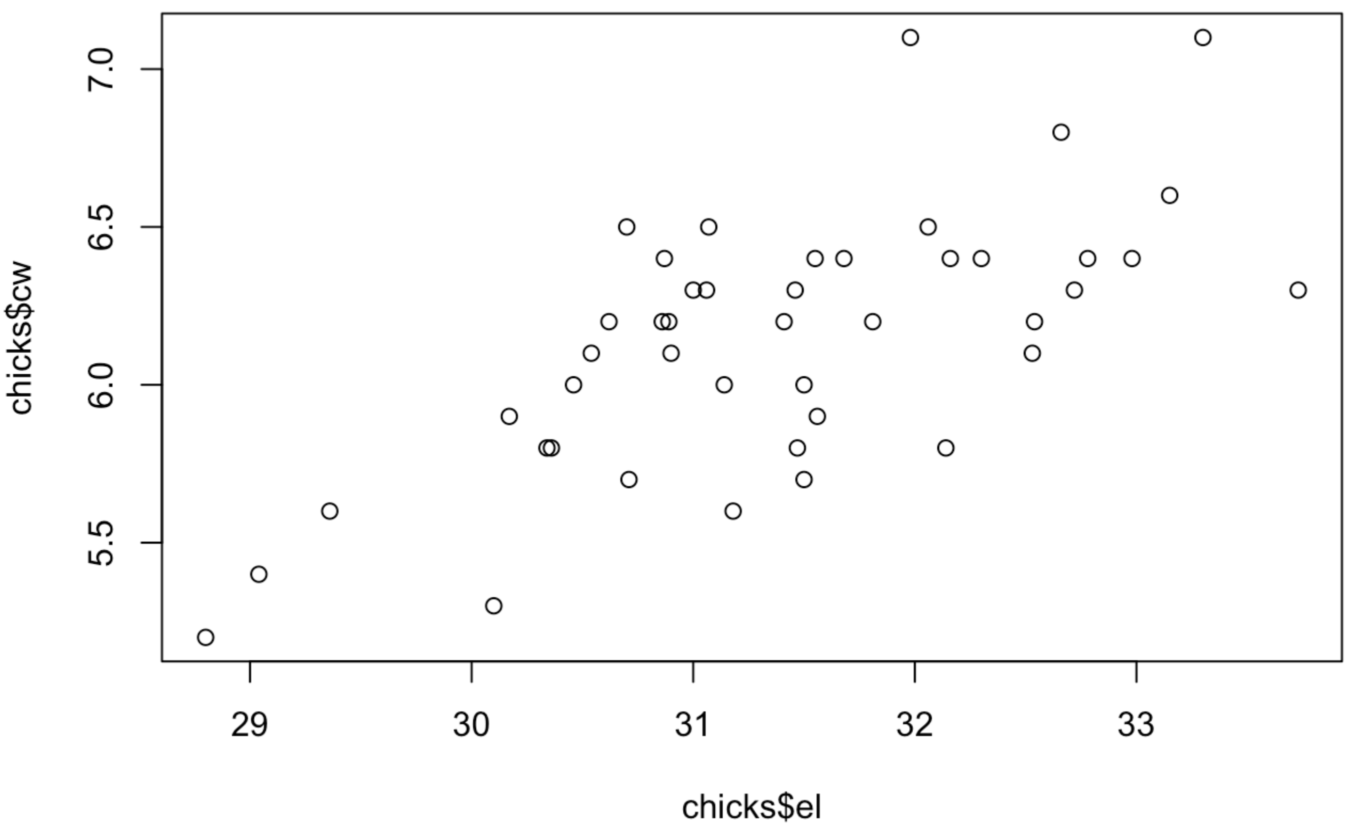
hw10-roni-shen

Roni Shen

11/20/2018

Problem 10A: part a)

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

# standard regression model: cwi = β0 + β1 * eli + εi, where cwi is the weight of the ith chick, eli the length of the egg from which it hatched, and εi the normal error.

plot(chicks$el, chicks$cw) # plot looks relatively linear and homoscedastic
```



```
# mean chick weight
mean(chicks$cw)
```

```
## [1] 6.145455
```

```
# standard deviation of chick weight
sd(chicks$cw)
```

```
## [1] 0.4105892
```

```
# mean egg length
mean(chicks$el)
```

```
## [1] 31.38955
```

```
# standard deviation of egg length
sd(chicks$el)
```

```
## [1] 1.100892
```

```
# correlation between egg length and chick weight
cor(chicks$el, chicks$cw)
```

```
## [1] 0.6761419
```

```
# slope of regression line
chicks_slope <- cor(chicks$el, chicks$cw) * sd(chicks$cw) / sd(chicks$el)

# intercept of regression line
chicks_inter <- mean(chicks$cw) - chicks_slope * mean(chicks$el)

# equation of the regression line: estimated chick weight = 0.2522 * egg length - 1.7
702
```

Problem 10A: part b)

```
lm(chicks$cw ~ chicks$el) # same as in part a)
```

```
##
## Call:
## lm(formula = chicks$cw ~ chicks$el)
##
## Coefficients:
## (Intercept)    chicks$el
##     -1.7702       0.2522
```

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

```
##
## Call:
## lm(formula = chicks$cw ~ chicks$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
## chicks$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
```

```
# t-test for intercept
# Null: intercept is 0
# Alternate: intercept is not 0
# Conclusion: p-value = .19, fail to reject that the null of the intercept is 0

# t-test for slope
# Null: Slope is 0
# Alternate: Slope is not 0
# Conclusion: t = 5.947, p-value is close to 0, slope is not 0

# F-test for slope
# Null: Slope is 0
# Alternate: Slope is not 0
# Conclusion: F - 35.37, p-value is close to 0, slope is not 0
```

Problem 10A: part c)

```
# finding the best predictor
cor(chicks$el, chicks$cw)
```

```
## [1] 0.6761419
```

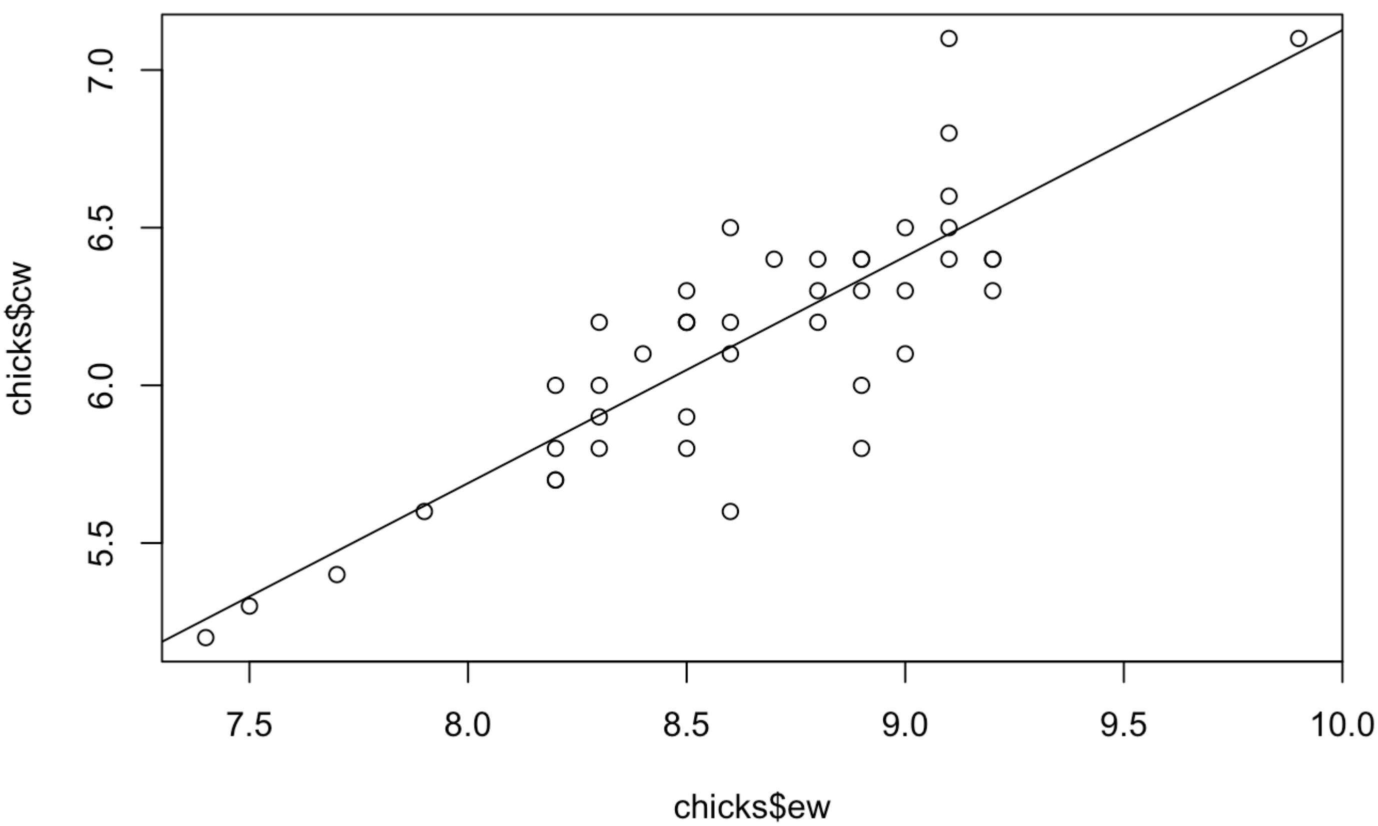
```
cor(chicks$eb, chicks$cw)
```

```
## [1] 0.7336866
```

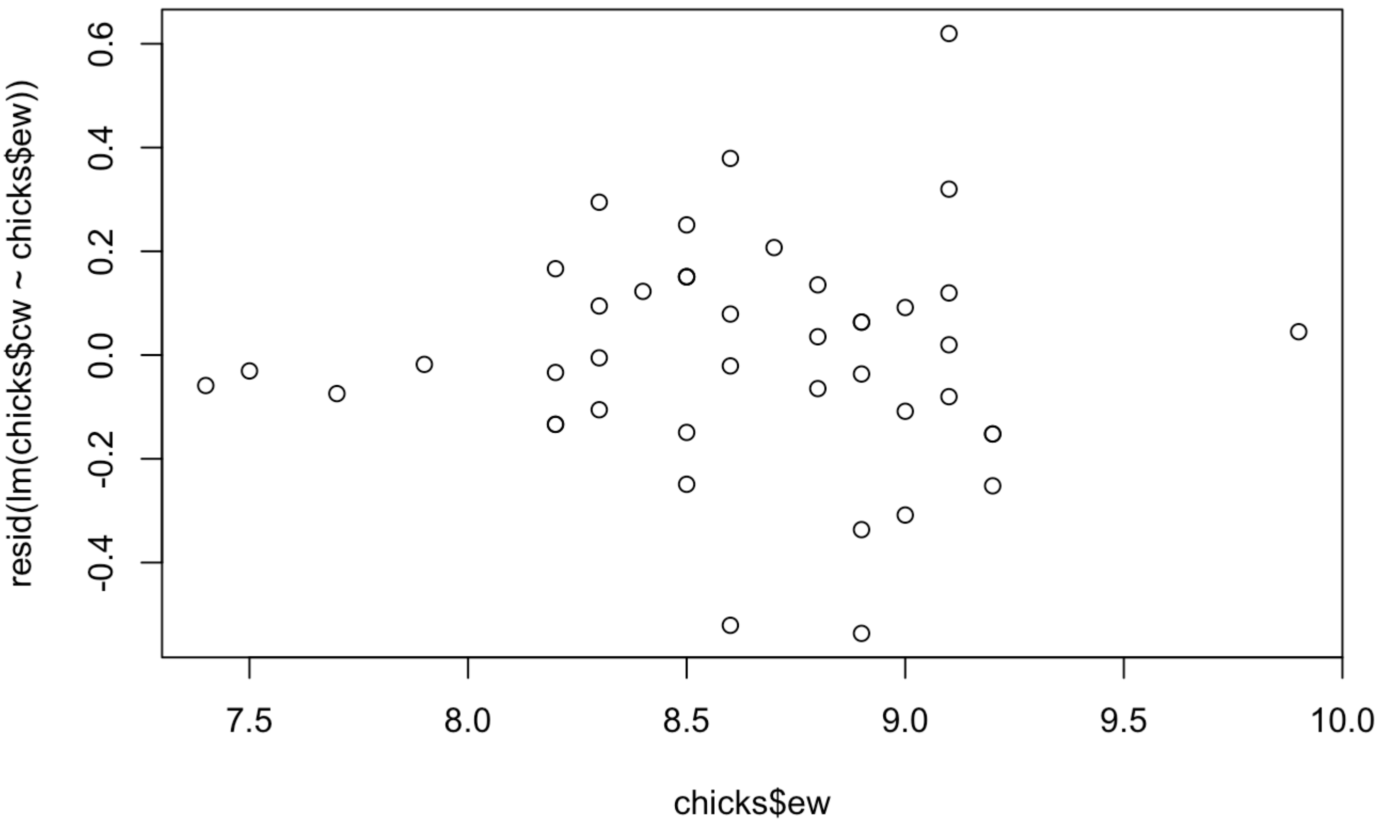
```
cor(chicks$ew, chicks$cw)
```

```
## [1] 0.8472275
```

```
# scatterplot with linear regression line
plot(chicks$ew, chicks$cw)
abline(lm(chicks$cw~chicks$ew))
```




```
# residual plot
plot(chicks$ew, resid(lm(chicks$cw~chicks$ew))) # The plot is relatively linear, hete
roscedasticity is more noticable in the middle values.
```



Problem 10A: part d)

```
a_d_mean <- (lm(chicks$cw~chicks$ew)[[1]][2] * 8.5 + lm(chicks$cw~chicks$ew)[[1]][1])
[[1]]
a_d_se <- 0.2207 * sqrt((1 / 44) + (8.5 - mean(chicks$ew)) ^ 2 / (43 * var(chicks$ew)
))

qt(.975, df = 42)
```

```
## [1] 2.018082
```

```
# 95% CI: (5.98, 6.12)
```

Problem 10A: part e)

```
a_e_se <- 0.2207 * sqrt((1 / 44) + (8.5 - mean(chicks$ew)) ^ 2 / (43 * var(chicks$ew)
) + 1)

# 95% PI: (6.00, 6.50)
```

Problem 10A: part f)

```
a_f_mean <- (lm(chicks$cw~chicks$ew)[[1]][2] * 12 + lm(chicks$cw~chicks$ew)[[1]][1])[
[1]]
a_f_se <- 0.2207 * sqrt((1 / 44) + (12 - mean(chicks$ew)) ^ 2 / (43 * var(chicks$ew))
)

# 95% CI: (8.09, 9.04)

a_f_se2 <- 0.2207 * sqrt((1 / 44) + (12 - mean(chicks$ew)) ^ 2 / (43 * var(chicks$ew)
) + 1)

