Bruce Campell ST 503 HW 2

Problems 1, 4, 7 Chapter 2 Faraway, Julian J. Linear Models with R, Second Edition. CRC Press.

Bruce Campbell

05 September, 2017

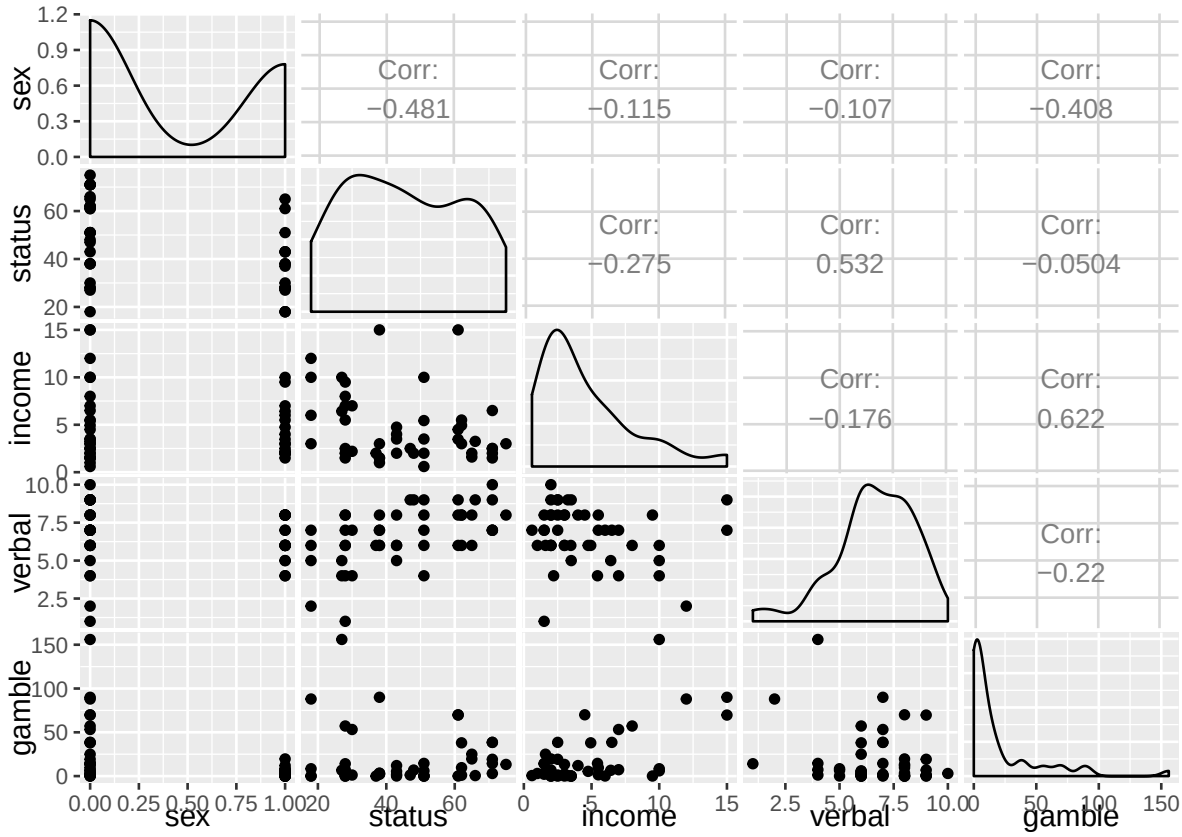
This document was rendered in Rmarkdown. Some of the code is not displayed. The markdown used to generate this is located on github at

https://github.com/brucebcampbell/applied-regression-with-R/blob/master/BruceCampbell_ST503_HW2_FarawayCh2_Pblms_1_4_7.Rmd

Problem 2.1

The dataset teengamb concerns a study of teenage gambling in Britain. Fit a regression model with the expenditure on gambling as the response and the sex, status, income and verbal score as predictors. Present the output.

- (a) What percentage of variation in the response is explained by these predictors?
- (b) Which observation has the largest (positive) residual? Give the case number.
- (c) Compute the mean and median of the residuals.
- (d) Compute the correlation of the residuals with the fitted values.
- (e) Compute the correlation of the residuals with the income.
- (f) For all other predictors held constant, what would be the difference in predicted expenditure on gambling for a male compared to a female?




```
##
## 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
```

(a) What percentage of variation in the response is explained by these predictors?

