Bruce Campell ST 503 HW 1

Problems 1,3 Chapter 2 Faraway, Julian J.. Linear Models with R, Second Edition (Chapman & Hall/CRC Texts in Statistical Science). CRC Press.

Bruce Campbell

26 August, 2017

Sat Aug 26 18:34:13 2017

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

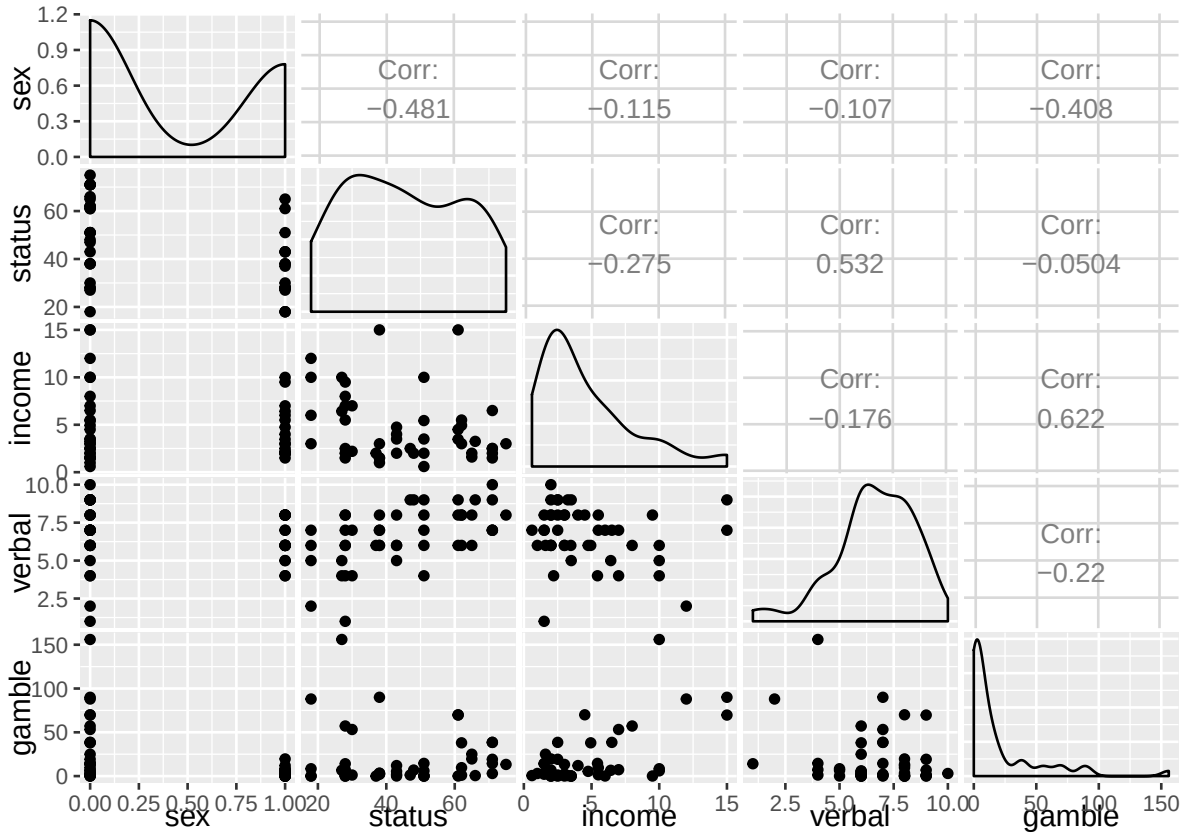
- (a) What percentage of variation in the response is explained by these predictors?
- (b) Which observation has the largest (positive) residual? Give the case number.
- (c) Compute the mean and median of the residuals.
- (d) Compute the correlation of the residuals with the fitted values.
- (e) Compute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

```
if (!require(faraway)) {
    install.packages("faraway")
    library(faraway)
}
```

```
library(pander)
library(ggplot2)
library(GGally)
data(teengamb, package = "faraway")
head(teengamb)
```

```
##   sex status income verbal gamble
## 1   1     51   2.00      8    0.0
## 2   1     28   2.50      8    0.0
## 3   1     37   2.00      6    0.0
## 4   1     28   7.00      4    7.3
## 5   1     65   2.00      8   19.6
## 6   1     61   3.47      6    0.1
```

