HW10

Code

Problem 10A

The dataset chicks was obtained from BLSS: The Berkeley Interactive Statistical System by Abrahams and Rizzardi. Each observation corresponds to an egg (and the resulting chick) of a bird called the Snowy Plover. The data were taken at Point Reyes Bird Observatory. Column 1 contains the egg length in millimeters, Column 2 the egg breadth in millimeters, Column 3 the egg weight in grams, and Column 4 the chick weight in grams. The object is to estimate the size of the chick based on dimensions of the egg.

a)

Hide

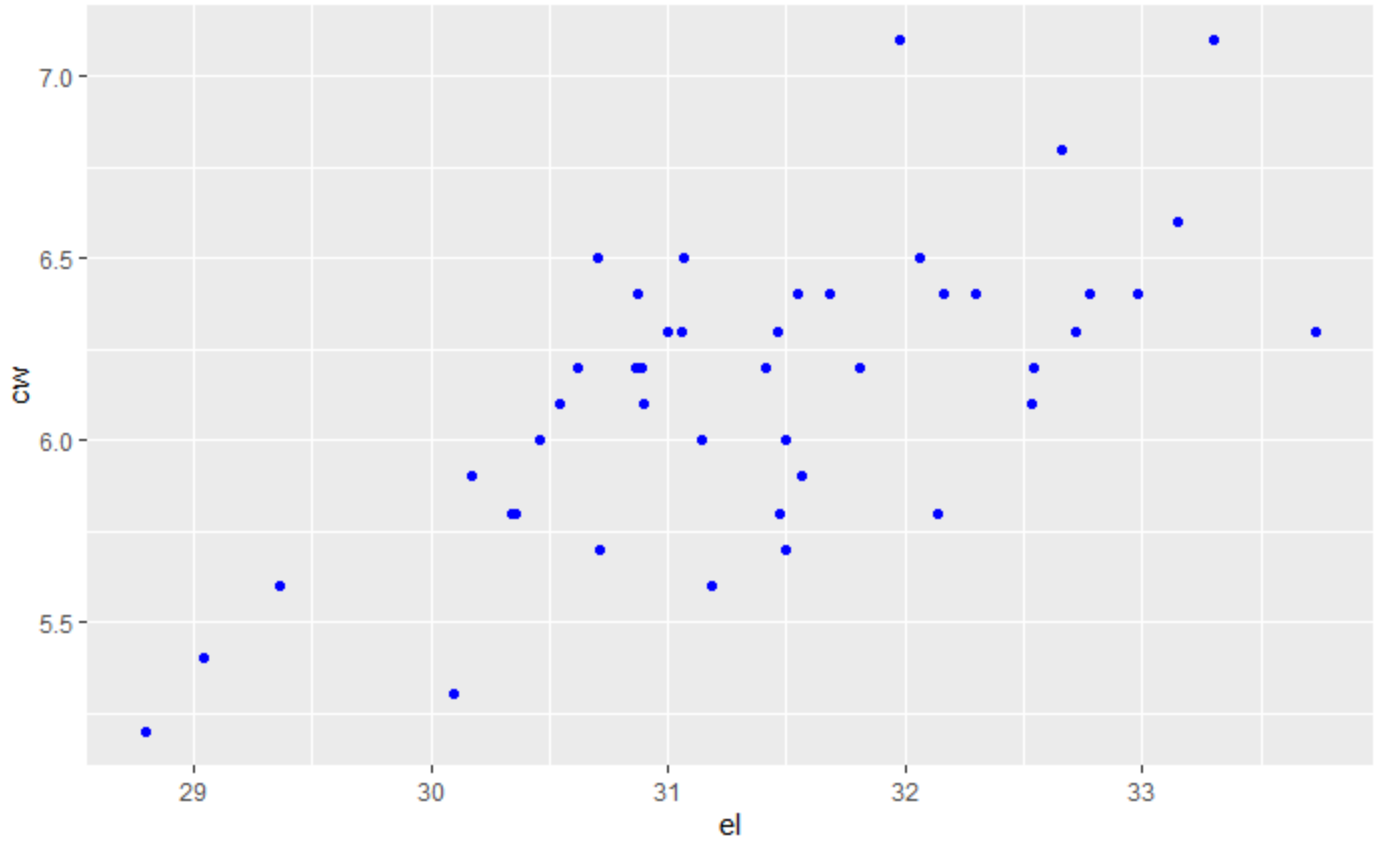
```
library(dplyr)
library(ggplot2)

chicks <- read.table("hw10/chicks.txt", header=TRUE)
chicks
```

el <dbl>	eb <dbl>	ew <dbl>	cw <dbl>
28.80	21.84	7.4	5.2
29.04	22.45	7.7	5.4
29.36	22.48	7.9	5.6
30.10	21.71	7.5	5.3
30.17	22.75	8.3	5.9
30.34	22.84	8.5	5.8
30.36	22.50	8.2	5.8
30.46	22.72	8.3	6.0
30.54	23.31	9.0	6.1
30.62	22.94	8.5	6.2

1-10 of 44 rows Previous **1** 2 3 4 5 Next

Hide

```
cw <- chicks$cw
el <- chicks$el
ew <- chicks$ew
eb <- chicks$eb

chicks %>% ggplot(aes(x=el,y=cw)) +
  geom_point(color='blue')
```



Plot looks linear and homoscedastic. The model $cw_i = \beta_0 + \beta_1 el_i + \epsilon_i$ should be good.

Hide

```
cw_mean <- mean(cw)
cw_mean
```

```
[1] 6.145455
```

Hide

```
sd(cw)
```

```
[1] 0.4105892
```

Hide

```
el_mean <- mean(el)
el_mean
```

```
[1] 31.38955
```

Hide

```
sd(el)
```

```
[1] 1.100892
```

Hide

```
cor(cw,el)
```

```
[1] 0.6761419
```

Hide

```
slope <- cov(cw,el)/var(el)
slope
```

```
[1] 0.2521743
```

Hide

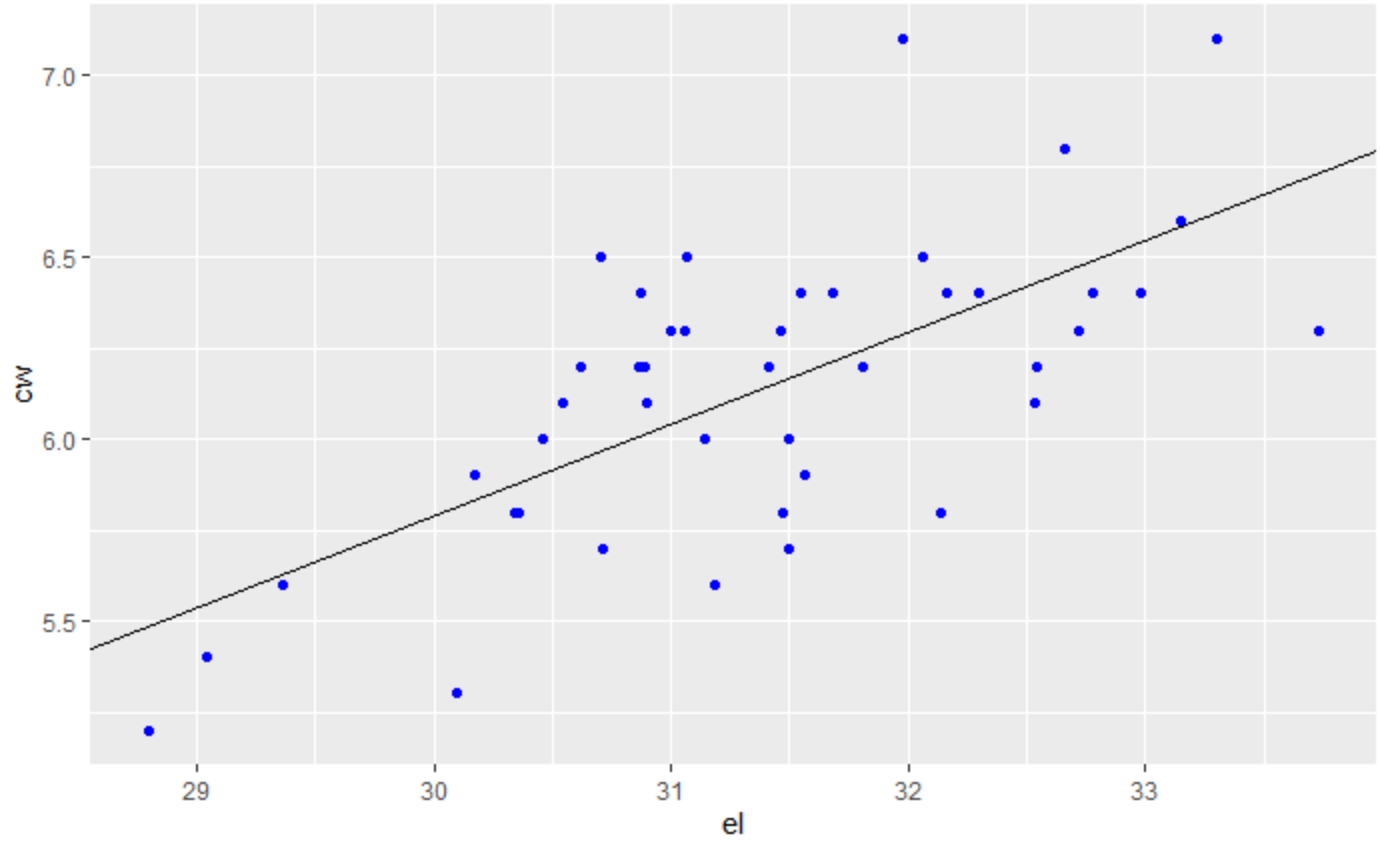
```
intercept <- cw_mean - slope*el_mean
intercept
```

```
[1] -1.770182
```