# 95% PI: (7.91, 9.22)

# Warning: Due to extrapolation, making estimates outside the range of data can lead
to inaccurate estimates.
```

Problem 10B: part a)

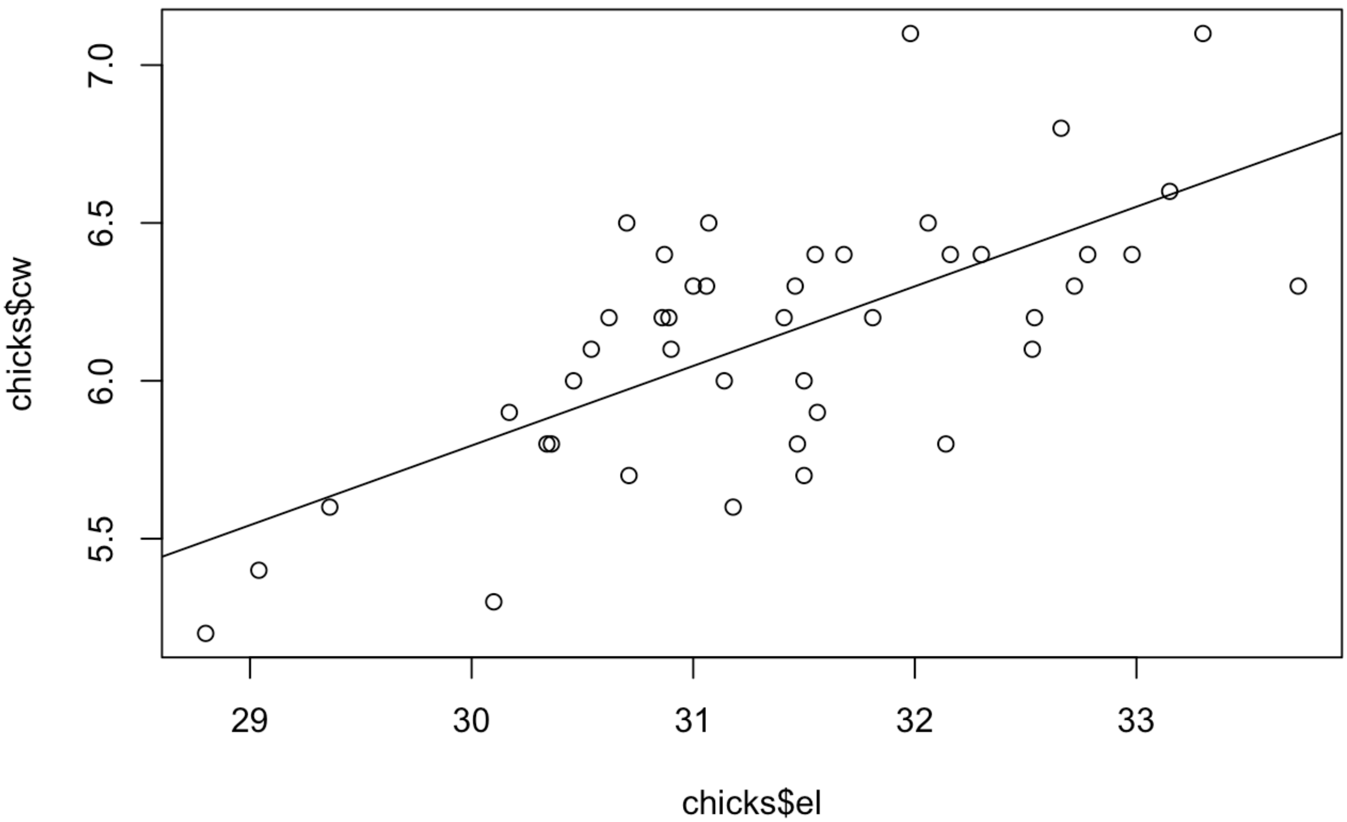
```
# regression with weight of chicks and length of egg
lm(chicks$cw ~ chicks$el)
```

```
##
## Call:
## lm(formula = chicks$cw ~ chicks$el)
##
## Coefficients:
## (Intercept)    chicks$el
##     -1.7702       0.2522
```

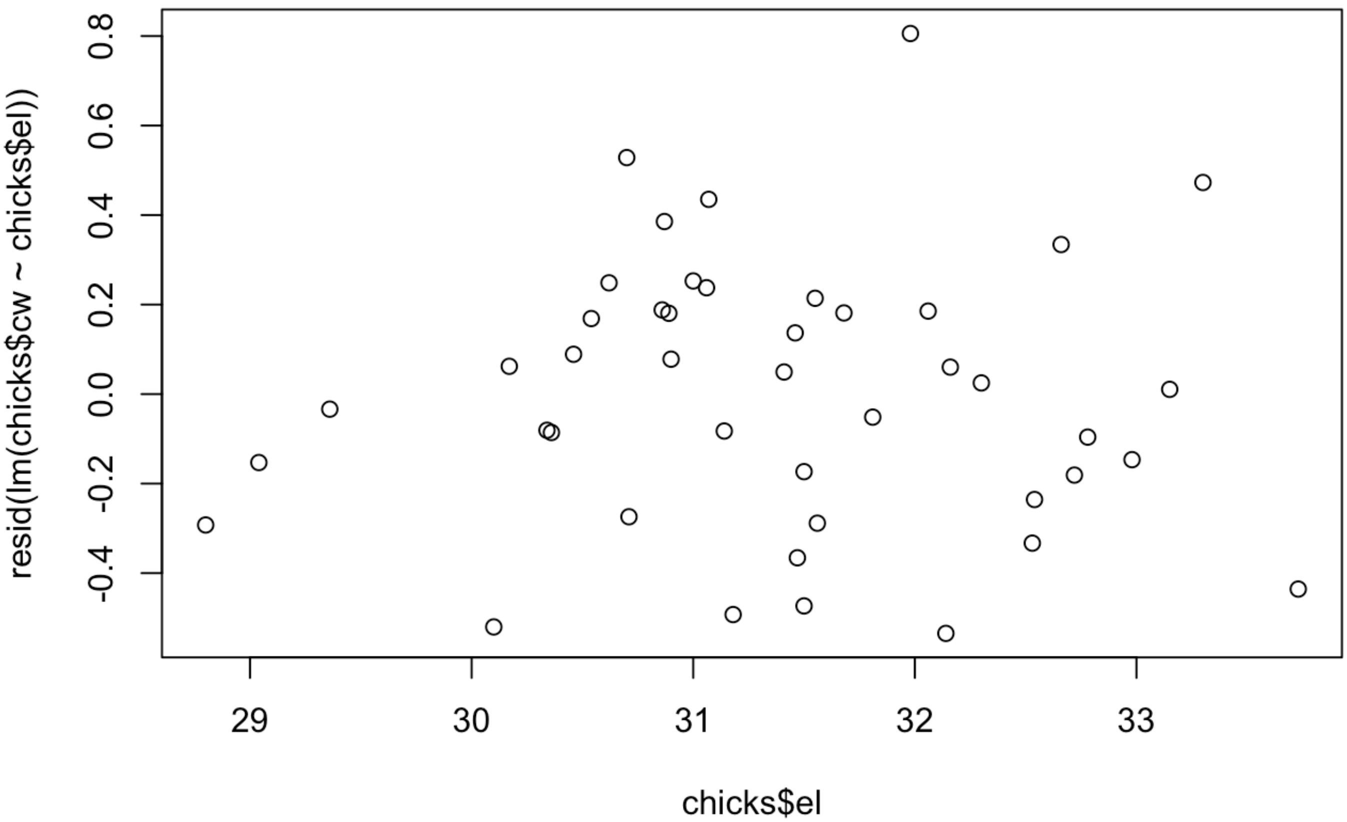
```
summary(lm(chicks$cw ~ chicks$el))
```

```
##
## Call:
## lm(formula = chicks$cw ~ chicks$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
## chicks$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
```

```
plot(chicks$el, chicks$cw)
abline(lm(chicks$cw~chicks$el))
```




```
plot(chicks$el, resid(lm(chicks$cw~chicks$el)))
```



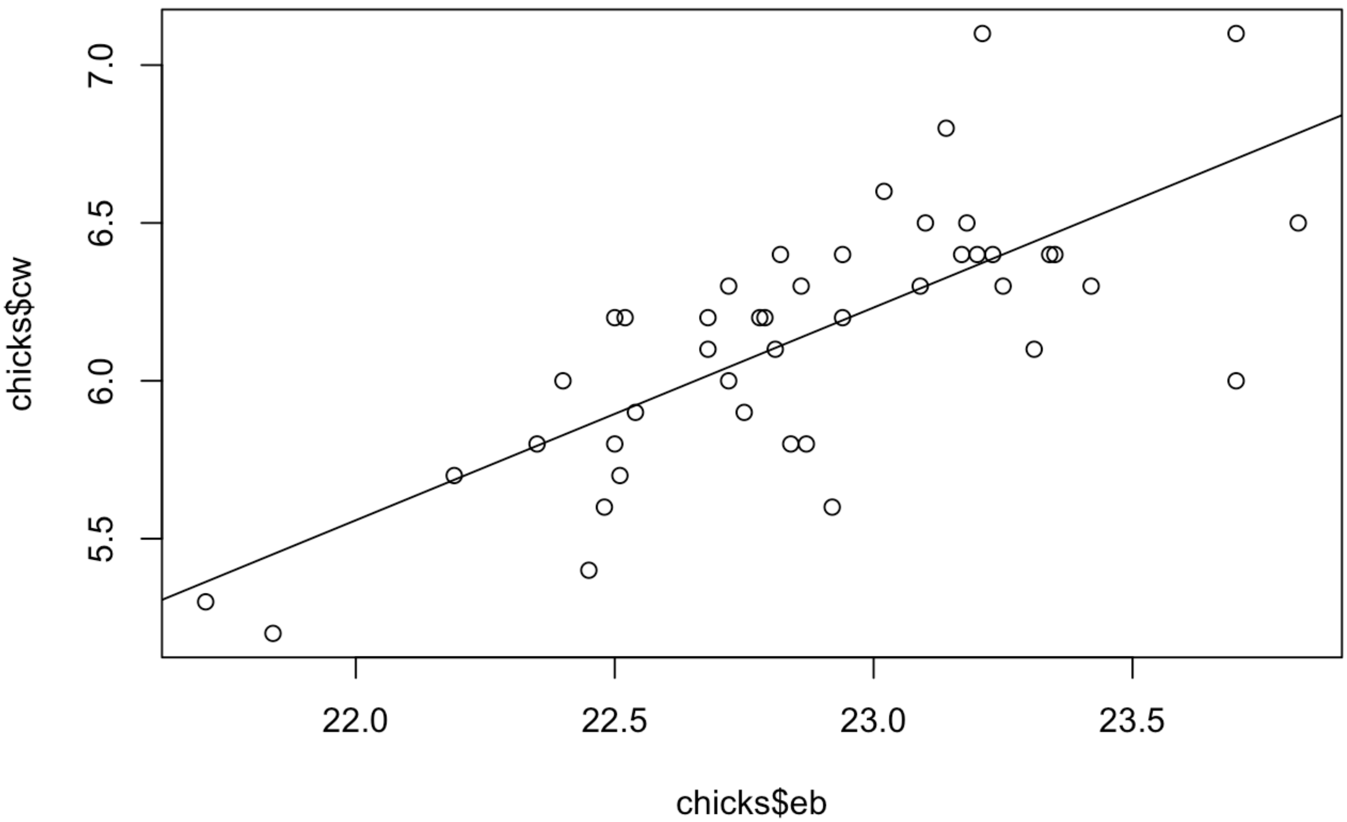

```
# regression with weight of chicks and breadth of egg
lm(chicks$cw ~ chicks$eb)
```

```
##
## Call:
## lm(formula = chicks$cw ~ chicks$eb)
##
## Coefficients:
## (Intercept)    chicks$eb
##     -9.2626       0.6737
```

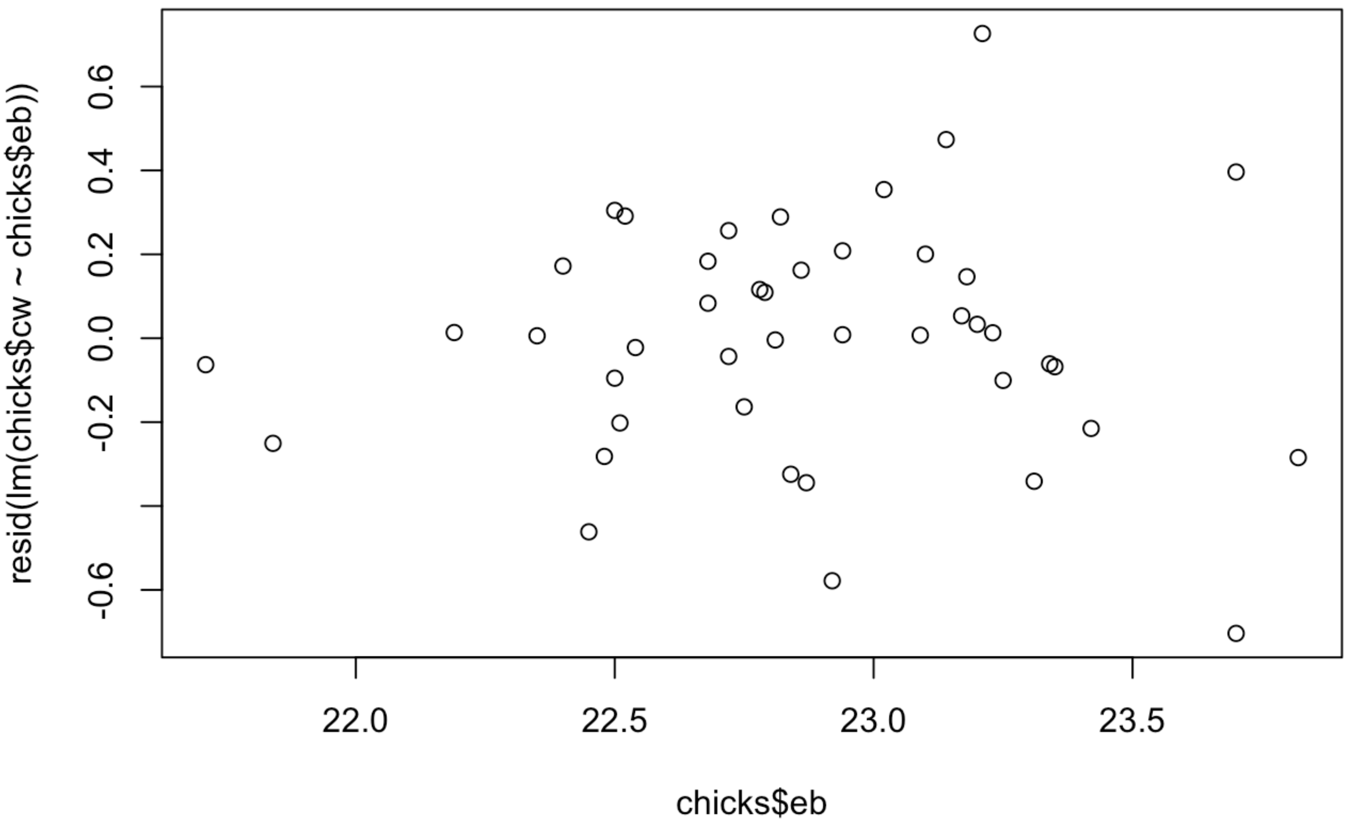
```
summary(lm(chicks$eb ~ chicks$cw))
```

```
##
## Call:
## lm(formula = chicks$eb ~ chicks$cw)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.48582 -0.23754 -0.03009  0.16000  0.94486
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  17.9609     0.7033  25.540  < 2e-16 ***
## chicks$cw     0.7990     0.1142   6.998 1.46e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3074 on 42 degrees of freedom
## Multiple R-squared:  0.5383, Adjusted R-squared:  0.5273
## F-statistic: 48.97 on 1 and 42 DF,  p-value: 1.465e-08
```