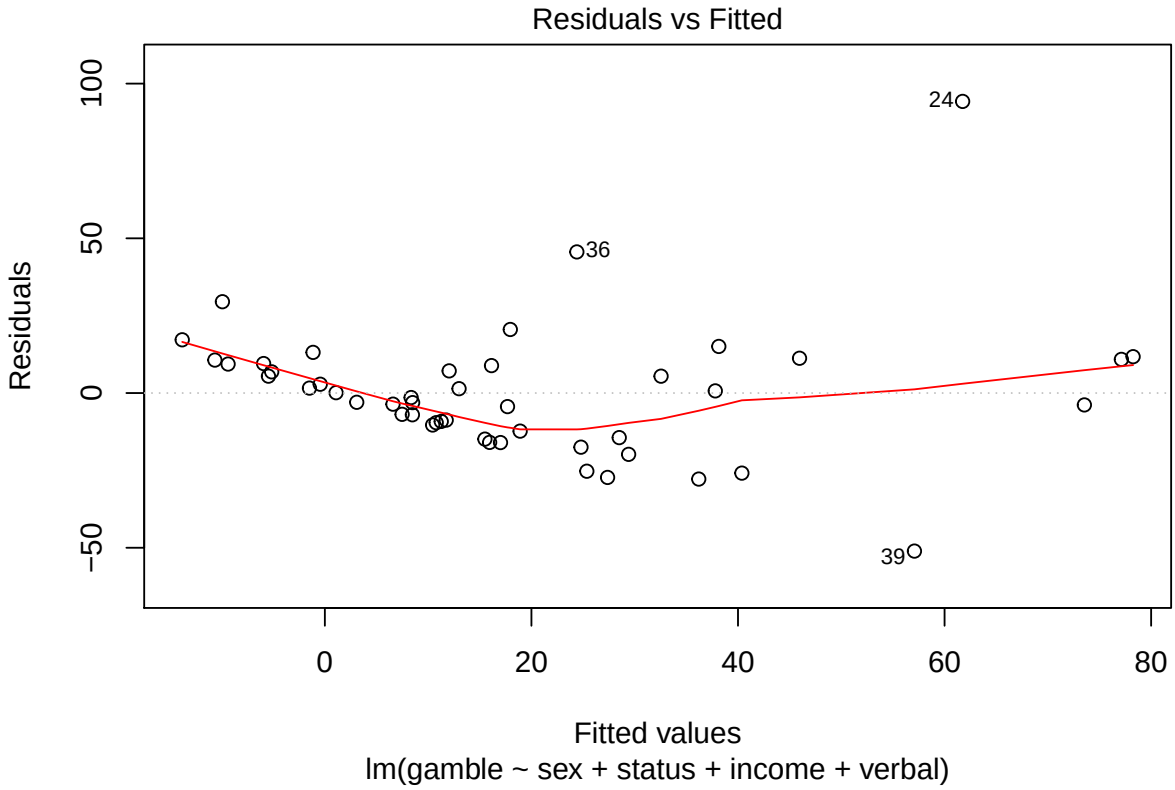
```
ggpairs(teengamb)
```

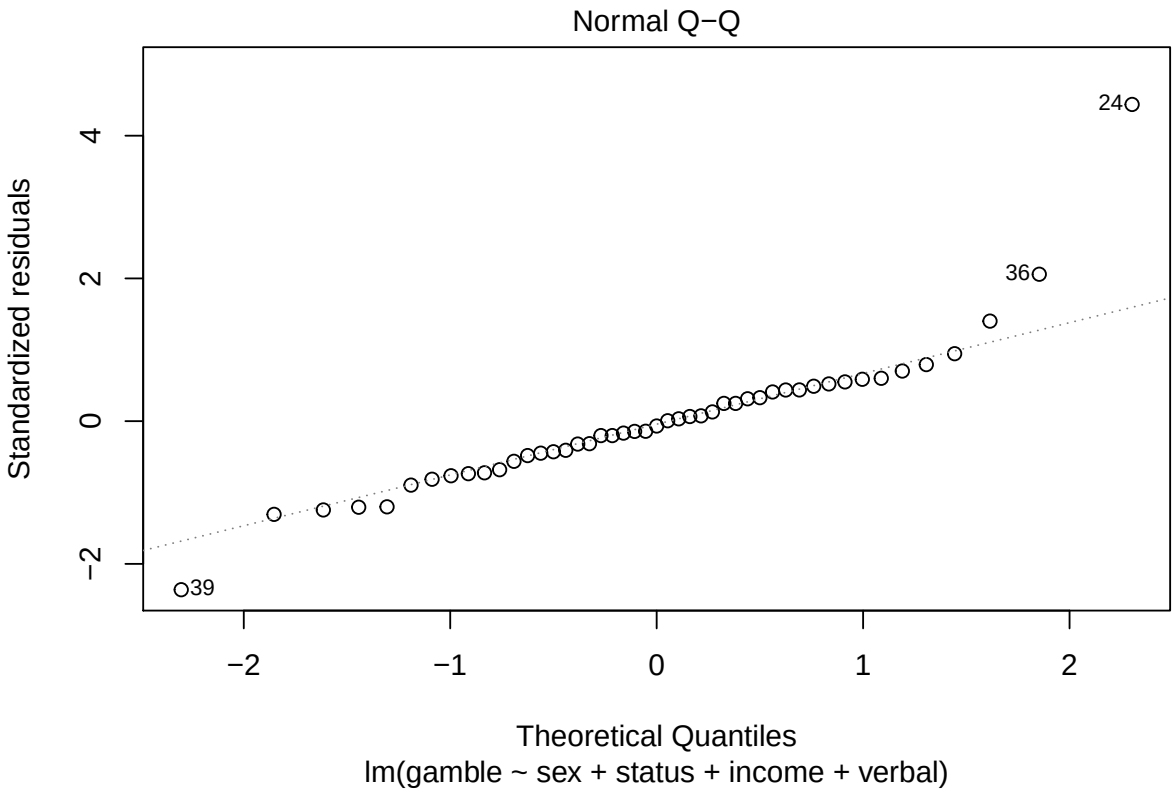