Therefore, the regression line is $\hat{y} = 0.252x - 1.77$

Hide

```
chicks %>% ggplot(aes(x=el,y=cw)) +
  geom_point(color='blue')+
  geom_abline(slope=0.252,intercept=-1.77)
```



b)

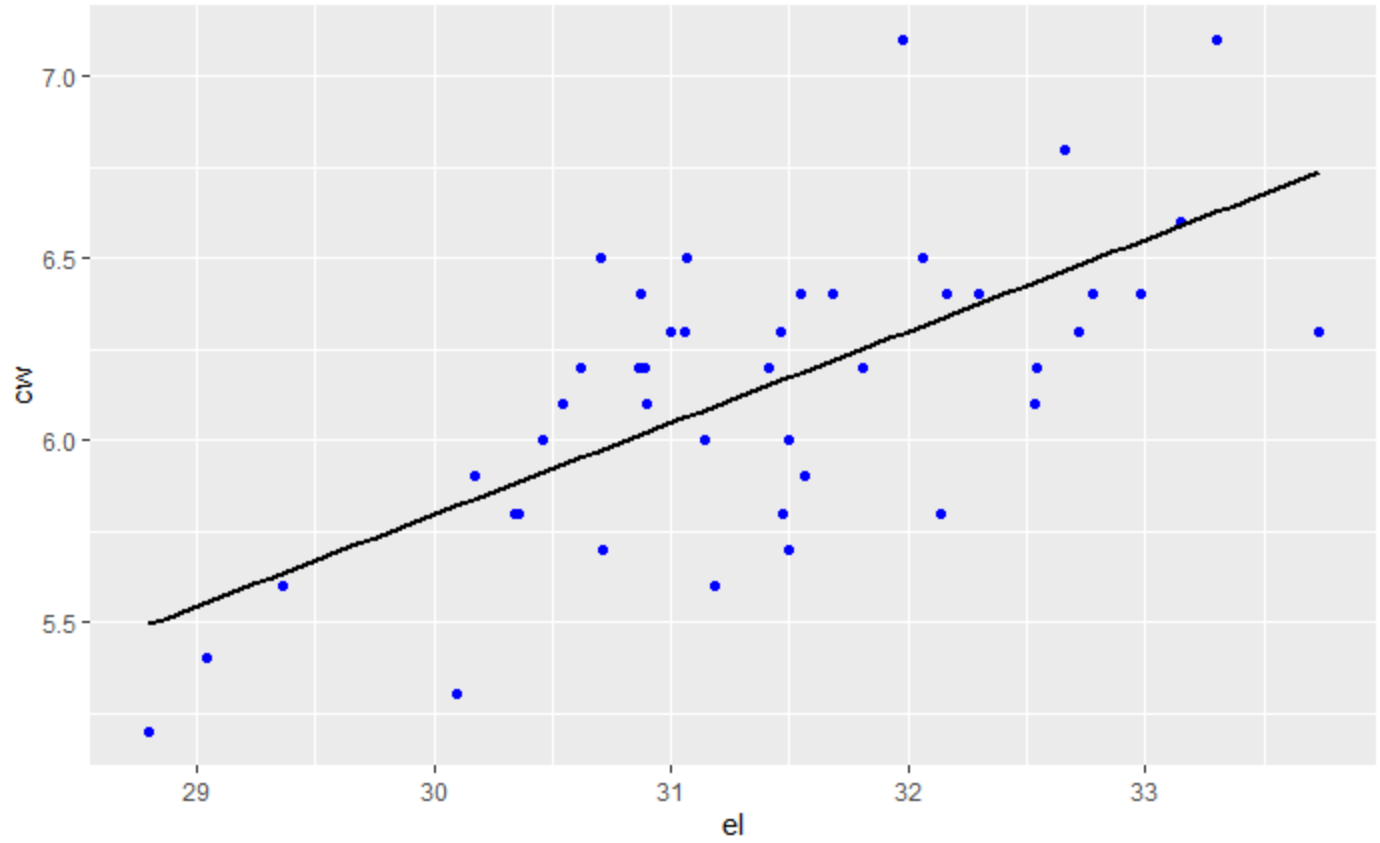
Hide

```
summary(lm(cw~el))
```

```
Call:
lm(formula = cw ~ el)

Residuals:
     Min       1Q   Median       3Q      Max
-0.53470 -0.19461  0.01778  0.18613  0.80565

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.7702     1.3317  -1.329    0.191
el            0.2522     0.0424   5.947 4.73e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3061 on 42 degrees of freedom
Multiple R-squared:  0.4572,    Adjusted R-squared:  0.4442
F-statistic: 35.37 on 1 and 42 DF,  p-value: 4.727e-07
```

Hide

```
chicks %>% ggplot(aes(x=el,y=cw)) +
  geom_point(color='blue')+
  geom_smooth(method=lm, se=FALSE,color='black')
```



R estimates the slope and the intercept at the same values that we computed in part a). Given a p-value of 0.191 for the intercept, we fail to reject the null hypothesis in this case. However, given the p-value for the slope of 4.727e-07, we reject the null hypothesis.

c)

Hide

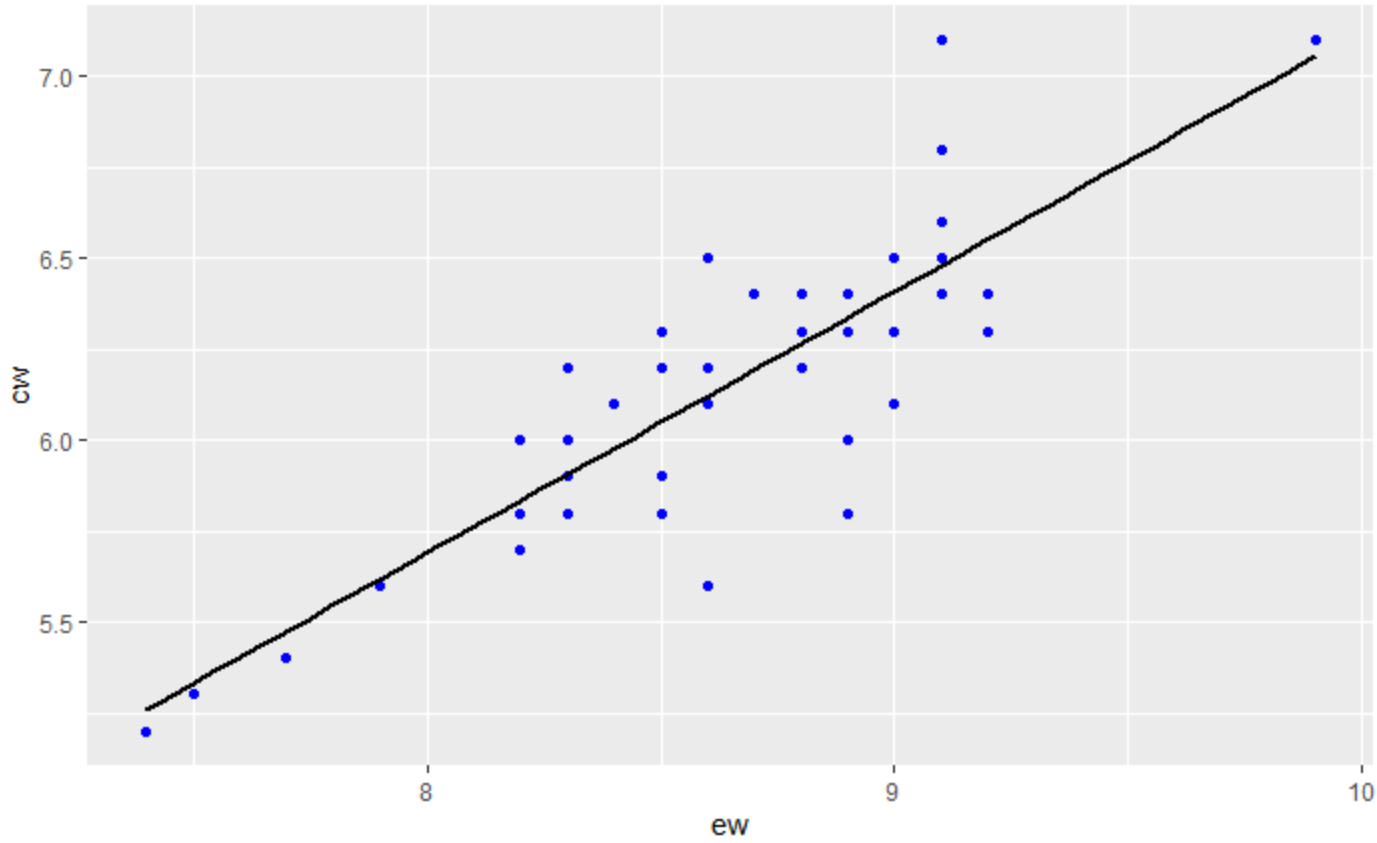
```
cor(chicks)
```

```
          el        eb        ew        cw
el 1.0000000 0.4027642 0.7924492 0.6761419
eb 0.4027642 1.0000000 0.8390767 0.7336866
ew 0.7924492 0.8390767 1.0000000 0.8472275
cw 0.6761419 0.7336866 0.8472275 1.0000000
```

