Statistical modeling in logistics

MMS075

Lecture 5a – Reviewing diagnostics, optimal advertising strategy, Classification problems

Acknowledgement: Some of the figures in this presentation are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani

Department of Mechanics and Maritime Sciences
Division of Vehicle Safety

Outline

- Reviewing regression diagnostics
  - Studentized vs standardized residuals
  - Outliers, high leverage points, influential points
  - Homoscedasticity and normality
  - Collinearity example from computer lab
- Optimal advertising strategy
  - Reviewing advertisement example
  - Which budget distribution gives highest sales?
- Preparation for logistic regression
  - The classification problem
  - Reviewing the exponential function and logarithm
- Feedback

Recommended resources

Reading in ISL: Ch 4 introduction and sections 4.1-4.3 for theory, 4.6.1, 4.6.2 for R codes

Online resources for discussion of regression diagnostics (examples):

https://online.stat.psu.edu/stat462/node/87/

http://www.sthda.com/english/articles/39-regression-model-diagnostics/161-linear-regression-assumptions-and-diagnostics-in-r-essentials/

Online resources for exponential function and logarithm (examples):

Nykamp DQ, "Basic idea and rules for logarithms." From Math Insight. https://mathinsight.org/logarithm_basics

https://www.mathsisfun.com/algebra/exponents-logarithms.html

http://tutorial.math.lamar.edu/Classes/CalcI/ExpLogEqns.aspx

The videos from the Statistical Learning course are available at this link. Relevant videos for the new material today:

- **Introduction to Classification** (10:25)
- **Logistic Regression and Maximum Likelihood** (9:07)
- **Lab: Logistic Regression** (10:14)

Reviewing regression diagnostics

Studentized vs standardized residuals

Outliers, high leverage points, influential points

Homoscedasticity and normality

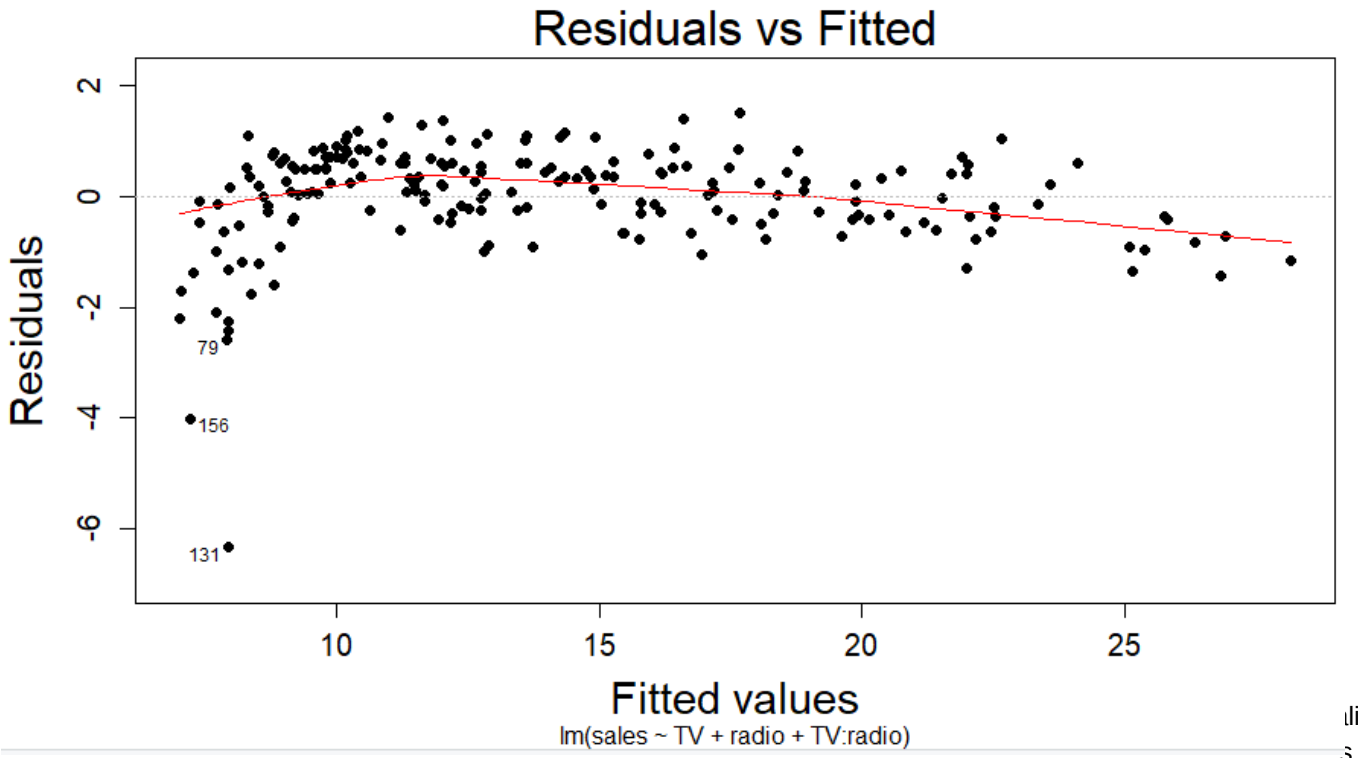
Collinearity example from computer lab

Recall: the residual plot

- For each data point i, we have an observed response y_i, a predicted (/fitted) value $\hat{y}_i$ and their difference is the residual: $e_i = y_i - \hat{y}_i$