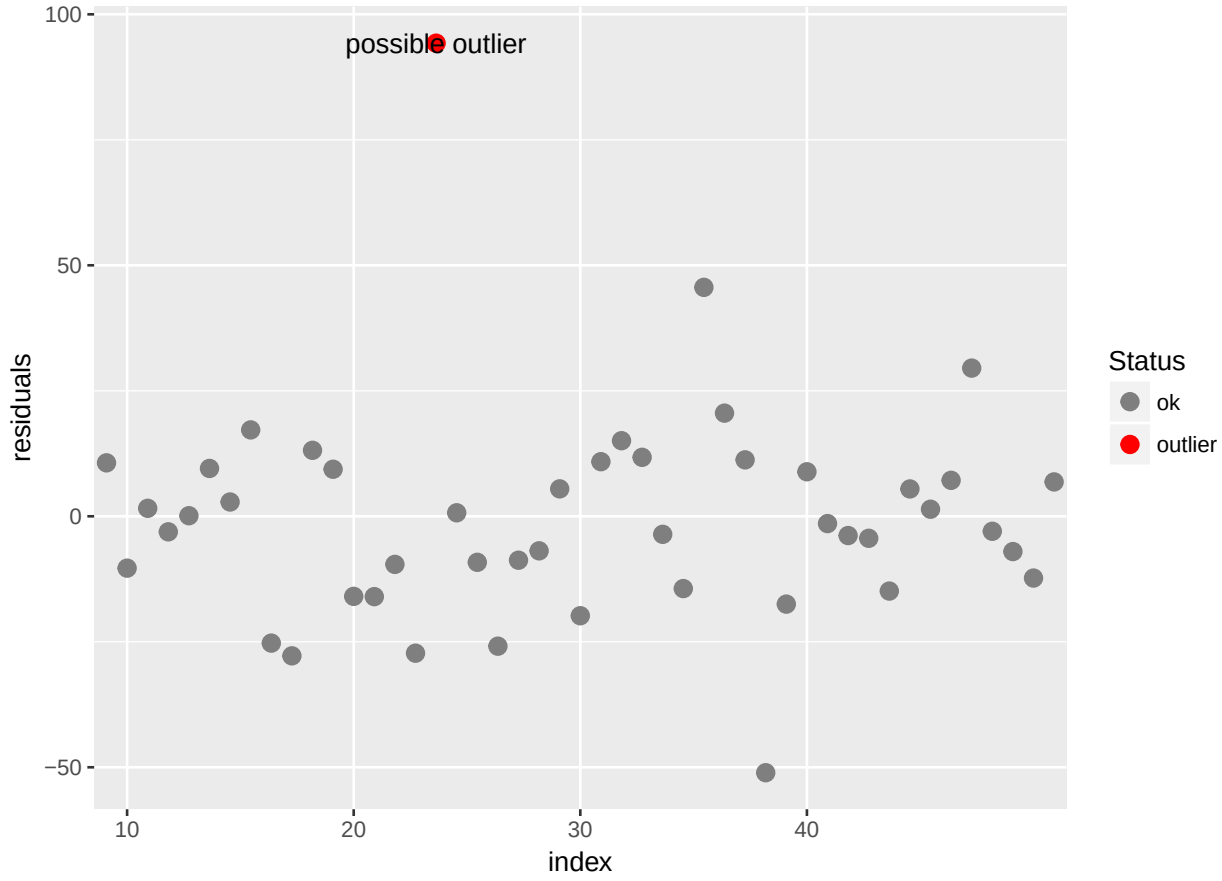
Here we calculate the proportion of explained and unexplained variance in the response that is given by the predictors in the mode we fit.

Table 1: Proportion of Variance Explained

var.explained.proportion
0.5267

(b) Which observation has the largest (positive) residual? Give the case number.

We're not sure if the question seeks the larges residual in absolute value or the largest of the positive residuals. We suspect that we're looking for the largest residual in absolute values since this may be an outlier that needs investigation, but we'll report both.



The largest residual occurs at index 24 of the dataframe. This is the associated case data.

Table 2: Potential outlier.

