STAT151 HW4

Xuanpei Ouyang 3032360371

March 6, 2017

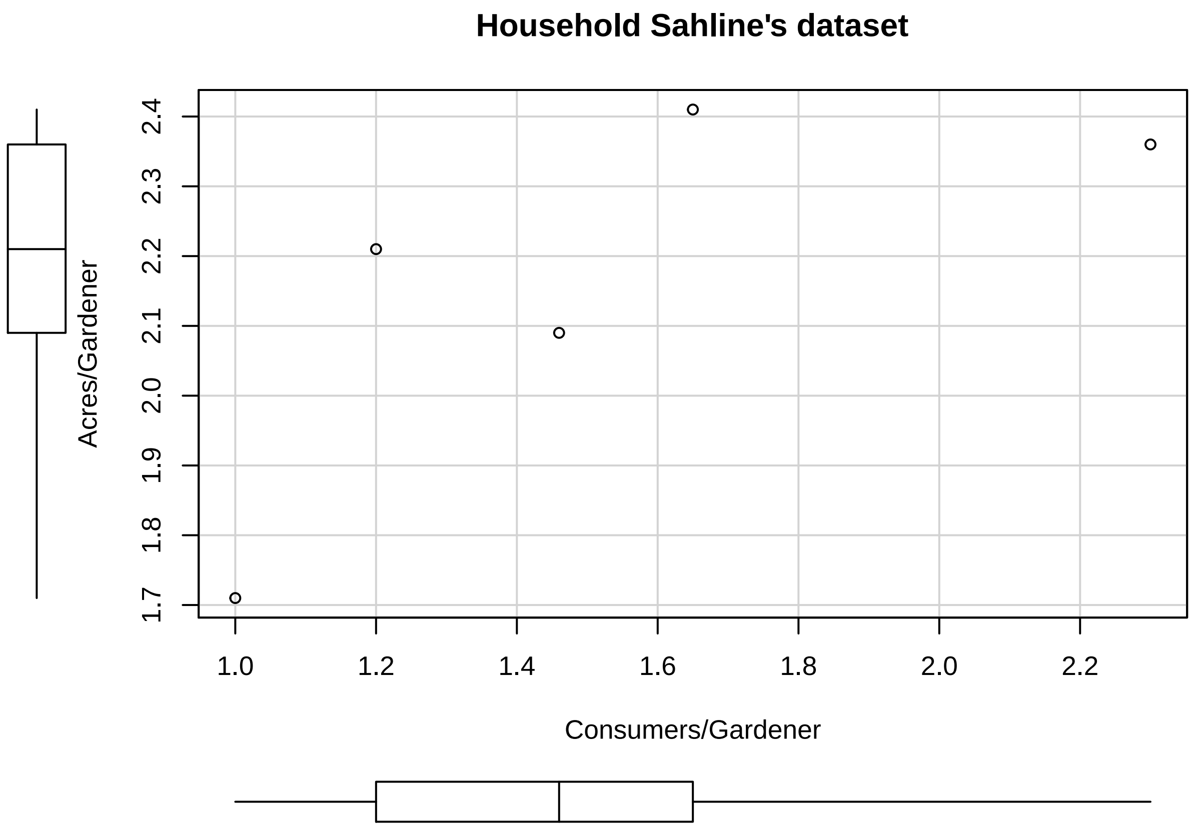
```r
# Some seld defined function used for calculation
find_coeff = function(x, y){
  n = length(x)
  B = (sum(x * y)-n*mean(x)*mean(y))/(sum((x - mean(x))^2))
  A = mean(y) - B*mean(x)
  return(data.frame(
    A = A,
    B = B))
}
find_Sxx = function(x) {
  return(sum((x - mean(x))^2))
}
find_RSS = function(x, y, A, B){
  y_hat = A + B*x
  return(sum((y - y_hat)^2))
}
find_RegSS = function(x, y, A, B){
  y_hat = A + B*x
  return(sum((y_hat - mean(y))^2))
}
find_TSS = function(y){
  return(sum((y - mean(y))^2))
}
find_RMS = function(RSS,x) {
  y_hat = RSS/(length(x)-2)
}
find_R_squared = function(RegSS, TSS){
  return(RegSS/TSS)
}
find_SE_B = function(RMS, Sxx){
  return(sqrt(RMS/Sxx))
}
```

