Chapter 8

8.1

For the following data, use R to verify that the least squares regresion line is $\hat{Y} = 1.8X - 8.5$

X: 5, 8, 9, 7, 14 Y: 3, 1, 6, 7, 19

```
dat <- data.table(
   X = c(5, 8, 9, 7, 14),
   Y = c(3, 1, 6, 7, 19))

lm(Y ~ X, data = dat)
```

```
Call:
lm(formula = Y ~ X, data = dat)

Coefficients:
(Intercept)            X
     -8.478        1.823
```

Also verify that the Theil-Sen estimator, the slope is estimated to be 1.746 and the intercept is estimated to be -7.968.

```
tsreg(dat$X, dat$Y)$coef
```

```
Intercept
-5.730159  1.746032
```

8.2

Using the R function *lsfit*, compute the residuals using the data in E1,

Verify that if you square and sum the residusls, you get 46.585.

```
res <- lsfit(dat$X, dat$Y)$resid

sum(res^2)
```

```
[1] 46.58407
```

8.3

Verify that for the data in E1, if you use $\hat{Y} = 2X - 9$, the sum of the squared residuals is greater than 46.584.

```
Yhat <- 2*dat$X - 9
res <- sum( (dat$Y - Yhat)^2 )
```

```
stopifnot(res > 46.583)
res
```

```
[1] 53
```

Why would you expect a value greater than 46.584?

The x coefficent increased.

8.4

Suppose that based on $n = 25$ values, $s_x^2 = 12$ and $\sum(X_i - \bar{X})(Y_i - \bar{Y}) = 144$.

What is the slope of the least squares regression?

$A = 144, C = (n-1)s_x^2 = 288, b_1 = A/C = 144/288 = .5$

8.5

The following table reports breast cancer rates plus levels of solar radiation (in calories per day) for various cities in the United States. The data are stored in the file cancer_rate_dat.txt.

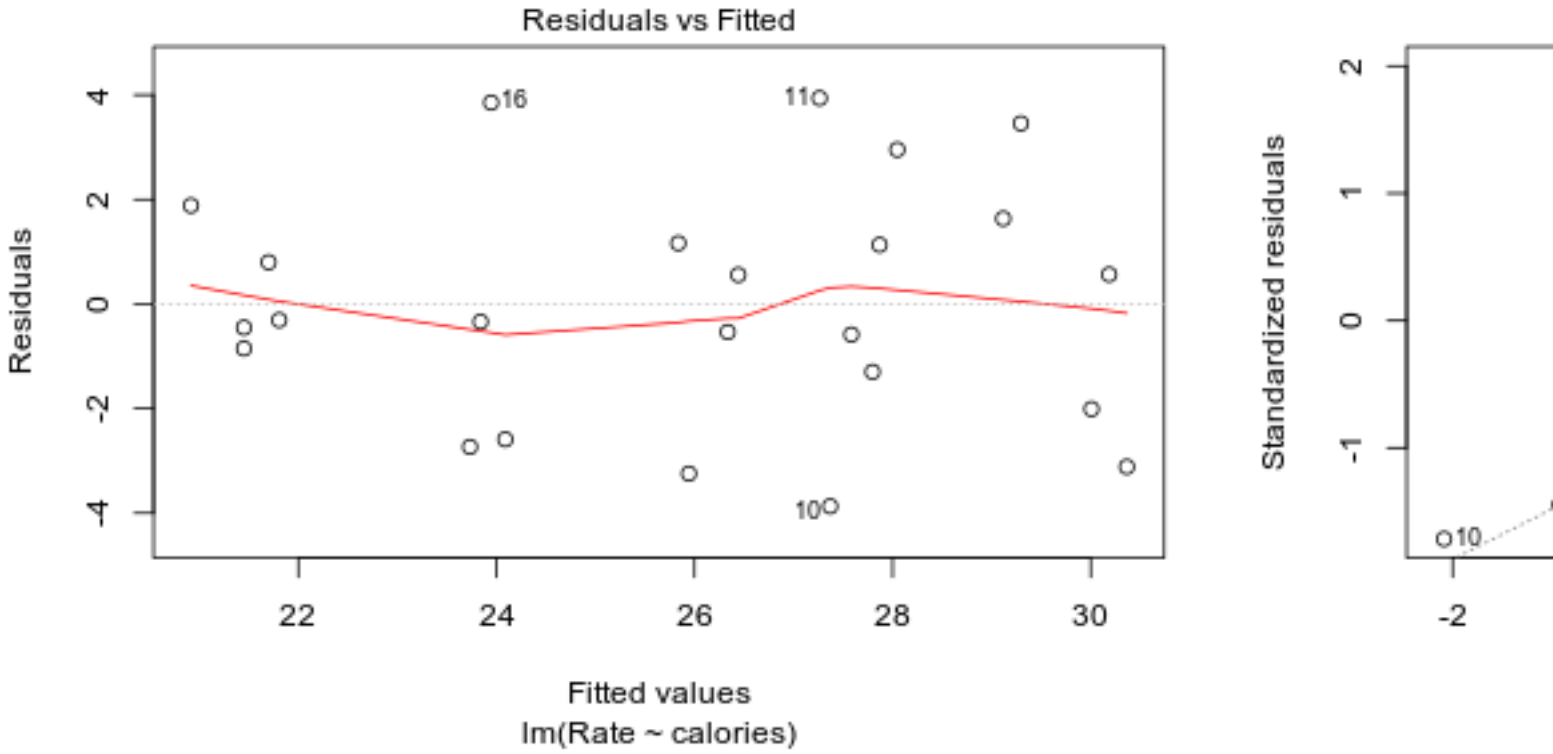
```
dat <- data.table::fread(paste0(data.dir, "cancer_rate_dat.txt"), fill = T, sep = "&")

dat
```