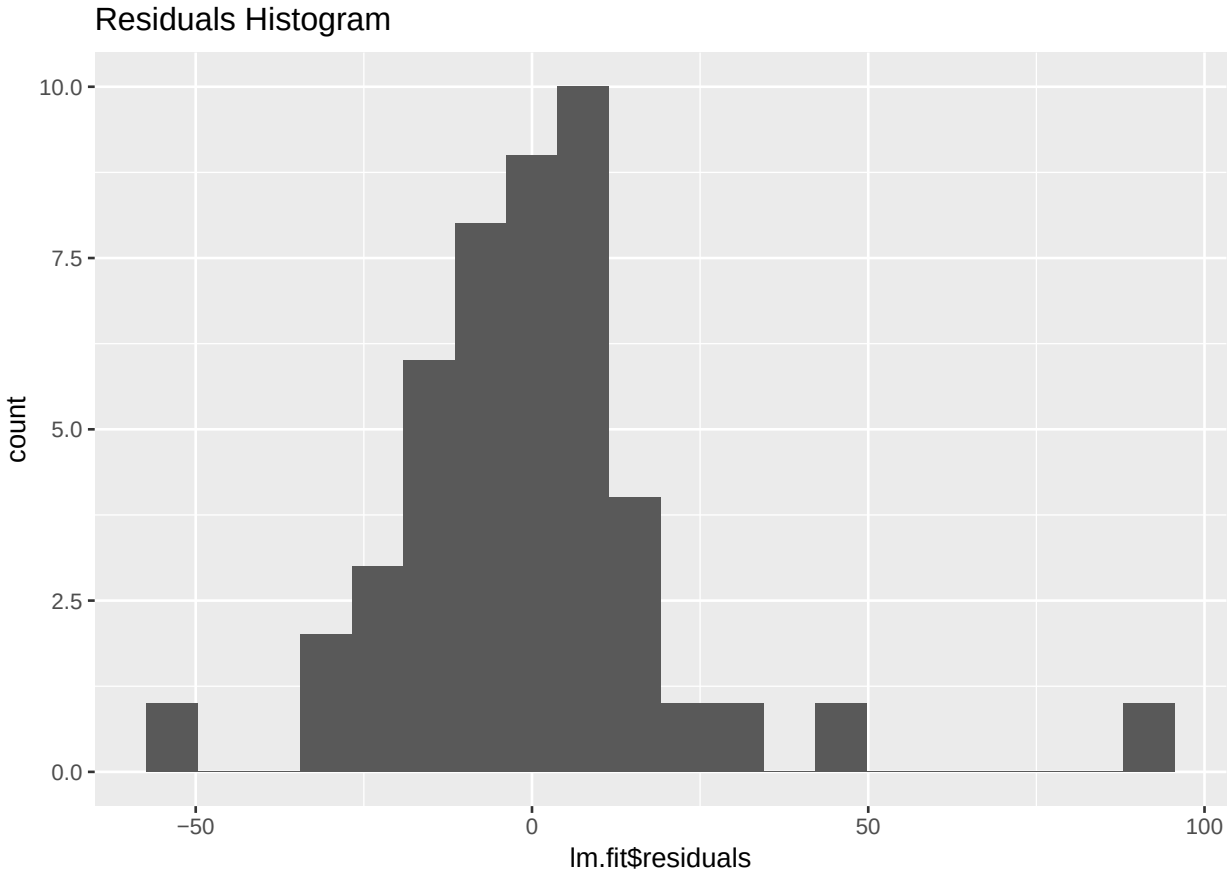
	sex	status	income	verbal	gamble
24	0	27	10	4	156

	sex	status	income	verbal	gamble

(c) Compute the mean and median of the residuals.