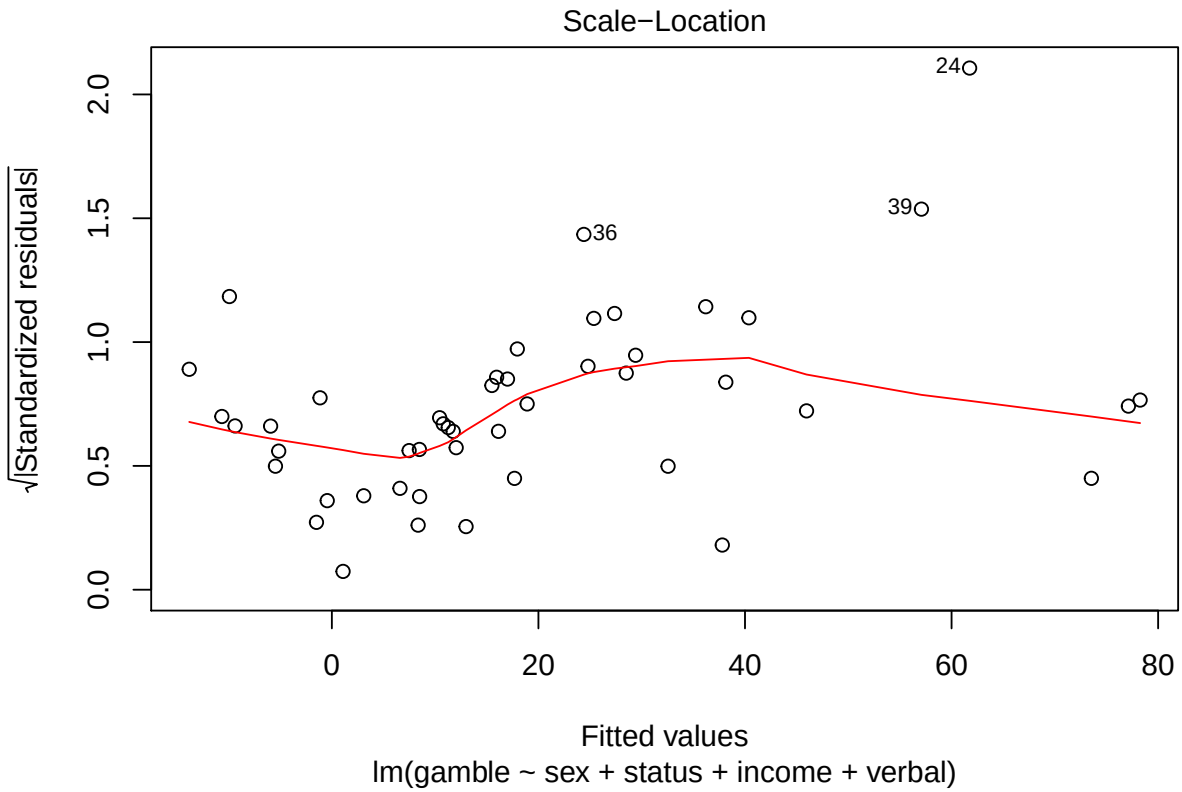

```
lm.fit <- lm(gamble ~ sex + status + income + verbal, data = teengamb)

summary(lm.fit)
```