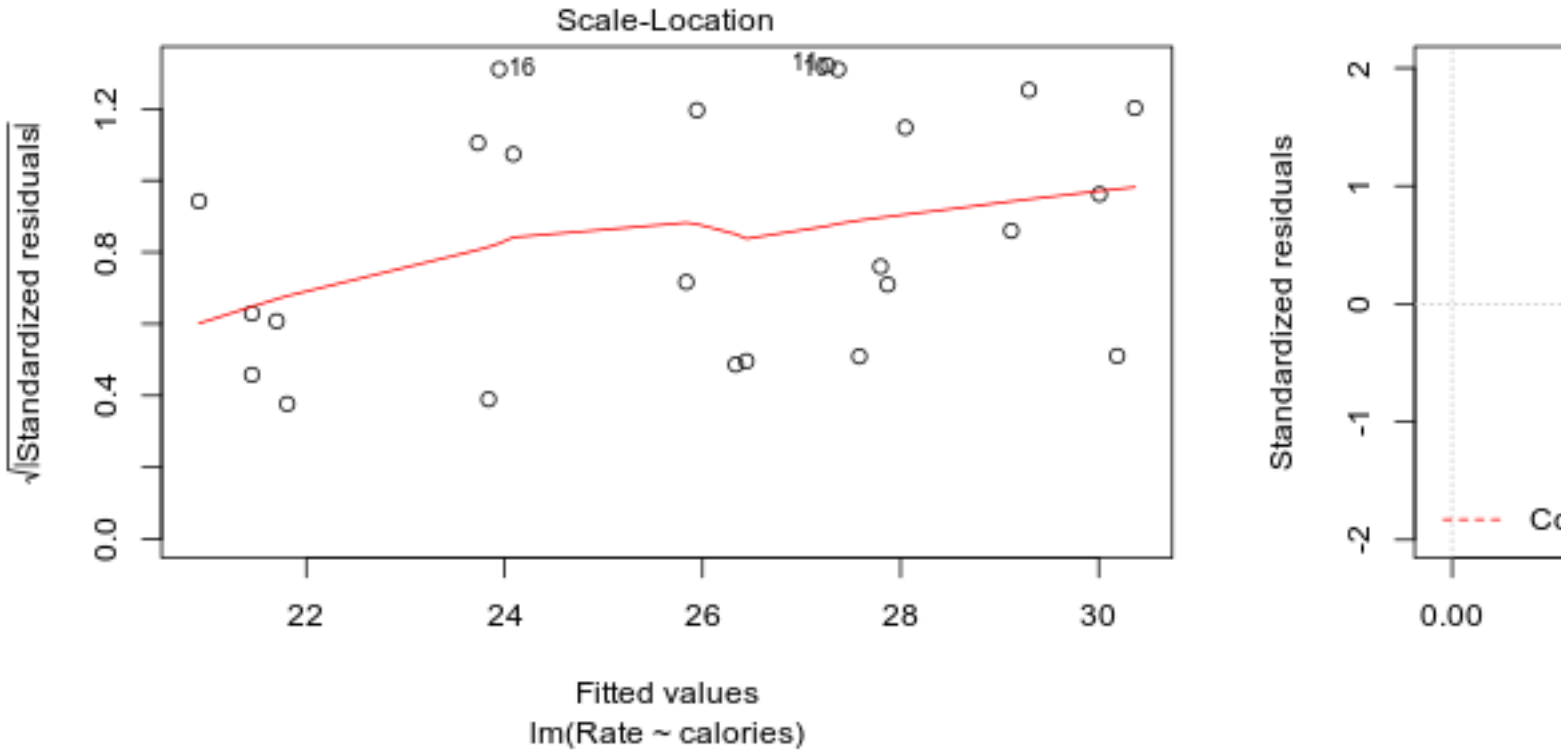
```
             City  Rate calories
 1:       New York 32.75      300
 2:        Chicago 30.75      275
 3:     Pittsburgh 28.00      280
 4:        Seattle 27.25      270
 5:         Boston 30.75      305
 6:      Cleveland 31.00      335
 7:       Columbus 29.00      340
 8:   Indianapolis 26.50      342
 9:    New Orleans 27.00      348
10:      Nashville 23.50      354
11: Washington, DC 31.20      357
12: Salt Lake City 22.70      394
13:          Omaha 27.00      380
14:      San Diego 25.80      383
15:        Atlanta 27.00      397
16:    Los Angeles 27.80      450
17:          Miami 23.50      453
18:     Fort Worth 21.50      446
19:          Tampa 21.00      456
```

```
20:    Albuquerque 22.50      513
21:      Las Vegas 21.50      510
22:       Honolulu 20.60      520
23:        El Paso 22.80      535
24:        Phoenix 21.00      520
              City  Rate calories
```

Fit a OLS regression to predict cancer rates and comment on what this suggests.

```
fit <- lm(Rate ~ calories, data = dat)

plot(fit)
```





8.6

For the following data, use R to compute the least squares regresion line for predicting GPA given SAT.

