$E(y) = X\beta$, $r(X) = p$, $h_{ii} = \left(X(X'X)^{-1}X'\right)_{ii}$. $S^2_{(i)} = \hat{\sigma}$ without ith case.

t_i = externally studentized residual

r_i = internally studentized residual

1. Mean shift model: $EY = X\beta + \theta \cdot \phi_i$, $\phi_i = [0, \dots 1, \dots 0]'$

Derive a test to determine if ith case is outlying, and show: the test statistic is the i^{th} externally Studentized statistic; and its square $\sim F(1, n-p-1)$

Soln: full model: $EY = (X \mid \phi_i)\begin{pmatrix}\beta \\ \theta\end{pmatrix} + \varepsilon$, H_0: ith observation is outlier.

reduced model: $EY = X\beta + \varepsilon$, H_1: without outlier, linear restriction $\theta = 0$

general F-test: under H_0: $f = \dfrac{(RSS_H - RSS)/1}{RSS/(n-p-1)}$, $f \sim F(1, n-p-1)$

$RSS_H = Y'(I - P_{(X)})Y = Y'Q_{(X)}Y$

$RSS = Y'(I - P_{(X|\phi_i)})Y$

$I - P_{(X|\phi_i)} = I - P_{(X)} - Q_{(X)}\phi_i(\phi_i'Q_{(X)}\phi_i)^{-1}\phi_i'Q_{(X)}$

$= I - P_{(X)} - Q_{(X)}\phi_i(1-h_{ii})^{-1}\phi_i'Q_{(X)}$

$= Q_{(X)} - Q_{(X)}\,\mathrm{diag}(0 \dots \overset{(i)}{1/(1-h_{ii})} \dots 0)\cdot Q_{(X)}$

$RSS_H - RSS = Y'(Q_{(X)} \cdot \mathrm{diag}(0 \dots 1/(1-h_{ii}), \dots 0))Y$

$= e_i^2/(1-h_{ii})$

$RSS = Y'Q_{(X)}Y - e_i^2/(1-h_{ii})$

$= (n-p)\cdot S^2 - e_i^2/(1-h_{ii})$

since: $(n-p)S^2 - e_i^2/(1-h_{ii}) = (n-p-1)S^2_{(i)}$

$$f = \frac{RSS_H - RSS}{RSS/(n-p-1)} = \frac{e_i^2/(1-h_{ii})}{(n-p-1)S^2_{(i)}/(n-p-1)} = \left(\frac{e_i}{S_{(i)}\sqrt{1-h_{ii}}}\right)^2 = t^2_{(i)}$$

since $f \sim F(1, n-p-1)$, $t^2_{(i)} \sim F(1, n-p-1)$ #

2. Establish an algebraic relationship between internally and externally studentized residual

Soln: internally: $r_{(i)} = \dfrac{e_i}{s\sqrt{1-h_{ii}}}$ $\Rightarrow \dfrac{r_i}{t_i} = \dfrac{s_{(i)}}{s} = \sqrt{\dfrac{s_{(i)}^2}{s^2}}$

externally: $t_{(i)} = \dfrac{e_i}{s_{(i)}\sqrt{1-h_{ii}}}$

that is: we can first derive the algebraic relationship between $s_{(i)}^2$ and s^2

$$s^2(n-p) = (Y - X\hat{\beta})'(Y - X\hat{\beta})$$

$$s_{(i)}^2(n-1-p) = (Y_{(i)} - X_{(i)}\hat{\beta}_{(i)})'(Y_{(i)} - X_{(i)}\hat{\beta}_{(i)})_{(n-1)}$$

because. $\hat{\beta} - \hat{\beta}_{(i)} = (X'X)^{-1}x_i e_i / (1-h_{ii}) = \dfrac{1}{1-h_{ii}}(X'X)^{-1}x_i(y_i - x_i'\hat{\beta})$

and $s_{(i)}^2(n-1-p) = \left\| \begin{pmatrix} y_1 - x_1'\hat{\beta}_i \\ \vdots \\ y_{i-1} - x_{i-1}'\hat{\beta}_i \\ y_{i+1} - x_{i+1}'\hat{\beta}_i \\ \vdots \\ y_n - x_n'\hat{\beta}_i \end{pmatrix} \right\| = \| (y - X\hat{\beta}_i) - (y_i - x_i'\hat{\beta}_i)e_i \|^2$