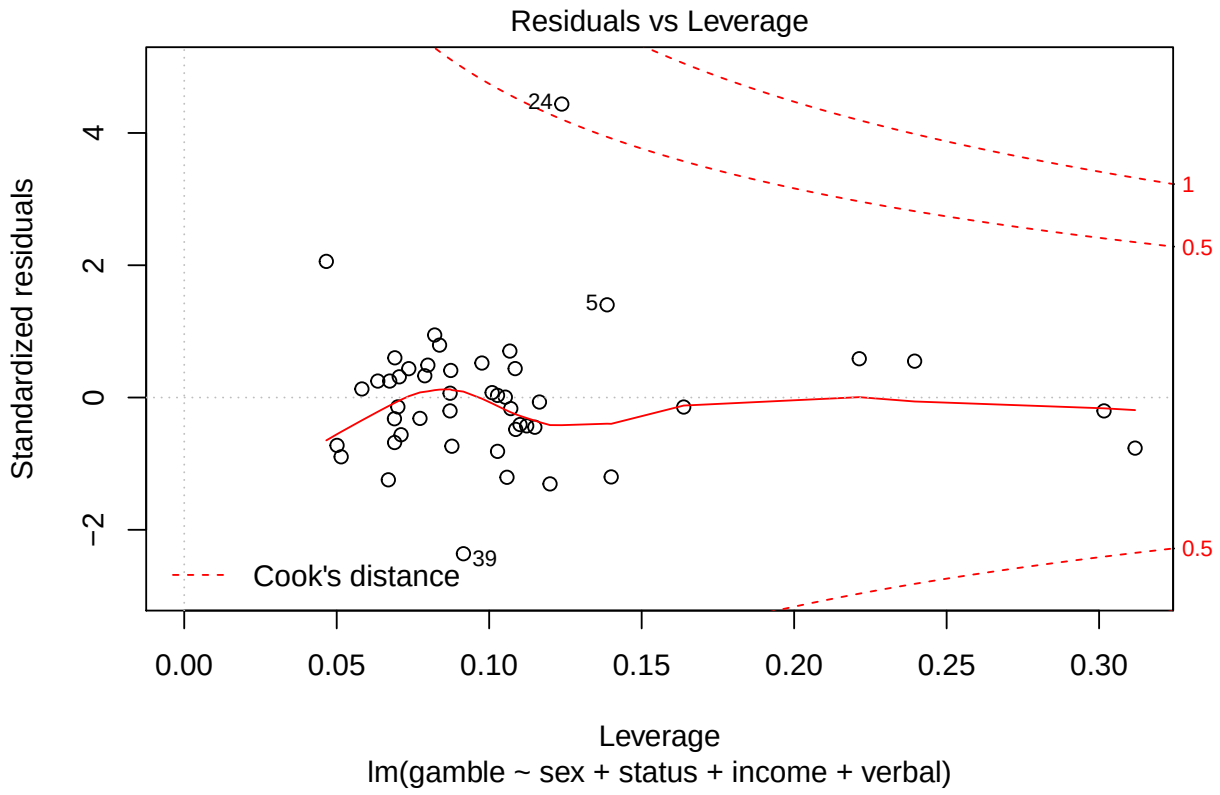
```
##
## Call:
## lm(formula = gamble ~ sex + status + income + verbal, data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.082 -11.320  -1.451   9.452  94.252
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.55565   17.19680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