Table 3: mean and median of the residuals

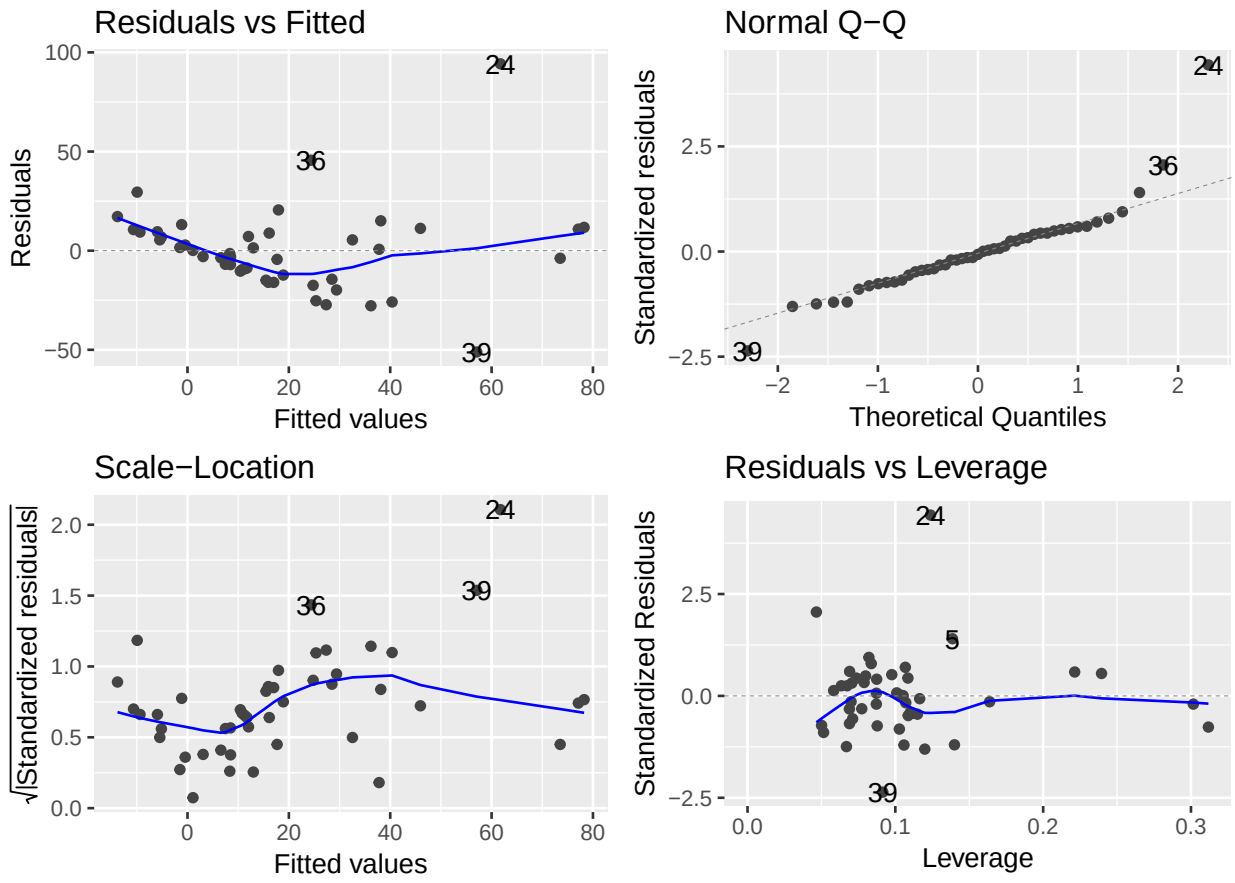
residuals.mean	residuals.median
-3.065e-17	-1.451




```
## [1] 21.68137
```

The mean residual is a very small number! We'd need to think through the implications of this - possibly it is an artifact of data that was generated.

Regression diagnostics are plotted below.