Exercise D5.1

(a) Construct a scatterplot for Y and X.

```r
# create the data table for the 5 sample
sample = data.frame(
  consumers = c(1, 1.2, 1.46, 1.65, 2.3),
  acres = c(1.71, 2.21, 2.09, 2.41, 2.36)
)

# plot the scatter plot
scatterplot(acres ~ consumers, data = sample, main = "Household Sahline's dataset", ylab = "Acres/Garden
```



(b) Find A and B for the least-squares regression of Y on X, and draw the least-squares line on the scatterplot. Interpret A and B.

```
# Compute by self-defined function
coeff = find_coeff(sample$consumers, sample$acres)
A = coeff$A
A
```

```
## [1] 1.531902
```

```
B = coeff$B
B
```

```
## [1] 0.4100511
```

```
# Use R build in function
coeff = lm(acres ~ consumers, data = sample)
summary = summary(coeff)
A = summary$coefficients["(Intercept)","Estimate"]
A
```

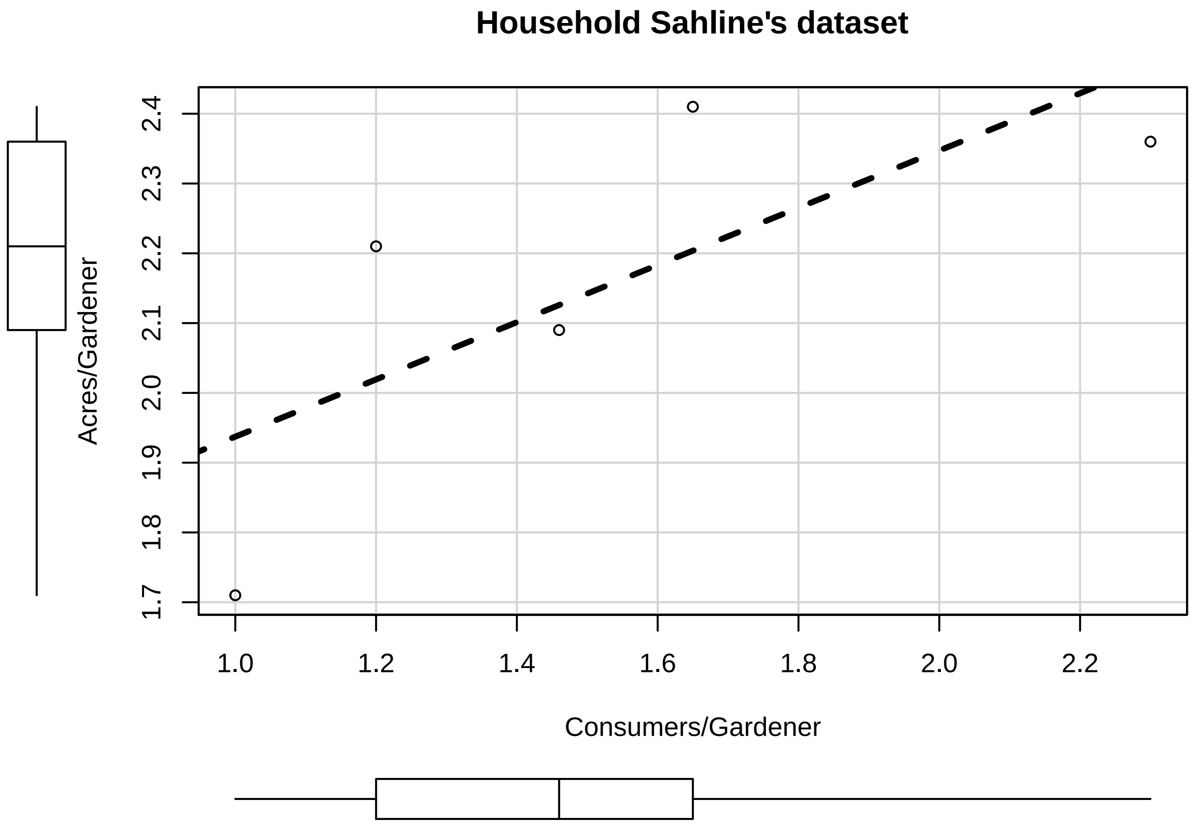
```
## [1] 1.531902
```

```
B = summary$coefficients["consumers","Estimate"]
B
```

```
## [1] 0.4100511
```

```
with(sample, {
  scatterplot(acres ~ consumers, data = sample, main = "Household Sahline's dataset", ylab = "Acres/Gard
```

```
  abline(lm(acres ~ consumers), lwd=3, lty="dashed")
})
```