```r
plot(lm.fit)
```


Residuals vs Fitted
Residuals
100
50
0
−50
24
36
39
0
20
40
60
80
Fitted values
lm(gamble ~ sex + status + income + verbal)


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(gamble ~ sex + status + income + verbal)
24
36
39
−2
0
2
4
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
24
39
36
0
20
40
60
80
Fitted values
lm(gamble ~ sex + status + income + verbal)



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

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

```
var.explained.proportion <- summary(lm.fit)$r.squared
var.unexplaned.proportion <- 1 - summary(lm.fit)$r.squared

pander(data.frame(var.explained.proportion = var.explained.proportion), caption = "Explained")
```

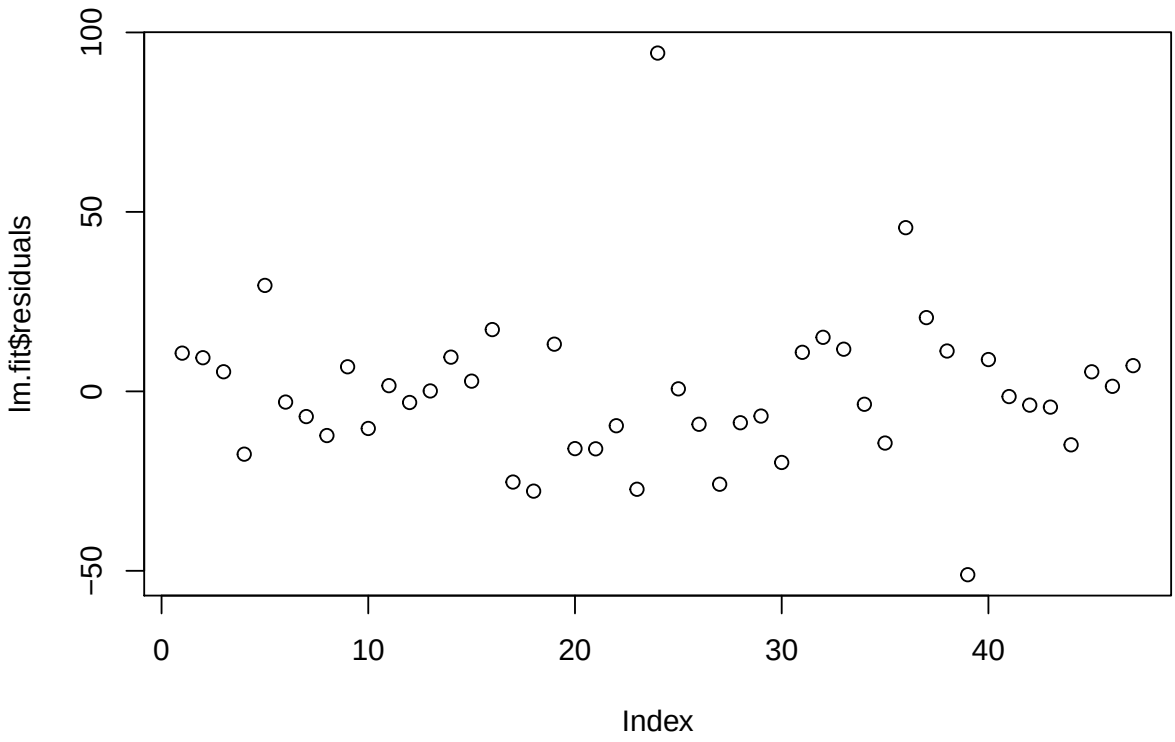
Table 1: Explained

var.explained.proportion
0.5267

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

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