$$\Rightarrow s_{(i)}^2(n-1-p) = \|y - X\hat{\beta}_i\|^2 - 2(y_i - x_i'\hat{\beta}_i)e_i'(y - X\hat{\beta}_i) + (y_i - x_i'\hat{\beta}_i)^2$$

$$= \|y - X\hat{\beta} + \frac{1}{1-h_{ii}}X(X'X)^{-1}x_i e_i\|^2 - \left(y_i - \frac{1}{1-h_{ii}}x_i'(X'X)^{-1}x_i e_i\right)^2$$

$$= \|y - X\hat{\beta}\|^2 + \frac{2e_i}{1-h_{ii}}x_i'(X'X)^{-1}X'(y - X\hat{\beta}) + \left(\frac{e_i}{1-h_{ii}}\right)^2 \cdot x_i'(X'X)^{-1}x_i - \left(y_i - \frac{h_{ii}}{1-h_{ii}}e_i\right)^2$$

$$= \|y - X\hat{\beta}\|^2 - e_i^2/(1-h_{ii})$$

$$= s^2(n-p) - e_i^2/(1-h_{ii})$$

$$\frac{s_{(i)}^2}{s^2}(n-1-p) = (n-p) - \frac{e_i^2}{s^2(1-h_{ii})}$$

$$\Rightarrow \left(\frac{r_{(i)}}{t_{(i)}}\right)^2(n-1-p) = (n-p) - r_{(i)}^2$$

3. Show: $r_i^2/(n-p) \sim Beta(1/2, (n-p-1)/2)$ when r_i is the ith internally studentized residual.

Soln: $r_i^2/(n-p) = \frac{e_i^2}{s^2(1-h_{ii})\cdot(n-p)} = \frac{e_i^2}{1-h_{ii}} \cdot \frac{1}{s^2(n-p)} = \frac{e_i^2}{1-h_{ii}} \cdot \frac{1}{(y-x\hat{\beta})'(y-x\hat{\beta})}$

$e_i = (y_i - x_i'\hat{\beta}) = (0,\dots 1,\dots 0)'(y - x\hat{\beta})$

$= c_i' \cdot (I - X(X'X)^{-1}X')y$, let $c_i \triangleq (0,\dots \overset{(i)}{1},\dots 0)'$

$= c_i'(I - X(X'X)^{-1}X')(X\beta + \varepsilon)$

$= c_i'(I-H)\varepsilon$, $\varepsilon \sim N(0, I_n\sigma^2)$

therefore: $\frac{r_i^2}{n-p} = \frac{\varepsilon'(I-H)c_ic_i'(I-H)\varepsilon}{1-h_{ii}} \cdot \frac{1}{\varepsilon'(I-H)\varepsilon}$

$= \frac{\varepsilon' Q \varepsilon}{\varepsilon'(I-H)\varepsilon} \cdot \frac{\sigma^2}{\sigma^2}$, in which $Q = \frac{1}{1-h_{ii}}(I-H)c_ic_i'(I-H)$

$= \frac{z'Qz}{z'(I-H)z}$, in which $z = \frac{\varepsilon}{\sigma} \sim N(0, I_n)$

Q is a projection matrix as:

$Q' = \frac{1}{1-h_{ii}}(I-H)'c_ic_i'(I-H)' = \frac{1}{1-h_{ii}}(I-H)c_ic_i'(I-H) = Q$

$Q^2 = (\frac{1}{1-h_{ii}})^2 (I-H)c_ic_i'(I-H)c_ic_i'(I-H) = \frac{1}{1-h_{ii}}(I-H)c_ic_i'(I-H) = Q \Rightarrow z'Qz \sim \chi^2(1)$

therefore $(I-H-Q)$ also projection matrix as:

$(I-H-Q)' = I - H' - Q' = I - H - Q$

$(I-H-Q)^2 = (I-H)^2 + Q^2 + 2(I-H)\cdot\frac{1}{1-h_{ii}}(I-H)c_ic_i'(I-H)$

$= (I-H) - Q \Rightarrow z'(I-H-Q)z \sim \chi^2(n-p-1)$

also: $z'Qz \perp z'(I-H-Q)z$ as:

$Q(I-H-Q) = Q(I-H) - Q^2$

$= \frac{1}{1-h_{ii}}(I-H)c_ic_i'(I-H) - \frac{1}{1-h_{ii}}(I-H)c_ic_i'(I-H)$

$= 0$

therefore: $\frac{r_i^2}{n-p} = \frac{z'Qz}{z'Qz + z'(I-H-Q)z} \sim Beta(\frac{1}{2}, \frac{1}{2}(n-p-1))$

4. Model $E(y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$. an added variable plot is constructed by first regression y on $x_1 \cdots x_k$, then regressing x_k on $x_1 \cdots x_{k-1}$, then regressing residuals e_1 on e_2

(i) $e_1 \sim \alpha_0 + \alpha_1 e_2$, what is $\hat{\alpha} = \begin{pmatrix} \hat{\alpha}_0 \\ \hat{\alpha}_1 \end{pmatrix}$

(ii) what is $\text{corr}(e_1, e_2)$: relation with the partial correlation of $r_{y, x_k \cdot (x_1 \cdots x_{k-1})}$

(iii) coefficients from the original linear model: $\hat{\beta} = (X'X)^{-1}X'Y$

$y \sim x_1 + \cdots + x_{k-1}$, $e_1 = (I - P_{C(X_{(k)})})y = Q_{C(X_{(k)})}y$

$x_k \sim x_1 + \cdots + x_{k-1}$. $e_2 = (I - P_{C(X_{(k)})})x_k = Q_{C(X_{(k)})}x_k$

$$P_{C(X)} = P_{C(X_{(k)} | x_k)} = P_{C(X_{(k)})} + Q_{C(X_{(k)})}x_k (x_k' Q_{C(X_{(k)})} x_k)^{-1} x_k' Q_{C(X_{(k)})}$$

$e_1 \sim \alpha_0 + \alpha_1 e_2$

then: $\hat{\alpha}_1 = \dfrac{\sum (e_{1i} - \bar{e}_1)(e_{2i} - \bar{e}_2)}{\sum (e_{2i} - \bar{e}_2)^2} = \dfrac{(e_1, e_2)}{\|e_2\|^2}$

$$= \frac{x_k' Q_{C(X_{(k)})} y}{x_k' Q_{C(X_{(k)})} x_k}$$

$= \hat{\beta}_k$ (if we consider $\hat{\beta}_k$ as the coefficient for added variable x_k to the reduced model $y \sim x_1 + \cdots + x_{k-1}$)

$\hat{\alpha}_0 = \bar{e}_1 - \bar{e}_2 \hat{\alpha}_1 = 0$

(ii) $\text{corr}(e_1, e_2) = \dfrac{(e_1, e_2)}{\|e_1\| \|e_2\|}$

because $e_1 = (I - P_{C(X_{(k)})})y = y^{\perp}$: $y - \hat{y}$ projected on space spanned by $(x_1 \cdots x_{k-1})$

$e_2 = (I - P_{C(X_{(k)})})x_k = x_k^{\perp}$: $x_k - \hat{x}_k$ projected on ——

$\Rightarrow r_{x_k, y \cdot x_1 \cdots x_{k-1}} = \dfrac{(y^{\perp}, x_k^{\perp})}{\|y^{\perp}\| \|x_k^{\perp}\|} = \dfrac{(e_1, e_2)}{\|e_1\| \|e_2\|} = \text{corr}(e_1, e_2)$

5. The cook's distance of the ith case is defined as $\frac{(\hat{y}-\hat{y}_{(i)})'(\hat{y}-\hat{y}_{(i)})}{p \cdot s^2}$

express it in terms of p. h_{ii} and r_i

Soln:

$$\hat{\beta}-\hat{\beta}_{(i)} = \frac{1}{1-h_{ii}}(X'X)^{-1}x_i(y_i-x_i'\hat{\beta})$$

$$\Rightarrow \hat{y}-\hat{y}_{(i)} = \frac{1}{1-h_{ii}}X(X'X)^{-1}x_i(y_i-x_i'\hat{\beta})$$

$$(\hat{y}-\hat{y}_{(i)})'(\hat{y}-\hat{y}_{(i)}) = \frac{1}{(1-h_{ii})^2}\cdot(y_i-x_i'\hat{\beta})^2\cdot x_i'(X'X)^{-1}X'X(X'X)^{-1}x_i$$

