Homework 2 Solutions

STAT 3032

PROBLEMS: 2.15, 2.20, 3.1, 3.3(1,3)

CHAPTER 2

2.15

1.

```r
library("alr4")
```

```
## Loading required package: car

## Loading required package: effects

##
## Attaching package: 'effects'

## The following object is masked from 'package:car':
##
##     Prestige
```

```r
data = wblake

fitted = lm(Length ~ Age, data = data)
predict(fitted, data.frame(Age = c(2, 4, 6)), interval = "confidence")
```

```
##        fit      lwr      upr
## 1 126.1749 122.1643 130.1856
## 2 186.8227 184.1217 189.5237
## 3 247.4705 243.8481 251.0929
```

2. The default value for prediction intervals is 95% so we don't need to specify the level that we want:

```r
predict(fitted, data.frame(Age = c(9)), interval = "confidence")
```

```
##        fit      lwr      upr
## 1 338.4422 331.4231 345.4612
```

This is an extrapolation outside the range of the data, as there were no fish older than 8 years in the sample. Therefore, we do not know if the straight-line mean function applies at age 9.

2.20

1. In the summary, we have provided prediction intervals and scatter plots which are not required.

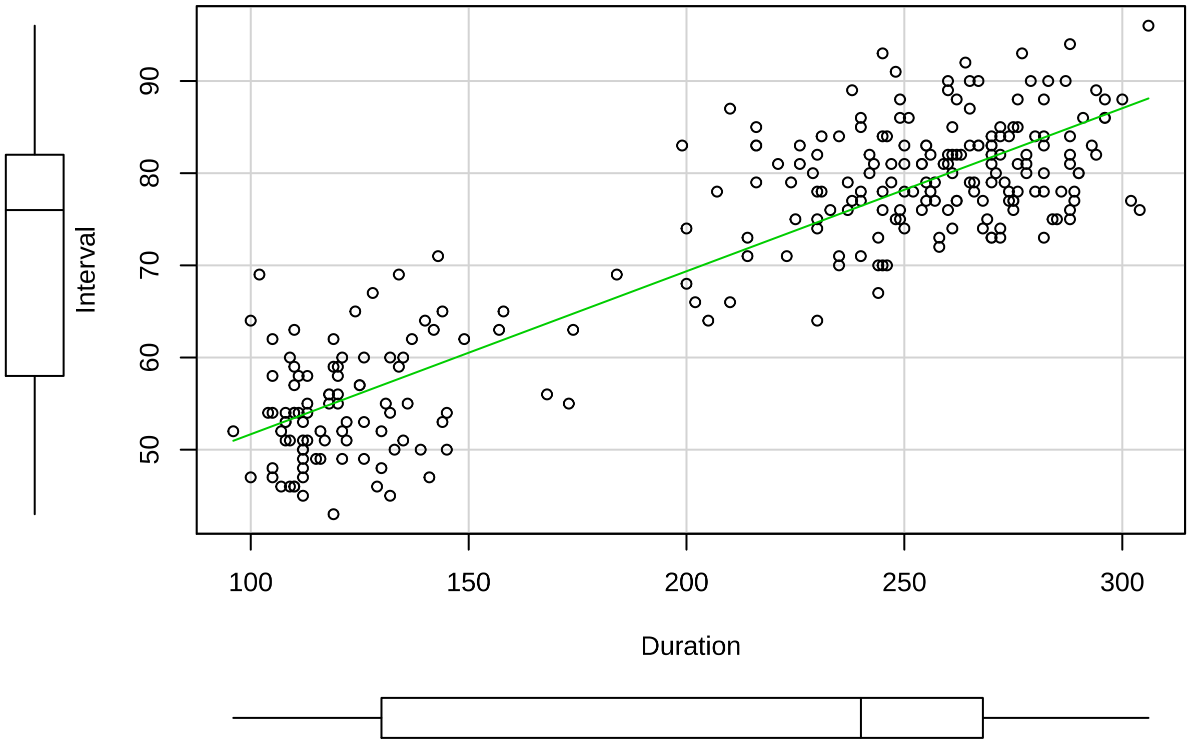
```
data = oldfaith

fitted = lm(Interval ~ Duration, data = data)
summary(fitted)
```

```
##
## Call:
## lm(formula = Interval ~ Duration, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.3337  -4.5250   0.0612   3.7683  16.9722
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.987808   1.181217   28.77   <2e-16 ***
## Duration     0.176863   0.005352   33.05   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.004 on 268 degrees of freedom
## Multiple R-squared:  0.8029, Adjusted R-squared:  0.8022
## F-statistic:  1092 on 1 and 268 DF,  p-value: < 2.2e-16
```

```
predict(fitted, data.frame(Duration = c(130, 240, 300)), interval = "prediction")
```

```
##        fit      lwr      upr
## 1 56.97999 45.10812 68.85185
## 2 76.43491 64.58867 88.28115
## 3 87.04669 75.16668 98.92669
```

```
scatterplot(Interval ~ Duration, data = data, smooth = FALSE)
```



2.

```
predict(fitted, data.frame(Duration = c(250)), interval = "prediction")
```

```
##        fit      lwr      upr
## 1 78.20354 66.35401 90.05307
```

3. The `predict` method in `R` can be used for this:

```
predict(fitted, data.frame(Duration = c(250)), interval = "confidence", level = 0.8)
```

```
##        fit      lwr    upr
## 1 78.20354 77.65908 78.748
```

The 80% interval cuts off 10% of the probability below `lwr` and 10% above `upr`. Consequently, the value `upr` estimate the 0.90 quantile.

