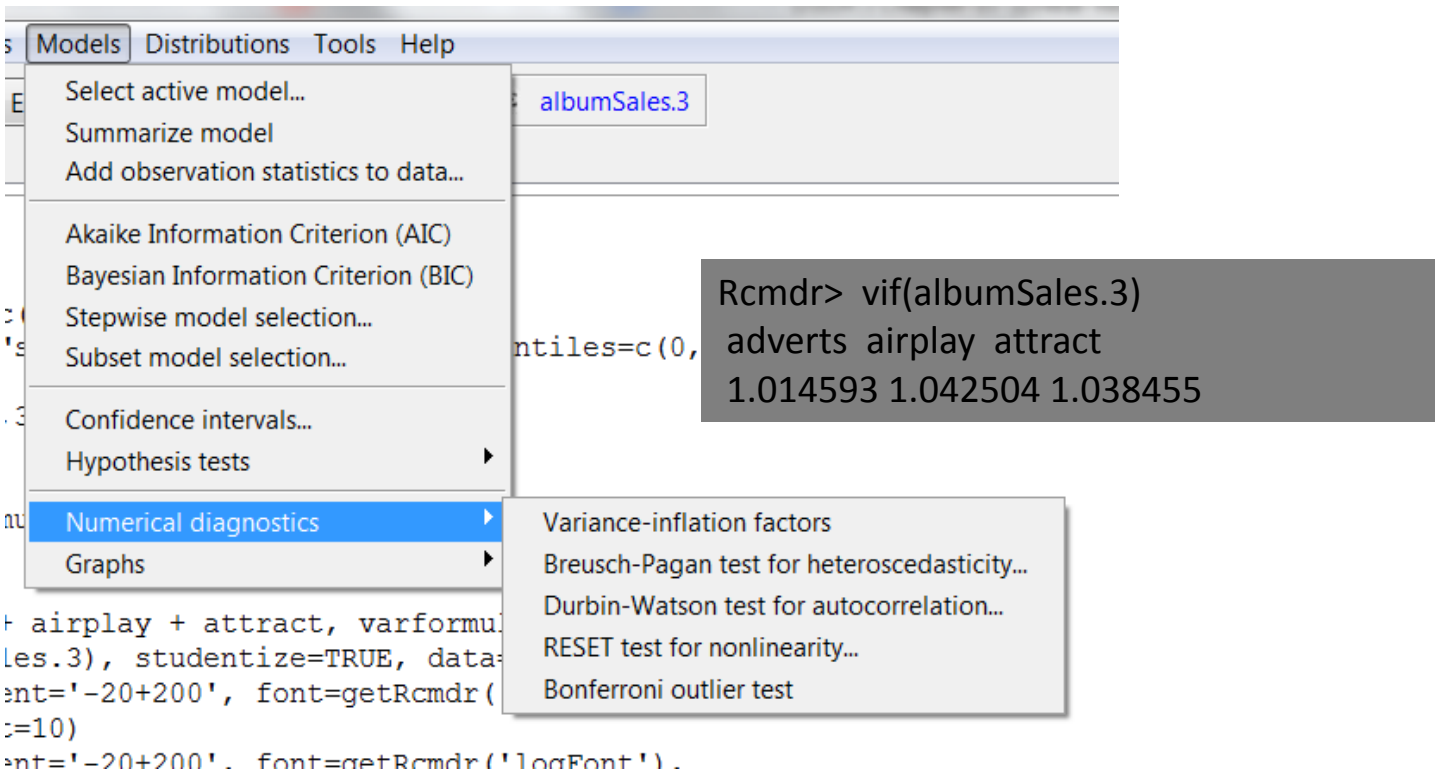


- Circle size proportional to Cook's D
- Hat value average here 4/200
- Looking for the combination of large Hat-values to go with large Cook's D

