Classical and Bayesian Statistics

Problems 9

Problem 9.1

The library `ISLR` contains the data set `Carseats`. We want `Sales` (number of child car seats) based on different predictors in 400 different locations.

The data set contains qualitative predictors, such as `ShelveLoc` as an indicator of the location in the rack, i.e. the space in a shop where the car seat is displayed. The predictor assumes the three values `Bath`, `Medium` and `Good`. For qualitative variables `R` generates dummy variables automatically.

(∗) a) Examine the data set with `head(Carseat)` and `?Carseat`.

(∗∗) b) Find a multiple regression model with `lm()` to predict `Sales` from `Price`, `Urban` and `US`.

(∗∗) c) Interpret the coefficients in this model. Be aware that some variables are qualitative.

(∗) d) Write the model as an equation. Make sure that you treat the qualitative variables correctly.

(∗) e) For which predictors can the null hypothesis $H_0 : \beta_j = 0$ be rejected?

(∗∗) f) Based on the previous question, find a smaller model that only uses predictors for which there is evidence of a relationship with the response variable.

(∗∗) g) How exactly do the models in b) and f) fit the data?

Problem 9.2

In a third world country, 1 % of the people suffer from a certain infectious disease. A test correctly indicates the disease in those actually ill with a probability of 98 %. Unfortunately, the test also indicates that 3 % of healthy people are sick.

Let S denote a sick person and T a person who tested positive.

(∗) a) Interpret (do not calculate!) the probabilities

$$P(S), \qquad P(\overline{T}), \qquad P(S \mid T), \qquad P(T \mid S), \qquad P(\overline{T} \mid \overline{S})$$

(∗) b) Denote the given probabilities in the problem definition using the notation as in a).

(∗) c) Calculate $P(\overline{S})$.

(∗) d) What is the probability that the test show a positive result for a randomly selected person?

Hint: Use the law of total probability.

(∗) e) What is the probability that a person who tested positive is actually sick? Interpret the result.

Hint: Use the Bayes' Theorem.

(∗∗) f) What is the probability that a person tested negative is actually healthy? Interpret the result.

Hint: Use the Bayes' Theorem.

Problem 9.3

A doping test is carried out at a sporting event. If an athlete has doped, the test is positive with a probability of 99 %.

However, if an athlete has not doped, the test will still show a positive result with a probability of 5 %. From experience we know that 20 % of the athletes are doped.

(∗∗) a) What is the probability that a doping test is positive.

(∗∗) b) What is the probability that the test is negative even though the athlete has doped?

(∗∗) c) What is the probability that an athlete has doped, if the doping test is negative.

Problem 9.4

An polygraph (lie detector) test is routinely performed on employees who work in sensitive positions. Let $+$ denote the event that the test is positive, i.e., that the polygraph indicates that the employee has lied. With W we denote the event that the employee told the truth and with L that the employee lied.

From former investigations of polygraphs test, we know that

$$P(+ \mid L) = 0.88 \quad \text{and} \quad P(- \mid W) = 0.86$$

Furthermore, we know that

$$P(W) = 0.99$$

(∗) a) Interpret the probabilities $P(W)$ and $P(+ \mid L)$.

(∗∗) b) For a person, the detector indicates that a lie has been told. What is the effective probability that this person has lied?

(∗∗) c) Interpret the result from b) in 2-3 sentences. How significant do you hold polygraph test?

Problem 9.5

(∗∗) The serum test examines pregnant women for babies with Down syndrome. The test is a very good but not perfect test. About 1 % of the babies have Down syndrome. If the baby has Down syndrome, there is a 90 % probability that the result will be positive. If the baby is not affected, there is still a 1 % probability that the result will be positive. A pregnant woman has been tested and the result is positive. What is the probability that your baby actually has Down syndrome?

Problem 9.6

The smoke sensors in a factory report a fire with a probability of 0.95. On a day without fire, they will give a false alarm with a probability of 0.01. One fire is expected per year.