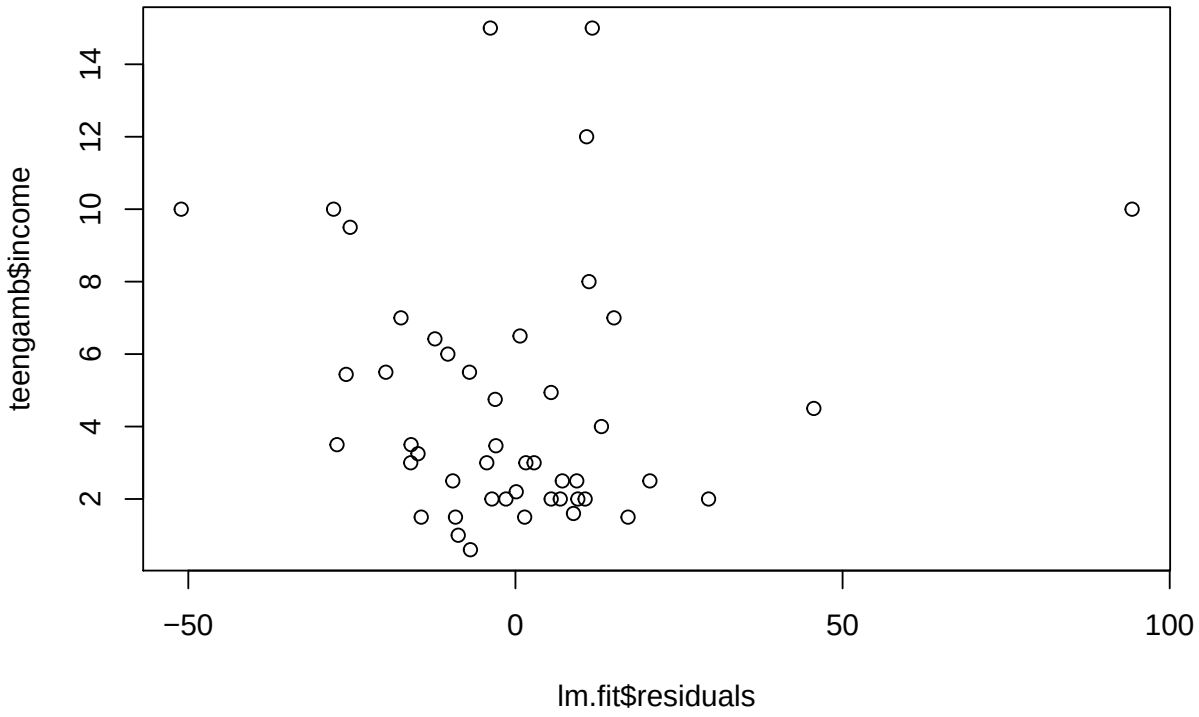


(d) Compute the correlation of the residuals with the fitted values.

corr.residuals.vs.fitted
-1.071e-16

(e) Compute the correlation of the residuals with the income.

corr.residuals.income
-7.242e-17



(f) For all other predictors held constant, what would be the difference in predicted expenditure on gambling for a male compared to a female?

This should be the value of the coefficient for gender. We need to be careful about the encoding and understanding whether this was treated as a factor in the regression. Querying the data `?teengamb` tells us that sex is encoded as so `0=male, 1=female`. Looking at the data frame teengamb we see that the class of the variable is integer and not a factor so we can now interpret the coefficient properly.

	gender.coefficient
sex	-22.12

This value represents the change in the response when there is a unit change in the predictor. In this case since female is encoded as `1` we can say that females have that much less gamble response (less because the coefficient is negative).

We can apply the model by hand to a element of the data set to see this in practice.

```
data.sample <- sample(nrow(teengamb), 1)
data.element <- teengamb[data.sample, ]
data.element$gamble <- NULL
```

```
data.element <- as.matrix(cbind(intercept = 1, data.element))
beta.hat <- as.matrix(lm.fit$coefficients)

pander(data.frame(data.element), caption = "Data sample")
```

Table 7: Data sample

	intercept	sex	status	income	verbal
43	1	0	75	3	8

```
response.orig <- (data.element) %*% beta.hat

# change the gender of our data element
data.element[1, 2] <- ifelse(data.element[1, 2] == 1, 0, 1)


pander(data.frame(data.element), caption = "Data sample with gender modified")
```

Table 8: Data sample with gender modified

	intercept	sex	status	income	verbal
43	1	1	75	3	8

```
response.gendermod <- (data.element) %*% beta.hat

pander(data.frame(response.difference = (response.orig - response.gendermod)))
```

	response.difference
43	22.12

Problem 2.4

The dataset prostate comes from a study on 97 men with prostate cancer who were due to receive a radical prostatectomy. Fit a model with lpsa as the response and lcavol as the predictor. Record the residual standard error and the R^2. Now add lweight, svi, lbph, age, lcp, pgg45 and gleason to the model one at a time. For each model record the residual standard error and the R^2. Plot the trends in these two statistics.

Load data and fit the models

Fit lpsa ~ lcavol +lweight

```
##
## Call:
## lm(formula = lpsa ~ lcavol, data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.67625 -0.41648  0.09859  0.50709  1.89673
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.50730    0.12194   12.36   <2e-16 ***
## lcavol       0.71932    0.06819   10.55   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7875 on 95 degrees of freedom
## Multiple R-squared:  0.5394, Adjusted R-squared:  0.5346
## F-statistic: 111.3 on 1 and 95 DF,  p-value: < 2.2e-16
```

```
##          term  estimate  std.error statistic      p.value
## 1 (Intercept) 1.5072979 0.12193682  12.36130 1.722234e-21
## 2      lcavol 0.7193201 0.06819288  10.54832 1.118616e-17
```

We will only print the models for $n = 2$ and $n = 8$ predictors.