```
plot(lm.fit$residuals)
```




```
index.largest.pos.residual <- which.max(lm.fit$residuals)

index.largest.abs.residual <- which.max(abs(lm.fit$residuals))
```

The largets residual occurs at index 24 of the dataframe. This is the associated cases data.

```
pander(teengamb[24, ], caption = "Potential outlier.")
```

Table 2: Potential outlier.

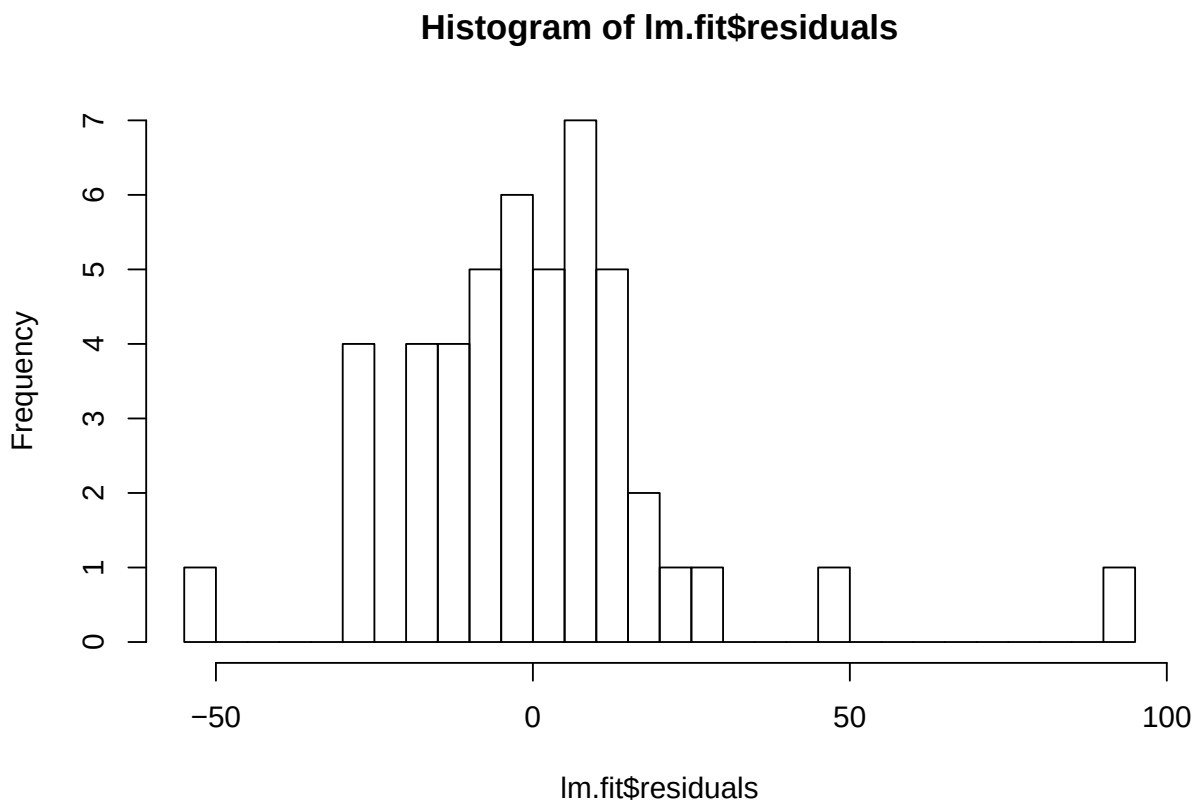
	sex	status	income	verbal	gamble
24	0	27	10	4	156

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