The value of A means the value of acres per gardener is 1.5319022 when the consumers per gardener is 0. The value of B means the increase in acres per gardener is 0.4100511 with every one unit of increase of the acres per gardener.

(c) Calculate the standard error of the regression, SE, and the correlation coeffcient, r. Interpret these statistics.

```
r_squared = summary$r.squared
r = sqrt(r_squared)
r
```

```
## [1] 0.7343492
```

```
standard_error = summary$sigma
standard_error
```

```
## [1] 0.2190096
```

The correlation coefficient is 0.7343492 between acres per gardener and consumers per gardener and it indicates that there is a fairly strong relationship between acres per gardener and consumers per gardener.

The standard error is about 0.2190096 and the small value of standard error indicates that the fitted linear regression line is good and the prediction made by the regression line is pretty accurate.

Exercise D5.2

```r
# read the related dataset
household = read.table("~/Desktop/STAT 151A/STAT-151A/hw/hw4/Sahlins.txt")

# fit a linear regresison model with build in function
# and calculate A, B, r , standard error
coeff = lm(acres ~ consumers, data = household)
summary = summary(coeff)
A = summary$coefficients["(Intercept)","Estimate"]
A
```

```
## [1] 1.375645
```

```r
B = summary$coefficients["consumers","Estimate"]
B
```

```
## [1] 0.5163201
```

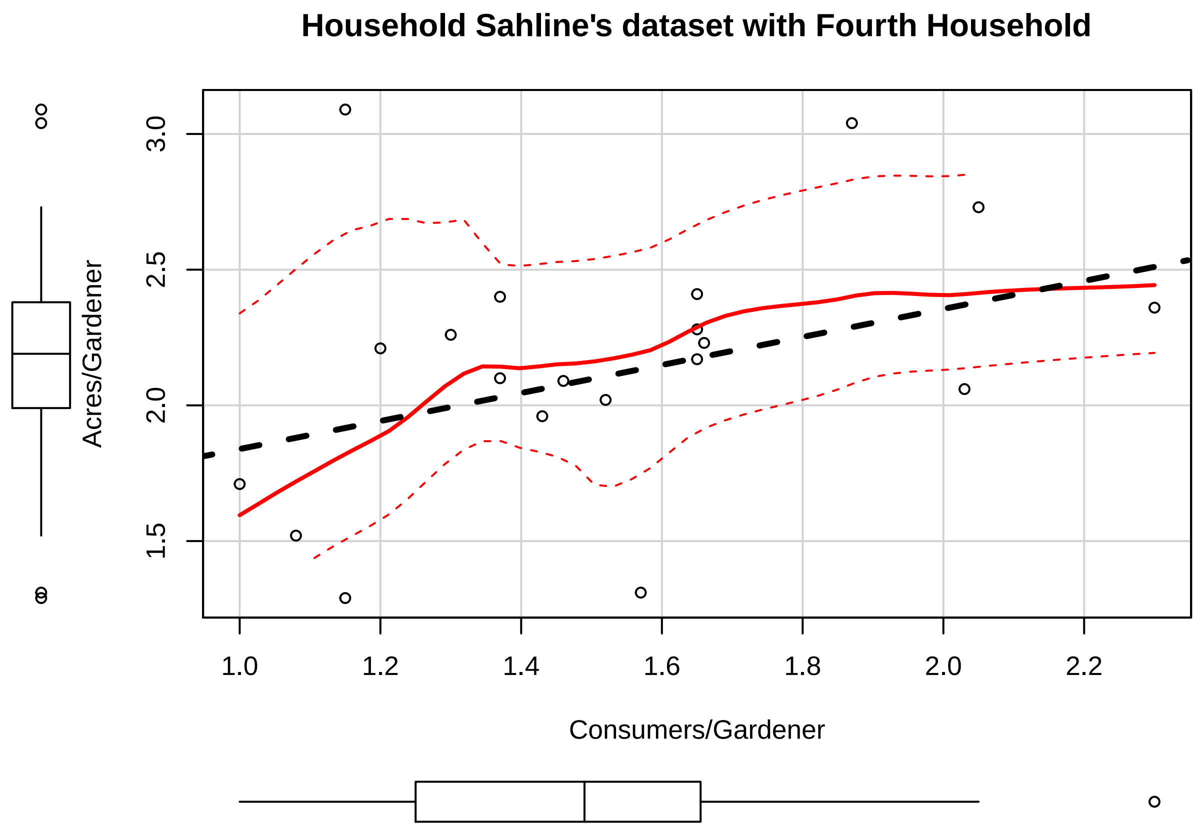
```r
r_squared = summary$r.squared
r = sqrt(r_squared)
r
```

```
## [1] 0.3756561
```

```r
standard_error = summary$sigma
standard_error
```

```
## [1] 0.4543179
```

```r
# plot the scatter plot with regression line
with(household, {
  scatterplot(acres ~ consumers, data = household, main = "Household Sahline's dataset with Fourth Hous
  abline(lm(acres ~ consumers), lwd=3, lty="dashed")
})
```



