Chapter 2 Problem 6 Faraway

Faraway, Julian J.. Linear Models with R, Second Edition (Chapman & Hall/CRC Texts in Statistical Science). CRC Press.

Bruce Campbell

25 August, 2017

Thirty samples of cheddar cheese were analyzed for their content of acetic acid, hydrogen sulfide and lactic acid. Each sample was tasted and scored by a panel of judges and the average taste score produced. Use the cheddar data to answer the following:

- (a) Fit a regression model with taste as the response and the three chemical contents as predictors. Report the values of the regression coefficients.
- (b) Compute the correlation between the fitted values and the response. Square it. Identify where this value appears in the regression output.
- (c) Fit the same regression model but without an intercept term. What is the value of R 2 reported in the output? Compute a more reasonable measure of the goodness of fit for this example.

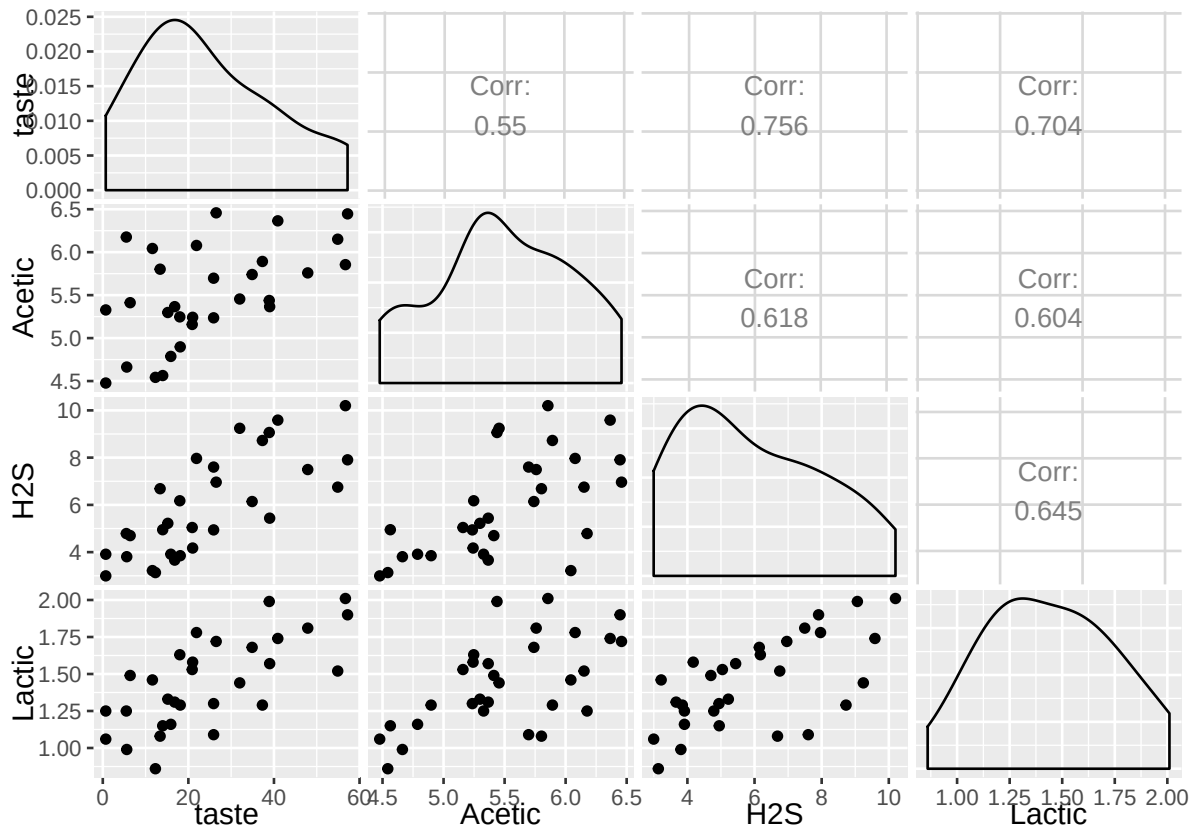
Fit the linear model $taste \sim Acetic + H2S + Lactic$