SAT: 500, 530, 590, 660, 610, 700, 570, 640 GPA: 2.3, 3.1, 2.6, 3.0, 2.4, 3.3, 2.6, 3.5

```
dat <- data.table(
   SAT = c(500, 530, 590, 660, 610, 700, 570, 640),
   GPA = c(2.3, 3.1, 2.6, 3.0, 2.4, 3.3, 2.6, 3.5)
)

fit <- lm(GPA ~ SAT, data = dat)
coef(fit)
```

```
(Intercept)         SAT
0.484615385 0.003942308
```

8.7

Compute the residuals for the data used in the previous problem and verify that sum to zero.

```
round(sum((dat$GPA - fit$fitted.values)), 5)
```

```
[1] 0
```

8.8

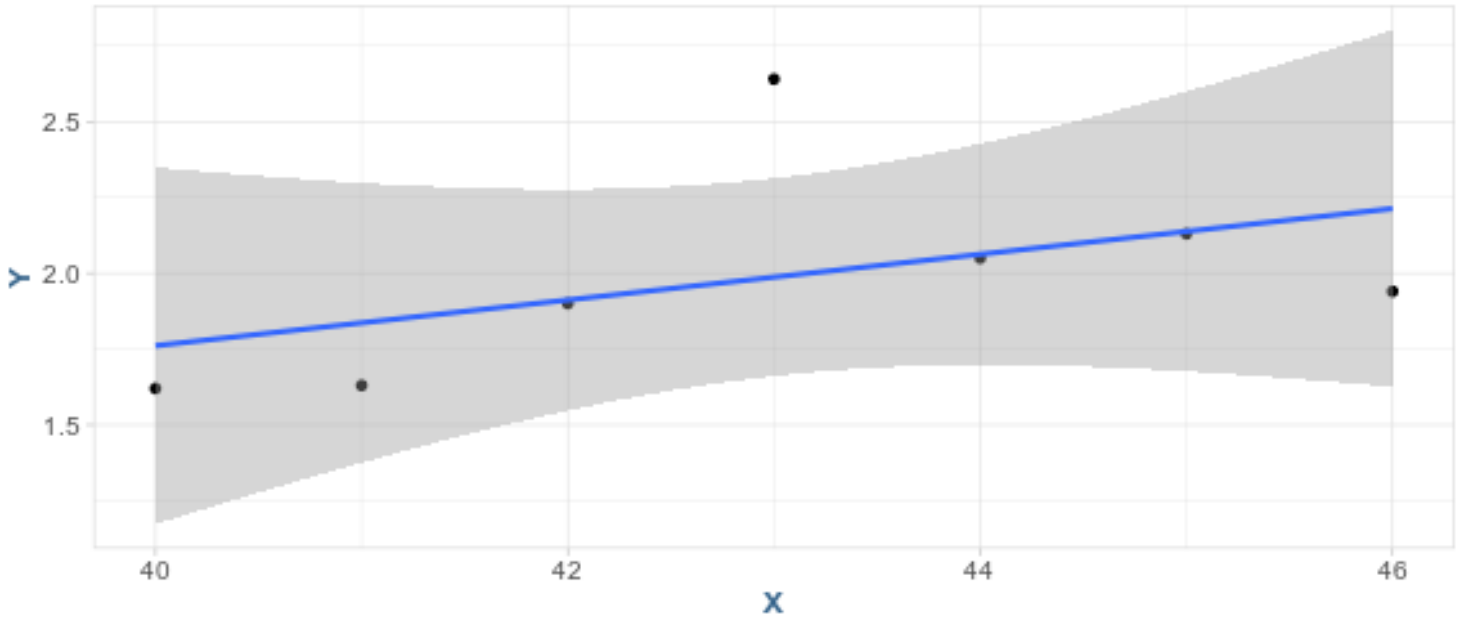
For the following data, use R to compute the least squares regression line for predicting Y from X.

X: 40, 41, 42, 43, 44, 45, 46 Y: 1.62, 1.63, 1.90, 2.64, 2.05, 2.13, 1.94

```
dat <- data.table(
   X = c(40, 41, 42, 43, 44, 45, 46),
   Y = c(1.62, 1.63, 1.90, 2.64, 2.05, 2.13, 1.94)
)

summary(fit <- lm(Y ~ X, data = dat))
```

```
Call:
lm(formula = Y ~ X, data = dat)

Residuals:
        1         2         3         4         5         6         7
-0.141071 -0.206429 -0.011786  0.652857 -0.012500 -0.007857 -0.273214

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.25321    2.73157  -0.459    0.666
X            0.07536    0.06346   1.188    0.288

Residual standard error: 0.3358 on 5 degrees of freedom
Multiple R-squared:   0.22, Adjusted R-squared:  0.064
F-statistic:  1.41 on 1 and 5 DF,  p-value: 0.2883
```

