Chapter 3 Problem 7

Faraway, Julian J. Linear Models with R, Second Edition (Chapman & Hall/CRC Texts

Bruce Campbell

10 September, 2017

In the punting data, we find the average distance punted and hang times of 10 punts of an American football as related to various measures of leg strength for 13 volunteers.

- (a) Fit a regression model with Distance as the response and the right and left leg strengths and flexibilities as predictors. Which predictors are significant at the 5% level?
- (b) Use an F-test to determine whether collectively these four predictors have a relationship to the response.
- (c) Relative to the model in (a), test whether the right and left leg strengths have the same effect.
- (d) Construct a 95% confidence region for (??RStr,??LStr). Explain how the test in (c) relates to this region - not required
- (e) Fit a model to test the hypothesis that it is total leg strength defined by adding the right and left leg strengths that is sufficient to predict the response in comparison to using individual left and right leg strengths.
- (f) Relative to the model in (a), test whether the right and left leg flexibilities have the same effect.
- (g) Test for left-right symmetry by performing the tests in (c) and (f) simultaneously.
- (h) Fit a model with Hang as the response and the same four predictors. Can we make a test to compare this model to that used in (a)? Explain.

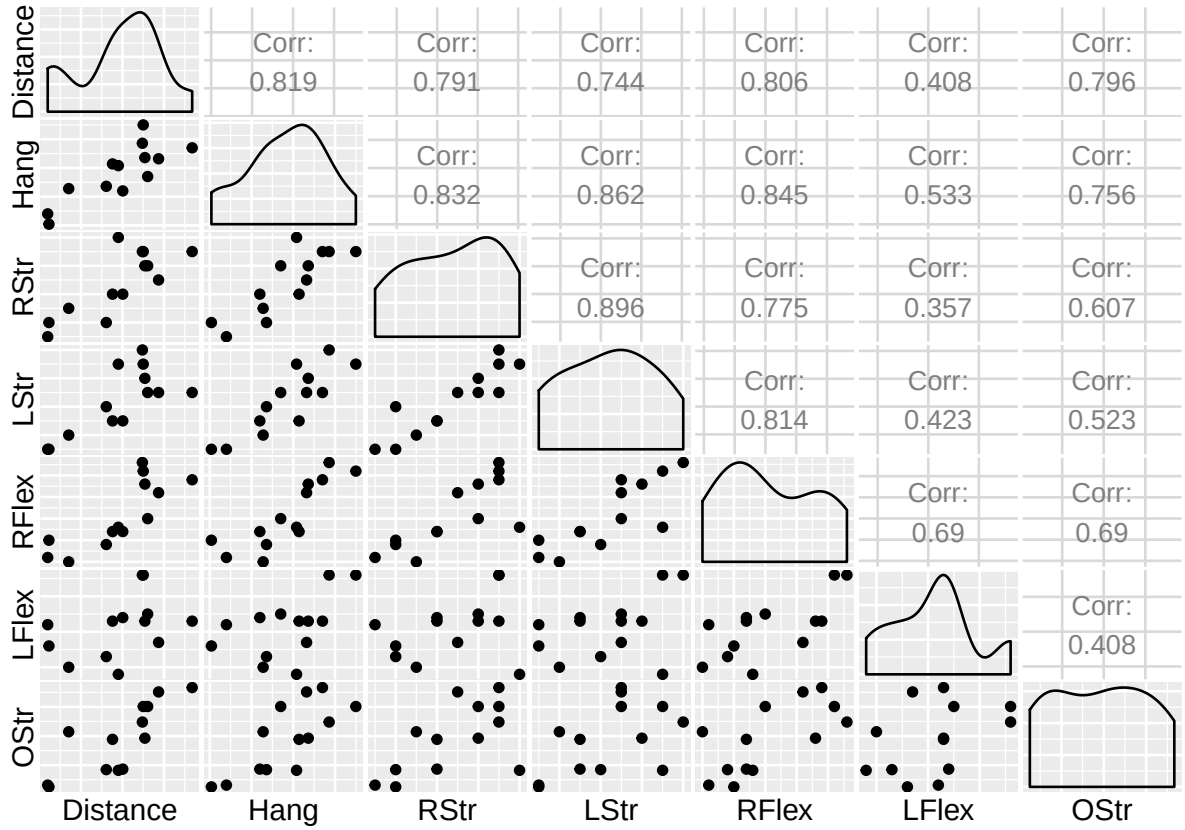
First we load and inspect the data.

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

```
##   Distance Hang RStr LStr RFlex LFlex   OStr
## 1   162.50 4.75  170  170   106   106 240.57
## 2   144.00 4.07  140  130    92    93 195.49
```

```
## 3   147.50 4.04  180  170    93    78 152.99
## 4   163.50 4.18  160  160   103    93 197.09
## 5   192.00 4.35  170  150   104    93 266.56
## 6   171.75 4.16  150  150   101    87 260.56
```

```
ggpairs(data = punting, axisLabels = "none")
```



a) Fit a regression model with Distance as the response and the right and left leg strengths and flexibilities as predictors. Which predictors are significant at the 5% level

```
lm.fit <- lm(Distance ~ RStr + LStr + RFlex + LFlex, data = punting)

summary(lm.fit)
```

```
##
## Call:
## lm(formula = Distance ~ RStr + LStr + RFlex + LFlex, data = punting)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.941  -8.958  -4.441  13.523  17.016
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -79.6236    65.5935  -1.214    0.259
## RStr          0.5116     0.4856   1.054    0.323
## LStr         -0.1862     0.5130  -0.363    0.726
## RFlex         2.3745     1.4374   1.652    0.137
## LFlex        -0.5277     0.8255  -0.639    0.541
##
## Residual standard error: 16.33 on 8 degrees of freedom
## Multiple R-squared:  0.7365, Adjusted R-squared:  0.6047
## F-statistic:  5.59 on 4 and 8 DF,  p-value: 0.01902
```

```
# Uncomment for diagnostic plots. plot(lm.fit)
```

We see that none of the predictors are significant at the $5% level for this model.

b) Use an F-test to determine whether collectively these four predictors have a relationship to the response

The test we want to perform is

$$H_0 : \beta_{Rstr} = \beta_{LStr} = \beta_{RFlex} = \beta_{LFlex} = 0$$

versus the alternative that one or more of the coefficients is not zero. The likelihood ratio test for the full model versus the null model $Y \sim \beta_0 + \epsilon$ works out to be an F-test.

```
lm.fit.null <- lm(Distance ~ 1, data = punting)

anova(lm.fit.null, lm.fit)
```

```
## Analysis of Variance Table
##
## Model 1: Distance ~ 1
## Model 2: Distance ~ RStr + LStr + RFlex + LFlex
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     12 8093.3
## 2      8 2132.6  4    5960.7 5.5899 0.01902 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the p-value we have enough evidence to reject the null hypothesis at a significance of 5% in this case and claim that collectively the four predictors have a predictive relationship

with the response.

(c) Relative to the model in (a), test whether the right and left leg strengths have the same effect.

The test we want to perform in this case is

$$H_0 : \beta_{Rstr} = \beta_{LStr}$$

versus the alternative that the effect is not the same.

```
lm.fit.subspace <- lm(Distance ~ I(RStr + LStr) + RFlex + LFlex, data = punting)

anova(lm.fit.subspace, lm.fit)
```

```
## Analysis of Variance Table
##
## Model 1: Distance ~ I(RStr + LStr) + RFlex + LFlex
## Model 2: Distance ~ RStr + LStr + RFlex + LFlex
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1      9 2287.4
## 2      8 2132.6  1    154.72 0.5804  0.468
```

Based on this p-value we do not have enough evidence to reject the null hypothesis that the right and left leg strength have the same effect.

(e) Fit a model to test the hypothesis that it is total leg strength defined by adding the right and left leg strengths that is sufficient to predict the response in comparison to using individual left and right leg strengths.

```
lm.fit.strength <- lm(Distance ~ RStr + LStr, data = punting)
summary(lm.fit.strength)
```

```
##
## Call:
## lm(formula = Distance ~ RStr + LStr, data = punting)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.280  -9.583   3.147  10.266  26.450
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.8490    33.0334   0.389    0.705
## RStr          0.7208     0.4913   1.467    0.173
## LStr          0.2011     0.4883   0.412    0.689
##
## Residual standard error: 17.24 on 10 degrees of freedom
## Multiple R-squared:  0.6327, Adjusted R-squared:  0.5592
## F-statistic: 8.611 on 2 and 10 DF,  p-value: 0.00669
```

```
lm.fit.strength.sum <- lm(Distance ~ I(RStr + LStr), data = punting)
summary(lm.fit.strength.sum)
```

```
##
## Call:
## lm(formula = Distance ~ I(RStr + LStr), data = punting)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.632 -11.531   2.171   8.443  30.672
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     14.0936    31.8838   0.442  0.66703
## I(RStr + LStr)   0.4601     0.1082   4.252  0.00136 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.68 on 11 degrees of freedom
## Multiple R-squared:  0.6217, Adjusted R-squared:  0.5874
## F-statistic: 18.08 on 1 and 11 DF,  p-value: 0.001361
```

```
anova(lm.fit.strength.sum, lm.fit.strength)
```

```
## Analysis of Variance Table
##
## Model 1: Distance ~ I(RStr + LStr)
## Model 2: Distance ~ RStr + LStr
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     11 3061.3
## 2     10 2973.1  1    88.281 0.2969 0.5978
```

(f) Relative to the model in (a), test whether the right and left leg flexibilities have the same effect.

```
lm.fit.subspace <- lm(Distance ~ RStr + LStr + I(RFlex + LFlex), data = punting)

anova(lm.fit.subspace, lm.fit)
```

```
## Analysis of Variance Table
##
## Model 1: Distance ~ RStr + LStr + I(RFlex + LFlex)
## Model 2: Distance ~ RStr + LStr + RFlex + LFlex
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1      9 2648.4
## 2      8 2132.6  1    515.72 1.9346 0.2017
```

Based on this p-value we do not have enough evidence to reject the null hypothesis that the right and left leg flexibility have the same effect.

(g) Test for left-right symmetry by performing the tests in (c) and (f) simultaneously

The test we want to perform is

$$H_0 : \beta_{Rstr} = \beta_{LStr} \ , \ \beta_{RFlex} = \beta_{LFlex}$$

```
lm.fit.subspace <- lm(Distance ~ I(RStr + LStr) + I(RFlex + LFlex), data = punting)

anova(lm.fit.subspace, lm.fit)
```

```
## Analysis of Variance Table
##
## Model 1: Distance ~ I(RStr + LStr) + I(RFlex + LFlex)
## Model 2: Distance ~ RStr + LStr + RFlex + LFlex
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1     10 2799.1
## 2      8 2132.6  2    666.43 1.25  0.337
```

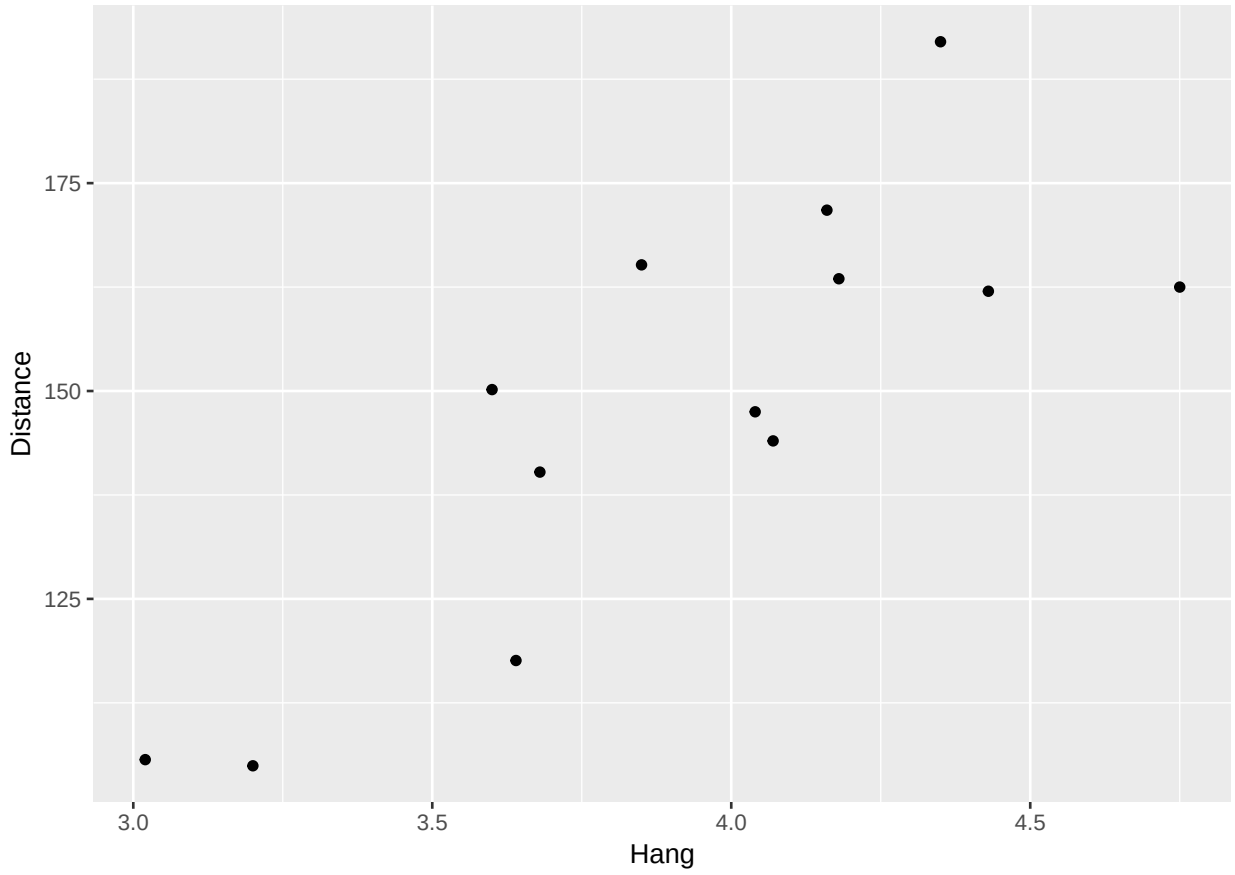
Based on this p-value we can not reject the null hypothesis of right-left symmetry.

(h) Fit a model with Hang as the response and the same four predictors. Can we make a test to compare this model to that used in (a)? Explain.

```
lm.fit <- lm(Hang ~ RStr + LStr + RFlex + LFlex, data = punting)

summary(lm.fit)
```

```
##
## Call:
## lm(formula = Hang ~ RStr + LStr + RFlex + LFlex, data = punting)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.36297 -0.13528 -0.07849  0.09938  0.35893
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.225239   1.032784  -0.218    0.833
## RStr         0.005153   0.007645   0.674    0.519
## LStr         0.007697   0.008077   0.953    0.369
## RFlex        0.019404   0.022631   0.857    0.416
## LFlex        0.004614   0.012998   0.355    0.732
##
## Residual standard error: 0.2571 on 8 degrees of freedom
## Multiple R-squared:  0.8156, Adjusted R-squared:  0.7235
## F-statistic: 8.848 on 4 and 8 DF,  p-value: 0.004925
```

We see a higher R^2 for this model. Here is a plot of hang verus distance



It is not clear what the criteria is for comparison in this case. We know we can't use an F-test - the models are not nested. We could build a full model with all the variables and look at interactions, but that's not a test. We also don't have enough data to consider all the interactions in $Distance \sim Hang * RStr * LStr * RFlex * LFlex$

NCSU ST 503 Discussion 5

Probem 4.5 Faraway, Julian J. Linear Models with R CRC Press.

Bruce Campbell

Comparing models of body fat measurement

For the fat data used in this chapter, a smaller model using only age, weight, height and abdom was proposed on the grounds that these predictors are either known by the individual or easily measured.

(a) Compare this model to the full thirteen-predictor model used earlier in the chapter. Is it justifiable to use the smaller model?

$brozek \sim age + weight + height + abdom$

```
##          term      estimate  std.error  statistic      p.value
## 1 (Intercept) -32.769635854 6.54190241 -5.0091906 1.041540e-06
## 2         age  -0.007051258 0.02434164 -0.2896789 7.723049e-01
## 3      weight  -0.123721774 0.02504553 -4.9398736 1.441726e-06
## 4      height  -0.116693958 0.08272693 -1.4105921 1.596231e-01
## 5       abdom   0.889704097 0.06726722 13.2264134 1.492006e-30
```

rsquared
0.7211

$brozek \sim age + weight + height + neck + chest + abdom + hip + thigh + knee + ankle + biceps + forearm + wrist$

```
##           term     estimate   std.error   statistic      p.value
## 1  (Intercept) -15.29254907 16.06992071 -0.95162567 3.422523e-01
## 2          age   0.05678616  0.02996465  1.89510481 5.929042e-02
## 3       weight  -0.08030986  0.04958051 -1.61978675 1.066023e-01
## 4       height  -0.06460028  0.08893033 -0.72641448 4.682985e-01
## 5         neck  -0.43754090  0.21533372 -2.03192006 4.327265e-02
## 6        chest  -0.02360333  0.09183940 -0.25700662 7.973957e-01
## 7        abdom   0.88542903  0.08007684 11.05724248 3.306570e-23
## 8          hip  -0.19841862  0.13515624 -1.46806848 1.434060e-01
## 9        thigh   0.23189542  0.13371812  1.73421094 8.417548e-02
```

```
## 10        knee  -0.01167679  0.22414282 -0.05209531 9.584964e-01
## 11       ankle   0.16353590  0.20514349  0.79717810 4.261422e-01
## 12      biceps   0.15279894  0.15851276  0.96395360 3.360476e-01
## 13     forearm   0.43048875  0.18445247  2.33387361 2.043567e-02
## 14       wrist  -1.47653692  0.49551887 -2.97977942 3.183449e-03
```

rsquared
0.749

The R^2 is slightly higher for the full model, but we claim that based on practical model deployment considerations it's justifiable to use the smaller model. If the measurements for the full model were made in a laboratory setting, one could imagine a scenario where the full model would perform worse in deployment due to poor measurement of the extra variables.

(b) Compute a 95% prediction interval for median predictor values and compare to the results to the interval for the full model. Do the intervals differ by a practically important amount?

Subset model prediction interval

```
##        fit      lwr      upr
## 1 17.84028 9.696631 25.98392
```

pi.width
16.29

Full model prediction interval

```
##        fit     lwr      upr
## 1 17.49322 9.61783 25.36861
```

pi.width
15.75

The full model does have a smaller prediction window. We don't see a big difference for the prediction intervals for the 2 models.

(c) For the smaller model, examine all the observations from case numbers 25 to 50. Which two observations seem particularly anomalous?

We plotted the features and examined the raw data and determined that case numbers 39 and 42 are potential anomalies or represent extreme values for the predictors.

Table 5: Possible outliers in dataset fat.

	age	weight	height	abdom
39	46	363.1	72.25	148.1
42	44	205	29.5	104.3

(d) Recompute the 95% prediction interval for median predictor values after these two anomalous cases have been excluded from the data. Did this make much difference to the outcome?

```
##       fit      lwr      upr
## 1 17.9033 9.887851 25.91874
```

pi.width
16.03

The prediction interval has gotten smaller but the removal of the outliers has not changed the size of the prediction interval by a lot. If we look at what happends when we perform this at the extreme values of the model parameters we might get another answer to this question.

Prediction interval for $brozek \sim age + weight + height + abdom$ at max of predictors

```
##        fit      lwr    upr
## 1 44.42187 35.41174 53.432
```

pi.width
18.02

Prediction interval for $brozek \sim age + weight + height + abdom$ at max of predictors. Model fit without outliers.

```
##        fit      lwr      upr
## 1 38.01869 29.61016 46.42723
```

pi.width
16.82

We see that the differnce prediction interval sizes is greater at the extremes of the predictors.

NCSU ST 503 Discussion 6

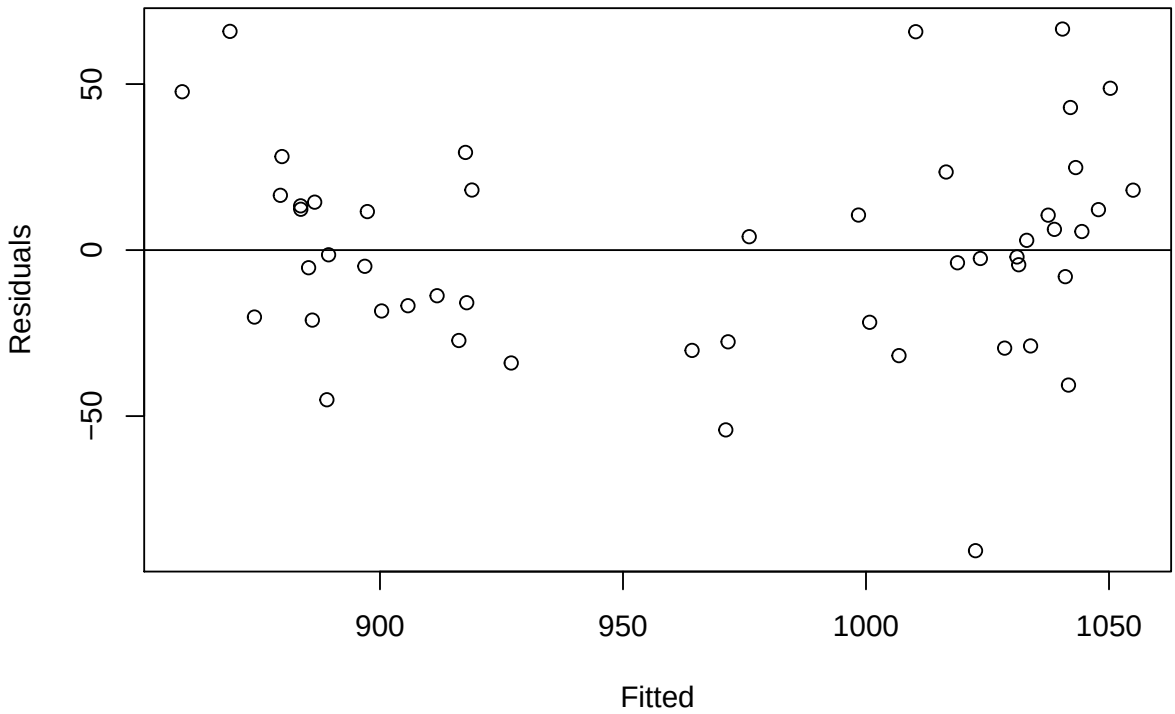
Probem 6.1 Faraway, Julian J. Linear Models with R CRC Press.

Bruce Campbell

Regression diagnostics with the SAT data set.

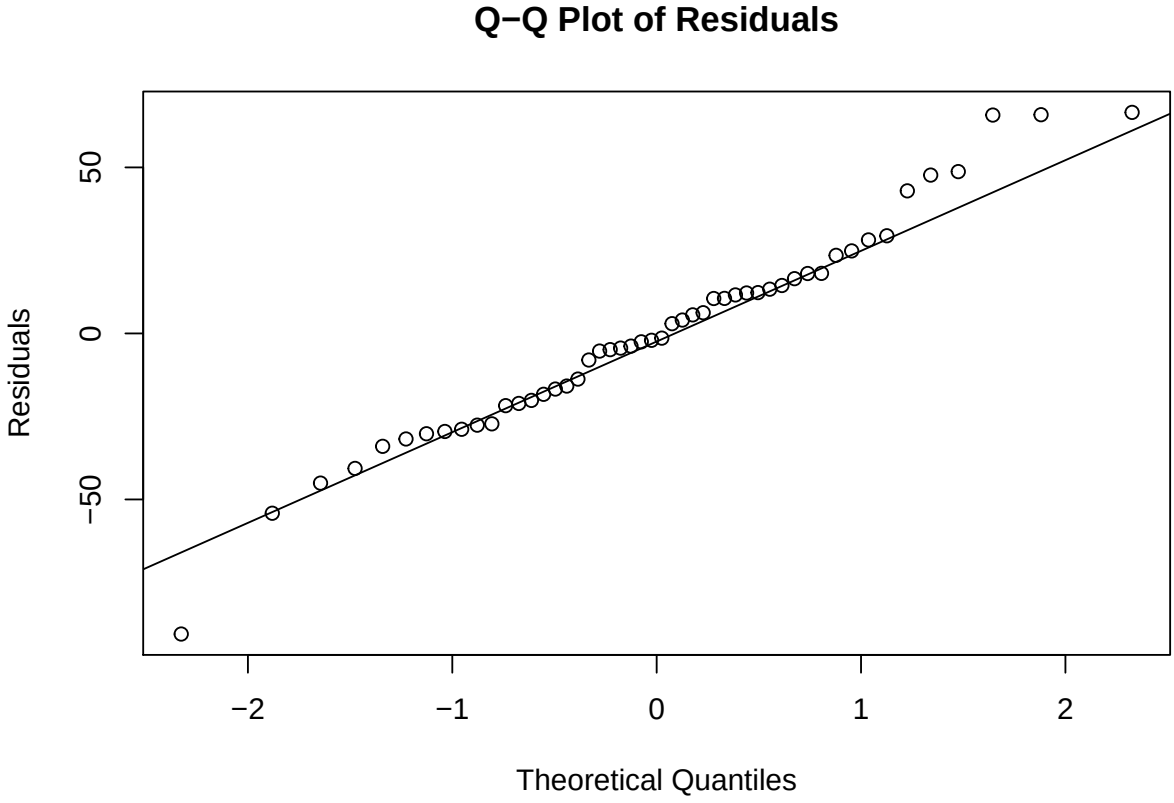
Using the sat dataset, fit a model with the total SAT score as the response and expend, salary, ratio and takers as predictors. Perform regression diagnostics on this model to answer the following questions. Display any plots that are relevant. Do not provide any plots about which you have nothing to say. Suggest possible improvements or corrections to the model where appropriate.

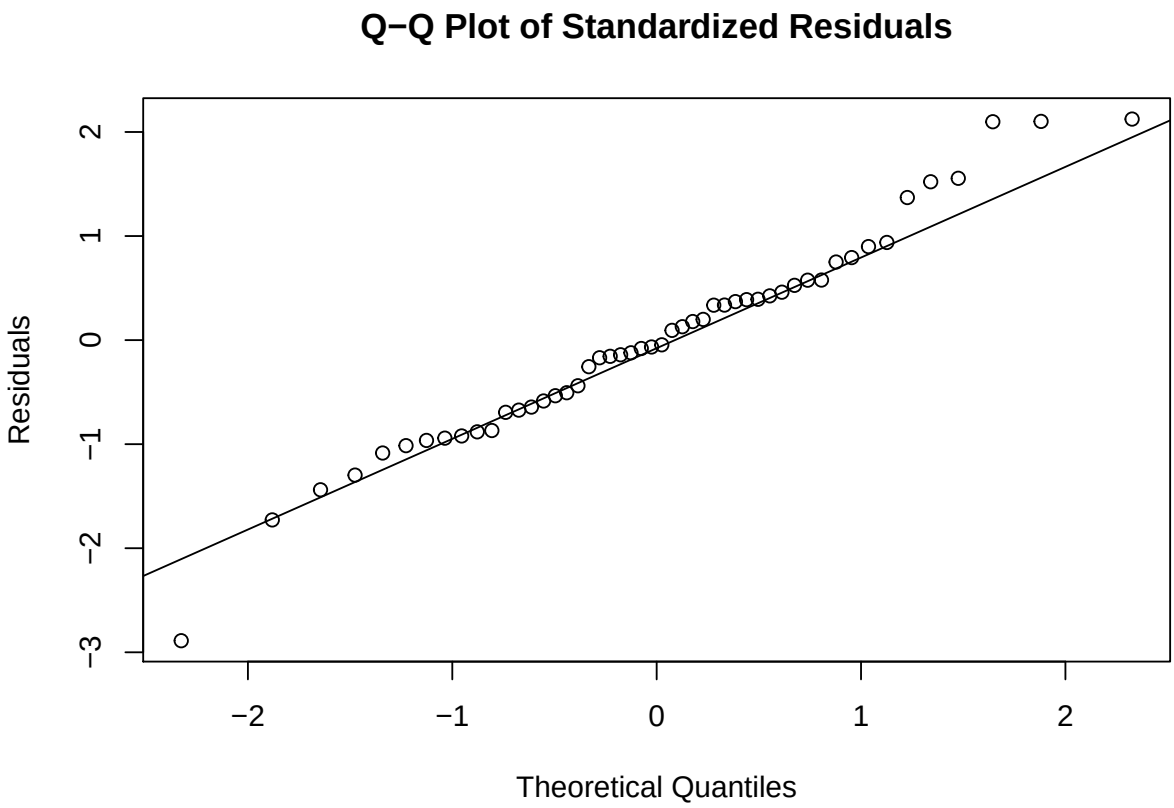
(a) Check the constant variance assumption for the errors.



To check the assumption of constant variance we plot fitted values against the residuals - looking for any structure in the distribution of values about the theoretical mean value line $E[\epsilon] = 0$. There is nothing alarming with this plot, the variance seems relatively constant along the range of the fitted values.

(b) Check the normality assumption.





Generally the residuals appear normally distributed in the middle of the range. The empirical distribution is slightly right skewed and there's a single point on the lower quantile that deviates from the theoretical distribution.

(c) Check for large leverage points.

Table 1: High Leverage Data Elements

	expend	ratio	salary	takers	verbal	math	total
California	4.992	24	41.08	45	417	485	902
Connecticut	8.817	14.4	50.05	81	431	477	908
New Jersey	9.774	13.8	46.09	70	420	478	898
Utah	3.656	24.3	29.08	4	513	563	1076

We've used the rule of thumb that points with a leverage greater than $\frac{2p}{n}$ should be looked at.

(d) Check for outliers.

Table 2: Range of Studentized residuals

range.residuals.left	range.residuals.right
-3.124	2.53

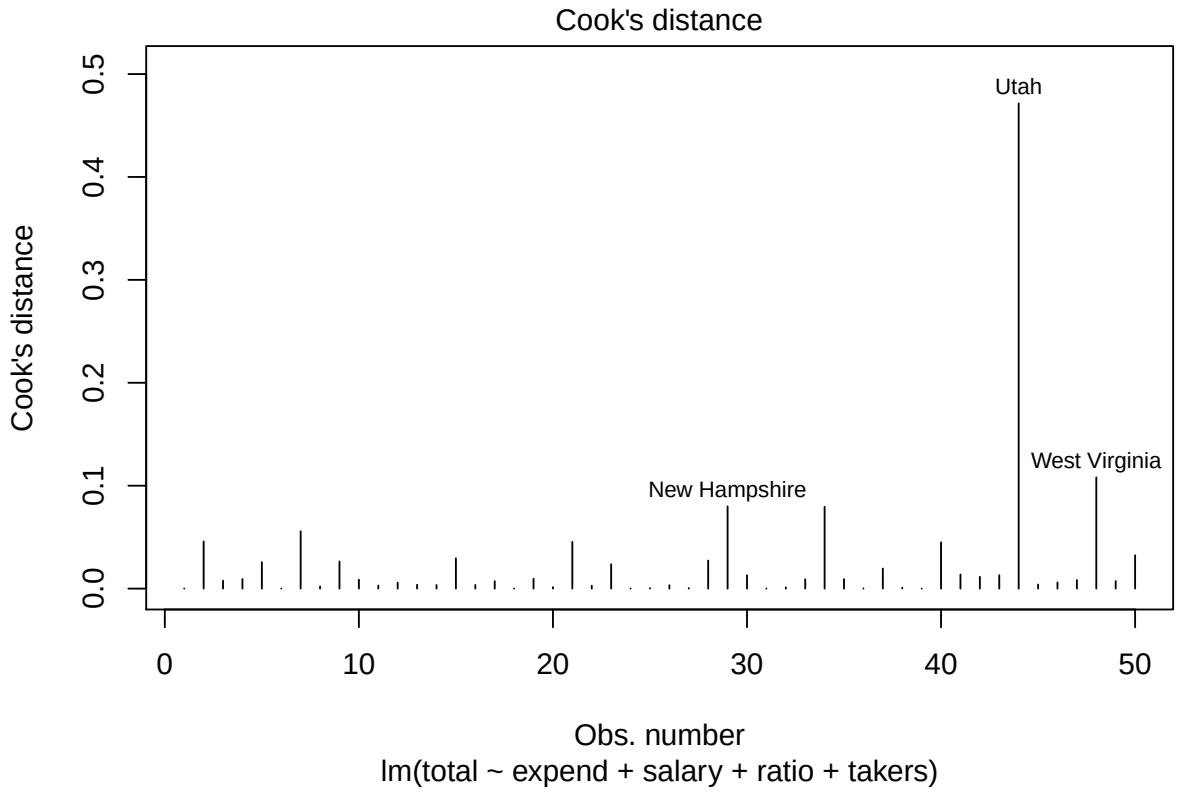
Table 3: Bonferroni corrected t-value

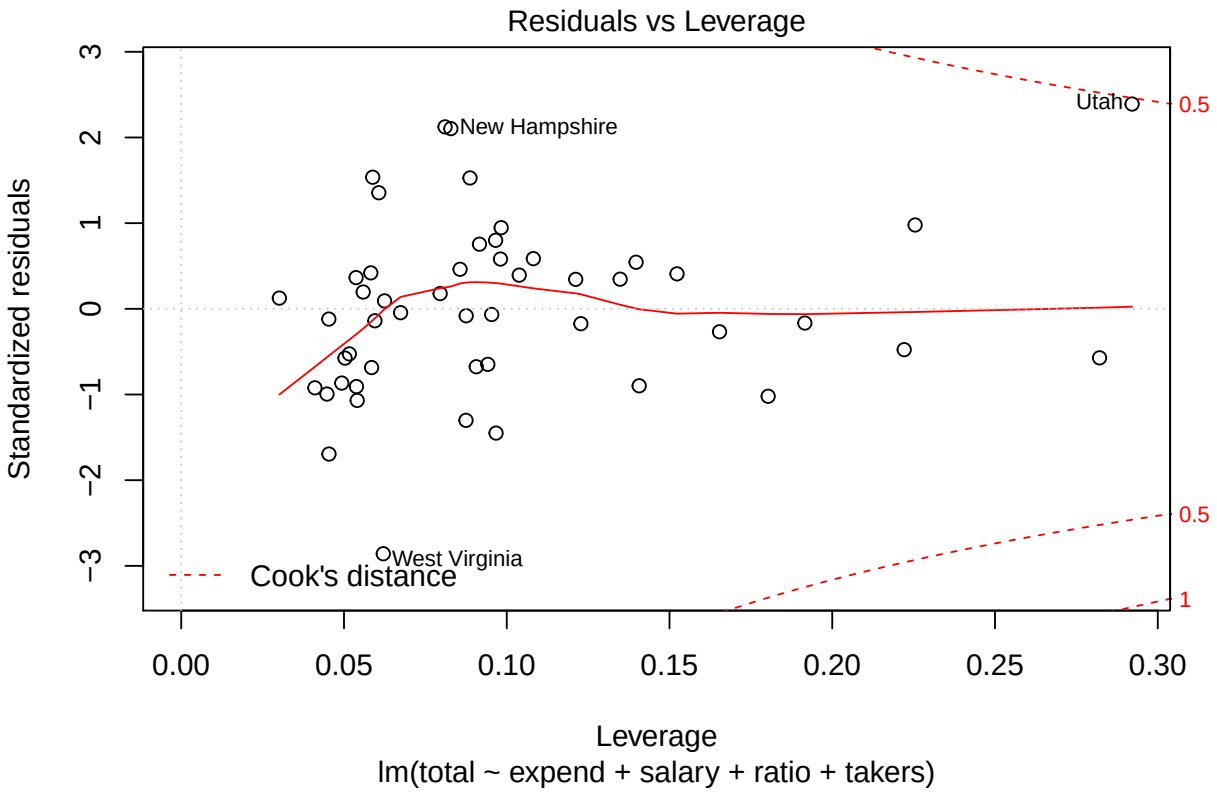
t.val.alpha
-3.526

Since none of the studentized residuals fall outside the interval given by the Bonferroni corrected t-values we claim there are no outliers in the dataset.

(e) Check for influential points.

We plot the Cook's distances and the residual-leverage plot with level set contours of the Cook distance.





We see the Utah, New Hampshire, and West Virginia are candidate influential points. The book does not discuss a criteria for selecting influential points from the Cook distances.

Some guidelines for selecting influential points; * points with a Cook distance more than three times the mean Cook distance
* points with a Cook distance greater than 4/n * points with a cook distance greater than 1

Here we select points with a Cook distance more than three times the mean Cook distance.

Table 4: Mean Cook distance

mean.cooks.distance
0.02575

Table 5: Points with Cook distance greater than three times the mean Cook distance.

	cook.distance
New Hampshire	0.07989
North Dakota	0.07954
Utah	0.4715

	cook.distance
West Virginia	0.1081

NCSU ST 503 Discussion 7

Probems 6.2,6.3,6.4,6.5 Parts c,d,e,f Faraway, Julian J. Linear Models with R CRC Press.

Bruce Campbell

6.2 Using the teengamb dataset, fit a model with gamble as the response and the other variables as predictors.

(c) Check for large leverage points.

```
rm(list = ls())
data(teengamb, package = "faraway")
```

```
df <- teengamb
numPredictors <- (ncol(df) - 1)
lm.fit <- lm(gamble ~ ., data = df)
hatv <- hatvalues(lm.fit)
lev.cut <- (numPredictors + 1) * 2 * 1/nrow(df)
high.leverage <- df[hatv > lev.cut, ]
pander(high.leverage, caption = "High Leverage Data Elements")
```

Table 1: High Leverage Data Elements

	sex	status	income	verbal	gamble
31	0	18	12	2	88
33	0	38	15	7	90
35	0	28	1.5	1	14.1
42	0	61	15	9	69.7

We've used the rule of thumb that points with a leverage greater than $\frac{2p}{n}$ should be looked at.

(d) Check for outliers.