$$= \frac{1}{(1-h_{ii})^2}\cdot h_{ii}(y_i-x_i'\hat{\beta})^2$$

$$\Rightarrow D_i = \frac{h_{ii}}{p}\cdot\frac{1}{(1-h_{ii})}\cdot\left(\frac{(y_i-x_i'\hat{\beta})^2}{(1-h_{ii})\cdot s^2}\right)$$

$$= \frac{h_{ii}}{p}\cdot\frac{1}{(1-h_{ii})}\cdot(r_i)^2$$

6. Model Diagnosis on Arsentic dataset

```
library(tidyverse)
library(ggplot2)
library(olsrr)
library(haven)
library(car)
```

```
# data preperation
arsenic <- read_dta("arsenic.dta")
arsenic$id <- c(1:21)
```

```
summary(arsenic)
```

```
##       age            sex               drinkuse        cookuse
##  Min.   : 8.00   Length:21          Min.   :1.000   Min.   :2.000
##  1st Qu.:41.00   Class :character   1st Qu.:4.000   1st Qu.:5.000
##  Median :45.00   Mode  :character   Median :5.000   Median :5.000
##  Mean   :47.57                      Mean   :4.333   Mean   :4.857
##  3rd Qu.:53.00                      3rd Qu.:5.000   3rd Qu.:5.000
##  Max.   :86.00                      Max.   :5.000   Max.   :5.000
##     arswater          arsnails            id
##  Min.   :0.00000   Min.   :0.0730   Min.   : 1
##  1st Qu.:0.00000   1st Qu.:0.1180   1st Qu.: 6
##  Median :0.00100   Median :0.1750   Median :11
##  Mean   :0.01624   Mean   :0.3664   Mean   :11
##  3rd Qu.:0.01800   3rd Qu.:0.3580   3rd Qu.:16
##  Max.   :0.13700   Max.   :2.2520   Max.   :21
```

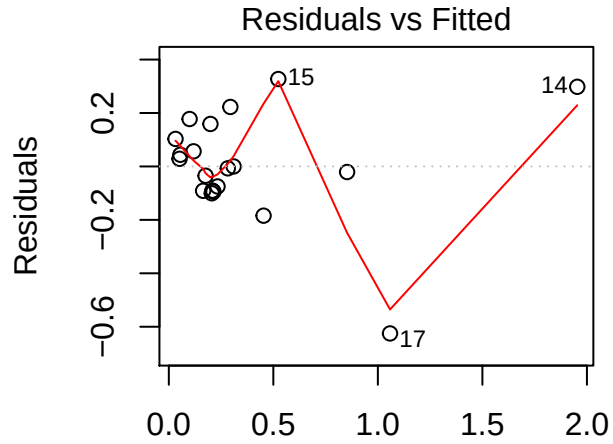
```
mod <- lm(arsnails ~ age + sex + drinkuse + cookuse + arswater + arswater, data = arsenic)
summary(mod)
```

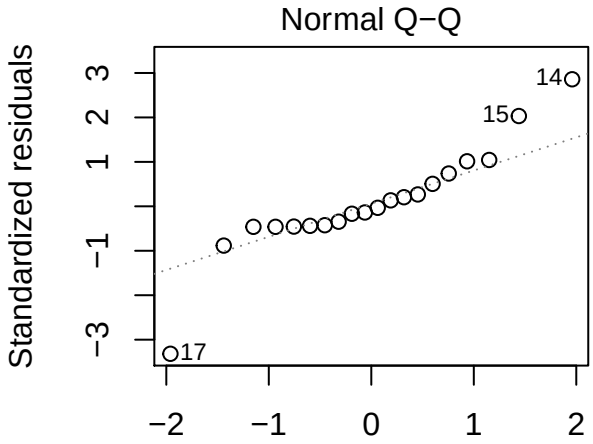
```
##
## Call:
## lm(formula = arsnails ~ age + sex + drinkuse + cookuse + arswater +
##     arswater, data = arsenic)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.62510 -0.09117 -0.00714  0.10297  0.32719
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.452972   0.418132   1.083    0.296
## age         -0.001290   0.003444  -0.374    0.713
## sexMale     -0.145038   0.107448  -1.350    0.197
## drinkuse    -0.011719   0.047010  -0.249    0.807
## cookuse     -0.027471   0.082861  -0.332    0.745
## arswater    13.195586   1.639792   8.047 8.01e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2302 on 15 degrees of freedom
## Multiple R-squared:  0.8323, Adjusted R-squared:  0.7764
## F-statistic: 14.89 on 5 and 15 DF,  p-value: 2.339e-05
```