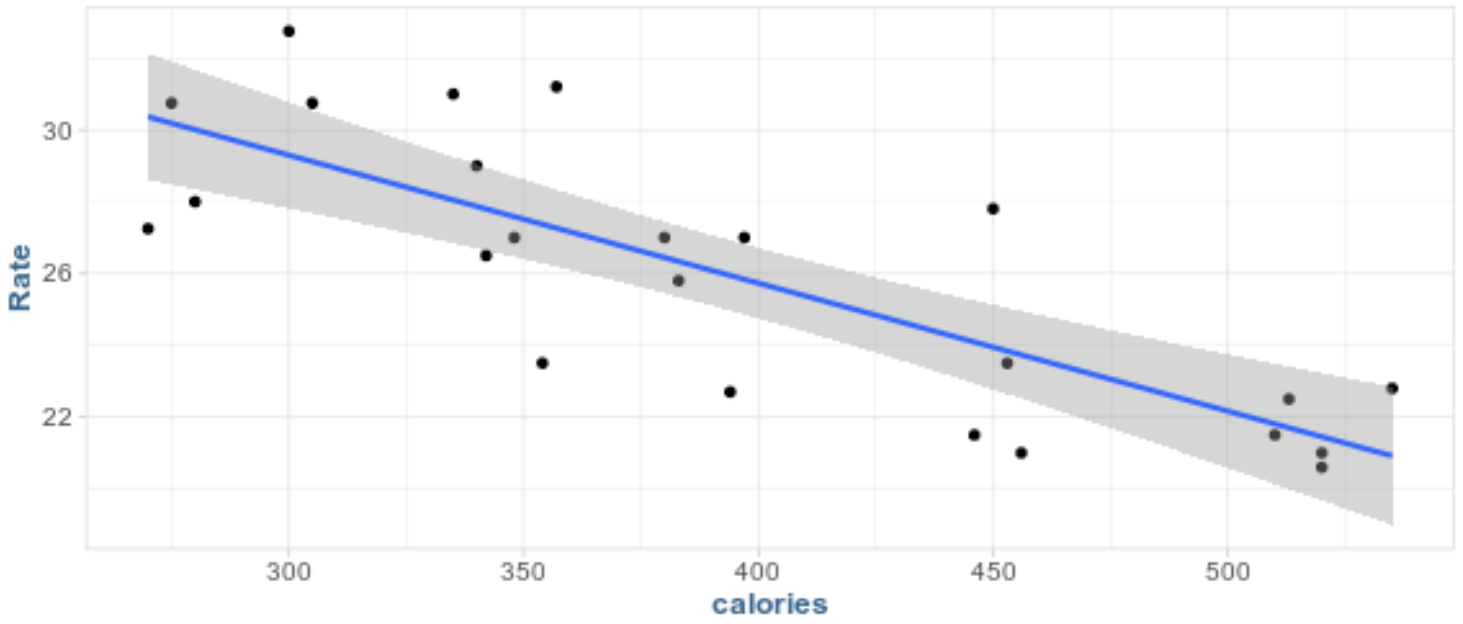
```
ggplot(dat, aes(X, Y)) +
   geom_point() +
   geom_smooth(method = "lm")
```



8.9

In exercise 5, what would be the least squares estimate of the cancer rate given a solar radiation of 600?

```
dat <- data.table::fread(paste0(data.dir, "cancer_rate_dat.txt"), fill = T, sep = "&")

ggplot(dat, aes(calories, Rate)) +
   geom_point() +
   geom_smooth(method = "lm")
```



```
fit <- lm(Rate ~ calories, data = dat)

coef(fit)
```

```
(Intercept)    calories
39.99094634 -0.03565283
```

```
39.99 -0.037*600
```

```
[1] 17.79
```

Why might this be unreasonable?

Because 600 is outside of the bounds of seen values (extrapolation).