```
studentized.residuals <- rstudent(lm.fit)
max.residual <- studentized.residuals[which.max(abs(studentized.residuals))]
range.residuals <- range(studentized.residuals)
```

```
names(range.residuals) <- c("left", "right")
pander(data.frame(range.residuals = t(range.residuals)), caption = "Range of Studentized
```

Table 2: Range of Studentized residuals

range.residuals.left	range.residuals.right
-2.506	6.016

```
p <- numPredictors + 1
n <- nrow(df)
t.val.alpha <- qt(0.05/(n * 2), n - p - 1)
pander(data.frame(t.val.alpha = t.val.alpha), caption = "Bonferroni corrected t-value")
```

Table 3: Bonferroni corrected t-value

t.val.alpha
-3.523

```
outlier.index <- abs(studentized.residuals) > abs(t.val.alpha)

outliers <- df[outlier.index == TRUE, ]

if (nrow(outliers) >= 1) {
    pander(outliers, caption = "outliers")
}
```

Table 4: outliers

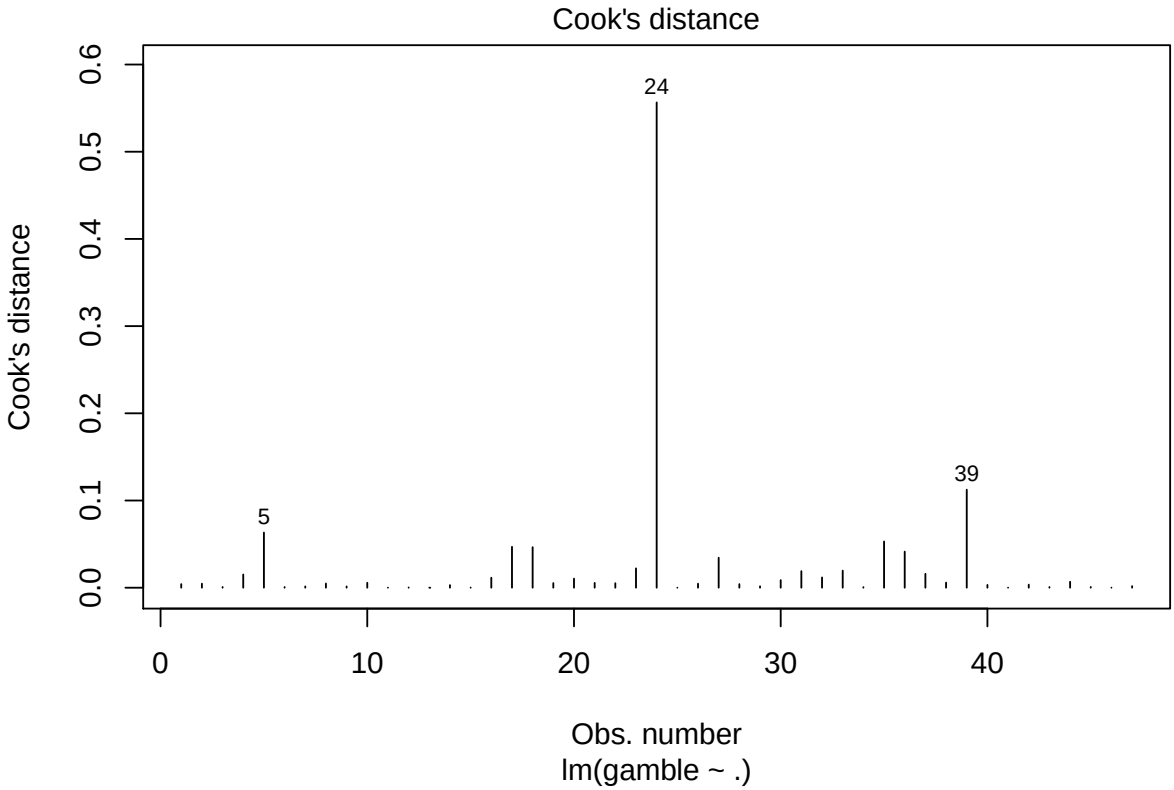
	sex	status	income	verbal	gamble
24	0	27	10	4	156

Here we look for studentized residuals that fall outside the interval given by the Bonferroni corrected t-values.

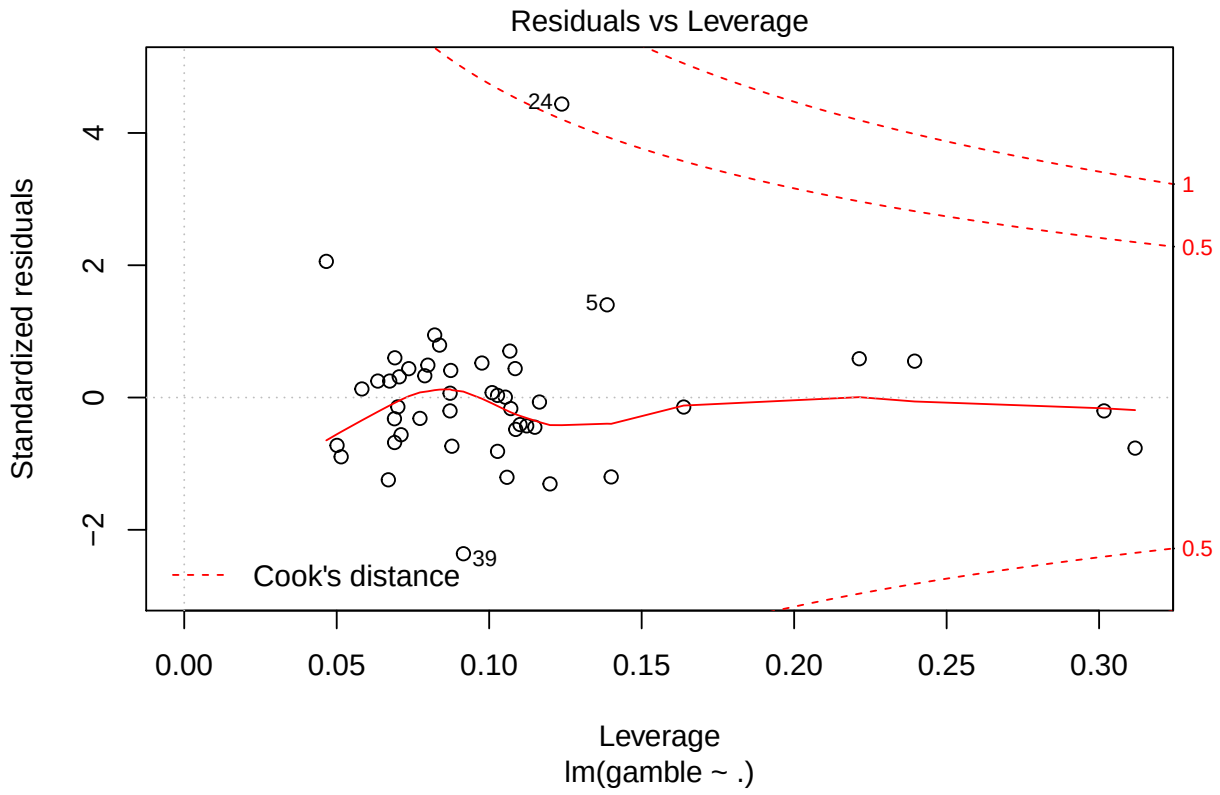
(e) Check for influential points.

We plot the Cook's distances and the residual-leverage plot with level set contours of the Cook distance.

```
plot(lm.fit, which = 4)
```



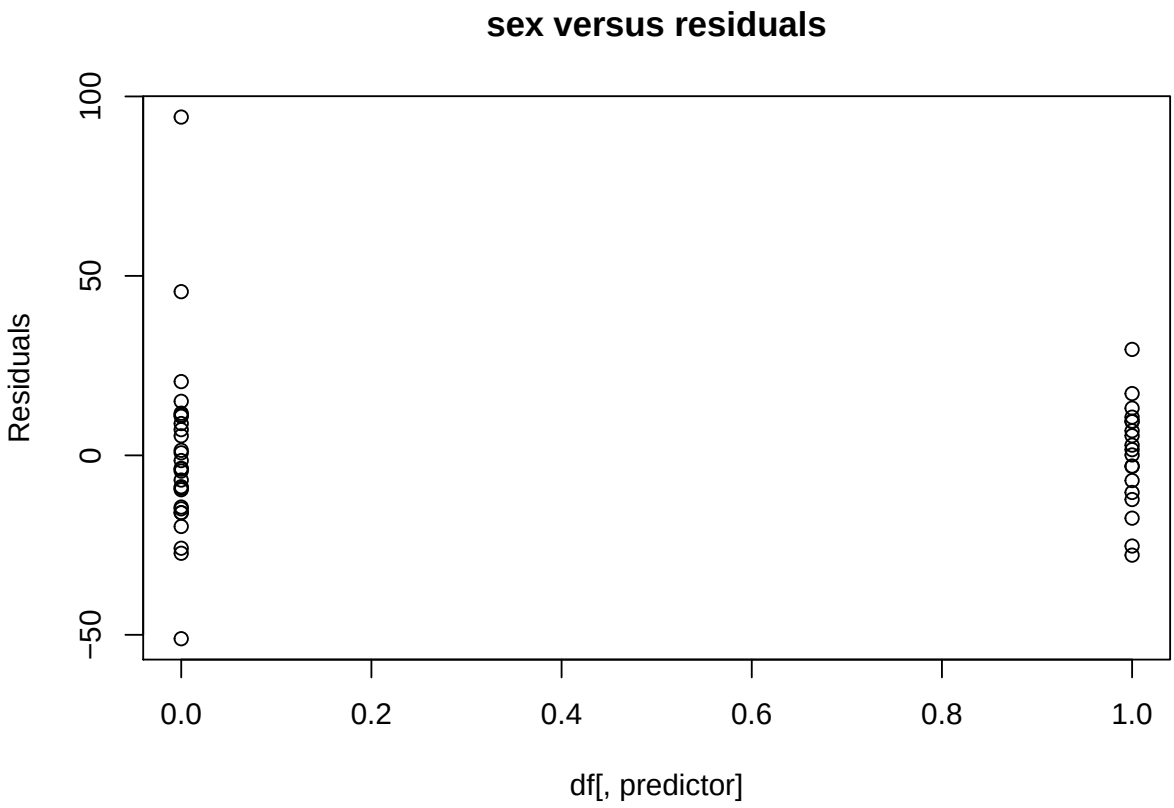

```
plot(lm.fit, which = 5)
```

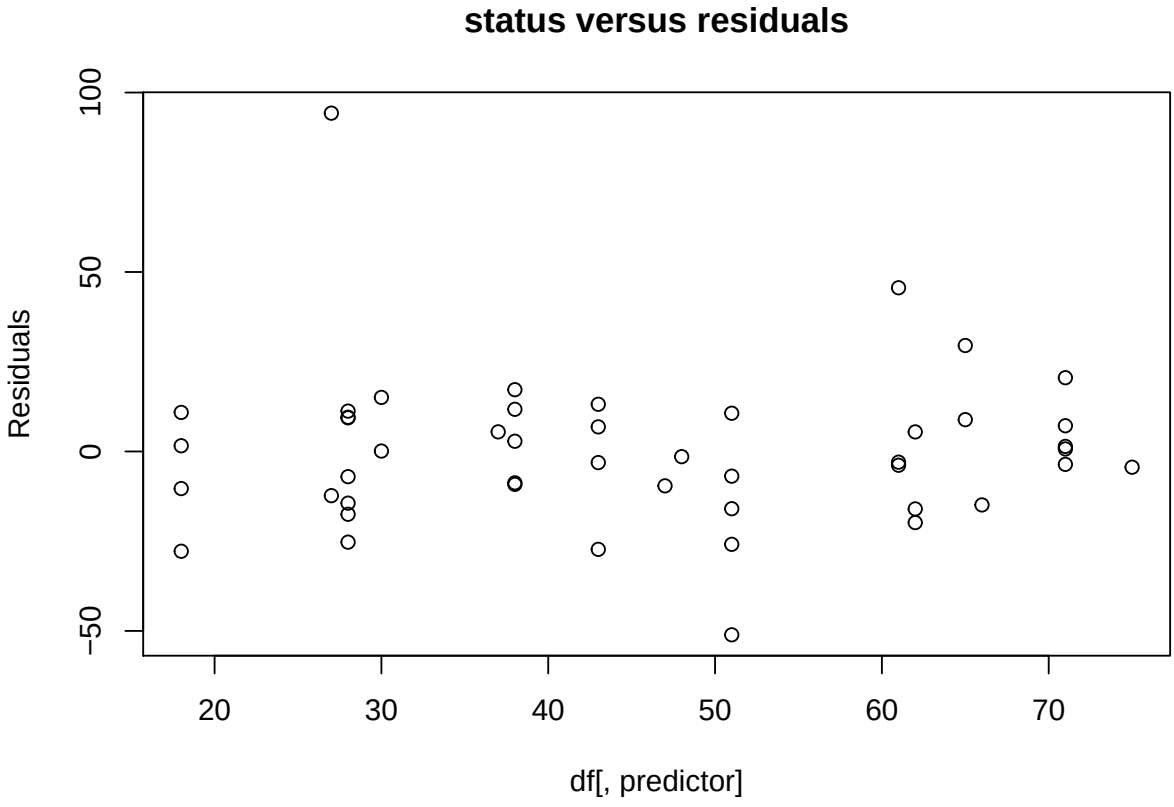


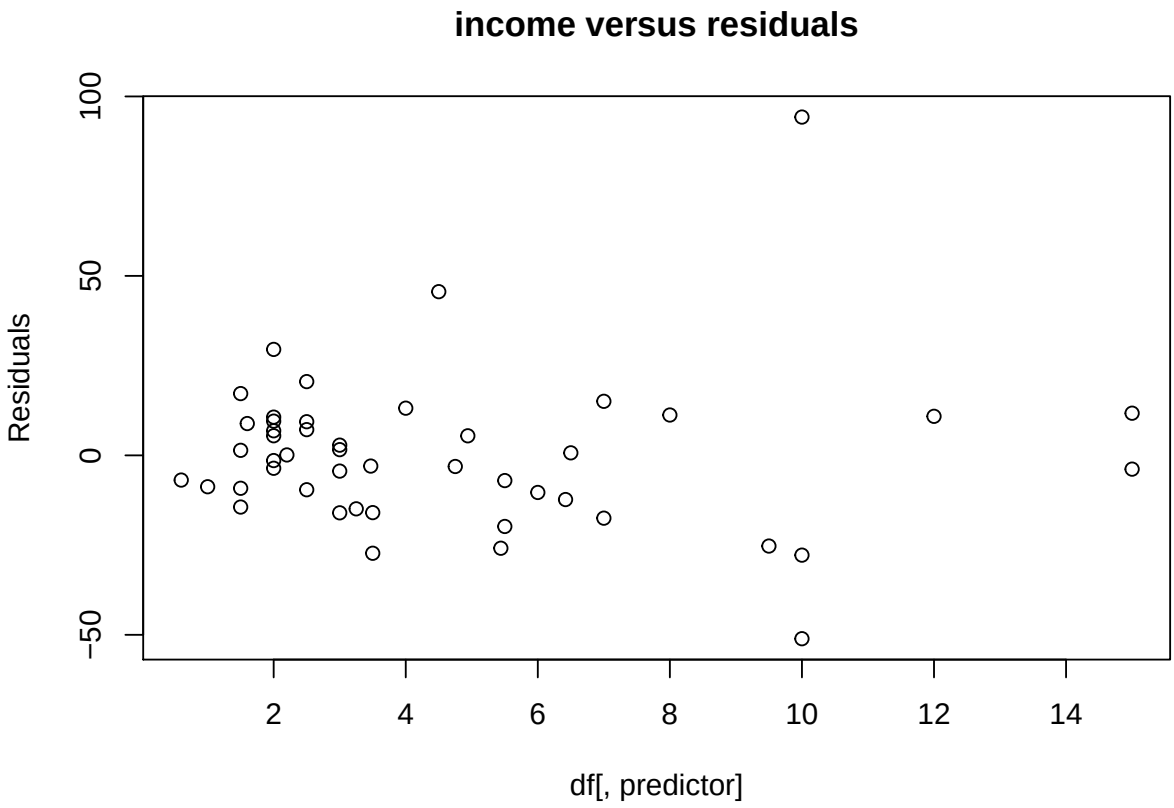
(f) Check for structure in the model.

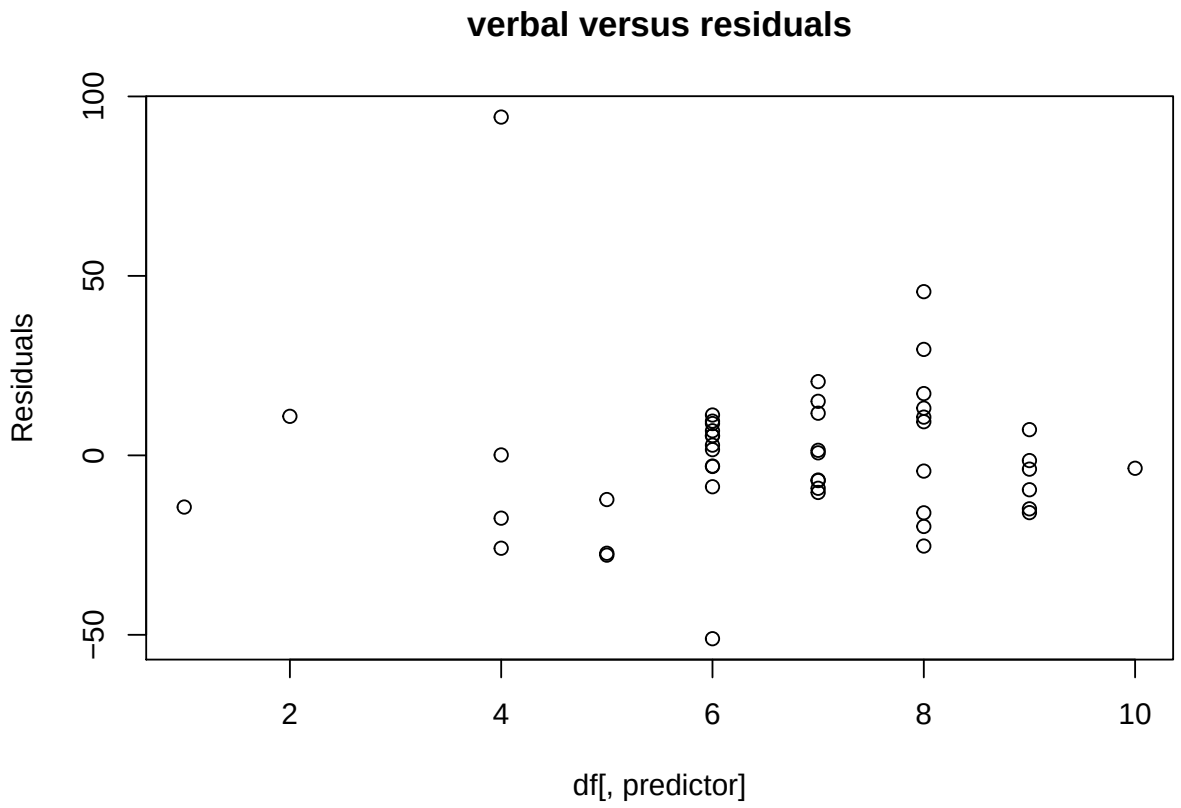
Plot residuals versus predictors

```
predictors <- names(lm.fit$coefficients)
predictors <- predictors[2:length(predictors)]

for (i in 1:length(predictors)) {
    predictor <- predictors[i]

    plot(df[, predictor], residuals(lm.fit), xlab = , ylab = "Residuals", main = paste(p
        " versus residuals", sep = ""))

}
```


sex versus residuals
Residuals
100
50
0
−50
0.0
0.2
0.4
0.6
0.8
1.0
df[, predictor]


status versus residuals
Residuals
100
50
0
−50
20
30
40
50
60
70
df[, predictor]

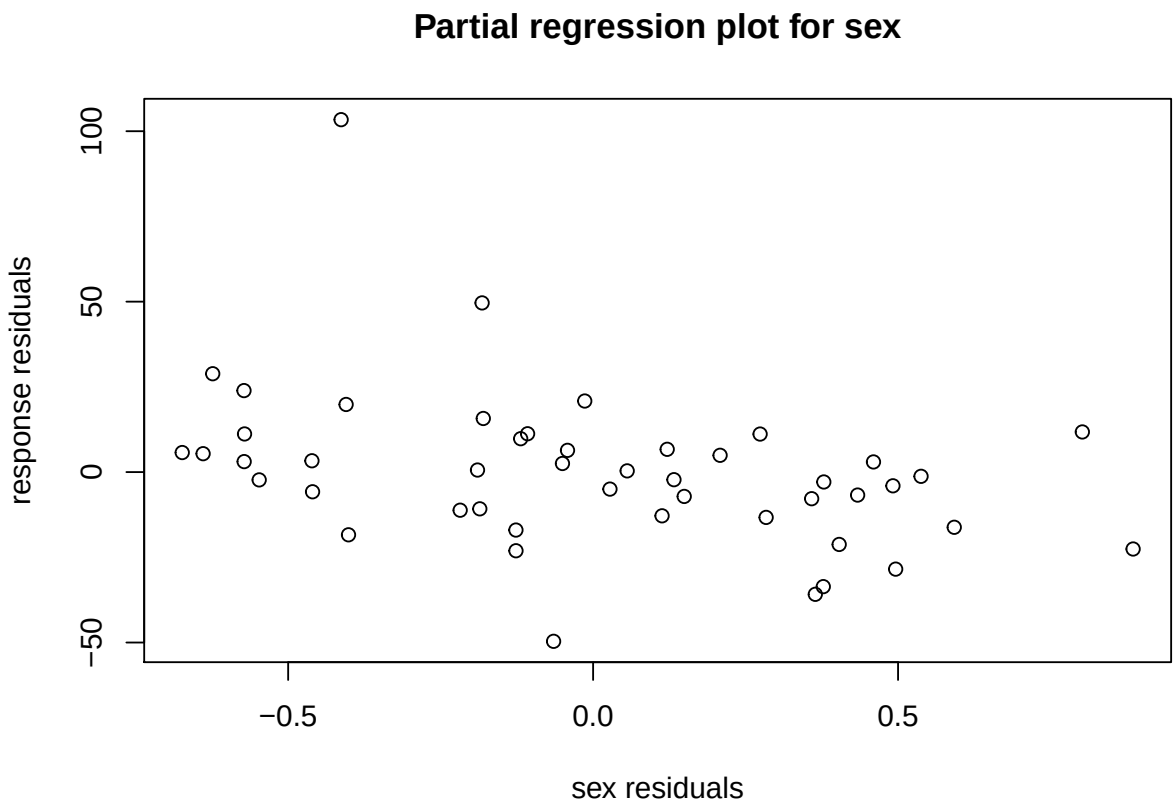

income versus residuals
Residuals
100
50
0
−50
2
4
6
8
10
12
14
df[, predictor]

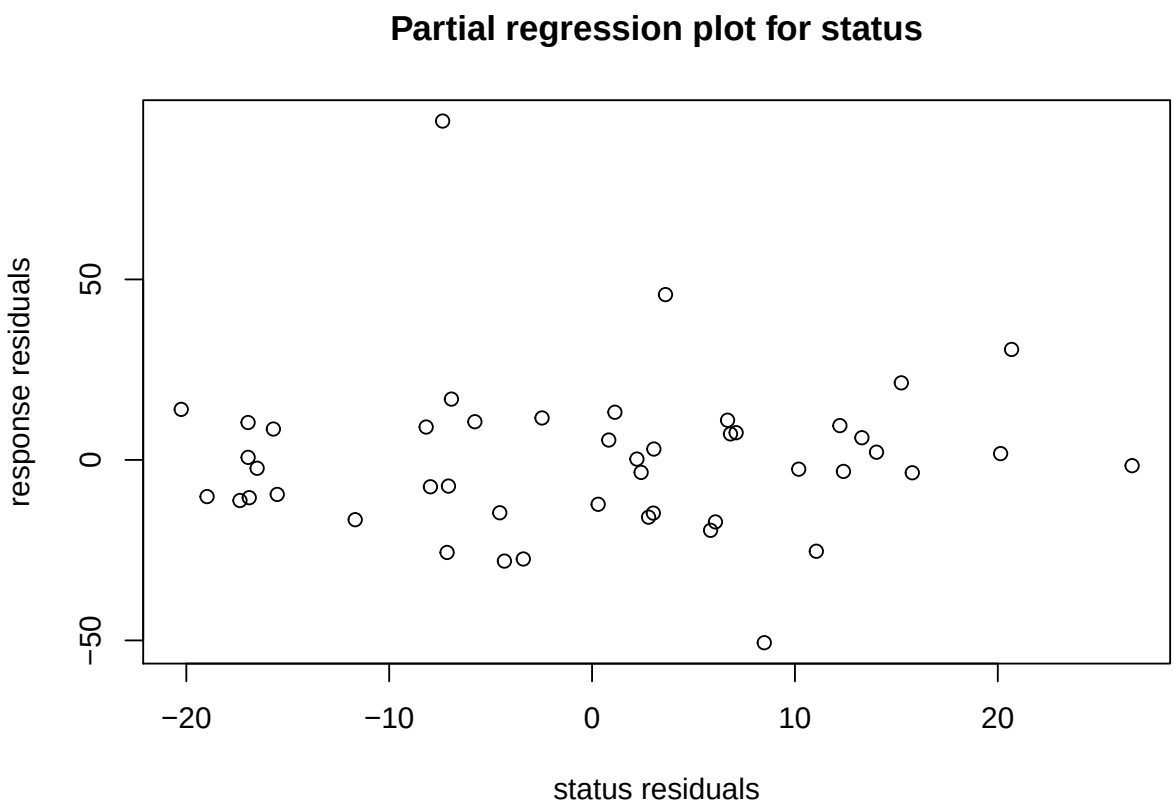


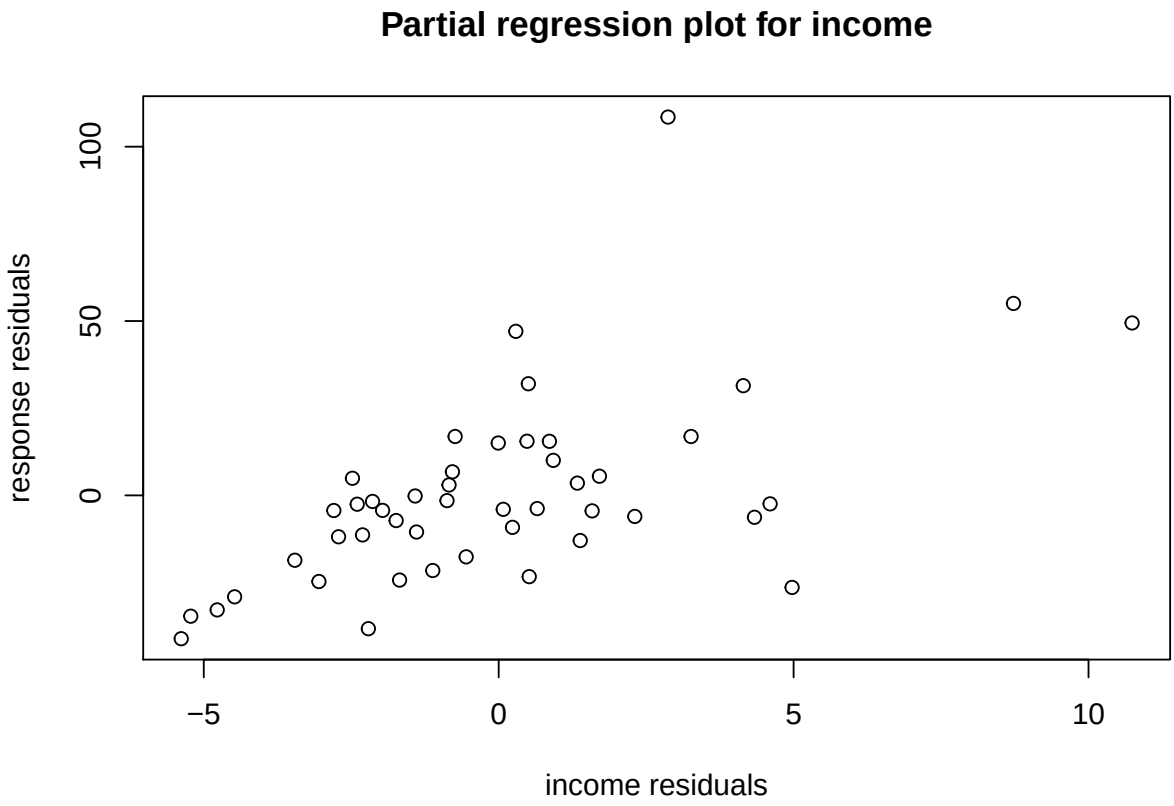
Perform partial regression

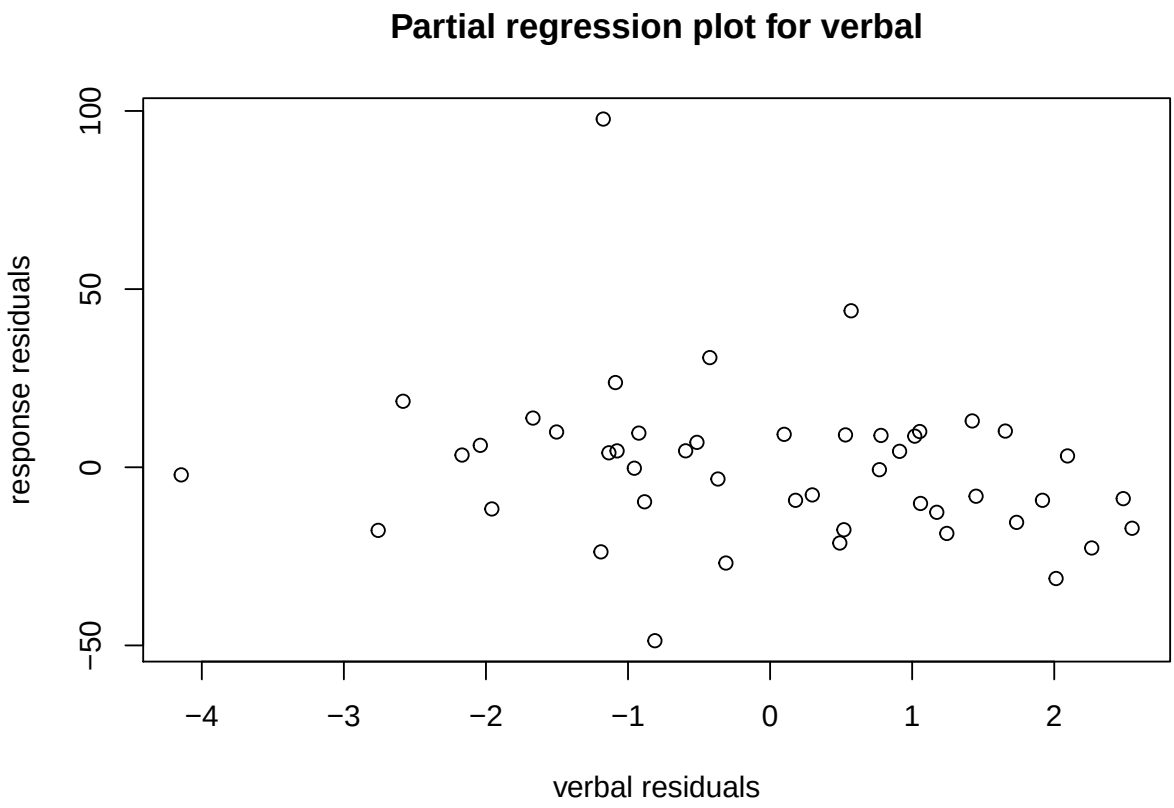
```
predictors <- names(lm.fit$coefficients)
predictors <- predictors[2:length(predictors)]

lm.formula <- formula(lm.fit)
response <- lm.formula[[2]]

for (i in 1:length(predictors)) {
    predictor <- predictors[i]
    others <- predictors[which(predictors != predictor)]
    d.formula <- paste(response, " ~ ", sep = "")
    m.formula <- paste(predictor, " ~ ", sep = "")

    for (j in 1:(length(others) - 1)) {
        d.formula <- paste(d.formula, others[j], " + ", sep = "")
        m.formula <- paste(m.formula, others[j], " + ", sep = "")
    }
    d.formula <- paste(d.formula, others[length(others)], sep = "")
    d.formula <- formula(d.formula)
```

```
    m.formula <- paste(m.formula, others[length(others)], sep = "")
    m.formula <- formula(m.formula)

    d <- residuals(lm(d.formula, df))

    m <- residuals(lm(m.formula, df))

    plot(m, d, xlab = paste(predictor, " residuals", sep = ""), ylab = "response residua
        main = paste("Partial regression plot for ", predictor, sep = ""))

}
```




Partial regression plot for status
response residuals
50
0
−50
−20
−10
0
10
20
status residuals


Partial regression plot for income
response residuals
100
50
0
−5
0
5
10
income residuals



6.3 For the prostate data, fit a model with lpsa as the response and the other variables as predictors.

```
rm(list = ls())
data(prostate, package = "faraway")
lm.fit <- lm(lpsa ~ ., data = prostate)
```

```
df <- prostate
numPredictors <- (ncol(df) - 1)
hatv <- hatvalues(lm.fit)
lev.cut <- (numPredictors + 1) * 2 * 1/nrow(df)
high.leverage <- df[hatv > lev.cut, ]
pander(high.leverage, caption = "High Leverage Data Elements")
```

Table 5: High Leverage Data Elements

	lcavol	lweight	age	lbph	svi	lcp	gleason	pgg45	lpsa
32	0.1823	6.108	65	1.705	0	-1.386	6	0	2.008

	lcavol	lweight	age	lbph	svi	lcp	gleason	pgg45	lpsa
37	1.423	3.657	73	-0.5798	0	1.658	8	15	2.158
41	0.6206	3.142	60	-1.386	0	-1.386	9	80	2.298
74	1.839	3.237	60	0.4383	1	1.179	9	90	3.075
92	2.533	3.678	61	1.348	1	-1.386	7	15	4.13

We've used the rule of thumb that points with a leverage greater than $\frac{2p}{n}$ should be looked at.

(d) Check for outliers.

```
studentized.residuals <- rstudent(lm.fit)
max.residual <- studentized.residuals[which.max(abs(studentized.residuals))]
range.residuals <- range(studentized.residuals)
names(range.residuals) <- c("left", "right")
pander(data.frame(range.residuals = t(range.residuals)), caption = "Range of Studentized
```

Table 6: Range of Studentized residuals

range.residuals.left	range.residuals.right
-2.617	2.554

```
p <- numPredictors + 1
n <- nrow(df)
t.val.alpha <- qt(0.05/(n * 2), n - p - 1)
pander(data.frame(t.val.alpha = t.val.alpha), caption = "Bonferroni corrected t-value")
```

Table 7: Bonferroni corrected t-value

t.val.alpha
-3.607

```
outlier.index <- abs(studentized.residuals) > abs(t.val.alpha)

outliers <- df[outlier.index == TRUE, ]

if (nrow(outliers) >= 1) {
    pander(outliers, caption = "outliers")
```

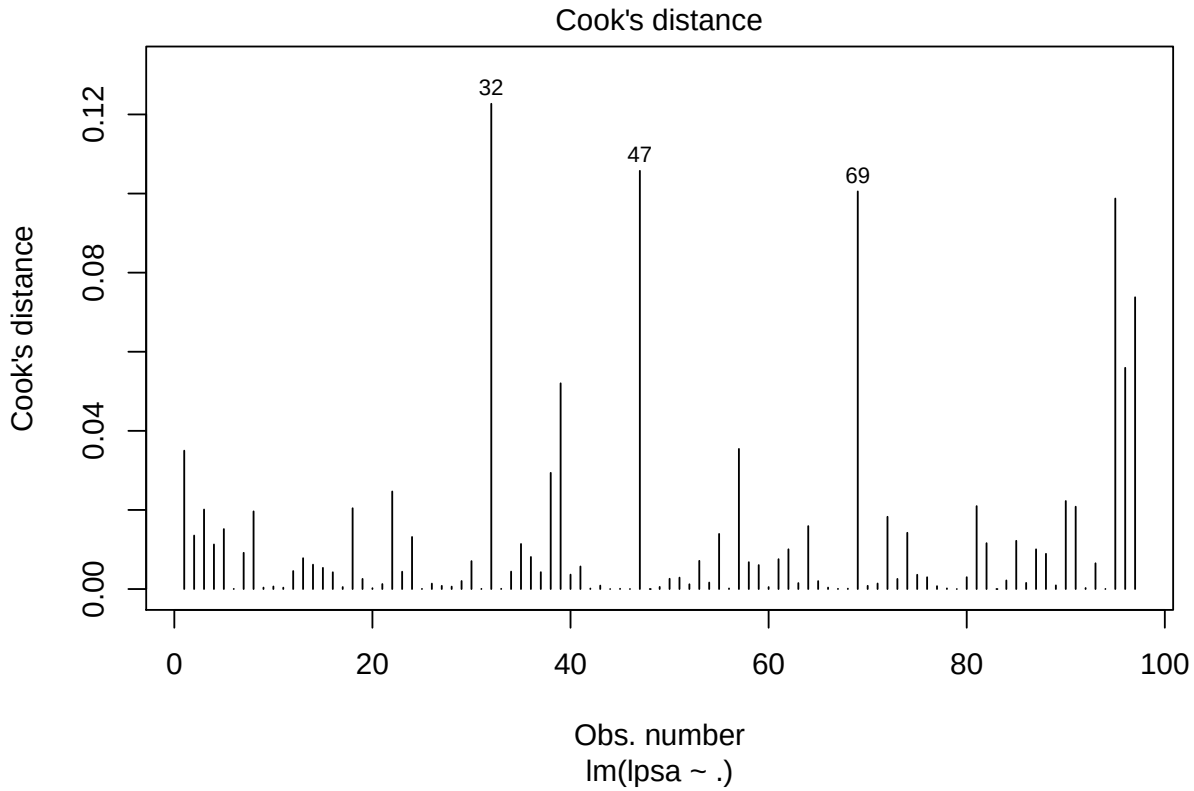
```
}
```

Here we look for studentized residuals that fall outside the interval given by the Bonferroni corrected t-values.

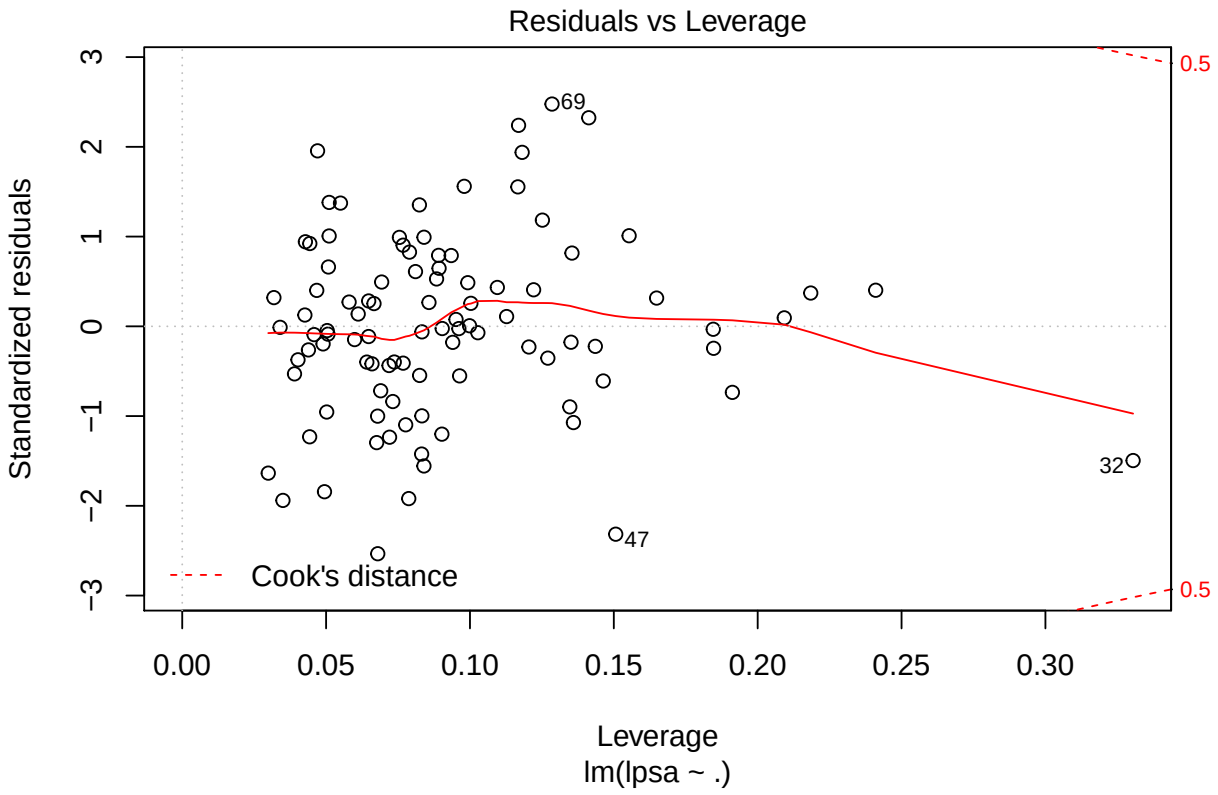
(e) Check for influential points.

We plot the Cook's distances and the residual-leverage plot with level set contours of the Cook distance.