Diagnositic plots

```
plot(mod)
```

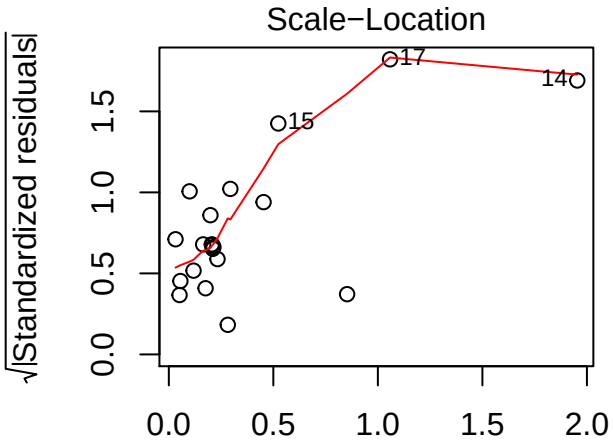
```
## Warning: not plotting observations with leverage one:
##   9
```

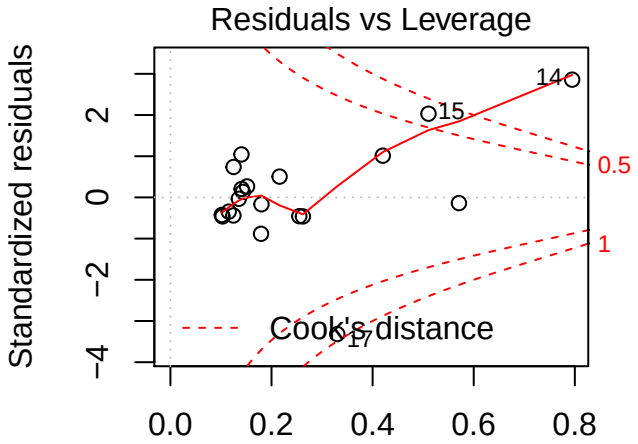




ails ~ age + sex + drinkuse + cookuse + arswate ails ~ age + sex + drinkuse + cookuse + arswate

```
## Warning: not plotting observations with leverage one:
##   9
```





ails ~ age + sex + drinkuse + cookuse + arswate ails ~ age + sex + drinkuse + cookuse + arswate

From the residuals - fitted value plot, we can see a existence of heteroscedasticity, the residual variance tends to be larger when the fitted values are larger. Also, we can see three abnormal cases, 14, 15, and 17, with both larger studentized residuals and cook's distance, indicating them to be outlying and highly influencial. Besides, there the 9th case has leverage 1, being an abnormal case as well.

```
arsenic[c(9,14,15,17),]
```

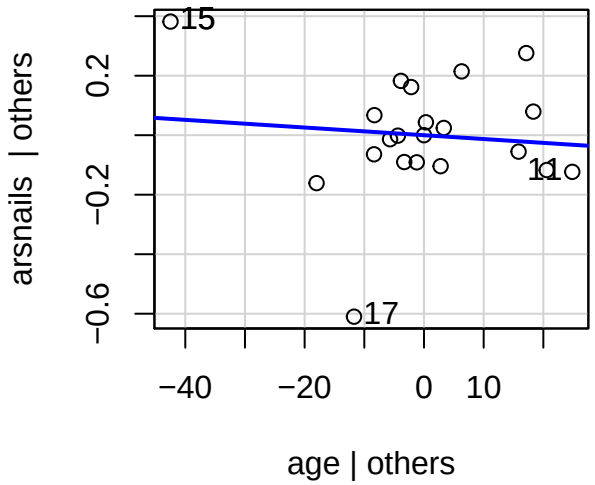
```
## # A tibble: 4 x 7
##     age sex    drinkuse cookuse arswater arsnails    id
##   <dbl> <chr>     <dbl>   <dbl>    <dbl>    <dbl> <int>
```

```
## 1    41 Female        3       2   0         0.310     9
## 2    86 Female        5       5   0.137     2.25     14
## 3     8 Female        5       5   0.0210    0.851    15
## 4    44 Male          5       5   0.0760    0.433    17
```

