# Classical and Bayesian Statistics

## Problems 8

### Problem 8.1

We want to perform a multiple linear regression for Auto.

First remove the variable `name`, as it is qualitative (model of the cars).

```
head(Auto)

  mpg cylinders displacement horsepower weight acceleration year origin                      name
1  18         8          307        130   3504         12.0   70      1 chevrolet chevelle malibu
2  15         8          350        165   3693         11.5   70      1         buick skylark 320
3  18         8          318        150   3436         11.0   70      1        plymouth satellite
4  16         8          304        150   3433         12.0   70      1             amc rebel sst
5  17         8          302        140   3449         10.5   70      1               ford torino
6  15         8          429        198   4341         10.0   70      1          ford galaxie 500

Auto_1 <- within(Auto, rm(name))

head(Auto_1)

  mpg cylinders displacement horsepower weight acceleration year origin
1  18         8          307        130   3504         12.0   70      1
2  15         8          350        165   3693         11.5   70      1
3  18         8          318        150   3436         11.0   70      1
4  16         8          304        150   3433         12.0   70      1
5  17         8          302        140   3449         10.5   70      1
6  15         8          429        198   4341         10.0   70      1
```

(∗∗) a) Produce with `pairs` scatterplot containing all variables of the data set.

(∗∗) b) Calculate the correlation matrix between the variables with `cor()`.

Interpret the values for `horsepower` and `displacement` using the scatter plot above.

c) We use `lm()` to perform a multiple regression with the target variable `mpg` and all other variables (except `name`) as predictors. Use again to interpret the output of the `summary()` command.

i) Is there a relationship between the predictors and the response variable? Justify this with the $p$ value to the $F$ value.

(∗∗) ii) Which predictors seem to have a statistically significant influence on the target variable?

(∗∗) iii) What does the coefficient for `year` indicate?

(∗∗) d) Examine the model from c) still for interaction effects.

## Problem 8.2

We investigate further the data set Boston.

In order to fit a multiple linear regression model using least squares, we again use the lm() function. The syntax `lm(y ~ x1 + x2 + x3)` is used to fit a model with three predictors, `x1` , `x2` , and `x3` . The `summary()` function now outputs the regression coefficients for all the predictors.

(∗∗) a) Fit a multiple linear regression model with response variable `medv` and predictors `lstat` and `age`.

Define the model and interpret all values in the `summary()` output which we discussed in class (coefficients, its $P$ values, $R^2$ value, $P$ value of the $F$-statistics).

(∗∗) b) The Boston data set contains 13 variables, and so it would be cumbersome to have to type all of these in order to perform a regression using all of the predictors. Instead, we can use the following short-hand `lm(medv ~ ., data = Boston)`.

In the `summary()` output interpret the coefficient of `age` and the corresponding $p$-value compare this with the output in a) and explain the difference.

(∗∗) c) The $R^2$ value is bigger than the one calculated in a). Explain.

(∗∗) d) It is easy to include interaction terms in a linear model using the `lm()` function. The syntax `lstat:black` tells `R` to include an interaction term between `lstat` and `black`.

The syntax `lstat * age` simultaneously includes `lstat`, `age`, and the interaction term `lstat × age` as predictors; it is a shorthand for `lstat + age + lstat:age`.

Again, discuss all the values in the `summary()` of `lstat*age` as in a).

## Problem 8.3

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?