```
plot(lm.fit, which = 4)
```



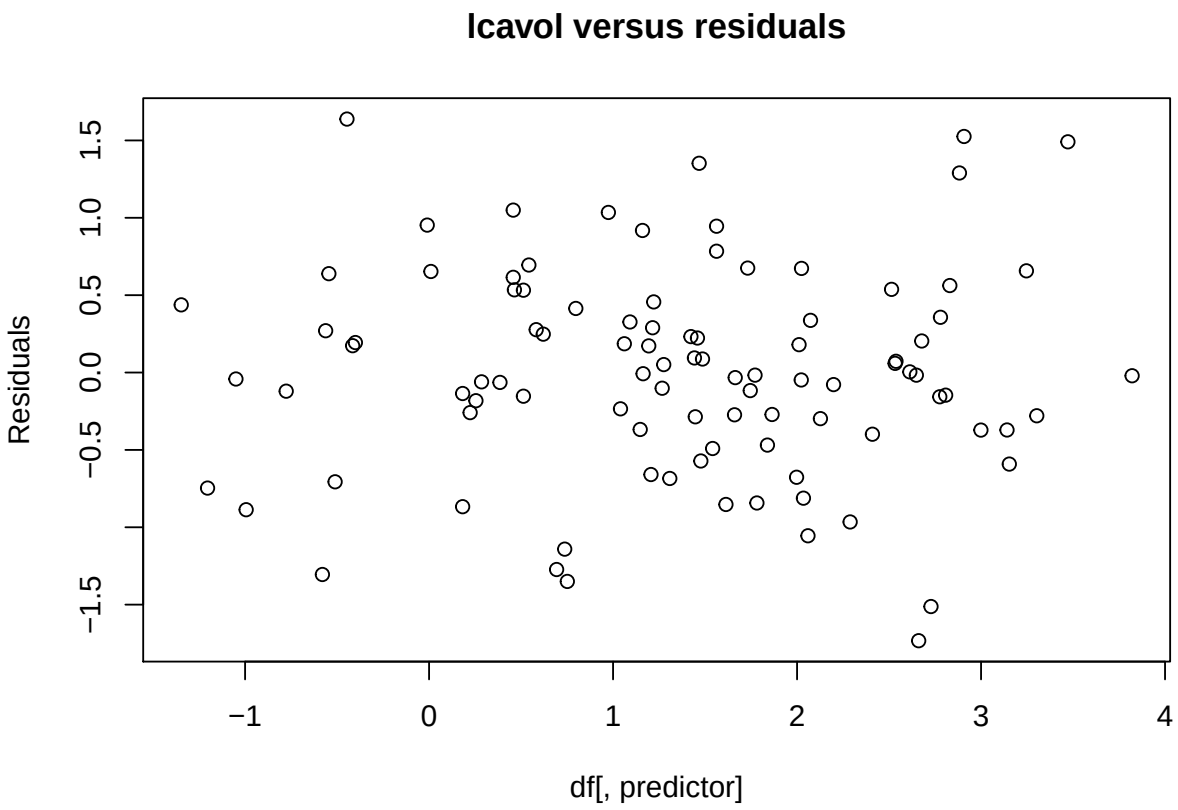

```
plot(lm.fit, which = 5)
```

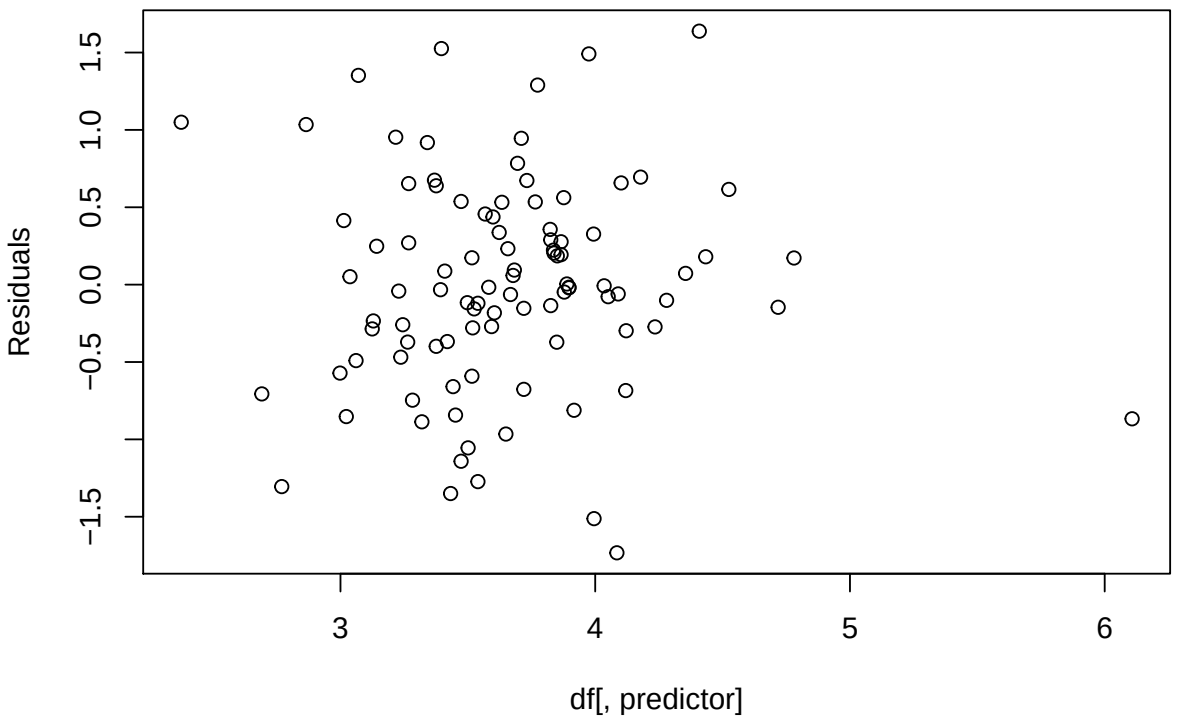


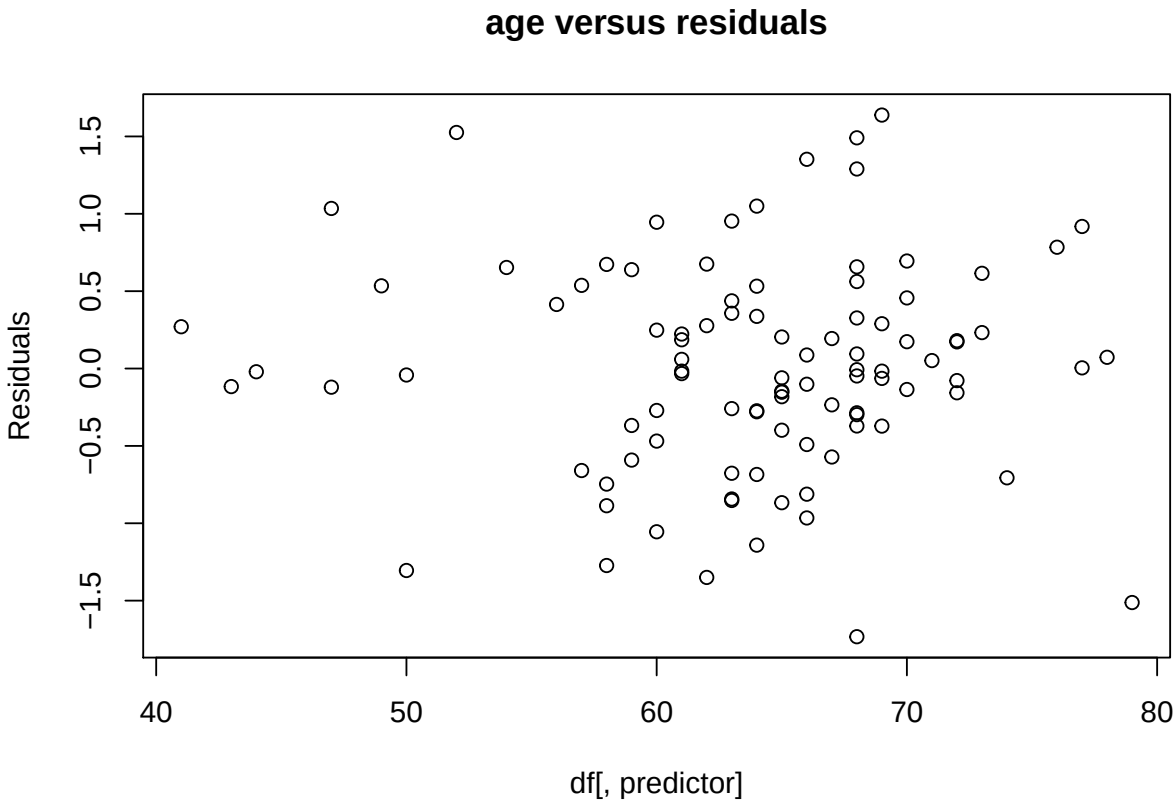
(f) Check for structure in the model.

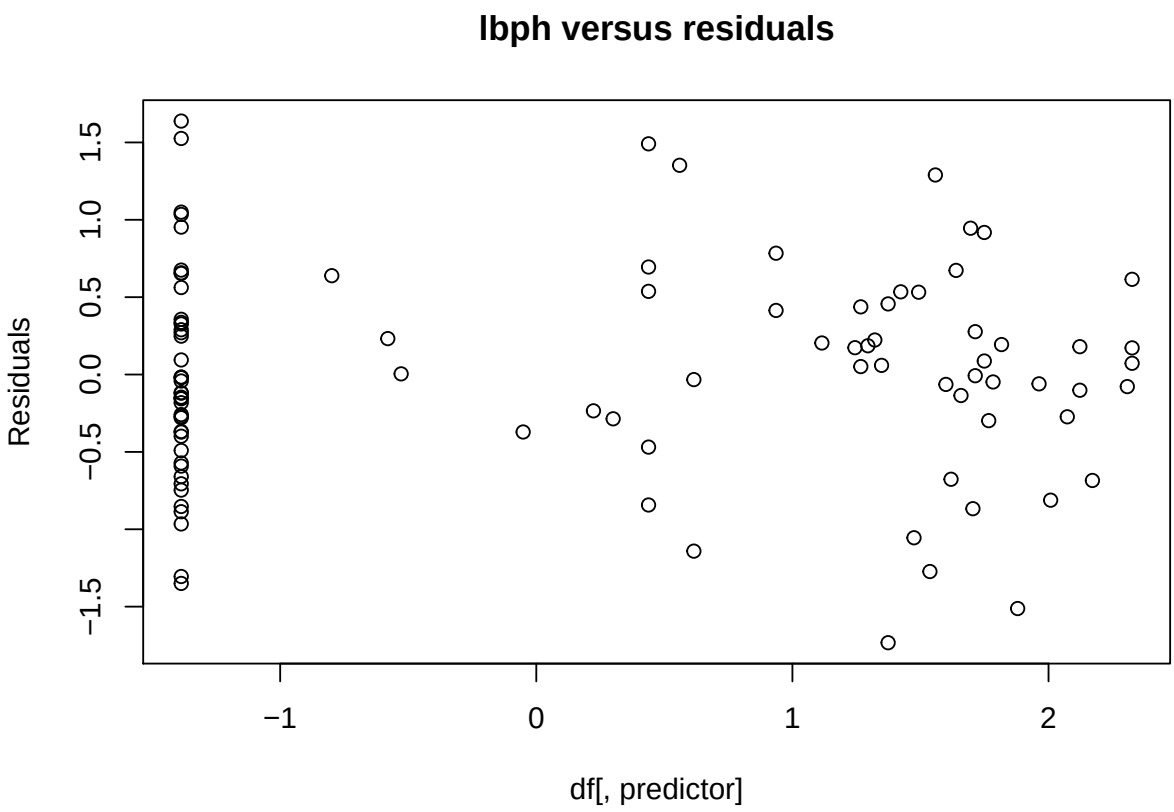
Plot residuals versus predictors

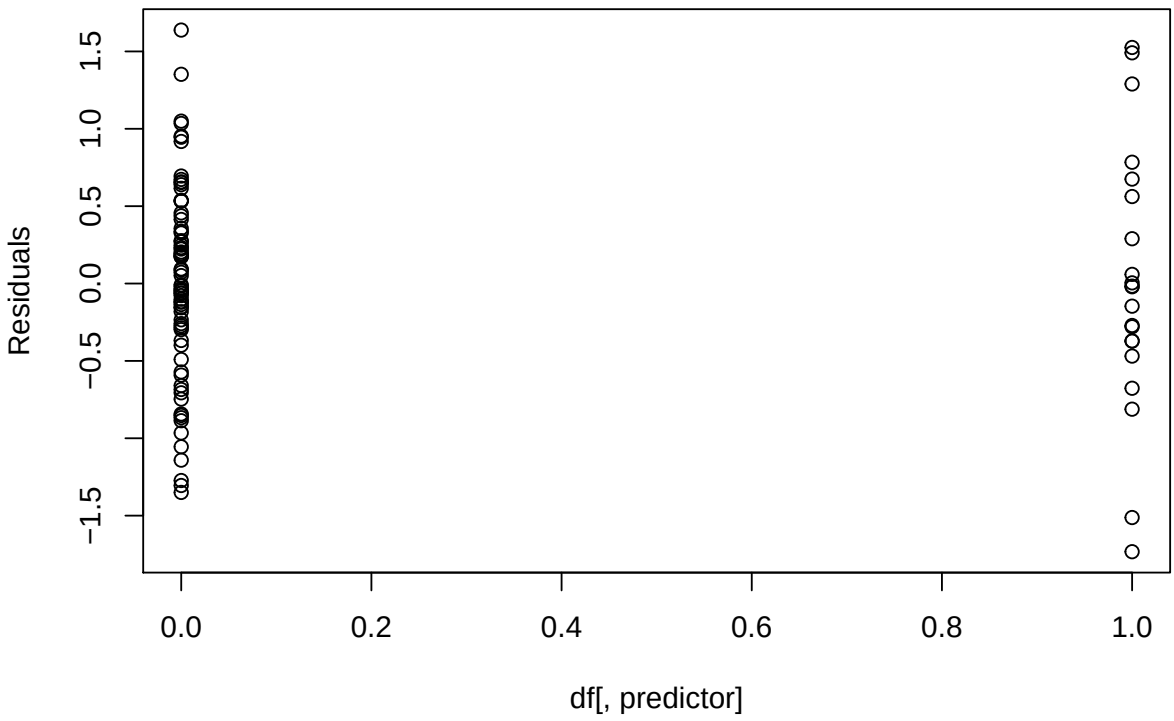
```
predictors <- names(lm.fit$coefficients)
predictors <- predictors[2:length(predictors)]

for (i in 1:length(predictors)) {
    predictor <- predictors[i]

    plot(df[, predictor], residuals(lm.fit), xlab = , ylab = "Residuals", main = paste(p
        " versus residuals", sep = ""))

}
```

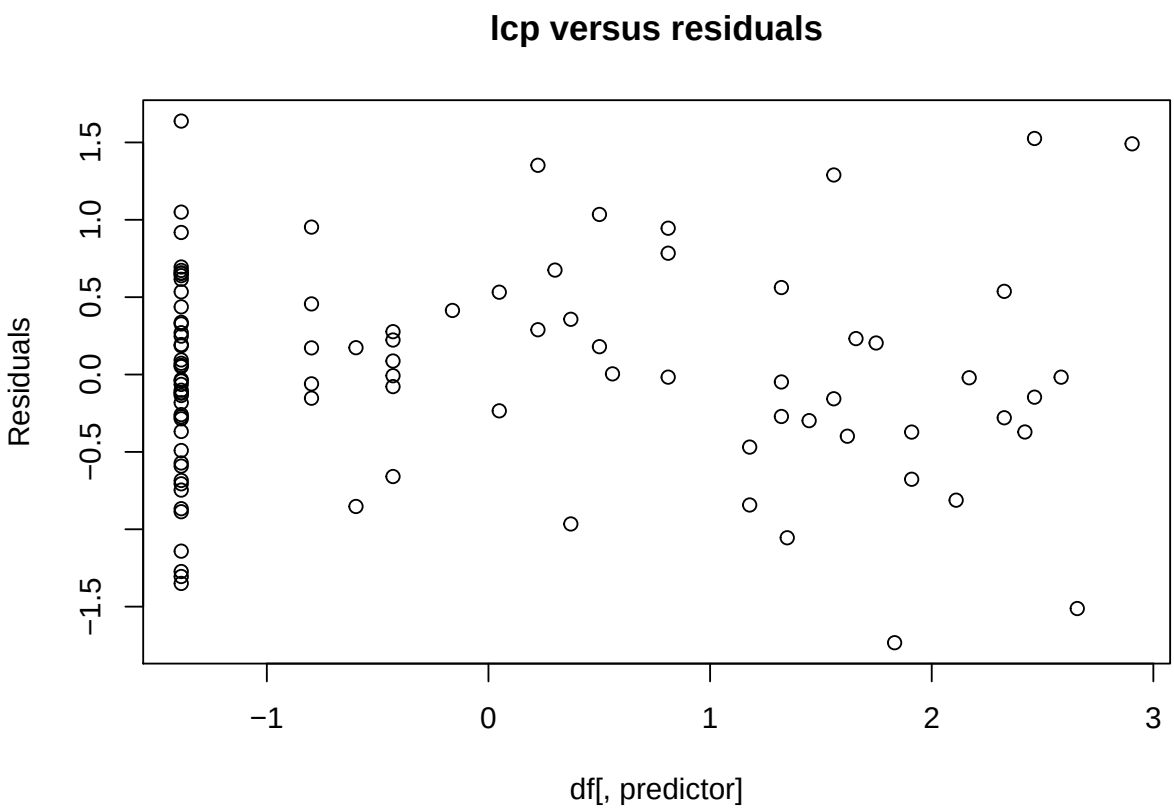

lcavol versus residuals
Residuals
1.5
1.0
0.5
0.0
−0.5
−1.5
−1
0
1
2
3
4
df[, predictor]

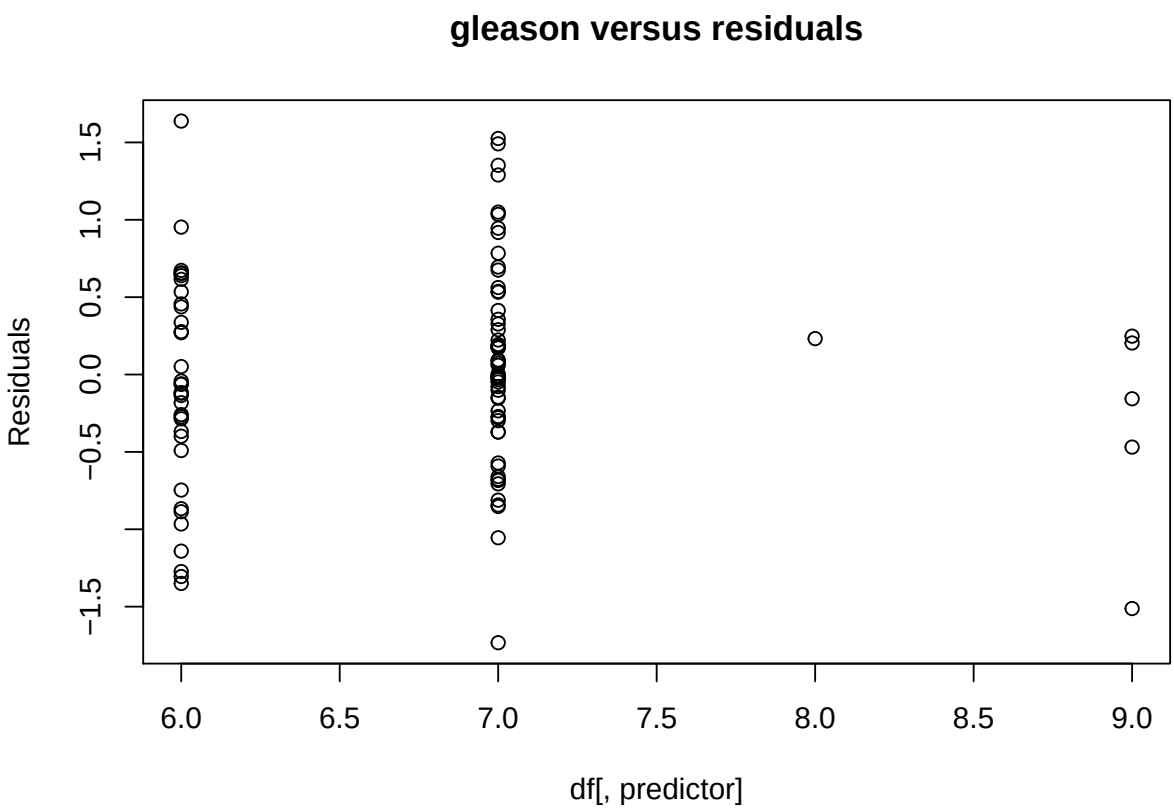

lweight versus residuals
Residuals
1.5
1.0
0.5
0.0
−0.5
−1.5
3
4
5
6
df[, predictor]

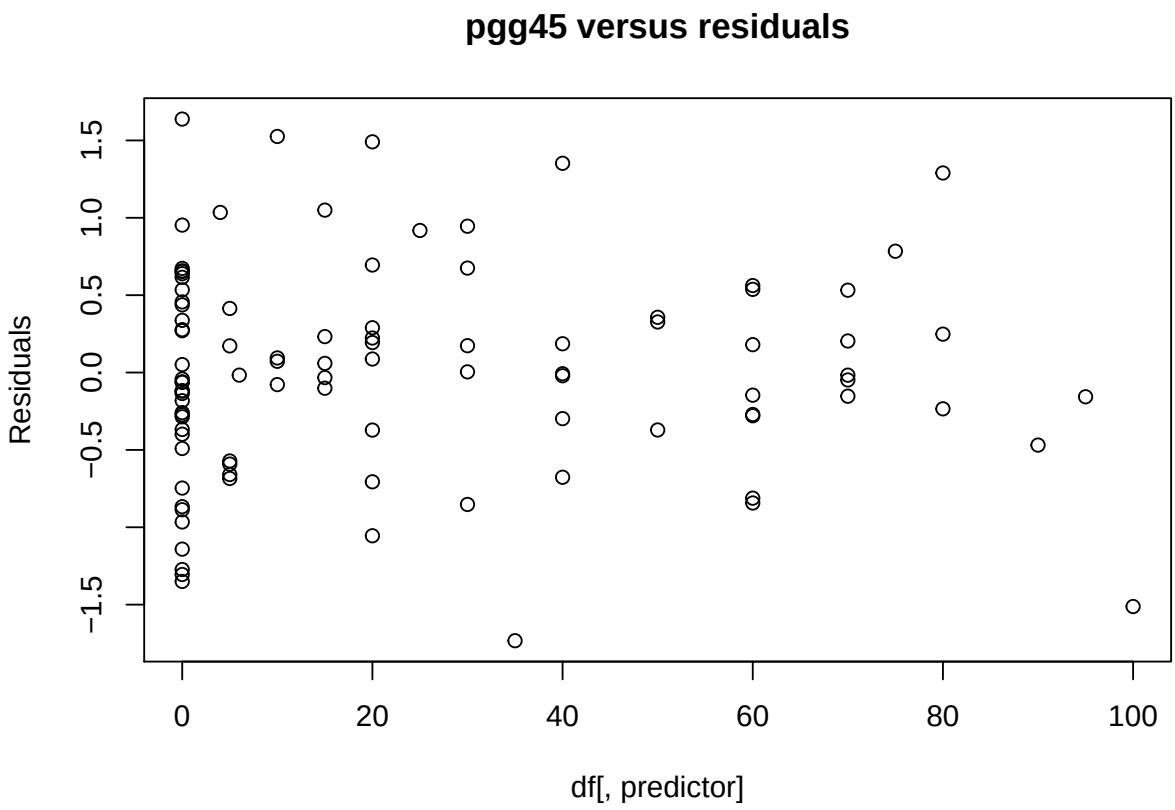

age versus residuals
Residuals
1.5
1.0
0.5
0.0
−0.5
−1.5
40
50
60
70
80
df[, predictor]


lbph versus residuals
Residuals
1.5
1.0
0.5
0.0
−0.5
−1.5
−1
0
1
2
df[, predictor]


svi versus residuals
Residuals
−1.5
−0.5
0.0
0.5
1.0
1.5
0.0
0.2
0.4
0.6
0.8
1.0
df[, predictor]


lcp versus residuals
Residuals
−1.5
−0.5
0.0
0.5
1.0
1.5
−1
0
1
2
3
df[, predictor]

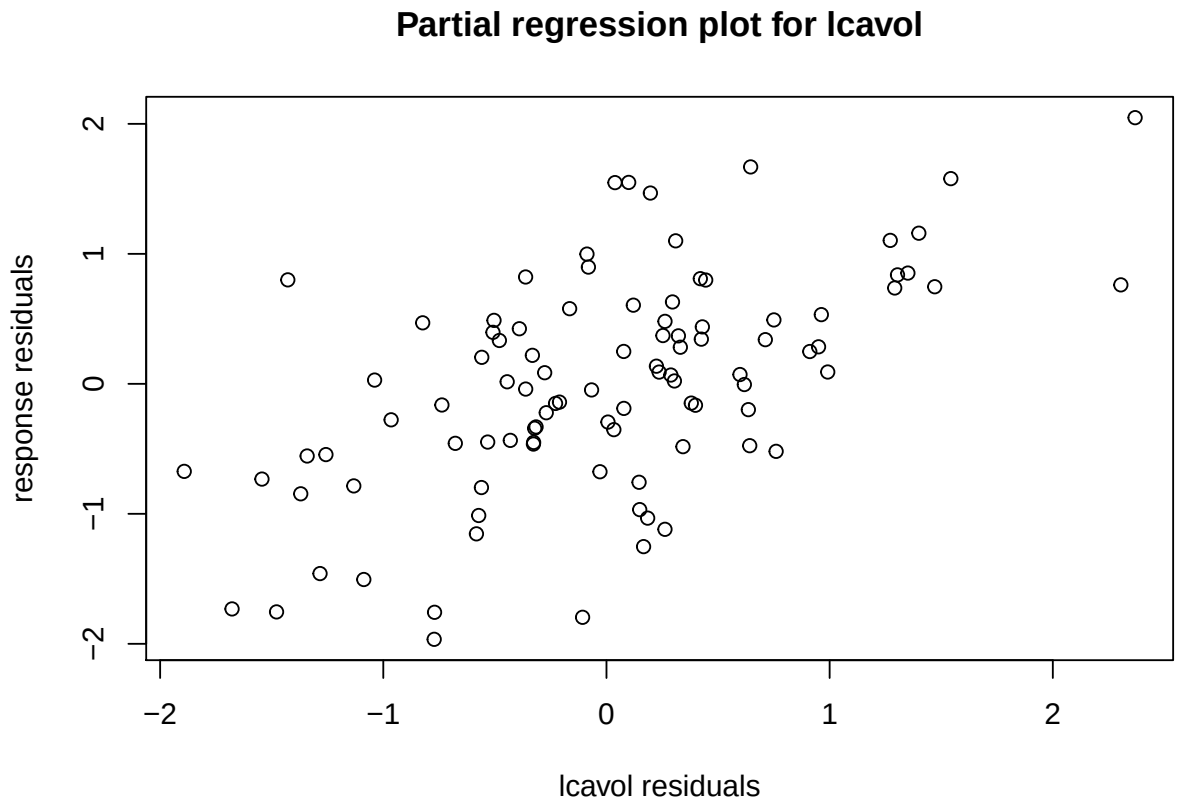

gleason versus residuals
Residuals
1.5
1.0
0.5
0.0
−0.5
−1.5
6.0
6.5
7.0
7.5
8.0
8.5
9.0
df[, predictor]

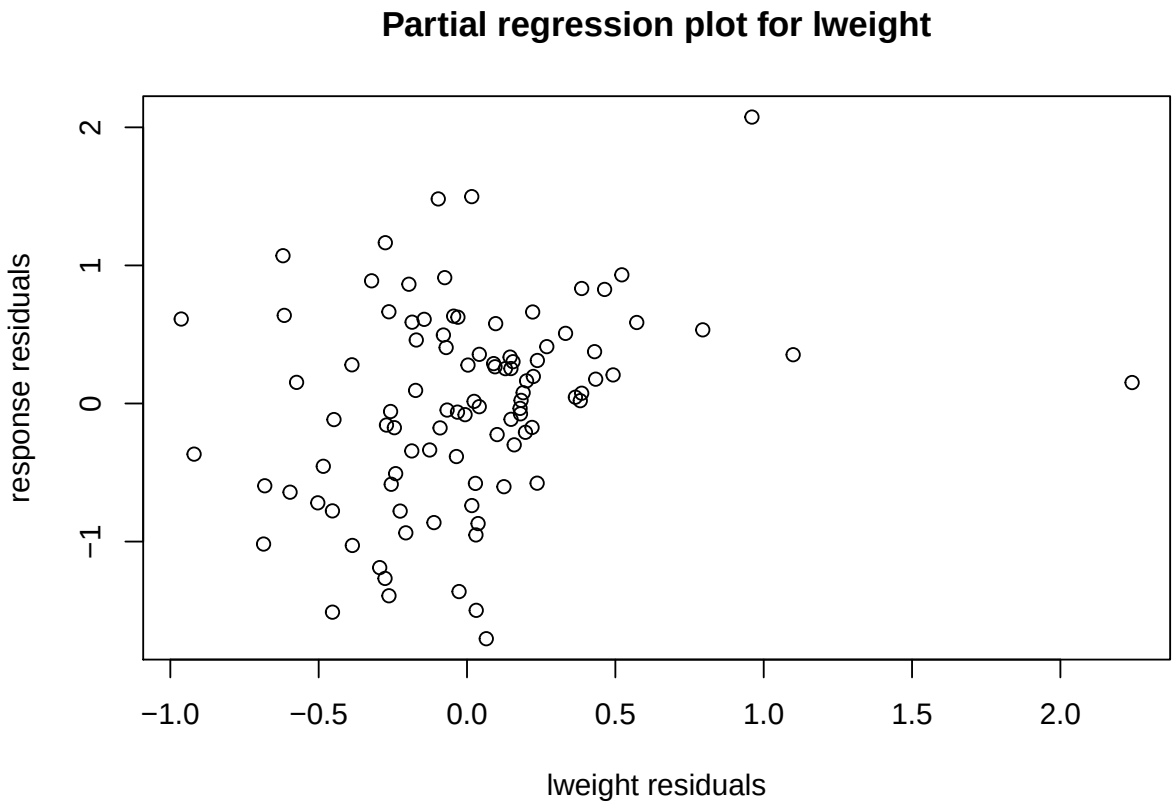


Perform partial regression

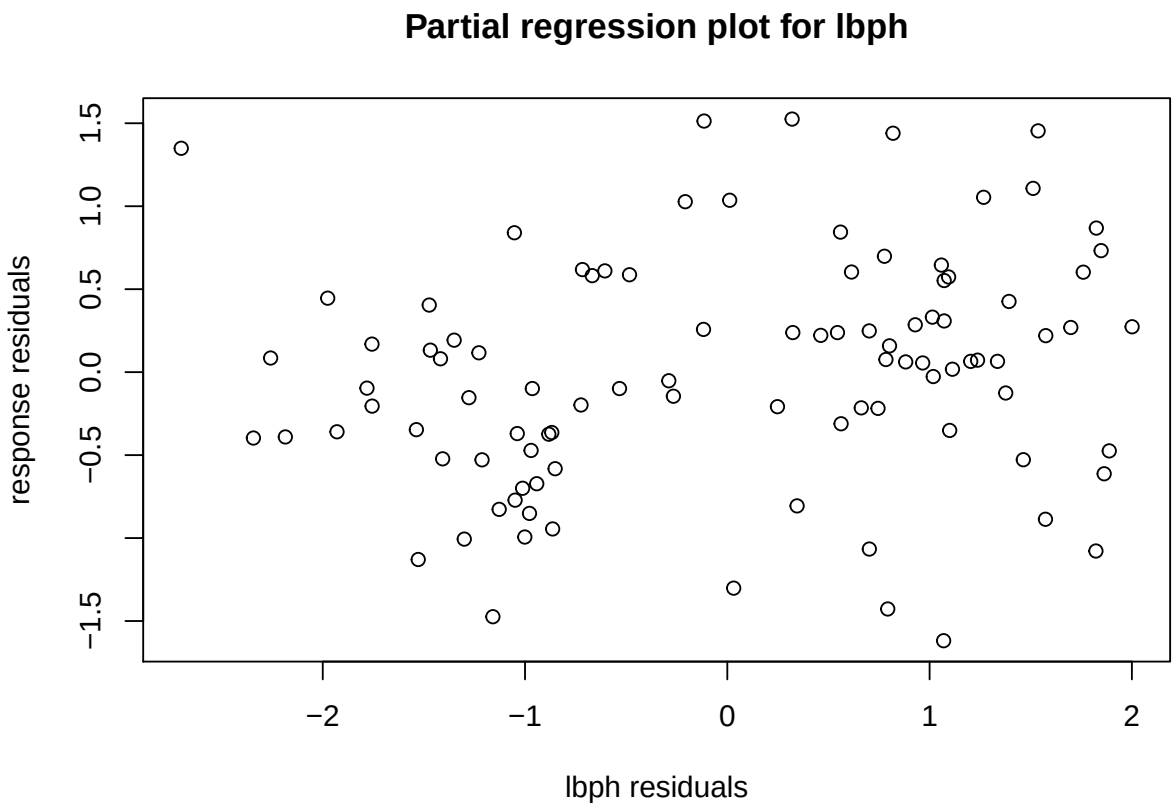
```
predictors <- names(lm.fit$coefficients)
predictors <- predictors[2:length(predictors)]

lm.formula <- formula(lm.fit)
response <- lm.formula[[2]]

for (i in 1:length(predictors)) {
    predictor <- predictors[i]
    others <- predictors[which(predictors != predictor)]
    d.formula <- paste(response, " ~ ", sep = "")
    m.formula <- paste(predictor, " ~ ", sep = "")

    for (j in 1:(length(others) - 1)) {
        d.formula <- paste(d.formula, others[j], " + ", sep = "")
        m.formula <- paste(m.formula, others[j], " + ", sep = "")
    }
    d.formula <- paste(d.formula, others[length(others)], sep = "")
    d.formula <- formula(d.formula)
```

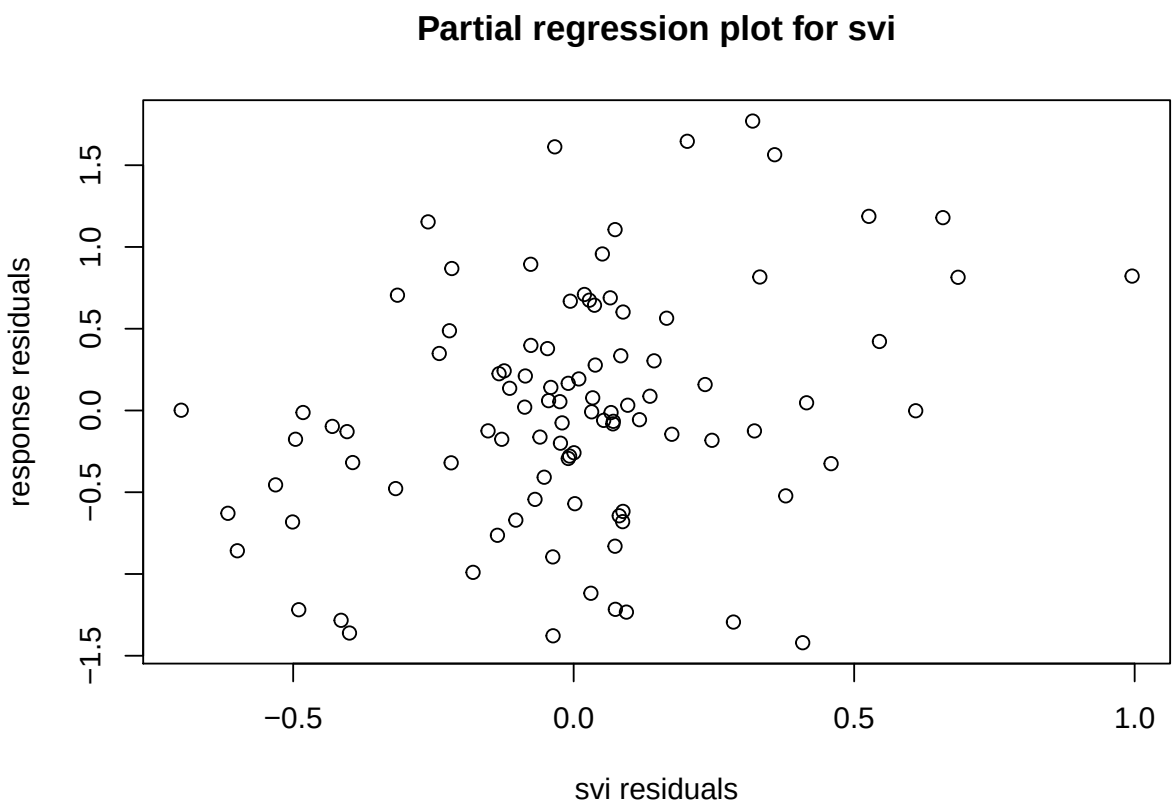
```
    m.formula <- paste(m.formula, others[length(others)], sep = "")
    m.formula <- formula(m.formula)

    d <- residuals(lm(d.formula, df))

    m <- residuals(lm(m.formula, df))

    plot(m, d, xlab = paste(predictor, " residuals", sep = ""), ylab = "response residua
        main = paste("Partial regression plot for ", predictor, sep = ""))

