Faraway - Chapter 2

J. A. Kilgallen

10/21/2020

Linear Models - Chapter 2 Exercises

Question 1

The following R code fits a regression model to the data from a study on teenage gambling in Britain. Expenditure on gambling is the response variable and sex, status, income, and verbal score are predictor variables.

```
library(faraway)
```

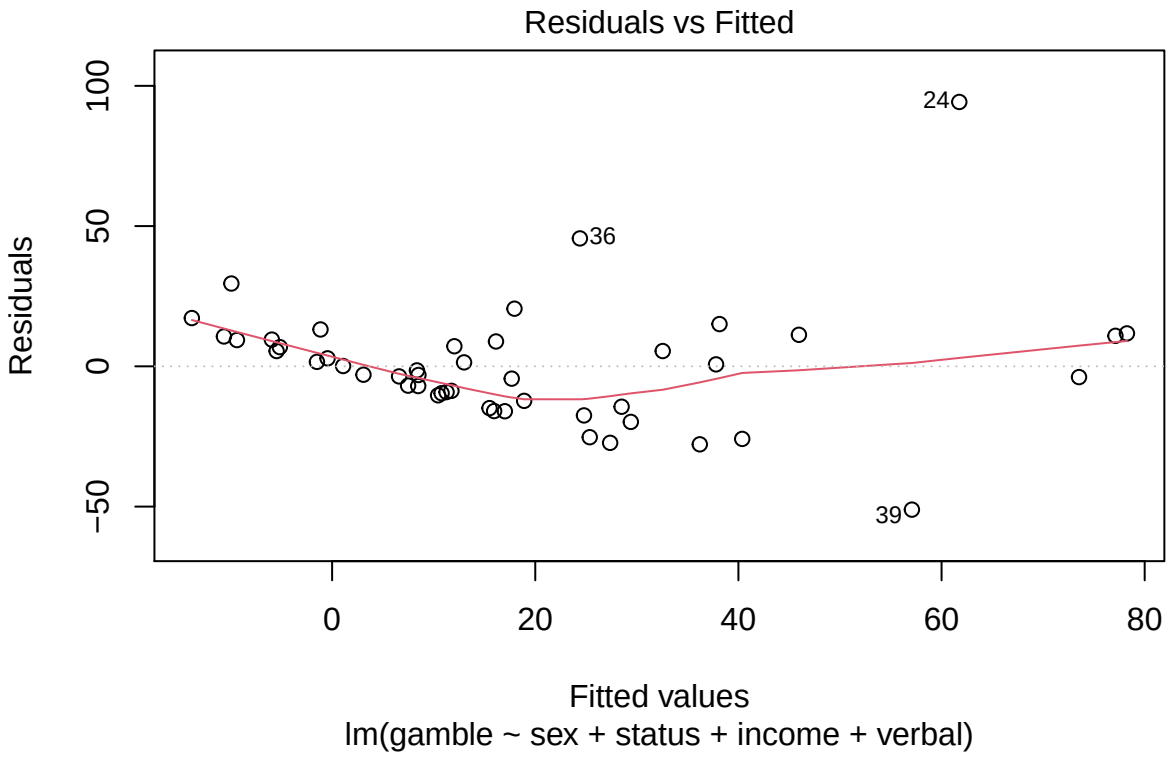
```
## Warning: package 'faraway' was built under R version 4.0.3
```

```
data(teengamb)
fit <- lm(gamble ~ sex + status + income + verbal, data=teengamb)
summary(fit)
```

```
##
## Call:
## lm(formula = gamble ~ sex + status + income + verbal, data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.082 -11.320  -1.451   9.452  94.252
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.55565   17.19680   1.312   0.1968
## sex         -22.11833    8.21111  -2.694   0.0101 *
## status        0.05223    0.28111   0.186   0.8535
## income        4.96198    1.02539   4.839 1.79e-05 ***
## verbal       -2.95949    2.17215  -1.362   0.1803
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.69 on 42 degrees of freedom
## Multiple R-squared:  0.5267, Adjusted R-squared:  0.4816
## F-statistic: 11.69 on 4 and 42 DF,  p-value: 1.815e-06
```

1. Our multiple R^2 indicates that 52.67% of variation in the response variable is explained by our predictor variables.

```
plot(fit, which =1)
```



2. As the graph above shows, the largest positive residual is case 24.

```
mean(fit$residuals)
```

```
## [1] -3.065293e-17
```

3. As the output above shows, the mean value of the residuals is $-3.065293e^{-17}$.

```
cor(fit$residuals,fit$fitted.values)
```

```
## [1] -1.070659e-16
```

4. As the output above shows the correlation between the residuals and the fitted values is $-1.070659e^{-16}$.

```
cor(fit$residuals, teengamb$income)
```

```
## [1] -7.242382e-17
```

5. As the output above shows the correlation between the residuals and the income is $-7.242382e^{-17}$.
6. In our summary output above we can see that the contribution sex makes to the model is -22.11833. So for all other predictors held constant we would expect a male to spend approximately £22.12 more than a female on gambling.