- **Residual plot:**
  - predicted values on x-axis
  - residuals on y-axis

- Pattern indicates problem



Residuals in standard error units

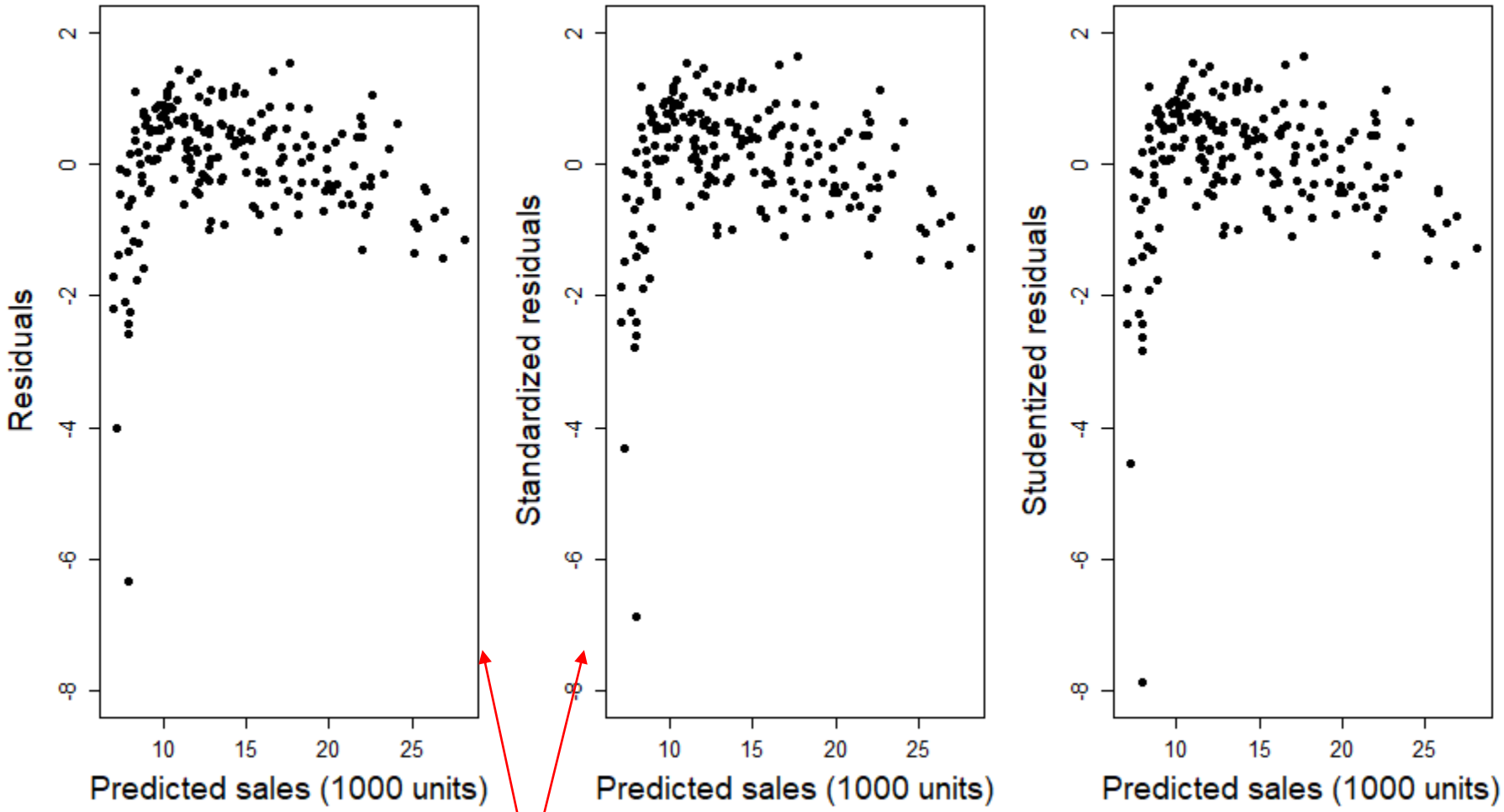
- Studentized and standardized residuals both divide each residual value with an estimated standard error

- For **standardized residuals**, the error estimate is based on *all residuals*. For **studentized residuals**, the estimate for a specific residual is based on *all other residuals except the one which is considered.*

- Studentized residuals are better for detecting outliers that deviate a lot from their estimate. Thresholds are -3 and 3.

- Sometimes, standardized residuals with the same thresholds are used to identify outliers → you should specify which measure you use

Residual plot types for advertising model



RSE=0.94 makes differences small

Reviewing regression diagnostics

Studentized vs standardized residuals

Outliers, high leverage points, influential points

Homoscedasticity and normality

Collinearity example from computer lab

Definitions

- Those points are called **outliers** whose response value y_i is far from the predicted value $\hat{y}_i$

- Those points are called **high leverage points** whose predictor values are outside the usual predictor ranges or have an unusual combination of usual predictor values

- Those points are called **influential points** that have a large influence on the prediction model; high leverage points that are also outliers are often influential points

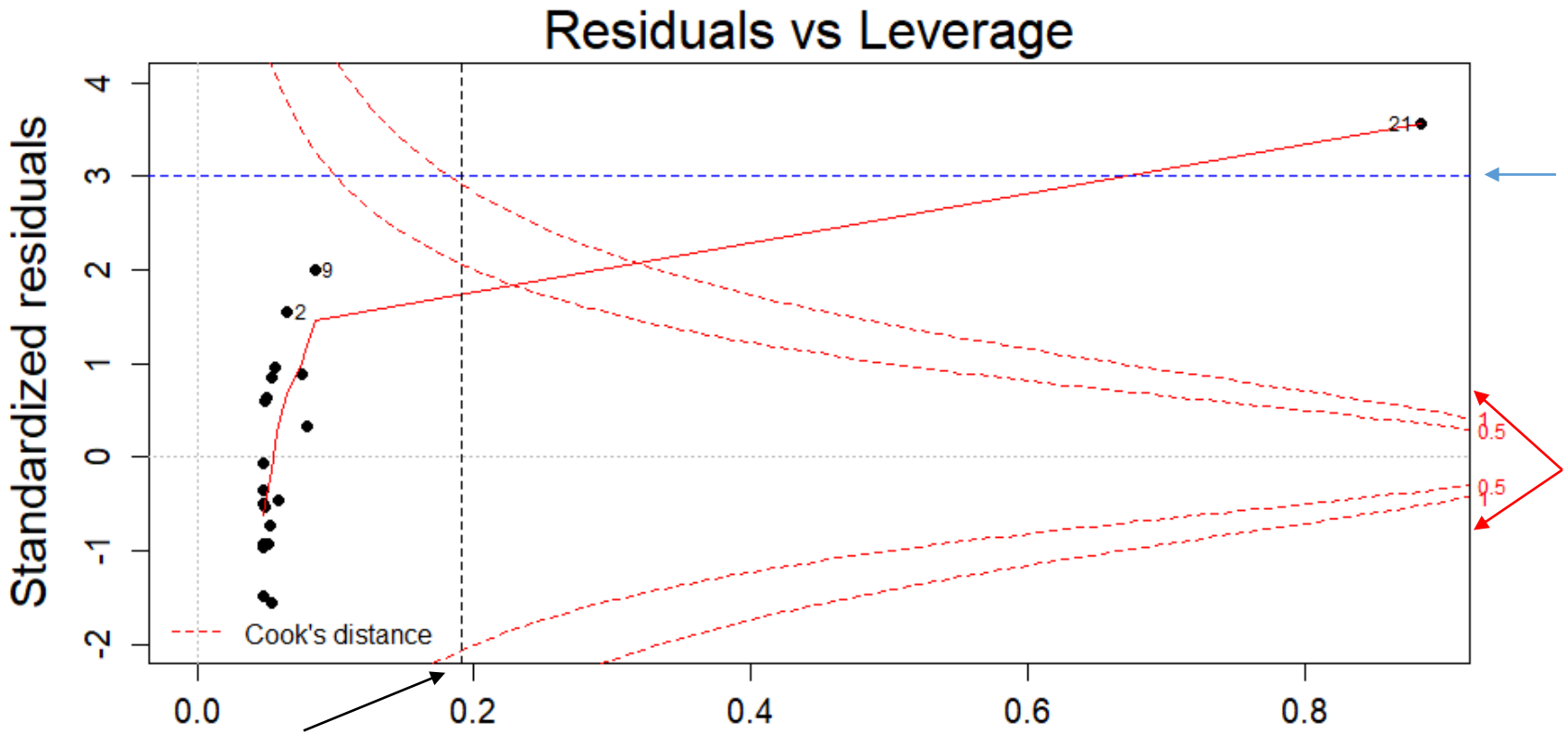
How to characterize them?