}
```

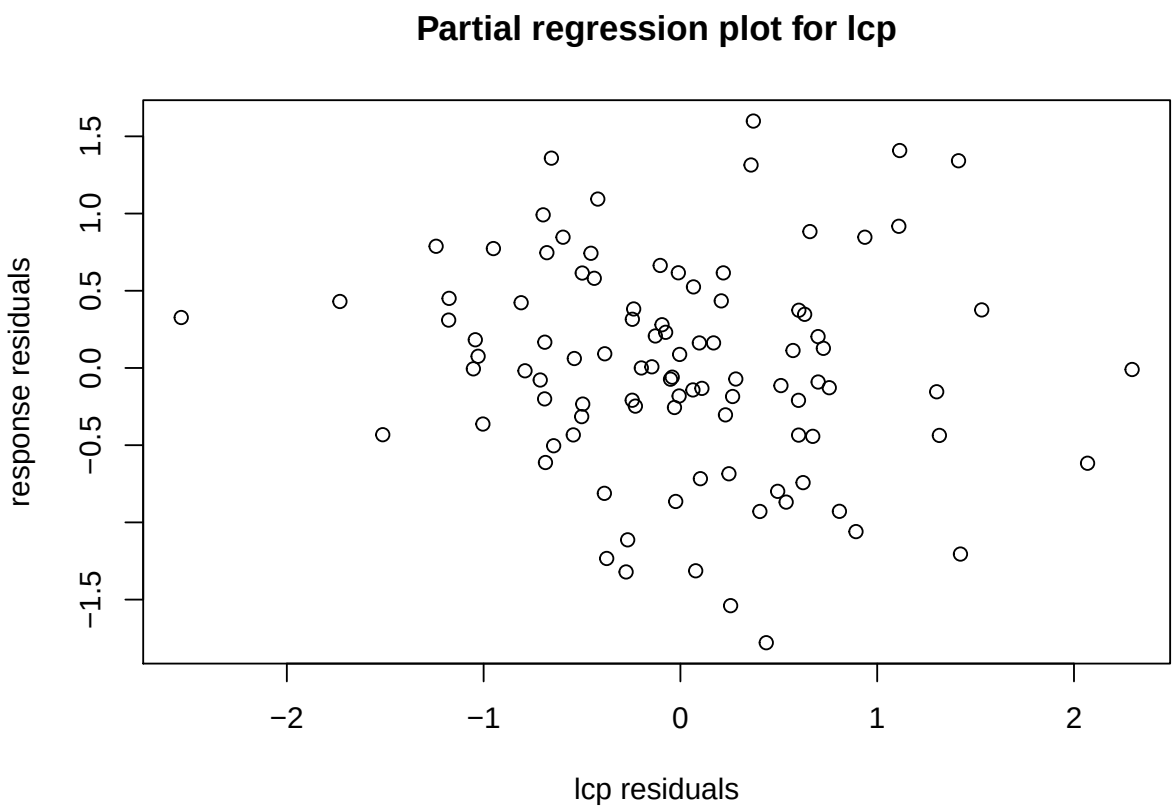


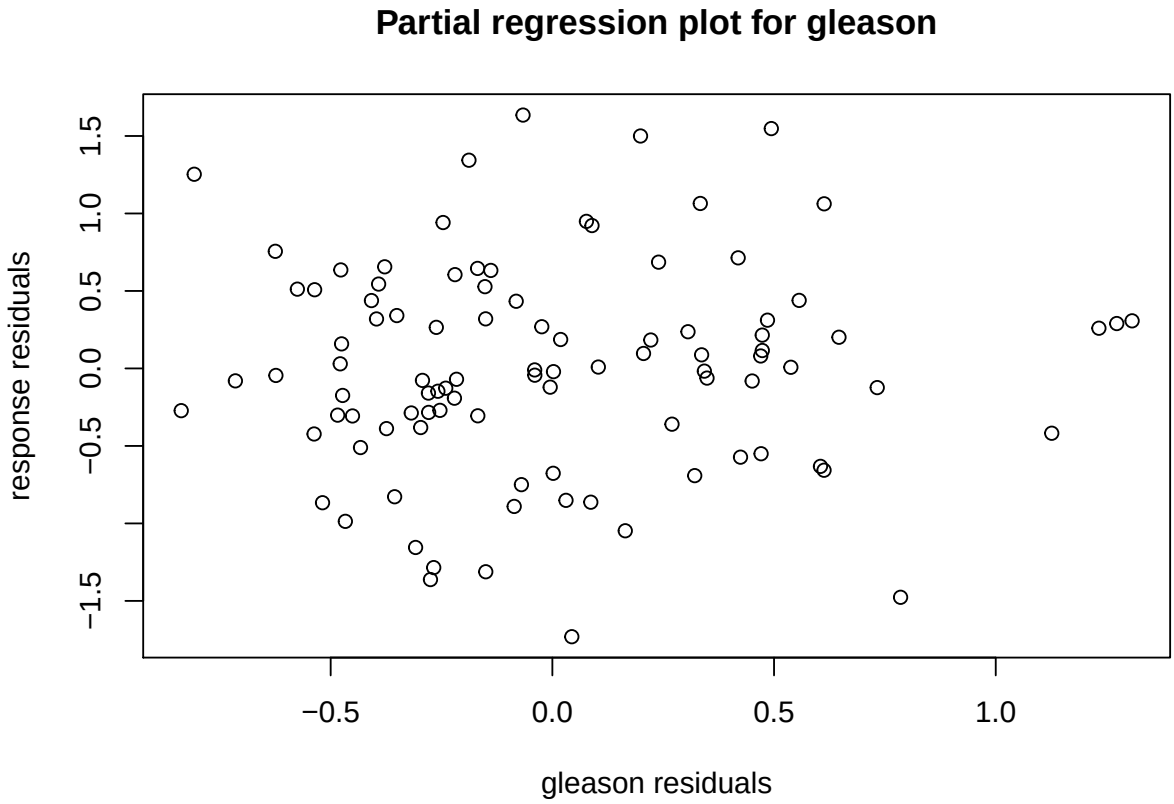

Partial regression plot for lweight
response residuals
2
1
0
−1
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
lweight residuals

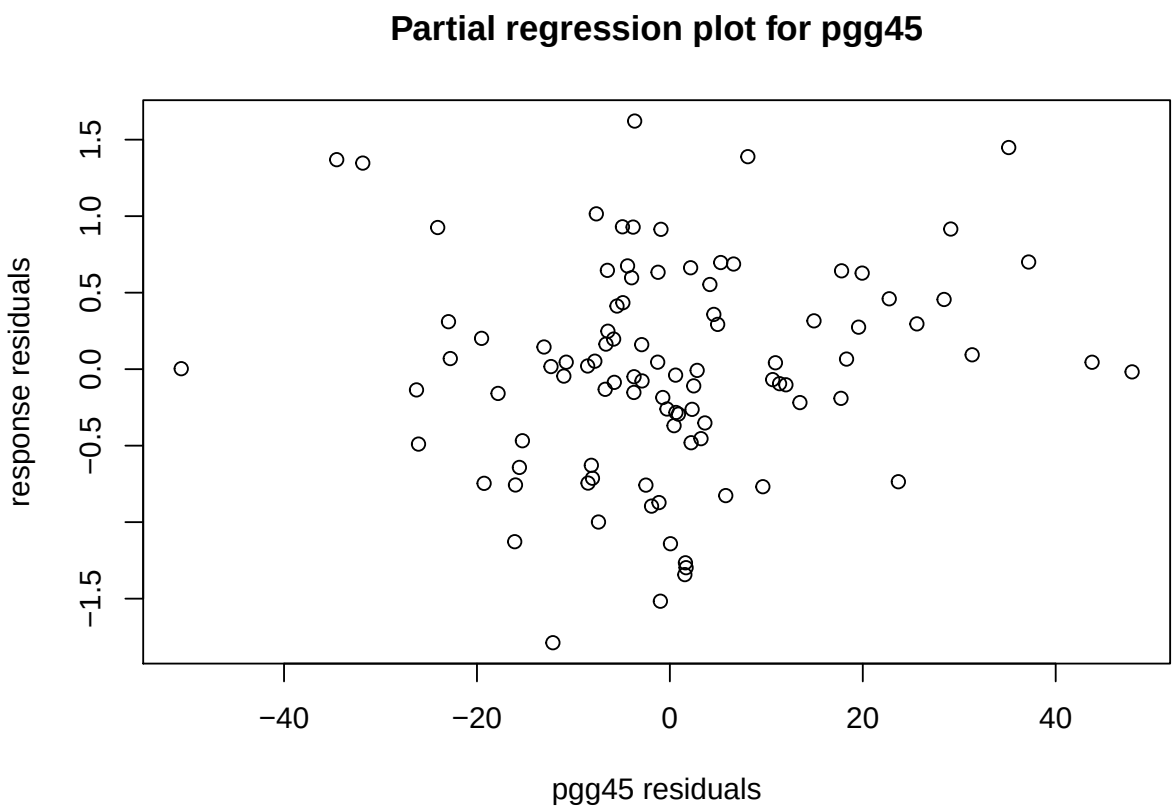

Partial regression plot for age
response residuals
age residuals
−1.5
−0.5
0.0
0.5
1.0
1.5
−20
−15
−10
−5
0
5
10
15


Partial regression plot for lbph
response residuals
lbph residuals
1.5
1.0
0.5
0.0
−0.5
−1.5
−2
−1
0
1
2


Partial regression plot for svi
response residuals
svi residuals
−1.5
−0.5
0.0
0.5
1.0
1.5
−0.5
0.0
0.5
1.0


Partial regression plot for lcp
response residuals
−1.5
−0.5
0.0
0.5
1.0
1.5
−2
−1
0
1
2
lcp residuals


Partial regression plot for gleason
response residuals
1.5
1.0
0.5
0.0
−0.5
−1.5
−0.5
0.0
0.5
1.0
gleason residuals



6.4 For the swiss data, fit a model with Fertility as the response and the other variables as predictors.

```
rm(list = ls())
data(swiss, package = "faraway")
lm.fit <- lm(Fertility ~ ., data = swiss)
```

```
df <- swiss
numPredictors <- (ncol(df) - 1)
hatv <- hatvalues(lm.fit)
lev.cut <- (numPredictors + 1) * 2 * 1/nrow(df)
high.leverage <- df[hatv > lev.cut, ]
pander(high.leverage, caption = "High Leverage Data Elements")
```

Table 8: High Leverage Data Elements (continued below)

	Fertility	Agriculture	Examination	Education
La Vallee	54.3	15.2	31	20
V. De Geneve	35	1.2	37	53

	Fertility	Agriculture	Examination	Education

	Catholic	Infant.Mortality
La Vallee	2.15	10.8
V. De Geneve	42.34	18

We've used the rule of thumb that points with a leverage greater than $\frac{2p}{n}$ should be looked at.

(d) Check for outliers.

```
studentized.residuals <- rstudent(lm.fit)
max.residual <- studentized.residuals[which.max(abs(studentized.residuals))]
range.residuals <- range(studentized.residuals)
names(range.residuals) <- c("left", "right")
pander(data.frame(range.residuals = t(range.residuals)), caption = "Range of Studentized
```

Table 10: Range of Studentized residuals

range.residuals.left	range.residuals.right
-2.394	2.445

```
p <- numPredictors + 1
n <- nrow(df)
t.val.alpha <- qt(0.05/(n * 2), n - p - 1)
pander(data.frame(t.val.alpha = t.val.alpha), caption = "Bonferroni corrected t-value")
```

Table 11: Bonferroni corrected t-value

t.val.alpha
-3.529

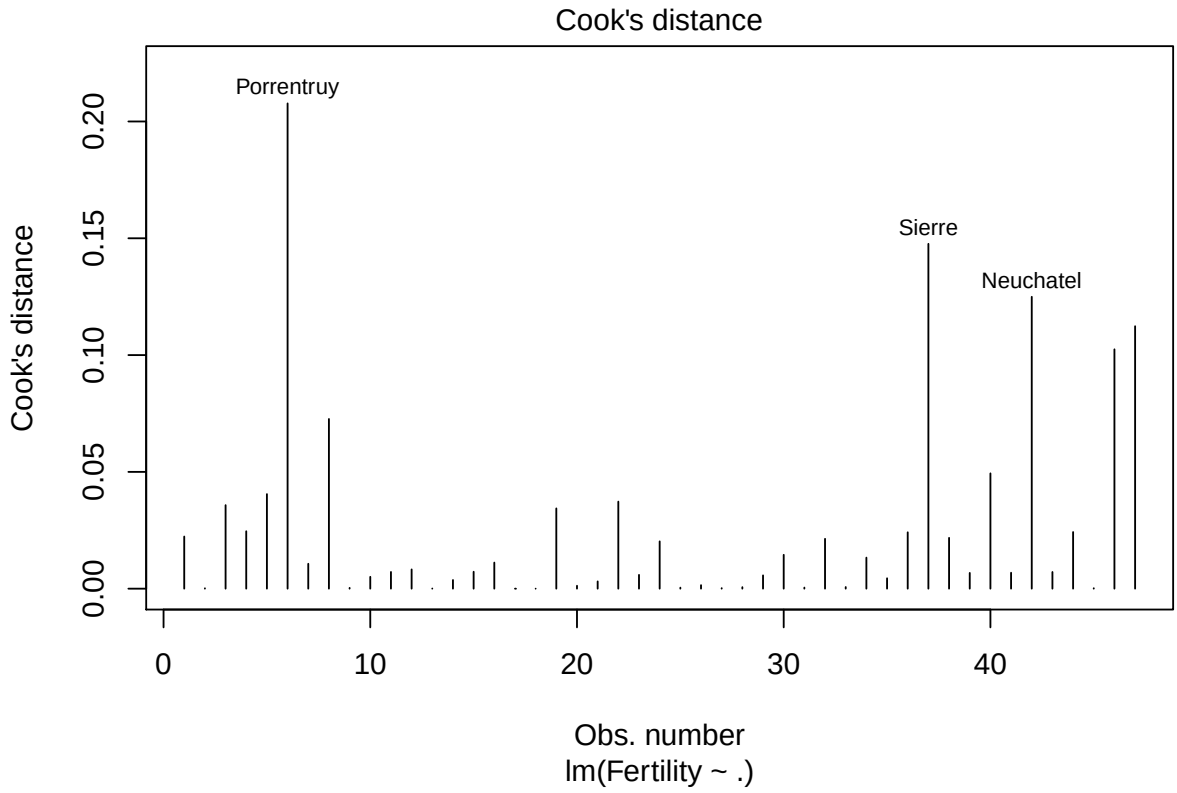
```
outlier.index <- abs(studentized.residuals) > abs(t.val.alpha)

outliers <- df[outlier.index == TRUE, ]

if (nrow(outliers) >= 1) {
    pander(outliers, caption = "outliers")
}
```

Here we look for studentized residuals that fall outside the interval given by the Bonferroni corrected t-values.

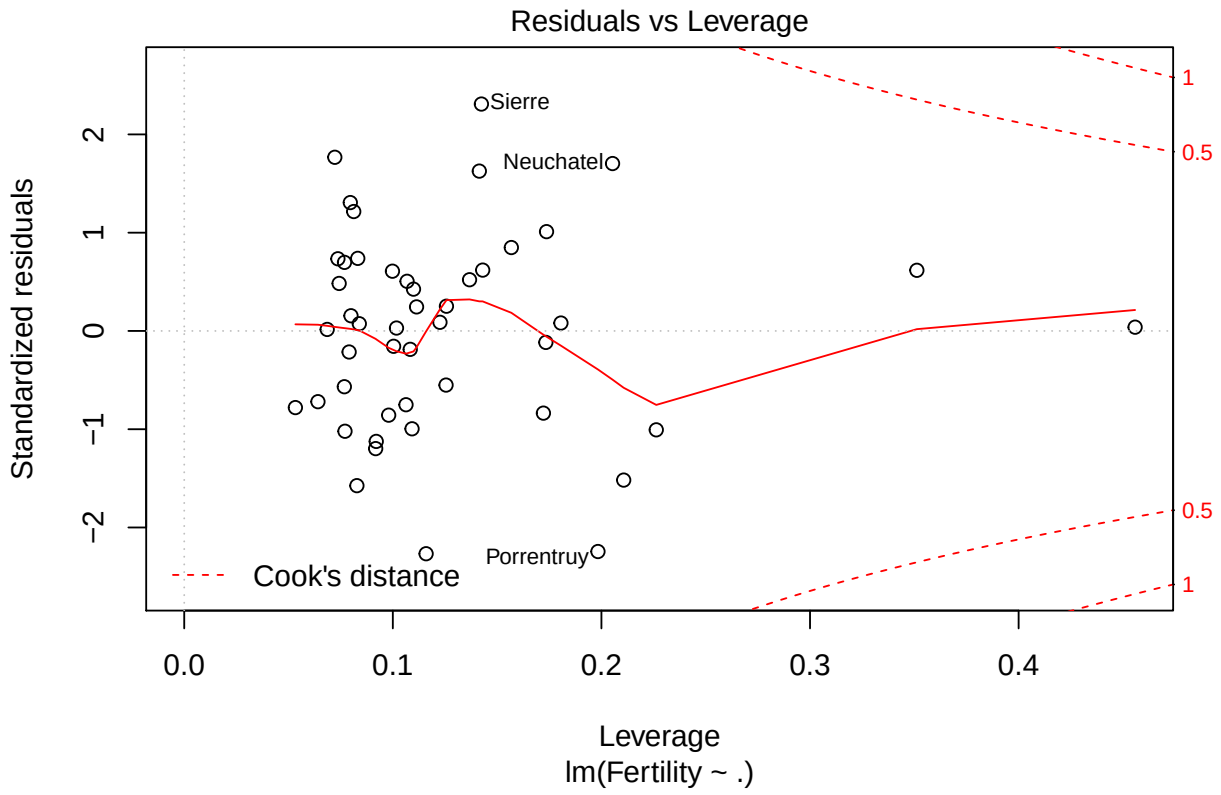
(e) Check for influential points.

We plot the Cook's distances and the residual-leverage plot with level set contours of the Cook distance.

```
plot(lm.fit, which = 4)
```



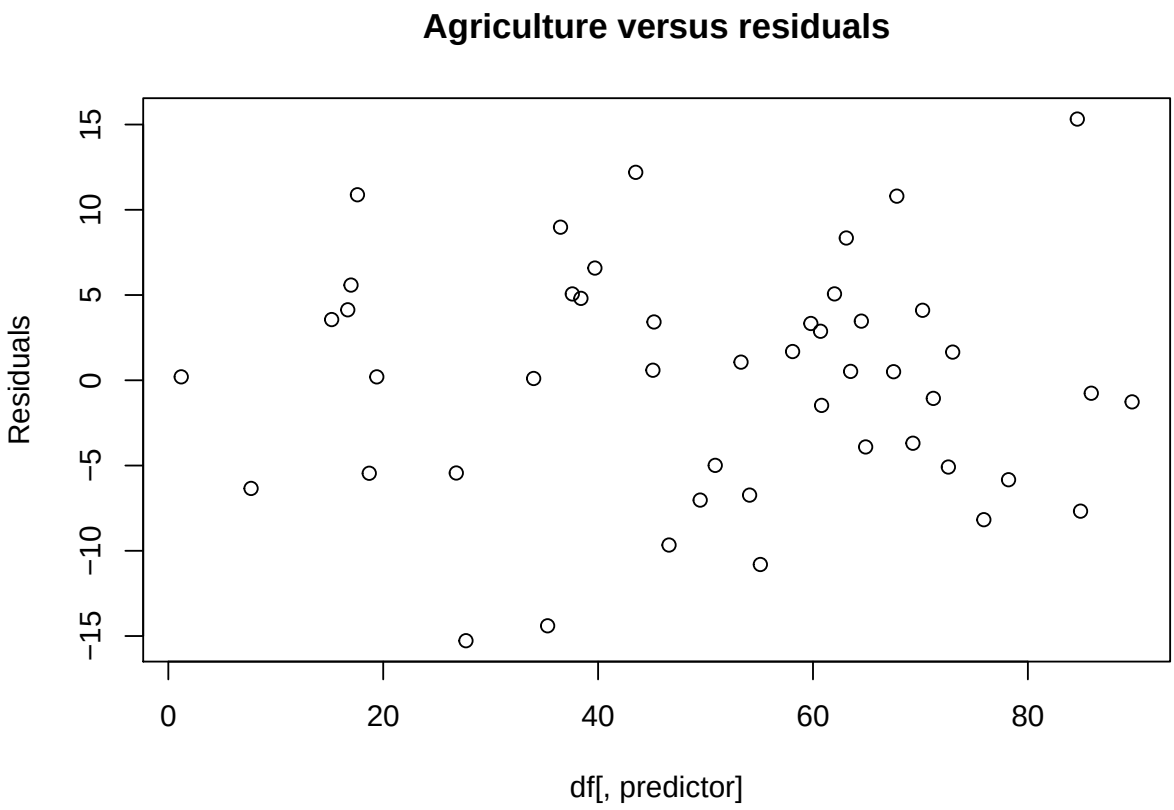

```
plot(lm.fit, which = 5)
```

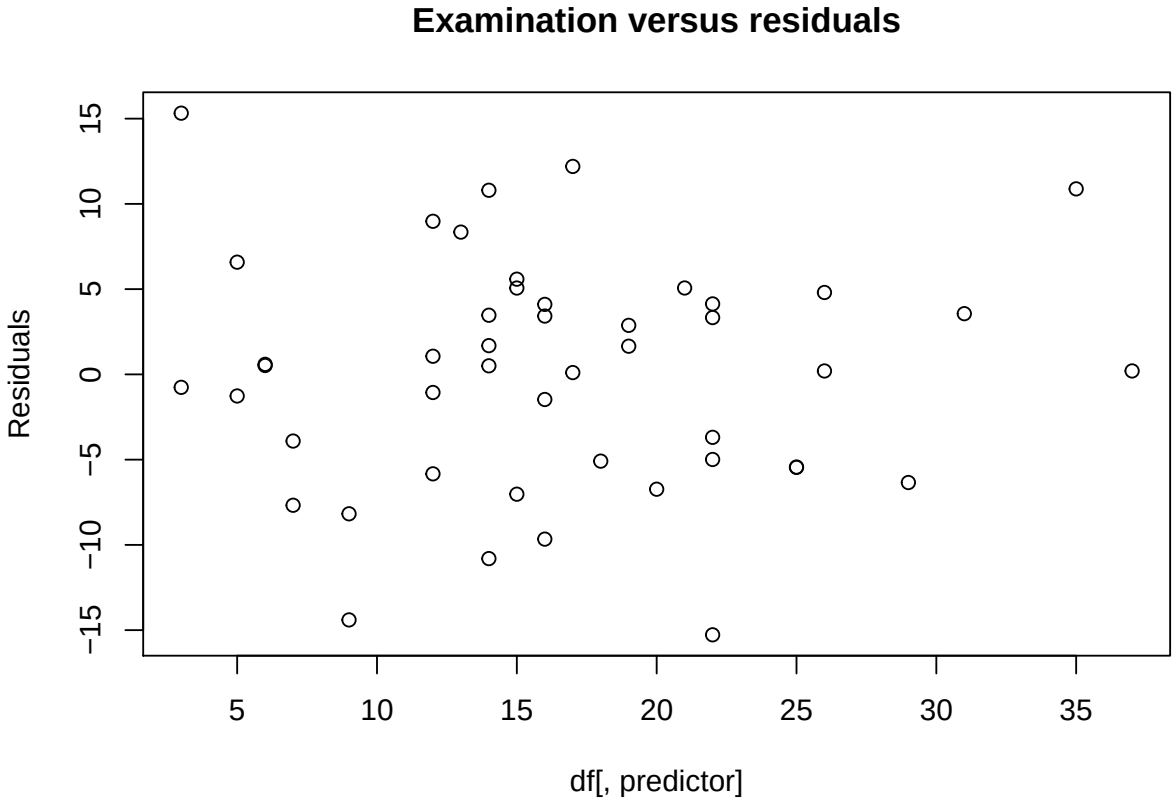


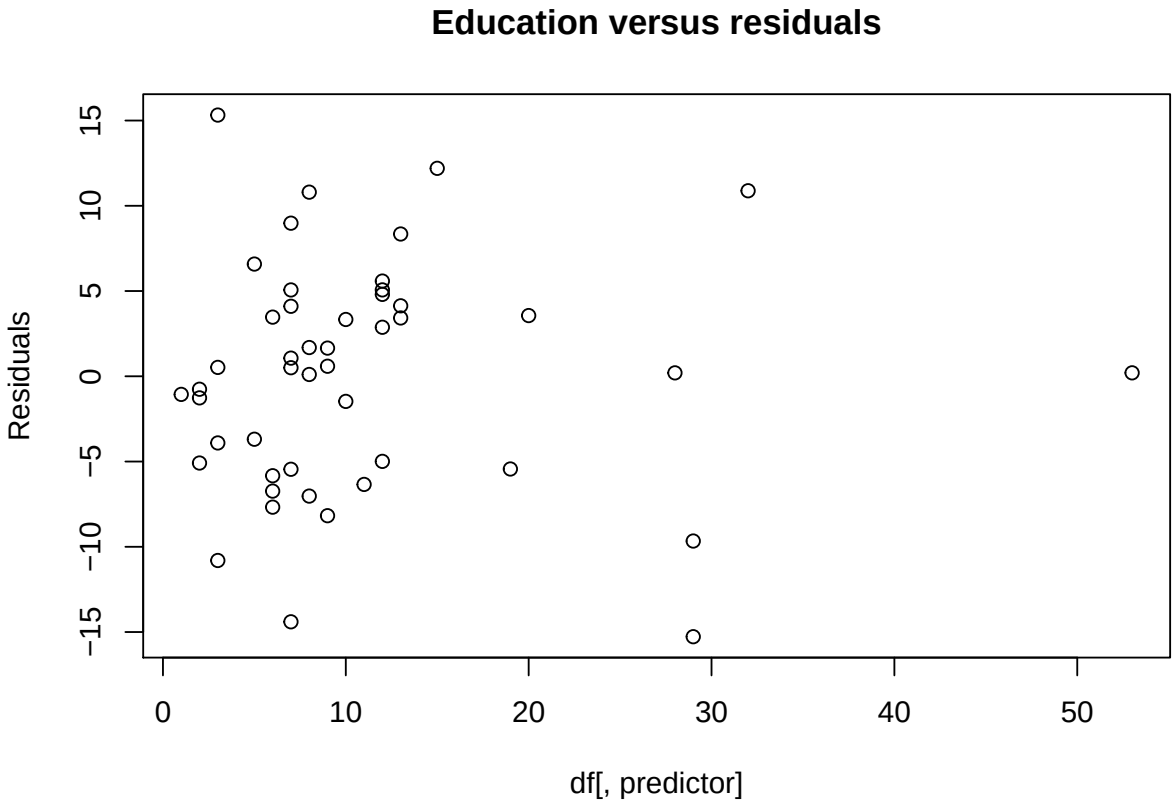
(f) Check for structure in the model.

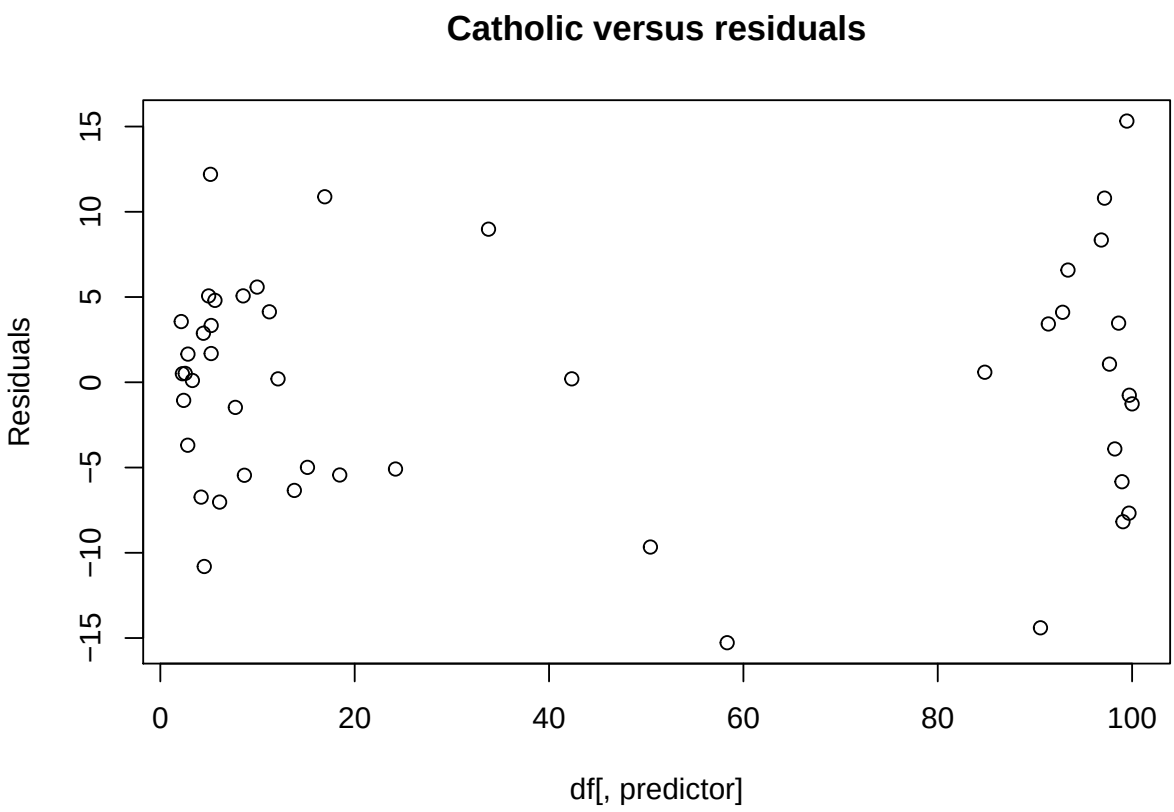
Plot residuals versus predictors

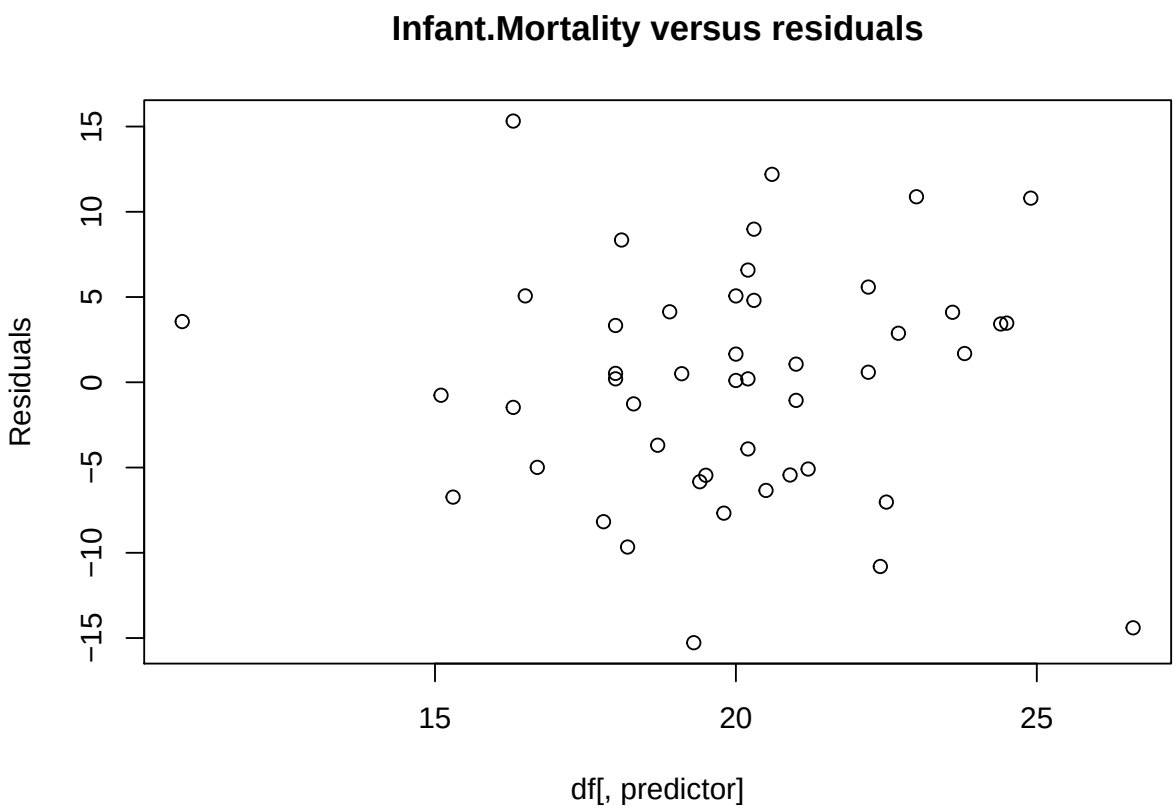
```
predictors <- names(lm.fit$coefficients)
predictors <- predictors[2:length(predictors)]

for (i in 1:length(predictors)) {
    predictor <- predictors[i]

    plot(df[, predictor], residuals(lm.fit), xlab = , ylab = "Residuals", main = paste(p
        " versus residuals", sep = ""))

