DISCOVERING STATISTICS USING R

Regression

Part 1: Simple linear regression

1

DISCOVERING STATISTICS USING R

Aims and Objectives

- Understand linear regression with one predictor
- Understand how we assess the fit of a regression model
  - Total Sum of Squares
  - Model Sum of Squares
  - Residual Sum of Squares
  - R^2
  - F
- Know how to perform linear regression and interpret the model

2

DISCOVERING STATISTICS USING R

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

3

DISCOVERING STATISTICS USING R

What is regression?

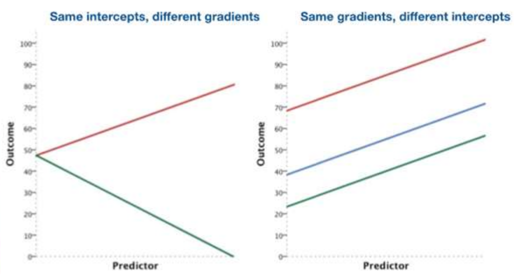
$$\boxed{Y_i = b_0 + b_i X_i + \varepsilon_i}$$

- Describing a straight line (the Linear Model)
  - b_0
    - Intercept (value of Y when $X = 0$)
    - Point at which the regression line crosses the Y-axis
  - b_i
    - Regression coefficient for the predictor
    - Gradient (slope) of the regression line
    - Direction/strength of relationship
  - ε_i
    - The model error

4

DISCOVERING STATISTICS USING R

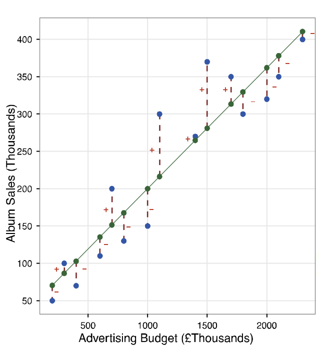
What is regression?



5

DISCOVERING STATISTICS USING R

The method of least squares



A scatterplot of some data with a line representing the general trend. The vertical lines (dotted) represent the differences (or residuals) between the line and the actual data

6

DISCOVERING STATISTICS USING R

How good is the model?

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

7

DISCOVERING STATISTICS USING R

Sums of Squares

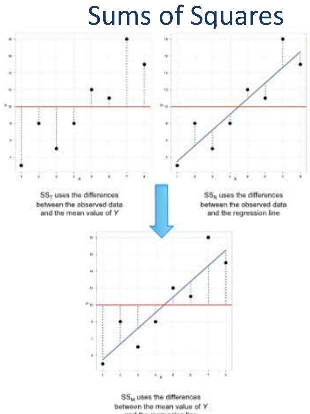


Diagram showing from where the regression sums of squares derive

8

DISCOVERING STATISTICS USING R

Sums of Squares

- SS_T
  - Total variability
    - Variability between scores and the mean
- SS_R
  - Residual/error variability
    - Variability between the regression model and the actual data
- SS_M
  - Model variability
    - Difference in variability between the model and the mean

9

DISCOVERING STATISTICS USING R

Testing the model

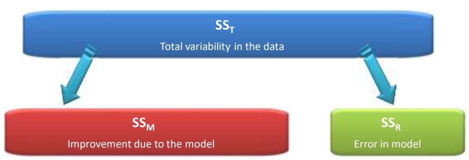
- How well does our model 'fit' the data?
  - R^2
    - The proportion of the total variability in the data accounted for by the regression model
    - Equivalent to the Pearson correlation coefficient squared

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

10

DISCOVERING STATISTICS USING R

Testing the model



- If the model results in better prediction than using the mean, we expect: $SS_M > SS_R$

11

DISCOVERING STATISTICS USING R

Testing the model

- However, recall from Chapter 2
  - Sums of squares are total values
    - Good measures of variability, but dependent on sample size
  - To overcome this, we can calculate Mean Squares (MS)
    - Compare to how we calculated variance!!
- To assess whether our regression model fits the data better than the mean, we can thus perform an analysis of variance (ANOVA):