- Outliers are those points with studentized residual values (or standardized residual values) below -3 or above 3
- High leverage points are those whose leverage statistic exceeds (p+1)/n by a lot – e.g. 2(p+1)/n can be a threshold*
- Influential points can be characterized in different ways; a characterization based on Cook's distance is used in R

*Bruce, P. & Bruce, A. (2017), Practical Statistics for Data Scientists

Graphical representation

- All these point types can be easily spotted in a default graph in R for regression models after adding some extra lines:



Points above this line or below -3 are **outliers**

Points in these corners are **influential**

Reviewing regression diagnostics

Studentized vs standardized residuals

Outliers, high leverage points, influential points

Homoscedasticity and normality

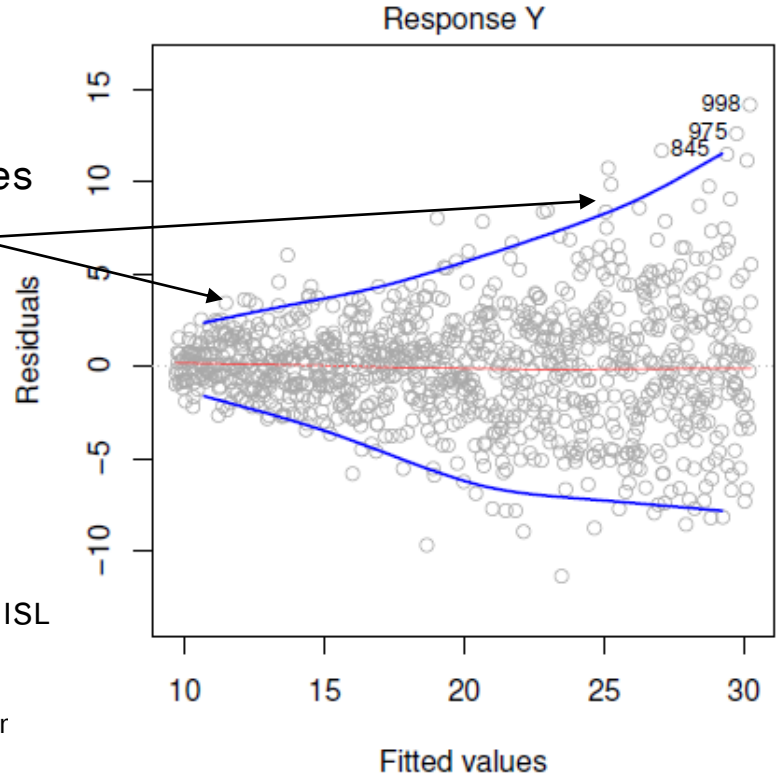
Collinearity example from computer lab

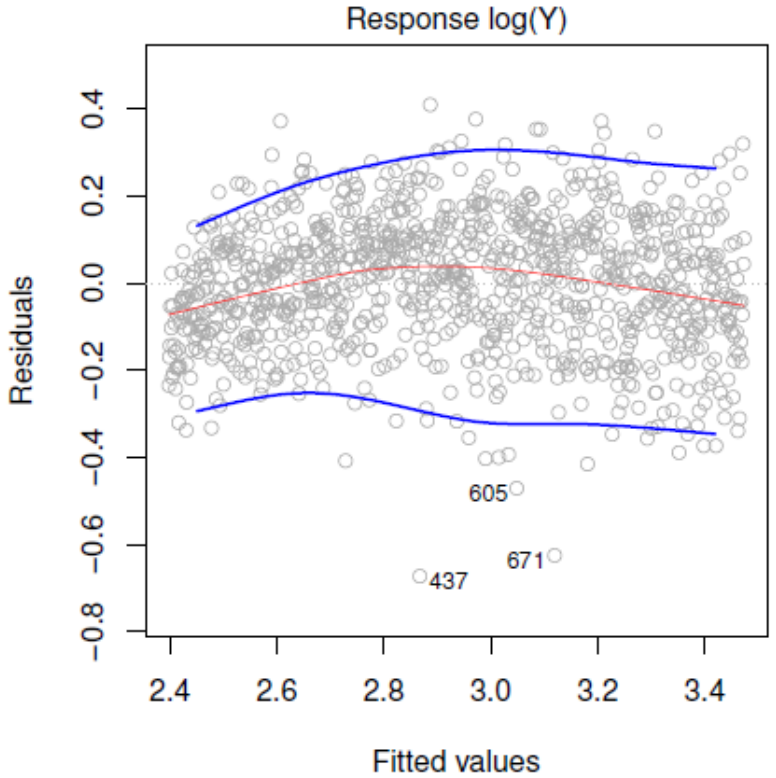
Recall: spot & handle heteroscedasticity

- Check residual plots for patterns (e.g. funnel shape, bow tie shape).