}
```


Agriculture versus residuals
Residuals
15
10
5
0
−5
−10
−15
0
20
40
60
80
df[, predictor]


Examination versus residuals
Residuals
15
10
5
0
−5
−10
−15
5
10
15
20
25
30
35
df[, predictor]


Education versus residuals
Residuals
15
10
5
0
−5
−10
−15
0
10
20
30
40
50
df[, predictor]

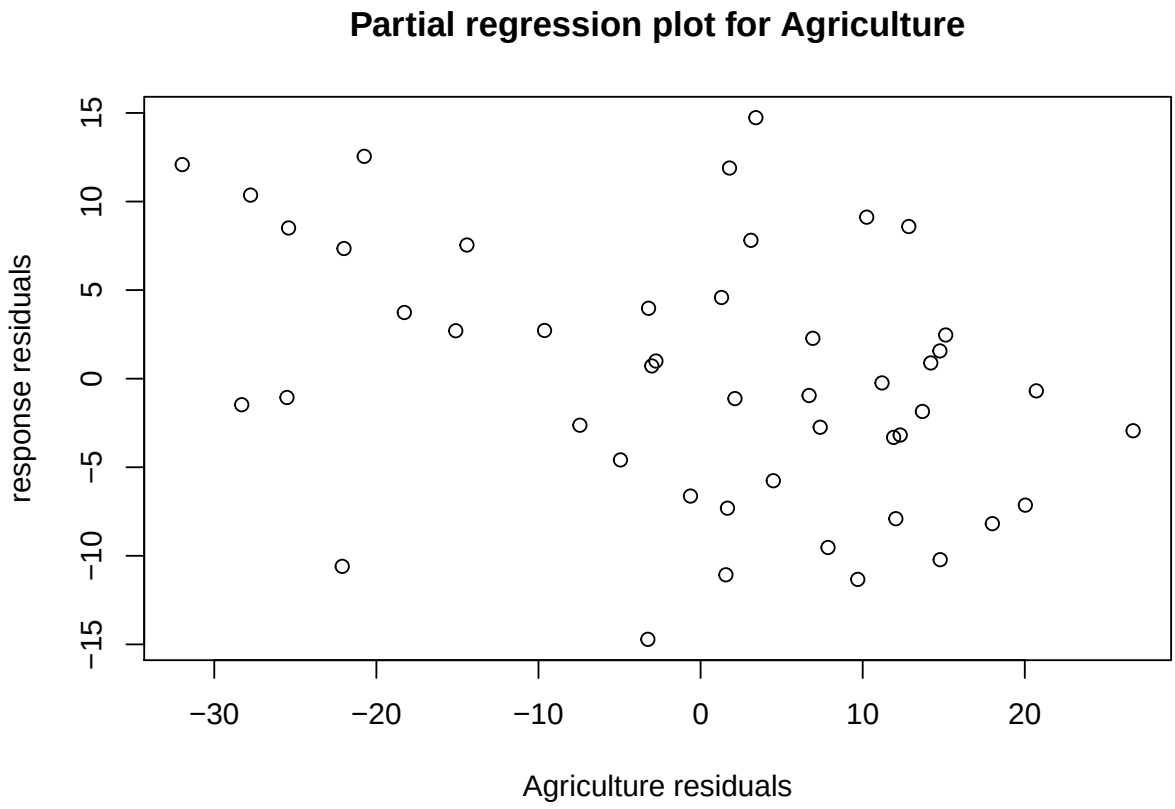

Catholic versus residuals
Residuals
15
10
5
0
−5
−10
−15
0
20
40
60
80
100
df[, predictor]

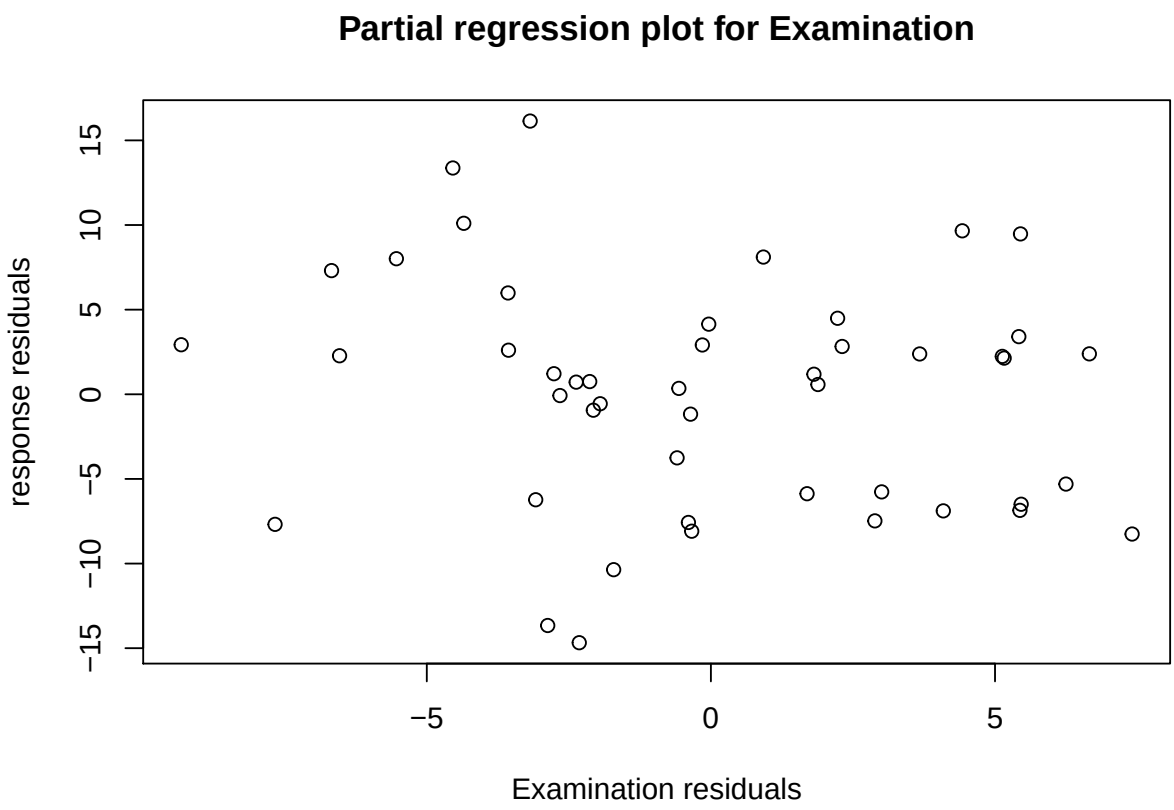


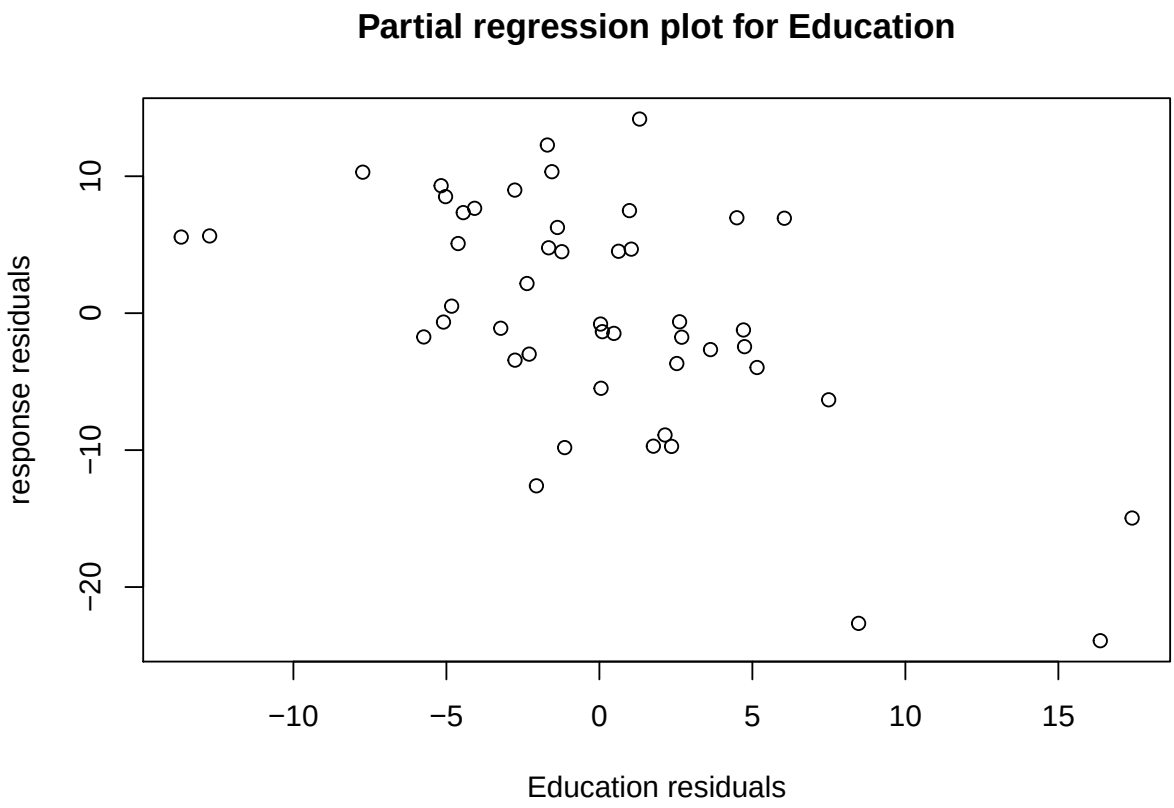
Perform partial regression

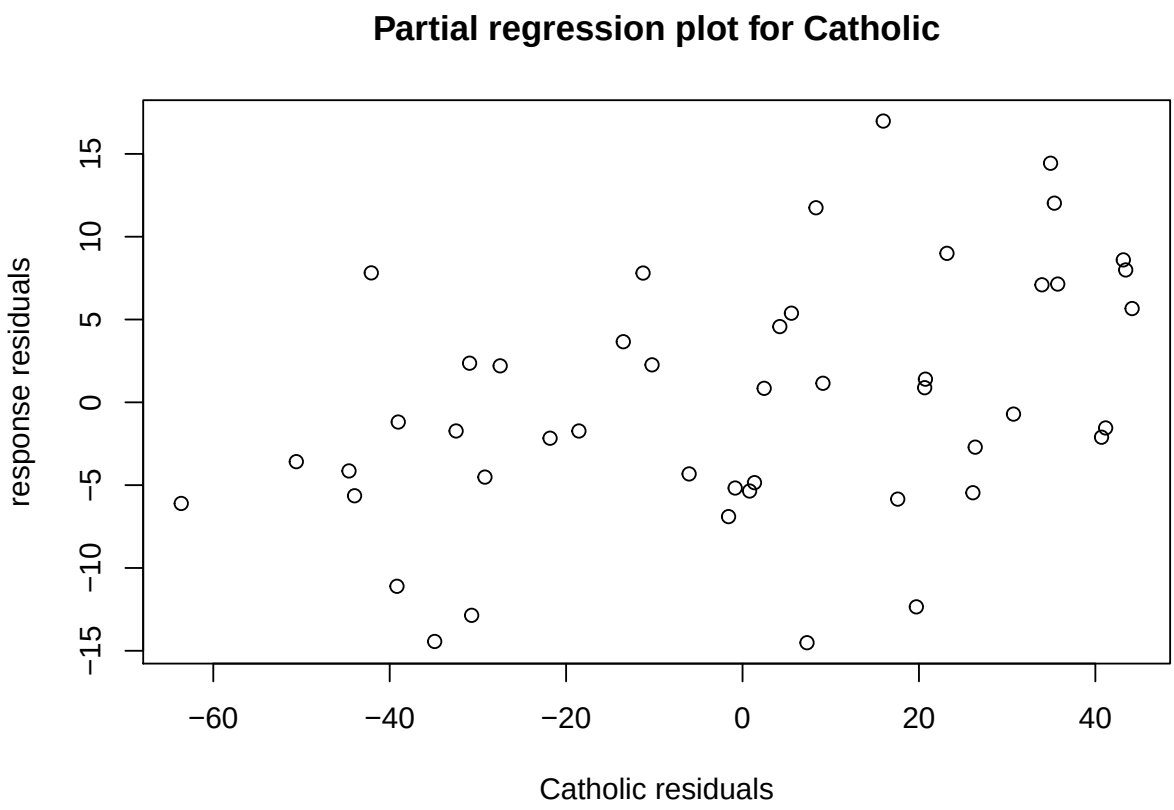
```
predictors <- names(lm.fit$coefficients)
predictors <- predictors[2:length(predictors)]

lm.formula <- formula(lm.fit)
response <- lm.formula[[2]]

for (i in 1:length(predictors)) {
    predictor <- predictors[i]
    others <- predictors[which(predictors != predictor)]
    d.formula <- paste(response, " ~ ", sep = "")
    m.formula <- paste(predictor, " ~ ", sep = "")

    for (j in 1:(length(others) - 1)) {
        d.formula <- paste(d.formula, others[j], " + ", sep = "")
        m.formula <- paste(m.formula, others[j], " + ", sep = "")
    }
    d.formula <- paste(d.formula, others[length(others)], sep = "")
    d.formula <- formula(d.formula)
```

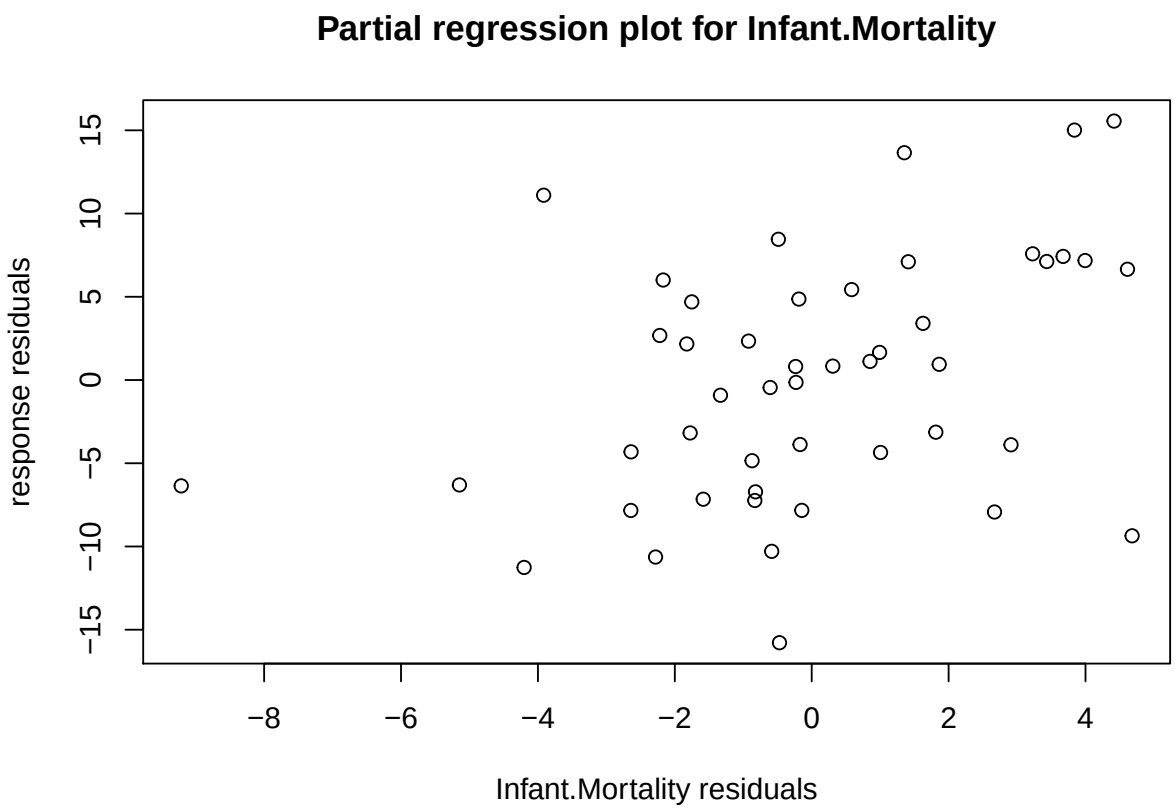
```
    m.formula <- paste(m.formula, others[length(others)], sep = "")
    m.formula <- formula(m.formula)

    d <- residuals(lm(d.formula, df))

    m <- residuals(lm(m.formula, df))

    plot(m, d, xlab = paste(predictor, " residuals", sep = ""), ylab = "response residua
        main = paste("Partial regression plot for ", predictor, sep = ""))

}
```




Partial regression plot for Examination
response residuals
15
10
5
0
−5
−10
−15
−5
0
5
Examination residuals


Partial regression plot for Education
response residuals
10
0
−10
−20
−10
−5
0
5
10
15
Education residuals


Partial regression plot for Catholic
response residuals
Catholic residuals
15
10
5
0
−5
−10
−15
−60
−40
−20
0
20
40



6.5 Using the cheddar data, fit a model with taste as the response and the other three variables as predictors.

```
rm(list = ls())
data(cheddar, package = "faraway")
lm.fit <- lm(taste ~ ., data = cheddar)
```

```
df <- cheddar
numPredictors <- (ncol(df) - 1)
hatv <- hatvalues(lm.fit)
lev.cut <- (numPredictors + 1) * 2 * 1/nrow(df)
high.leverage <- df[hatv > lev.cut, ]
pander(high.leverage, caption = "High Leverage Data Elements")
```

Table 12: High Leverage Data Elements

taste	Acetic	H2S	Lactic

We've used the rule of thumb that points with a leverage greater than $\frac{2p}{n}$ should be looked at.

(d) Check for outliers.

```
studentized.residuals <- rstudent(lm.fit)
max.residual <- studentized.residuals[which.max(abs(studentized.residuals))]
range.residuals <- range(studentized.residuals)
names(range.residuals) <- c("left", "right")
pander(data.frame(range.residuals = t(range.residuals)), caption = "Range of Studentized
```

Table 13: Range of Studentized residuals

range.residuals.left	range.residuals.right
-1.878	3.015

```
p <- numPredictors + 1
n <- nrow(df)
t.val.alpha <- qt(0.05/(n * 2), n - p - 1)
pander(data.frame(t.val.alpha = t.val.alpha), caption = "Bonferroni corrected t-value")
```

Table 14: Bonferroni corrected t-value

t.val.alpha
-3.523

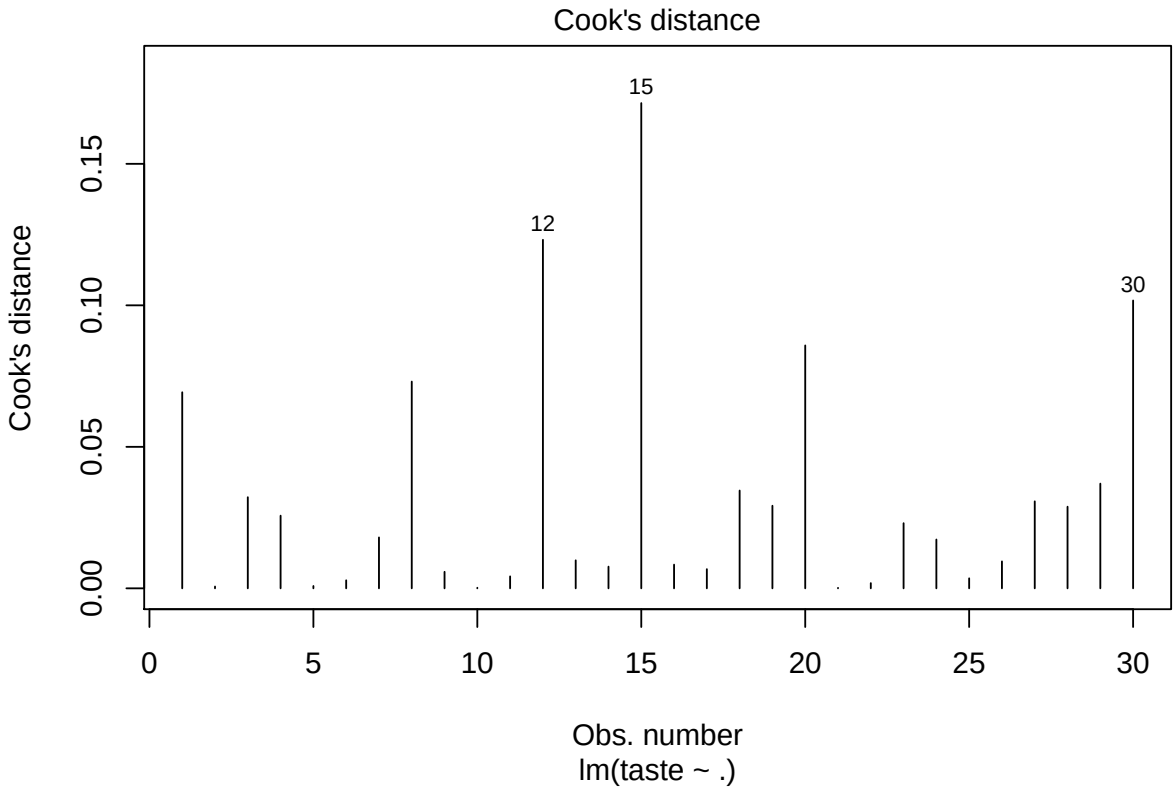
```
outlier.index <- abs(studentized.residuals) > abs(t.val.alpha)

outliers <- df[outlier.index == TRUE, ]

if (nrow(outliers) >= 1) {
    pander(outliers, caption = "outliers")
}
```

Here we look for studentized residuals that fall outside the interval given by the Bonferroni corrected t-values.

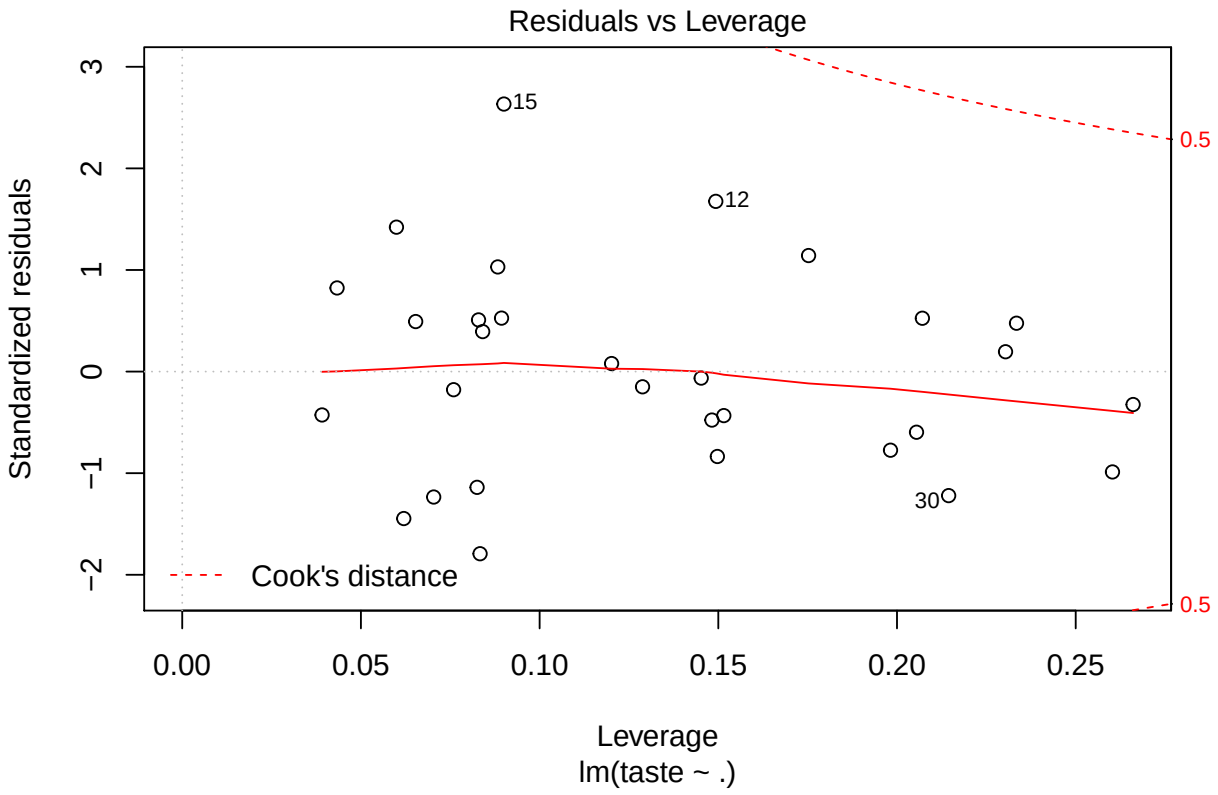
(e) Check for influential points.

We plot the Cook's distances and the residual-leverage plot with level set contours of the Cook distance.

```
plot(lm.fit, which = 4)
```



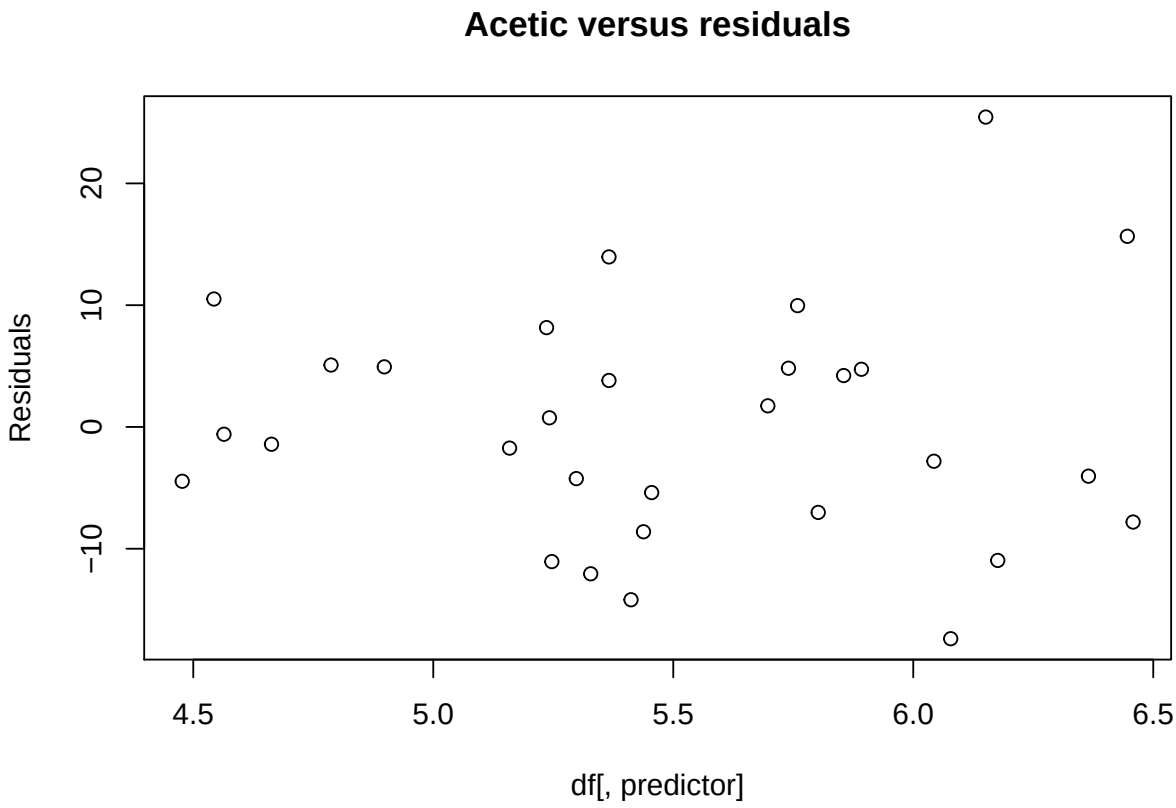

```
plot(lm.fit, which = 5)
```

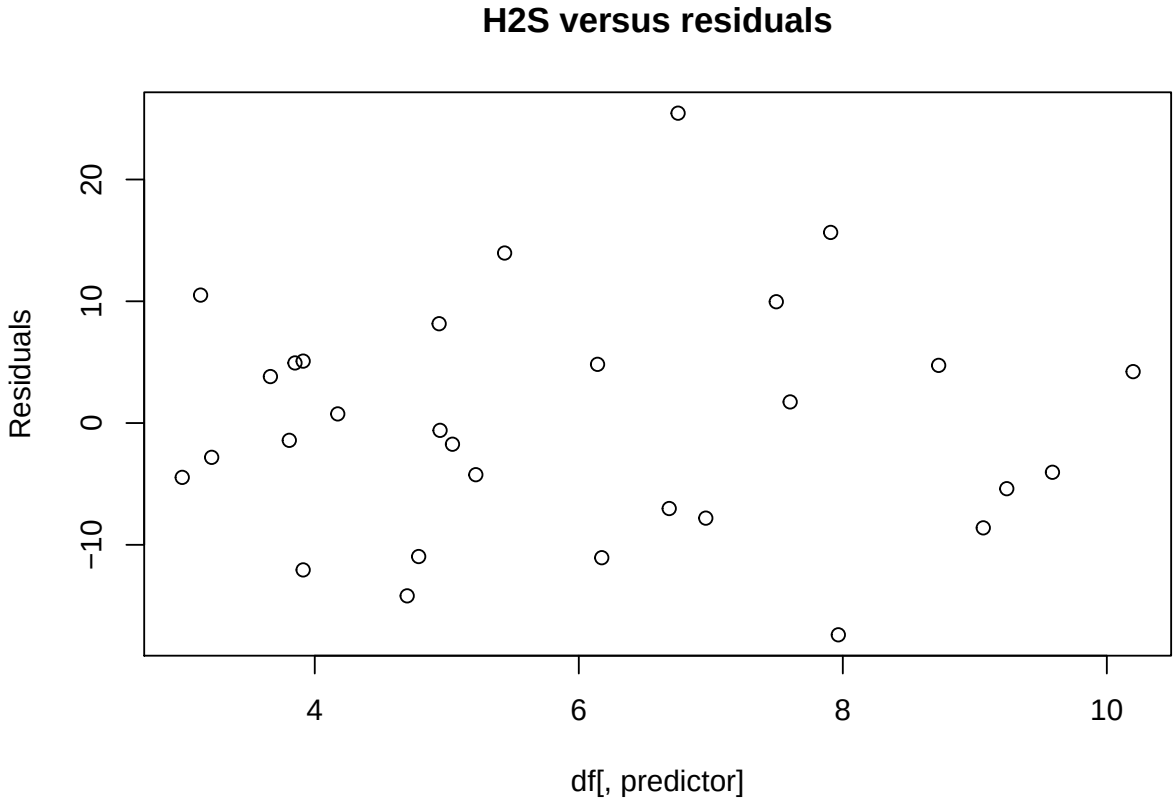


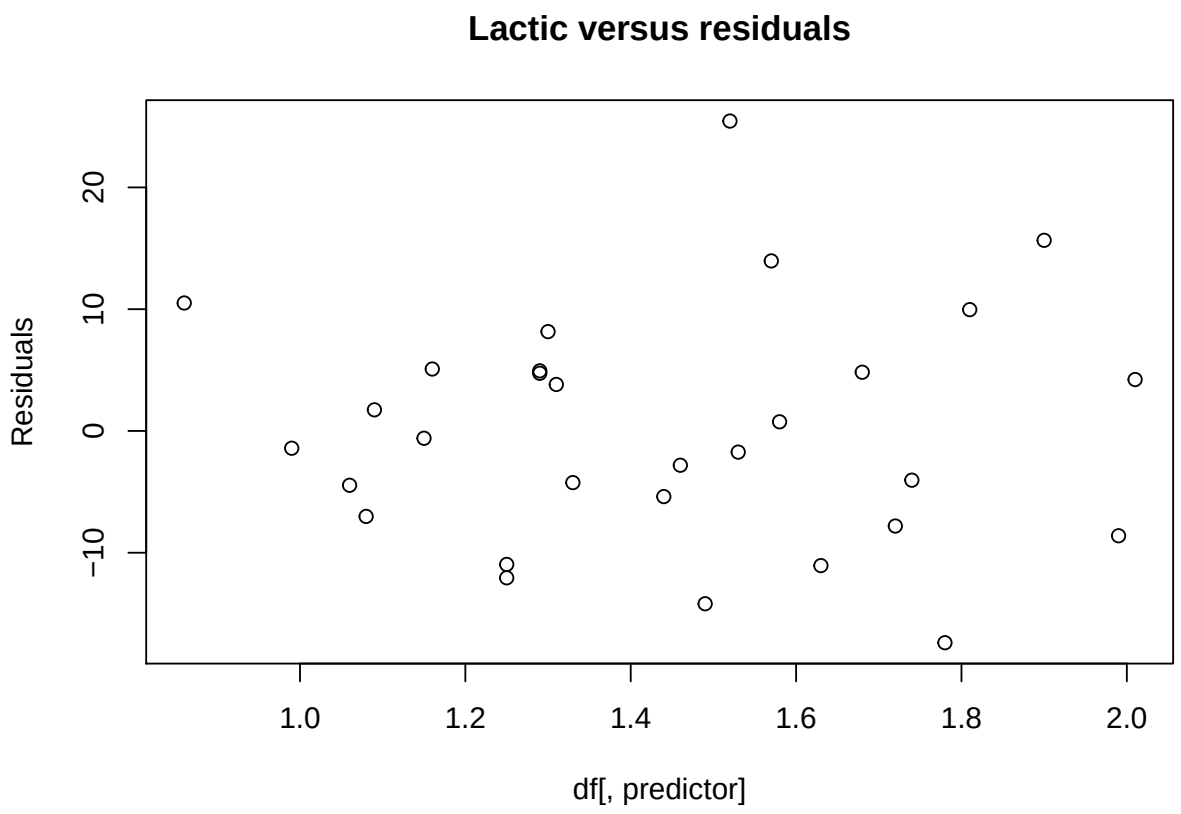
(f) Check for structure in the model.

Plot residuals versus predictors

```
predictors <- names(lm.fit$coefficients)
predictors <- predictors[2:length(predictors)]

for (i in 1:length(predictors)) {
    predictor <- predictors[i]

    plot(df[, predictor], residuals(lm.fit), xlab = , ylab = "Residuals", main = paste(p
        " versus residuals", sep = ""))

}
```


Acetic versus residuals
Residuals
20
10
0
−10
4.5
5.0
5.5
6.0
6.5
df[, predictor]

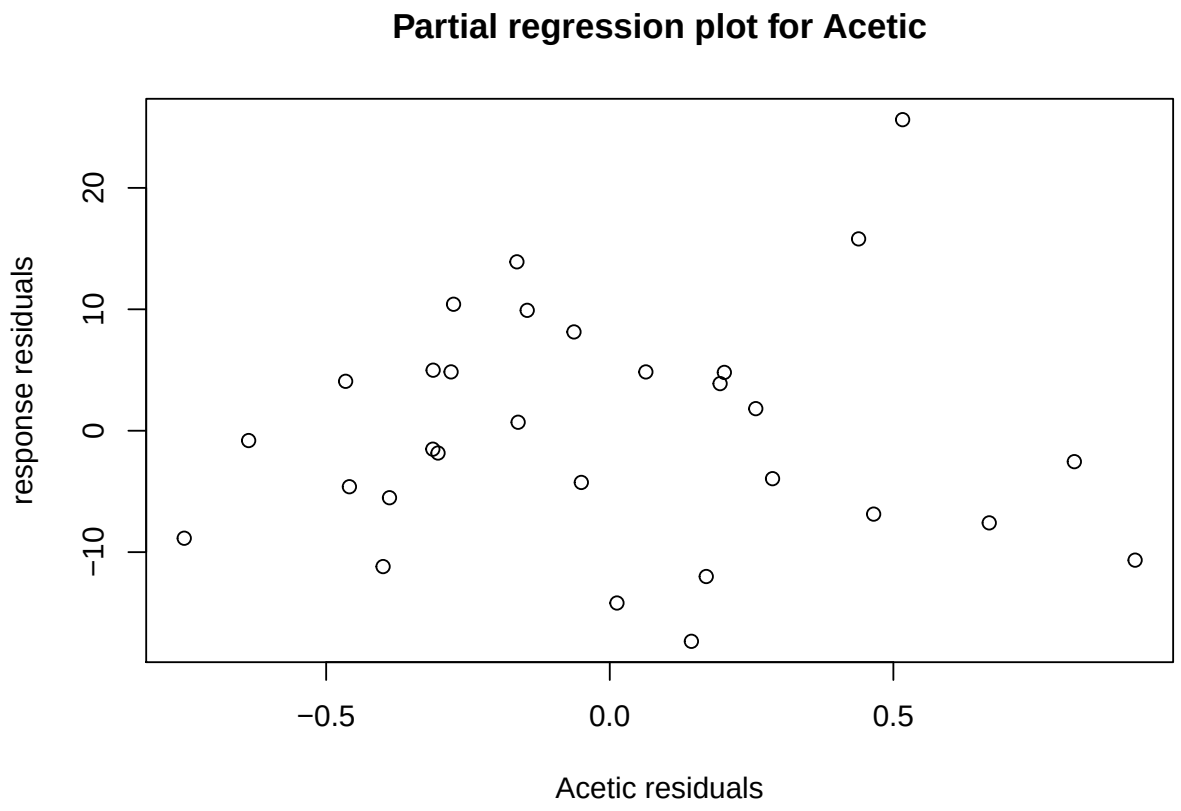

H2S versus residuals
Residuals
20
10
0
−10
4
6
8
10
df[, predictor]

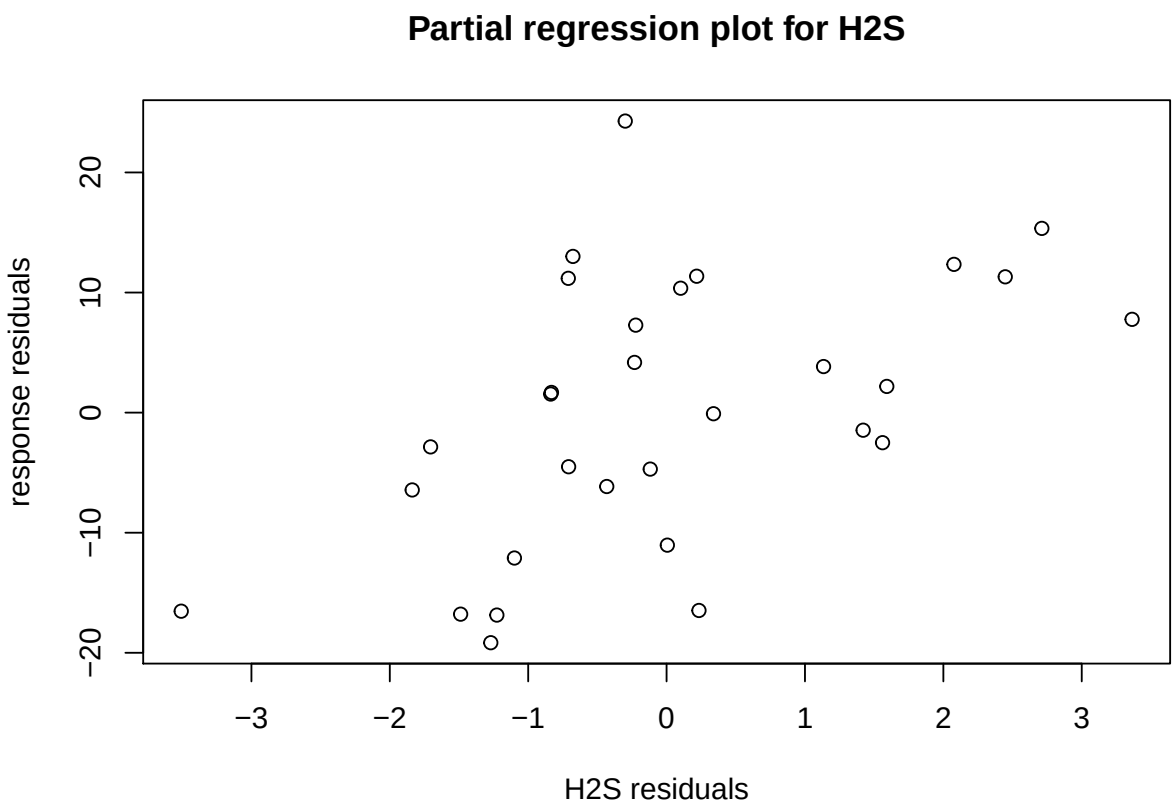


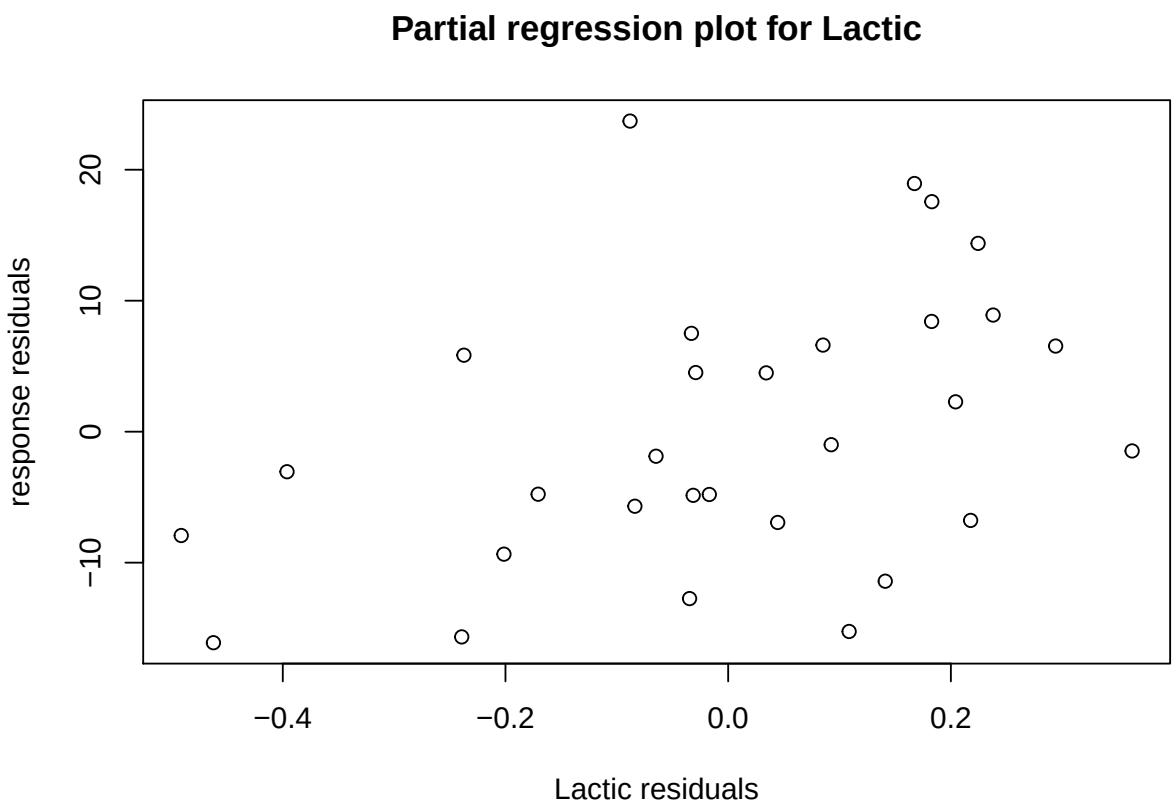
Perform partial regression