The value of A means the value of acres per gardener is 1.3756445 when the consumers per gardener is 0. The value of B means the increase in acres per gardener is 0.5163201 with every one unit of increase of the consumers per gardener.

The correlation coefficient is 0.3756561 between acres per gardener and consumers per gardener and it indicates that the relationship between acres per gardener and consumers per gardener is very weak.

The standard error is about 0.4543179 and it indicates that the fitted linear regression line is not very accurate for predicting the response variable acres per gardener.

The regression line here has a positive slope and a intercept around 0. This suggests that the redistribution in this society is purely through the market, each household should have to work in proportion to its consumption needs.

```
household_wo_4th = household[-4,]
coeff = lm(acres ~ consumers, data = household_wo_4th)
summary = summary(coeff)
A = summary$coefficients["(Intercept)","Estimate"]
A
```

```
## [1] 1.000004
```

```
B = summary$coefficients["consumers","Estimate"]
B
```

```
## [1] 0.7215941
```

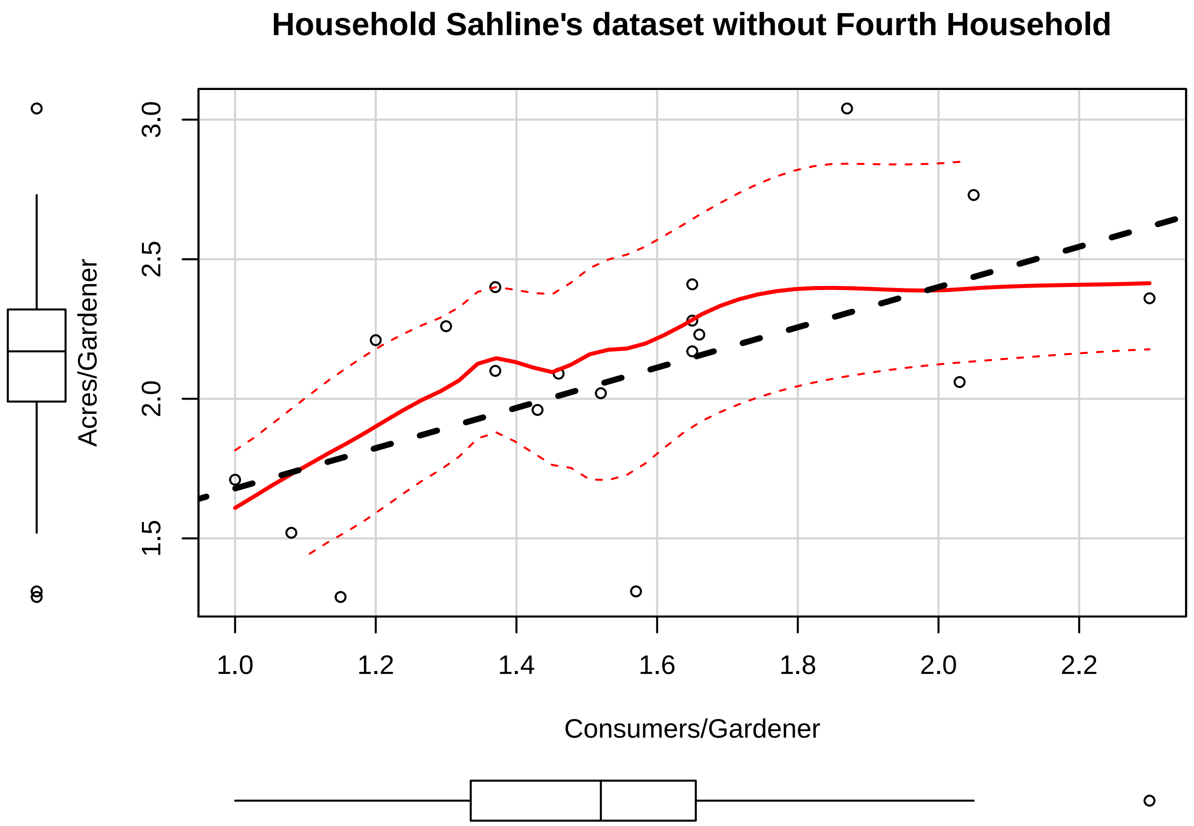
```
r_squared = summary$r.squared
r = sqrt(r_squared)
r
```

```
## [1] 0.5713188
```

```
standard_error = summary$sigma
standard_error
```

```
## [1] 0.3680763
```

```
with(household_wo_4th, {
  scatterplot(acres ~ consumers, data = household_wo_4th, main = "Household Sahline's dataset without Fo
  abline(lm(acres ~ consumers), lwd=3, lty="dashed")
})
```



Now, the value of A means the value of acres per gardener is 1.000004 when the consumers per gardener is 0 and the value of B means the increase in acres per gardener is 0.7215941 with every one unit of increase of the acres per gardener.