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

12

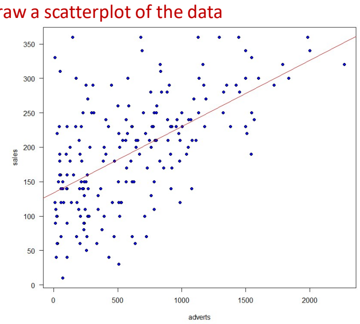
Example

- 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 units of £1000) spent promoting the record before release

13

Example

- Draw a scatterplot of the data



14

Example

- We run a regression analysis using the '*lm()*' function (lm stands for 'linear model')

```
> newModel<- lm(outcome~ predictor(s), data=
               dataFrame, na.action= an action))
```

- For our record sales example we type

```
> albumSales.1<- lm(sales~ adverts, data= album1)
```

15

DISCOVERING STATISTICS USING R

Example

- Use '*anova()*' to look at the *SS* and *MS* of the model

```
> anova(albumSales.1)
```

- Use '*summary()*' to look at the parameter estimates of the model

```
> summary(albumSales.1)
```

16

DISCOVERING STATISTICS USING R

Example

$$\text{Record Sales}_i = b_0 + b_1 \text{Advertising Budget}_i = 134.14 + (0.09612 \times \text{Advertising Budget}_i)$$

$$\text{Record Sales}_i = 134.14 + (0.09612 \times \text{Advertising Budget}_i) = 134.14 + (0.09612 \times 100) = 143.75$$

17

DISCOVERING STATISTICS USING R

Regression

Part 2: Multiple linear regression

18

DISCOVERING STATISTICS USING R

Aims and Objectives

- Understand the multiple regression equation
- Understand different methods of regression
  - Hierarchical
  - Forced entry
  - Stepwise
- Know how to perform a multiple regression and interpret the model
- Understand the assumptions of multiple regression and know how to test them

19

DISCOVERING STATISTICS USING R

What is multiple regression?

- Simple regression is a model to predict the value of one variable from another
- Multiple regression is a natural extension of this:
  - Used to predict values of an outcome from *several* predictors
  - It is a hypothetical model of the relationship between several variables

20

DISCOVERING STATISTICS USING R

Example

- A record company boss was interested in predicting album sales from advertising **and air time**
- Data
  - 200 different album releases
- Outcome variable
  - Sales (CDs and downloads) in the week after release
- Predictor variables
  - The amount (in units of £1000) spent promoting the record
  - Air time on the radio

21

DISCOVERING STATISTICS USING R

Multiple regression as an equation

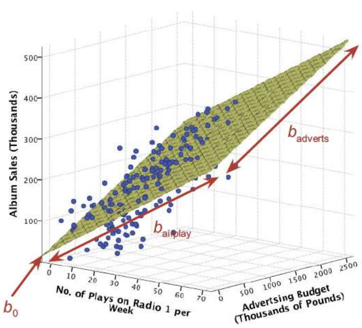
$$y_i = b_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n + \varepsilon_i$$

- In simple regression our model is described by a straight line
- In multiple regression we extend this equation
  - b_0
    - Intercept (value of Y when all Xs= 0)
    - Point at which the regression line crosses the Y-axis (vertical)
  - $b_{1\text{-}n}$
    - Regression coefficients for predictors 1 to n
    - Gradient (slope) of the regression line
    - Direction/strength of relationship

22

DISCOVERING STATISTICS USING R

A model with two predictors



Scatterplot of the relationship between album sales, advertising budget and radio play

23

DISCOVERING STATISTICS USING R

Methods of regression

- How to incorporate multiple predictor variables into a model?
  - Hierarchical
    - Experimenter decides the order in which variables are entered into the model
  - Forced entry
    - All predictors are entered simultaneously
  - Stepwise
    - Predictors are selected [illegible]eir semi-partial correlation with the outcome
    - Forward, backward, [illegible] all-subsets

24

DISCOVERING STATISTICS USING R

Model fit

- Sums of Squares
  - Although the computation of these values is more complex than in simple regression, their use and interpretation is the same
- R^2
  - This is a multiple R^2, *i.e.* the squared correlation coefficient between observed and fitted values based on all predictors
  - As in simple regression, it is a measure of the proportion of the total variance in the outcome variable accounted for by the model
- Akaike Information Criterion (*AIC*)
  - Assess model fit, while penalizing for model complexity
  - Only makes sense when comparing different models of the same data, in which case: lower values represent a better fit