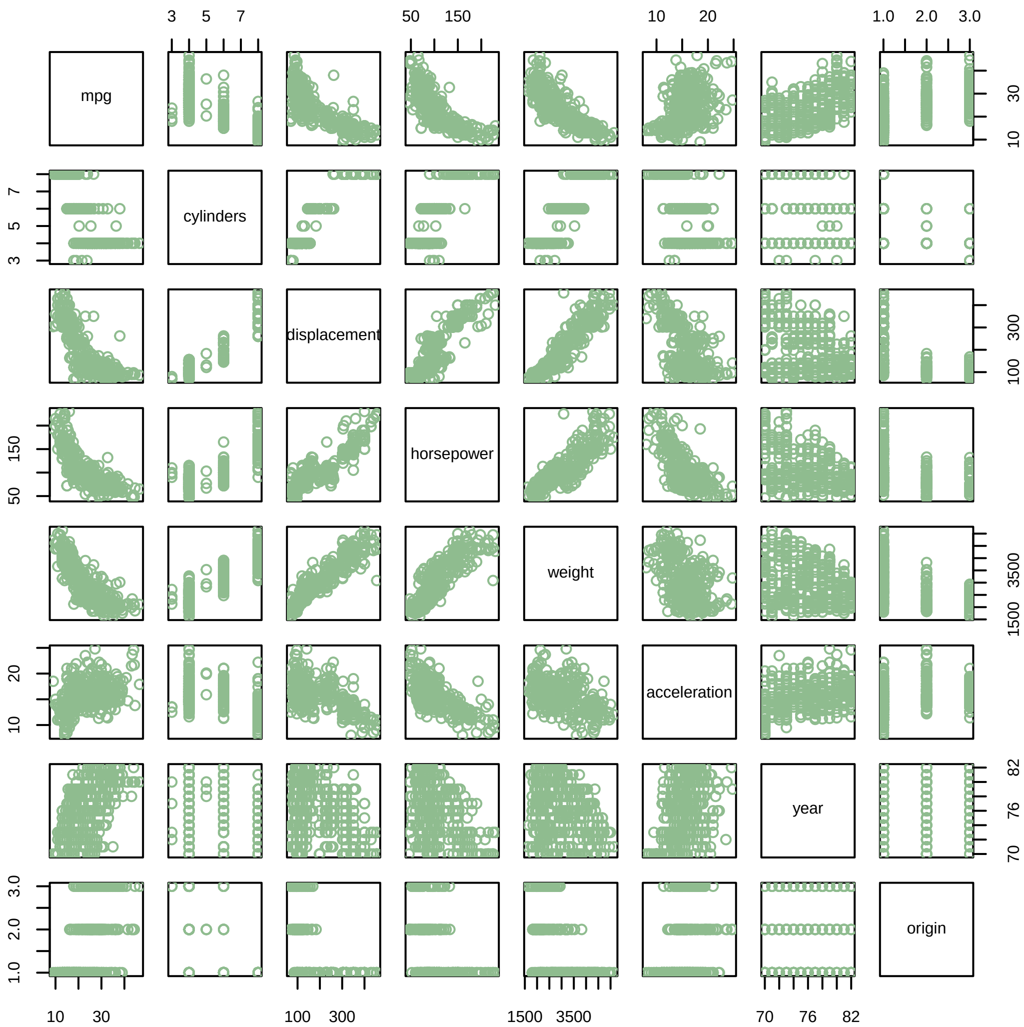
# Classical and Bayesian Statistic

## Sample solution for Problems 8

### Solution 8.1

a) Scatter diagram:

```
pairs(Auto_1, col="darkseagreen")
```




b) Correlation matrix

```
round(cor(Auto_1),3)

               mpg cylinders displacement horsepower weight acceleration   year origin
mpg          1.000    -0.778       -0.805     -0.778 -0.832        0.423  0.581  0.565
cylinders   -0.778     1.000        0.951      0.843  0.898       -0.505 -0.346 -0.569
displacement -0.805    0.951        1.000      0.897  0.933       -0.544 -0.370 -0.615
horsepower  -0.778     0.843        0.897      1.000  0.865       -0.689 -0.416 -0.455
```

```
weight        -0.832    0.898      0.933     0.865  1.000      -0.417 -0.309 -0.585
acceleration   0.423   -0.505     -0.544    -0.689 -0.417       1.000  0.290  0.213
year           0.581   -0.346     -0.370    -0.416 -0.309       0.290  1.000  0.182
origin         0.565   -0.569     -0.615    -0.455 -0.585       0.213  0.182  1.000
```