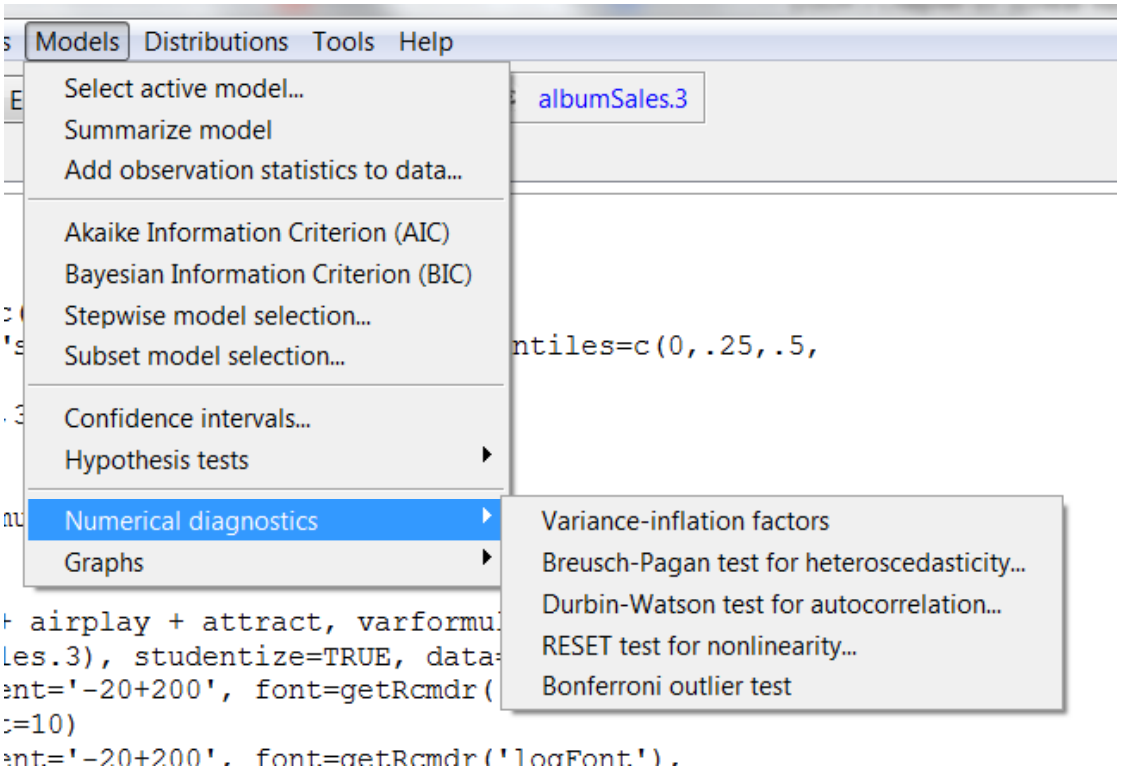
Residual Diagnostics

Variance-inflation factors -- multicollinearity

- Numerical diagnostics



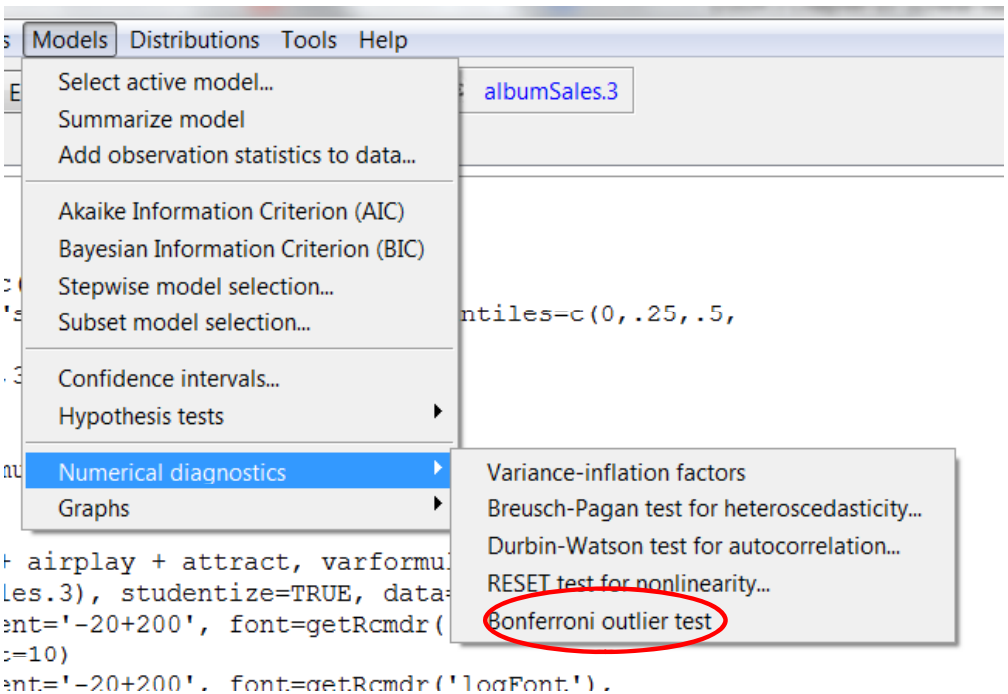
Residual Diagnostics




```
Rcmdr> bptest(sales ~ adverts + airplay + attract, varformula = ~ Rcmdr+
fitted.values(albumSales.3), studentize=TRUE, data=album2)
        studentized Breusch-Pagan test
data: sales ~ adverts + airplay + attract
BP = 0.28272, df = 1, p-value = 0.5949
```