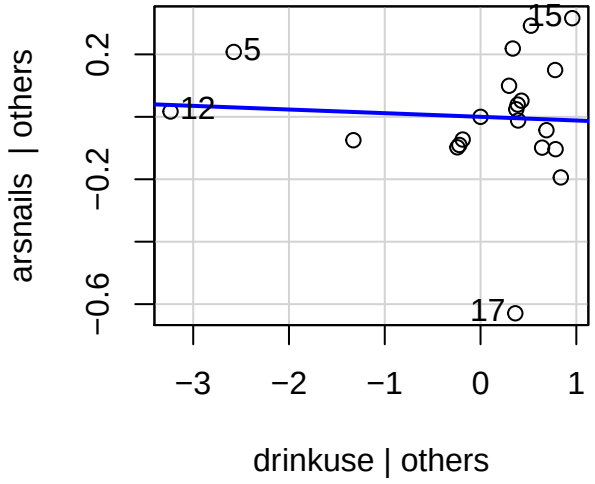
Partial Residual Plots

To analyze whether the predictors are useful, and if there is nonlinearity in each predictor, we then made partial residual plots for covariates. For the sex group, we also made a group wise boxplot to detect outliers.

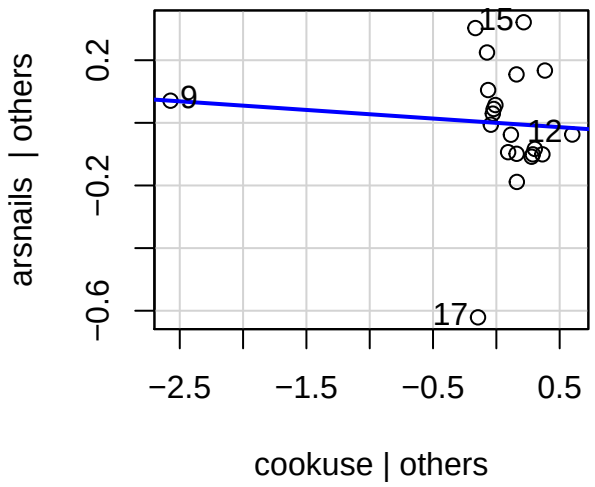
```
avPlots(mod, ~ age)
```



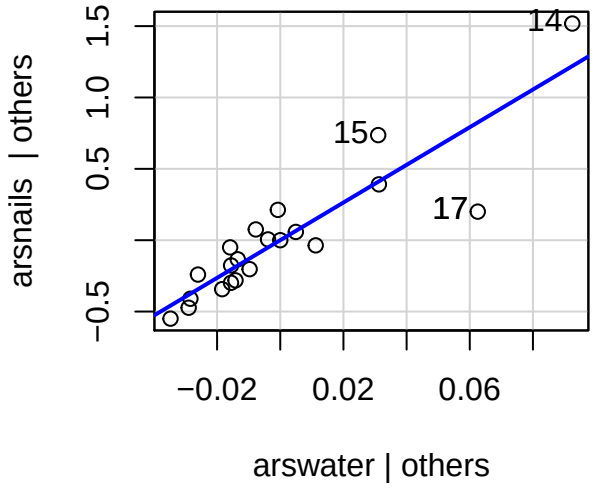

```
avPlots(mod, ~ drinkuse)
```



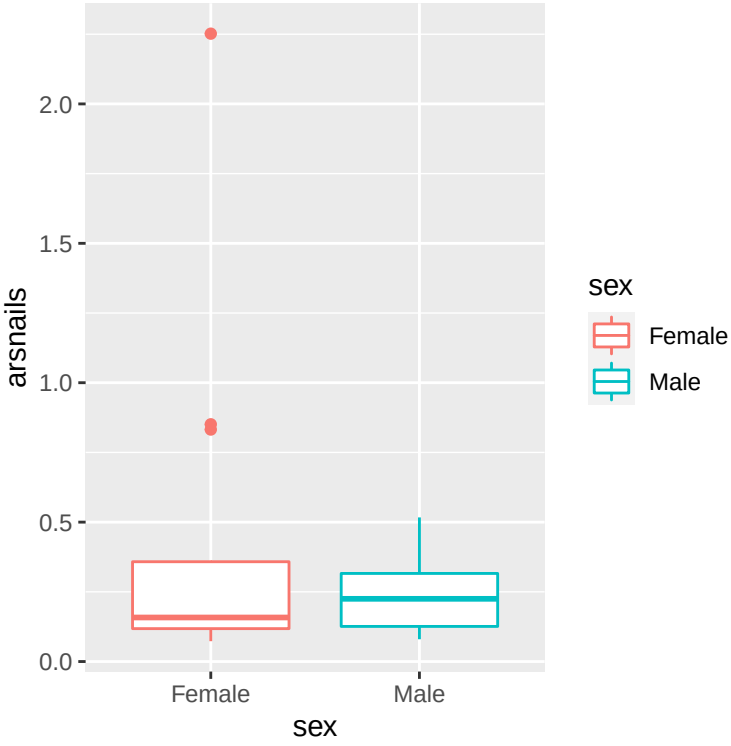

```
avPlots(mod, ~ cookuse)
```




```
avPlots(mod, ~ arswater)
```




```
ggplot(arsenic, aes(sex, arsnails, color = sex)) +
  geom_boxplot()
```



- Case 15 has an abonormally low age, also a high toe nail arsenic concentration, being an outlier with also high influence on model.
- Case 9 has an unusual frequency of cookuse, as it is the only case in the dataset to have 2 for cookuse, and the rest of the cases all have value 5.
- Case 14, 17 have an abnormally high outcome, which is the main reason why they are influencial. However, according to the partial residual plots, the positive linear relationship between toenail arsenic concentration with well water arsenic concentration is not completely driven only by these outliers.

Regression without outlier

```
mod1 <- lm(arsnails ~ age + sex + drinkuse + arswater + arswater, data = arsenic[-c(9,14,15,17),])
summary(mod1)
```

```
##
## Call:
## lm(formula = arsnails ~ age + sex + drinkuse + arswater + arswater,
##     data = arsenic[-c(9, 14, 15, 17), ])
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.084450 -0.040869 -0.015782  0.009797  0.239076
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.2681601  0.1493702   1.795   0.0978 .
## age         -0.0002232  0.0021761  -0.103   0.9200
## sexMale      0.0183767  0.0485070   0.379   0.7114
## drinkuse    -0.0278381  0.0224720  -1.239   0.2391
## arswater    12.8679620  2.2049209   5.836 8.01e-05 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09439 on 12 degrees of freedom
## Multiple R-squared:  0.8245, Adjusted R-squared:  0.766
## F-statistic:  14.1 on 4 and 12 DF,  p-value: 0.0001737
```