```
plot(chicks$eb, chicks$cw)
abline(lm(chicks$cw~chicks$eb))
```




```
plot(chicks$eb, resid(lm(chicks$cw~chicks$eb)))
```




```
# Both regressions are similar, and both have slight problems with the homoscedastici
ty assumptions. One is not noticably better than the other.
```

Problem 10B: part b)

```
lm(chicks$ew ~ chicks$el + chicks$eb)
```

```
##
## Call:
## lm(formula = chicks$ew ~ chicks$el + chicks$eb)
##
## Coefficients:
## (Intercept)    chicks$el    chicks$eb
##    -14.2220       0.2386       0.6719
```

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

```
##
## Call:
## lm(formula = chicks$ew ~ chicks$el + chicks$eb)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.231315 -0.076288 -0.004403  0.054513  0.273872
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -14.22199    0.87175  -16.31   <2e-16 ***
## chicks$el     0.23858    0.01667   14.31   <2e-16 ***
## chicks$eb     0.67190    0.04105   16.37   <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
```

```
# The R-squared value is 0.95, showing an almost perfectly linear relationship betwee
n egg length and egg breadth, which explains why the two regressions in part a are ve
ry similar.
```

Problem 10B: part c)

```
lm(chicks$cw ~ chicks$el + chicks$eb + chicks$ew)
```

```
##
## Call:
## lm(formula = chicks$cw ~ chicks$el + chicks$eb + chicks$ew)
##
## Coefficients:
## (Intercept)    chicks$el    chicks$eb    chicks$ew
##    -4.60567      0.06657      0.21591      0.43123
```

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

```
##
## Call:
## lm(formula = chicks$cw ~ chicks$el + chicks$eb + chicks$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
## chicks$el    0.06657    0.08286   0.803    0.426
## chicks$eb    0.21591    0.22872   0.944    0.351
## chicks$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
```

```
# The F-test suggests the slope is not 0, while the three t-tests suggests the slope
is in fact 0. Since the predicator values are very correlated with each other, the in
dividual slopes have no meaning. And also the R-squared value is a little higher, the
adjusted R-squared value is a little lower. In conclusion, linear regression should n
ot be computed on predictors that are highly correlated with each other.
```

Problem 10B: part d)

```
lm(chicks$cw ~ chicks$el + chicks$ew)
```

```
##
## Call:
## lm(formula = chicks$cw ~ chicks$el + chicks$ew)
##
## Coefficients:
## (Intercept)    chicks$el    chicks$ew
##   -0.133773     0.004769     0.709922
```

```
summary(lm(chicks$cw ~ chicks$el + chicks$ew))
```

```
##
## Call:
## lm(formula = chicks$cw ~ chicks$el + chicks$ew)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.53781 -0.12080 -0.00854  0.12614  0.62097
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.133773   1.007462  -0.133    0.895
## chicks$el    0.004769   0.050725   0.094    0.926
## chicks$ew    0.709922   0.115344   6.155 2.61e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2233 on 41 degrees of freedom
## Multiple R-squared:  0.7179, Adjusted R-squared:  0.7041
## F-statistic: 52.16 on 2 and 41 DF,  p-value: 5.428e-12
```

```
lm(chicks$cw ~ chicks$eb + chicks$ew)
```

```
##
## Call:
## lm(formula = chicks$cw ~ chicks$eb + chicks$ew)
##
## Coefficients:
## (Intercept)    chicks$eb    chicks$ew
##    -1.20273      0.07073      0.66370
```

```
summary(lm(chicks$eb ~ chicks$cw + chicks$ew))
```

```
##
## Call:
## lm(formula = chicks$eb ~ chicks$cw + chicks$ew)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.49425 -0.15734 -0.03488  0.12517  0.65217
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.18521    0.67653  23.924  < 2e-16 ***
## chicks$cw    0.08798    0.17363   0.507    0.615
## chicks$ew    0.71177    0.14725   4.834 1.91e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2483 on 41 degrees of freedom
## Multiple R-squared:  0.7059, Adjusted R-squared:  0.6915
## F-statistic:  49.2 on 2 and 41 DF,  p-value: 1.272e-11
```

```
lm(chicks$cw ~ chicks$el + chicks$eb)
```

```
##
## Call:
## lm(formula = chicks$cw ~ chicks$el + chicks$eb)
##
## Coefficients:
## (Intercept)    chicks$el    chicks$eb
##    -10.7386       0.1695       0.5057
```

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

```
##
## Call:
## lm(formula = chicks$cw ~ chicks$el + chicks$eb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.53454 -0.12055  0.01582  0.10292  0.68326
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -10.73860    1.78777  -6.007 4.23e-07 ***
## chicks$el     0.16945    0.03420   4.955 1.29e-05 ***
## chicks$eb     0.50566    0.08419   6.006 4.24e-07 ***
## ---
## 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
```

```
# Egg length and egg breadth are the best predictors. Using egg weight in combination
with other variables lead to an R-squared value of around .70.
```

Problem 10C: part a)

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

# parametric test
# Null: The means are the same.
# Alternate: The means are different.
t.test(tox$month15 - tox$base) # The t value is -6.1549, and the p-value is close to
zero, so we conclude that the means are different, and since the t-value is different
, the month15 values are on average less than the base values.
```

```
##
##  One Sample t-test
##
## data:  tox$month15 - tox$base
## t = -6.1549, df = 21, p-value = 4.167e-06
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -38.28457 -18.94725
## sample estimates:
## mean of x
## -28.61591
```

```
# non-parametric test
# Null: The underlying distributions are the same.
# Alternate: the underlying distributions are different.
wilcox.test(tox$base, tox$month15) # The value of the statistic is 398, and the p-val
ue is close to 0. We conclude that the underlying distributions are different, which
is consistent with the parametric test.
```

```
##
##  Wilcoxon rank sum test
##
## data:  tox$base and tox$month15
## W = 398, p-value = 0.0001396
## alternative hypothesis: true location shift is not equal to 0
```

Problem 10C: part b)

```
lm(tox$month15 ~ tox$height)
```

```
##
## Call:
## lm(formula = tox$month15 ~ tox$height)
##
## Coefficients:
## (Intercept)   tox$height
##    -88.6966       0.9963
```

```
summary(lm(tox$month15 ~ tox$height)) # r-squared = .15
```

```
##
## Call:
## lm(formula = tox$month15 ~ tox$height)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.697  -8.587  -0.273  11.826  49.673
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -88.6966    90.5847  -0.979   0.3392
## tox$height    0.9963     0.5251   1.897   0.0723 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.3 on 20 degrees of freedom
## Multiple R-squared:  0.1525, Adjusted R-squared:  0.1101
## F-statistic: 3.599 on 1 and 20 DF,  p-value: 0.07234
```

```
lm(tox$month15 ~ tox$rad)
```

```
##
## Call:
## lm(formula = tox$month15 ~ tox$rad)
##
## Coefficients:
## (Intercept)      tox$rad
##     80.4083       0.0064
```

```
summary(lm(tox$month15 ~ tox$rad)) # r-squared = .0017
```

```
##
## Call:
## lm(formula = tox$month15 ~ tox$rad)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.509 -13.062  -1.559   8.517  50.438
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 80.40831   14.82278   5.425 2.61e-05 ***
## tox$rad      0.00640    0.03521   0.182    0.858
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.95 on 20 degrees of freedom
## Multiple R-squared:  0.00165,    Adjusted R-squared:  -0.04827
## F-statistic: 0.03305 on 1 and 20 DF,  p-value: 0.8576
```

```
lm(tox$month15 ~ tox$chemo)
```

```
##
## Call:
## lm(formula = tox$month15 ~ tox$chemo)
##
## Coefficients:
## (Intercept)    tox$chemo
##     44.6182       0.2051
```

```
summary(lm(tox$month15 ~ tox$chemo)) # r-squared = .19
```

```
##
## Call:
## lm(formula = tox$month15 ~ tox$chemo)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.280 -12.953   3.406  11.773  39.697
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 44.61825   17.68205   2.523   0.0202 *
## tox$chemo    0.20508    0.09208   2.227   0.0376 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.77 on 20 degrees of freedom
## Multiple R-squared:  0.1987, Adjusted R-squared:  0.1587
## F-statistic:  4.96 on 1 and 20 DF,  p-value: 0.03758
```

```
lm(tox$month15 ~ tox$base)
```

```
##
## Call:
## lm(formula = tox$month15 ~ tox$base)
##
## Coefficients:
## (Intercept)     tox$base
##     32.1721       0.4553
```

```
summary(lm(tox$month15 ~ tox$base)) # r-squared = .31
```

```
##
## Call:
## lm(formula = tox$month15 ~ tox$base)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.910 -13.949  -0.343   9.690  42.409
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  32.1721    17.1511   1.876  0.07536 .
## tox$base      0.4553     0.1501   3.034  0.00656 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.35 on 20 degrees of freedom
## Multiple R-squared:  0.3151, Adjusted R-squared:  0.2809
## F-statistic: 9.203 on 1 and 20 DF,  p-value: 0.006559
```

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

```
##
## Call:
## lm(formula = tox$month15 ~ tox$base + tox$height + tox$chemo)
##
## Coefficients:
## (Intercept)     tox$base   tox$height    tox$chemo
##     22.8397       0.4515      -0.1677       0.2066
```

```
summary(lm(tox$month15 ~ tox$base + tox$height + tox$chemo)) # adjusted r-squared = .
40
```

```
##
## Call:
## lm(formula = tox$month15 ~ tox$base + tox$height + tox$chemo)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.045  -7.907  -1.643   8.358  31.856
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 22.83968   84.02626   0.272   0.7889
## tox$base     0.45150    0.14900   3.030   0.0072 **
## tox$height  -0.16765    0.57266  -0.293   0.7731
## tox$chemo    0.20659    0.09676   2.135   0.0468 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.83 on 18 degrees of freedom
## Multiple R-squared:  0.4871, Adjusted R-squared:  0.4016
## F-statistic: 5.697 on 3 and 18 DF,  p-value: 0.006351
```

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

```
##
## Call:
## lm(formula = tox$month15 ~ tox$base + tox$chemo)
##
## Coefficients:
## (Intercept)     tox$base    tox$chemo
##     -0.9992       0.4345       0.1898
```

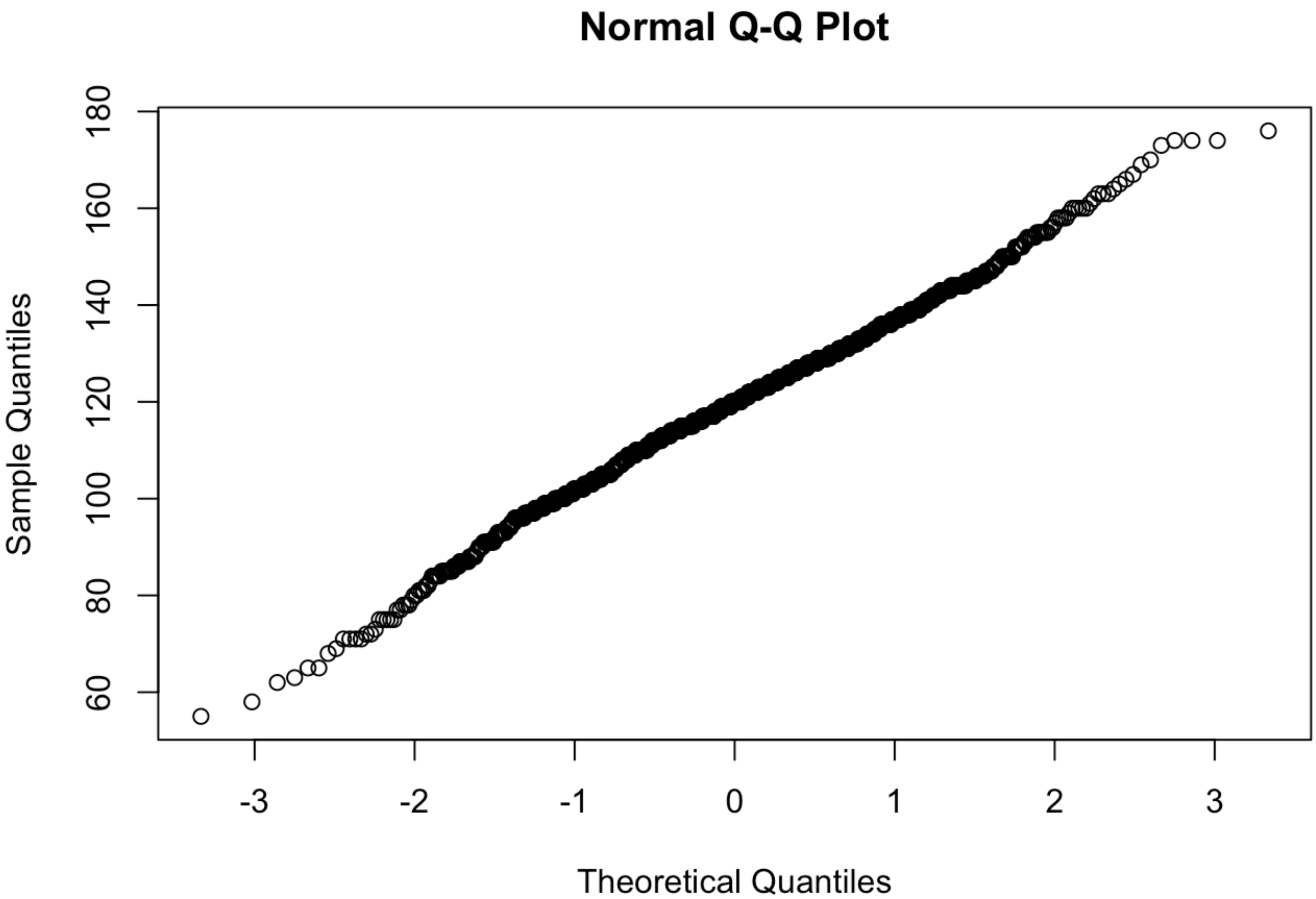
```
summary(lm(tox$month15 ~ tox$base + tox$chemo)) # adjusted r-squared = .43
```

```
##
## 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
```