The variable with the highest correlation to chick weight (cw) is egg weight (ew), 0.8472275.

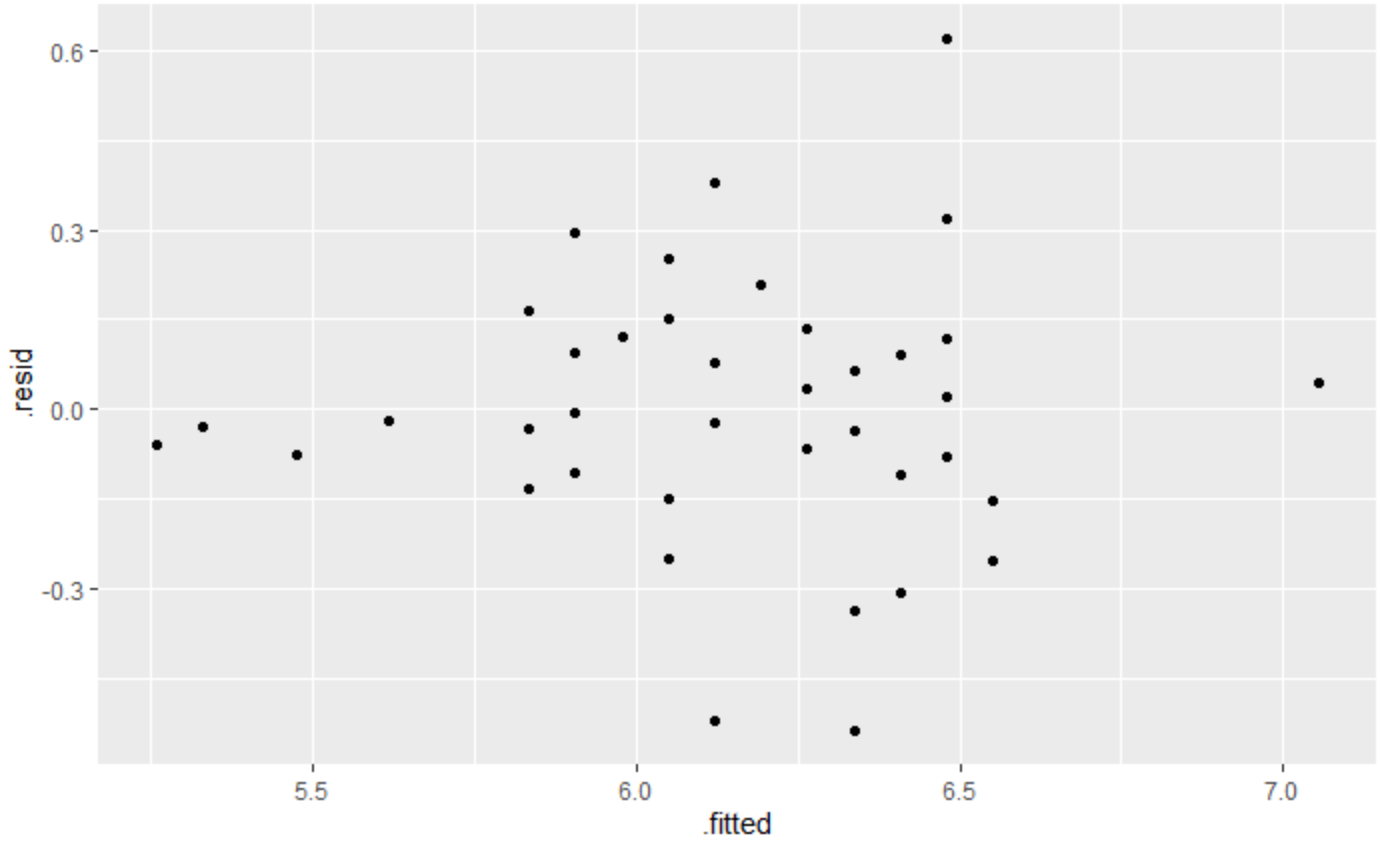
Hide

```
chicks %>% ggplot(aes(x=ew,y=cw)) +
  geom_point(color='blue')+
  geom_smooth(method=lm, se=FALSE,color='black')
```



Hide

```
ggplot(lm(cw~ew))+
  geom_point(aes(x=.fitted,y=.resid))
```



Linear plot but there is no heteroscedastcity.

d)

Hide

```
summary(lm(cw~ew))
```

```
Call:
lm(formula = cw ~ ew)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5365 -0.1147 -0.0117  0.1259  0.6198

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.05827    0.60114  -0.097    0.923
ew           0.71852    0.06952  10.336 4.15e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2207 on 42 degrees of freedom
Multiple R-squared:  0.7178,    Adjusted R-squared:  0.7111
F-statistic: 106.8 on 1 and 42 DF,  p-value: 4.148e-13
```

Hide

```
coef <- coefficients(summary(lm(cw~ew)))
coef
```

```
               Estimate Std. Error     t value     Pr(>|t|)
(Intercept) -0.05827226 0.60114179 -0.09693597 9.232383e-01
ew           0.71851534 0.06951749 10.33574908 4.147671e-13
```

Hide

```
# Therefore, the confidence interval is:
n <- length(cw)
se <- .2207+sqrt(1/n+(8.5-mean(ew))^2/((n-1)*var(ew)))
predict_weight <- coef[1]+8.5*coef[2]

CI <- c(predict_weight-qt(0.975,df=42)*se,predict_weight+qt(0.975,df=42)*se)
CI
```

```
[1] 5.287765 6.810451
```

e)

Hide

```
# new se
se <- .2207*sqrt(1+ 1/n + (8.5-mean(ew))^2/((n-1)*var(ew)))
PI <- c(predict_weight-qt(0.975,df=42)*se,predict_weight+qt(0.975,df=42)*se)
PI
```

```
[1] 5.598292 6.499924
```

f)

12 grams is too heavy, not seen as a part of our dataset, would not be a good model/it is unknown. (hint: bware of extrapolation)

Problem 10B

The object is still to find a good way to predict the weight of a chick given measurements on the egg, using linear regression as the only tool. The difference between this problem and Problem 10A is that now you are going to use a combination of variables to estimate the weights of the chicks.

a)