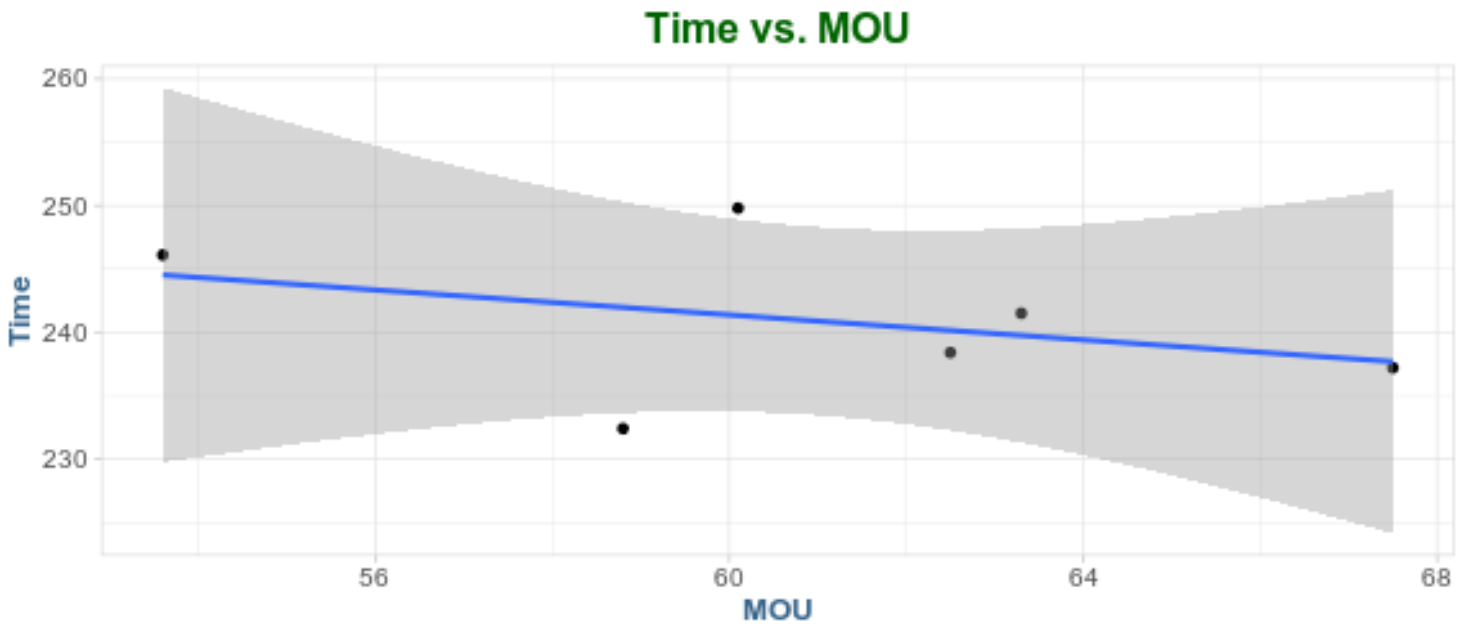
8.10

Maximal oxygen uptake (MOU) is a measure of an individual's physical fitness. You want to know how MOU is related to how fast someone can run a mile. Suppose you randomly sample six athletes and get:

MOU (kl/kg): 63.3, 60.1, 53.6, 58.8, 67.5, 62.5 Time (seconds): 241.5, 249.8, 246.1, 232.4, 237.2, 238.4

Compute the least squres regression line and comment on what the results suggest.

```
dat <- data.table(
   MOU = c(63.3, 60.1, 53.6, 58.8, 67.5, 62.5),
   Time = c(241.5, 249.8, 246.1, 232.4, 237.2, 238.4)
)

fit <- lm(Time ~ MOU, data = dat)

ggplot(dat, aes(MOU, Time)) +
   geom_point() +
   geom_smooth(method = "lm") +
   labs(title = "Time vs. MOU")
```



Generally, time decreases as MOU increases.

8.11

Verify that for the following pairs of points, the least squares regression line has a slope of zero. Plot the points and comment on the assumption that the regression line is straight.

X: 1, 2, 3, 4, 5, 6 Y: 1, 4, 7, 7, 4, 1

```
dat <- data.table(
   X = c(1, 2, 3, 4, 5, 6),
   Y = c(1, 4, 7, 7, 4, 1)
)

fit <- lm(Y ~ X, data = dat)

coef(fit)
```

```
  (Intercept)             X
 4.000000e+00 -5.838669e-16
```

8.12

Repeat the last exercise, only for points:

X: 1, 2, 3, 4, 5, 6 Y: 4, 5, 6, 7, 8, 2

```
dat <- data.table(
   X = c(1, 2, 3, 4, 5, 6),
   Y = c(4, 5, 6, 7, 8, 2)
)

fit <- lm(Y ~ X, data = dat)

coef(fit)
```

```
  (Intercept)             X
 5.333333e+00 -6.369458e-16
```

8.13

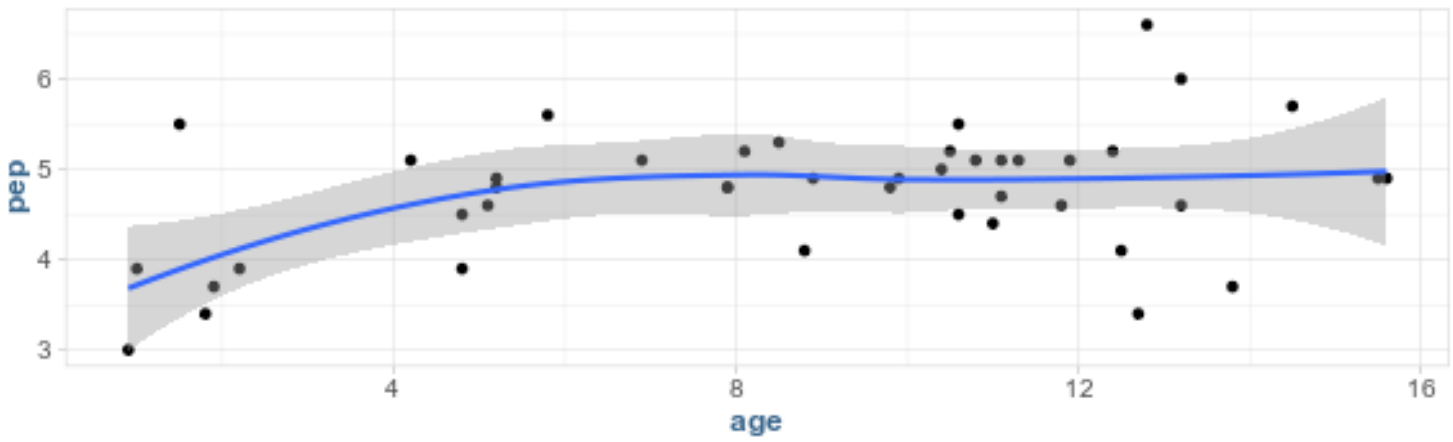
Vitamin A is required for good health. However, one bite of polar bear liver results in death becasue it contains a high concentration of vitamin A. Comment on this fact in terms of extrapolation.