c) Output

```
fit <- lm(mpg ~ . , data=Auto_1)
summary(fit)

Call:
lm(formula = mpg ~ ., data = Auto_1)

Residuals:
    Min      1Q  Median      3Q     Max
-9.5903 -2.1565 -0.1169  1.8690 13.0604

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
cylinders     -0.493376   0.323282  -1.526  0.12780
displacement   0.019896   0.007515   2.647  0.00844 **
horsepower    -0.016951   0.013787  -1.230  0.21963
weight        -0.006474   0.000652  -9.929  < 2e-16 ***
acceleration   0.080576   0.098845   0.815  0.41548
year           0.750773   0.050973  14.729  < 2e-16 ***
origin         1.426141   0.278136   5.127 4.67e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.328 on 384 degrees of freedom
Multiple R-squared:  0.8215,	Adjusted R-squared:  0.8182
F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

i) The $p$-value to the corresponding $F$ value is practically 0 and thus there is a statistically significant relationship between the response variable and the predictors.

ii) These are the coefficients with $**$ or $***$ (`displacement`, `weight`, `year` and `origin`)

iii) The coefficient for `year` is positive. This means that with younger cars you can get further per gallon of petrol. The newer cars are more fuel efficient in general.

d) Output:

```
fit <- lm(mpg ~ weight * year, data=Auto)
summary(fit)

Call:
lm(formula = mpg ~ weight * year, data = Auto)

Residuals:
    Min      1Q  Median      3Q     Max
-8.0397 -1.9956 -0.0983  1.6525 12.9896

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.105e+02  1.295e+01  -8.531 3.30e-16 ***
weight       2.755e-02  4.413e-03   6.242 1.14e-09 ***
year         2.040e+00  1.718e-01  11.876  < 2e-16 ***
weight:year -4.579e-04  5.907e-05  -7.752 8.02e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.193 on 388 degrees of freedom
Multiple R-squared:  0.8339,   Adjusted R-squared:  0.8326
F-statistic: 649.3 on 3 and 388 DF,  p-value: < 2.2e-16
```

The $p$ value of the interaction term is of the order of $8 \cdot 10^{-14}$, i.e. very close to 0, so the null hypothesis that there is no interaction is rejected.

This can be explained by the fact that the weight has become smaller and smaller with younger cars.

## Solution 8.2

a) Model:

$$\text{medv} = \beta_0 + \beta_1 \cdot \text{lstat} + \beta_2 \cdot \text{age}$$

```
library(MASS)

fit <- lm(medv ~ lstat + age, data = Boston)
summary(fit)

Call:
lm(formula = medv ~ lstat + age, data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-15.981  -3.978  -1.283   1.968  23.158

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.22276    0.73085  45.458  < 2e-16 ***
```

```
lstat       -1.03207    0.04819 -21.416  < 2e-16 ***
age          0.03454    0.01223   2.826  0.00491 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.173 on 503 degrees of freedom
Multiple R-squared:  0.5513,  Adjusted R-squared:  0.5495
F-statistic:   309 on 2 and 503 DF,  p-value: < 2.2e-16
```

The estimates are

$$\widehat{\beta}_0 = 33.22; \qquad \widehat{\beta}_1 = -1.03; \qquad \widehat{\beta}_2 = 0 - 03$$

We get for the model

$$\text{medv} = 33.22 - 1.03 \cdot \text{lstat} + 0.03 \cdot \text{age}$$

Interpretation of the estimates:

- $\widehat{\beta}_0 = 33.22$

  In neighborhoods where there is no population of lower status and no units build before 1940, the medium value of houses is \$ 33 220.

- $\widehat{\beta}_1 = -1.03$

  For each additional percent of population of lower status, the medium value decreases by \$ 1030.

- $\widehat{\beta}_2 = 0.03$

  For each additional percent of units build before 1949, the medium value increases by \$ 30.

- All $p$-values are significant (below the significance level of 5 %), so all estimates individually contribute significantly to the model.

- The $R^2$ value is 0.5513, therefore about 55 % of the variation is explained by the model.

- The $p$-value of the $F$ value is below the significance level and therefore significant. The null hypothesis $H_0$

  $$\beta_1 = \beta_2 = 0$$

  is rejected. One of $\beta$'s is significantly different from 0. At least one variables contributes significantly to the model.

b)
```
fit <- lm(medv ~ ., data = Boston)
summary(fit)

Call:
lm(formula = medv ~ ., data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-15.595  -2.730  -0.518   1.777  26.199