Residual Diagnostics

Outlier test for the largest residual




```
Rcmdr> outlierTest(albumSales.3)
 No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|: rstudent    unadjusted p-value   Bonferonni p
169                 3.163622    0.0018077            0.36154
```

Summary of Regression Module

- A gentle introduction to using R to obatin basic regression computations
- A few useful graphical tools used in the context of regression
- Some commonly used inference methods illustrated
- A brief overview of residual diagnostics

Introduction

1. One-way ANOVA
2. Multiple Comparison Procedures
3. Family wise Error Rates
4. False Discovery Rate

When and Why do we do ANOVA?

- We can use a t-test to compare means. But:
  - You can compare only 2 means, with just one grouping variable
  - Often we would like to compare means from 3 or more groups

- ANOVA
  - Compares several means
  - Can be used when there more than one grouping variable
  - It can be thought of as a multiple linear regression

Data Structure

- The data has to be in the long format. That is, data for different groups have to stacked.
- The data has to contain a group indicator
- The group indicator has to be read as a `factor` in R

Data Structure

vibration (microns)	Brand
13.1	1
15.0	1
14.0	1
14.4	1
14.0	1
11.6	1
16.3	2
15.7	2
17.2	2
14.9	2
14.4	2
17.2	2
13.7	3
13.9	3
12.4	3
13.8	3
14.9	3
13.3	3
15.7	4
13.7	4
14.4	4
16.0	4
13.9	4
14.7	4
13.5	5
13.4	5
13.2	5
12.7	5
13.4	5
12.3	5

Long (not wide)
Column indicating the group
(this column has to be a factor)

ANOVA Notations

- Notation

y_{ij} : the j^{th} sample observation selected from population i

n_i : the number of sample observations selected from population i

n_T : the total sample size; $n_T = \sum n_i$

$\bar{y}_{i.}$: the average of the n_i sample observations from population i

$\bar{y}_{..}$: the average of all sample observations; $\bar{y}_{..} = \sum_i \sum_j \frac{y_{ij}}{n_T}$

t : number of populations in the study

ANOVA model

Let μ_i denote the mean of the i^{th} population. i=1,2,...,T.

Let Y_{ij} denote the j^{th} measurement from the i^{th} population, j=1,2,...,n_i.

Then the ANOVA model is: $Y_{ij} = \mu_i + \varepsilon_{ij}$, where $\varepsilon_{ij} \sim \texttt{IID}\ \texttt{N}(0, \sigma^2)$

Objectives of ANOVA:

(1) Test $H_0 : \mu_1 = \mu_2 = \cdots = \mu_T$ vs H_0 : not H_0

(2) If H_0 is rejected, then order the population means (μ_i's)

The ANOVA F-test is used for (1). In order to accomplish (2), we need to test the hypotheses H_{0ij}:$\mu_i = \mu_j$ simultaneously.

Or obtain simultaneous confidence intervals for $\{\mu_i - \mu_j : i < j\}$

What Does ANOVA Tell Us?

- Null hyothesis:
  - Like a *t*-test, ANOVA tests the null hypothesis that the means are the same.
- Experimental hypothesis:
  - The means differ.
- ANOVA is an omnibus test
  - It test for an overall difference between groups.
  - It tells us that the group means are different.
  - It doesn't tell us exactly which means differ.

