Linear Regression

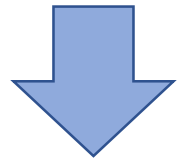
NCP Seminar with Rstudio

and Andy Field‘s DSUR Regression

general concepts of using R

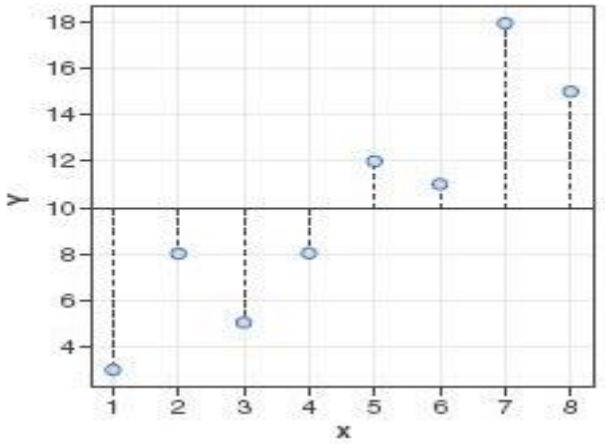
- Libraries – is there a solution to my problem / research question
- Datasets – evaluate open question / processes
- Documentation – PDF Docs / Stack Overflow, …
- Solutions: precise research, intuitive usage of R

Cook recipes & programming ???

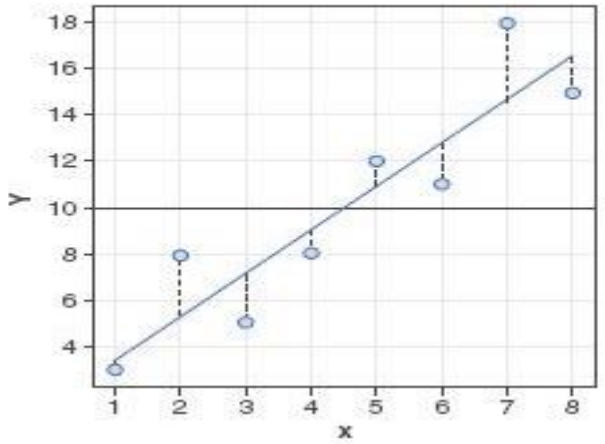


procedures

total sum of squares



SS_T uses the differences between the observed data and the mean value of Y

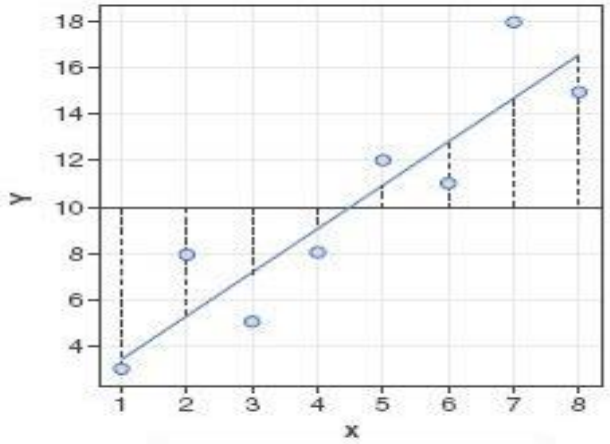


SS_R uses the differences between the observed data and the regression line

residual sum of squares

$$\sqrt{R^2} = \beta \rightarrow |r| \ (Pearson)$$

$$R^2 = \frac{model\ sum\ of\ squares}{total\ sum\ of\ squares}$$



SS_M uses the differences between the mean value of Y and the regression line

model sum of squares

improvement of model over residuals

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

$$t = \frac{b}{SE_b}$$

FIGURE 7.4 Diagram showing from where the regression sums of squares derive

simple linear regression

regression line: slope + intercept

regression slope

$$b_{yx} = \frac{cov\,(x,y)}{\sigma_x^2} = r * \frac{\sigma_y}{\sigma_x}$$

b equals derivative (slope)