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
crim        -1.080e-01  3.286e-02  -3.287 0.001087 **
zn           4.642e-02  1.373e-02   3.382 0.000778 ***
indus        2.056e-02  6.150e-02   0.334 0.738288
chas         2.687e+00  8.616e-01   3.118 0.001925 **
nox         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
rm           3.810e+00  4.179e-01   9.116  < 2e-16 ***
age          6.922e-04  1.321e-02   0.052 0.958229
dis         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
rad          3.060e-01  6.635e-02   4.613 5.07e-06 ***
tax         -1.233e-02  3.760e-03  -3.280 0.001112 **
ptratio     -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
black        9.312e-03  2.686e-03   3.467 0.000573 ***
lstat       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.745 on 492 degrees of freedom
Multiple R-squared:  0.7406,    Adjusted R-squared:  0.7338
F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

The $p$-value is almost 1, so not significant at all. But in a), the $p$-value is 0.005, which is significant. That means that the variable `age` must correlate strongly with other variables (see d)).

c) The more variables you have the bigger the $R^2$ value. That means that the $R^2$ is not a good indicator to compare different models.

d) Model:

$$\text{medv} = \beta_0 + \beta_1 \cdot \text{lstat} + \beta_2 \cdot \text{age} + \beta_{12} \cdot \text{lstat*age}$$

Remark: $*$ in `lstat*age` does *not* signify multiplication, it just means interaction.

```
fit <- lm(medv ~ lstat * age, data = Boston)
summary(fit)

Call:
```

```
lm(formula = medv ~ lstat * age, data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-15.806  -4.045  -1.333   2.085  27.552

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.0885359  1.4698355  24.553  < 2e-16 ***
lstat       -1.3921168  0.1674555  -8.313 8.78e-16 ***
age         -0.0007209  0.0198792  -0.036   0.9711
lstat:age    0.0041560  0.0018518   2.244   0.0252 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.149 on 502 degrees of freedom
Multiple R-squared:  0.5557,  Adjusted R-squared:  0.5531
F-statistic: 209.3 on 3 and 502 DF,  p-value: < 2.2e-16
```

The estimates are

$$\widehat{\beta}_0 = 36.10; \qquad \widehat{\beta}_1 = -1.39; \qquad \widehat{\beta}_2 = -0.0007; \qquad \widehat{\beta}_{12} = 0.004$$

We get for the model

$$\text{medv} = 36.10 - 1.39 \cdot \text{lstat} - 0.00072 \cdot \text{age} + 0.0041 \cdot \text{lstat*age}$$

Interpretation of the estimates:

- $\widehat{\beta}_0 = 36.10$

  In neighborhoods where there is no population of lower status and no units build before 1940, the medium value of houses is \$ 36 100.

- $\widehat{\beta}_1 = -1.39$

  For each additional percent of population of lower status, the medium value decreases by \$ 1930.

- $\widehat{\beta}_2 = -0.00072$

  For each additional percent of units build before 1949, the medium value decreases by \$ 0.27.

  As you can imagine, this value is not significant, as you can see from the output.

- $\widehat{\beta}_{12} = 0.004$

  This coefficient is somewhat difficult to interpret and we didn't do it in class.

- Not all $p$-values are significant (below the significance level of 5 %) anymore.

  The $p$-value for `age` is 0.97, so this not significance anymore, whereas without interaction it was. What is the reason for this?

  The $p$-value of the interaction term is 0.0252 which is below the significance level of 5 %. The null hypothesis $H_0$, that there is no interaction, is rejected. There is statistically significant interaction.

  Now, let's take a look at the correlation coefficient of the two explanatory variables `lstat` and `age`.

  ```
  cor(Boston["lstat"],Boston["age"])
              age
  lstat 0.6023385
  ```

  This value is quite high. An explanation *could* be that in the poorer neighborhoods, people didn't have the money to build new houses, so there are more houses built before 1940.

- The $R^2$ value is 0.56, therefore about 56 % of the variation is explained by the model.

- The $p$-value of the $F$ value is below the significance level and therefore significant. The null hypothesis $H_0$

  $$\beta_1 = \beta_2 = \beta_{12} = 0$$

  is rejected. One of $\beta$'s is significantly different from 0. At least one variables contributes significantly to the model.

## Solution 8.3

a) Data set:

```
library(ISLR)

head(Carseats)

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