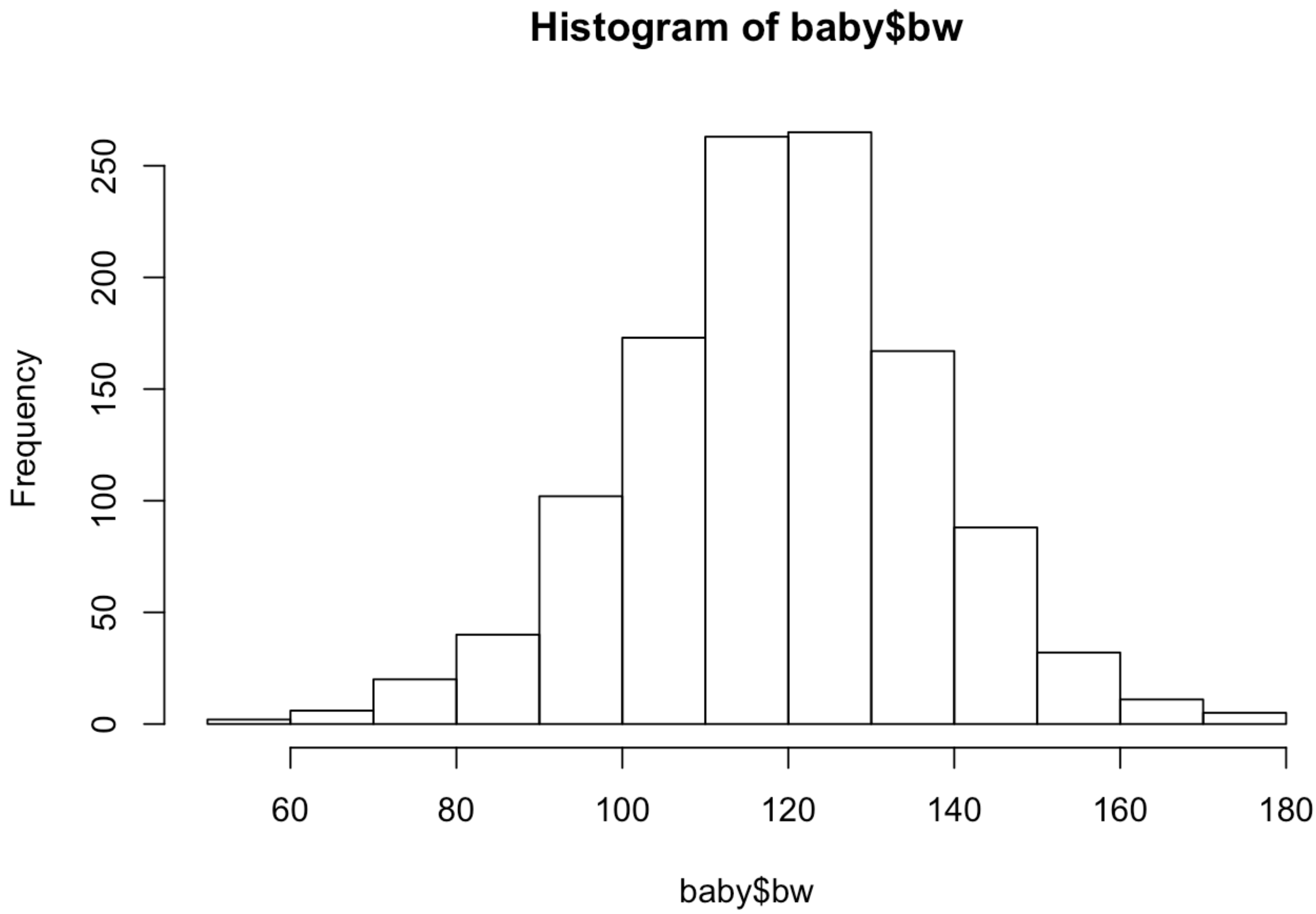
```
# The best predictors are base and chemo, and both slopes are significantly different
from 0.
```

Problem 10D: part a)

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

qqnorm(baby$bw)
```


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
60
80
100
120
140
160
180
-3
-2
-1
0
1
2
3

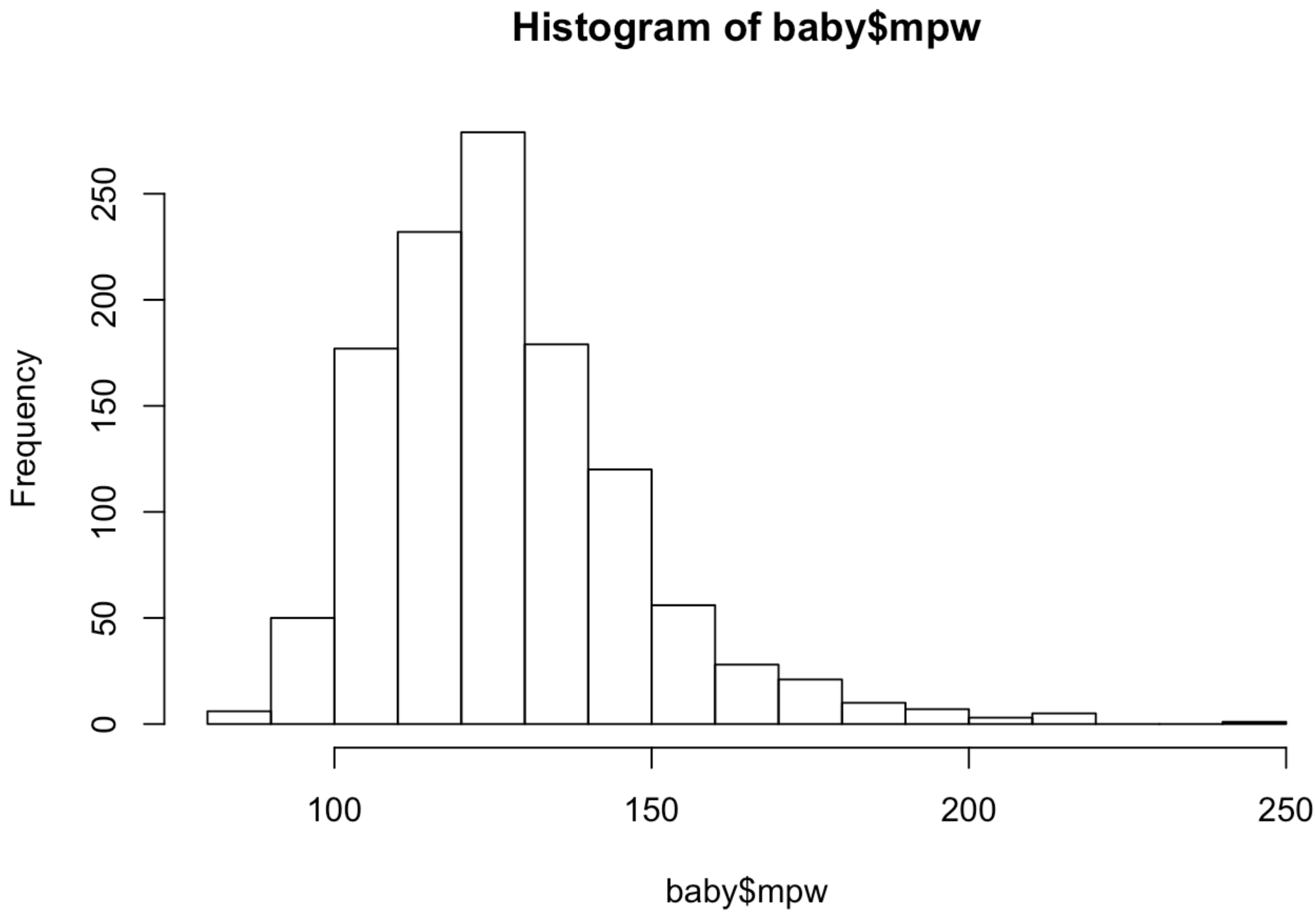
```
hist(baby$bw)
```



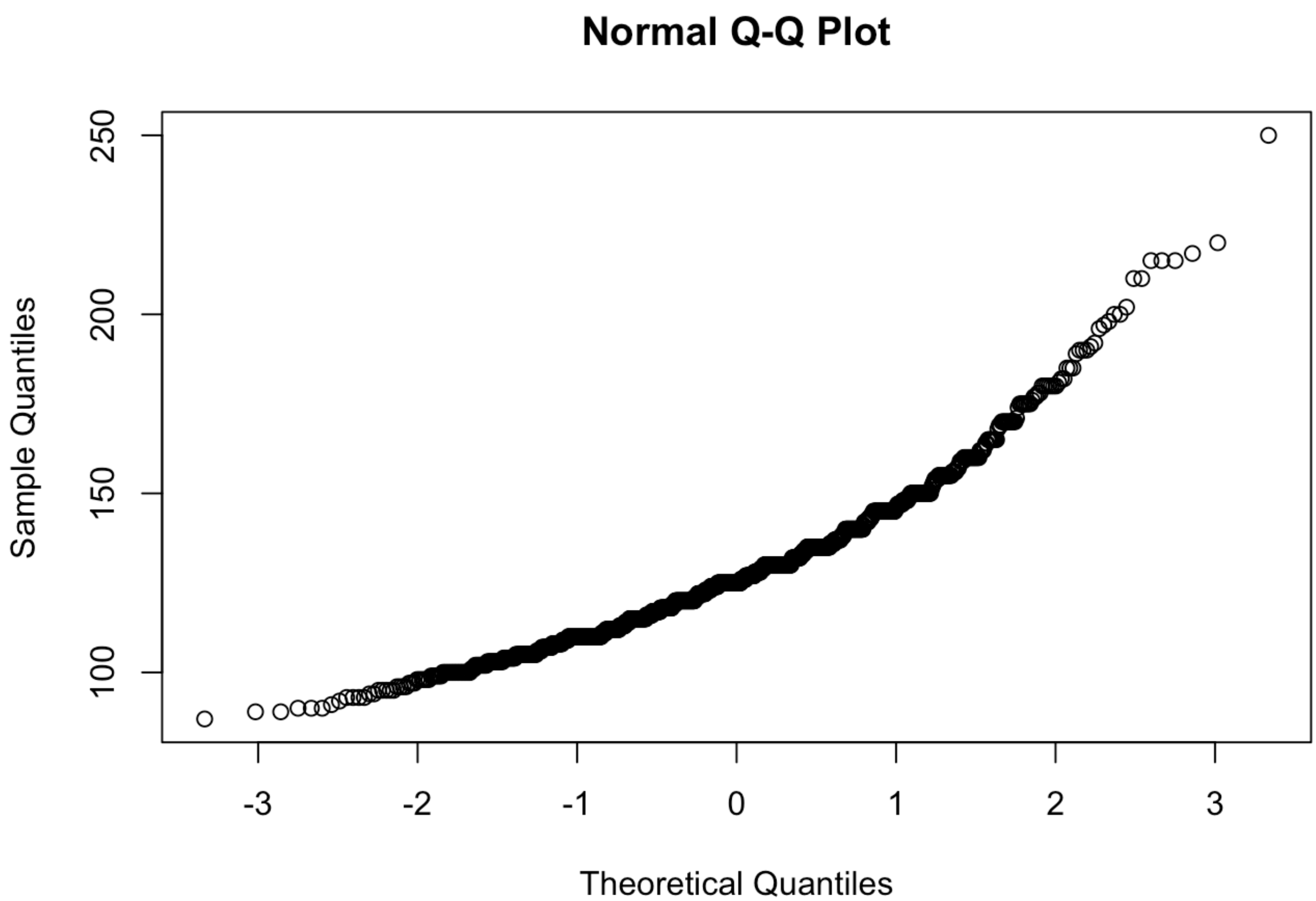

```
# Both plots show normal pattern data. The historgram shows symmetry, and the scatter
plot shows linear pattern.
```

Problem 10D: part b)

```
hist(baby$mpw)
```




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




```
# The ditribtuion is slightly skewed to the right. If the histogram was skewed the ot
her way, the shape of the qqnorm plot would be concave instead of convex.
```

Problem 10D: part c)

```
summary(lm(baby$bw ~ baby$gd)) # r-squared = .16
```

```
##
## Call:
## lm(formula = baby$bw ~ baby$gd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.348 -11.065   0.218  10.101  57.704
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -10.75414    8.53693   -1.26    0.208
## baby$gd       0.46656    0.03054   15.28   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.74 on 1172 degrees of freedom
## Multiple R-squared:  0.1661, Adjusted R-squared:  0.1654
## F-statistic: 233.4 on 1 and 1172 DF,  p-value: < 2.2e-16
```

```
summary(lm(baby$bw ~ baby$ma)) # r-squared close to 0
```

```
##
## Call:
## lm(formula = baby$bw ~ baby$ma)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -65.123 -11.172   0.387  11.472  57.237
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 117.14791    2.56126  45.738   <2e-16 ***
## baby$ma       0.08501    0.09199   0.924    0.356
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.33 on 1172 degrees of freedom
## Multiple R-squared:  0.0007281,  Adjusted R-squared:  -0.0001245
## F-statistic: 0.8539 on 1 and 1172 DF,  p-value: 0.3556
```

```
summary(lm(baby$bw ~ baby$mh)) # r-squared = .04
```

```
##
## Call:
## lm(formula = baby$bw ~ baby$mh)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -65.868 -10.433   0.654  11.436  59.045
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  24.7963    13.3004   1.864   0.0625 .
## baby$mh       1.4780     0.2075   7.123 1.84e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.95 on 1172 degrees of freedom
## Multiple R-squared:  0.0415, Adjusted R-squared:  0.04068
## F-statistic: 50.74 on 1 and 1172 DF,  p-value: 1.838e-12
```

```
summary(lm(baby$bw ~ baby$mpw)) # r-squared = .02
```

```
##
## Call:
## lm(formula = baby$bw ~ baby$mpw)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -66.051 -10.916   0.328  11.026  56.084
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 101.75393    3.31927  30.655  < 2e-16 ***
## baby$mpw      0.13783    0.02551   5.404 7.89e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.11 on 1172 degrees of freedom
## Multiple R-squared:  0.02431,    Adjusted R-squared:  0.02348
## F-statistic:  29.2 on 1 and 1172 DF,  p-value: 7.887e-08
```

```
summary(lm(baby$bw ~ baby$sm)) # r-squared = .06
```

```
##
## Call:
## lm(formula = baby$bw ~ baby$sm)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -68.085 -11.085   0.915  11.181  52.915
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 123.0853     0.6645 185.221   <2e-16 ***
## baby$sm      -9.2661     1.0628  -8.719   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.77 on 1172 degrees of freedom
## Multiple R-squared:  0.06091,    Adjusted R-squared:  0.06011
## F-statistic: 76.02 on 1 and 1172 DF,  p-value: < 2.2e-16
```

```
summary(lm(baby$bw ~ baby$gd + baby$mh + baby$mpw + baby$sm)) # r-squared = 0.25
```

```
##
## Call:
## lm(formula = baby$bw ~ baby$gd + baby$mh + baby$mpw + baby$sm)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -56.630 -10.387  -0.348   9.794  51.891
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -77.25871   14.05139  -5.498 4.71e-08 ***
## baby$gd       0.43718    0.02909  15.028  < 2e-16 ***
## baby$mh       1.09733    0.20463   5.363 9.88e-08 ***
## baby$mpw      0.05981    0.02491   2.401   0.0165 *
## baby$sm      -8.34833    0.95453  -8.746  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.88 on 1169 degrees of freedom
## Multiple R-squared:  0.2519, Adjusted R-squared:  0.2493
## F-statistic: 98.39 on 4 and 1169 DF,  p-value: < 2.2e-16
```

```
# All the predictors can be used except mother's height, which has almost no correlat
ion to baby's weight. The r-squared value of all the other predictors is .25.
```

Problem 10D: part d)

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

```
##
## Call:
## lm(formula = baby$bw ~ baby$gd + baby$mh + baby$mpw + baby$sm)
##
## Coefficients:
## (Intercept)      baby$gd      baby$mh     baby$mpw      baby$sm
##   -77.25871      0.43718      1.09733      0.05981     -8.34833
```