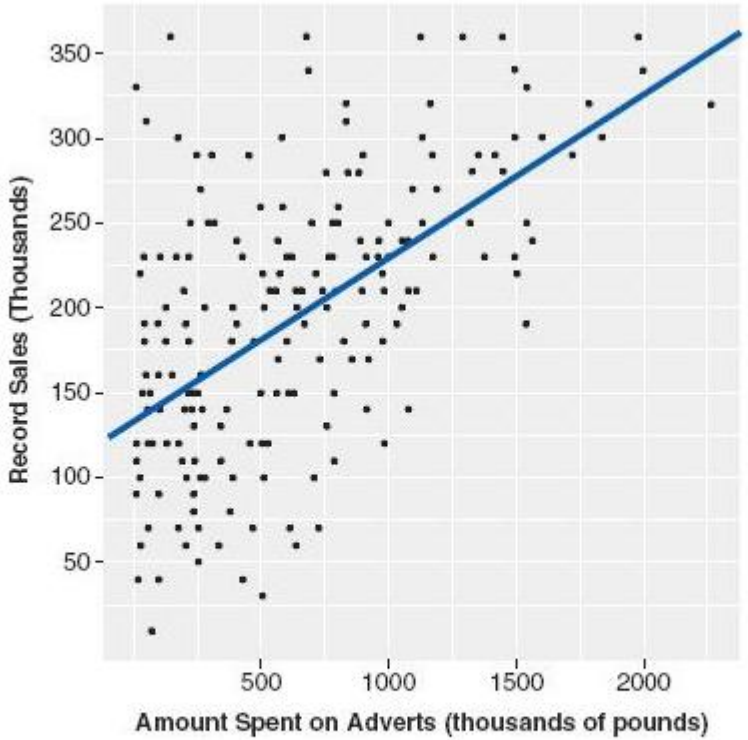


FIGURE 7.6 Scatterplot showing the relationship between album sales and the amount spent promoting the album

intercept

$$a_{yx} = \bar{y} - b_{yx} * \bar{x}$$

- method of least squares
- goodness of fit
- R -> R^2 and R^2 -> R (sqrt)

summary(albumSales.1)

```
Call:
lm(formula = sales ~ adverts, data = album1)

Residuals:
     Min       1Q   Median       3Q      Max
-152.949  -43.796   -0.393   37.040  211.866

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.341e+02  7.537e+00  17.799   <2e-16 ***
adverts     9.612e-02  9.632e-03   9.979   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 65.99 on 198 degrees of freedom
Multiple R-squared:  0.3346,    Adjusted R-squared:  0.3313
F-statistic: 99.59 on 1 and 198 DF,  p-value: < 2.2e-16
```

sig. coefficient with a = 134.1 & b = 0.096

Explained variance

Our regression model results in significantly better prediction of album sales

exercise one:

- use the functions: lm(), summary(), predict()
- how many records more than baseline were sold for certain investments?
- investment in €: c(0,150,1000,1E6, 45032)
- please report integer values.

- Solution:
- 0, 14, 96, 96124, 4238 records more.
- floor(predict(albumSales.1, investment) -
  albumSales.1$Coefficients[1])

multiple linear regression

$$Y_i = (b_0 + b_1X_{1i} + b_2X_{2i} + b_3X_{3i} + \ldots + b_nX_{ni}) + \varepsilon_i$$

methods:
- hierarchical
- forced entry
- stepwise methods

evaluate best model

finally generalization!?