After removing the fourth household, the correlation coefficient r increase from 0.376 to 0.571 and standard error decrease from 0.454 to 0.368. After removing the fourth household, we basically have the same conclusion as when we keep the fourth household, the fitted regression model is more accurate and reveals stronger linear relationship since the fourth household data point is a outlier.

```
coeff1 = lm(acres ~ consumers, data = household)
summary1 = summary(coeff1)
summary1
```

```
##
## Call:
## lm(formula = acres ~ consumers, data = household)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.8763 -0.1873 -0.0211  0.2135  1.1206
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.3756     0.4684   2.937  0.00881 **
## consumers     0.5163     0.3002   1.720  0.10263
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4543 on 18 degrees of freedom
## Multiple R-squared:  0.1411, Adjusted R-squared:  0.0934
## F-statistic: 2.957 on 1 and 18 DF,  p-value: 0.1026
```

```
coeff2 = lm(acres ~ consumers, data = household_wo_4th)
summary2 = summary(coeff2)
summary2
```

```
##
## Call:
## lm(formula = acres ~ consumers, data = household_wo_4th)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.82291 -0.16808  0.03215  0.23505  0.69061
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.0000     0.3969   2.519   0.0221 *
## consumers     0.7216     0.2514   2.870   0.0106 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3681 on 17 degrees of freedom
## Multiple R-squared:  0.3264, Adjusted R-squared:  0.2868
## F-statistic: 8.238 on 1 and 17 DF,  p-value: 0.01061
```