25

DISCOVERING STATISTICS USING R

Model diagnostics

- To assess a model's accuracy within the sample
  - Look for outliers
    - Calculate the standardized residual for each case
    - 95% of data should lie between -1.96 and +1.96
    - 99% of data should lie between -2.58 and + 2.58
    - An absolute value greater than 3 is indicative of an outlier
  - Look for influential cases
    - There are several residual statistics to assess the influence of a particular case on model parameter estimates (Cook's distance, DFBeta, DFFit, leverage and the covariance ratio)
    - "Quick and dirty" rule of thumb: any case with an absolute value of its Cook's distance > 1 may be cause for concern

26

DISCOVERING STATISTICS USING R

Model generalization

!!Dummy coding!!

- To assess how well a model generalizes from the sample to the population, several assumptions must hold
  - Pretty straightforward ones:
    - Variable type
      - Outcome must be a continuous variable
      - Predictors can be continuous or dichotomous variables
    - Non-zero variance
      - Predictors must not have zero variance
    - Linearity
      - The relationship we model is, in reality, linear
    - Independence
      - All values of the outcome should come from a different entity

27

DISCOVERING STATISTICS USING R

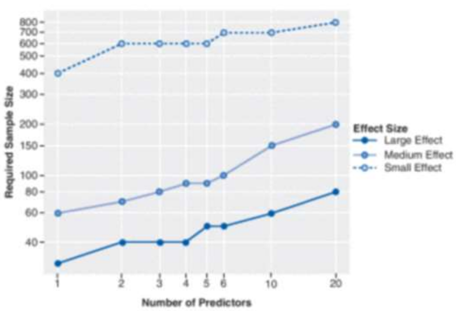
Model generalization

- To assess how well a model generalizes from the sample to the population, several assumptions must hold
  - More tricky ones:
    - No multicollinearity
      - Predictors must not be highly correlated
    - Homoscedasticity
      - For each value of the predictors the variance of the error term should be constant
    - Independent errors
      - For any pair of observations, the error terms should be uncorrelated
    - Normally distributed errors
      - Visual inspection of histograms and QQ plots

28

DISCOVERING STATISTICS USING R

Sample size



29

DISCOVERING STATISTICS USING R

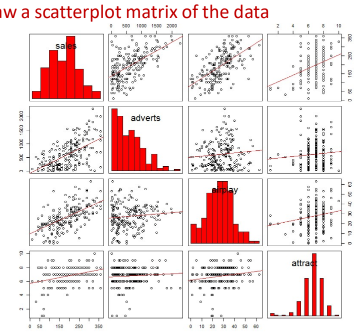
Example

- A record company boss was interested in predicting album sales from advertising, air time and attractiveness of the band
- Data
  - 200 different album releases
- Outcome variable
  - Sales (CDs and downloads) in the week after release
- Predictor variables
  - The amount (in units of £1000) spent promoting the record
  - Air time on the radio
  - Attractiveness

30

Example

- Draw a scatterplot matrix of the data



31

Example

- Let's say we want to use the 'forced entry' method (unlike the example in the book)

```
> albumSales.F<- lm(sales~ adverts + airplay +
                    attract, data= album2)
```

- Inspect *SS* and *MS*, using the '*anova()*' function

```
> anova(albumSales.F)
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 9.49e-06***
Residuals 196 434575    2217
```

32

Example

- Next look at the model parameters, using '*summary()*'

```
> summary(albumSales.F)

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
```

33

DISCOVERING STATISTICS USING R

Example

- We can now write the equation of our regression model

$$sales_i = b_0 + b_1\, adverts + b_2\, airplay + b_3\, attract + \varepsilon_i$$

$$sales_i = -26.61 + 0.08{*}adverts + 3.37{*}airplay + 11.09{*}attract + \varepsilon_i$$

- To be able to directly compare the influence of different predictors, we can calculate standardized parameters (β_i instead of b_i) using '*lm.beta()*'

```
> lm.beta(albumSales.F)
  adverts   airplay   attract
0.5108462 0.5119881 0.1916834
```