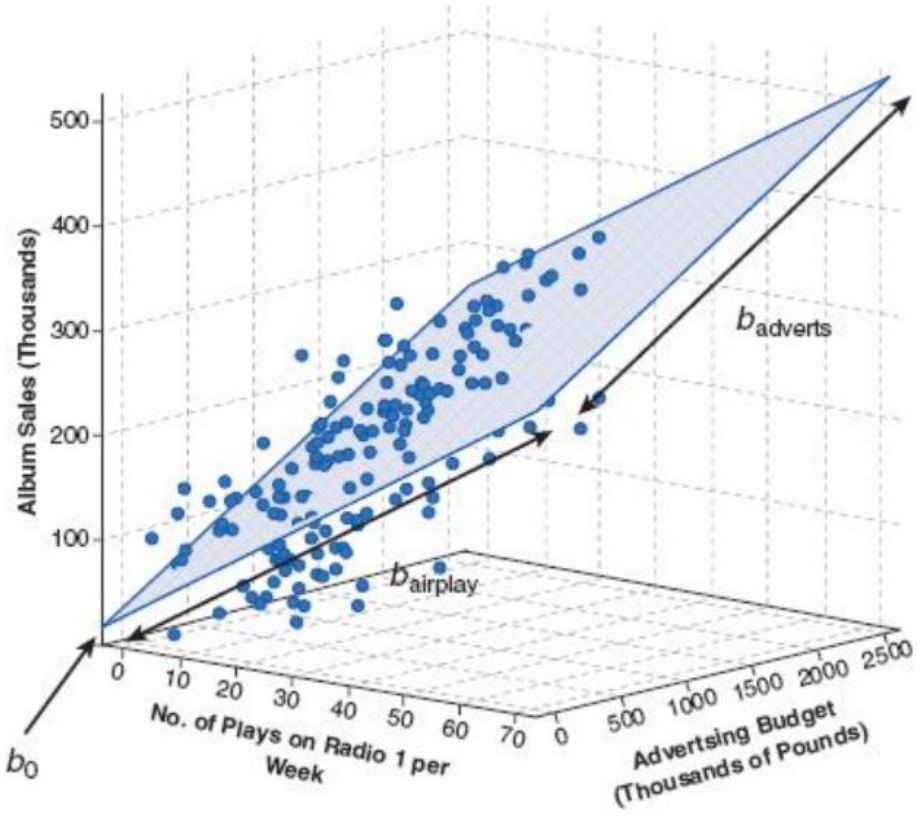


FIGURE 7.8 Scatterplot of the relationship between album sales, advertising budget and radio play

linearization of complex functions

$$Y = b_0 + b_1 f(x) + \varepsilon$$

parsimony-adjusted measures of fit

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

Akaike information criterion (AIC)

exercise two:

- use the functions: lm(), summary(), lm.beta(), confint(), anova()
- what model is the best? test all possible multiple regressions
- are all three variables necessary?
- How is the AIC for each model?

- Solution:
- Model with 3 Predictors has best explanation and lowest AIC!

how accurate is my regression model?

- Outliers, residuals and influential cases
- R functions
  - resid() - residuals
  - rstandard() - standardized residuals
  - rstudent() - studentized residuals
  - cooks.distance() - influential cases
  - dfbeta() - excluding subject-wise
  - dffits() - excluding subject-wise
  - hatvalues() - leverage

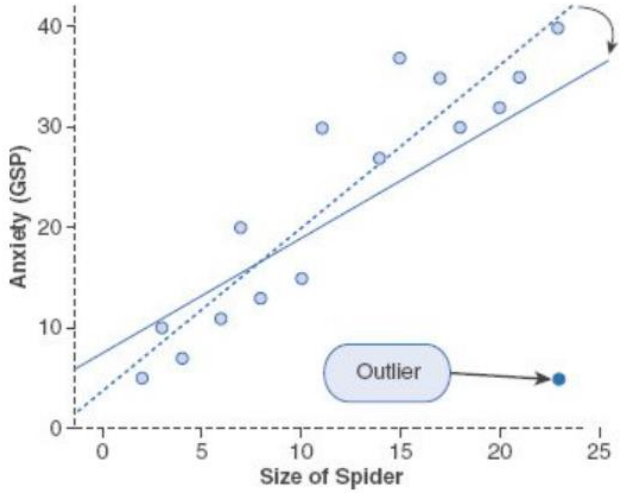


FIGURE 7.9 Graph demonstrating the effect of an outlier. The dashed line represents the original regression line for these data (see Figure 7.3), whereas the solid line represents the regression line when an outlier is present

how accurate is my regression model?

- Generalization
  - variable types (quantitative predictors)
  - non-zero variance
  - no high multicollinearity - vif(model)
  - predictors are uncorrelated with ‘external variables’
  - homoscdasticity
  - independent-errors (Durbin-Watson test)
  - normally distributed errors
  - independence
  - linearity