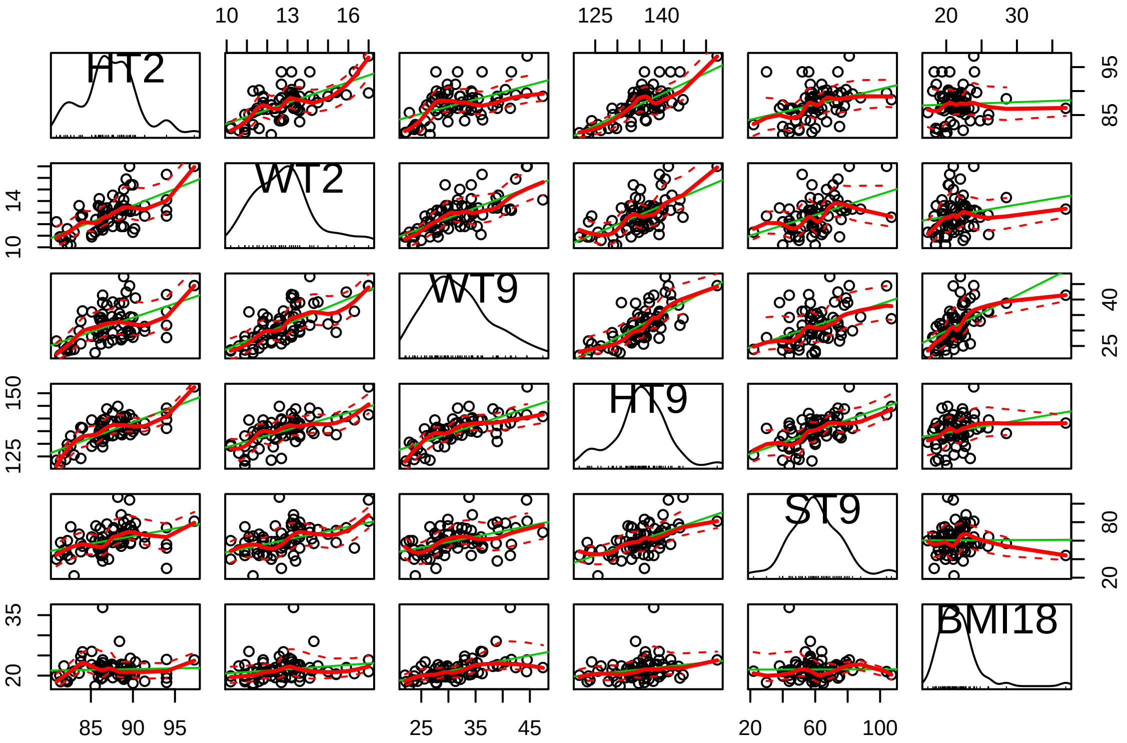
3.1

Use any point-identifying software (such as `scatterplot` in the `car` package in `R` with the argument `id.n` set to about 10) you can discover all the odd points correspond to countries in Africa, apart possibly from Nauru, which is an island nation in the South Pacific.