```
# The coefficient of the indicator variable is -8.35. It represents the average diffe
rence in weight of a baby of a smoker and a nonsmoker. A baby born to a mother who sm
okes on average weigh 8.35 ounces less.
```

Problem 10E: part a)

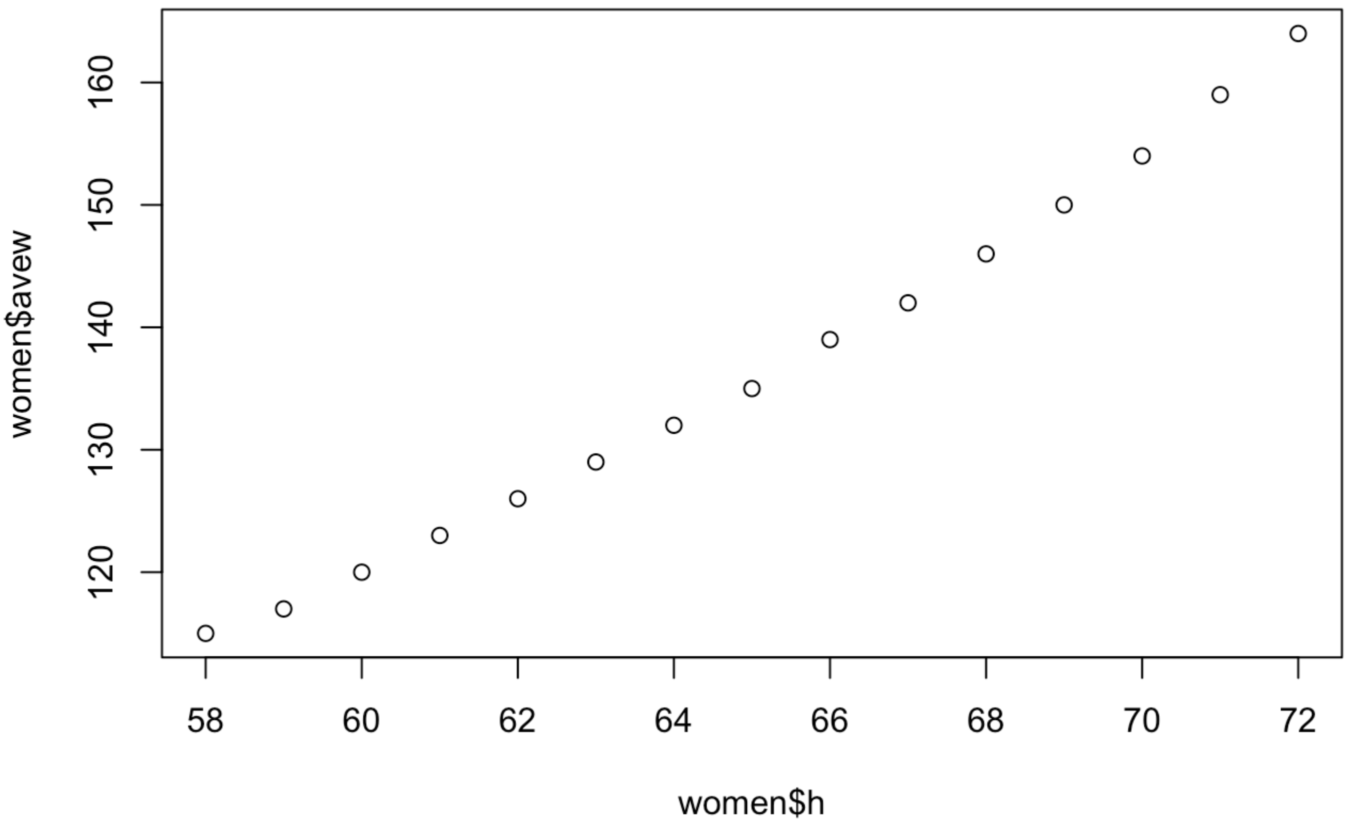
```
women <- read.table("women.txt", header = TRUE)

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

```
##
## Call:
## lm(formula = women$avew ~ women$h)
##
## Coefficients:
## (Intercept)      women$h
##      -87.52         3.45
```

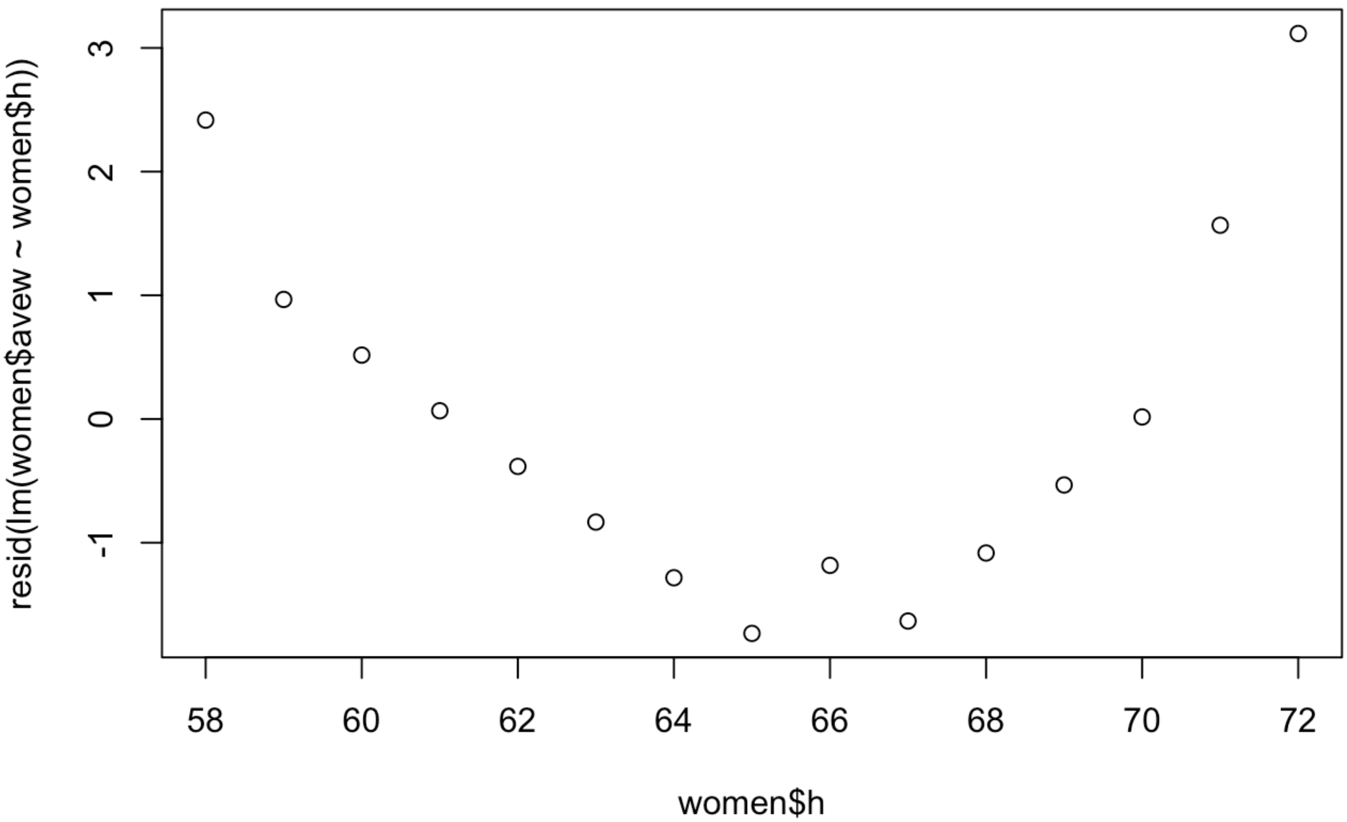
```
summary(lm(women$avew ~ women$h))
```

```
##
## Call:
## lm(formula = women$avew ~ women$h)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7333 -1.1333 -0.3833  0.7417  3.1167
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -87.51667    5.93694  -14.74 1.71e-09 ***
## women$h       3.45000    0.09114   37.85 1.09e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.525 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
```

```
plot(women$h, women$avew)
```




```
plot(women$h, resid(lm(women$avew ~ women$h)))
```



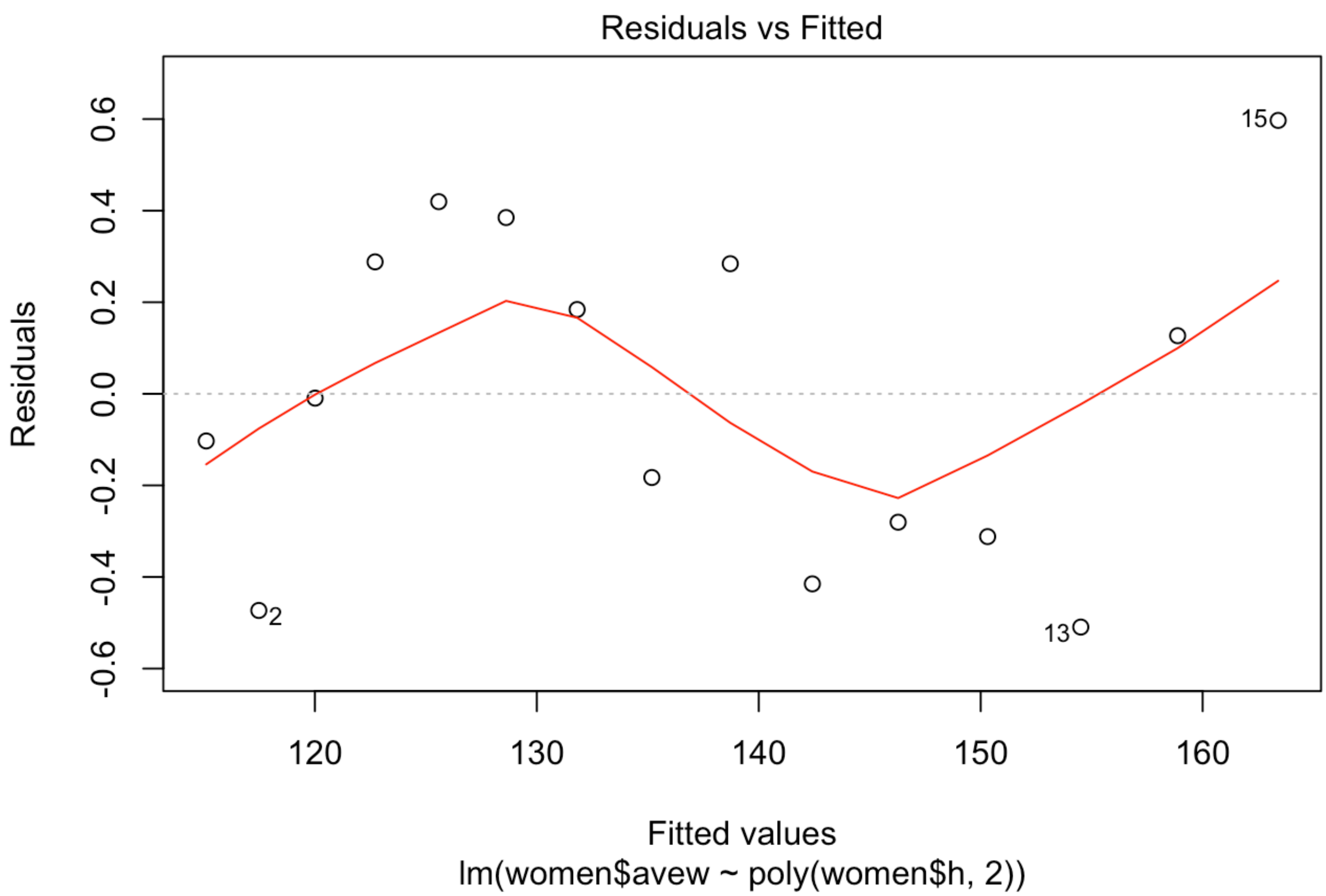

```
# The r-squared = .99. The correlation is high, and the slope is significant. However
, the residual plot shows a strong non-linear pattern, so a straight line is not fit
for this data.
```

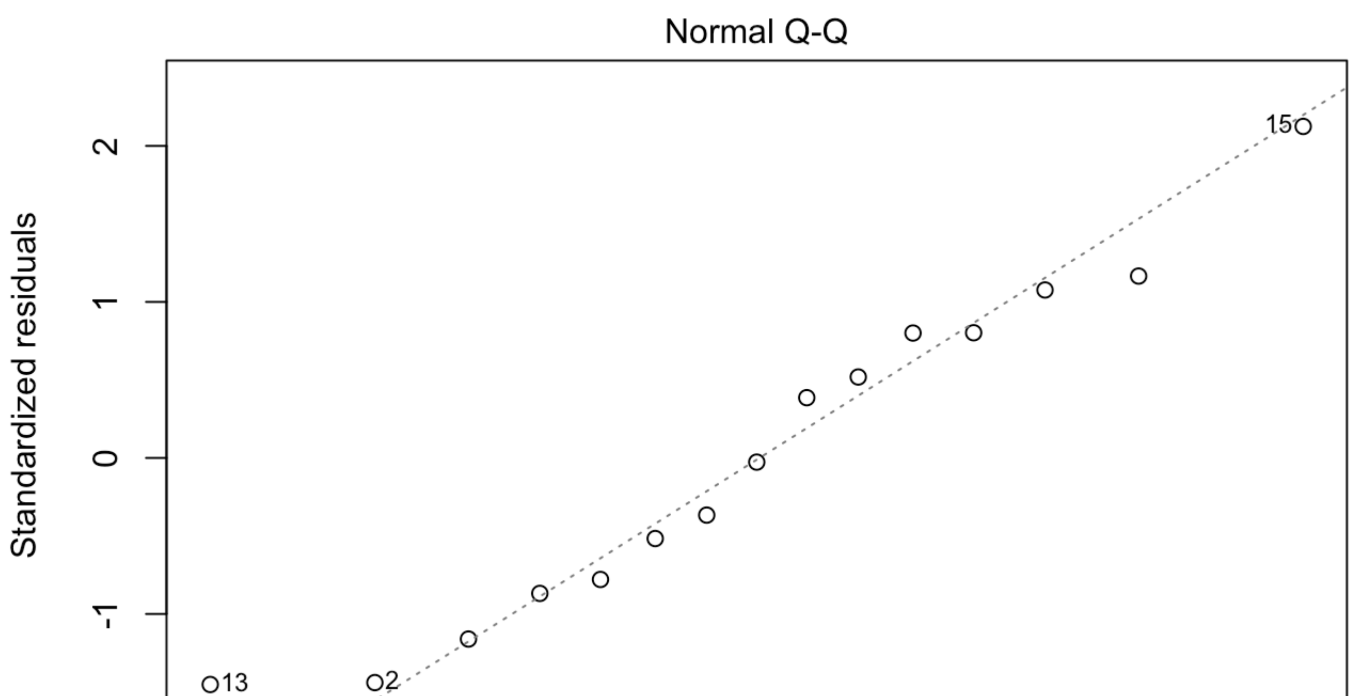
Problem 10E: part b)

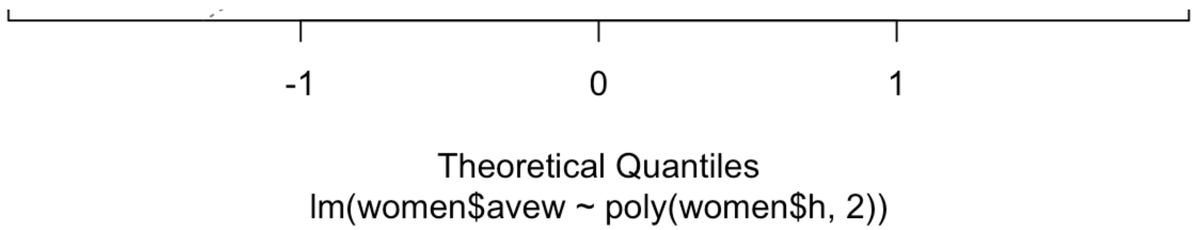
```
# Each point in the dataset women represent many women, so if the points were to be b
roken into their individual women, the data would be more wide-spread instead of summ
arized and condensed and so the correlation will drop.
```