Vitaim A as a function of health is a bounded variable. Predicting a response when the independent variables are far from what has been observed can lead to undefined results (death).

8.14

Socket et al. (1987) report data related to patterns of residual insulin secretion in children. A portion of the study was concerned with whether age can be used to predict the logarithm of C-peptide concentrations at diagnosis. The observed values are (data file), Replicate the *LOESS* smoothed curve in fig 8.4.1.

```
dat <- data.table::fread(paste0(data.dir, "diabetes_sockett_dat.txt"), fill = T)

ggplot(dat, aes(age, pep)) +
   geom_point() +
   geom_smooth(method = "loess")
```



8.15

For the data in the last exercise, use R to verify that a least squares regression line using only X values (age) less than or equal to 7 yields a p-value equal to 0.026 when using the R function *olsch4*. Also verify that the p-value, when using the Theil-Sen estimator, is 0.0233.

```
dat2 <- dat[age <= 7]

WRS::olshc4(dat2$age, dat2$pep)
```

```
$n
[1] 14

$n.keep
[1] 14

$ci
            Coef. Estimates   ci.lower  ci.upper      p-value  Std.Error
(Intercept)     0 3.5148814 2.45961431 4.5701484 1.004728e-05 0.48433121
Slope           1 0.2474008 0.03559377 0.4592078 2.570292e-02 0.09721213

$cov
            [,1]         [,2]
[1,]  0.2345767 -0.045083903
[2,] -0.0450839  0.009450198
```

```
WRS::regci(dat2$age, dat2$pep)
```

```
[1] "Duplicate values detected; tshdreg might have more power than tsreg"
[1] "Taking bootstrap samples. Please wait."
$regci
              ci.low     ci.up  Estimate      S.E.    p-value
Intercept 2.55336867 4.7290698 3.0666667 0.6468266 0.01669449
Slope 1   0.02325581 0.4871795 0.3333333 0.1278130 0.03338898

$n
[1] 14

$n.keep
[1] 14
```

8.16

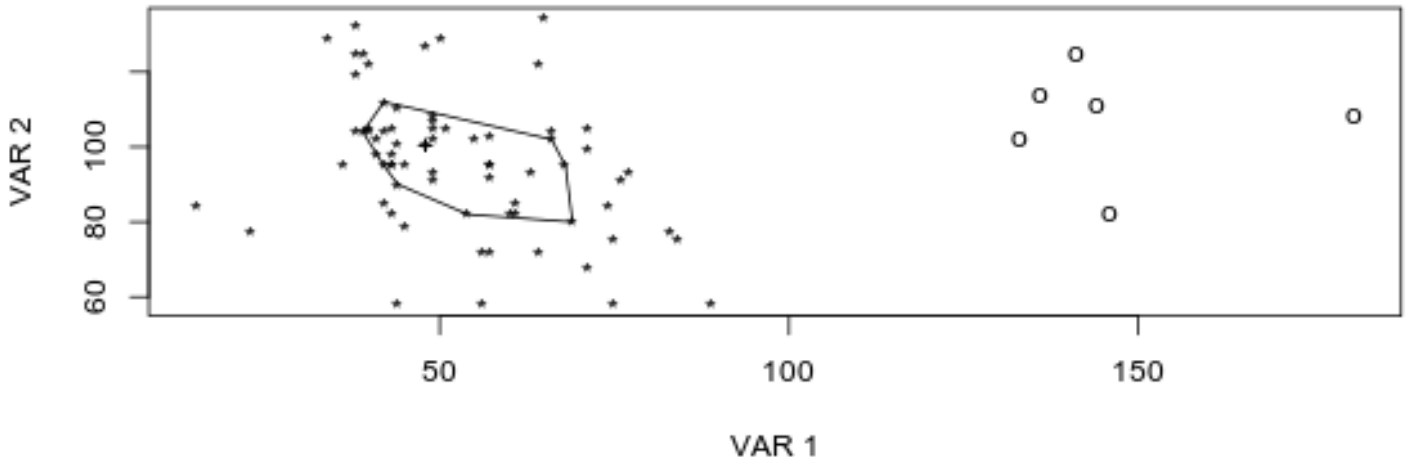
For the reading data in Table 8.5, verify that the R function spearci returns a p-value equal to 0.014 and that scorci returns a p-value equal to 0.002. Based on the plot returned by scorci, why is it not suprising that these two function give similar results?

```
dat <- data.table::fread(paste0(data.dir, "read_dat.txt"), fill = T, skip = 13)

spearci( dat[,4], dat[,8])
```

```
$cor.ci
[1] -0.48330768 -0.01914648

$p.value
[1] 0.038

$cor.est
[1] -0.2687277
```

```
scorciMC( dat[,4], dat[,8])
```

```
Attaching package: 'MASS'

The following object is masked from 'package:WRS':

    ltsreg

The following object is masked from 'package:formattable':

    area

The following object is masked from 'package:dplyr':

    select
```



```
$cor.ci
[1] -0.6900937 -0.1184465

$p.value
[1] 0.01

$cor.est
[1] -0.3868361
```

