MAS programmes - Statistical Analysis (Autumn Term)

Exercises 3 - SOLUTIONS

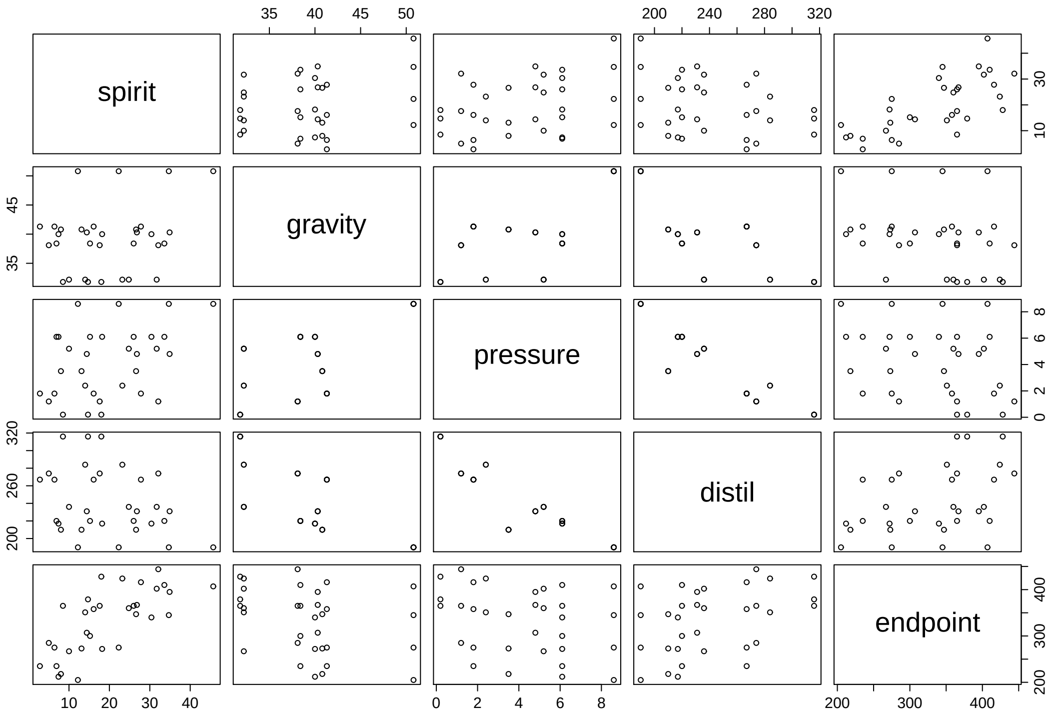
1. You may need to reload the data into R as the data frame `oil`. [The following code assumes you have the file `oil.txt` in your *working directory*].

```
> oil <- read.table("oil.txt")
> names(oil) <- c("spirit", "gravity", "pressure", "distil", "endpoint")
```

(a) Using the suggested commands, we find the following:

```
> pairs(oil)

> cor(oil)
             spirit    gravity   pressure     distil   endpoint
spirit    1.0000000  0.2463260  0.3840706 -0.3150243  0.7115262
gravity   0.2463260  1.0000000  0.6205867 -0.7001539 -0.3216782
pressure  0.3840706  0.6205867  1.0000000 -0.9062248 -0.2979843
distil   -0.3150243 -0.7001539 -0.9062248  1.0000000  0.4122466
endpoint  0.7115262 -0.3216782 -0.2979843  0.4122466  1.0000000
```



The top row of plots (in the *scatterplot matrix*) shows the relationship between the response variable `spirit` and each of the explanatory (regressor) variables. The strongest relationship appears to be with `endpoint` so that we would expect to see this variable in a good linear regression model, plus possibly one from `pressure` and `distil`. [Since there is a very strong correlation between these variables (-0.91, the strongest between all the explanatory variables), we might not expect to see both in the same model].