- Typical situation: variance of error terms increases with the value of the response → transforming the response by the log function may help:

Funnel shape indicates changing variance



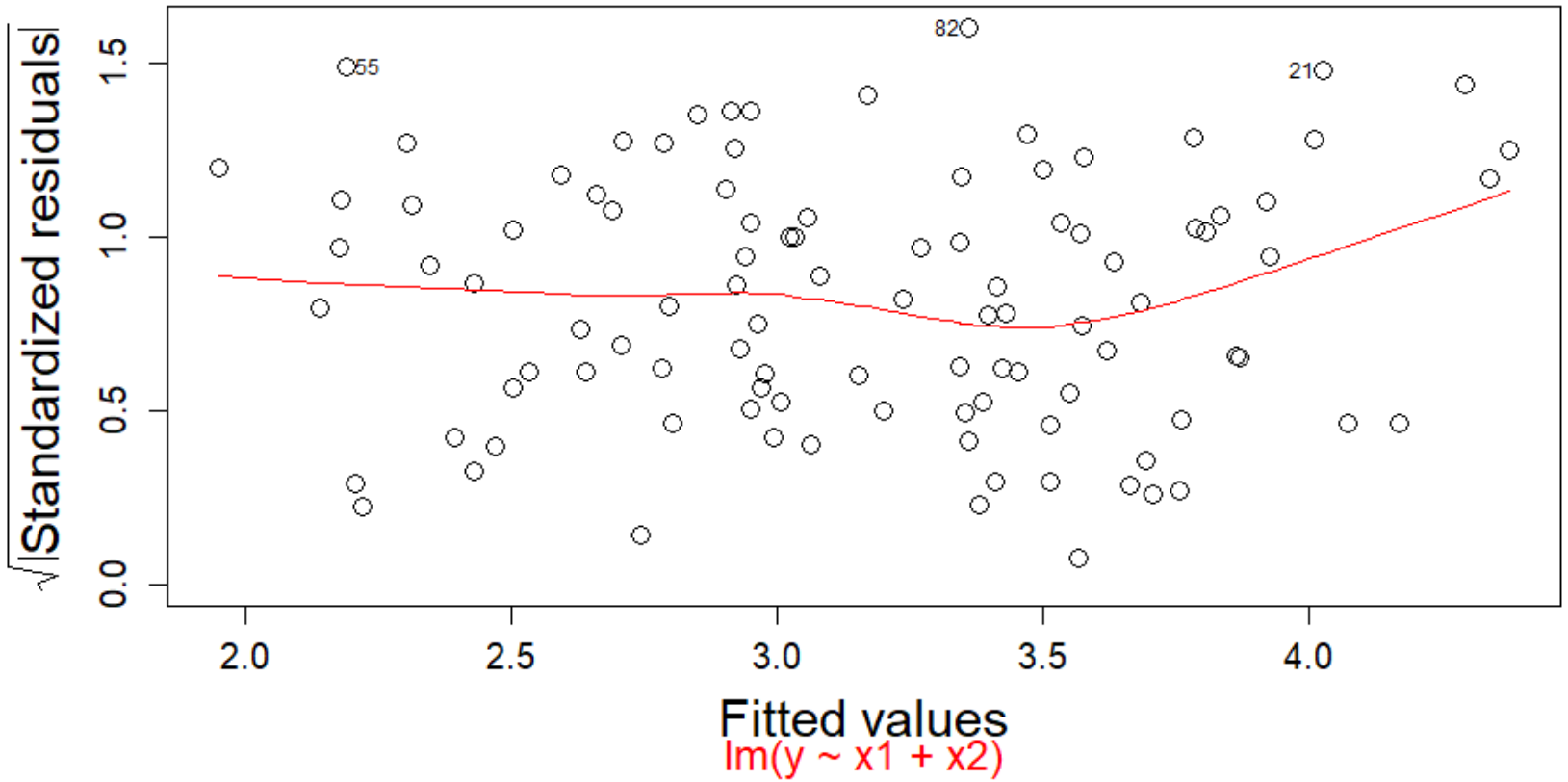


No evidence of changing variance in residual plot with tranformed response

Source: Figure 3.11 in ISL

Alternative in R: Scale-Location plot

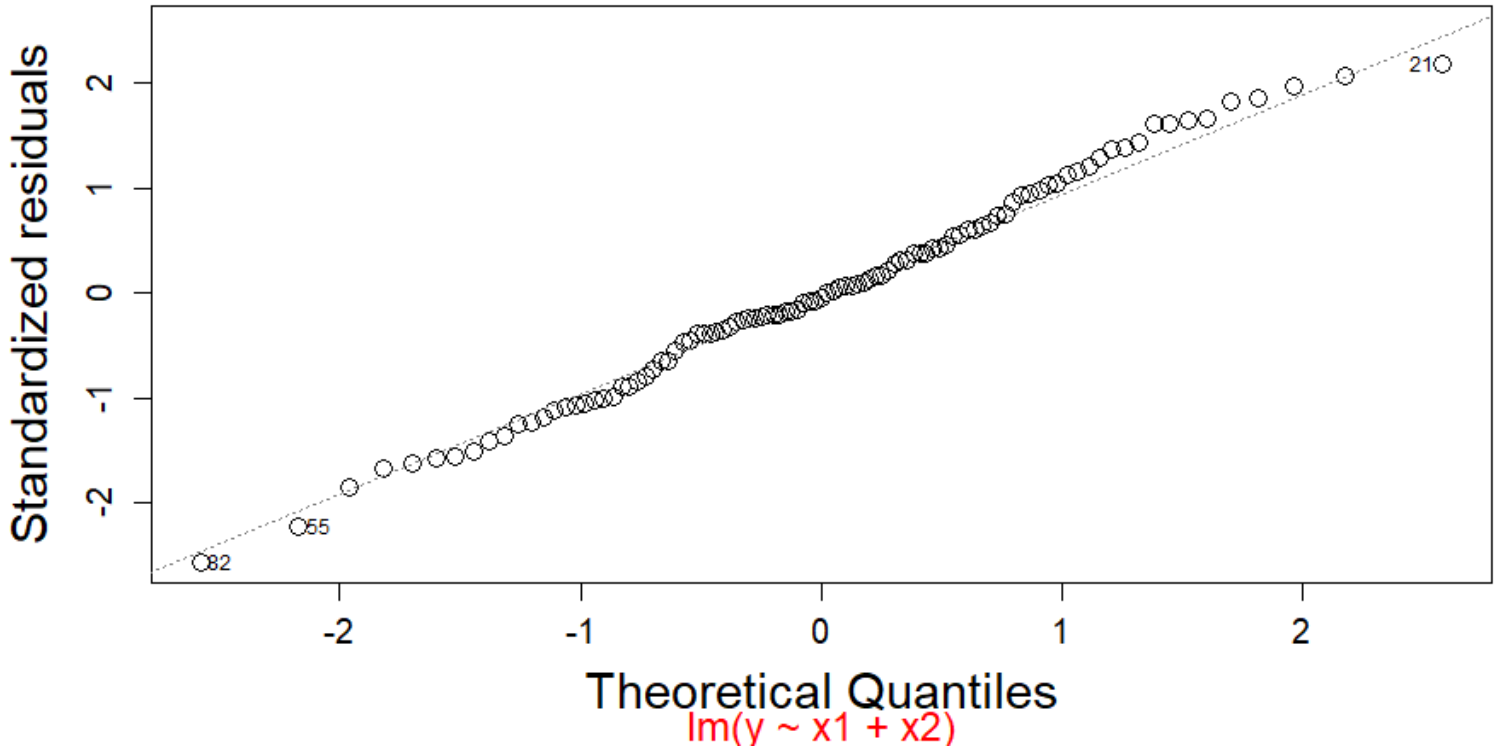
- Straight line indicates equal variances