First we load and inspect the data

```
data(cheddar, package = "faraway")
head(cheddar)
```

```
##   taste Acetic   H2S Lactic
## 1  12.3  4.543 3.135   0.86
## 2  20.9  5.159 5.043   1.53
## 3  39.0  5.366 5.438   1.57
## 4  47.9  5.759 7.496   1.81
## 5   5.6  4.663 3.807   0.99
## 6  25.9  5.697 7.601   1.09
```

```
ggpairs(data = cheddar)
```



Fit the model and display the regression coefficients

```
lm.fit <- lm(taste ~ Acetic + H2S + Lactic, data = cheddar)

summary(lm.fit)
```

```
##
## Call:
## lm(formula = taste ~ Acetic + H2S + Lactic, data = cheddar)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.390  -6.612  -1.009   4.908  25.449
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -28.8768    19.7354  -1.463  0.15540
## Acetic        0.3277     4.4598   0.073  0.94198
## H2S           3.9118     1.2484   3.133  0.00425 **
## Lactic       19.6705     8.6291   2.280  0.03108 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.13 on 26 degrees of freedom
## Multiple R-squared:  0.6518, Adjusted R-squared:  0.6116
## F-statistic: 16.22 on 3 and 26 DF,  p-value: 3.81e-06
```

Compute the correlation between the fitted values and the response. Square it. Identify where this value appears in the regression output.

```
yhat <- lm.fit$fitted.values

y <- cheddar$taste

corr.fitted_response <- cor(yhat, y)

pander(data.frame(r_sq = corr.fitted_response^2), caption = "Squared Correlation between actual and pred
```

Table 1: Squared Correlation between actual and predicted

r_sq
0.6518

The value we calculated is the multiple R-squared in the regression output. Note the adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. We use the adjusted R-squared when comparing models - it increases only if the new terms improve the model more than would be expected by chance.

Diagnotic plots for this model - uncomment for display.

```
# plot(lm.fit)
```

Fit the model without an intercept term.

In R a formula has an implied intercept term. To remove this use either $y \sim x - 1$ or $y \sim 0 + x$ in the formula. For our case we will fit $taste \sim Acetic + H2S + Lactic - 1$. We'll need to account for the lack of intercept term when evaluating the quality of the fit. The default R^2 calculation in R assumes a null model with an intercept.

```
lm.fit.nointercept <- lm(taste ~ Acetic + H2S + Lactic - 1, data = cheddar)

summary(lm.fit.nointercept)
```

```
##
## Call:
## lm(formula = taste ~ Acetic + H2S + Lactic - 1, data = cheddar)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.4521  -6.5262  -0.6388   4.6811  28.4744
##
## Coefficients:
##        Estimate Std. Error t value Pr(>|t|)
## Acetic   -5.454      2.111  -2.583  0.01553 *
## H2S       4.576      1.187   3.854  0.00065 ***
## Lactic   19.127      8.801   2.173  0.03871 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.34 on 27 degrees of freedom
## Multiple R-squared:  0.8877, Adjusted R-squared:  0.8752
```

```
## F-statistic: 71.15 on 3 and 27 DF,  p-value: 6.099e-13
```

The reported R^2 in this case is 0.8877 We see the jump in calculated value that the textbook mentions. Now we'll calculate $R^2 = corr^2(y, \hat{y})$

```
yhat <- lm.fit.nointercept$fitted.values

y <- cheddar$taste

corr.fitted_response.nointercept <- cor(yhat, y)


pander(data.frame(r_sq = corr.fitted_response.nointercept^2), caption = "Squared Correlation between ac
```

Table 2: Squared Correlation between actual and predicted

r_sq
0.6244

We see this value is commensurate with the value we obtained when there was an intercept.

Diagnotic plots for this model - uncomment for display.

```
# plot(lm.fit.nointercept)
```