Fit lpsa ~ lcavol +lweight

```
##          term   estimate  std.error  statistic      p.value
## 1 (Intercept) -0.3026179 0.56904195 -0.5318024 5.961175e-01
## 2      lcavol  0.6775253 0.06626223 10.2249086 6.120248e-17
## 3     lweight  0.5109495 0.15725697  3.2491371 1.606370e-03
```

Fit lpsa ~ lcavol +lweight + svi + lbph + age + lcp + pgg45+ gleason

```
##          term     estimate   std.error  statistic      p.value
## 1 (Intercept)  0.669336698 1.296387471  0.5163091 6.069335e-01
## 2      lcavol  0.587021826 0.087920303  6.6767493 2.110698e-09
## 3     lweight  0.454467424 0.170012435  2.6731423 8.955363e-03
## 4         svi  0.766157326 0.244309148  3.1360157 2.328749e-03
## 5        lbph  0.107054031 0.058449214  1.8315735 7.039846e-02
## 6         age -0.019637176 0.011172725 -1.7575995 8.229321e-02
## 7         lcp -0.105474263 0.091013487 -1.1588861 2.496377e-01
## 8       pgg45  0.004525231 0.004421179  1.0235350 3.088604e-01
```

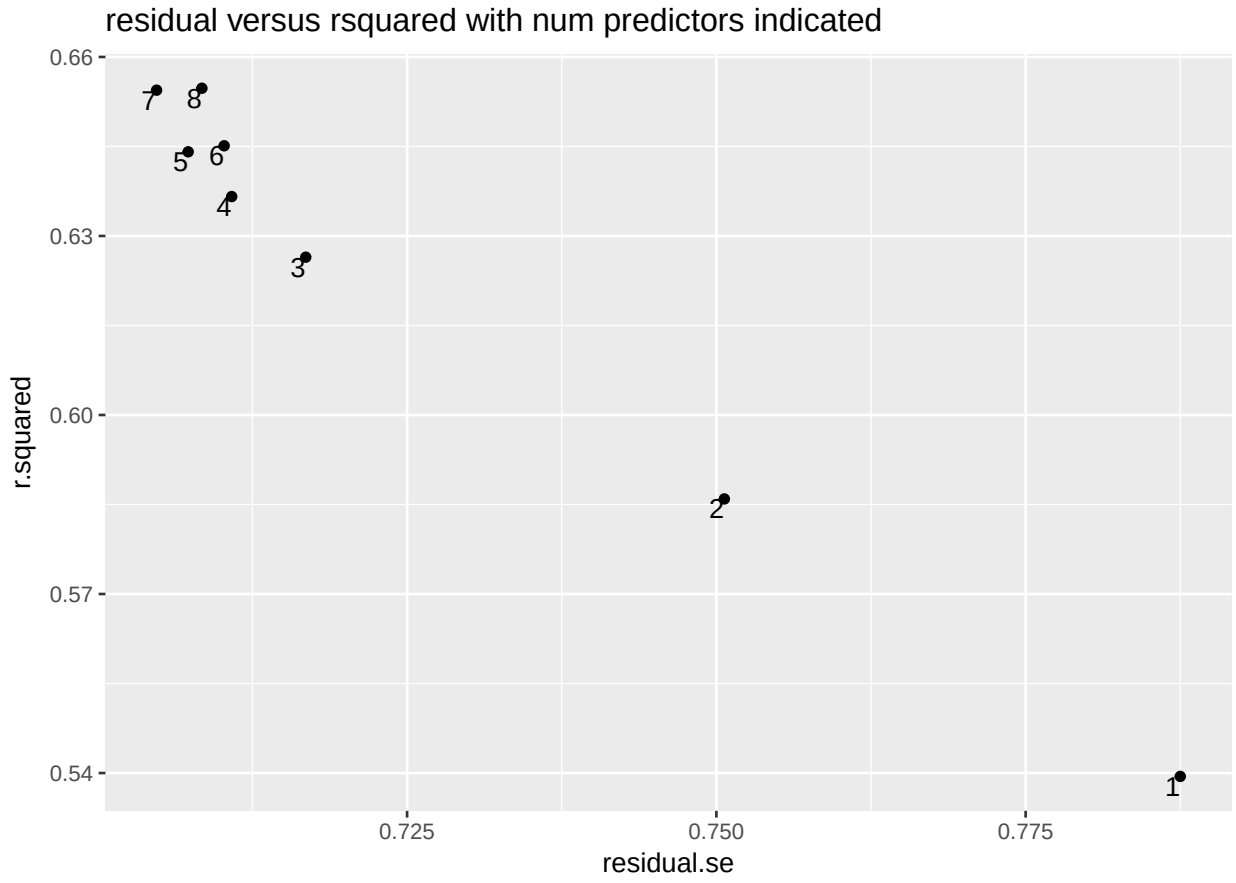
```
## 9     gleason  0.045141598 0.157464523  0.2866779 7.750328e-01
```