When we excluded all the high influencial cases, the regression result stays the same, indicating a robust positive linear relationship between toenail arsenic concentration and the arsenic concentration in their well-water. And the other factors, age, sex, frequency of drink and cook use remain insignificant.

Summary

In the analysis above, we studied the ingestion of arsenic-containing water of 21 people and what factors have impact on the ingestion. We built up multiple linear regression model with toenail arsenic concentration as outcome, and had age, sex, frequncy of private well water for drinking and cooking, and the arsenic concentration in their well water as predictors. The model revealed a significant positive linear relationship between the arsenic concentration in well water and toenail arsenic concentration.

Within the 21 participants, 13 of them are female and 8 are male. Most of the participants are aged from 40 - 60, while case 9 is very young of age 8, and case 11 is very old of age 86. Most participants except for 3 reported a high drinkuse of 4 or 5, and all participants except for 1 reported a very high cookuse of 5. Arsenic concentration were detected in 15 of the 21 well-water samples from the participants, but we still include the 6 cases with no arsenic concentration detected, imputing a value of 0. 3 participants had abnormally high townail arsenic concentration, and all of them are female.

The model results and direction of linear relationship remained valid when we excluded all influencial outlying cases. Therefore, despite the extreme data points and the small sample size, our conclusion that toenail arsenic concentration has a significant positvie linear relationship with the arsenic concentration of participants' well water arsenic concentration is very stable.