Data Structure

ANOVA Table

Source	Sum of Squares	Degrees of Freedom	Mean Square	F Test
Between Samples	SSB	$t-1$	$s_B^2 = \frac{SSB}{t-1}$	$\frac{s_B^2}{s_W^2}$
Within Samples	SSW	$n_T - t$	$s_W^2 = \frac{SSW}{n_T - t}$	
Totals	SST	$n_T - 1$		

ANOVA

- The null hypothesis of the equality of t population means is rejected if

$$F = \frac{s_B^2}{s_W^2} > F_{\alpha,t-1,n_T-t}$$

- Summarize the results in an ANOVA table. . .

ANOVA

- Total Sum of Squares (SST)

$$SST = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2$$

- Within-sample Sum of Squares (SSW)

$$\begin{aligned} SSW &= \sum_{i=1}^{t} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2 \\ &= (n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2 + \ldots + (n_t - 1)\, s_t^2 \end{aligned}$$

- Between-sample Sum of Squares (SSB)

$$SSB = \sum_{i=1}^{t} n_i \, (\bar{y}_{i.} - \bar{y}_{..})^2$$

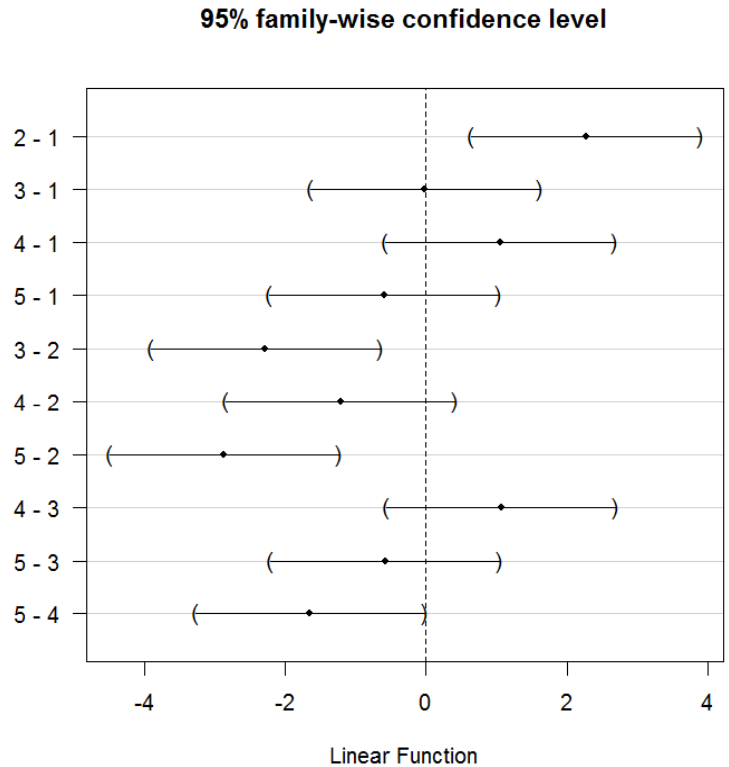
Motor Data Set

- Five brands of motors were compared for noise
- Six units were tested in each group
- Data is in motor.xls, can be loaded easily for any of the packages
- Dataframe is also named motor

Why Use Follow-Up Tests?

- The *F*-ratio tells us only that the experiment was successful
  - i.e. group means were different
- It does not tell us specifically which group means differ from which.
- We need additional tests to find out where the group differences lie.

Pairwise Comparisons of Means



Pairwise Comparisons of Means: Fisher's Least Significant Difference

Fit: aov(formula = vibration..microns. ~ Brand.fac, data = motor)

Linear Hypotheses:

	Estimate	Std. Error	t value	Pr(>|t|)	
2 - 1 == 0	2.26667	0.55183	4.108	0.00313	**
3 - 1 == 0	-0.01667	0.55183	-0.030	1.00000	
4 - 1 == 0	1.05000	0.55183	1.903	0.34183	
5 - 1 == 0	-0.60000	0.55183	-1.087	0.81132	
3 - 2 == 0	-2.28333	0.55183	-4.138	0.00291	**
4 - 2 == 0	-1.21667	0.55183	-2.205	0.21070	
5 - 2 == 0	-2.86667	0.55183	-5.195	< 0.001	***
4 - 3 == 0	1.06667	0.55183	1.933	0.32684	
5 - 3 == 0	-0.58333	0.55183	-1.057	0.82620	
5 - 4 == 0	-1.65000	0.55183	-2.990	0.04449	*

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

(Adjusted p values reported -- single-step method)

What do we mean by "simultaneous coverage"?