Individual regressions of the response on each of the explanatory variables (not shown) suggest `endpoint` to be strongly significant and `distil` and `pressure` significant; `gravity` does not seem to have a linear relationship with `spirit`. Also because `pressure` and `distil` are highly collinear it comes as no surprise that the final model includes `endpoint` and `distil`.

(b) The confidence and predictive intervals can be found in R using the following code.

```
> oil2.lm <- lm(spirit ~ distil + endpoint, data = oil)
> x <- data.frame(distil = 200, endpoint = 400)
predict.result.c <-predict(oil2.lm,x,se.fit=T, interval = c("confidence"))
predict.result.c
$fit
       fit      lwr      upr
1 38.92724 37.00562 40.84885

$se.fit
[1] 0.9395607

$df
[1] 29

$residual.scale
[1] 2.425522

predict.result.p <-predict(oil2.lm,x,se.fit=T, interval = c("prediction"))
predict.result.p
$fit
       fit      lwr      upr
1 38.92724 33.60731 44.24717

$se.fit
[1] 0.9395607

$df
[1] 29

$residual.scale
[1] 2.425522
```

2. If necessary, the following code can be run in R to obtain a data frame `sugar` which contains the variables `price`, `consump` and `lconsump`. [Again this assumes that you have the file `sugar.txt` in your *working directory*].

```
> sugar <- read.table("sugar.txt", header = T)

> sugar$lconsump <- log(sugar$consump)
```

(a) The linear model object is recreated in R as follows:

```
> sugar.lm <- lm(lconsump ~ price, data = sugar)
```

We use the `anova` function to find the sums of squares.

```
> anova(sugar.lm)
Analysis of Variance Table

Response: lconsump
          Df Sum Sq Mean Sq F value    Pr(>F)
price      1 4.4940  4.4940  584.34 < 2.2e-16 ***
Residuals 53 0.4076  0.0077
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

which (for the simple linear regression) has the required form directly.

ANOVA Table

Source of Variation	d.f.	Sum of Squares (SS)	Mean Square (MS)
Regression	1	4.4940	4.4940
Residual	53	0.4076	0.0077
Total	54	4.9016	

The F-statistic is hence $\dfrac{4.4940}{0.0077} = 584.34$, with p-value

```
> pf(584.39, 1, 53, lower.tail = F)
[1] 2.706843e-30

> summary(sugar.lm)
Call:
lm(formula = lconsump ~ price, data = sugar)

Residuals:
      Min        1Q    Median        3Q       Max
-0.206114 -0.068467  0.004681  0.059175  0.235160

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.081099   0.038202  133.00   <2e-16 ***
price       -0.138536   0.005731  -24.17   <2e-16 ***
---
```

```
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.0877 on 53 degrees of freedom
Multiple R-squared:  0.9168,    Adjusted R-squared:  0.9153
F-statistic: 584.3 on 1 and 53 DF,  p-value: < 2.2e-16
```

This test is reported in the final line of the `summary` output given above. Also since we have only a single explanatory variable the ANOVA is exactly equivalent to the t-test for the slope parameter `price` given above ($(-24.17)^2 = 584.3$).

(b) Using the R command noted in (a), the 95% predictive interval is found as follows:

```
> pred <-predict(sugar.lm,data.frame(price = 6),se.fit=T, interval = c("prediction"))
> pred
$fit
       fit      lwr      upr
1 4.249886 4.072354 4.427417

$se.fit
[1] 0.01198314

$df
[1] 53

$residual.scale
[1] 0.08769635
```

Recalling that the model is for `lconsump` which is the *log* of consumption, the predicted value for `consump` (consumption) is

$$\exp(4.249886) = 70.097 \quad \text{(pounds per capita)}$$

with 95% predictive interval $(\exp(4.072354), \exp(4.427417)) = (58.69, 83.71)$.

```
> exp(pred$fit)
       fit    lwr      upr
1 70.09739 58.695 83.71488
```