(∗ ∗ ∗) a) The alarm system reports a fire. What is the probability that there is in fact a fire? Interpret the result.

(∗ ∗ ∗) b) In one night it is quiet (no alarm). What is the probability that there is really a fire? Interpret the result.

Problem 9.7

An insurance company believes that you can divide people into two classes - unlucky and others. Their statistics show that an unlucky person will have an accident within one year with a probability of 0.4, while for all others the probability is only 0.2. We assume that 30 percent of the population is unlucky.

(∗∗) a) What is the probability that a new customer will have an accident within one year after signing the contract?

(∗∗) b) A new customer has an accident within one year. What is the probability that he is an unlucky person?

Classical and Bayesian Statistic

Sample solution for Problems 9

Solution 9.1

a) Data set:

```
library(ISLR)
head(Carseats)

##   Sales CompPrice Income Advertising Population Price ShelveLoc Age
## 1  9.50       138     73          11        276   120       Bad  42
## 2 11.22       111     48          16        260    83      Good  65
## 3 10.06       113     35          10        269    80    Medium  59
## 4  7.40       117    100           4        466    97    Medium  55
## 5  4.15       141     64           3        340   128       Bad  38
## 6 10.81       124    113          13        501    72       Bad  78
##   Education Urban  US
## 1        17   Yes Yes
## 2        10   Yes Yes
## 3        12   Yes Yes
## 4        14   Yes Yes
## 5        13   Yes  No
## 6        16    No Yes
```

b) Output:

```
fit <- lm(Sales ~ Price + Urban + US, data = Carseats)
summary(fit)

##
## Call:
## lm(formula = Sales ~ Price + Urban + US, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9206 -1.6220 -0.0564  1.5786  7.0581
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.043469   0.651012  20.036  < 2e-16 ***
## Price       -0.054459   0.005242 -10.389  < 2e-16 ***
## UrbanYes    -0.021916   0.271650  -0.081    0.936
## USYes        1.200573   0.259042   4.635 4.86e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.472 on 396 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2335
## F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

c) Interpretation of the coefficients:

- The coefficient 13.04 is a bit difficult to interpret. According to the model under d), this is the average sales figures in shops reached in rural areas outside the USA, with the price of child seats still being \$0 (not very realistic).
- The coefficient -0.05 indicates that for an increase of one dollar, an average of 0.05 units of child seats are sold less.
- The coefficient -0.021 means that on average 0.021 less units are sold in urban areas compared to rural areas. However, the p value is very high, so this is more of a random variation.
- The 1.2 coefficient means that 1.2 more units are sold within the US compared to shops outside the USA. Perhaps child seats are compulsory in the USA.

d) Model: For `Urban` we choose the dummy variable:

$$x_{2i} = \begin{cases} 1 & \text{if } i\text{th person lives in urban area} \\ 0 & \text{if } i\text{-th person lives in rural area} \end{cases}$$

For `US` we choose the dummy variable

$$x_{3i} = \begin{cases} 1 & \text{if } i\text{th person lives in the USA} \\ 0 & \text{if } i\text{-th person does not live in the USA} \end{cases}$$

The model is then

$$\begin{aligned} y_i &= \beta_0 + \beta_1 \cdot \texttt{Price} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i \\ &= \beta_0 + \beta_1 \cdot \texttt{Price} + \begin{cases} \beta_2 + \beta_3 + \varepsilon_i & \text{if } i\text{-th person lives in urban are in the USA} \\ \beta_2 + \varepsilon_i & \text{if } i\text{th person lives in urban area outside the USA} \\ \beta_3 + \varepsilon_i & \text{if } i\text{th person lives in rural area in the USA} \\ \varepsilon_i & \text{if } i\text{th person lives in rural area outside the USA} \end{cases} \end{aligned}$$

e) For all except `Urban`

f) Output:

```
fit <- lm(Sales ~ Price + US, data = Carseats)
summary(fit)

##
## Call:
```

```
## lm(formula = Sales ~ Price + US, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9269 -1.6286 -0.0574  1.5766  7.0515
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.03079    0.63098  20.652  < 2e-16 ***
## Price       -0.05448    0.00523 -10.416  < 2e-16 ***
## USYes        1.19964    0.25846   4.641 4.71e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.469 on 397 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2354
## F-statistic: 62.43 on 2 and 397 DF,  p-value: < 2.2e-16
```

Model: For `US` we choose the dummy variable

$$x_{2i} = \begin{cases} 1 & \text{if } i\text{th person lives in the USA} \\ 0 & \text{if } i\text{-th person does not live in the USA} \end{cases}$$

The model is then

$$\begin{aligned} y_i &= \beta_0 + \beta_1 \cdot \texttt{Price} + \beta_2 x_{2i} + \varepsilon_i \\ &= \beta_0 + \beta_1 \cdot \texttt{Price} + \begin{cases} \beta_2 + \varepsilon_i & \text{if } i\text{th person lives in the USA} \\ \varepsilon_i & \text{if } i\text{-th person does not live in the USA} \end{cases} \\ &= 13.03 - 0.055 \cdot \texttt{Price} + \begin{cases} 1.2 + \varepsilon_i & \text{if } i\text{-th person lives in the USA} \\ \varepsilon_i & \text{if } i\text{-th person does not live in the USA} \end{cases} \end{aligned}$$

g) In both models the correlation is proven (p-value for F-value practically 0), but if we look at the R^2-values, the one with 0.2393 is relatively bad. That means that although the correlation is verified the fit is bad, because only 23 % of the variability of the `Sales` can be explained by the model.

Solution 9.2

a)
- $P(S)$: Probability that a randomly selected person in this country is really sick.
- $P(\overline{T})$: Probability that a person is tested negative (illness is not displayed).
- $P(S \mid T)$: Probability that a person who tested positive is actually sick.

Or: A person is tested positive. Probability that she is really ill.

- $P(T \mid S)$: Probability that a sick person is tested positive.

  Or: A person is sick. Probability that she is tested positive.

- $P(\overline{T} \mid \overline{S})$: Probability that a healthy person is tested negative.

  Or: A person is healthy. Probability that she is also tested negative.

b)
- In a third world country 1 % of the people suffer from a certain infectious disease:
$$P(S) = 0.01$$

- A test correctly indicates the disease in those actually ill with a probability of 98 %:
$$P(T \mid S) = 0.98$$

- Unfortunately, the test also indicates that 3 % of healthy individuals are ill:
$$P(T \mid \overline{S}) = 0.03$$

c) It follows
$$P(\overline{S}) = 1 - P(S) = 1 - 0.01 = 0.99$$

d) Sought: $P(T)$

$$\begin{aligned} P(T) &= P(T \mid S) \cdot P(S) + P(T \mid \overline{S}) \cdot P(\overline{S}) \\ &= P(T \mid S) \cdot P(S) + P(T \mid \overline{S}) \cdot (1 - P(S)) \\ &= 0.98 \cdot 0.01 + 0.03 \cdot 0.99 \\ &= 0.0395 \end{aligned}$$

This means that 3.95 % of the tested persons are tested positive.

e) Wanted: $P(S \mid T)$