Present the model stats

Table 10: model statistics

num.predictors	r.squared	residual.se	model.string
8	0.6548	0.7084	lpsa ~ lcavol +lweight + svi + lbph + age + lcp + pgg45+ gleason
7	0.6544	0.7048	lpsa ~ lcavol +lweight + svi + lbph + age + lcp + pgg45
6	0.6451	0.7102	lpsa ~ lcavol +lweight + svi + lbph + age + lcp
5	0.6441	0.7073	lpsa ~ lcavol +lweight + svi + lbph + age
4	0.6366	0.7108	lpsa ~ lcavol +lweight + svi + lbph
3	0.6264	0.7168	lpsa ~ lcavol +lweight + svi
2	0.5859	0.7506	lpsa ~ lcavol +lweight
1	0.5394	0.7875	lpsa ~ lcavol

Plot SE versus R^2



We see that generally the proportion of variance explained by the model increases and the residual standard error decreases as the dimension of the model increases. The effect becomes less pronounced as we get to 6+ predictors. One could argue that inclusion of gleason to the model does not add much explanatory power. This may make empirical sense since the gleason score is assigned by a pathologist based on a stained tissue slide. It could be the case that this feature summaries or is a weak proxy for the biochemical variables.

Problem 2.7

An experiment was conducted to determine the effect of four factors on the resistivity of a semiconductor wafer. The data is found in wafer where each of the four factors is coded as - or + depending on whether the low or the high setting for that factor was used.

Fit the linear model $resist \sim x1 + x2 + x3 + x4$

```
## [1] "Inspect Data"
```

```
##   x1 x2 x3 x4 resist
## 1  -  -  -  -  193.4
```

```
## 2  +  -  -  -  247.6
## 3  -  +  -  -  168.2
## 4  +  +  -  -  205.0
## 5  -  -  +  -  303.4
## 6  +  -  +  -  339.9
```

```
## [1] "check the class of the columns"
```

```
## $x1
## [1] "factor"
##
## $x2
## [1] "factor"
##
## $x3
## [1] "factor"
##
## $x4
## [1] "factor"
##
## $resist
## [1] "numeric"
```

```
## [1] "Fi the model"
```

(a) Extract the X matrix using the model.matrix function. Examine this to determine how the low and high levels have been coded in the model.

```
##    (Intercept) x1+ x2+ x3+ x4+
## 1            1   0   0   0   0
## 2            1   1   0   0   0
## 3            1   0   1   0   0
## 4            1   1   1   0   0
## 5            1   0   0   1   0
## 6            1   1   0   1   0
## 7            1   0   1   1   0
## 8            1   1   1   1   0
## 9            1   0   0   0   1
## 10           1   1   0   0   1
## 11           1   0   1   0   1
## 12           1   1   1   0   1
## 13           1   0   0   1   1
## 14           1   1   0   1   1
## 15           1   0   1   1   1
## 16           1   1   1   1   1
## attr(,"assign")
## [1] 0 1 2 3 4
## attr(,"contrasts")
```

```
## attr(,"contrasts")$x1
## [1] "contr.treatment"
##
## attr(,"contrasts")$x2
## [1] "contr.treatment"
##
## attr(,"contrasts")$x3
## [1] "contr.treatment"
##
## attr(,"contrasts")$x4
## [1] "contr.treatment"
```

Now let's look at the data matrix to see how the factors are coded

x1	x2	x3	x4	resist
-	-	-	-	193.4
+	-	-	-	247.6
-	+	-	-	168.2
+	+	-	-	205
-	-	+	-	303.4
+	-	+	-	339.9
-	+	+	-	226.3
+	+	+	-	208.3
-	-	-	+	220
+	-	-	+	256.4
-	+	-	+	165.7
+	+	-	+	203.5
-	-	+	+	285
+	-	+	+	268
-	+	+	+	169.1
+	+	+	+	208.5

Comparing the model matrix to the original dataframe we see that low level $- \longrightarrow 0$ and high level $+ \longrightarrow 1$

(b) Compute the correlation in the X matrix. Why are there some missing values in the matrix?