Normality of errors

- Normality of errors is an assumption, but the model is robust to small deviations / symmetric errors

- Normal Q-Q plot in R: check that points are close to the diagonal



Reviewing regression diagnostics

Studentized vs standardized residuals

Outliers, high leverage points, influential points

Homoscedasticity and normality

Collinearity example from computer lab

Code for Exercise 14 on page 125 of ISL (1)

- Definition of the model

```
set.seed(1)                          # Get same random numbers each time we run the code
x1=runif(100)                        # Generate 100 random numbers between 0 and 1
x2=0.5*x1+rnorm(100)/10              # x2 is constant times x1 + small random error
y=2+2*x1+0.3*x2+rnorm(100)           # Define relationship between predictors and response
```

- These lines define that the **true model** is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon, \text{ with } \beta_0 = 2,\ \beta_1 = 2,\ \beta_2 = 0.3,\ \varepsilon \sim N(0, 1)$$

Code for Exercise 14 on page 125 of ISL (2)

- Fitting multiple and simple linear regression models

```
> summary(lm(y~x1+x2))$coefficients
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 2.130500  0.2318817 9.1878742 7.606713e-15
x1          1.439555  0.7211795 1.9961126 4.872517e-02
x2          1.009674  1.1337225 0.8905831 3.753565e-01
> summary(lm(y~x1))$coefficients
             Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 2.112394  0.2307448 9.154676 8.269388e-15
x1          1.975929  0.3962774 4.986227 2.660579e-06
> summary(lm(y~x2))$coefficients
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 2.389949  0.1949307 12.260508 1.682395e-21
x2          2.899585  0.6330467  4.580365 1.366430e-05
```

Estimates are not all close to true coefficients of

$$\beta_0 = 2,\ \beta_1 = 2,\ \beta_2 = 0.3$$

Both predictors have much smaller p-values in the simple linear regression model

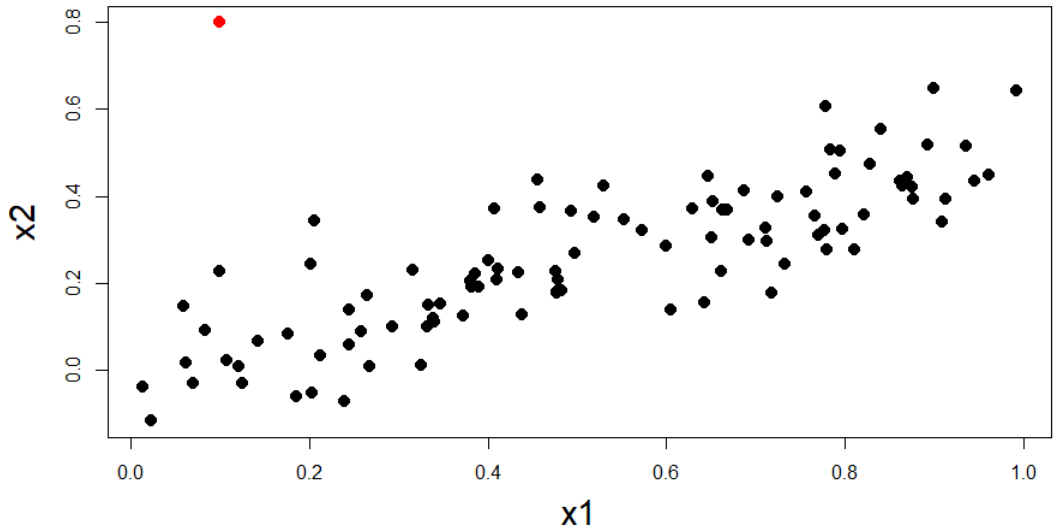
- In the presence of x1, the other predictor does not improve the model

Code for Exercise 14 on page 125 of ISL (3)

- Adding a new, mismeasured observation

```
x1=c(x1,0.1)    # Adding a new value of 0.1 to x1
x2=c(x2,0.8)    # Adding a new value of 0.8 to x2
y=c(y,6)        # Adding a new response value of 6 that will correspond to the new predictor values
```

- The new observation has a somewhat unusual x2 value and very unusual combination of (x1,x2):



Code for Exercise 14 on page 125 of ISL (4)