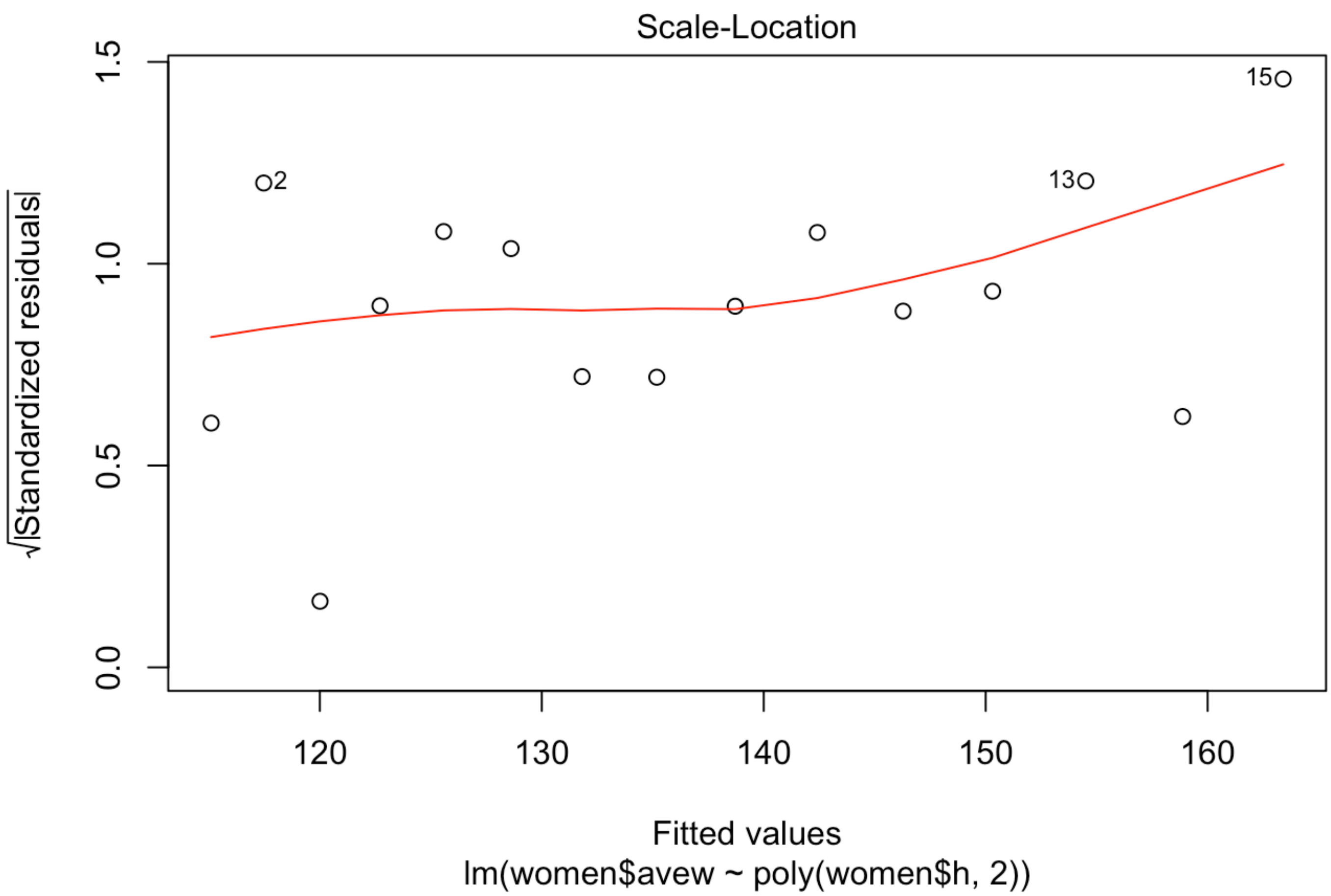
Problem 10E: part c)

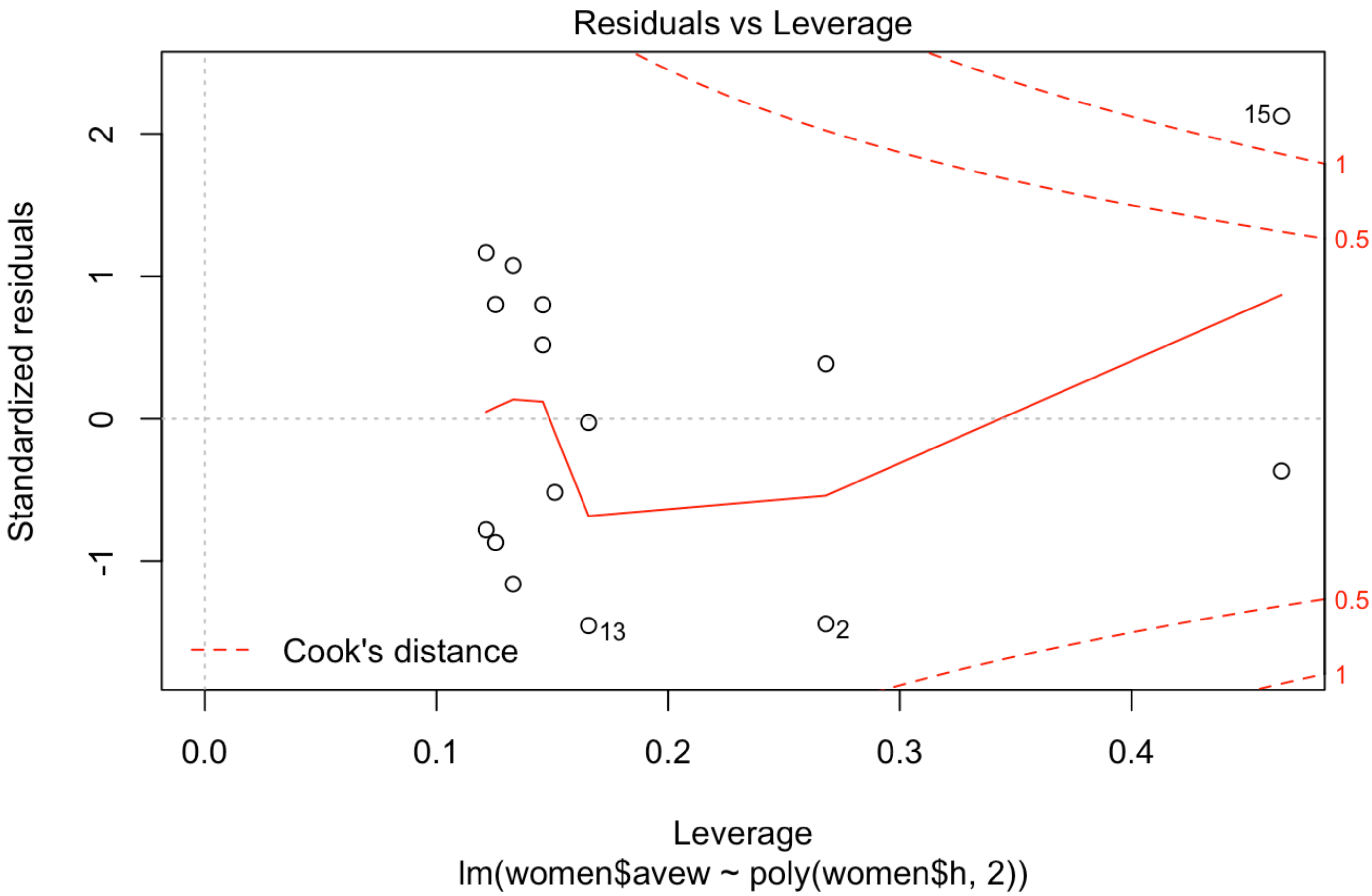
```
plot(lm(women$avew ~ poly(women$h, 2)))
```


Residuals vs Fitted
Residuals
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
15
2
13
120
130
140
150
160
Fitted values
lm(women$avew ~ poly(women$h, 2))

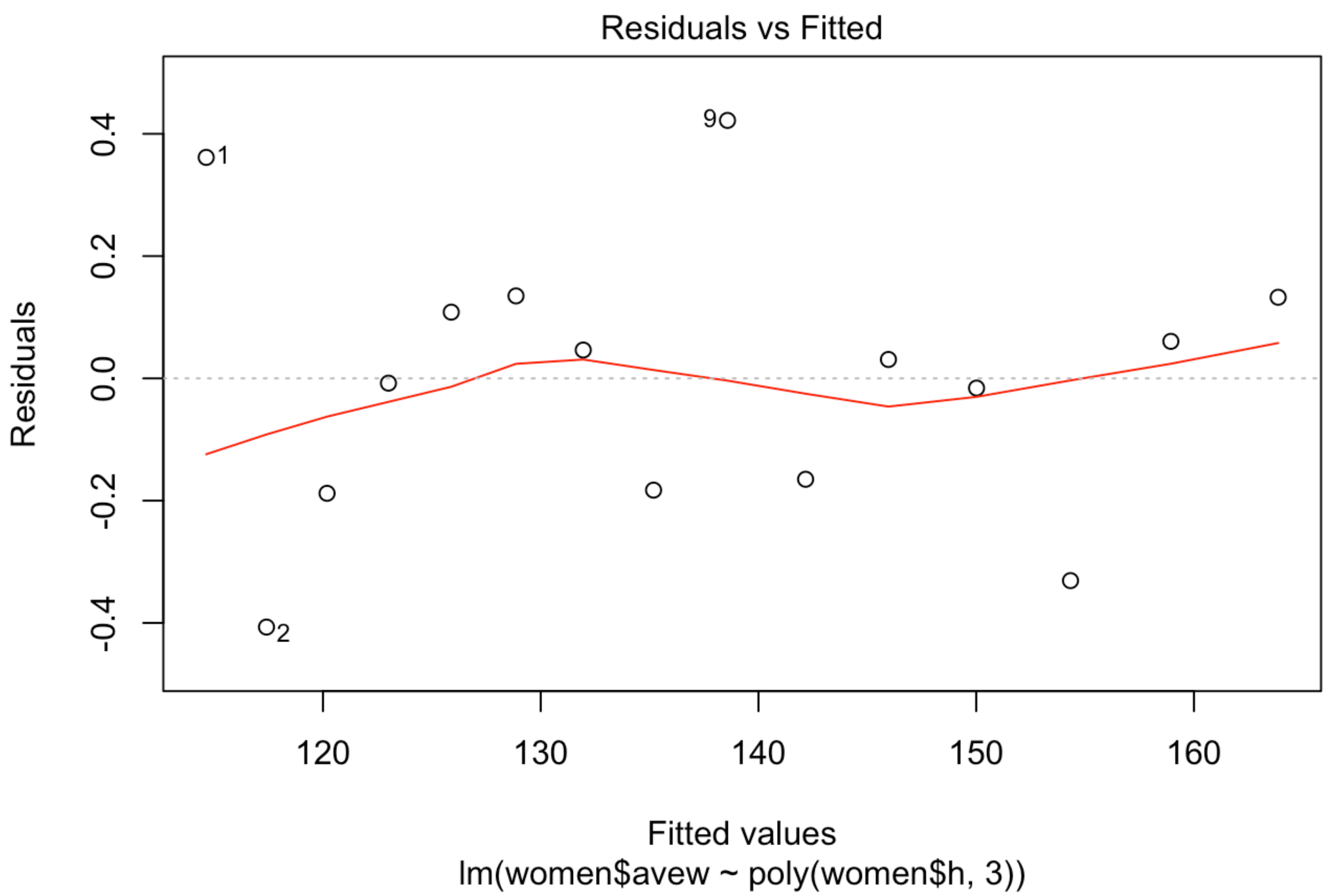

Normal Q-Q
Standardized residuals
2
1
0
-1
15
13
2

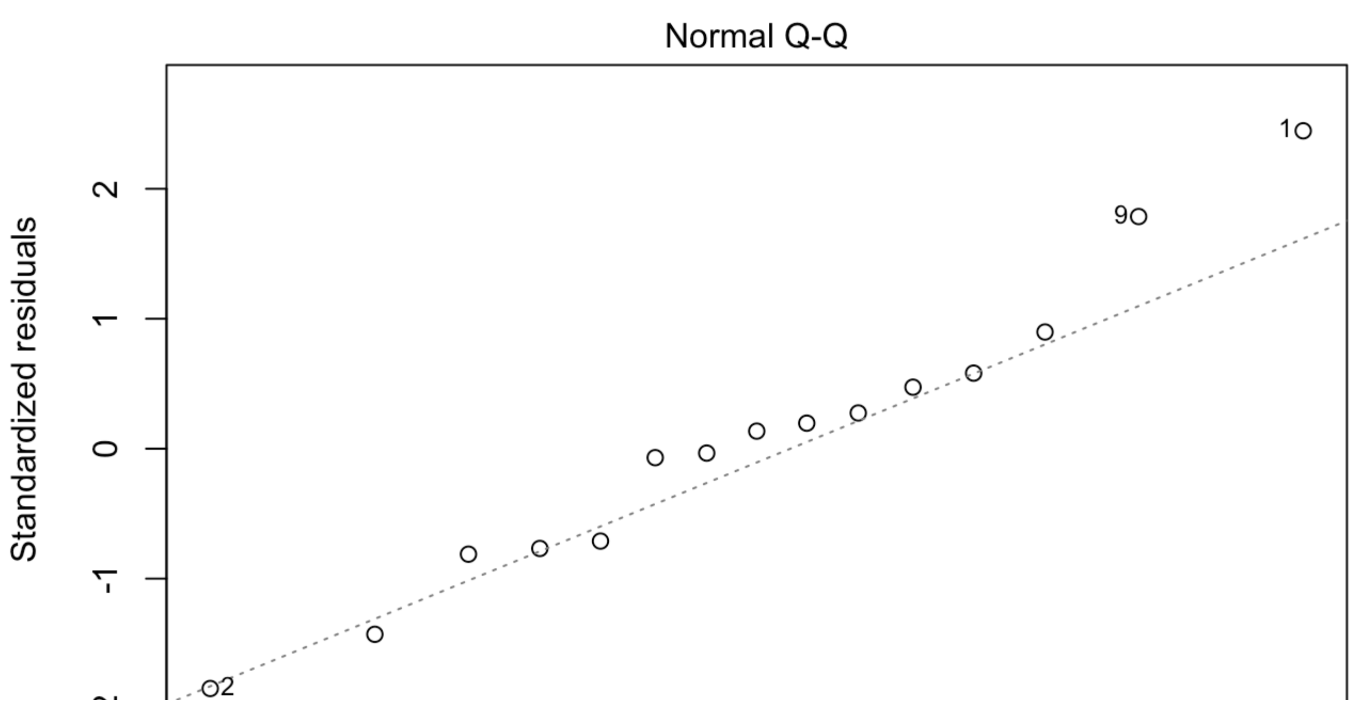

-1
0
1
Theoretical Quantiles
lm(women$avew ~ poly(women$h, 2))