```
predictors <- names(lm.fit$coefficients)
predictors <- predictors[2:length(predictors)]

lm.formula <- formula(lm.fit)
response <- lm.formula[[2]]

for (i in 1:length(predictors)) {
    predictor <- predictors[i]
    others <- predictors[which(predictors != predictor)]
    d.formula <- paste(response, " ~ ", sep = "")
    m.formula <- paste(predictor, " ~ ", sep = "")

    for (j in 1:(length(others) - 1)) {
        d.formula <- paste(d.formula, others[j], " + ", sep = "")
        m.formula <- paste(m.formula, others[j], " + ", sep = "")
    }
    d.formula <- paste(d.formula, others[length(others)], sep = "")
    d.formula <- formula(d.formula)
```

```
    m.formula <- paste(m.formula, others[length(others)], sep = "")
    m.formula <- formula(m.formula)

    d <- residuals(lm(d.formula, df))

    m <- residuals(lm(m.formula, df))

    plot(m, d, xlab = paste(predictor, " residuals", sep = ""), ylab = "response residua
        main = paste("Partial regression plot for ", predictor, sep = ""))

}
```




Partial regression plot for H2S
response residuals
H2S residuals
20
10
0
−10
−20
−3
−2
−1
0
1
2
3


Partial regression plot for Lactic
response residuals
20
10
0
−10
−0.4
−0.2
0.0
0.2
Lactic residuals

NCSU ST 503 Discussion 8

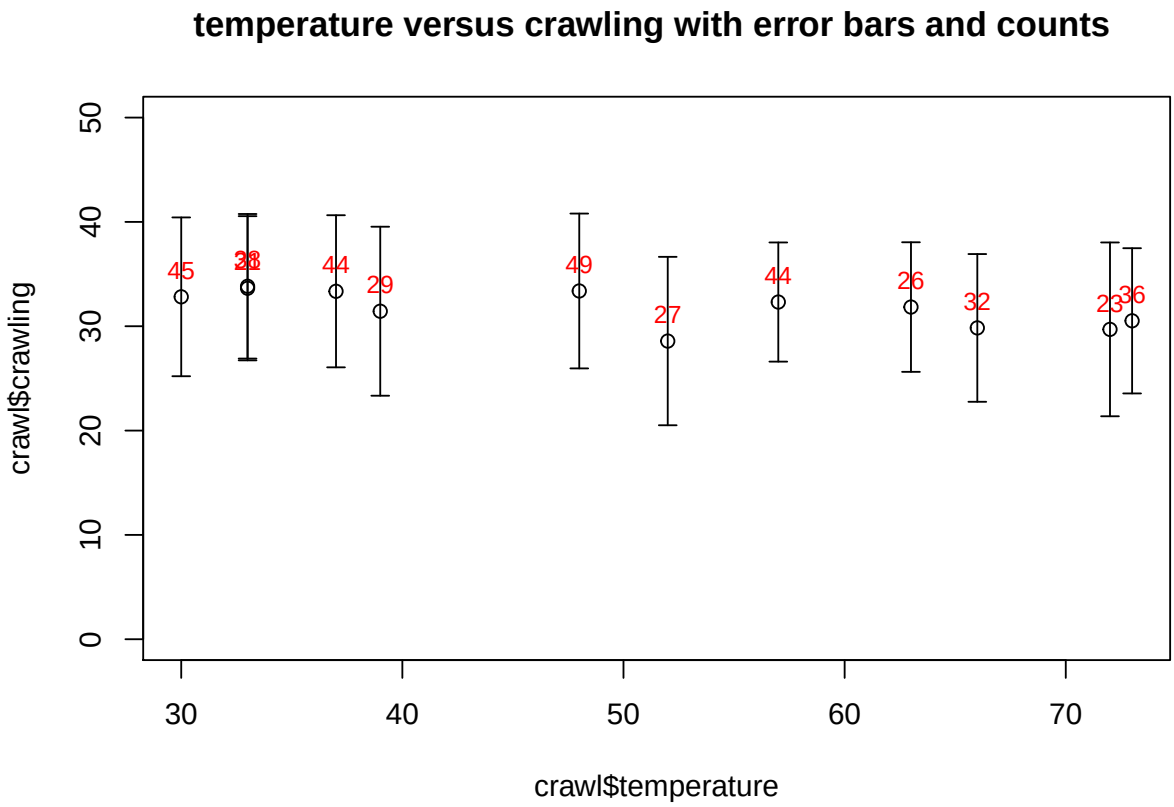
Probem 8.7 Faraway, Julian J. Linear Models with R CRC Press.

Bruce Campbell

8.7 crawl data analysis

The crawl dataset contains data on a study looking at the age when babies learn to crawl as a function of ambient temperatures. There is additional information about the number of babies studied each month and the variation in the response. Make an appropriate choice of weights to investigate the relationship between crawling age and temperature.

First we plot the dat along with the error and count information



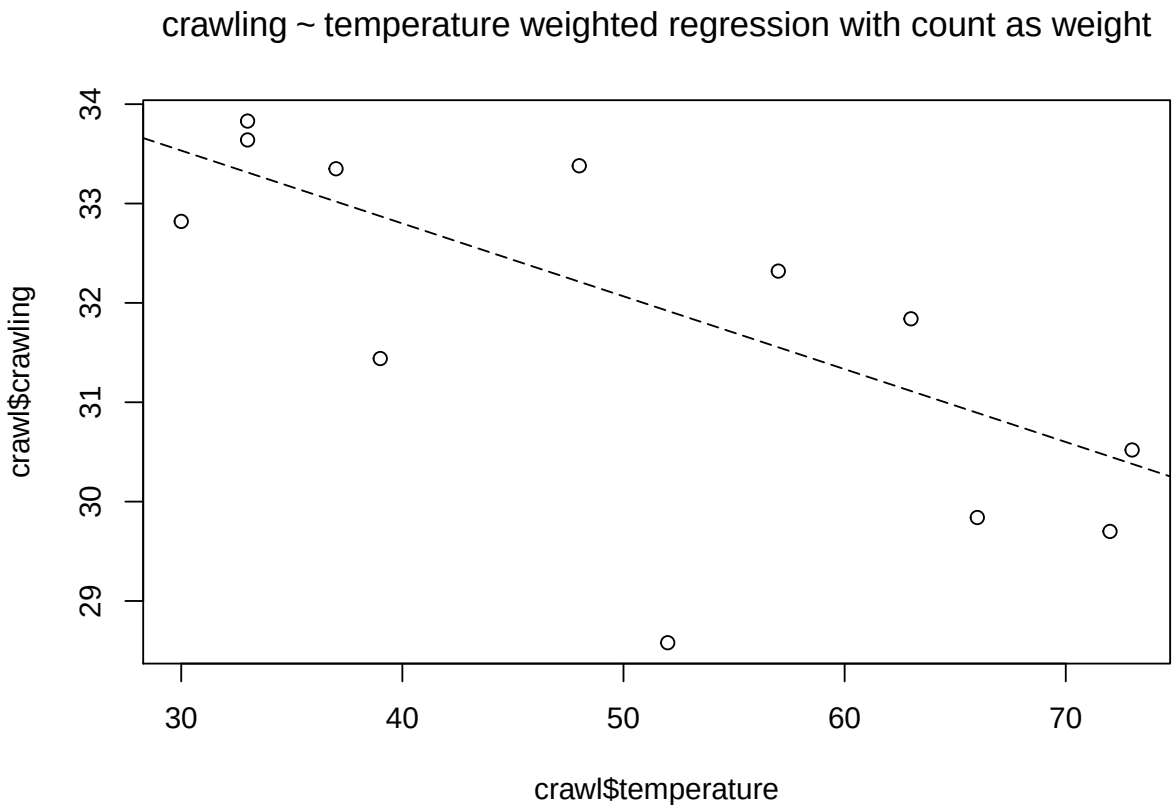
We fit a weighted least squares model with `lm` using weights $w_i = \frac{n_i}{SD_i^2}$.

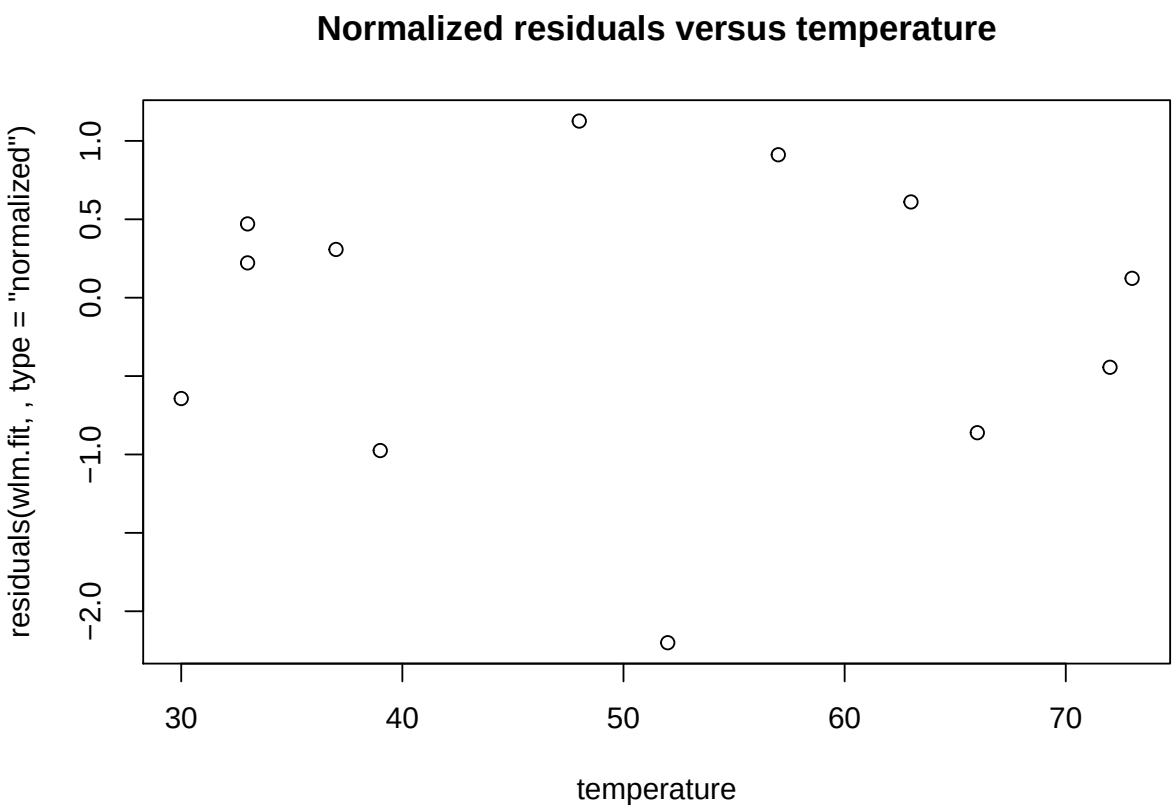
```
##
```

```
## Call:
## lm(formula = crawling ~ temperature, data = crawl, weights = wts)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1504 -0.6817  0.1688  0.4941  1.1009
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 35.73262    1.21153   29.49 4.69e-11 ***
## temperature -0.07332    0.02328   -3.15   0.0103 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9772 on 10 degrees of freedom
## Multiple R-squared:  0.4981, Adjusted R-squared:  0.4479
## F-statistic: 9.923 on 1 and 10 DF,  p-value: 0.01033
```

We fit a weighted least squares model with gls using weights $w_i = \frac{n_i}{SD_i^2}$.

```
## Generalized least squares fit by REML
##   Model: crawling ~ temperature
##   Data: crawl
##        AIC      BIC    logLik
##   48.76397 49.67173 -21.38199
##
## Variance function:
##  Structure: fixed weights
##  Formula: ~SD^2/n
##
## Coefficients:
##                Value Std.Error   t-value p-value
## (Intercept) 35.73262 1.2115286 29.493832  0.0000
## temperature -0.07332 0.0232771 -3.150053  0.0103
##
##  Correlation:
##             (Intr)
## temperature -0.96
##
## Standardized residuals:
##        Min         Q1        Med         Q3        Max
## -2.2007008 -0.6976329  0.1727593  0.5057059  1.1266156
##
## Residual standard error: 0.9771564
## Degrees of freedom: 12 total; 10 residual
```


crawling ~ temperature weighted regression with count as weight
crawl$crawling
34
33
32
31
30
29
30
40
50
60
70
crawl$temperature



This data may be amenable to the lack of fit analysis discussed in the text. We don't have the original $y_i i$ in the formaula for

$$SS_{pe} = \sum_j \sum_i (y_{ji} - \bar{y})^2$$

But we know n is the sum of the counts, that the mean response for each temperature as the crawl variable, and that $SE_j = \frac{1}{n_j} \sum_i (y_{ii} - \bar{y_j})^2$

So

$$\hat{\sigma}^2 = \frac{SS_{pe}}{(\sum_j n_i) - j} = \frac{1}{(\sum_j n_i) - j} \sum_j \sum_i (y_{ji} - \bar{y})^2 = \frac{1}{(\sum_j n_i) - j} \sum_j n_j SE_j$$

Calculating this for our data set we have

Table 1: estimated SD from the repeated predictor values

se
2.718

NCSU ST 503 Discussion 8

Probem 10.6 Faraway, Julian J. Linear Models with R CRC Press.

Bruce Campbell

10.6 Model Selection with the hipcenter data.

Use the seatpos data with hipcenter as the response.

(a) Fit a model with all eight predictors. Comment on the effect of leg length on the response.

```
##
## Call:
## lm(formula = hipcenter ~ ., data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -73.827 -22.833  -3.678  25.017  62.337
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 436.43213  166.57162   2.620   0.0138 *
## Age           0.77572    0.57033   1.360   0.1843
## Weight        0.02631    0.33097   0.080   0.9372
## HtShoes      -2.69241    9.75304  -0.276   0.7845
## Ht            0.60134   10.12987   0.059   0.9531
## Seated        0.53375    3.76189   0.142   0.8882
## Arm          -1.32807    3.90020  -0.341   0.7359
## Thigh        -1.14312    2.66002  -0.430   0.6706
## Leg          -6.43905    4.71386  -1.366   0.1824
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.72 on 29 degrees of freedom
## Multiple R-squared:  0.6866, Adjusted R-squared:  0.6001
## F-statistic:  7.94 on 8 and 29 DF,  p-value: 1.306e-05
```

We note that leg length is significant at a level of $\alpha = 0.182$ and it has a negative association with the response.

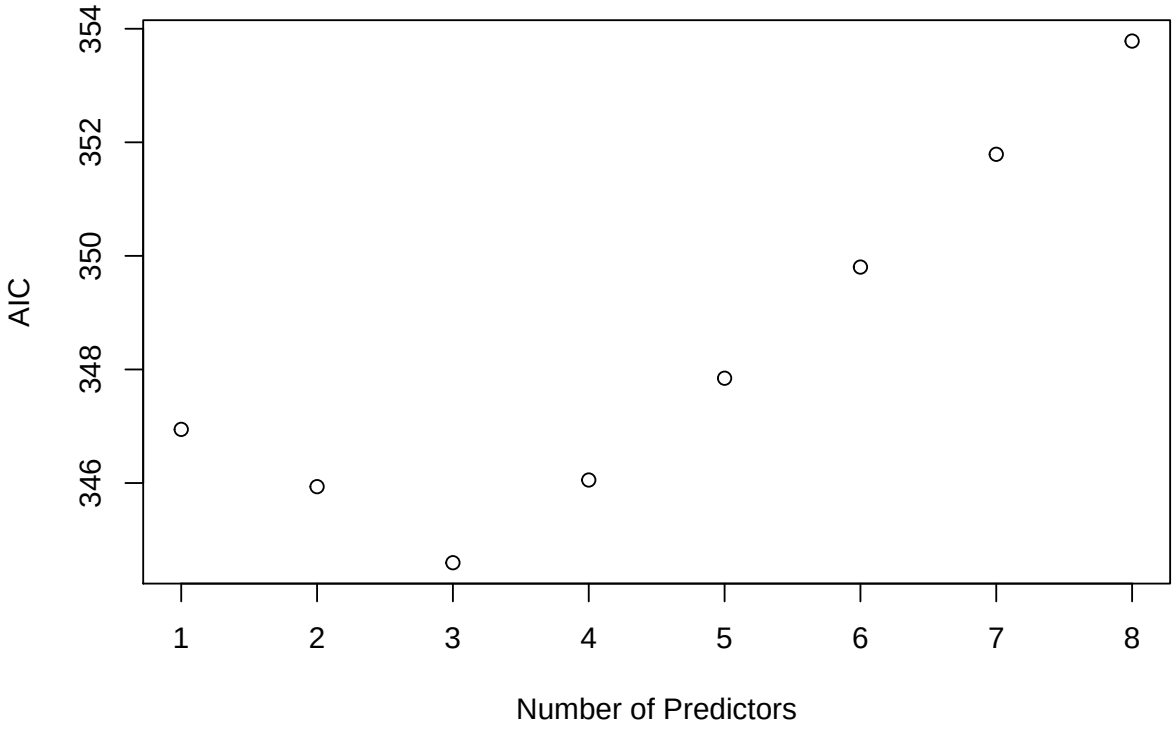
(b) Compute a 95% prediction interval for the mean value of the predictors.

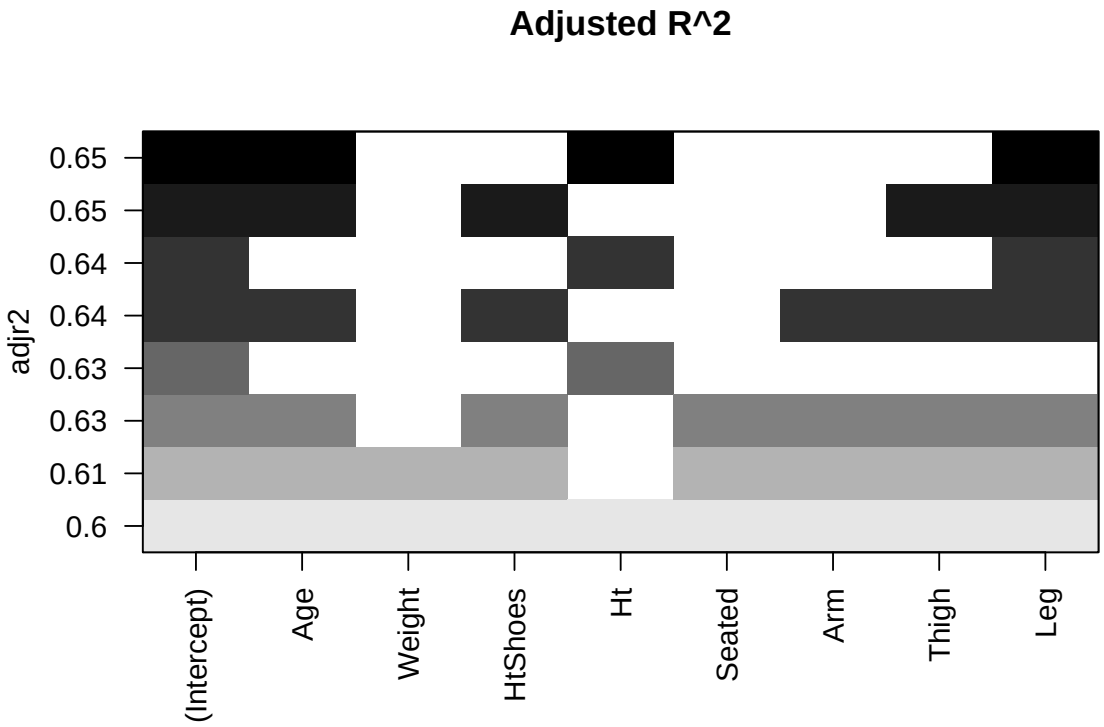
```
##         fit     lwr       upr
## 1 -164.8849 -243.04 -86.72972
```

pi.width
156.3

(c) Use AIC to select a model. Now interpret the effect of leg length and compute the prediction interval. Compare the conclusions from the two models.

```
##   (Intercept)   Age Weight HtShoes    Ht Seated   Arm Thigh   Leg
## 1        TRUE FALSE  FALSE   FALSE  TRUE  FALSE FALSE FALSE FALSE
## 2        TRUE FALSE  FALSE   FALSE  TRUE  FALSE FALSE FALSE  TRUE
## 3        TRUE  TRUE  FALSE   FALSE  TRUE  FALSE FALSE FALSE  TRUE
## 4        TRUE  TRUE  FALSE    TRUE FALSE  FALSE FALSE  TRUE  TRUE
## 5        TRUE  TRUE  FALSE    TRUE FALSE  FALSE  TRUE  TRUE  TRUE
## 6        TRUE  TRUE  FALSE    TRUE FALSE   TRUE  TRUE  TRUE  TRUE
## 7        TRUE  TRUE   TRUE    TRUE FALSE   TRUE  TRUE  TRUE  TRUE
## 8        TRUE  TRUE   TRUE    TRUE  TRUE   TRUE  TRUE  TRUE  TRUE
```


AIC
346
348
350
352
354
1
2
3
4
5
6
7
8
Number of Predictors



We see that $hipcenter \sim +age + ht + Leg$ is the model with the lowest AIC. We also plot the Adjusted R^2 of the models.

```
##
## Call:
## lm(formula = hipcenter ~ Age + Ht + Leg, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -79.715 -22.758  -4.102  21.394  60.576
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 452.1976   100.9482   4.480 8.04e-05 ***
## Age           0.5807     0.3790   1.532   0.1347
## Ht           -2.3254     1.2545  -1.854   0.0725 .
## Leg          -6.7390     4.1050  -1.642   0.1099
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.12 on 34 degrees of freedom
```

```
## Multiple R-squared:  0.6814, Adjusted R-squared:  0.6533
## F-statistic: 24.24 on 3 and 34 DF,  p-value: 1.426e-08
```

Leg now has a p-value of 0.1099

```
##         fit      lwr       upr
## 1 -164.8849 -237.192 -92.57771
```

pi.width
144.6

As expected our prediction interval has decresed in width. Ht is now significant at 0.07 which is a dramatic change. We presume this is due to linear association among the predictors. We note that the predictions of the two models are similar.

NCSU ST 503 Discussion 10

Probem 11.6 Faraway, Julian J. Linear Models with R CRC Press.

Bruce Campbell

10.6 PCA analysis of kanga dataset

The dataset kanga contains data on the skulls of historical kangaroo specimens.

(a) Compute a PCA on (the 18 skull measurements. You will need to exclude observations with missing values. What percentage of variation is explained by the first principal component?

```
## Importance of components:
##                             PC1      PC2      PC3      PC4      PC5
## Standard deviation     288.0382 69.51124 30.74720 27.85580 21.73015
## Proportion of Variance   0.9003  0.05243  0.01026  0.00842  0.00512
## Cumulative Proportion    0.9003  0.95269  0.96295  0.97136  0.97649
##                             PC6      PC7     PC8      PC9     PC10
## Standard deviation     19.42356 17.28247 16.6247 14.52310 13.98826
## Proportion of Variance  0.00409  0.00324  0.0030  0.00229  0.00212
## Cumulative Proportion   0.98058  0.98382  0.9868  0.98911  0.99123
##                            PC11     PC12     PC13     PC14    PC15    PC16
## Standard deviation     12.35253 12.07402 11.94245 10.82939 10.0735 8.46081
## Proportion of Variance  0.00166  0.00158  0.00155  0.00127  0.0011 0.00078
## Cumulative Proportion   0.99289  0.99447  0.99602  0.99729  0.9984 0.99917
##                            PC17    PC18
## Standard deviation     7.16825 5.01246
## Proportion of Variance 0.00056 0.00027
## Cumulative Proportion  0.99973 1.00000
```

47 data elements were removed due to missing values in the measurement dimensions. We see that %90 of variance in the measurements is explained by the first principal component.

(b) Provide the loadings for the first principal component. What variables are prominent?

The loadings for a principal component $\mathbf{u}_i$ are the values of the dimensions u_{ij} , in our case the measurements. We note that the textbook uses r method prcomp to perform principal

components and that it gets loadings from the rot matrix. There is another r function in common use for pca - princomp. This method has a loading structure in the output. We tested this method and got a vector similar to the other method except all the signs were reversed. This is OK because the direction is the same - i.e. if we project all the data points on the first version, we'll get the same points as if we had projected on the second version. We also note there is some confusion on the difference between eigenvectors and loadings. The rot matrix is orthogonal - we checked this for a few values

```
t(pca.kanga$rotation[,1]) %*%  pca.kanga$rotation[,1]
t(pca.kanga$rotation[,3]) %*%  pca.kanga$rotation[,4]
t(pca.kanga$rotation[,1]) %*%  pca.kanga$rotation[,2]
```

Table 1: First Principal Component

	first.pc.loadings
basilar.length	0.484
occipitonasal.length	0.456
palate.length	0.366
palate.width	0.084
nasal.length	0.248
nasal.width	0.075
squamosal.depth	0.064
lacrymal.width	0.119
zygomatic.width	0.207
orbital.width	0.014
.rostral.width	0.106
occipital.depth	0.178
crest.width	-0.082
foramina.length	0.01
mandible.length	0.436
mandible.width	0.03
mandible.depth	0.058
ramus.height	0.209

We note that the following measurements all have loadings greater than .2

{*basilar.length*, *occipitonasal.length*, *palate.length*, *nasal.length*, *mandible.length*, *zygomatic.width*

(c) Repeat the PCA but with the variables all scaled to the same standard deviation. How do the percentage of variation explained and the first principal component differ from those found in the previous PCA?

PCA of scaled measurements