If we conduct the t test for these two models above, we can see that the p value - 0.01 for the response variable in the model fitted without the fourth household is much less than that - 0.1 in the model fitted with the fourth household. Thus, if we use alpha = 0.05, we can reject the null hypothesis of no linear regression in the model2 but fail to reject the null hypothesis in model1. I think the regression model fitted without the fourth household does a better job of summarizing the relationship between acres per gardener and consumers per gardener and tßhe fitted regression model without the fourth household reveals a linear relationship between acres per gardener and consumers per gardener.

Exercise D6.1 Assuming that the observations were independently sampled, and the standard error SE(B) of the slope coeffcient, and calculate a 90-percent confidence interval for the population slope.

```
coeff = lm(acres ~ consumers, data = sample)
summary = summary(coeff)
B = summary$coefficients["consumers","Estimate"]
SE_B = summary$coefficients["consumers","Std. Error"]
SE_B
```

```
## [1] 0.2188259
```

```r
df = 3
# to find 90% confidence interval, we find the quantile when prob = 0.95
t_c = qt(0.95, df)
t_c
```

```
## [1] 2.353363
```

```r
lower_bound = B - t_c*SE_B
upper_bound = B + t_c*SE_B
lower_bound
```

```
## [1] -0.1049257
```

```r
upper_bound
```

```
## [1] 0.9250279
```

Assuming that the observations are independently sampled, the standard error SE(B) is 0.2188259 and the 90% Confidence Interval for beta is [-0.1049257,0.9250279]

Exercise D6.2 Find the standard errors of the least-squares intercept and slope. Can we conclude that the population slope is greater than zero? Can we conclude that the intercept is greater than zero? Repeat these computations omitting the fourth household.

```r
coeff = lm(acres ~ consumers, data = household)
summary = summary(coeff)

# standard errors of the intercept
SE_A = summary$coefficients["(Intercept)","Std. Error"]
SE_A
```

```
## [1] 0.4684047
```

```r
# standard errors of the slope
SE_B = summary$coefficients["consumers","Std. Error"]
SE_B
```

```
## [1] 0.3002335
```

```r
# find value A
A = summary$coefficients["(Intercept)","Estimate"]
A
```

```
## [1] 1.375645
```

```r
# find value B
B = summary$coefficients["consumers","Estimate"]
B
```

```
## [1] 0.5163201
```

```r
# t statistics for A
t_A = A/SE_A
t_A
```

```
## [1] 2.936872
```

```r
# t statistics for B
t_B = B/SE_B
t_B
```

```
## [1] 1.719728
```

```r
df = length(household$consumers) - 2
# find the t for quantile = 0.05 since it's one sided test
t_c_0.95 = qt(0.95, df)
t_c_0.95
```

```
## [1] 1.734064
```

```r
summary
```

```
##
## Call:
## lm(formula = acres ~ consumers, data = household)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8763 -0.1873 -0.0211  0.2135  1.1206
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.3756     0.4684   2.937  0.00881 **
## consumers     0.5163     0.3002   1.720  0.10263
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4543 on 18 degrees of freedom
## Multiple R-squared:  0.1411, Adjusted R-squared:  0.0934
## F-statistic: 2.957 on 1 and 18 DF,  p-value: 0.1026
```

1. To check if the population slope is greater than zero, we conduct a one sided t-test with alpha = 0.05.

- Null Hypothesis: beta = 0, the population slope is equal 0.
- Alternative Hypothesis: beta > 0, the population slope is greater than 0.

From the output of summary, since the t-statistics for slope B is 1.7197283 and is less than the critical t value 1.7340636. Therefore, we fail to reject the null hypothesis and conclude that the population slope is not greater than 0 and there is no linear relationship between consumers per gardener and acres per gardener.

2. To check if the intercept is greater than zero, we conduct a one sided t-test with alpha = 0.05.

- Null Hypothesis: alpha = 0, the intercept is equal to 0.
- Alternative Hypothesis: alpha > 0, the intercept is greater than 0.