Let A_i = event that $\theta_i \in [L_i, U_i]$. Coverage probability (Confidence Level) for each interval is $P(A_i)$. Simultaneous coverage probability=$P\left(\bigcap_{i=1}^{m} A_i\right)$.

$$\min_{i=1,\ldots,m} \{P(A_i)\} \geq P\left(\bigcap_{i=1}^{m} A_i\right) = 1 - P\left[\left(\bigcap_{i=1}^{m} A_i\right)^c\right] = 1 - P\left[\bigcup_{i=1}^{m} A^c{}_i\right] \geq 1 - \sum_{i=1}^{m} P\left[A^c{}_i\right]$$

If $P(A_i) = 1 - \alpha$, for each i,

$$1 - \alpha \geq P\left(\bigcap_{i=1}^{m} A_i\right) \geq 1 - \sum_{i=1}^{m} P\left[A^c{}_i\right] = 1 - m\alpha$$

Simultaneous cofidence is [usually] LESS THAN individual confidence !!

What do we mean by "family-wise error rate"?

Consider simultaneous testing of several hypotheses:

$H_{0i}:\theta_i = 0$ vs $H_{1i}:\theta_i \neq 0$. Let $R_i = \{\text{Reject } H_{0i}\}$, the rejection region.

Experimentwise Error Rate $= P\left(\bigcup_{i=1}^{m} R_i\right)$, when all H_{0i} are true.

In general, $\max_{i=1,\ldots,m} \left\{P(R_i)\right\} \leq P\left(\bigcup_{i=1}^{m} R_i\right) \leq \sum_{i=1}^{m} P[R_i]$.

Suppose, $H_{0i}:\theta_i = 0$ for all i and $P(R_i) = \alpha$. Then,

$\alpha \leq P\left(\bigcup_{i=1}^{m} R_i\right) \leq m\alpha$, but in practice it can be much larger than α.

One popular solution: Bonferroni

- Perform each test at level α/m so that

$$P\left(\bigcup_{i=1}^{m} R_i\right) \leq \sum_{i=1}^{m} P[R_i] = m\frac{\alpha}{m} = \alpha.$$

  - Works, but tests become too conservative

- Construct each confidence interval with confidence level $1-\frac{\alpha}{m}$
  - Works, but intervals are too wide

Family-wise Error Rates

- Bonferroni method
- Fisher's LSD
- Tukey's W Procedure
- Student-Newman-Keuls Procedure
- Dunnett's Procedure
- Scheffe's S method

- Since testing procedure is a "dual" of computing confidence intervals, each method above applies to control simultaneous coverage also.

Tukey's W Procedure

1. Rank the *t* sample means
2. Two population means are different if

$$|\bar{y}_{i.} - \bar{y}_{j.}| \geq W, \quad \text{where } W = q_\alpha(t, \nu)\sqrt{\frac{s_W^2}{n}}$$

note that sample sizes must be the same!

Tukey's procedure in R

```
aov(formula = vibration..microns. ~ Brandfac, data = motor)
> TukeyHSD(AnovaModel.2)
Tukey multiple comparisons of means
95% family-wise confidence level
Fit: aov(formula = vibration..microns. ~ Brandfac, data = motor)
```

$Brandfac

	diff	lwr	upr	p adj
2-1	2.26666667	0.6460270	3.8873064	0.0031588
3-1	-0.01666667	-1.6373064	1.6039730	0.9999998
4-1	1.05000000	-0.5706397	2.6706397	0.3418272
5-1	-0.60000000	-2.2206397	1.0206397	0.8112981
3-2	-2.28333333	-3.9039730	-0.6626936	0.0029299
4-2	-1.21666667	-2.8373064	0.4039730	0.2106883
5-2	-2.86666667	-4.4873064	-1.2460270	0.0002024
4-3	1.06666667	-0.5539730	2.6873064	0.3268245
5-3	-0.58333333	-2.2039730	1.0373064	0.8262091
5-4	-1.65000000	-3.2706397	-0.0293603	0.0445279

Simint in R

R Documentation

simint {multcomp}

Description

Computes simultaneous intervals for several multiple procedures.