Question 2

The following R code fits a regression model to the data from a study on US wages. The response variable is weekly wages and the predictor variables are years of education and years of experience.

```r
data("uswages")
fit <- lm(wage ~ educ + exper, data= uswages)
summary(fit)
```

```
##
## Call:
## lm(formula = wage ~ educ + exper, data = uswages)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1018.2  -237.9   -50.9   149.9  7228.6
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -242.7994    50.6816  -4.791 1.78e-06 ***
## educ          51.1753     3.3419  15.313  < 2e-16 ***
## exper          9.7748     0.7506  13.023  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 427.9 on 1997 degrees of freedom
## Multiple R-squared:  0.1351, Adjusted R-squared:  0.1343
## F-statistic:   156 on 2 and 1997 DF,  p-value: < 2.2e-16
```

This model suggests that for two individuals with the same number of years of experience we would expect a worker with an extra n years of education to earn approximately $n \cdot \$51.18$ more per week.

```r
fit <- lm(log(wage) ~ educ + exper, data= uswages)
summary(fit)
```

```
##
## Call:
## lm(formula = log(wage) ~ educ + exper, data = uswages)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7533 -0.3495  0.1068  0.4381  3.5699
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.650319   0.078354   59.35   <2e-16 ***
## educ        0.090506   0.005167   17.52   <2e-16 ***
## exper       0.018079   0.001160   15.58   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6615 on 1997 degrees of freedom
## Multiple R-squared:  0.1749, Adjusted R-squared:  0.174
## F-statistic: 211.6 on 2 and 1997 DF,  p-value: < 2.2e-16
```

This model suggests that for two individuals with the same number of years of experience we would expect a worker with an n extra years of education to earn approximately $n \cdot 9.0473$ more per week.

I believe the second model to be a more natural interpretation as it puts the relationship that the predictor variables have to the response variable in more comparative terms. Additionally under the wrong circumstances the first interpretation might imply that someone with sufficiently fewer years of education earns a negative

wage, and we know that this cannot occur in reality.

Question 3

The following R code fits a regression model to some generated data, both by directly computing coefficients, and by using the R language's lm command.

```
x <- 1:20
y <- x + rnorm(20)
lm(y~x+I(x^2))
```

```
##
## Call:
## lm(formula = y ~ x + I(x^2))
##
## Coefficients:
## (Intercept)            x       I(x^2)
##    0.224557     1.028977    -0.002781
```

```
x1 <- model.matrix(~ x + I(x^2))
solve(t(x1) %*% x1) %*% t(x1) %*% y
```

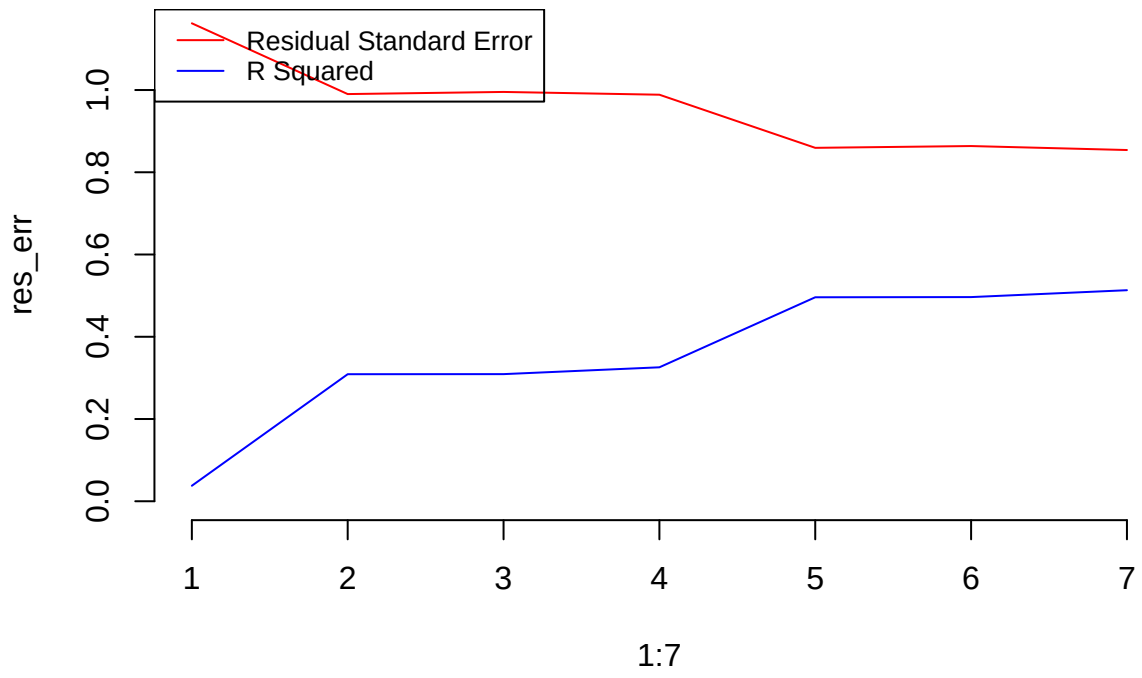
```
##                    [,1]
## (Intercept)  0.22455710
## x            1.02897682
## I(x^2)      -0.00278077
```

We can see that both methods of finding coefficients for a simple linear regression model give the same results. If we were to use a polynomial of degree 20 our direct calculation would fail as we would be attempting to estimate more coefficients than we have data points for. Which means that our matrix x1 would not be of full rank, and hence would not have a unique solution.