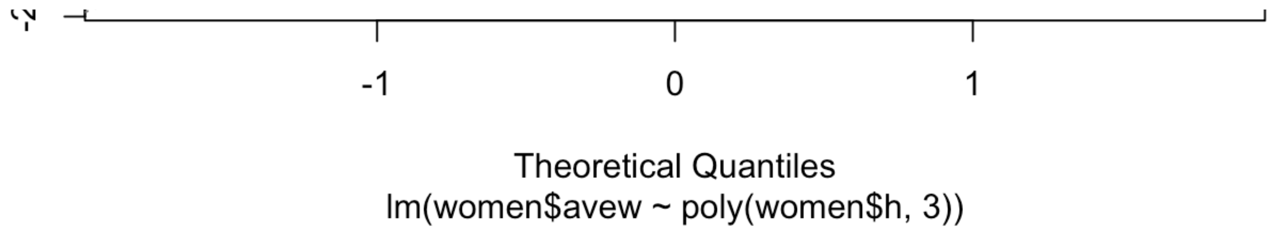

Scale-Location
1.5
1.0
0.5
0.0
√|Standardized residuals|
2
13
15
120
130
140
150
160
Fitted values
lm(women$avew ~ poly(women$h, 2))

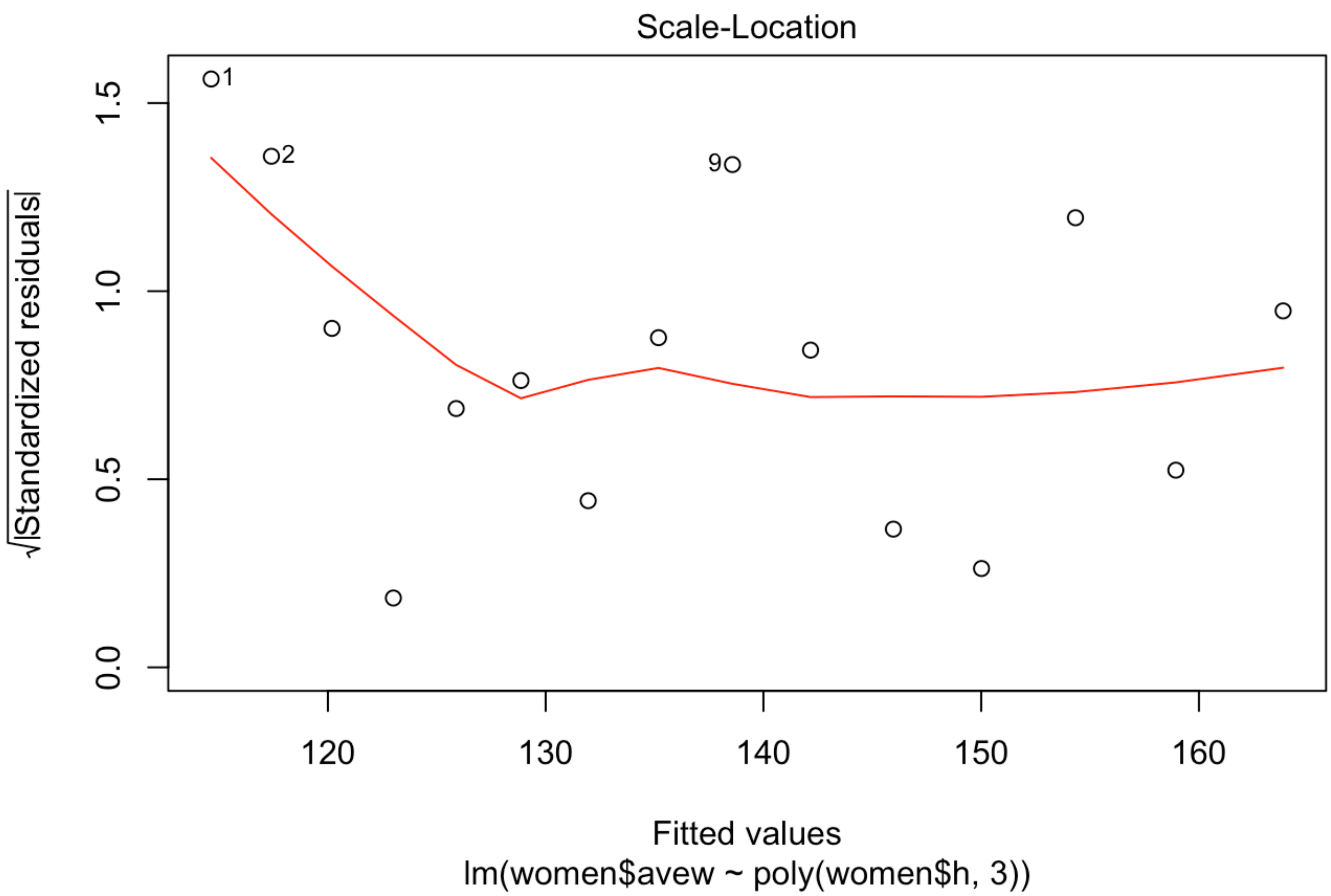


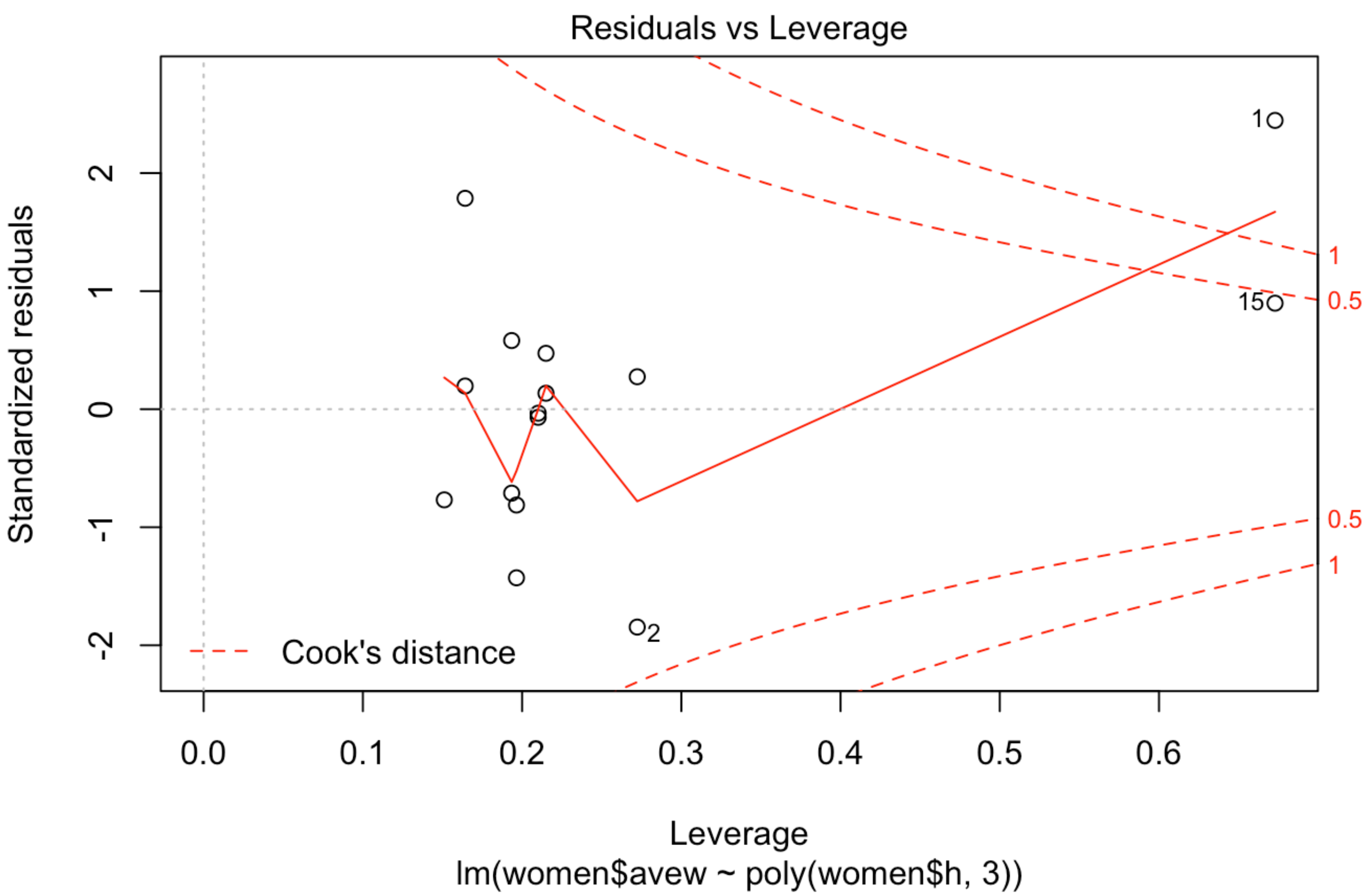

```
plot(lm(women$avew ~ poly(women$h, 3)))
```


Residuals vs Fitted
Residuals
0.4
0.2
0.0
-0.2
-0.4
120
130
140
150
160
Fitted values
lm(women$avew ~ poly(women$h, 3))
1
2
9


Normal Q-Q
Standardized residuals
2
1
0
-1
1
9
2


-2
-1
0
1
Theoretical Quantiles
lm(women$avew ~ poly(women$h, 3))

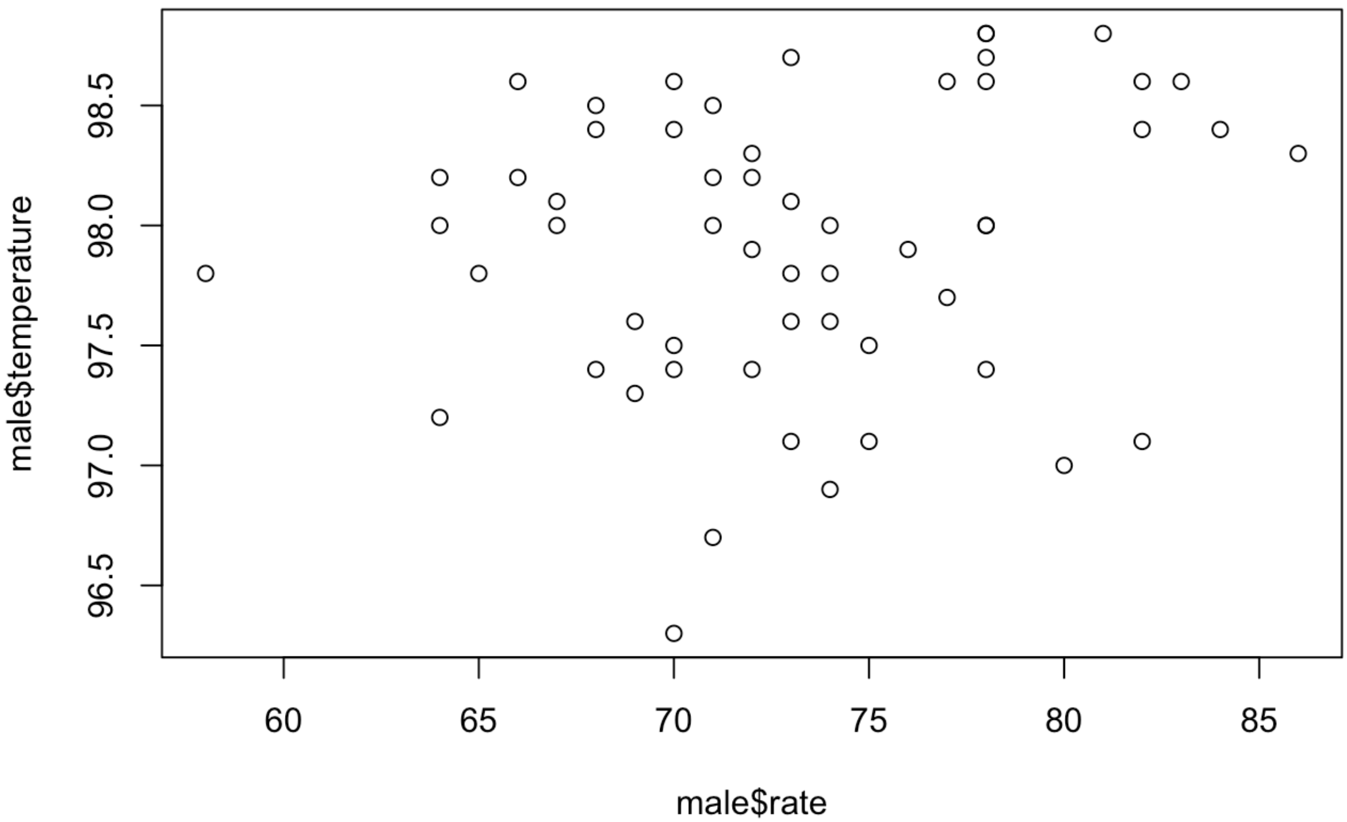

Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
1
2
9
120
130
140
150
160
Fitted values
lm(women$avew ~ poly(women$h, 3))



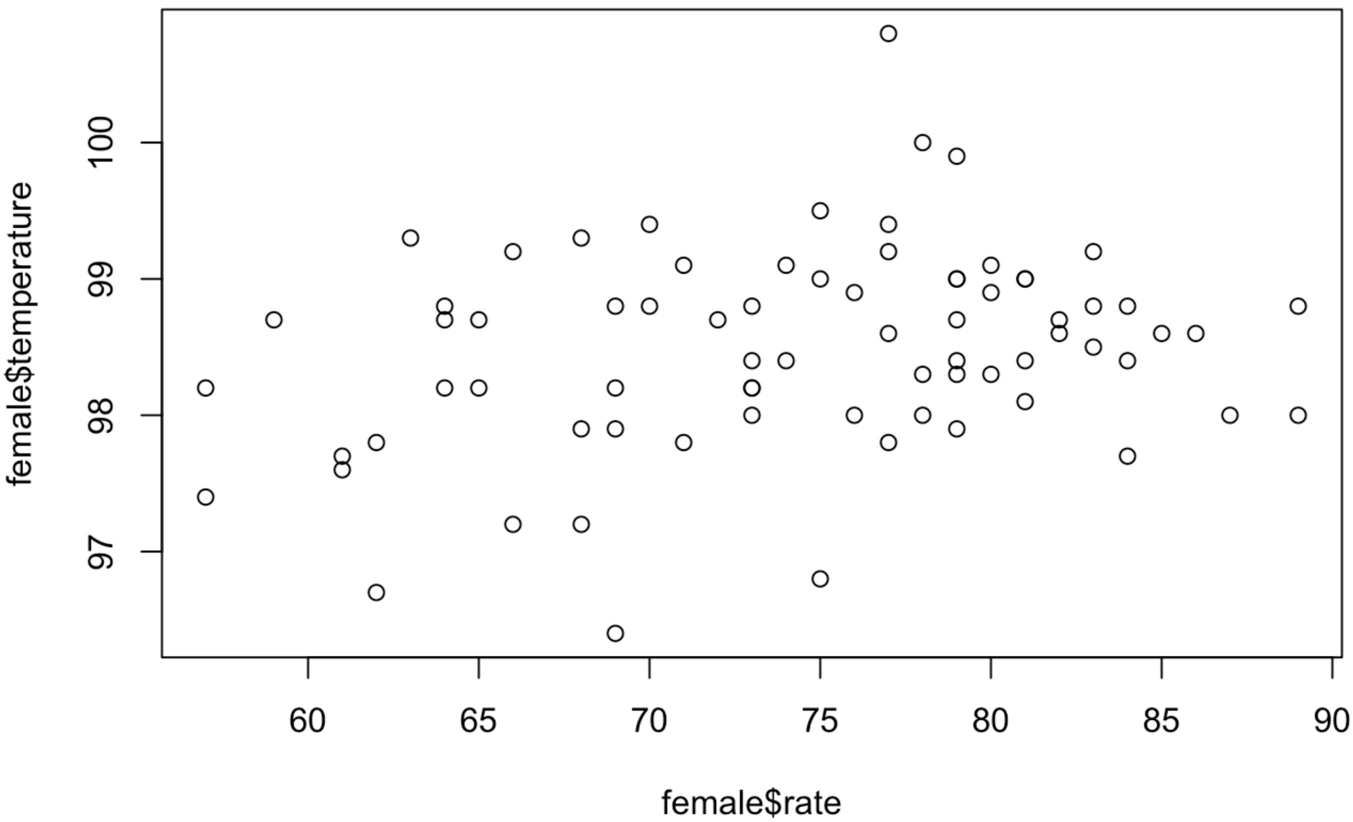

```
# A polynomial of degree 2 looks like a good fit, however the residual plots show tha
t it's not, since there is a pattern. A polynomial of degree 3 is a better fit, as ca
n be seen through the residual plot.
```