8.17

Given that $b_1 = -1.5, n = 10, s^2_{y,x} = 35, \sum(X_i - \bar{X})^2 = 140$, assume normality and homosecedasiticy and compute a 0.95 confidence interval for slope, β_1.

```
b1 <- -1.5; n <- 10; se <- sqrt(35 / 140)
alpha <- .05

b1 + qt(c(Lower = alpha/2, Upper = 1 - alpha/2), df = n - 2) * se
```

```
     Lower      Upper
-2.6530021 -0.3469979
```

8.18

Repeat the previous problem, only compute a 0.98 confidence interval.

```
b1 <- -1.5; n <- 10; se <- sqrt(35 / 140)
alpha <- 1 - .98

b1 + qt(c(Lower = alpha/2, Upper = 1 - alpha/2), df = n - 2) * se
```

```
      Lower       Upper
-2.94822972 -0.05177028
```

8.19

Based on results convered in the previous chapters, speculate about why the confidence intervals computed in the last two problems might be inaccurate.

Least squares regression can be negatively impacted by non-normality, heteroscedasticity and outliers.

8.20

Assume normality and homoscedasticity and suppose $n = 30, \sum(X_i) = 15, \sum(Y_i) = 30, \sum(X_i - \bar{X})(Y_i - \bar{Y}) = 30, \sum(X_i - \bar{X})^2 = 10$

```
n <- 30; A <- 30; C <- 10;
xbar <- 15 / 30; ybar <- 30 / 30

slope <- A / C
intercept <- ybar - xbar*slope
```

Determine the least squares estimates of the slope and intercept.

$\beta_1 = 3, \beta_0 = -0.5$

8.21

Assume normality and homoscedasticity and suppose $n = 38, \bar{Y} = 20, \sum X_i^2 = 1,922, \bar{X} = 7, \sum(X_i - \bar{X})(Y_i - \bar{Y}) = 180, \sum(X_i - \bar{X})^2 = 60, s^2_{X.Y} = 121.$

a.) Determine the least squares estimate of the slope and intercept.

```
n <- 38; ybar <- 20; xbar <- 7
variance <- 121; A <- 180; C <- 60
ssx <- 1922

slope <- A / C
intercept <- ybar - xbar*slope
```

$\beta_1 = 3, \beta_0 = -1$

b.) Test the hypothesis $H_0 : \beta_0 = 0, \alpha = 0.02$

```
alpha <- 0.02

slope + qt(c(Lower = alpha/2, Upper = 1 - alpha/2), df = n - 2) * sqrt(variance / n)
```

```
    Lower     Upper
-1.344198  7.344198
```

```
Tval <- intercept * sqrt( ((n - 2) * C) / (variance  * ssx) )

crit <- qt(alpha/2, df = n - 2)

ifelse(abs(Tval) < crit, "Reject Null", "Fail to Reject")
```

```
[1] "Fail to Reject"
```

c.) Compute a 0.9 confidence interval for β_1

```
alpha <- 1 - .9

slope + qt(c(Lower = alpha/2, Upper = 1 - alpha/2), df = n - 2) * sqrt( variance / C)
```

```
    Lower     Upper
0.6024587 5.3975413
```

8.22

Assume normality and homoscedasticity and suppose $n = 41, \bar{Y} = 10, \bar{X} = 12, \sum(X_i - \bar{X})(Y_i - \bar{Y}) = 100, \sum(X_i - \bar{X})^2 = 400, s^2_{Y.X} = 144.$

a.) Determine the least squares regression line

```
n <- 41; ybar <- 10; xbar <- 12;
ssr <- 100; ssx <- 400; variance <- 144

b1 <- ssr / ssx

b0 <- ybar - xbar*b1
```

$\beta_0 = 7, \beta_1 = 0.25$

b.) Compute a .9 confidence interval for β_1

```
alpha <- 1 - .9

b1 + qt(c( Lower = alpha/2, Upper = 1 - alpha/2), df = n - 2) * sqrt( variance / ssx)
```

```
     Lower      Upper
-0.7609251  1.2609251
```

8.23

Assume normality and homoscedasticity and suppose $n = 18, \beta_1 = 3.1, \sum (X_i - \bar{X})^2 = 144, s^2_{X.Y} = 36.$

Compute a 0.95 confidence interval for β_1.

```
n <- 18; b1 <- 3.1; ssx <- 144; variance <- 36
alpha <- 1 - .95

b1 + qt(c(Lower = alpha/2, Upper = 1 - alpha/2), df = n - 2) * sqrt( variance / ssx )
```

```
   Lower    Upper
2.040047 4.159953
```

Would you conclude β_1 > 2?

Given the confidence interval above, it seems reasonable.