- R^2 explains our sample

- Adjusted R^2 for population
- prevents overfitting

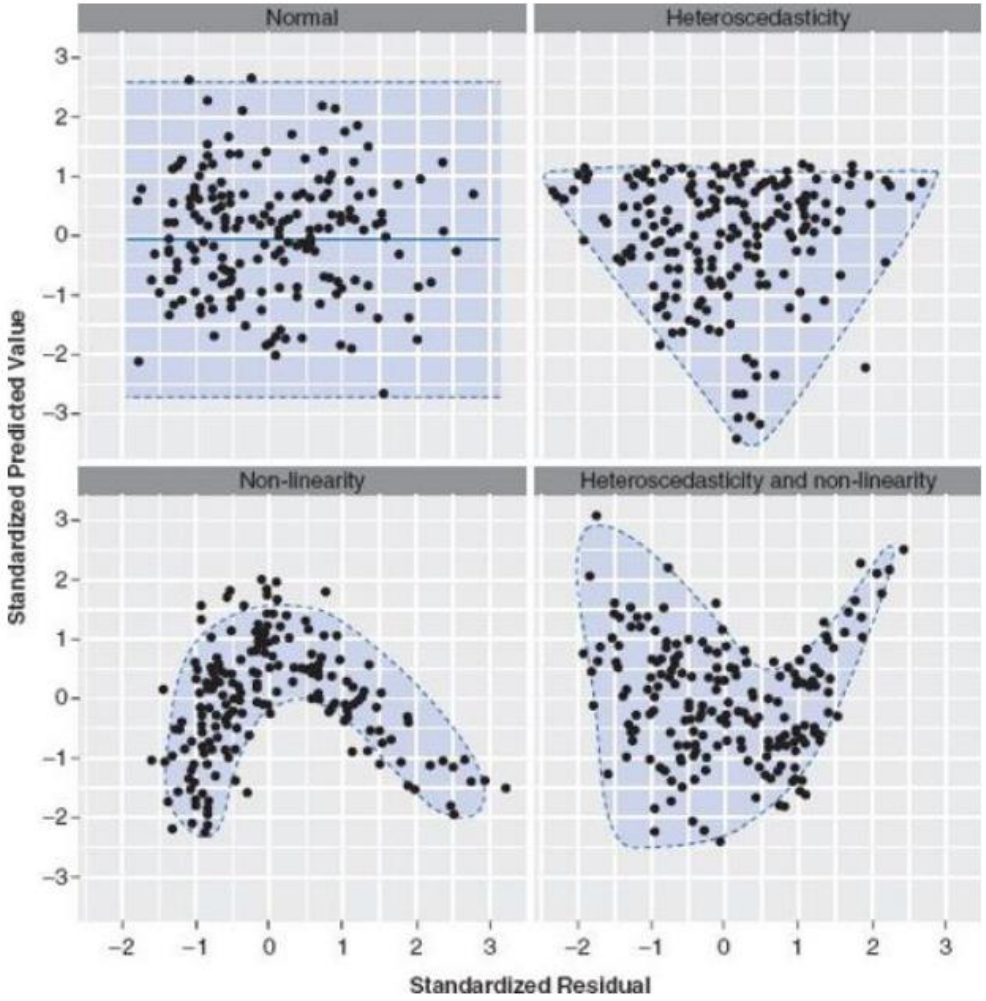
Violated assumptions: you cannot generalize the model!

Bootstrapping: resampling (with repetitions)

plot inspection

GOOD!

PROBLEM!



PROBLEM!

PROBLEM!

dummy coding (one hot encoding)

- special case multiple regression: categorical predictors (gender, …)
- biserial correlation (zero and one)
  - influence of a binary variable can be estimated with correlation coefficient
- what about more than two categories?
- dummy variables (one hot) (N-1)
- baseline group (control, majority)
- compares influence to baseline
- R dummy codes automatically!

Sidenote:
xgboost (extreme gradient boosting)
For data analysis and Kaggle‘s
winning choice

further analyses

mind basic variable naming errors!!!

Smart Alex's tasks

- **Task 1:** Run a simple regression for the pubs.dat data in Jane Superbrain Box 7.1, predicting mortality from number of pubs. Try repeating the analysis but bootstrapping the regression parameters. ①

- **Task 2:** A fashion student was interested in factors that predicted the salaries of catwalk models. She collected data from 231 models. For each model she asked them their salary per day on days when they were working **(salary),** their age **(age),** how many years they had worked as a model **(years),** and then got a panel of experts from modelling agencies to rate the attractiveness of each model as a percentage, with 100% being perfectly attractive **(beauty).** The data are in the file **Supermodel. dat.** Unfortunately, this fashion student bought some substandard statistics text and so doesn't know how to analyse her data. ☺ Can you help her out by conducting a multiple regression to see which variables predict a model's salary? How valid is the regression model? ②

further analyses

- **Task 3:** Using the Glastonbury data from this chapter, which you should've already analysed, comment on whether you think the model is reliable and generalizable. ③
- **Task 4:** A study was carried out to explore the relationship between **Aggression** and several potential predicting factors in 666 children who had an older sibling. Variables measured were **Parenting_Style** (high score = bad parenting practices), **Computer_Games** (high score = more time spent playing computer games), **Television** (high score = more time spent watching television), **Diet** (high score = the child has a good diet low in additives), and **Sibling_Aggression** (high score = more aggression seen in their older sibling). Past research indicated that parenting style and sibling aggression were good predictors of the level of aggression in the younger child. All other variables were treated in an exploratory fashion. The data are in the file **ChildAggression.dat.** Analyse them with multiple regression. ②

Answers can be found on the companion website.

https://studysites.uk.sagepub.com/dsur/study/default.htm