Hide

```
summary(lm(cw~eb+el))
```

```

Call:
lm(formula = cw ~ eb + el)

Residuals:
     Min       1Q   Median       3Q      Max
-0.53454 -0.12055  0.01582  0.10292  0.68326

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.73860    1.78777  -6.007 4.23e-07 ***
eb            0.50566    0.08419   6.006 4.24e-07 ***
el            0.16945    0.03420   4.955 1.29e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.226 on 41 degrees of freedom
Multiple R-squared:  0.7112,    Adjusted R-squared:  0.6972
F-statistic: 50.49 on 2 and 41 DF,  p-value: 8.732e-12
```

The regressions seem to be fairly similar in most aspects, noticeably the R^2 of 0.7112.

b)

Hide

```
summary(lm(ew~eb+el))
```

```

Call:
lm(formula = ew ~ eb + el)

Residuals:
      Min        1Q    Median        3Q       Max
-0.231315 -0.076288 -0.004403  0.054513  0.273872

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -14.22199    0.87175  -16.31   <2e-16 ***
eb            0.67190    0.04105   16.37   <2e-16 ***
el            0.23858    0.01667   14.31   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1102 on 41 degrees of freedom
Multiple R-squared:  0.9506,    Adjusted R-squared:  0.9482
F-statistic: 394.6 on 2 and 41 DF,  p-value: < 2.2e-16
```

R^2 has gone up to 0.9506, which essentially means that ew is a linear function of egg length and breadth. This also explains why the regressions are similar from before since ew is roughly the same as eb+el.

c)

Hide

```
summary(lm(cw~eb+el+ew))
```

```
Call:
lm(formula = cw ~ eb + el + ew)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.52731 -0.12047 -0.00941  0.11040  0.64121 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.60567    4.84329  -0.951    0.347
eb           0.21591    0.22872   0.944    0.351
el           0.06657    0.08286   0.803    0.426
ew           0.43123    0.31701   1.360    0.181

Residual standard error: 0.2236 on 40 degrees of freedom
Multiple R-squared:  0.724, Adjusted R-squared:  0.7033 
F-statistic: 34.98 on 3 and 40 DF,  p-value: 2.903e-11
```

This regression shows odd results. Even though the R^2 slightly went up, the test itself does not make sense. Even though every coef has p-values that suggest that we fail to reject the slope in the null hypothesis, the F-stat shows us that we should reject it. This just means that the variables that we are using to predict are too correlated with one another.

d)

Just use the two that we did in a) to prevent what happpened here in part c). Prevents minimal correlation and is easier to understand.

Problem 10C

This problem concerns the dataset tox. The data are observations on a simple random sample of Hodgkins disease patients at Stanford Hospital, taken as part of a study of the toxicity of the treatment to the patients lungs.

Hide

```
tox <- read.table("hw10/tox.txt", header=TRUE)
tox
```

height <dbl>	rad <dbl>	chemo <dbl>	base <dbl>	month15 <dbl>
164	679	180	160.57	87.77
168	311	180	98.24	67.62
173	388	239	129.04	133.33
157	370	168	85.41	81.28
160	468	151	67.94	79.26
170	341	96	150.51	80.97
163	453	134	129.88	69.24
175	529	264	87.45	56.48
185	392	240	149.84	106.99
178	479	216	92.24	73.43

1-10 of 22 rows Previous **1** 2 3 Next

a)

Parametric t-test value of -6.15, 21 df. Non-parametric test of the f-stat of 246. Because of these numbers, it shows that the means are different and that the distributions are different from one another. Furthermore, since the t-test value in the first parametric test is negative, this means that at month15 the scores are lower than the base scores.

b)

Hide

```
cor(tox)
```

```
            height          rad        chemo       base    month15
height   1.0000000 -0.305206183  0.576824659 0.35422852 0.39052708
rad     -0.3052062  1.000000000 -0.003739408 0.09643241 0.04061608
chemo    0.5768247 -0.003739408  1.000000000 0.06218682 0.44578818
base     0.3542285  0.096432406  0.062186822 1.00000000 0.56137060
month15  0.3905271  0.040616084  0.445788177 0.56137060 1.00000000
```

Hide

```
summary(lm(tox$month15~tox$base+tox$chemo))
```

```
Call:
lm(formula = tox$month15 ~ tox$base + tox$chemo)

Residuals:
    Min      1Q  Median      3Q     Max
-30.611  -7.823  -2.261   8.782  32.914

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.99921   20.22704  -0.049  0.96112
tox$base     0.43447    0.13383   3.246  0.00425 **
tox$chemo    0.18975    0.07592   2.500  0.02176 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.44 on 19 degrees of freedom
Multiple R-squared:  0.4846,    Adjusted R-squared:  0.4304
F-statistic: 8.933 on 2 and 19 DF,  p-value: 0.001842
```

Given the correlation matrix, I decided to use base and chemo to estimate month15. Given this, the R^2 is 0.4846 and the adjusted is 0.4304. Given this, the prediction model would most likely be usable.

Problem 10D

The dataset baby contains observations on mothers and their newborns at Kaiser Hospital (data courtesy of D. Nolan).

Hide

```
baby <- read.table("hw10/baby.txt", header=TRUE)
baby
```

bw <int>	gd <int>	ma <int>	mh <int>	mpw <int>	sm <int>
120	284	27	62	100	0
113	282	33	64	135	0
128	279	28	64	115	1
108	282	23	67	125	1
136	286	25	62	93	0
138	244	33	62	178	0
132	245	23	65	140	0
120	289	25	62	125	0
143	299	30	66	136	1

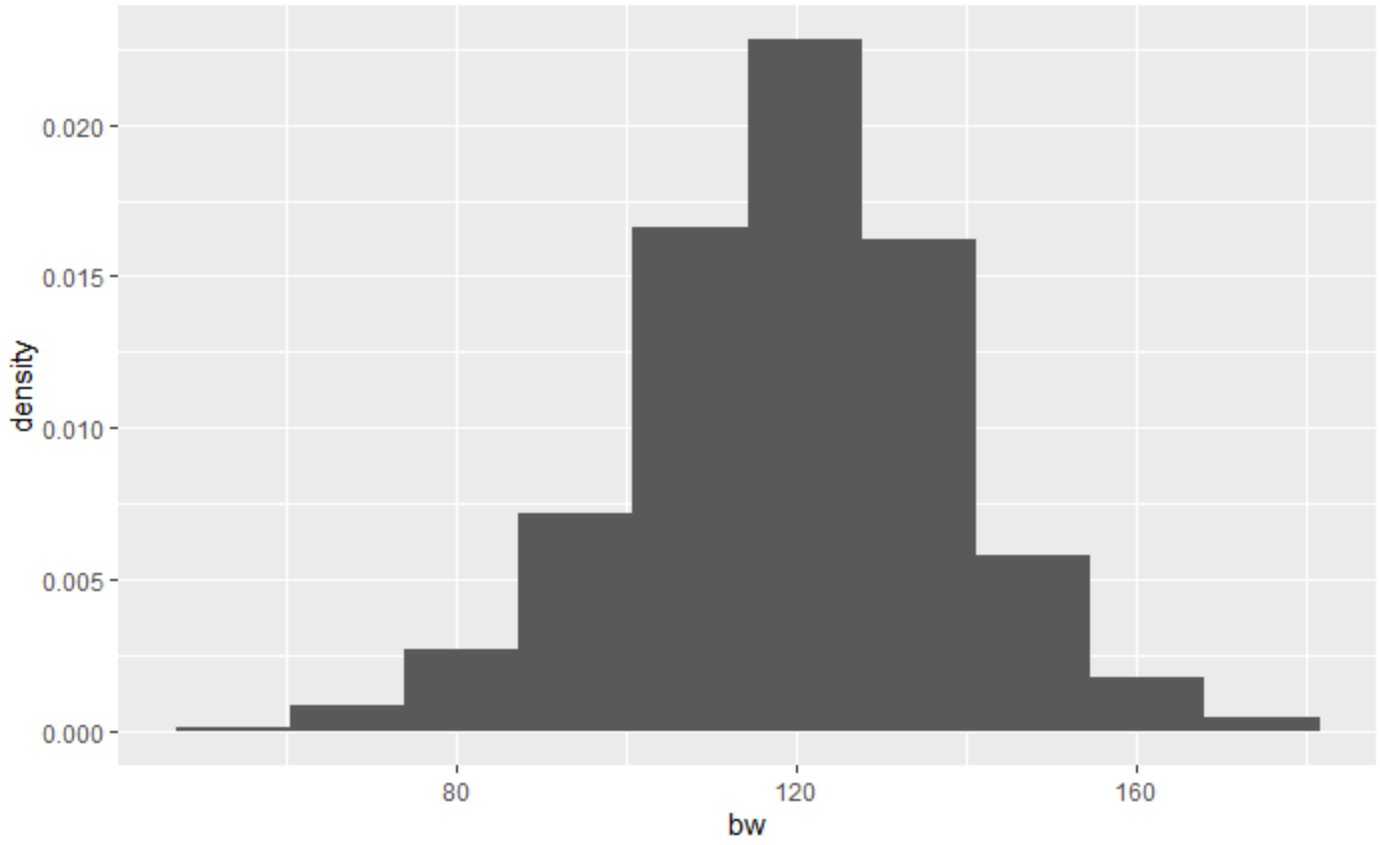
bw <int>	gd <int>	ma <int>	mh <int>	mpw <int>	sm <int>
140	351	27	68	120	0