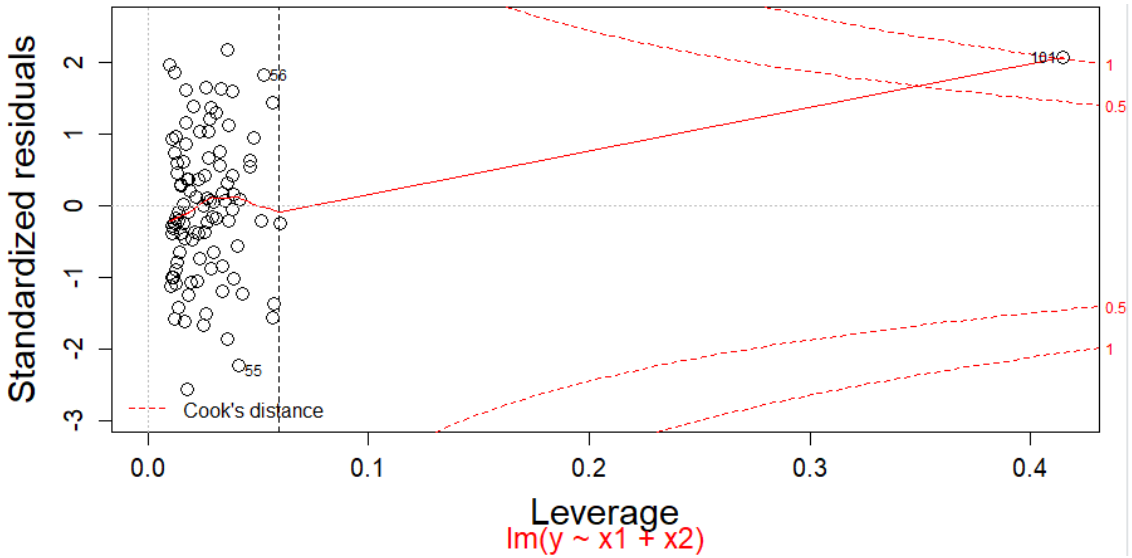
- Fitting multiple and simple linear regression models

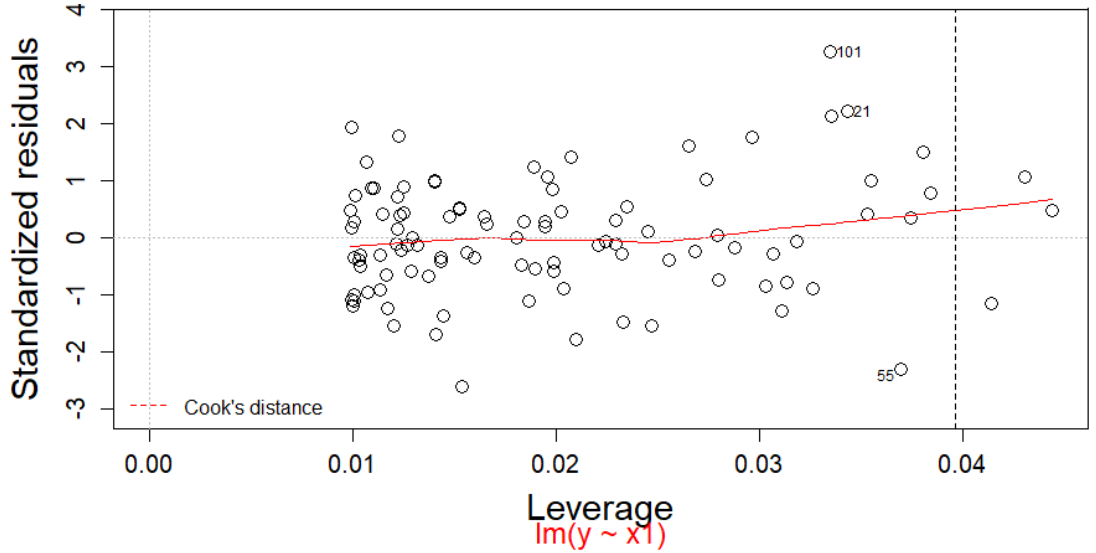
```
> summary(lm(y~x1+x2))$coefficients
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 2.2266917  0.2313578 9.6244495 7.909712e-16
x1          0.5394397  0.5921970 0.9109125 3.645766e-01
x2          2.5145694  0.8976915 2.8011508 6.135787e-03
> summary(lm(y~x1))$coefficients
             Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 2.256927  0.2389635 9.444654 1.780749e-15
x1          1.765695  0.4123781 4.281739 4.294817e-05
> summary(lm(y~x2))$coefficients
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 2.345107  0.1912183 12.264029 1.403060e-21
x2          3.119050  0.6040352  5.163689 1.253125e-06
```

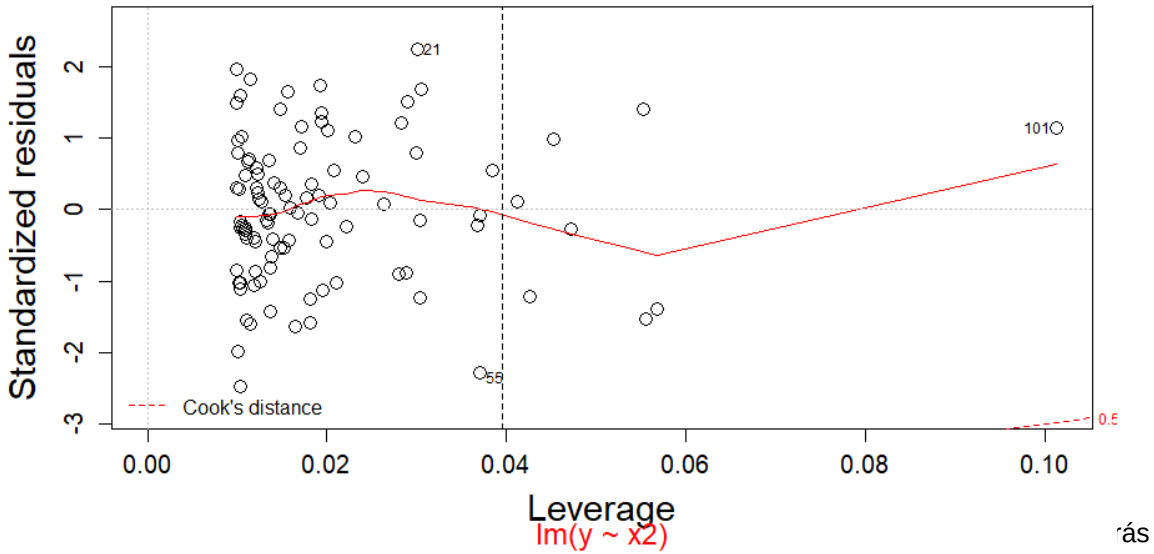
Estimates in multiple regression model change very substantially

With the new observation, x2 has smaller p-values than x1

Influential and high leverage points







Optimal advertising strategy

Reviewing advertisement example

Which budget distribution gives highest sales?