Problem 10F: part a)

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

male <- bodytemp[1:56,]
female <- bodytemp[57:130,]

plot(male$rate, male$temperature)
```



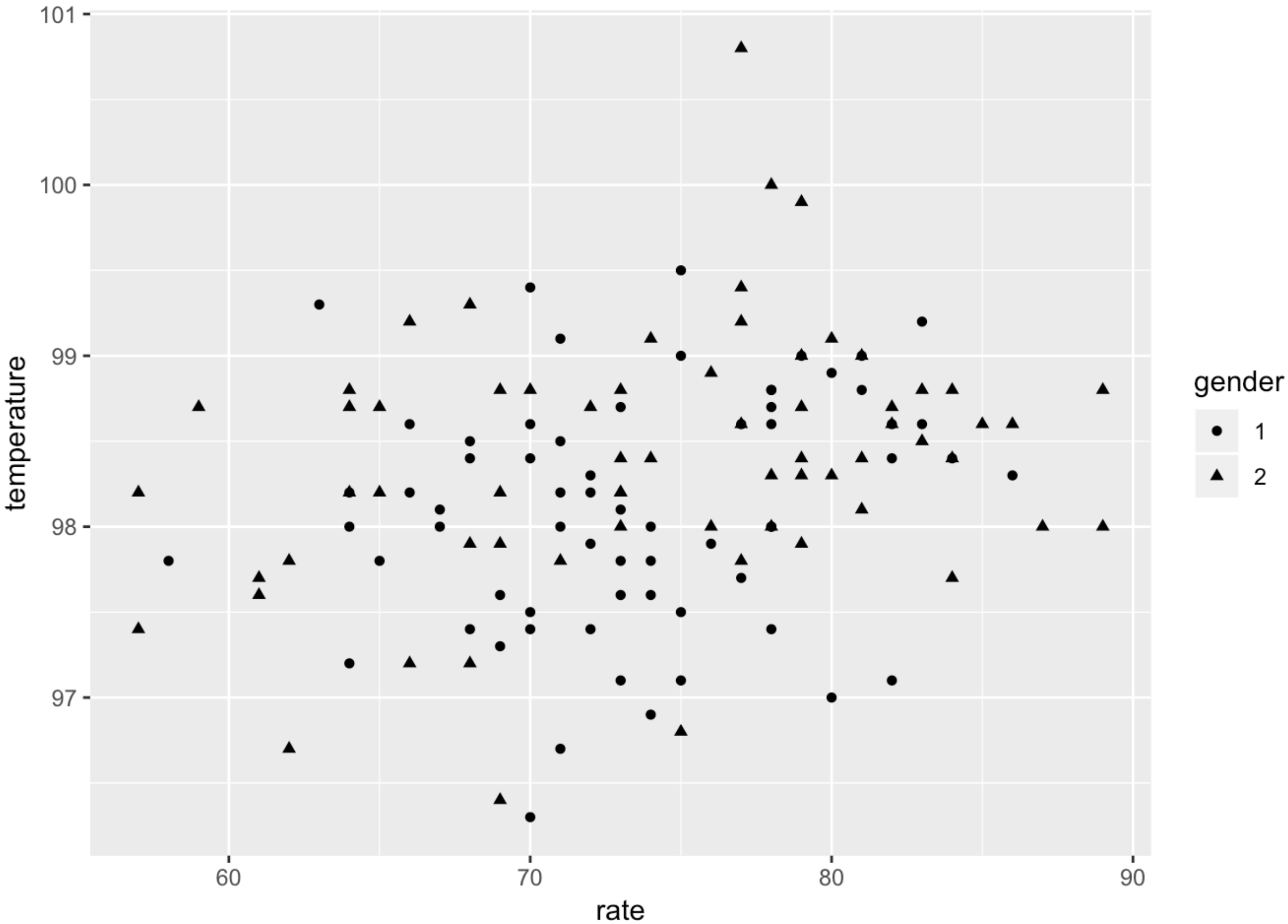

```
plot(female$rate, female$temperature)
```




```
# The graphs look similar and both lack an obvious pattern of any kind. The points ap
pear to be random.
```

Problem 10F: part b)

```
library(ggplot2)

bodytemp$gender <- as.factor(bodytemp$gender)

ggplot(bodytemp, aes(x = rate, y = temperature, group = gender)) + geom_point(aes(sha
pe = gender))
```




```
# The data points of men versus women are similar, except men's heart rate seem to ha
ve less variability.
```

Problem 10F: part c)

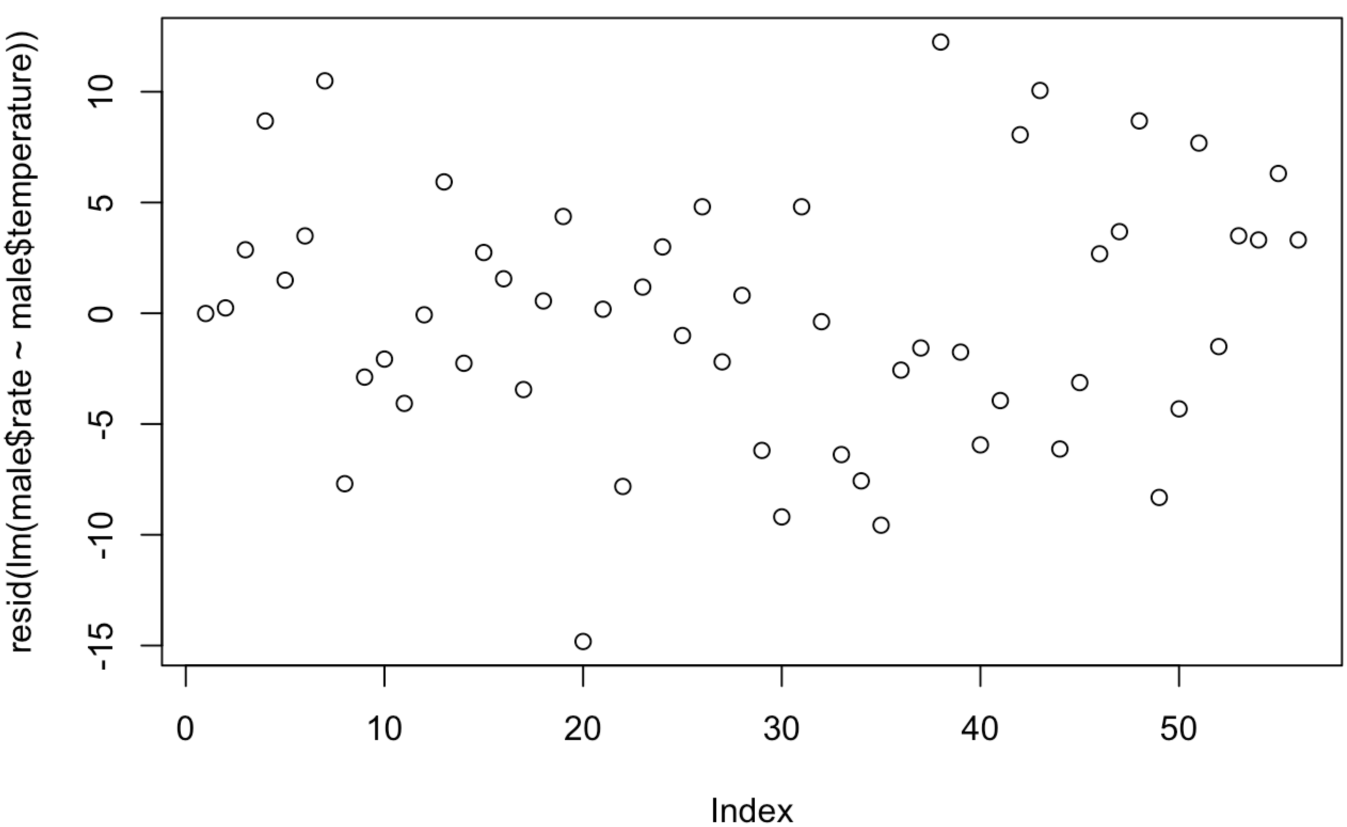
```
lm(male$rate ~ male$temperature) # slope =  1.872
```

```
##
## Call:
## lm(formula = male$rate ~ male$temperature)
##
## Coefficients:
##      (Intercept)  male$temperature
##         -110.260             1.872
```

```
summary(lm(male$rate ~ male$temperature)) # standard error = 1.305
```

```
##
## Call:
## lm(formula = male$rate ~ male$temperature)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.8174  -3.5674   0.0866   3.4942  12.2466
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -110.260    127.760  -0.863    0.392
## male$temperature    1.872      1.305   1.435    0.157
##
## Residual standard error: 5.741 on 54 degrees of freedom
## Multiple R-squared:  0.03673,    Adjusted R-squared:  0.01889
## F-statistic: 2.059 on 1 and 54 DF,  p-value: 0.1571
```

```
plot(resid(lm(male$rate ~ male$temperature)))
```



```
# The residual plot shows no obvious patterns, which is good, showing that a linear p
attern can be seen.
```

Problem 10F: part d)

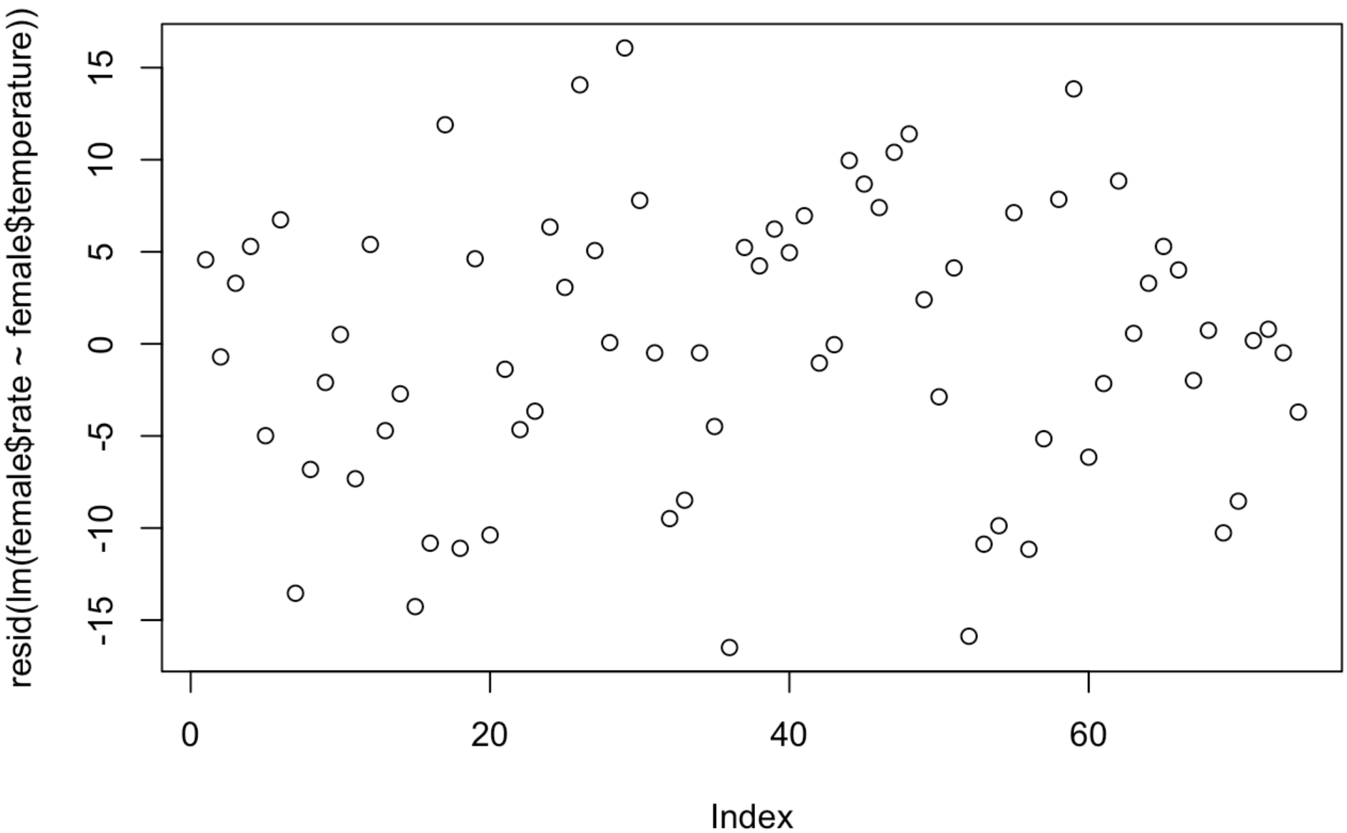
```
lm(female$rate ~ female$temperature) # slope =  2.776
```

```
##
## Call:
## lm(formula = female$rate ~ female$temperature)
##
## Coefficients:
##        (Intercept)  female$temperature
##           -199.098               2.776
```

```
summary(lm(female$rate ~ female$temperature)) # standard error = 1.207
```

```
##
## Call:
## lm(formula = female$rate ~ female$temperature)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.4885  -4.9183   0.1235   5.2908  16.0666
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -199.098    118.885  -1.675   0.0983 .
## female$temperature    2.776      1.207   2.300   0.0244 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.663 on 72 degrees of freedom
## Multiple R-squared:  0.06842,    Adjusted R-squared:  0.05548
## F-statistic: 5.288 on 1 and 72 DF,  p-value: 0.02437
```

```
plot(resid(lm(female$rate ~ female$temperature)))
```




```
# Similar to the male plot, he residual plot shows no obvious patterns, which is good
, showing that a linear pattern can be seen.
```

Problem 10F: part e)

```
# difference in slope = 2.776 - 1.872 = .904
# se = sqrt(1.305 ^ 2 + 1.207 ^ 2) = 1.778
# 95%CI = (-2.58, 4.39)
# Since the CI contains 0, the slopes can be concluded to be equal at a 5% level.
```

Problem 10F: part f)

```
# difference in intercept: -110.260 - (-199.098) = 88.838
# se = sqrt(127.760 ^ 2 + 118.885 ^ 2) = 174.517
# 95%CI = (-253.215, 430.891)
# Since the CI contains 0, the intercepts can be concluded to be equal at a 5% level.
```