1-10 of 1,174 rows Previous **1** 2 3 4 5 6 … 100 Next

a)

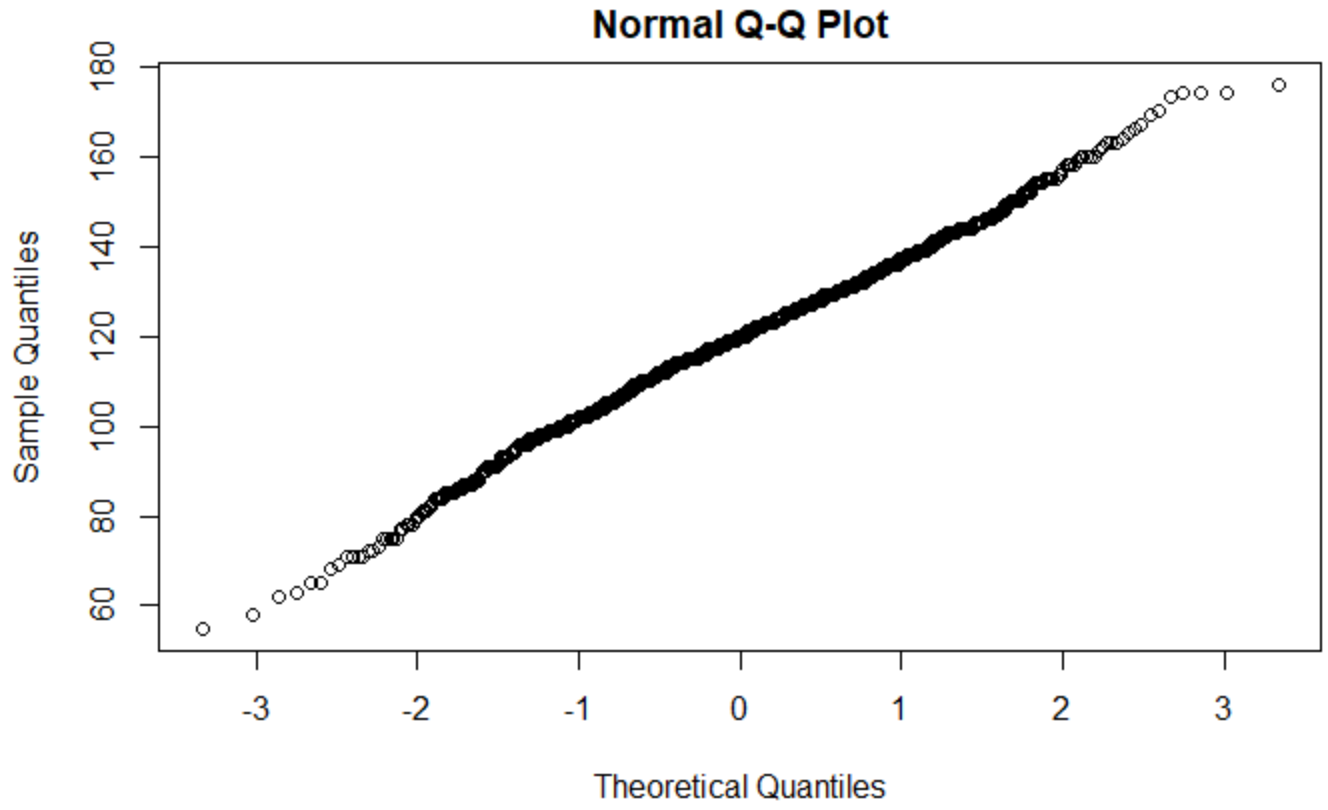
Hide

```
baby %>% ggplot(aes(x=bw,y=..density..))+
  geom_histogram(bins=10)
```



Hide

```
qqnorm(baby$bw)
```

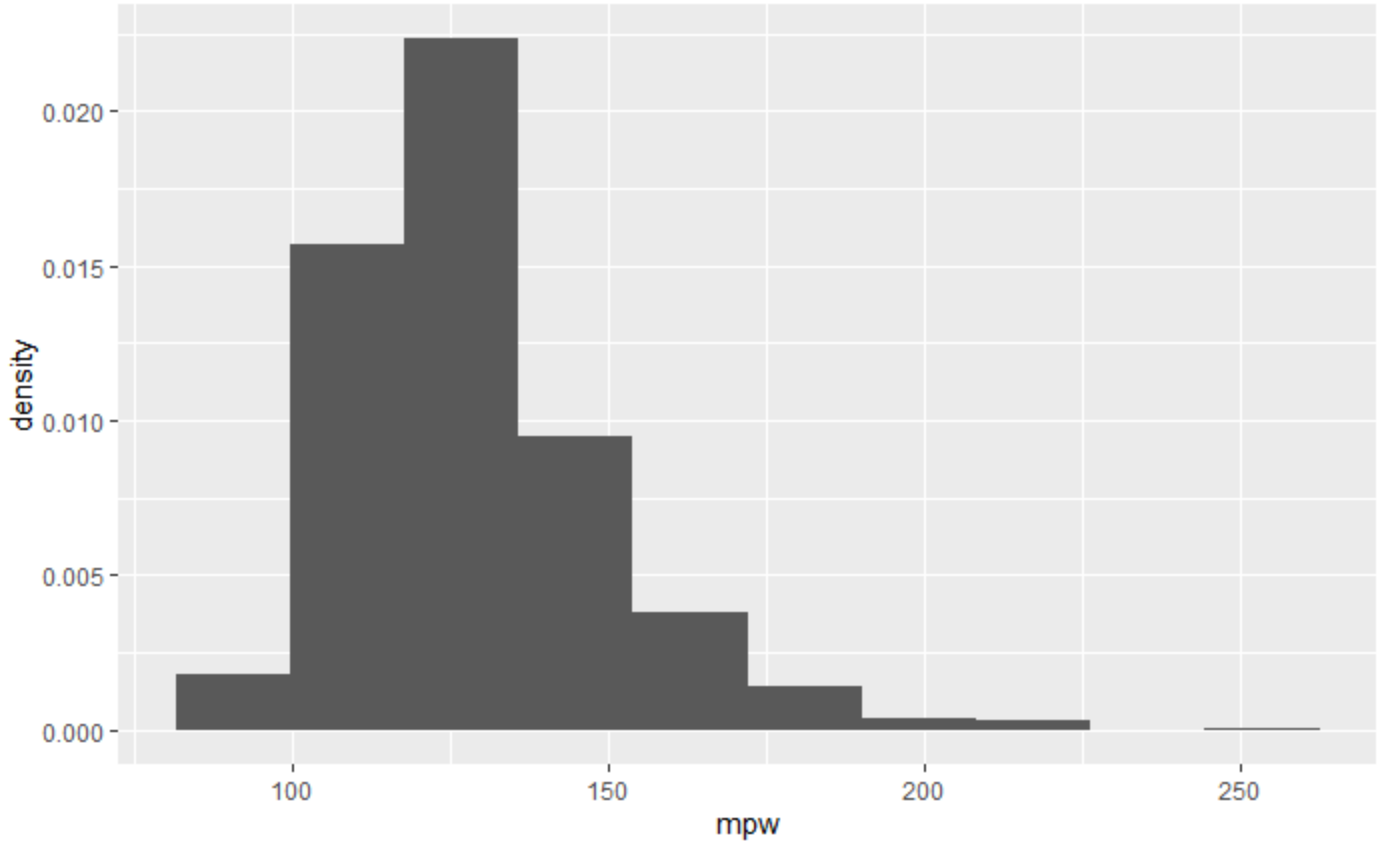


Plot(s) looks normal.

b)