```
residuals.mean <- mean(lm.fit$residuals)

residuals.median <- median(lm.fit$residuals)

pander(data.frame(residuals.mean = residuals.mean, residuals.median = residuals.median))
```

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

```
hist(lm.fit$residuals, 30)
```



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

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

```
corr.residuals.vs.fitted <- cor(lm.fit$residuals, lm.fit$fitted.values)
pander(data.frame(corr.residuals.vs.fitted = corr.residuals.vs.fitted))
```

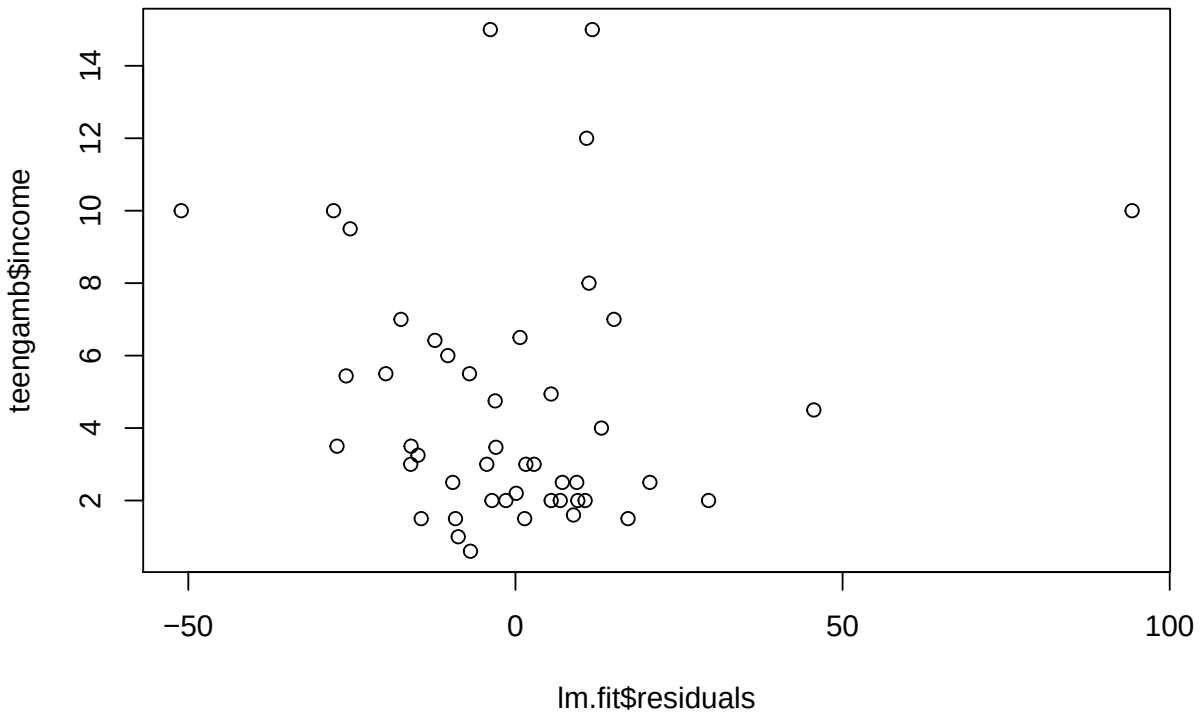
corr.residuals.vs.fitted
-1.071e-16

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

```
corr.residuals.income <- cor(lm.fit$residuals, teengamb$income)
pander(data.frame(corr.residuals.income = corr.residuals.income))
```

corr.residuals.income
-7.242e-17

```
plot(lm.fit$residuals, teengamb$income)
```



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

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

```
gender.coefficient <- lm.fit$coefficients["sex"]

pander(data.frame(gender.coefficient = gender.coefficient))
```

	gender.coefficient
sex	-22.12

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

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

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

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

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

Table 7: Data sample

	intercept	sex	status	income	verbal
30	1	0	62	5.5	8

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

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


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

Table 8: Data sample with gender modified

	intercept	sex	status	income	verbal
30	1	1	62	5.5	8

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


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

	response.difference
30	22.12