```
## Importance of components:
##                            PC1     PC2    PC3    PC4    PC5     PC6     PC7
## Standard deviation     3.5321 1.30672 1.1006 0.8443 0.6463 0.56426 0.51064
## Proportion of Variance 0.6931 0.09486 0.0673 0.0396 0.0232 0.01769 0.01449
## Cumulative Proportion  0.6931 0.78796 0.8553 0.8949 0.9181 0.93575 0.95024
##                            PC8     PC9   PC10    PC11   PC12    PC13
## Standard deviation     0.45185 0.43863 0.3723 0.30491 0.2815 0.24345
## Proportion of Variance 0.01134 0.01069 0.0077 0.00517 0.0044 0.00329
## Cumulative Proportion  0.96158 0.97227 0.9800 0.98514 0.9895 0.99283
##                           PC14    PC15    PC16    PC17    PC18
## Standard deviation     0.22317 0.18583 0.15031 0.11849 0.08949
## Proportion of Variance 0.00277 0.00192 0.00126 0.00078 0.00044
## Cumulative Proportion  0.99560 0.99752 0.99878 0.99956 1.00000
```

After scaling the proportion of variance explained by the first principal component has dropped to .69

(d) Give an interpretation of the second principal component.

Table 2: Second Principal Component

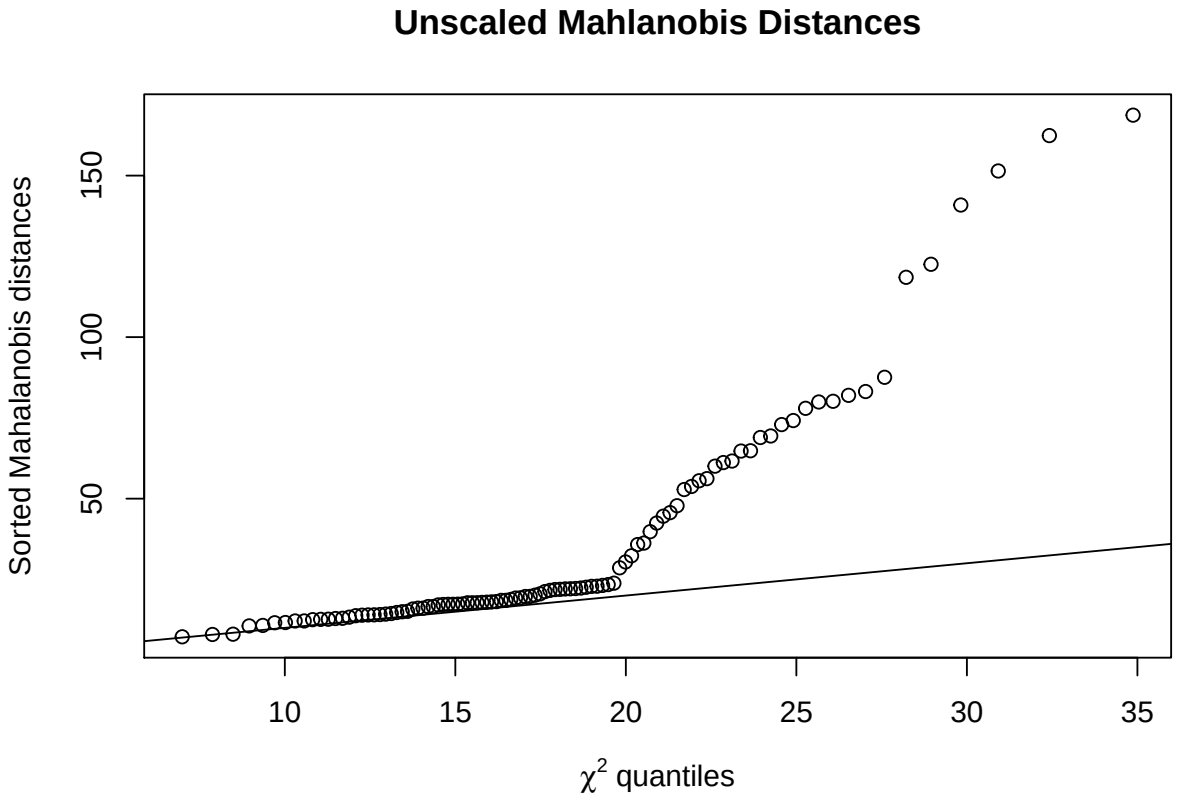
	first.pc.loadings
basilar.length	-0.138
occipitonasal.length	0.414
palate.length	-0.002
palate.width	-0.023
nasal.length	0.584
nasal.width	0.127
squamosal.depth	-0.105
lacrymal.width	-0.04
zygomatic.width	-0.41
orbital.width	0.001
.rostral.width	-0.063
occipital.depth	-0.079
crest.width	-0.245
foramina.length	0.061
mandible.length	-0.212
mandible.width	-0.092
mandible.depth	-0.106
ramus.height	-0.365

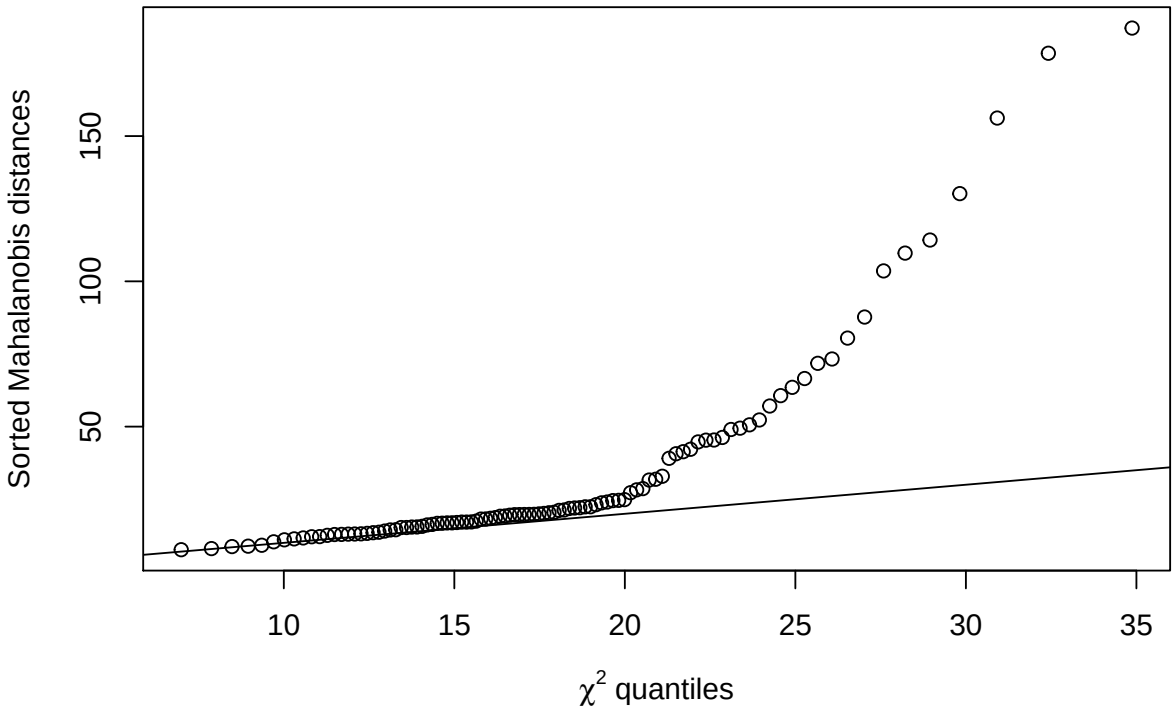
As evidenced by the loadings, the first principal component mainly account for variation in length feature. The second principal component is primarily a contrast between the variables

$\{occipitonasal.length,\ \ nasal.length\}$ and $\{ramus.height,\ \ crest.width\}$

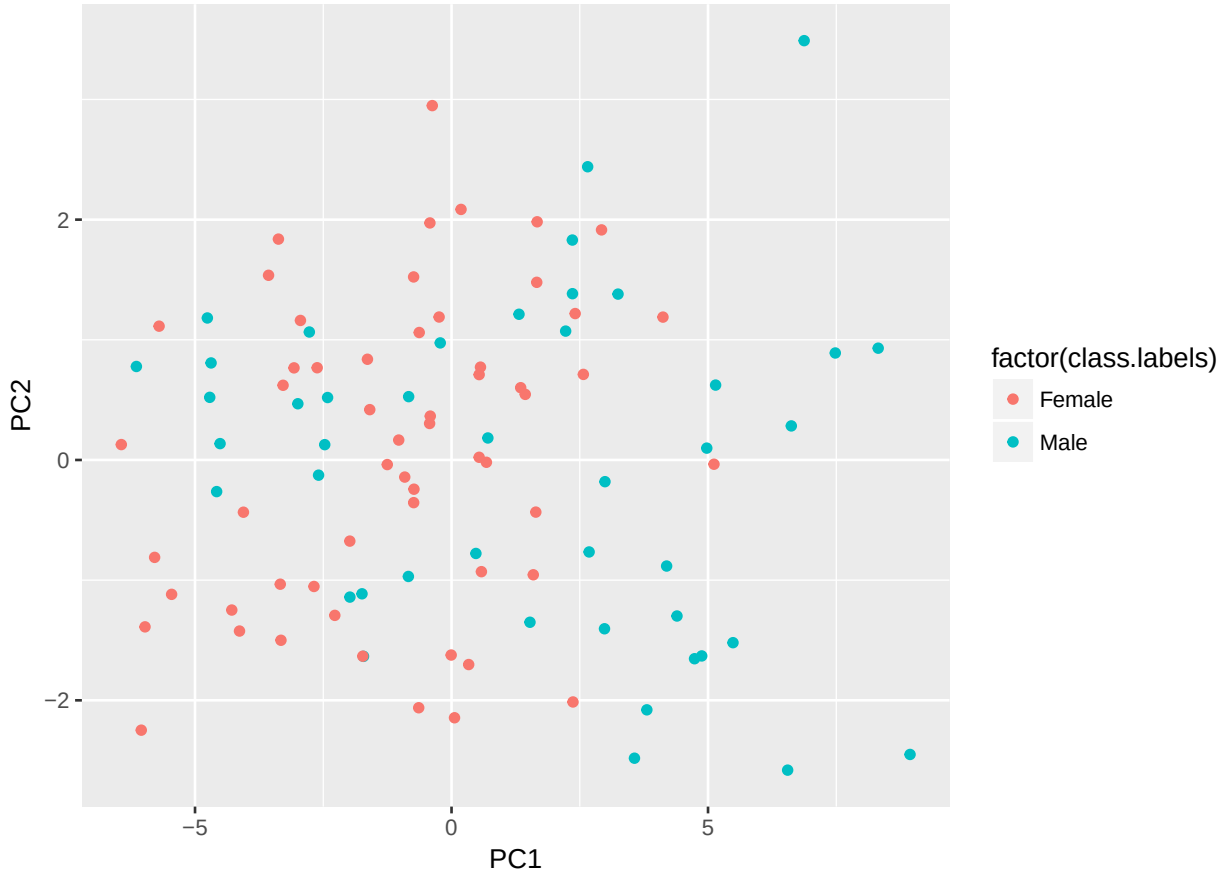
(e) Compute the Mahalanobis distances and plot appropriately to check for outliers.

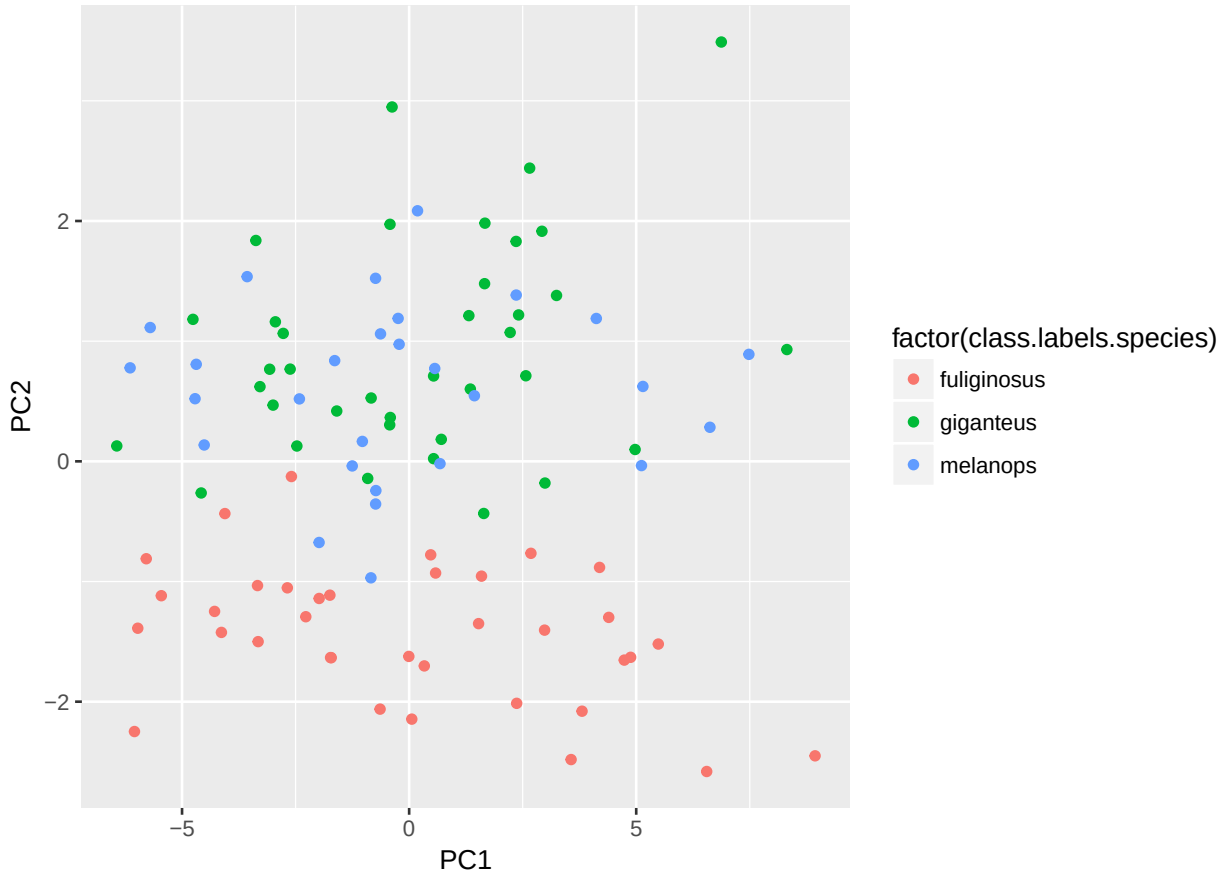
We calculate the distances in the unscaled and scaled data




Sorted Mahalanobis distances
50
100
150
10
15
20
25
30
35
χ^2 quantiles

(f) Make a scatterplot of the first and second principal components using a different plotting symbol depending on the sex of the specimen. Do you think these two components would be effective in determining the sex of a skull?





We see that the measurements do not allow for a linear classifier for discriminating sex via the first two PCA projections. We can discriminate the species fulignosus from giganteus and melanops.

NCSU ST 503 Discussion 11

Probem 2.1 Faraway, Julian J. Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models CRC Press.

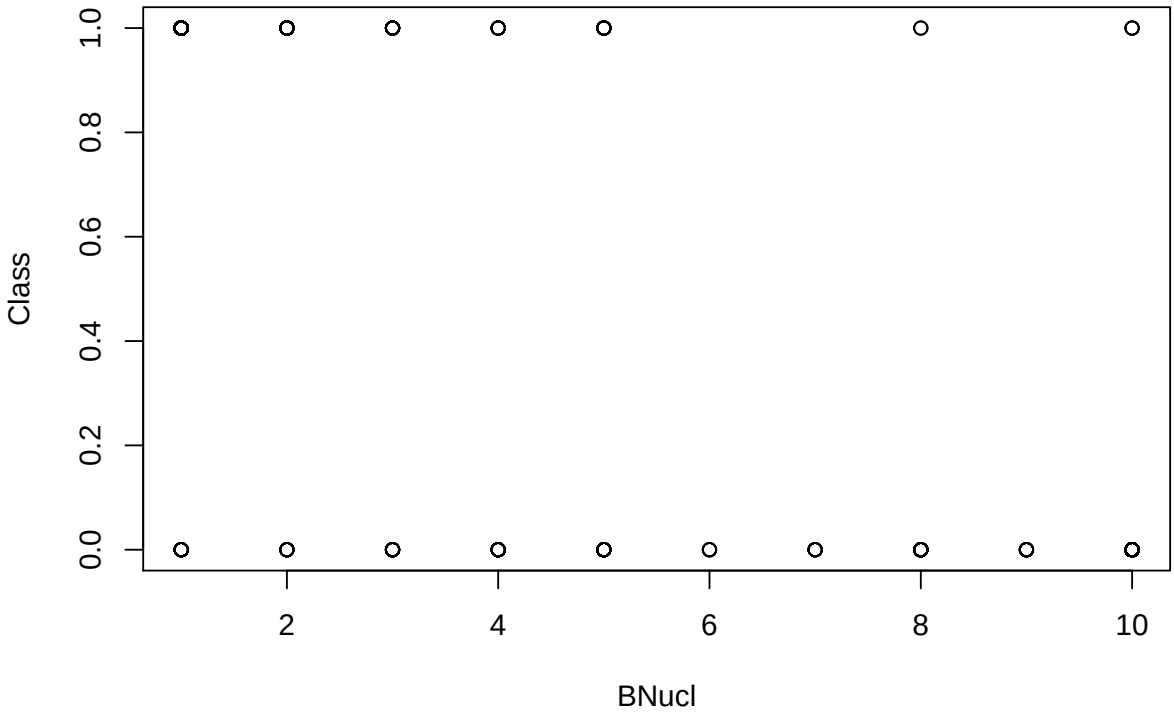
Bruce Campbell

2.1 wbca analysis

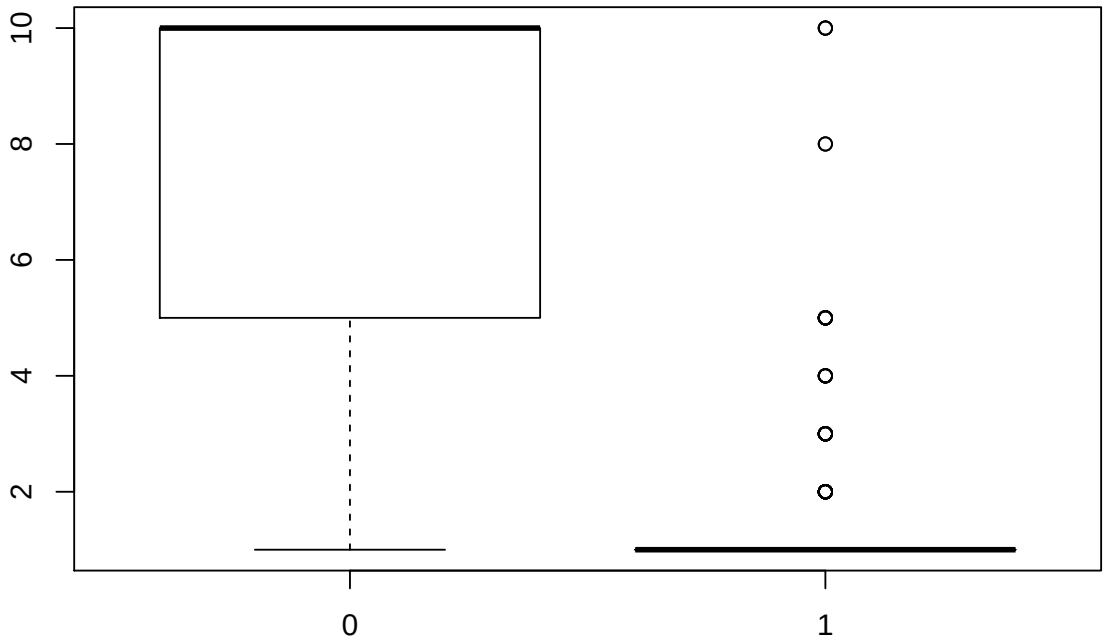
The dataset wbca comes from a study of breast cancer in Wisconsin. There are 681 cases of potentially cancerous tumors of which 238 are actually malignant. Determining whether a tumor is really malignant is traditionally determined by an invasive surgical procedure. The purpose of this study was to determine whether a new procedure called fine needle aspiration, which draws only a small sample of tissue, could be effective in determining tumor status.

(a) Plot the relationship between the classification and BNucl. i. Explain why plot(Class ~ BNucl, wbca) does not work well. ii. Create a factor version of the response and produce a version of the first panel of Figure 2.1. Comment on the shape of the boxplots. iii. Produce a version of the second panel of Figure 2.1. What does this plot say about the distribution? iv. Produce a version of the interleaved histogram shown in Figure 2.2 and comment on the distribution.

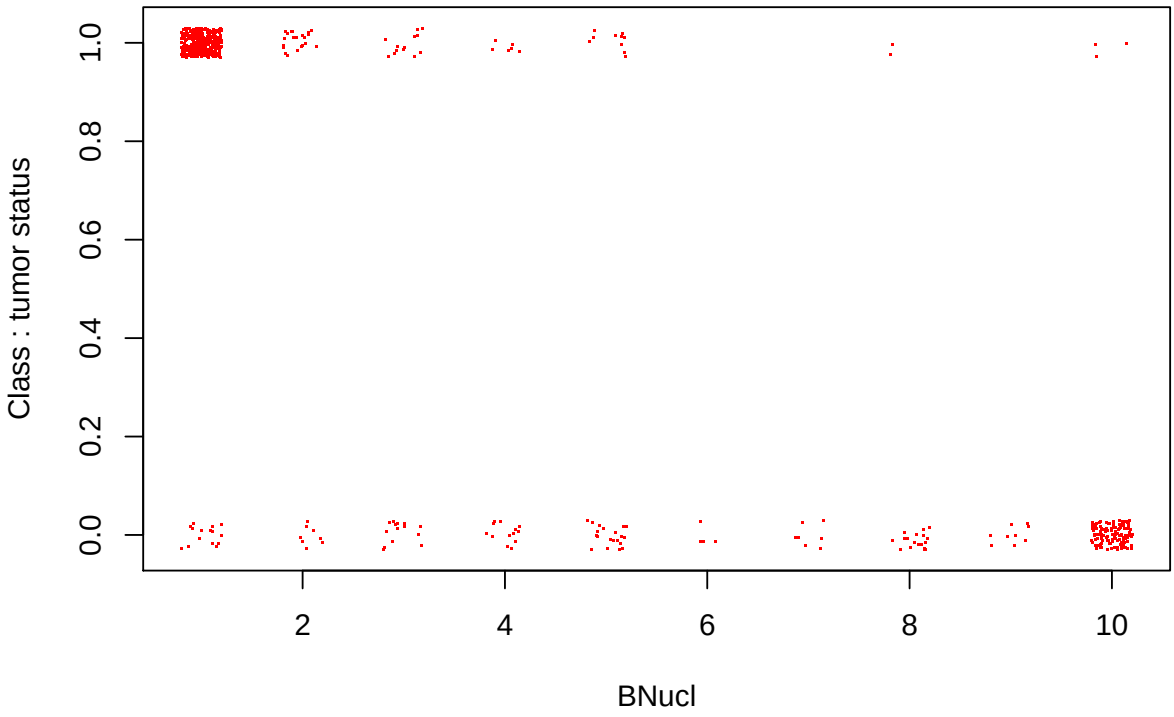
Here we plot $Class \sim BNucl$



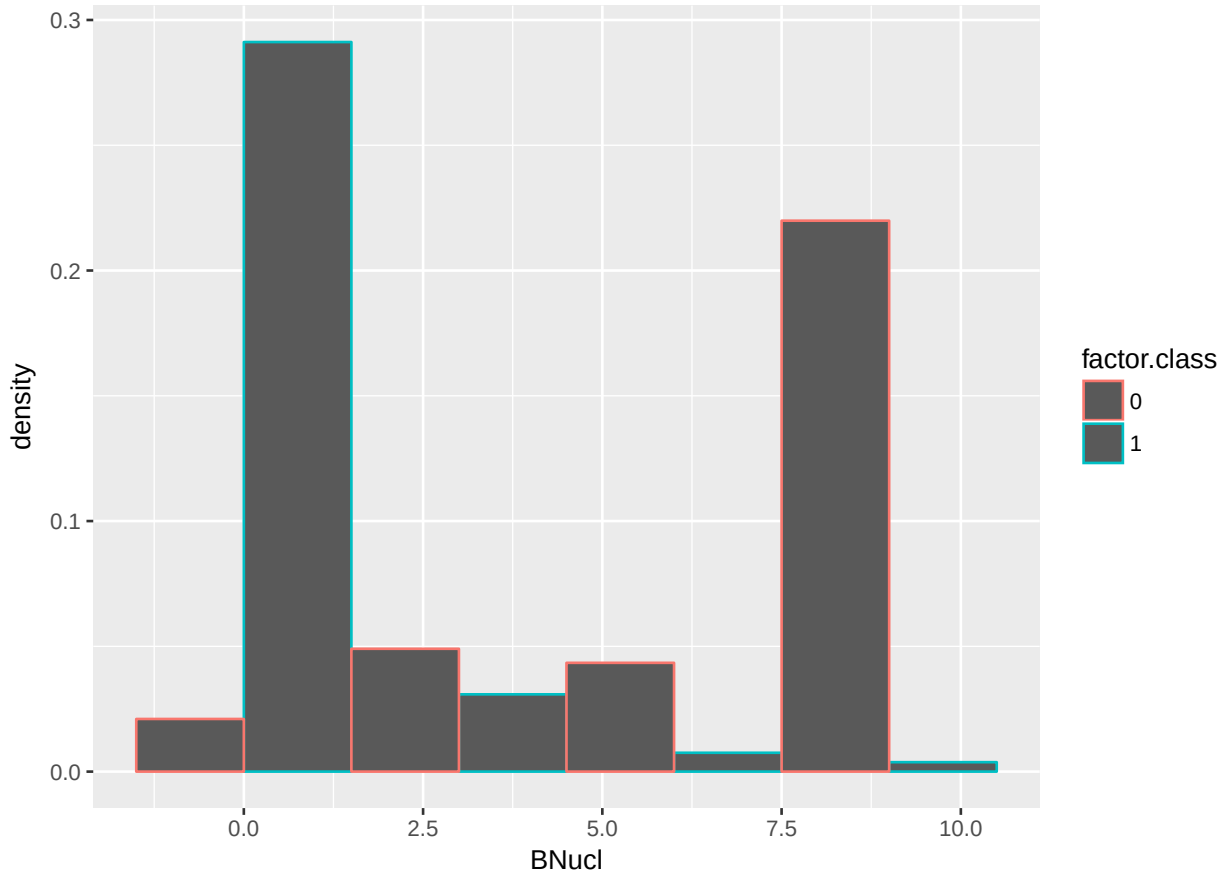
We see that since $BNucl$ is discrete we don't have a sense of how the variable is distributed by class well since the points overlap on the plot. A box plot provides a better visualization of the distribution by class.



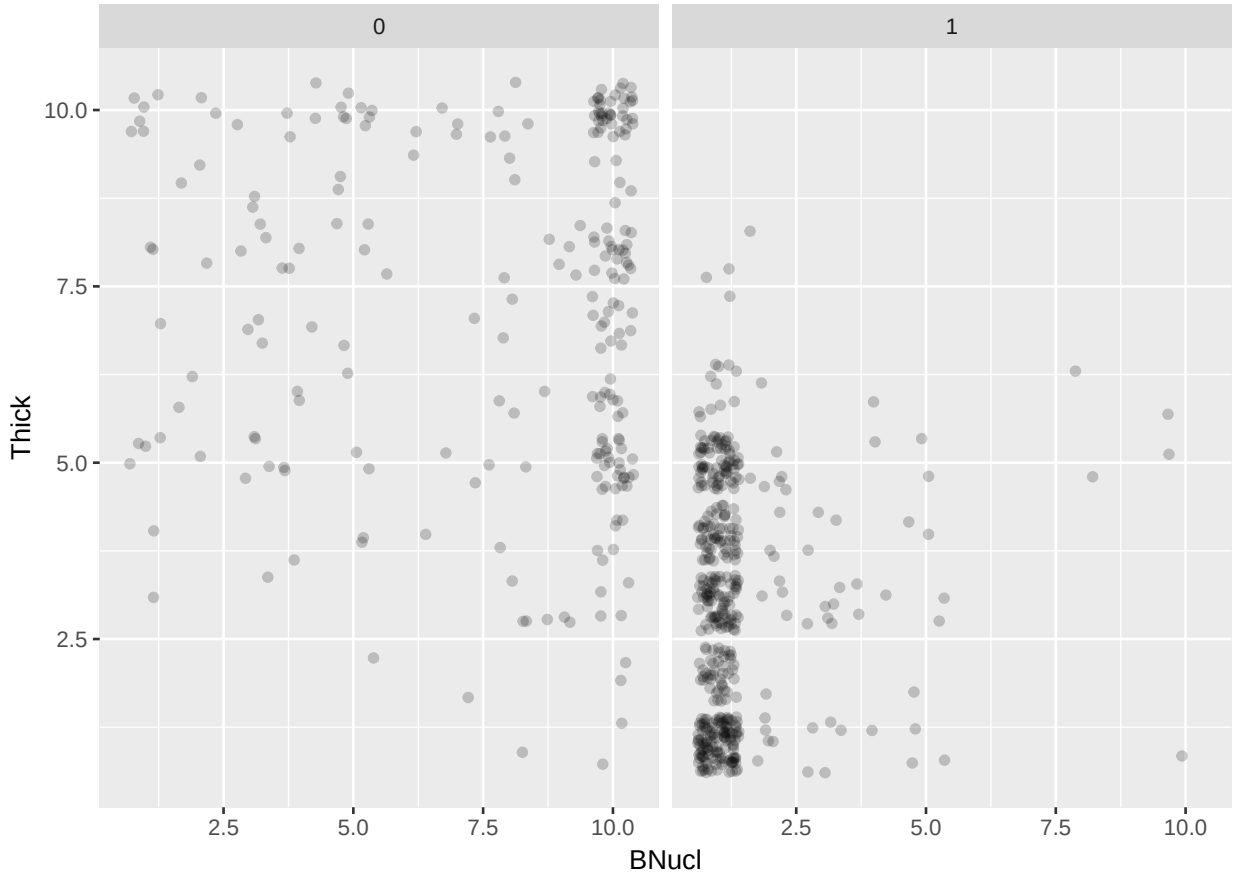
The boxplot show us that the $BNucl$ feature is a viable candidate for predicting cancer status. We can also add noise to the $Class \sim BNucl$ plot to remove the overlap in the points.

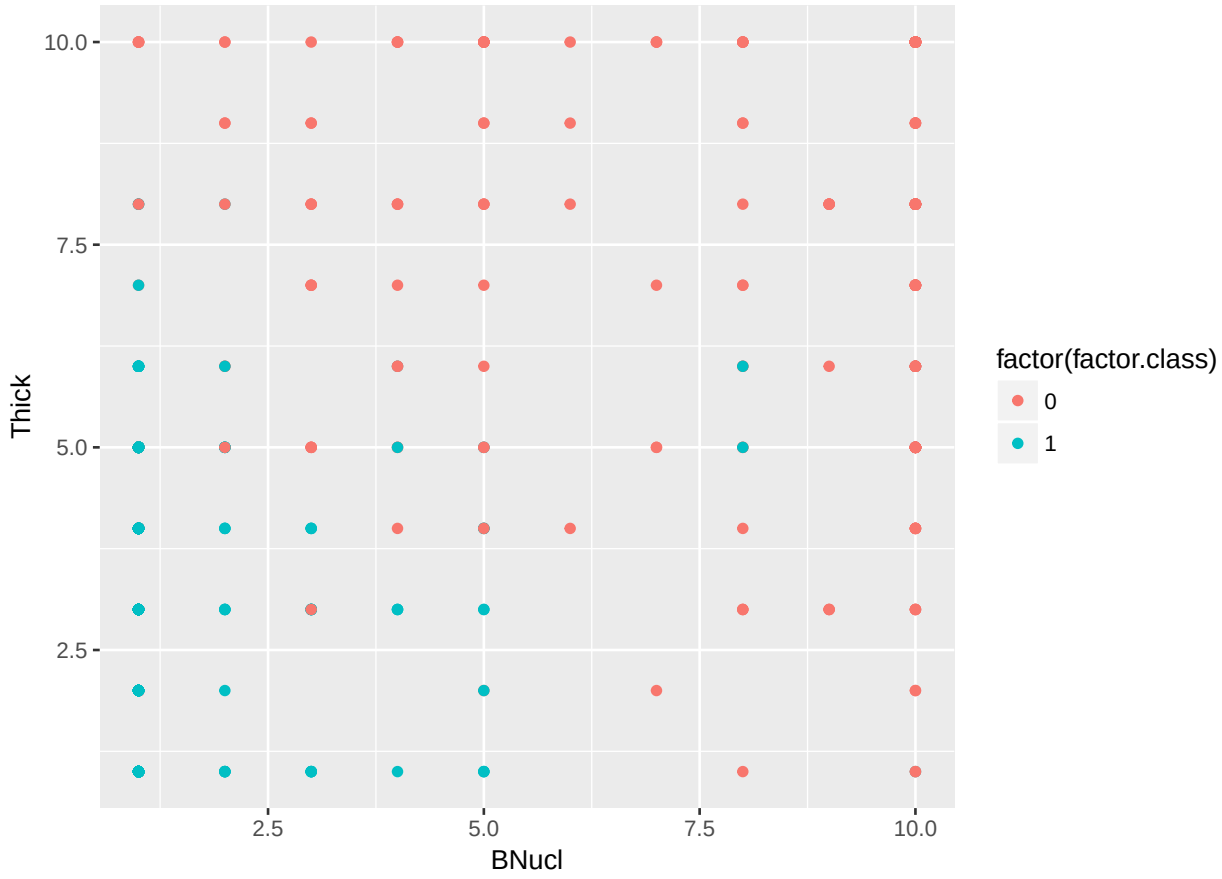


It looks like the $BNucl$ feature may be conditionally (on the class) modeled as a multinomial distribution. Most of the mass for the positive class is located at $BNucl = 1$ while most of the mass for the negative class is located at $BNucl = 10$.


0.3
0.2
0.1
0.0
density
0.0
2.5
5.0
7.5
10.0
BNucl
factor.class
0
1

(b) Produce a version of Figure 2.3 for the predictors BNucl and Thick. Produce an alternative version with only one panel but where the two types are plotted differently. Compare the two plots and describe what they say about the ability to distinguish the two types using these two predictors.





We see that a higher value of $BNucl$ is associated with an elevated value of $Thick$ and that a lower value of $BNucl$ is associated with a lower value of $Thick$. $Thick$ is a good candidate for inclusion in a mode using $BNucl$ to discriminate cancer status.

(c) Fit a binary regression with Class as the response and the other nine variables as predictors. Report the residual deviance and associated degrees of freedom. Can this information be used to determine if this model fits the data? Explain.

```
##
## Call:
## glm(formula = Class ~ ., family = binomial, data = wbca)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.48282  -0.01179   0.04739   0.09678   3.06425
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 11.16678    1.41491   7.892 2.97e-15 ***
## Adhes       -0.39681    0.13384  -2.965  0.00303 **
## BNucl       -0.41478    0.10230  -4.055 5.02e-05 ***
```

```
## Chrom       -0.56456    0.18728  -3.014  0.00257 **
## Epith       -0.06440    0.16595  -0.388  0.69795
## Mitos       -0.65713    0.36764  -1.787  0.07387 .
## NNucl       -0.28659    0.12620  -2.271  0.02315 *
## Thick       -0.62675    0.15890  -3.944 8.01e-05 ***
## UShap       -0.28011    0.25235  -1.110  0.26699
## USize        0.05718    0.23271   0.246  0.80589
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 881.388  on 680  degrees of freedom
## Residual deviance:  89.464  on 671  degrees of freedom
## AIC: 109.46
##
## Number of Fisher Scoring iterations: 8
```

The deviance is used for hypothesis testing in model comparison. Since our response is Bernoulli we can not use the deviance for evaluating goodness of fit. To use the deviance in this setting we would bin the responses to approximate a binomially distributed response

(e) Suppose that a cancer is classified as benign if p $>$ 0.5 and malignant if p $<$ 0.5. Compute the number of errors of both types that will be made if this method is applied to the current data with the reduced model.

```
##    class.predicted
##     FALSE TRUE
##   0   228   10
##   1     9  434
```

We see that the false positive rate is $10/(228 + 10) = 0.04201681$ and the false negative rate is $9/(434 + 9) = 0.02031603$

(f) Suppose we change the cutoff to 0.9 so that p $<$ 0.9 is classified as malignant and p $>$ 0.9 as benign. Compute the number of errors in this case.

```
##    class.predicted
##     FALSE TRUE
##   0   237    1
##   1    16  427
```

We see that the false positive rate is $1/(237 + 1) = 0.004201681$ and the false negative rate is $16/(427 + 16) = 0.03611738$

(h) It is usually misleading to use the same data to fit a model and test its predictive ability. To investigate this, split the data into two parts - assign every third observation to a test set and the remaining two thirds of the data to a training set. Use the training set to determine the model and the test set to assess its predictive performance. Compare the outcome to the previously obtained results.

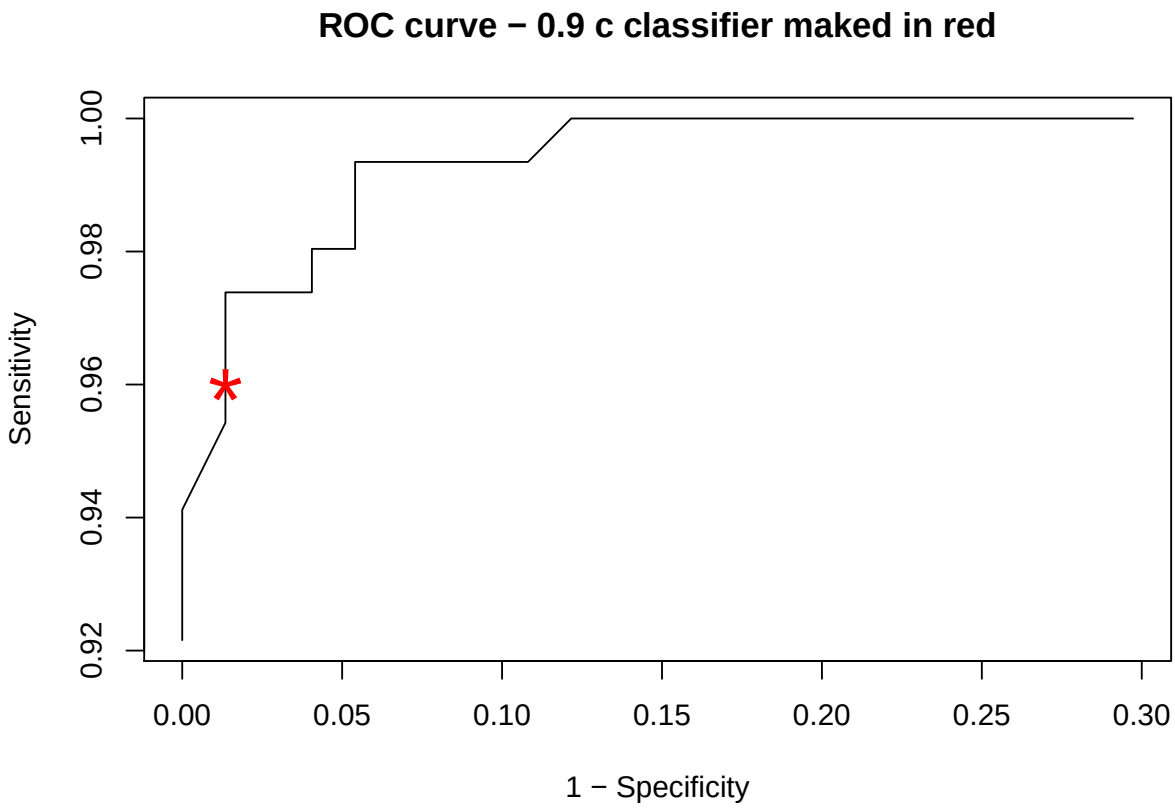
```
##
## Call:
## glm(formula = Class ~ ., family = binomial, data = DFTrain)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3520  -0.0123   0.0406   0.0969   3.1771
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 11.65211    1.84168   6.327  2.5e-10 ***
## Adhes       -0.39057    0.19726  -1.980 0.047708 *
## BNucl       -0.46793    0.13566  -3.449 0.000562 ***
## Chrom       -0.65578    0.23207  -2.826 0.004716 **
## Epith       -0.02961    0.24675  -0.120 0.904469
## Mitos       -0.57934    0.51120  -1.133 0.257094
## NNucl       -0.25630    0.15143  -1.693 0.090543 .
## Thick       -0.80067    0.21174  -3.781 0.000156 ***
## UShap       -0.23802    0.26885  -0.885 0.375988
## USize        0.17773    0.24495   0.726 0.468109
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 593.945  on 453  degrees of freedom
## Residual deviance:  62.044  on 444  degrees of freedom
## AIC: 82.044
##
## Number of Fisher Scoring iterations: 9
```

Table 1: Confusion matrix p=0.9

	FALSE	TRUE
0	73	1
1	6	147

Table 2: Confusion matrix p=0.5

	FALSE	TRUE
0	70	4
1	3	150



We see that we have error rate of $(1+6)/(73+1+6+147) = 0.030837$ using the $p = 0.9$ threshold. Using the threshold $p = 0.5$ we have an accuracy $(4+3)/(70+4+3+150) = 0.030837$ In this case we have very good evidence that the classifier will perform well on new data. We note that our total error rates for the 2 models are the same - we may prefer one over the other based on the class conditional error rate. A ROC curve may help us in model tuning.

NCSU ST 503 Discussion 12

Probem 2.5 Faraway, Julian J. Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models CRC Press.

Bruce Campbell

2.5 spector data analysis

We investigate the efficacy of a new method for teaching economics. The data has the following variables;

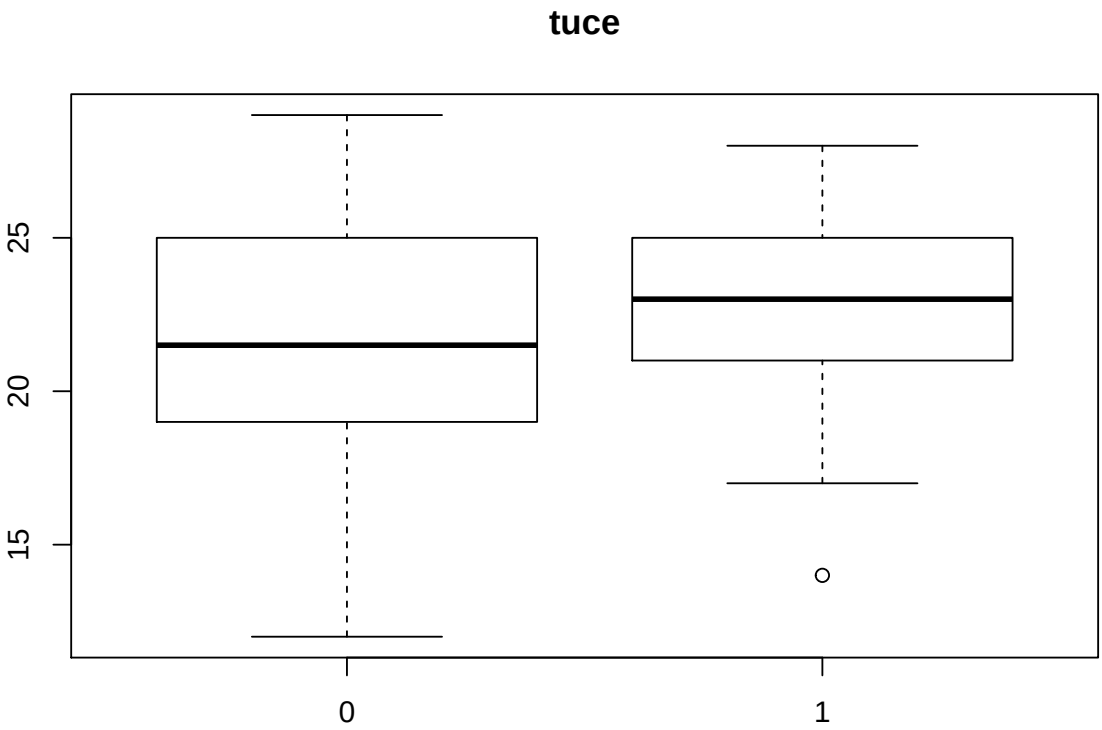
- grade 1 = exam grades improved, 0 = not improved
- psi 1 = student exposed to PSI (a new teach method), 0 = not exposed
- tuce a measure of ability when entering the class
- gpa grade point average

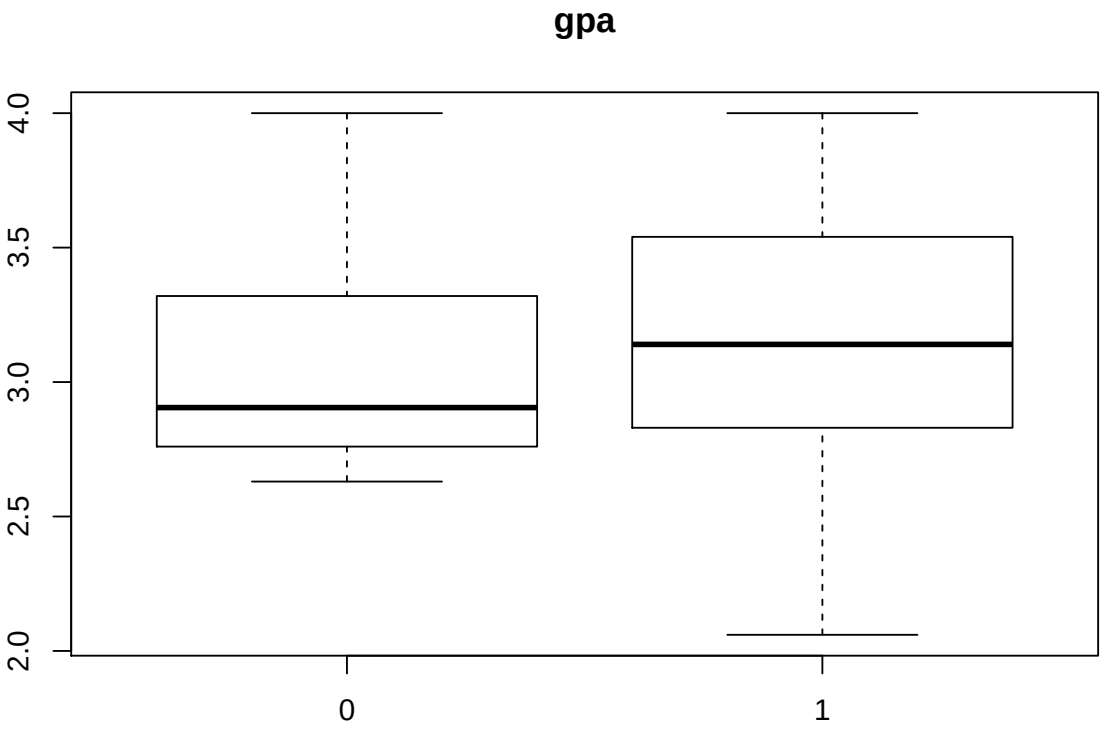
The data originates from

Spector, L. and Mazzeo, M. (1980), "Probit Analysis and Economic Education", Journal of Economic Education, 11, 37 - 44.

We will fit a logistic model with response *grade* and predictors $psi, tuce, gpa$

Below are box plots of the variables $tuce, gpa$ by the category *psi* We expect that the *tuce* and *gpa* are equally distributed among the psi class . We also display a pivot of the grade by psi. The association between psi and grade is not prefect and we anticipate that the tuce and gpa predictors will help explain the relationship between psi and grade..


tuce
25
20
15
0
1


gpa
4.0
3.5
3.0
2.5
2.0
0
1

	0	1
0	15	6

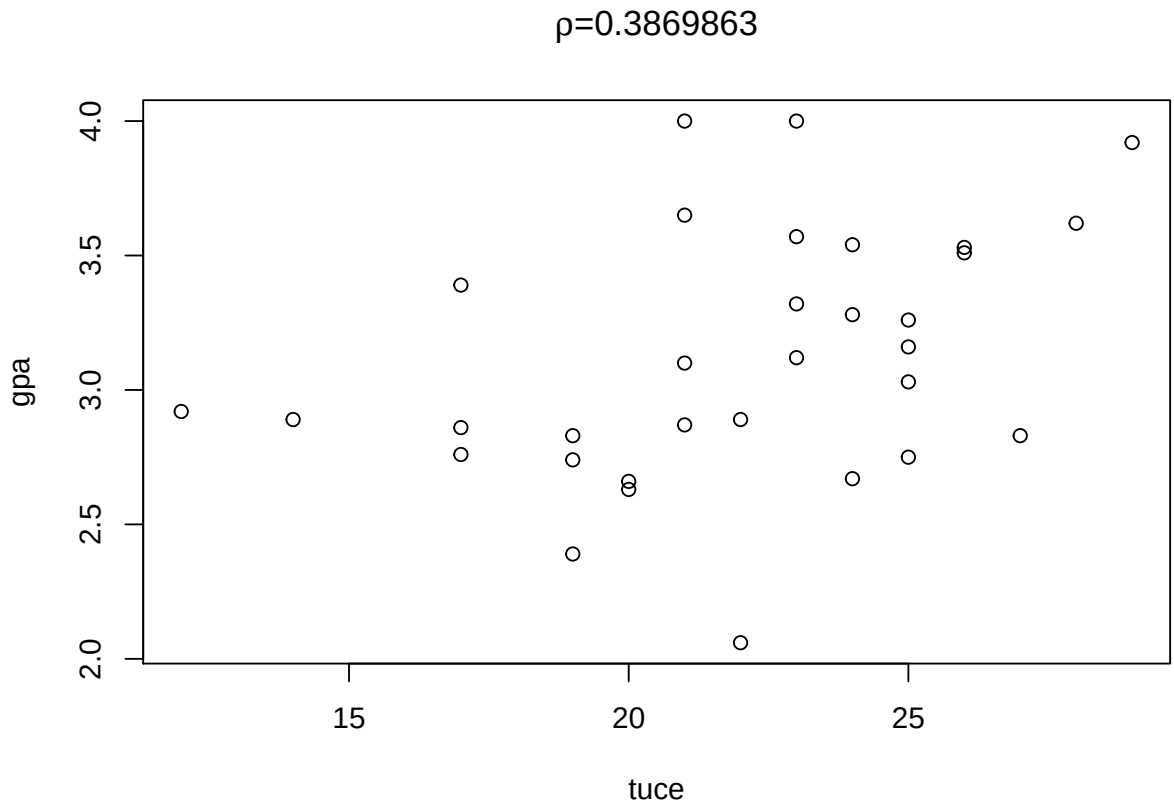
1 3 8

Table: pivot showing improved or not by the psi variable

We observe that the levels of tuce and gpa for the students exposed to the new method are systematically higher than those for students not exposed to the new teaching method. This may affect our conclusions. We might look into the possibility of weighting to alleviate any bias from the design.