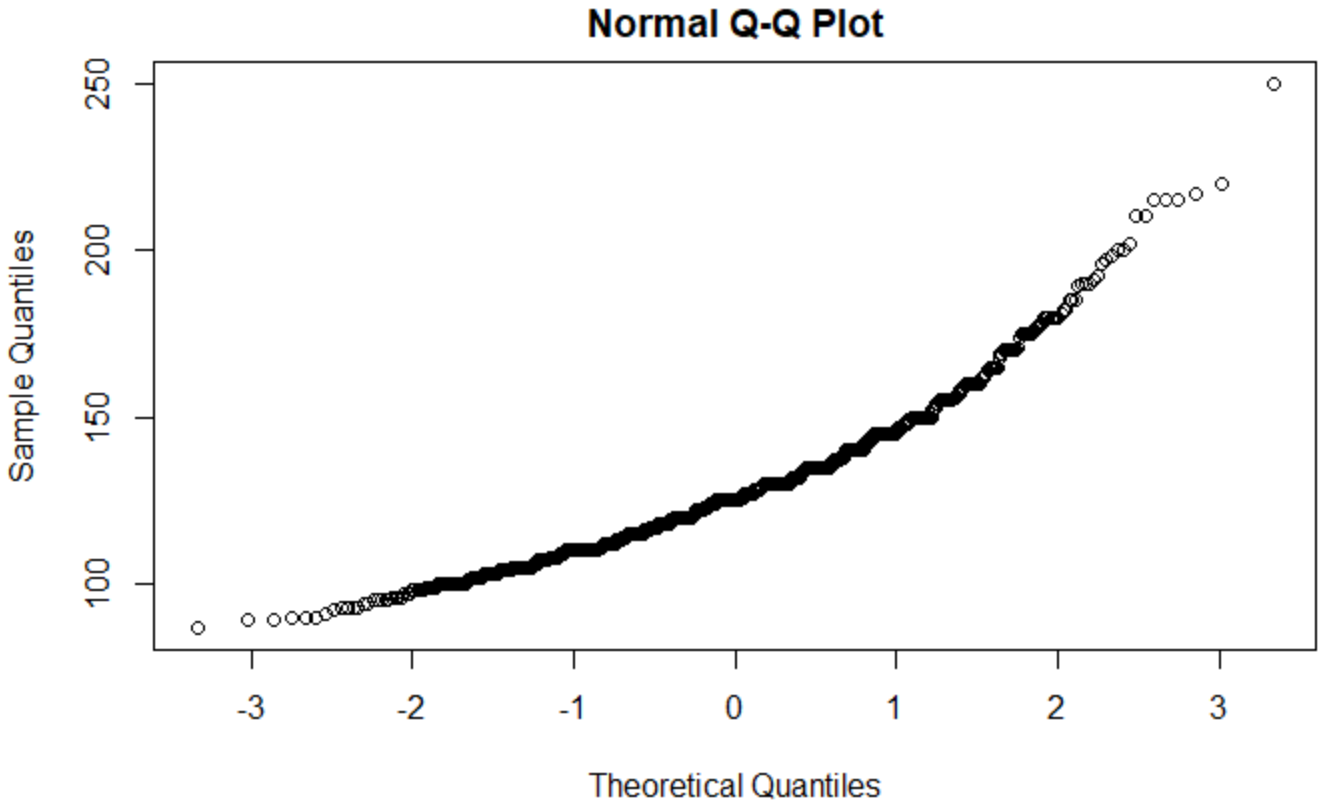
Hide

```
baby %>% ggplot(aes(x=mpw,y=..density..))+
  geom_histogram(bins=10)
```



Hide

```
qqnorm(baby$mpw)
```



The plot is skewed right. The qqplot has an almost convex shape to it. If the histo plot was skewed left, it would have more of a concave shape to it.

c)

Hide

```
cor(baby)
```

```
             bw          gd           ma           mh         mpw          sm
bw   1.00000000  0.40754279  0.026982911  0.203704177  0.15592327 -0.24679951
gd   0.40754279  1.00000000 -0.053424774  0.070469902  0.02365494 -0.06026684
ma   0.02698291 -0.05342477  1.000000000 -0.006452846  0.14732211 -0.06777194
mh   0.20370418  0.07046990 -0.006452846  1.000000000  0.43528743  0.01750660
mpw  0.15592327  0.02365494  0.147322111  0.435287428  1.00000000 -0.06028140
sm  -0.24679951 -0.06026684 -0.067771942  0.017506595 -0.06028140  1.00000000
```

Given the correlation matrix, I would use gd, mh, and sm as my prediction variables.

Hide

```
summary(lm(baby$bw~baby$gd+baby$mh+baby$sm))
```

```

Call:
lm(formula = baby$bw ~ baby$gd + baby$mh + baby$sm)

Residuals:
    Min      1Q  Median      3Q     Max
-57.373 -10.365  -0.450   9.853  53.250

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -83.01125   13.87372  -5.983 2.90e-09 ***
baby$gd       0.43631    0.02915  14.969  < 2e-16 ***
baby$mh       1.31198    0.18443   7.114 1.96e-12 ***
baby$sm      -8.52262    0.95370  -8.936  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.91 on 1170 degrees of freedom
Multiple R-squared:  0.2482,    Adjusted R-squared:  0.2463
F-statistic: 128.7 on 3 and 1170 DF,  p-value: < 2.2e-16
```

d)

Even though the other variables have a coef of 0.43 and 1.31 respectively, the smoking indicator variable is -8.52. Because of this, we can conclude that if a mother smoked, their baby would be on average -8.52 oz lighter than if the mother didn’t smoke (assuming all else equal).

Problem 10E

The dataset women contains the average weight in pounds (Column 2) for American women whose heights, correct to the nearest inch, are given in Column 1.

Hide

```
women <- read.table("hw10/women.txt", header=TRUE)
women
```

h <int>	avew <int>
58	115
59	117
60	120
61	123
62	126
63	129
64	132
65	135
66	139
67	142

1-10 of 15 rows Previous **1** 2 Next

a)

Hide

```
summary(lm(women$h~women$avew))
```

```
Call:
lm(formula = women$h ~ women$avew)

Residuals:
     Min       1Q   Median       3Q      Max
-0.83233 -0.26249  0.08314  0.34353  0.49790

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 25.723456   1.043746   24.64 2.68e-12 ***
women$avew   0.287249   0.007588   37.85 1.09e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.44 on 13 degrees of freedom
Multiple R-squared:  0.991, Adjusted R-squared:  0.9903
F-statistic:  1433 on 1 and 13 DF,  p-value: 1.091e-14
```

In this case, the value of R^2 is 0.991, essentially being linear. Because of this, this indicates that height and weight are essentially identical to one another in terms of regression.

b)

The correlation would be less. Each dataset represents many women, if the point is replaced by all the data of the women of those heights it would decrease the accuarcy of the data/make the data harder to read and reduce the correlation.

c)

Residual plot of linear regression suggests that the df that should be used is 2.

Problem 10F

Hide

```
bodytemp <- read.csv("hw10/bodytemp.csv")
bodytemp
```

	temperature <dbl>	gender <int>	rate <int>
	96.3	1	70
	96.7	1	71
	96.9	1	74
	97.0	1	80
	97.1	1	73