Table 12: Correlation of X

	X.Intercept.	x1.	x2.	x3.	x4.
(Intercept)	1	NA	NA	NA	NA
x1+	NA	1	0	0	0
x2+	NA	0	1	0	0
x3+	NA	0	0	1	0
x4+	NA	0	0	0	1

The correlation in the X matrix is the pairwise values of the column correlations. The correlation is the covariance divided by the sqaure root of the product of the two variances.

If X and Y are jointly distributed random variables and the variances and covariances of both X and Y exist and the variances are nonzero, then the correlation of X and Y , denoted by , is $\rho = Cov(X,Y)/\sqrt{Var(X)Var(Y)}$

In this case we're dealing with samples and calculating the sample correlations.

There are NaN's in the due to the intercept. The variance of this vector is 0 and when R attempts to divide by 0 in the calculation of $\rho_{i,j}$ the results is a NaN. Note that R sets the diagonal of the correlation to one - it does not calculate the value - that's why we see $\rho_{1,1} = 1$.

We noted that the $i =\neq j$ terms for $i, j > 1$ were zero - this cause concern and we wrote some test code to validate the entries. The values were verified to be correct.

(d) Refit the model without x4 and examine the regression coefficients and standard errors? What stayed the the same as the original fit and what changed?

Reduced Model

```
##
## Call:
## lm(formula = resist ~ x1 + x2 + x3, data = wafer)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.137 -20.550   3.575  18.462  41.013
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   229.54      13.32  17.231 7.88e-10 ***
## x1+            25.76      13.32   1.934 0.077047 .
## x2+           -69.89      13.32  -5.246 0.000206 ***
## x3+            43.59      13.32   3.272 0.006677 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.64 on 12 degrees of freedom
## Multiple R-squared:  0.7777, Adjusted R-squared:  0.7221
## F-statistic: 13.99 on 3 and 12 DF,  p-value: 0.0003187
```

Full Model

```
##
## Call:
## lm(formula = resist ~ x1 + x2 + x3 + x4, data = wafer)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -43.381 -17.119   4.825  16.644  33.769
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   236.78      14.77  16.032 5.65e-09 ***
## x1+            25.76      13.21   1.950 0.077085 .
## x2+           -69.89      13.21  -5.291 0.000256 ***
## x3+            43.59      13.21   3.300 0.007083 **
## x4+           -14.49      13.21  -1.097 0.296193
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.42 on 11 degrees of freedom
## Multiple R-squared:  0.7996, Adjusted R-squared:  0.7267
## F-statistic: 10.97 on 4 and 11 DF,  p-value: 0.0007815
```

We note that the p value for X4 was not significant in the first model and that removing it resulted in a model where the explained variance is not significantly changes. We could do a LRT on the two models to further understand if adding X4 enhances the model.

(e) Explain how the change in the regression coefficients is related to the correlation matrix of X.

When the model matrix is orthogonal the covariance matrix of the sampling distribution of the regression parameters will be diagonal - when the error are iid $N(0, \sigma)$.

$$\hat{\beta} \sim N(\beta, \sigma(\mathbf{X}^\intercal \mathbf{X})^{-1})$$

This means the regression parameters are independent. That's why we did not see a change in the estimates of the coefficients for $X1\,X2$, $X3$ when we removed $X4$ from the model.

We can verify

$$(\mathbf{X}^\intercal \mathbf{X})^{-1}(\mathbf{X}^\intercal \mathbf{X}) = I$$

for our model matrix

```
##             (Intercept) x1+ x2+ x3+ x4+
## (Intercept)           1   0   0   0   0
## x1+                   0   1   0   0   0
## x2+                   0   0   1   0   0
## x3+                   0   0   0   1   0
## x4+                   0   0   0   0   1
```