From the output of summary, since the t-statistics for slope A is 2.9368719 and is greater than the critical t value 1.7340636. Therefore, we reject the null hypothesis and conclude that the intercept is greater than 0.

```r
coeff = lm(acres ~ consumers, data = household_wo_4th)
summary = summary(coeff)

# standard errors of the intercept
SE_A = summary$coefficients["(Intercept)","Std. Error"]
SE_A
```

```
## [1] 0.3969254
```

```r
# standard errors of the slope
SE_B = summary$coefficients["consumers","Std. Error"]
SE_B
```

```
## [1] 0.251414
```

```r
# find value A
A = summary$coefficients["(Intercept)","Estimate"]
A
```

```
## [1] 1.000004
```

```r
# find value B
B = summary$coefficients["consumers","Estimate"]
B
```

```
## [1] 0.7215941
```

```r
# t statistics for A
t_A = A/SE_A
t_A
```

```
## [1] 2.519375
```

```r
# t statistics for B
t_B = B/SE_B
t_B
```

```
## [1] 2.870143
```

```r
df = length(household_wo_4th$consumers) - 2
t_c_0.95 = qt(0.95, df)
t_c_0.95
```

```
## [1] 1.739607
```

```r
summary
```

```
##
## Call:
## lm(formula = acres ~ consumers, data = household_wo_4th)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.82291 -0.16808  0.03215  0.23505  0.69061
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.0000     0.3969   2.519   0.0221 *
## consumers     0.7216     0.2514   2.870   0.0106 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3681 on 17 degrees of freedom
## Multiple R-squared:  0.3264, Adjusted R-squared:  0.2868
## F-statistic: 8.238 on 1 and 17 DF,  p-value: 0.01061
```

3. To check if the population slope is greater than zero, we conduct a one sided t-test with alpha = 0.05.

- Null Hypothesis: beta = 0, the population slope is equal 0.

- Alternative Hypothesis: beta > 0, the population slope is greater than 0.

From the output of summary, since the t-statistics for slope B is 2.8701432 and is greater than the critical t value 1.7396067. Therefore, we reject the null hypothesis and conclude that the population slope is greater than 0 and there is a linear relationship between consumers per gardener and acres per gardener.

4. To check if the intercept is greater than zero, we conduct a one sided t-test with alpha = 0.05.

- Null Hypothesis: alpha = 0, the intercept is equal to 0.
- Alternative Hypothesis: alpha > 0, the intercept is greater than 0.

From the output of summary, since the t-statistics for slope A is 2.5193754 and is greater than the critical t value 1.7396067. Therefore, we reject the null hypothesis and conclude that the intercept is greater than 0.