$$\begin{aligned} P(S \mid T) &= \frac{P(T \mid S) \cdot P(S)}{P(T)} \\ &= \frac{0.98 \cdot 0.01}{0.0395} \\ &= 0.2481 \end{aligned}$$

This means that only about 25 % of all those who test positive are effectively also ill.

f) Wanted: $P(\overline{S} \mid \overline{T})$:

$$\begin{aligned} P(\overline{S} \mid \overline{T}) &= \frac{P(\overline{T} \mid \overline{S}) \cdot P(\overline{S})}{P(\overline{T})} \\ &= \frac{(1 - P(T \mid \overline{S})) \cdot P(\overline{S})}{P(\overline{T})} \\ &= \frac{(1 - 0.03) \cdot 0.99}{1 - 0.0395} \\ &= 0.999792 \end{aligned}$$

This means that if the test is negative, we are very sure that we are healthy.

While a positive test does not say very much about having the disease, a negative test says tells very much.

Solution 9.3

Notation:

- D: Doped
- T: Tested positive

The following applies from the task definition

$$P(D) = 0.2, \qquad P(T \mid D) = 0.99, \qquad P(T \mid \overline{D}) = 0.05$$

a) Sought: $P(T)$.

We use the law of total probability:

$$\begin{aligned} P(T) &= P(T \mid D) \cdot P(D) + P(T \mid \overline{D}) \cdot P(\overline{D}) \\ &= P(T \mid D) \cdot P(D) + P(T \mid \overline{D}) \cdot (1 - P(D)) \\ &= 0.99 \cdot 0.2 + 0.05 \cdot 0.8 \\ &= 0.238 \end{aligned}$$

This means that 23.8 % of the tested will test positive.

b) Wanted: $P(\overline{T} \mid D)$

$$P(\overline{T} \mid D) = 1 - P(T \mid D) = 1 - 0.99 = 0.01$$

c) Sought: $P(D \mid \overline{T})$

We use the theorem of Bayes:

$$\begin{aligned} P(D \mid \overline{T}) &= \frac{P(\overline{T} \mid D) \cdot P(D)}{P(\overline{T})} \\ &= \frac{0.01 \cdot 0.2}{1 - 0.238} \\ &= 0.00262 \end{aligned}$$

Only 0.262 % of all negatively tested athletes are also doped. So a negative test is quite significant.

Solution 9.4

a) $P(W)$: Probability that a randomly selected employee is telling the truth.
$P(+ \mid L)$: Probability that for a person who lied, the polygraph indicates so.

b) Sought: $P(L \mid +)$

It follows

$$\begin{aligned} P(L \mid +) &= \frac{P(+ \mid L)P(L)}{P(+ \mid L)P(L) + P(+ \mid W)P(W)} \\ &= \frac{P(+ \mid L)(1 - P(W))}{P(+ \mid L)(1 - P(W)) + (1 - P(- \mid W))P(W)} \\ &= \frac{0.88 \cdot (1 - 0.99)}{0.88 \cdot (1 - 0.99) + (1 - 0.86) \cdot 0.99} \\ &= 0.0597 \end{aligned}$$

c) So, if the test indicates that the employee lied, then he is only a real liar with a probability of 6 %. Or in 94 % of positive tests, the employees, who allegedly lied, are telling the truth.

Thus, the test shows a very high percentage of false results and is therefore worthless.

This is also the reason why lie detector tests are not allowed in court.

Solution 9.5

Let D denote that the baby has Down syndrome and $+$ that the test is positive. Given are the following probabilities:

$$P(D) = 0.01. \qquad P(+ \mid D) = 0.9. \qquad P(+ \mid \overline{D}) = 0.01$$

Applying Bayes' theorem and the law of total probability, it follows

$$\begin{aligned} P(D \mid +) &= \frac{P(+ \mid D) \cdot P(D)}{P(+ \mid D) \cdot P(D) + P(+ \mid \overline{D}) \cdot P(\overline{D})} \\ &= \frac{0.9 \cdot 0.01}{0.9 \cdot 0.01 + 0.01 \cdot 0.99} \\ &= 0.4761905 \end{aligned}$$

If the test result is positive, there is a 48 % probability that the baby has Down syndrome.