Question 4

The following R code fits a regression model to data used in a study of men with prostate cancer due to receive a radical prostatectomy. Our response variable is lpsa and our predictor variable is lcavol. Additional predictor variables are added to examine the effect on the residual standard error and the R^2 of the model.

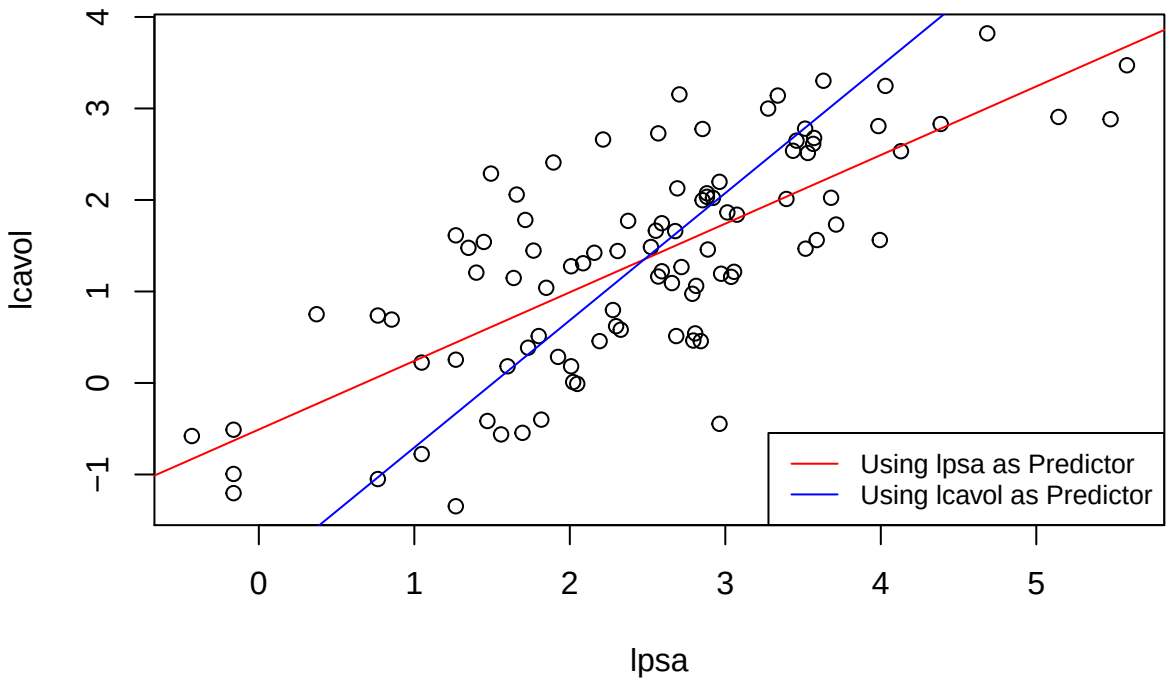
```
data("prostate")
prostate1 <- prostate[,c(1,2,5,4,3,6,8,7)]
res_err <- c()
r2 <- c()
for (val in 2:length(prostate1)) {
  fit <- lm(prostate1$lcavol ~., prostate1[2:val])
  res_err[val - 1] <- sqrt(sum(fit$residuals ** 2)/fit$df.residual)
  r2[val - 1] <- summary(fit)$r.squared
}
plot(1:7, res_err, type="l", col="red", frame = FALSE, ylim = c(0,1.15))
lines(1:7, r2, type="l", col="blue")
legend("topleft", legend=c("Residual Standard Error", "R Squared"),
       col=c("red", "blue"), lty = 1:1, cex=0.8)
```



Question 5

The following R code fits two regression models to data used in a study of men with prostate cancer due to receive a radical prostatectomy. We plot one line using lpsa as a predictor variable and lcavol as the response and another using lcavol as the predictor and lpsa as the response.

```
fit1 <- lm(lcavol ~ lpsa, data=prostate)
fit2 <- lm(lpsa ~ lcavol, data=prostate)
plot(lcavol ~ lpsa, data=prostate)
abline(fit1, col="red")
abline(-fit2$coefficients[1]/fit2$coefficients[2], 1/fit2$coefficients[2], col="blue")
legend("bottomright", legend=c("Using lpsa as Predictor", "Using lcavol as Predictor"),
       col=c("red", "blue"), lty = 1:1, cex=0.8)
```



The two lines representing the regressions models intercept at the point with respective x and y coordinates given by the mean of the two variables.