Main questions for advertisement model 1

- **Is there a relationship between advertising and sales?** Yes, multiple regression model with all variables is highly significant

- **How strong is the relationship?** R^2 value indicates that about 90% of the variability in sales can be explained by the predictors

- **Which ad types contribute to sales?** P-values of radio and TV indicate significant contribution for these ad types. However, in the presence of the other two variables, newspaper is not significant

- **How large is the effect of each medium on sales?** $1000 extra on TV ads → 43-49 extra units sold. $1000 extra on radio ads → 172-206 extra units sold. Estimates will be refined when considering synergies

Main questions for advertisement model 2

- **How accurately can we predict sales?** We can quantify uncertainty about average sales by confidence interval and about specific sales by prediction interval

- **Is the relationship linear?** Residual plot suggests nonlinear effects

- **Is there synergy between advertising media?**
  - Yes, there is synergy between TV and radio budget, the model has improved by including an interaction term
  - Increased TV budget improves effectiveness of radio advertisements & money on radio advertisement can increase effectiveness of TV advertisements

Optimal advertising strategy

Reviewing advertisement example

Which budget distribution gives highest sales?

Recall: Company goal was to optimize advertising

- A company wants to sell a specific product

- Advertising may increase sales – how to do this best? How much money to spend & how to spend that amount? How would you do this?

One reasonable plan:

- Identify different means of advertising
- Collect data on #(units sold) as a function of $ spent
- Model different solutions
- Choose the best available option

Dividing a fixed total budget

- The prediction model with the coefficient estimates:

$$\widehat{\text{sales}} = 6.7502 + 0.0191 \times \text{TV} + 0.0289 \times \text{radio} + 0.0011 \times \text{TV} \times \text{radio}$$

- If the total advertisement budget available is \$100 000, how should the management spend that amount?

- We have seen (in exercise class 4) that dividing the total budget equally between radio and TV gives higher estimated sales than spending it all on either. Can they do any better?

Finding optimal division of $100 000

$100 000 is divided between TV and radio → $\text{radio} = 100 - \text{TV}$ and

$$\widehat{\text{sales}} = 6.7502 + 0.0191 \times \text{TV} + 0.0289 \times (100 - \text{TV}) + 0.0011 \times \text{TV} \times (100 - \text{TV})$$
$$= 9.6402 + 0.01002 \times \text{TV} - 0.0011 \times (\text{TV})^2$$

How to find the value for TV that maximizes this? Either find value where derivative is 0 or rewrite equation as:

$$\widehat{\text{sales}} = c - 0.0011 \times \left(TV - \frac{0.01002}{2 \times 0.0011} \right)^2$$

This function takes maximum at $TV = \frac{0.01002}{0.0022} = 45.545$

→ $45545 on TV ads and $54455 on radio ads gives highest sales

Preparation for logistic regression

The classification problem

Reviewing the exponential function and logarithm

Qualitative response

So far we have considered:

- Quantitative response & quantitative predictors
- Quantitative response & categorical predictors (→ defined dummy variables to deal with those)

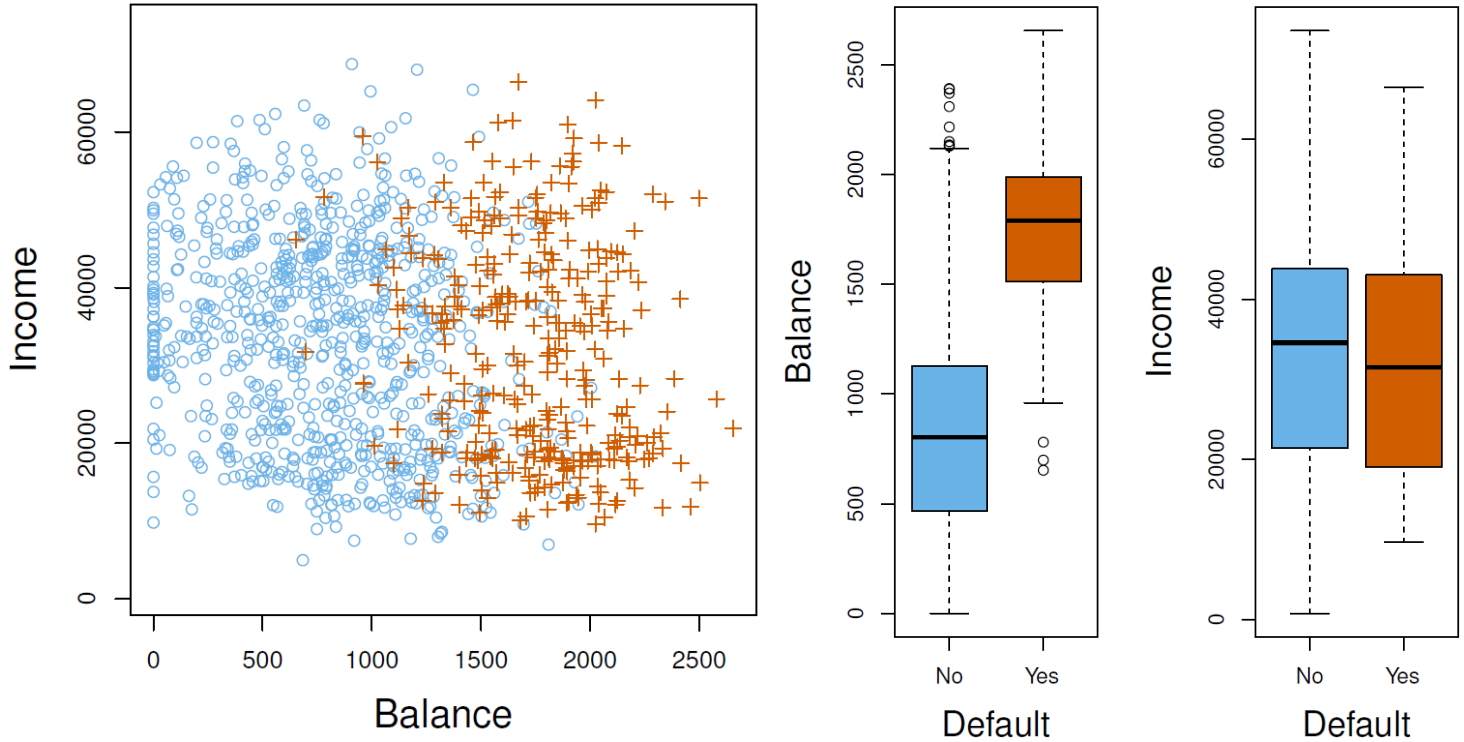
What to do if the **response is categorical**? Examples:

- Select most likely diagnosis based on symptoms
- Classify email as regular email or spam
- Predict most likely outcome of car crash at given speed

Example in ISL: credit card default

- How can a bank predict if an individual will default on their credit card payment (i.e. fail to pay the credit card debt by the due date)?