34

DISCOVERING STATISTICS USING R

Outliers and influential cases

- Add residual statistics to the original dataframe, *e.g.*

```
> album2$res<- resid(albumSales.F)
> album2$stz.r<- rstandard(albumSales.F)
> album2$stu.r<- rstudent(albumSales.F)
> album2$cd<- cooks.distance(albumSales.F)
> album2$dfbeta<- dfbeta(albumSales.F)
> album2$dffit<- dffits(albumSales.F)
> album2$leverage<- hatvalues(albumSales.F)
> album2$stz.cvr<- covratio(albumSales.F)
```

- Evaluate these against the criteria detailed in book

35

DISCOVERING STATISTICS USING R

Model assumptions

(the tricky ones)

- Independent errors
  - Durban-Watson test must not be significant

```
> dwt(albumSales.F)
```

- No multicollinearity
  - Value Inflation Factors (VIF)< 10
  - Mean VIF not substantially greater than 1

```
> vif(albumSales.F)
```

- Normally distributed errors and homoscedasticity
  - Inspect plots

```
> par(mfrow= c(2,2)); plot(albumSales.F)
```

36

Robust regression

- Use bootstrapping to estimate model parameters
  - Recall from Chapter 6

```
> object.B<- boot(data, function, replications)
```

  - Write the function to include in the bootstrap

```
> bootReg<- function(formula, data, i){
                         d<- data[i,]
                         fit<- lm(formula, data= d)
                         return(coef(fit))
                         }
```

37

Robust regression

- Use bootstrapping to estimate model parameters
  - Calculate the bootstrap estimates

```
> bootResults<- boot(statistic= bootReg, formula=
                         sales~ adverts + airplay +
                         attract, data= album2,
                         R= 10000)
```

  - Look at the confidence intervals around the estimates

```
> boot.ci(bootResults, type= "bca", index= 1)
> boot.ci(bootResults, type= "bca", index= 2)
> boot.ci(bootResults, type= "bca", index= 3)
> boot.ci(bootResults, type= "bca", index= 4)
```

38

Reporting results

- Typically in the form of a table slightly different from the one in the book

Table 1. Parameter estimates obtained for a linear model that expresses record sales as a function of advertising budget, air time and band attractiveness

	B	s.e.	*t*	*P*
Intercept	-26.61	17.35		
Advertising budget	0.08	0.01	12.26	< 0.001
Air time	3.37	0.28	12.12	< 0.001
Attractiveness	11.09	2.44	4.55	< 0.001
				$R^2_{Adj.}$ = 0.660, $F_{(3, 196)}$ = 129.5, p< 0.001

39

DISCOVERING STATISTICS USING R

Where do we go from here…

- Most of the remaining models we'll cover in this course are variations and generalizations of the regression (or linear model) framework
- We will explore ways in which to deal with:
  - categorical predictor variables that are not dichotomies (dummy coding -this chapter- and Chapters 10-12)
  - dependencies in the data, *e.g.* PGLS-models, repeated measures designs (Karin, Chapter 13)
- Biological data typically suffer from both non-normality and dependencies, and at the end of the course you should be sufficiently equipped to deal with these complications

40

DISCOVERING STATISTICS USING R

Rest of today…

- Practical Chapter 7
  - Read § 7.1, 7.2, "Cramming Sam's Tips" and "What Have I Discovered about Statistics?"
  - Skip sections on R commander:
    §7.4.1, § 7.8.2.1, §7.8.4.1, §7.9.1
  - Work through self-tests as you see fit (but skip self-tests in §7.12)
  - Solve Smart Alex's Tasks 1-2

41

DISCOVERING STATISTICS USING R

Errata

- p. 299

As we did for correlations, we need to write a function (R's Souls' Tip 6.2) we want to bootstrap. We'll write one called *bootReg()* – this function is a little more complex than the function we wrote for the correlation, because we are interested in more than one statistic (we have an intercept and three slope parameters to bootstrap). The function we need to execute is:

bootReg <- function (formula, data, i~~ndices~~)
{
d <- data [i,]

42