```
##
## Call:
## glm(formula = grade ~ ., family = binomial, data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9551  -0.6453  -0.2570   0.5888   2.0966
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -13.02135    4.93127  -2.641  0.00828 **
## psi           2.37869    1.06456   2.234  0.02545 *
## tuce          0.09516    0.14155   0.672  0.50143
## gpa           2.82611    1.26293   2.238  0.02524 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 41.183  on 31  degrees of freedom
## Residual deviance: 25.779  on 28  degrees of freedom
## AIC: 33.779
##
## Number of Fisher Scoring iterations: 5
```

We see that the tuce variable is not significant. We'll remove that variable from our model. The large s.e. suggests collinearity. A plot of *tuce gpa* confirms weak collinearity.

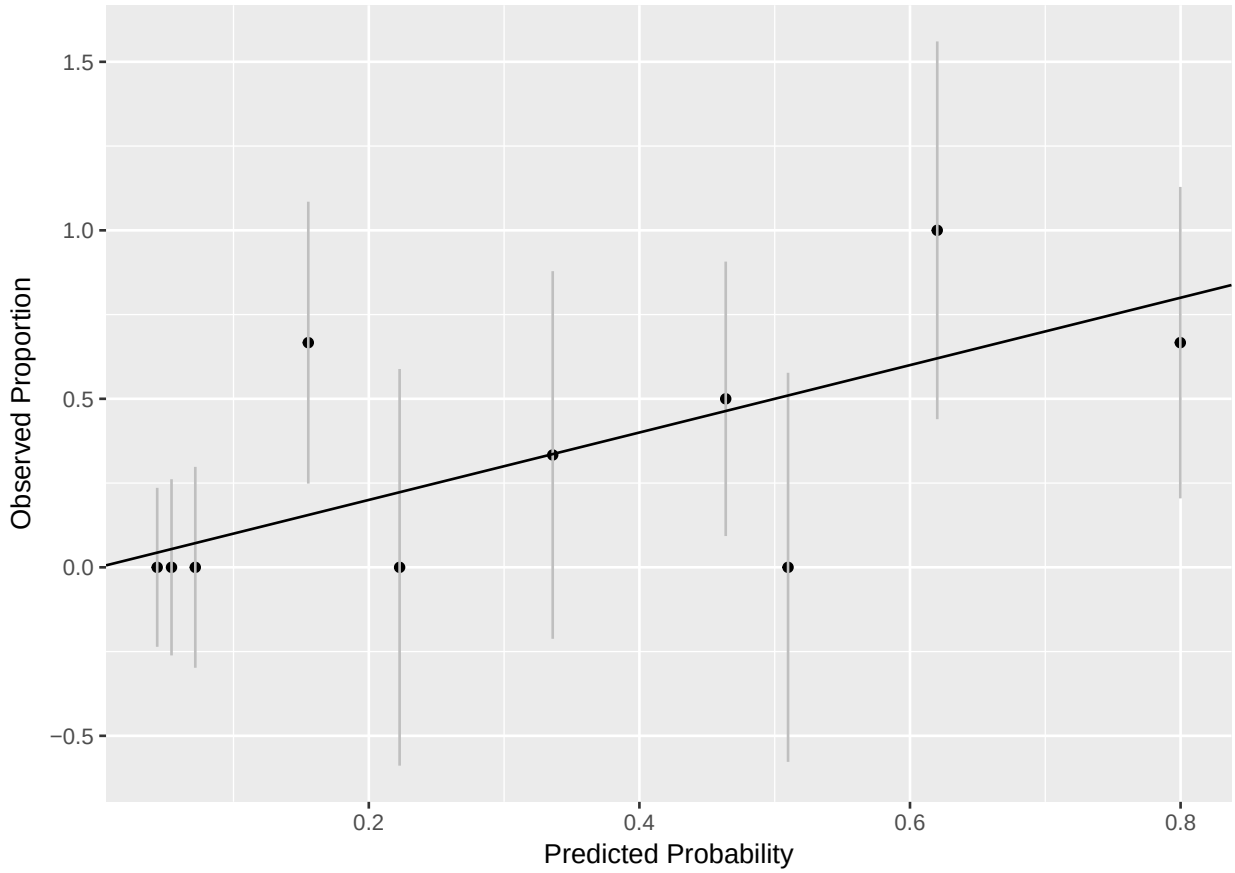


Refitting the model $grade \sim psi + gpa$

```
##
## Call:
## glm(formula = grade ~ psi + gpa, family = binomial, data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8396  -0.6282  -0.3045   0.5629   2.0378
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -11.602      4.213  -2.754  0.00589 **
## psi            2.338      1.041   2.246  0.02470 *
## gpa            3.063      1.223   2.505  0.01224 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 41.183  on 31  degrees of freedom
```

```
## Residual deviance: 26.253  on 29  degrees of freedom
## AIC: 32.253
##
## Number of Fisher Scoring iterations: 5
```

We now visualize the binned response and prepare to calculate the The Hosmer-Lemeshow statistic.

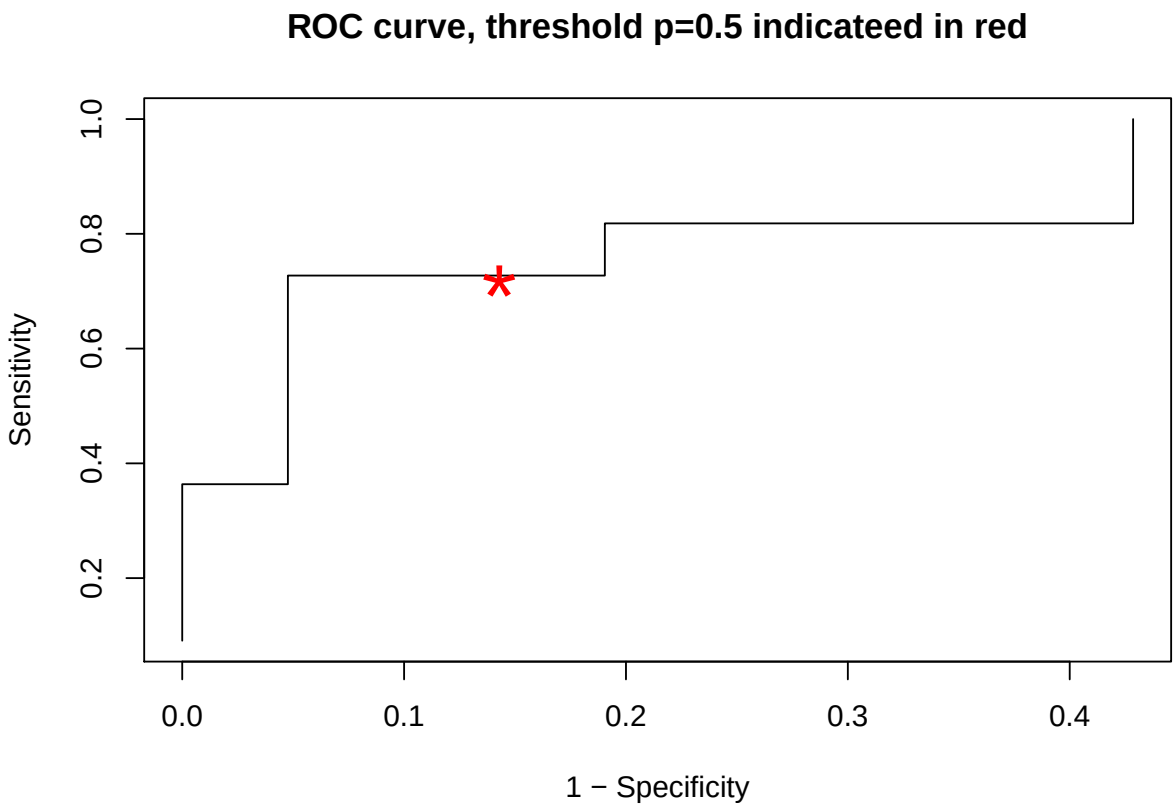


Hosmer.Lemeshow
0.1908

From the observed and predicted binned probabilities and the moderate value of the Hosmer Lemeshow statistic, we conclude that there is no evidence of a significant lack of fit.

Table 3: Training set accuracy

	FALSE	TRUE
0	18	3
1	3	8



We conclude that there is evidence that the new training method has a positive effect in grade outcome.

NCSU ST 503 Discussion 13

Probem 7.1 Faraway, Julian J. Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models CRC Press.

Bruce Campbell

7.1

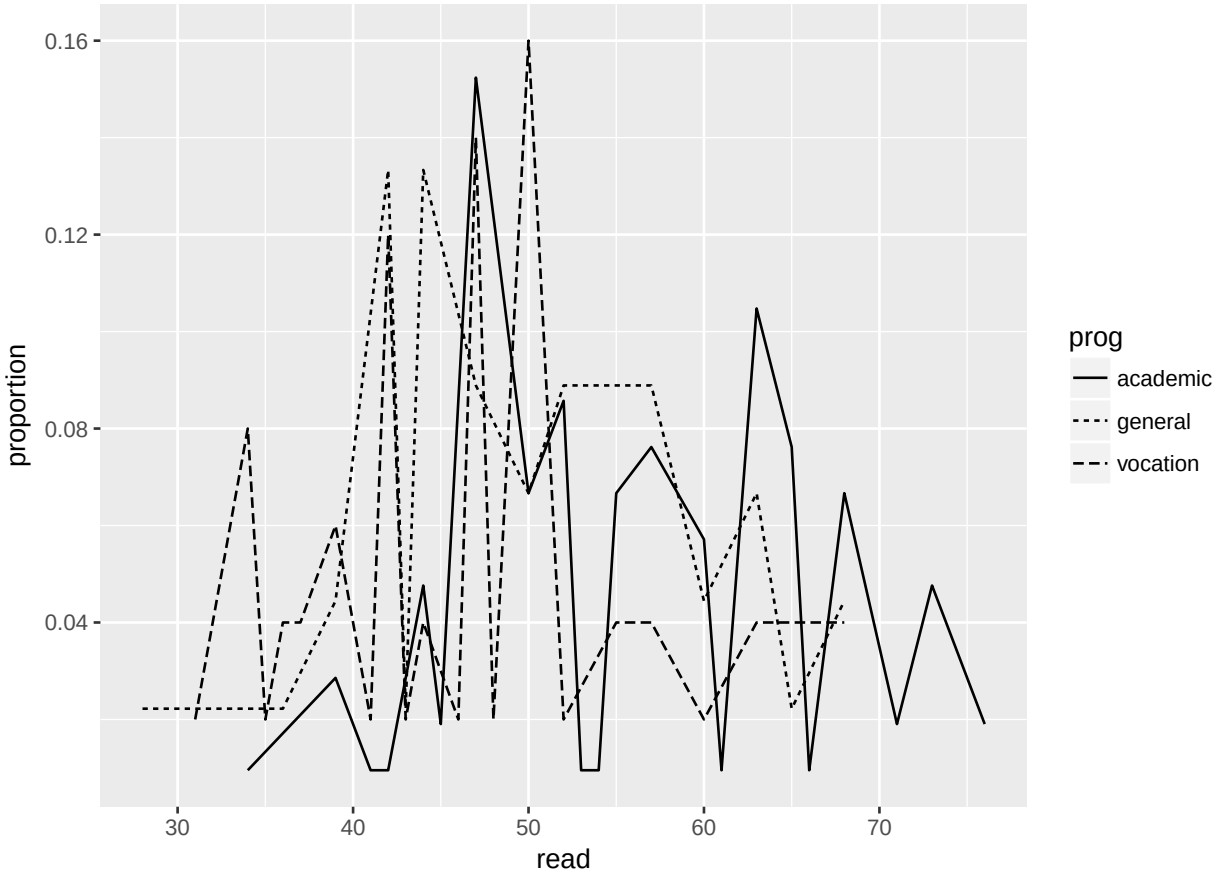
The hsb data was collected as a subset of the High School and Beyond study conducted by the National Education Longitudinal Studies program of the National Center for Education Statistics. The variables are gender; race; socioeconomic status (SES); school type; chosen high school program type; scores on reading, writing, math, science, and social studies. We want to determine which factors are related to the choice of the type of program - academic, vocational or general - that the students pursue in high school. The response is multinomial with three levels.

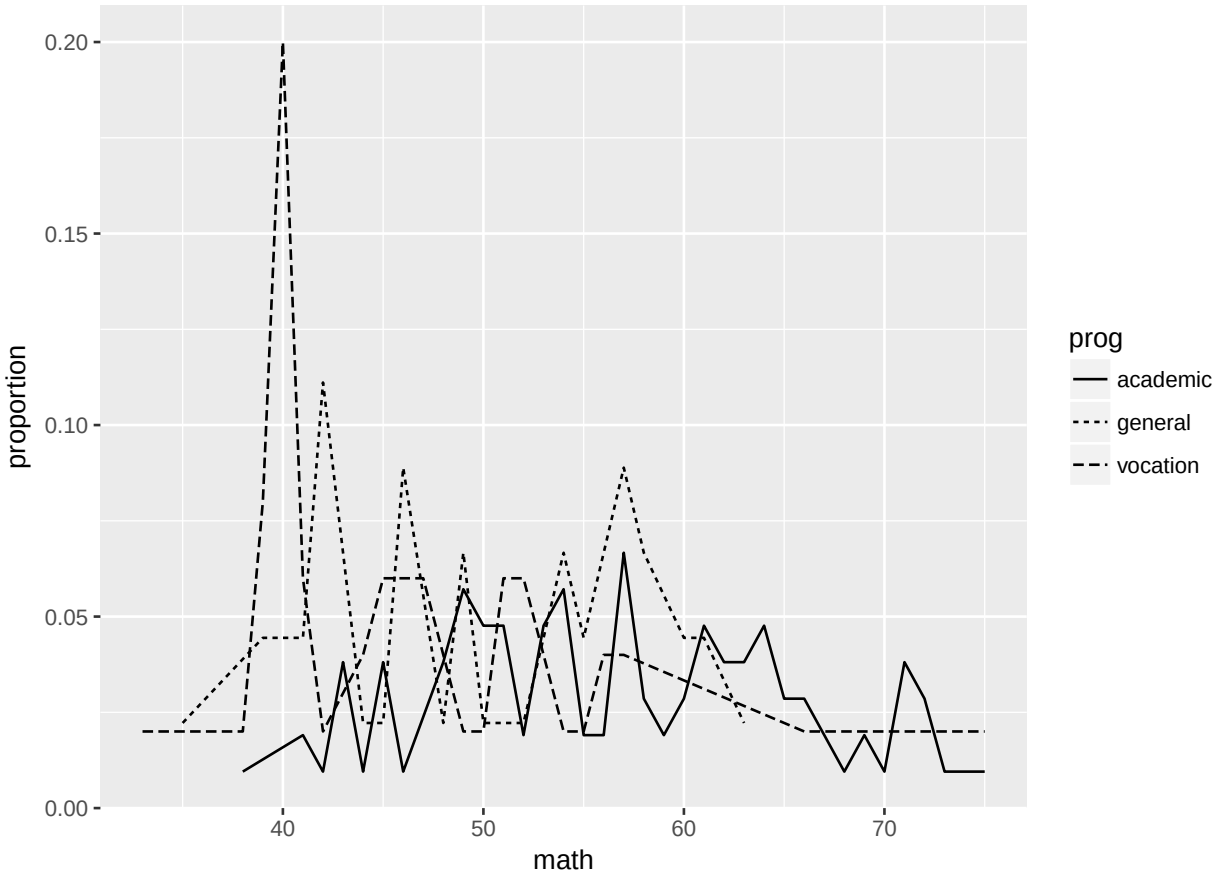
(a) Make a table showing the proportion of males and females choosing the three different programs. Comment on the difference. Repeat this comparison but for SES rather than gender.

```
##           gender
## prog       female male
##   academic     58   47
##   general      24   21
##   vocation     27   23
```

```
##           ses
## prog       high low middle
##   academic   42  19     44
##   general     9  16     20
##   vocation    7  12     31
```

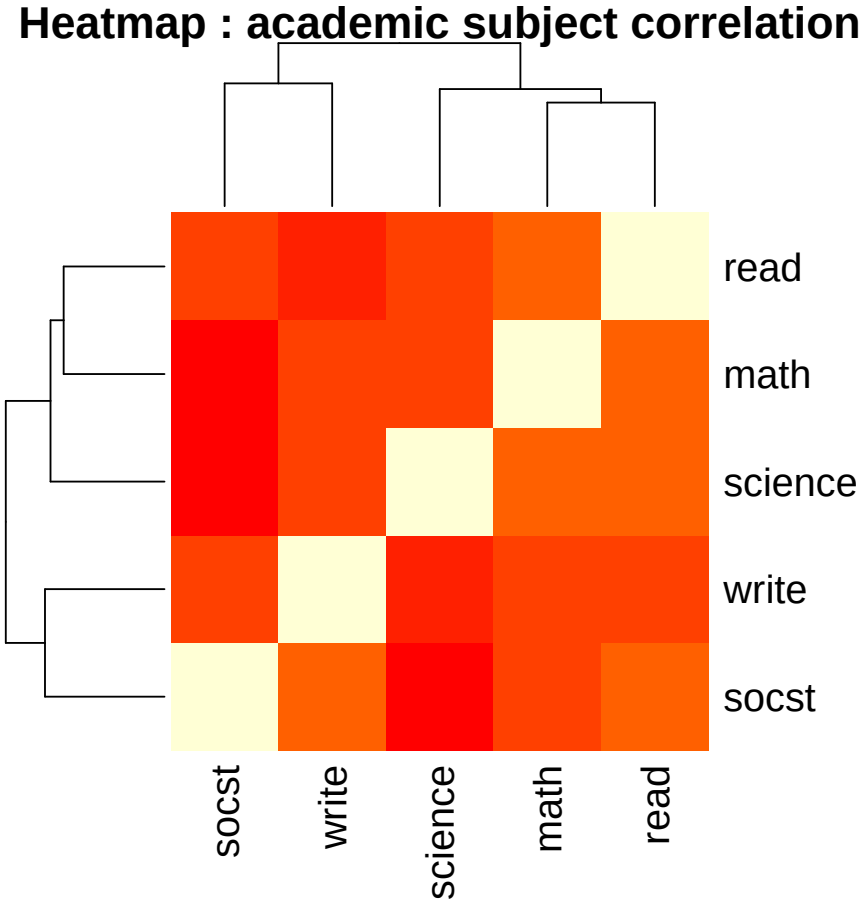
(b) Construct a plot like the right panel of Figure 7.1 that shows the relationship between program choice and reading score. Comment on the plot. Repeat for math in place of reading.





(c) Compute the correlation matrix for the five subject scores.

	read	write	math	science	socst
read	1	0.5968	0.6623	0.6302	0.6215
write	0.5968	1	0.6174	0.5704	0.6048
math	0.6623	0.6174	1	0.6307	0.5445
science	0.6302	0.5704	0.6307	1	0.4651
socst	0.6215	0.6048	0.5445	0.4651	1



(d) Fit a multinomial response model for the program choice and examine the fitted coefficients. Of the five subjects, one gives unexpected coefficients. Identify this subject and suggest an explanation for this behavior.

```
## # weights:  45 (28 variable)
## initial  value 219.722458
## iter  10 value 181.098338
## iter  20 value 154.577078
## iter  30 value 152.478856
## final  value 152.478368
## converged
```

```
## Call:
## multinom(formula = prog ~ ., data = df)
##
## Coefficients:
##           (Intercept)           id  gendermale  raceasian racehispanic
## general     4.263658 -0.007332836 -0.04666403  1.2170225   -0.8702109
## vocation    7.845921 -0.003680462 -0.29724832 -0.7863428   -0.3236628
##           racewhite    seslow sesmiddle schtyppublic        read
```

```
## general  0.8609754 1.1547399 0.7430976    0.1384853 -0.05445264
## vocation 0.6223190 0.0728241 1.1897765    1.8285649 -0.04078359
##                write       math   science       socst
## general  -0.03716360 -0.1037470 0.1065258 -0.01786542
## vocation -0.03220268 -0.1099712 0.0537472 -0.07959798
##
## Std. Errors:
##          (Intercept)          id gendermale raceasian racehispanic
## general     1.960941 0.007678009  0.4587870  1.064969    0.9286986
## vocation    2.288984 0.008408855  0.5048241  1.476435    0.8924359
##          racewhite    seslow sesmiddle schtyppublic       read      write
## general  0.9438010 0.6134530 0.5096129    0.7338284 0.03300204 0.03398842
## vocation 0.9519097 0.7067682 0.5739217    0.9981540 0.03583547 0.03597627
##                math    science      socst
## general  0.03556357 0.03331314 0.02737227
## vocation 0.03885464 0.03445137 0.02963317
##
## Residual Deviance: 304.9567
## AIC: 360.9567
```

(e) Construct a derived variable that is the sum of the five subject scores. Fit a multinomial model as before except with this one sum variable in place of the five subjects separately. Compare the two models to decide which should be preferred.

```
## # weights:  33 (20 variable)
## initial  value 219.722458
## iter  10 value 167.158173
## iter  20 value 164.141699
## final  value 164.130704
## converged
```

```
## Call:
## multinom(formula = prog ~ id + gender + race + ses + schtyp +
##     sum.subject, data = df.reduced)
##
## Coefficients:
##          (Intercept)           id  gendermale  raceasian racehispanic
## general     3.227335 -0.003708235  0.24883040  1.0243408   -0.5484976
## vocation    7.112010 -0.003220142 -0.09614882 -0.6015843   -0.1937564
##          racewhite    seslow sesmiddle schtyppublic sum.subject
## general   1.060033 1.0593830 0.6350558    0.3875245 -0.02052599
## vocation  1.098265 0.2517821 1.1874930    1.8098161 -0.04125543
##
```

```
## Std. Errors:
##          (Intercept)          id gendermale raceasian racehispanic
## general     1.798815 0.006823237  0.3941480 0.9439661    0.8799224
## vocation    2.157426 0.007659938  0.4364287 1.3769618    0.8411264
##          racewhite    seslow sesmiddle schtyppublic sum.subject
## general  0.8740777 0.5664146 0.4789630    0.6826598 0.005976099
## vocation 0.8970833 0.6797684 0.5566371    0.9568939 0.007225491
##
## Residual Deviance: 328.2614
## AIC: 368.2614
```

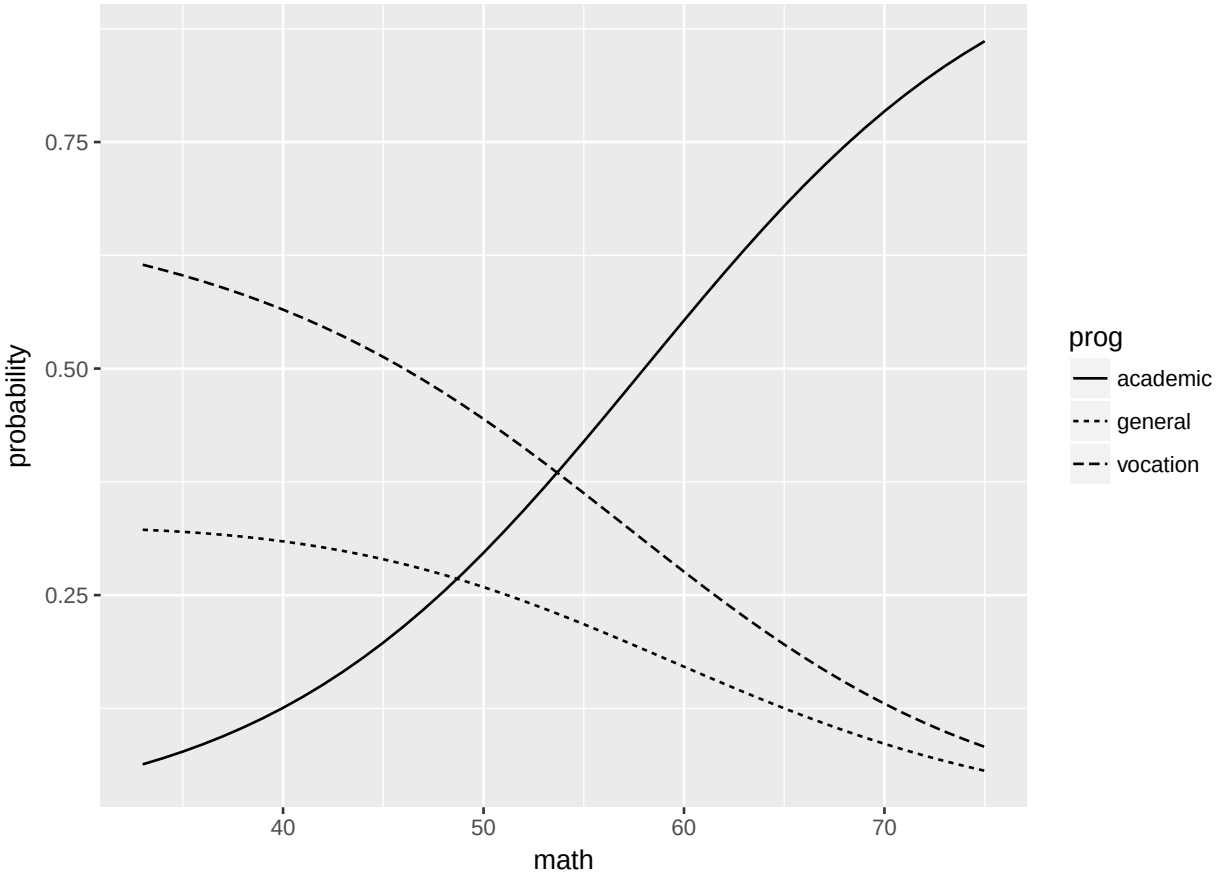
The s.e. for the combined subject variable is much lower than the single subjet variables. We suspect colliniearity may be the cause.

(f) Use a stepwise method to reduce the model. Which variables are in your selected model?

```
## Call:
## multinom(formula = prog ~ ses + schtyp + sum.subject, data = df.reduced)
##
## Coefficients:
##          (Intercept)    seslow sesmiddle schtyppublic sum.subject
## general     2.593944 0.8078324 0.5808536    0.5594952 -0.01635887
## vocation    6.372051 0.1330839 1.1517240    1.8490860 -0.03681150
##
## Std. Errors:
##          (Intercept)    seslow sesmiddle schtyppublic sum.subject
## general     1.587502 0.5386033 0.4720925    0.5219044 0.005422494
## vocation    1.877764 0.6468558 0.5465572    0.7974692 0.006553295
##
## Residual Deviance: 336.0554
## AIC: 356.0554
```

We see that there are 3 variables in the best model : $ses + schtyp + sum.subjec$

(g) Construct a plot of predicted probabilities from your selected model where the math score varies over the observed range. Other predictors should be set at the most common level or mean value as appropriate. Your plot should be similar to Figure 7.2. Comment on the relationship.



NCSU ST 503 Discussion 14

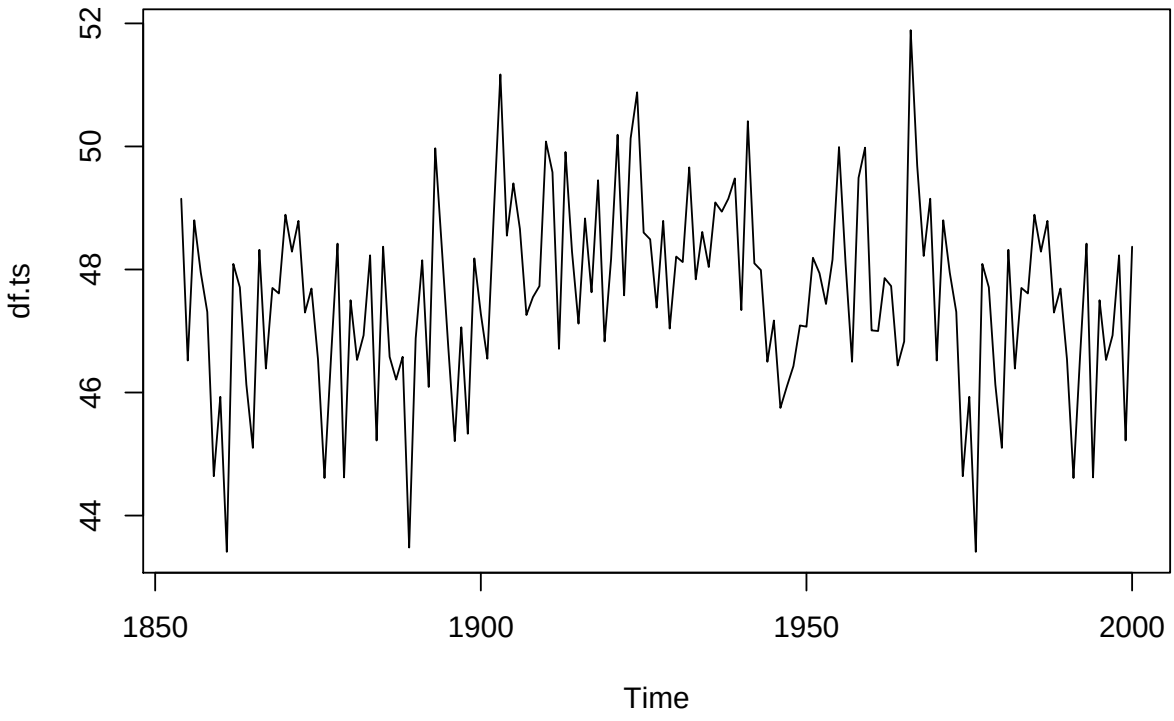
Probem 14.5 Faraway, Julian J. Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models CRC Press.

Bruce Campbell

14.5 temp data analysis

The aatemp data comes from the U.S. Historical Climatology network. They are the annual mean temperatures (in degrees Fahrenheit) in Ann Arbor, Michigan, going back about 150 years.

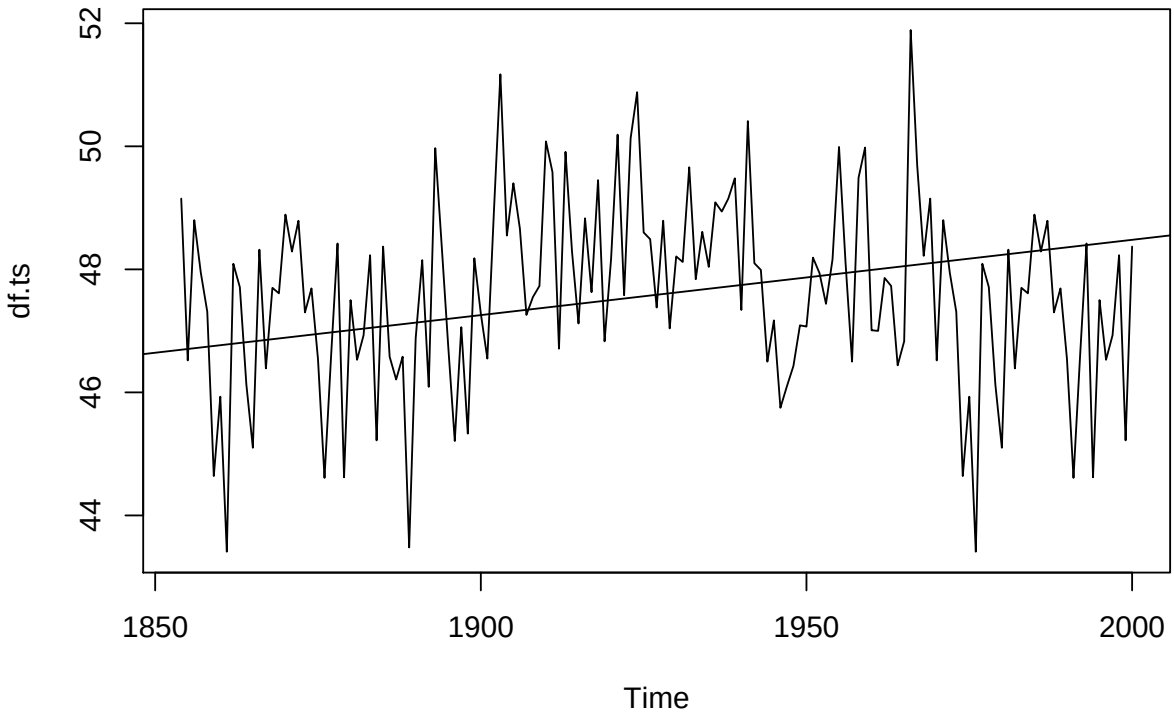
(a) Plot the temperature as a function of time and comment on the underlying trend.



It appears that the overall trend is rising and then falling. There are some higher frepency fluctuations as well.

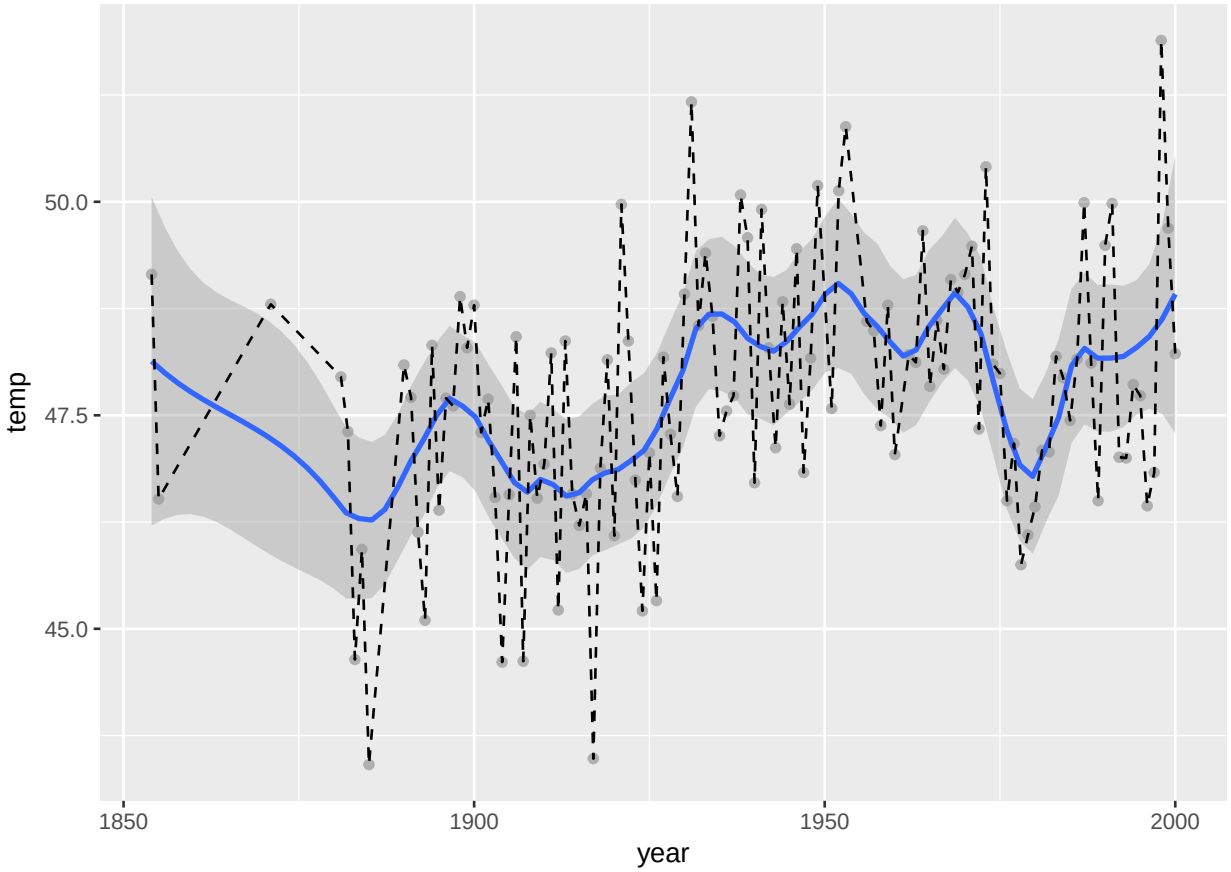
(b) Fit a least squares line to the data and test whether the slope of the line is different from zero. What is the main drawback of this modeling approach?

```
##
## Call:
## lm(formula = temp ~ year, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9843 -0.9113 -0.0820  0.9946  3.5343
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 24.005510   7.310781   3.284  0.00136 **
## year         0.012237   0.003768   3.247  0.00153 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.466 on 113 degrees of freedom
## Multiple R-squared:  0.08536,    Adjusted R-squared:  0.07727
## F-statistic: 10.55 on 1 and 113 DF,  p-value: 0.001533
```



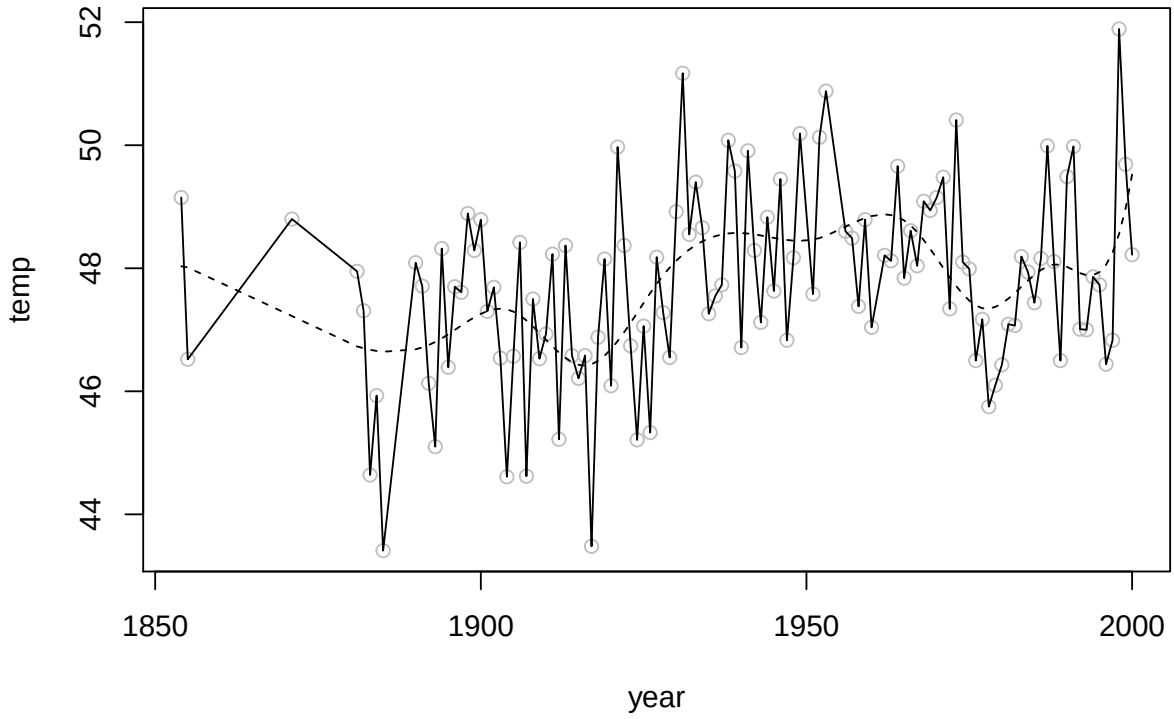
We see that the slope is significant, but none of the nonlinear components are captured in this model.

(c) Fit a Lowess curve to the data using the default amount of smoothing. Display the fit along with a 95% confidence band. What does this say about the underlying trend in the relationship?



This fit captures more of the nonlinearities in the data. The confidence bands are pointwise, the distance between them is determined by the variability of the points in the neighborhood about the band.

(d) Fit a regression spline basis to the data with 12 knots. Display the fit on the data.



(e) Compare this model to the linear fit using an F-test. Which model is preferred? What more needs to be explored with spline fit before drawing conclusions?

```
## Analysis of Variance Table
##
## Model 1: temp ~ year
## Model 2: temp ~ bs(year, 12)
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    113 242.94
## 2    102 199.67 11    43.275 2.0097 0.03476 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```