- Predictors considered: annual income & monthly credit card balance



Logistic regression equation

- Logistic regression is very commonly used to address classification problems

- The model equation for binary response is provided below, motivation & properties will be discussed in the next lectures

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

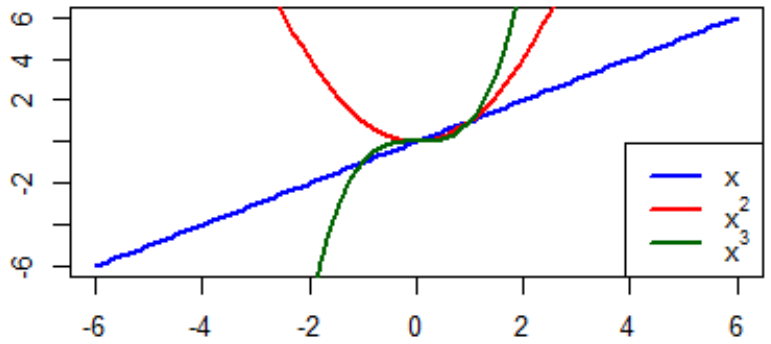
- Note: in line with the ISL book, we use ”log()” and not ”ln()” to denote natural logarithm (i.e. with base $e = 2.718...$)

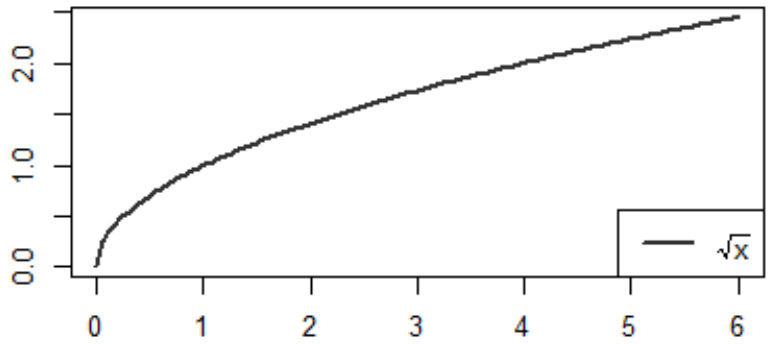
Preparation for logistic regression

The classification problem

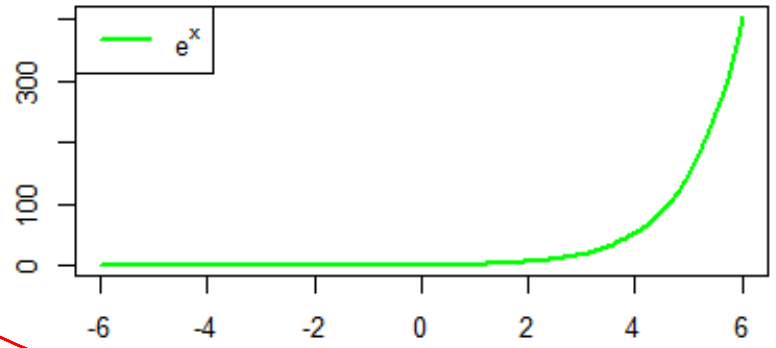
Reviewing the exponential function and logarithm

Recall: graphs of some important functions

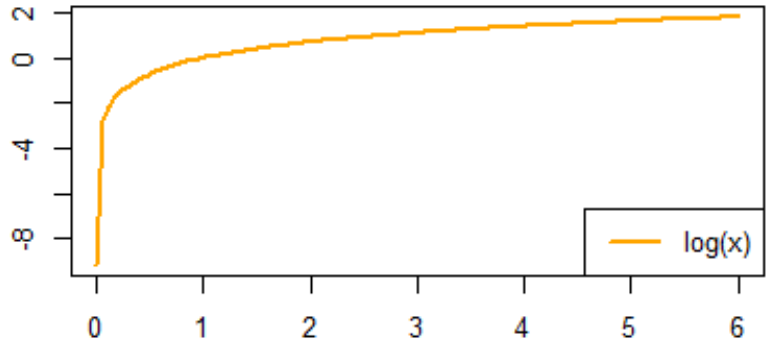




Exponential function



Logarithmic function



Basic properties of exp() and log()

- Some important properties are described below; detailed descriptions and exercises are available online (e.g. at the sites listed among recommended resources)

- $\exp(x) = e^x$ is defined for all values of x and is always positive

- $\log(x)$ is defined for positive values of x as the inverse of e^x; that is, $\log(x)$ is the number such that $e^{\log(x)} = x$

- Both functions are strictly increasing in x (i.e. take larger values when increasing x)

Basic properties of exp() and log() cont.

- $e^0 = 1, \log(1) = 0$

- The exponential function is greater than 1 for positive x and smaller than 1 for negative x

- Basic operations with the exponential function and logarithm:

$$e^{a+b} = e^a \cdot e^b \qquad \log(a \cdot b) = \log a + \log b$$

$$e^{a-b} = \frac{e^a}{e^b} \qquad \log(\frac{a}{b}) = \log a - \log b$$

$$\log a^b = b \cdot \log a$$

Feedback

Student representatives

- The names of student representatives for this course & contact information are published on the course homepage
- Please inform the student representatives about your impression of the course so far, e.g.:
  - Overall opinion
  - Potential issues
  - Improvement suggestions
- This is important even if you give feedback via www.menti.com

Feedback quiz - optional

Feedback is essential to me so that I can improve the lectures during the course. All comments about today's class or the course in general are welcome!

If you are willing to give feedback, please follow these steps:

1. Go to www.menti.com
2. Enter the code 32 21 69
3. Answer the questions or enter other comments related to the course