Solution 9.6

Notation:

F : Event that fire breaks out
A : Event that the alarm goes off

The probability that a fire will break out is

$$P(F) = \frac{1}{365}$$

The probability that the alarm will go off, given a fire breaks out, is

$$P(A \mid F) = 0.95$$

The probability that there is an alarm, given that no fire has broken out, is

$$P(A \mid \overline{F}) = 0.01$$

a) The probability that a fire has broken out, given there was an alarm, is

$$P(F \mid A) = \frac{P(A \mid F) \cdot P(F)}{P(A)}$$

The probability of an alarm is expressed by the law of total probability:

$$P(A) = P(A \mid F) \cdot P(F) + P(A \mid \overline{F}) \cdot P(\overline{F})$$

So:

$$\begin{aligned} P(F \mid A) &= \frac{P(A \mid F) \cdot P(F)}{P(A \mid F) \cdot P(F) + P(A \mid \overline{F}) \cdot P(\overline{F})} \\ &= \frac{0.95 \cdot \frac{1}{365}}{0.95 \cdot \frac{1}{365} + 0.01 \cdot (1 - \frac{1}{365})} \\ &= 0.207 \end{aligned}$$

In only 1 in 5 cases of alarm there is actually a fire.

b) The probability that no fire has broken out, if there was no alarm, is

$$P(\overline{F} \mid \overline{A})$$

We now apply the Bayes Theorem:

$$P(\overline{F} \mid \overline{A}) = \frac{P(\overline{A} \mid \overline{F}) \cdot P(\overline{F})}{P(\overline{A})}$$

The three probabilities on the right hand side of the equation are unknown, but we can calculate these from the known ones:

- For $P(\overline{A} \mid \overline{F})$ it follows

$$\begin{aligned} P(\overline{A} \mid \overline{F}) &= 1 - P(A \mid \overline{F}) \\ &= 1 - 0.01 \\ &= 0.99 \end{aligned}$$

- For $P(\overline{F})$ it follows

$$\begin{aligned} P(\overline{F}) &= 1 - P(F) \\ &= 1 - \frac{1}{365} \\ &= \frac{364}{365} \end{aligned}$$

- For $P(\overline{A})$ it follows

$$P(\overline{A}) = 1 - P(A)$$

We can calculate the probability $P(A)$ with the law of total probability:

$$\begin{aligned} P(A) &= P(A \mid F)P(F) + P(A \mid \overline{F})P(\overline{F}) \\ &= 0.95 \cdot \tfrac{1}{365} + 0.01 \cdot \tfrac{364}{365} \\ &= 0.01257534 \end{aligned}$$

So we find

$$\begin{aligned} P(\overline{F} \mid \overline{A}) &= \frac{P(\overline{A} \mid \overline{F}) \cdot P(\overline{F})}{P(\overline{A})} \\ &= \frac{\big(1 - P(A \mid \overline{F})\big) \cdot P(\overline{F})}{1 - P(A)} \\ &= \frac{(1 - 0.01) \cdot \frac{364}{365}}{1 - (0.95 \cdot \frac{1}{365} + 0.01 \cdot (1 - \frac{1}{365}))} \\ &= 0.9998613 \end{aligned}$$

That means, if there is no alarm, we are pretty sure that there is no fire.

Finally

$$P(F \mid \overline{A}) = 1 - P(\overline{F} \mid \overline{A}) = 1 - 0.999 = 0.0001387$$

Solution 9.7

a) Event A: The new customer is an unlucky one;

Event B: The new customer has an accident within one year

Known: $P(B \mid A) = 0.4$, $P(B \mid \overline{A}) = 0.2$, $P(A) = 0.3$, $P(\overline{A}) = 0.7$

$$P(B) = P(B \mid A) \cdot P(A) + P(B \mid \overline{A}) \cdot P(\overline{A}) = 0.4 \cdot 0.3 + 0.2 \cdot 0.7 = 0.26$$

b) We are looking for $P(A \mid B)$

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B \mid A) \cdot P(A)}{P(B)} = \frac{0.3 \cdot 0.4}{0.26} = 0.4615$$