8.24

Assume normality and homoscedasticity and suppose $n = 20, \beta_0 = 6, \sum X_i^2 = 169, S^2_{Y.X} = 25, \sum (X_i - \bar{X})^2 = 90.$

Compute a .95 confidence interval for β_0.

```
n <- 20; b0 <- 6; ssx <- 169
variance <- 25; xvar <- 90
alpha <- 1 - .95

b0 + qt(c(Lower = alpha/2, Upper = 1 - alpha/2), df = n - 2) * sqrt( (variance*ssx) / (n * xvar
```

```
   Lower    Upper
2.781252 9.218748
```

8.25

Given the following quantities, find the sample correlation coefficent, r, and test $H_0 : \rho = 0$ at the indicated value for α.

a.) $n = 27, \sum (Y_i - \bar{Y})^2 = 100, \sum (X_i - \bar{X})^2 = 625, \sum (X_i - \bar{X})(Y_i - \bar{Y})^2 = 200, \alpha = 0.01$

```
n <- 27; yvar <- 100; xvar <- 625
ssr <- 200; alpha <- 0.01

r <- ssr / sqrt( yvar * xvar )

test.stat.T <- r * sqrt( (n-2)/ (1 - r^2))

crit <- qt(1 - alpha/2, df = n - 2)

ifelse(abs(test.stat.T) >= crit, "Reject Null", "Fail to Reject")
```

```
[1] "Reject Null"
```

b.) $n = 5, \sum (Y_i - \bar{Y})^2 = 16, \sum (X_i - \bar{X})^2 = 25, \sum (X_i - \bar{X})(Y_i - \bar{Y})^2 = 10, \alpha = 0.05$

```
n <- 5; yvar <- 16; xvar <- 25
ssr <- 10; alpha <- 0.05

r <- ssr / sqrt( yvar * xvar )

test.stat.T <- r * sqrt( (n-2)/ (1 - r^2))

crit <- qt(1 - alpha/2, df = n - 2)
```

```
ifelse(abs(test.stat.T) >= crit, "Reject Null", "Fail to Reject")
```

```
[1] "Fail to Reject"
```

8.26

The high school grade-point average (X) and college grade-point (Y) for 29 randomly sampled college freshmen yielded the following results:

$\sum (Y_i - \bar{Y})^2 = 64, \sum (X_i - \bar{X})^2 = 100, \sum (X_i - \bar{X})(Y_i - \bar{Y}) = 40$

Test $H_0 : \rho = 0, \alpha = 0.1$ and interprest the results.

```
n <- 29; yvar <- 64; xvar <- 100; ssr <- 40
alpha <- 0.1

r <- ssr / sqrt( yvar * xvar )

crit <- qt(1 - alpha/2, df = n - 2)

test.stat.T <- r * sqrt( (n - 2) / (1 - r^2) )

ifelse( abs(test.stat.T) >= crit, "Reject null", "Fail to reject")
```

```
[1] "Reject null"
```

8.27

For the previous exercise, answer the following questions:

a.) Is it reasonable to conclude that the least squares regression line has a positive slope?

Yes, it appears there is a positive correlation.

b.) Is it possible that despite the value for r, as high school grade-point average increases, college grade-point average decreases? Explain.

It is possible, although unlikely.

c.) What might you do, beyond considering r, to decide it is reasonable to conclude that as high school grade-point averages increase, college grade point averages increase as well?

Plot the data to assess the visual indication of the relationship.

8.28

Using R, determine what happens to Pearson's correlation between X and Y if the Y values are multiplied by 3. Argue that if Y is multiplied by any constant c != 0, Pearson's correlation does not change.

```
set.seed(46)
x <- rnorm(30); y <- rnorm(30)
cor(x, y)
```

```
[1] 0.1888607
```

```
cor(3*x, y)
```

```
[1] 0.1888607
```

8.29

Repeat the previous problem, only determine what happens to the slope of the least squares regresion line.

```
set.seed(46)
x <- rnorm(30); y <- rnorm(30)

lsfit(x, y)$coef
```

```
    Intercept             X
-0.0005236425  0.2094865640
```

```
lsfit(x, 3*y)$coef
```

```
   Intercept            X
-0.001570928  0.628459692
```

The absolute value of the slope gets larger.

8.30

Consider a least squares regression line $Y = 0.52X + 2$, assume homoscedasiticy as consider the situation where the common variance is $\sigma^2 = 1$? What happens to the correlation coefficent between X and Y if instead $\sigma^2 = 2$?

It will increase the variance so the correlation will decrease.

8.31

The numerator of the coefficient of determination is $\sum (Y_i - \bar{Y}) - \sum (Y_i - \hat{Y})$. Based on the least squares principal, why is the value always greater than zero?

If $\bar{Y}$ is used to predict Y, $\sum (Y_i - \bar{Y})$ will be larger than $\sum (Y_i - \hat{Y})$.

8.32

Imagine a study where the correlation between some amount of an experimental drug and liver damage yields a value for r close to zero and the hypothesis $H_0 : \rho = 0$ is not rejected. Why might it be unreasonable to conclude that the two variabels are independent?

The magnitude of the residuals, curvature, outliers.

8.33

Suppose $r^2 = 0.95$. Explain why this does not provide convincing evidence that the least squares line provides a good fit to a scatterplot of the points.

Outliers can cause large r^2, however, be a poor fit.

8.34

Imagine a situation where points are removed for which the X values are judged to be outliers. Note that this restricts the range of X values. Without looking at the data, can someone predict whether Pearson's correlation will increase or decrease after these points are removed?

No. You need to look at the data. Restricting the range of X can increase as well as decrease r.

8.35

If the normality assumption is violated, what effect might this have when computing confidence intervals for the slope and intercept as described in 8.4.1?

The con

dence interval can be relatively long and is potentially inaccurate.

8.36

If the homoscedasticity assumption is violated, what effect might this have when computing confidence intervals as described in 8.4.1?

The con

dence interval can be relatively inaccurate due to using the wrong standard error.