3.3

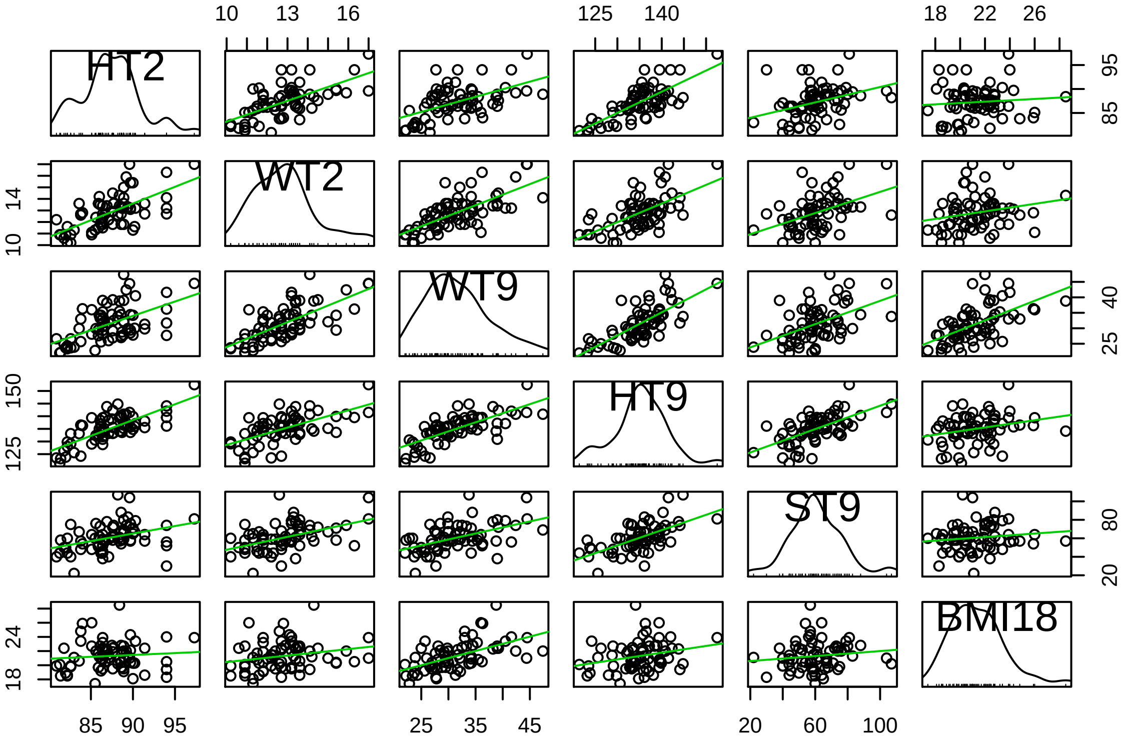
1. The scatterplot matrix below is enhanced by adding `OLS` line and a smoother.

```
data = BGSgirls

scatterplotMatrix(~ HT2 + WT2 + WT9 + HT9 + ST9 + BMI18, data = BGSgirls)
```



In virtually all of the frames that don't include BMI18, the regressions have linear mean functions, which means that the OLS fit and the smoother agree. This is the ideal case of multiple linear regression. The last row of the scatterplot matrix has the summary plots for the regression of BMI18 on each of the predictors individually. Examining this graph is difficult because resolution is lost due to a girl with BMI exess of 35 (values above 30 indicate obesity), and so getting a useful visual impression requires removing this point and re-plotting (not required):

```
scatterplotMatrix(~ HT2 + WT2 + WT9 + HT9 + ST9 + BMI18, data = BGSgirls, smooth = FALSE, subset = BMI18
```



We now see that `WT9` is most closely related to `BMI18`, and we can't really judge the role of the other predictors in a multiple regression from this plot.

Sample correlation matrix for all the girls:

```
print(cor(BGSgirls[, c("HT2", "WT2", "HT9", "WT9", "ST9", "BMI18")]), digits = 3)
```

```
##          HT2   WT2   HT9   WT9    ST9  BMI18
## HT2   1.0000 0.645 0.738 0.523 0.3617 0.0426
## WT2   0.6445 1.000 0.607 0.693 0.4516 0.1909
## HT9   0.7384 0.607 1.000 0.728 0.6034 0.2369
## WT9   0.5229 0.693 0.728 1.000 0.4530 0.5459
## ST9   0.3617 0.452 0.603 0.453 1.0000 0.0056
## BMI18 0.0426 0.191 0.237 0.546 0.0056 1.0000
```

From this scatterplot matrix we know to question the usefulness of the correlations with `BMI18`, because the one unusual point could distort the correlations. Deleting one unusual point (not required):

```
sel = BGSgirls$BMI18 < 35
print(with(BGSgirls[sel, ], cor(BMI18, cbind(HT2, WT2, HT9, WT9, ST9))), digits = 3)
```

```
##         HT2   WT2   HT9   WT9   ST9
## [1,] 0.0861 0.223 0.261 0.565 0.129
```

We see that in this particular case since the girl with large `BMI18` is near the middle of the range of the other variables it has little influence of the correlation, and so in this case the original correlation matrix would provide a sensible summary.

3.

```r
fitted = lm(BMI18 ~ HT2 + WT2 + HT9 + WT9 + ST9, data = BGSgirls)
summary(fitted)
```

```
## 
## Call:
## lm(formula = BMI18 ~ HT2 + WT2 + HT9 + WT9 + ST9, data = BGSgirls)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -5.0948 -1.2186 -0.2533  1.0090 10.4951 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 30.855335   8.781156   3.514 0.000817 ***
## HT2         -0.193997   0.130819  -1.483 0.142996    
## WT2         -0.317779   0.278736  -1.140 0.258505    
## HT9          0.008057   0.096344   0.084 0.933613    
## WT9          0.419762   0.075211   5.581  5.2e-07 ***
## ST9         -0.044416   0.022219  -1.999 0.049853 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.14 on 64 degrees of freedom
## Multiple R-squared:  0.4431, Adjusted R-squared:  0.3996 
## F-statistic: 10.19 on 5 and 64 DF,  p-value: 3.294e-07
```

The regression explains about 100 x $R^2 = 44\%$ of the variation in `BMI18`. The hypothesis tested by the t-values are that each of the $\beta_i = 0$ with the other β's arbitrary versus $\beta_i \neq 0$ with all the other β's arbitrary. For this test, only the intercept, WT9, and ST9 have t-values with p-values smaller than 0.05. We will reject all other variables.