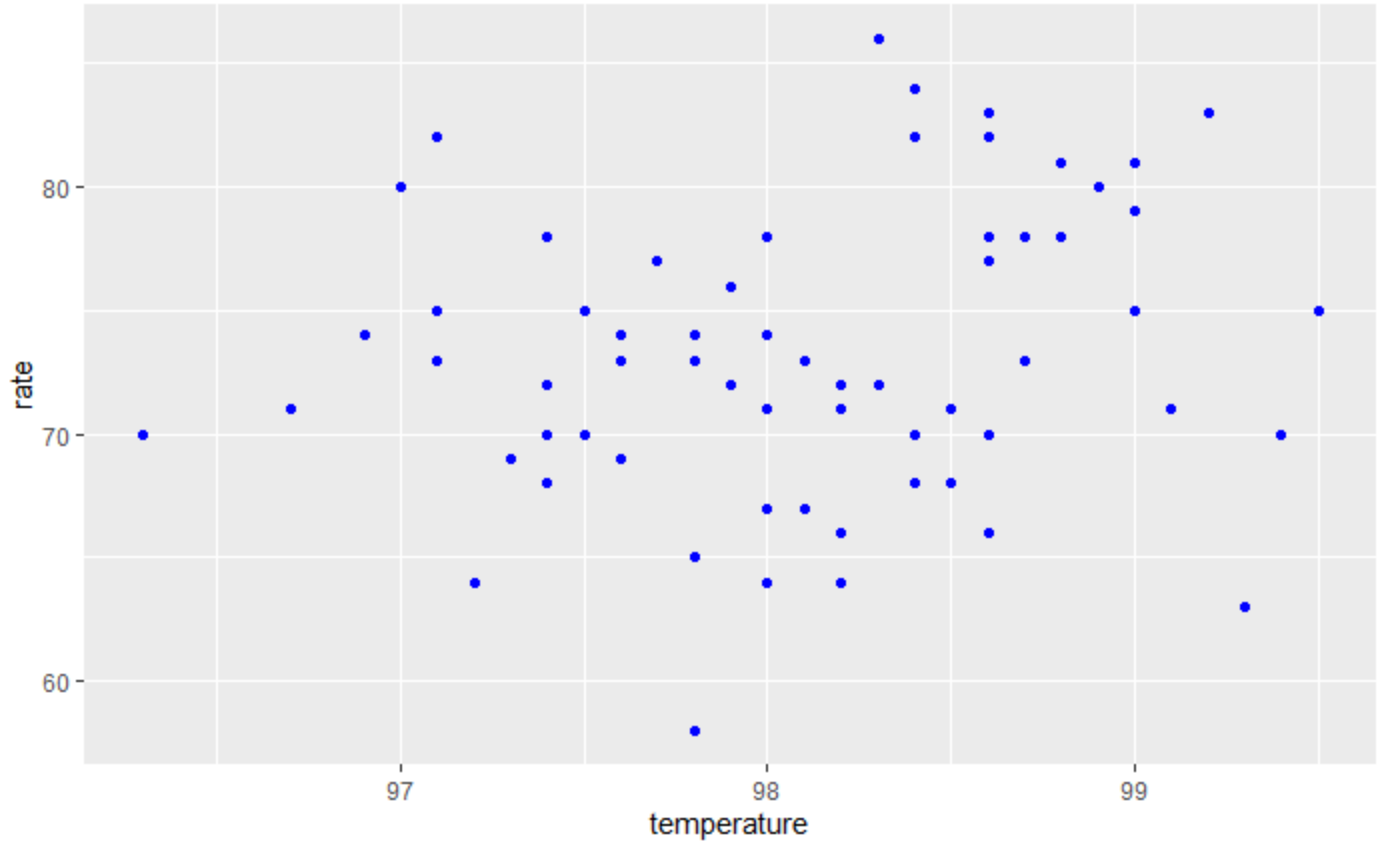
temperature <dbl>	gender <int>	rate <int>
97.1	1	75
97.1	1	82
97.2	1	64
97.3	1	69
97.4	1	70

1-10 of 130 rows Previous **1** 2 3 4 5 6 … 13 Next

a)

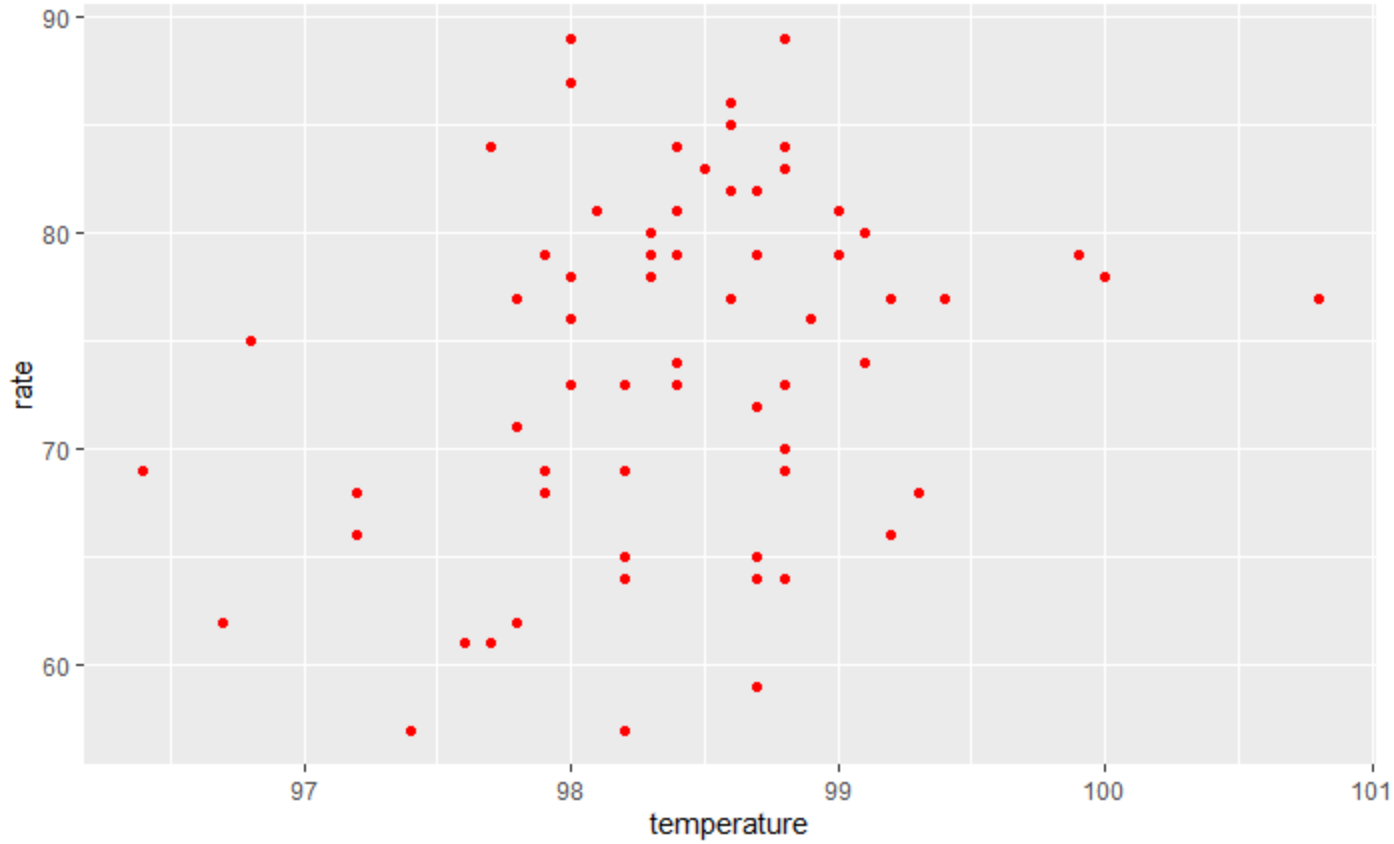
Hide

```
male_bt <- bodytemp[bodytemp$gender==1,]
male_bt %>% ggplot(aes(x=temperature,y=rate)) +
  geom_point(color='blue')
```



Hide

```
female_bt <- bodytemp[bodytemp$gender==2,]
female_bt %>% ggplot(aes(x=temperature,y=rate)) +
  geom_point(color='red')
```



Overall the plots are similar in shape, no clear relationship at all just plots of data.

b)

Comparing the two plots, it shows that the heartrate of men have less variance than the heartrate of women.

c)