Usage

```
## Default S3 method:
simint(y, x=NULL, type=c("Dunnett", "Tukey",
    "Sequen", "AVE", "Changepoint", "Williams", "Marcus",
    "McDermott","Tetrade"), cmatrix=NULL, conf.level=0.95,
    alternative=c("two.sided","less", "greater"),
    asympt=FALSE, eps=0.001, maxpts=1e+06, nlevel=NULL,
    nzerocol=c(0,0),...)
```

Family-wise Error Rate (FWER)

- Suppose we have performed m hypotheses tests:

  H_{01}vs H_{a1} P-value=p_1

  H_{02}vs H_{a2} P-value=p_2

  $\vdots$

  H_{0m}vs H_{am} P-value=p_m

- We reject H_{0j} if $p_j \leq \alpha_I$. Thus the j^{th} test has Type I Error α_I. That is, the probability of falsely rejecting H_{0j} is α_I
- α_I is the individual comparisons type I error rate.
- The family-wise error rate (FWER) is the probability of at least one false rejection. Let us denote this by α_F
- If A_j = the event that H_{0j} is falsely rejected, then

$$\alpha_I = P(A_j) \quad \text{and} \quad \alpha_F = P\left(\bigcup_{j=1}^{m} A_j\right)$$

- Generally, $\alpha_F >>> \alpha_I$ for large m

Family-wise Error Rate (FWER)

- Suppose we have performed m hypotheses tests:
- And, suppose the statistics used for testing these hypotheses are mutually independent.
- Then the events A_j's are mutually independent.
- Therefore, $\alpha_F = 1 - (1 - \alpha_I)^m$

	α_I = 0.100	0.050	0.010
1	0.100	0.050	0.010
m 5	0.410	0.226	0.049
10	0.651	0.401	0.096

Bonferroni method of controlling FWER

- Note

$$\alpha_F = P\left(\bigcup_{j=1}^{m} A_j\right) \leq \sum_{j=1}^{m} P(A_j) = m\alpha_I$$

- Thus, by taking $\alpha_I = \frac{\alpha}{m}$ (α being the desired level)

$$\alpha_F \leq m\alpha_I \leq m\frac{\alpha}{m} \leq \alpha.$$

- This is used when m is relatively small, as mentioned in the pairwise testing (and simultaneous interval estimation) problems as earlier.
- This is known to be too conservative (for large m), that is calls too few tests as signficant.

Bioinformatics example

- 12625 genes from a microarray study of radiation sensitivity
- 44 samples in the normal group and 14 in the radiation sensitivity group
- Goal: identify informative genes.
- Data structure:

Gene_#	Normal Group	Radiation Group	T-stat	P-value
Gene_1	X1,…,x44	Y1,..Y14	T_1	p_1
Gene_2	X1,…,x44	Y1,..Y14	T_2	p_2
….	….	….	….	….
Gene_m	X1,…,x44	Y1,..Y14	T_m	p_m

Bioinformatics example

- Suppose we perform a two sample t-test for each gene.
- Desired level is 0.05 (=α)
- One t-test per gene. That is, m=12,625.
- That is each test has to be conducted at level

  $\alpha_I \leq 0.05/(12625)=3.9 \times 10^{-6}$
- This level is too small and likely to not reject practically all genes!

False Discovery Rate (FDR) Approach

- Different apporach – DOES NOT control FWER, instead controls FDR.
- Instead looking at the proportion of falsely rejected hypotheses
- Possible outcomes for the m testing problems

	Accepted Null	Rejected Null	
Null is true	U	V	m_0
Null is false	T	S	m_1
	m-R	R	M

Note that

$$\text{FWER}=P(V \geq 1) \text{ and } \text{FDR}=E\left(\frac{V}{R}\right)$$

Benjamini-Hochberg Procedure

- Fix the desired FDR rate α.
- Order the p-values obtained from the m individual tests: $p_{(1)} \leq p_{(2)} \leq \ldots\ldots \leq p_{(m)}$
- Define $L = \text{Max}\left\{ j : p_{(j)} < \alpha . \frac{j}{m} \right\}$

- Reject all hypotheses H_{0j} for $j \leq L$.
- L is called the Benjamini-Hochberg rejection threshold. Graph the $p_{(1)}$ vs $\alpha(j/m)$ to see what is happening.

Bioinformatics example (Contd.)

- We have gene expression data on four groups of patients
- Each group represents a dosage level of the treatment
- Measurements of 12,625 genes are available for each group
- The scientist has performed 12,625 t-tests of comparing each group to a control group.
- Data are given in GeneTesting.xls