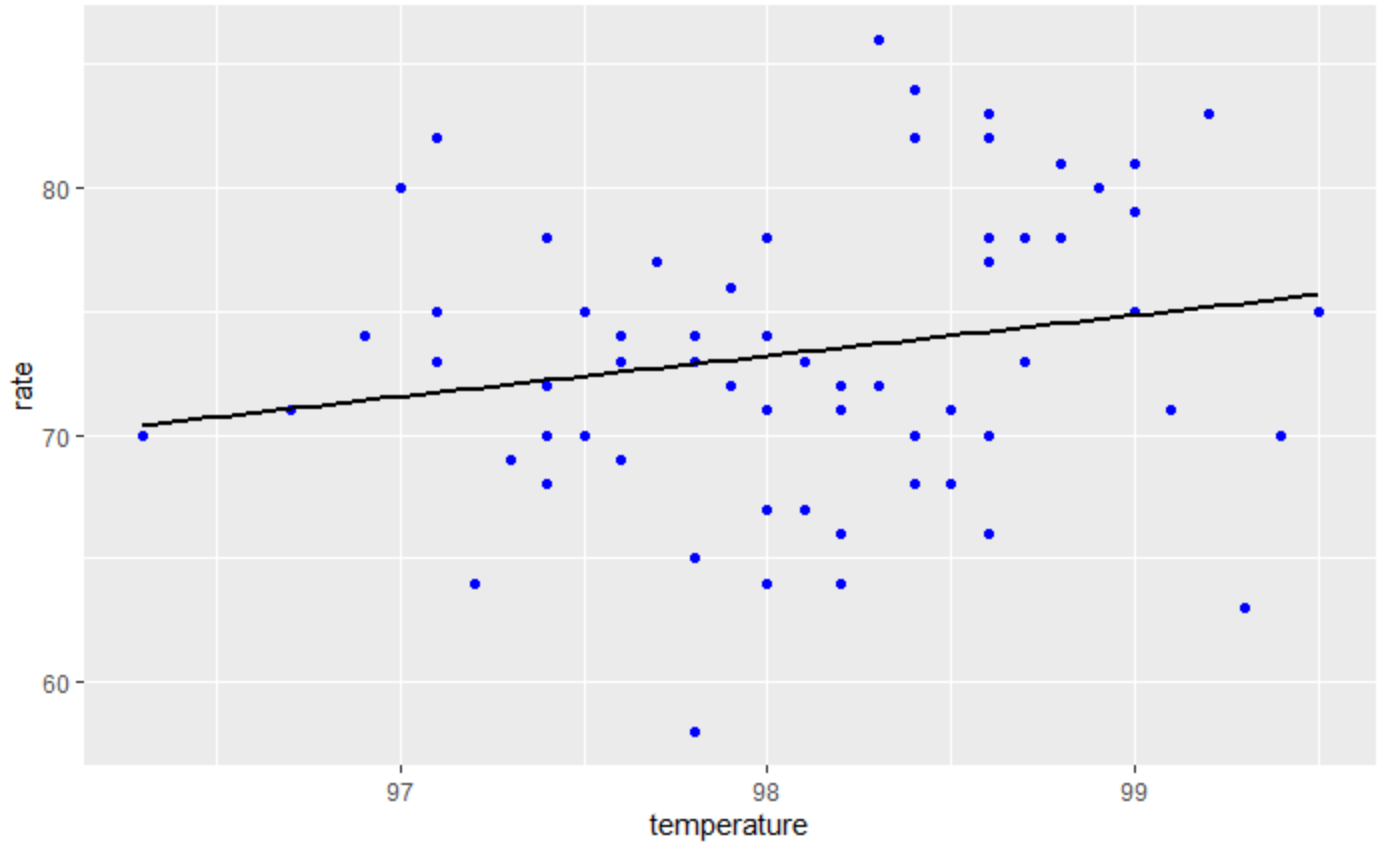
Hide

```
summary(lm(male_bt$temperature~male_bt$rate))
```

```
Call:
lm(formula = male_bt$temperature ~ male_bt$rate)

Residuals:
     Min       1Q   Median       3Q      Max
-1.72624 -0.49603  0.05291  0.48766  1.43659

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  96.39789    1.08154  89.130   <2e-16 ***
male_bt$rate  0.02326    0.01469   1.583    0.118
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6907 on 63 degrees of freedom
Multiple R-squared:  0.03826,   Adjusted R-squared:  0.02299
F-statistic: 2.506 on 1 and 63 DF,  p-value: 0.1184
```

Hide

```
male_bt %>% ggplot(aes(x=temperature,y=rate)) +
  geom_point(color='blue')+
  geom_smooth(method=lm, se=FALSE,color='black')
```



$R^2 is 0.038. The regression just shows the same stuff that we saw in part a).

d)

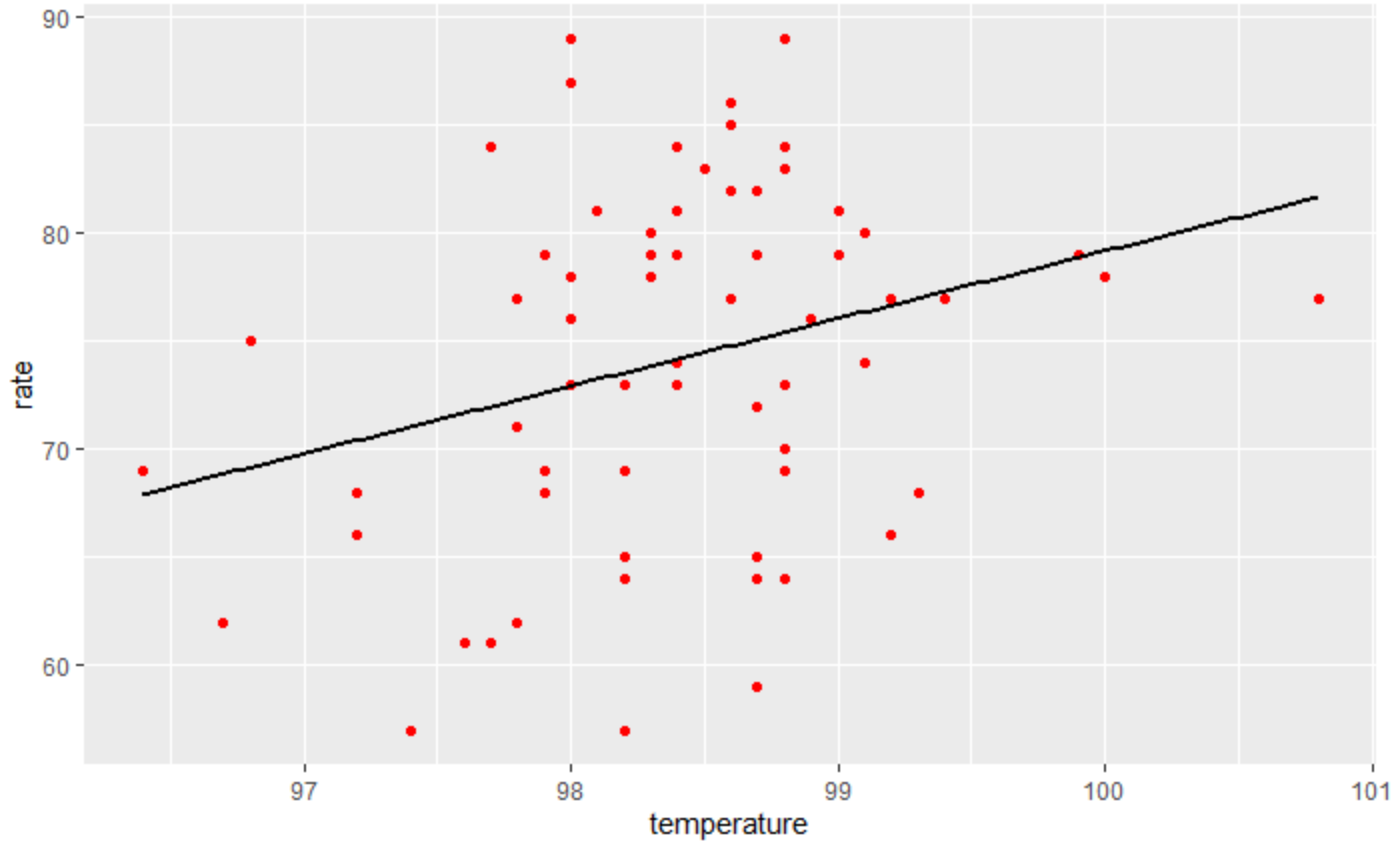
Hide

```
summary(lm(female_bt$temperature~female_bt$rate))
```

```

Call:
lm(formula = female_bt$temperature ~ female_bt$rate)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.8582 -0.3635 -0.0582  0.4576  2.3312 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)    96.44211    0.82576 116.792   <2e-16 ***
female_bt$rate  0.02632    0.01107   2.377   0.0205 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7179 on 63 degrees of freedom
Multiple R-squared:  0.08233,   Adjusted R-squared:  0.06776 
F-statistic: 5.652 on 1 and 63 DF,  p-value: 0.02048
```

Hide

```
female_bt %>% ggplot(aes(x=temperature,y=rate)) +
  geom_point(color='red')+
  geom_smooth(method=lm, se=FALSE,color='black')
```



While the R^2 is higher at around 0.082. It essentially shows the same as a).

e)

Slope for men is about 1.645 with an SE of 1.039. The slope for women was estimated 3.128 with an SE of 1.316. Calculate the difference, 1.483 with an SE of 1.68.

Hide

```
CI = c((1.483 - 2*1.68),(1.483 + 2*1.68))
CI
```

```
[1] -1.877  4.843
```

Since the CI contains 0, we can conclude that the slopes are equal.

f)

Given a difference in intercept of 145.657 and an SE of 164.767:

Hide

```
CI = c((145.657 - 2*164.767),(145.657 + 2*164.767))
CI
```

```
[1] -183.877  475.191
```

Since the CI contains 0, we can conclude that the intercepts are equal.