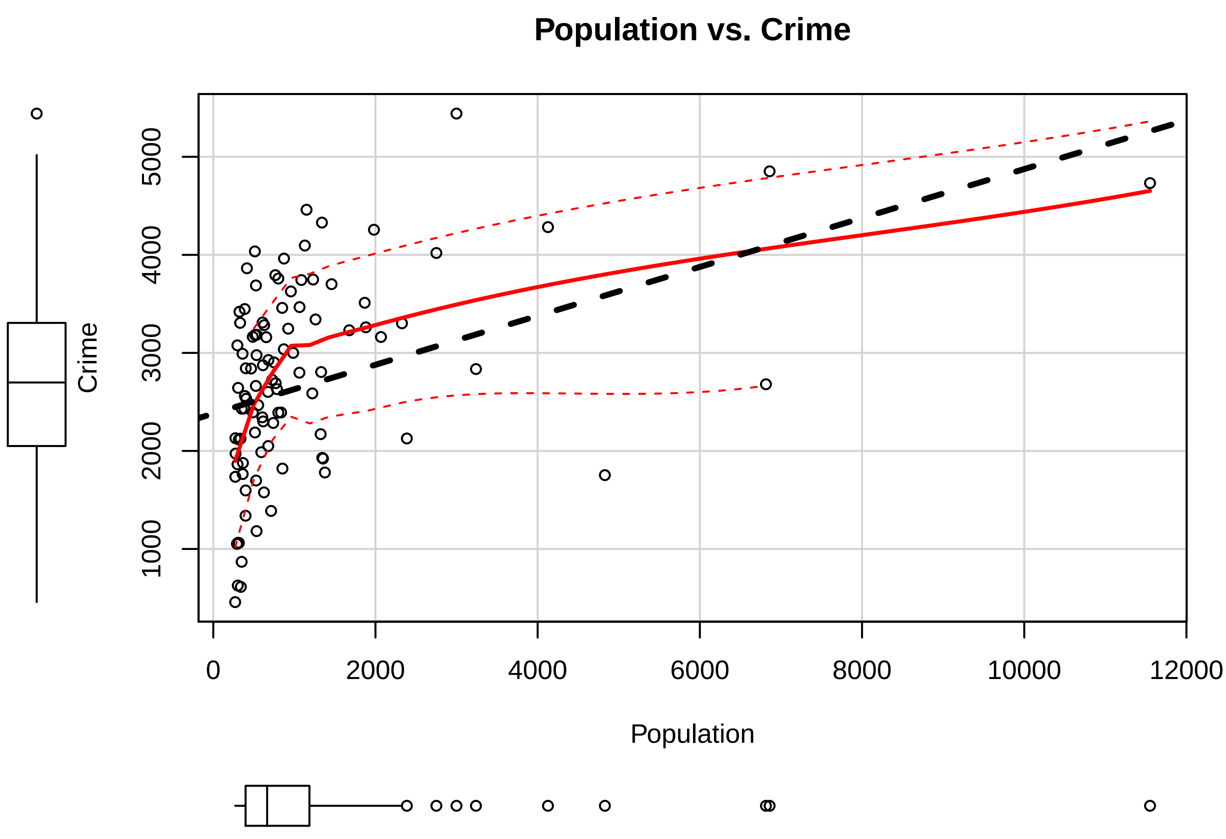
Exercise D6.3 Construct 95-percent confidence intervals for and in each of the least-squares simple-regression analyses that you performed in Exercise D5.3.

Here I choose Freedman as my dataset.

```
Freedman = read.table("~/Desktop/STAT 151A/STAT-151A/hw/hw4/Freedman.txt")
```

1. Simple regression on population and crime

```
# plot the scatterplot for population and crime
with(Freedman, {
  scatterplot(crime ~ population, data = Freedman, main = "Population vs. Crime", ylab = "Crime", xlab =
  abline(lm(crime ~ population), lwd=3, lty="dashed")
})
```



```r
# fitting the linear regression model for population and crime
coeff = lm(crime ~ population, data = Freedman)
summary = summary(coeff)
A = summary$coefficients["(Intercept)","Estimate"]
A
```

```
## [1] 2449.736
```

```r
B = summary$coefficients["population","Estimate"]
B
```

```
## [1] 0.2495038
```

```r
r_squared = summary$r.squared
r = sqrt(r_squared)
r
```

```
## [1] 0.3957592
```

```r
standard_error = summary$sigma
standard_error
```

```
## [1] 907.8697
```

```r
#standard errors of the intercept
SE_A = summary$coefficients["(Intercept)","Std. Error"]
SE_A
```

```
## [1] 112.5003
```

```r
#standard errors of the slope
SE_B = summary$coefficients["population","Std. Error"]
SE_B
```

```
## [1] 0.05848486
```

```r
# construct the 0.95 confidence interval for B
df = length(Freedman$population) - 2
t_c = qt(0.975, df)
t_c
```

```
## [1] 1.982173
```

```r
lower_bound_B = B - t_c*SE_B
upper_bound_B = B + t_c*SE_B
lower_bound_B
```

```
## [1] 0.1335767
```

```r
upper_bound_B
```

```
## [1] 0.365431
```

```r
# construct the 0.95 confidence interval for A
lower_bound_A = A - t_c*SE_A
upper_bound_A = A + t_c*SE_A
lower_bound_A
```

```
## [1] 2226.741
```

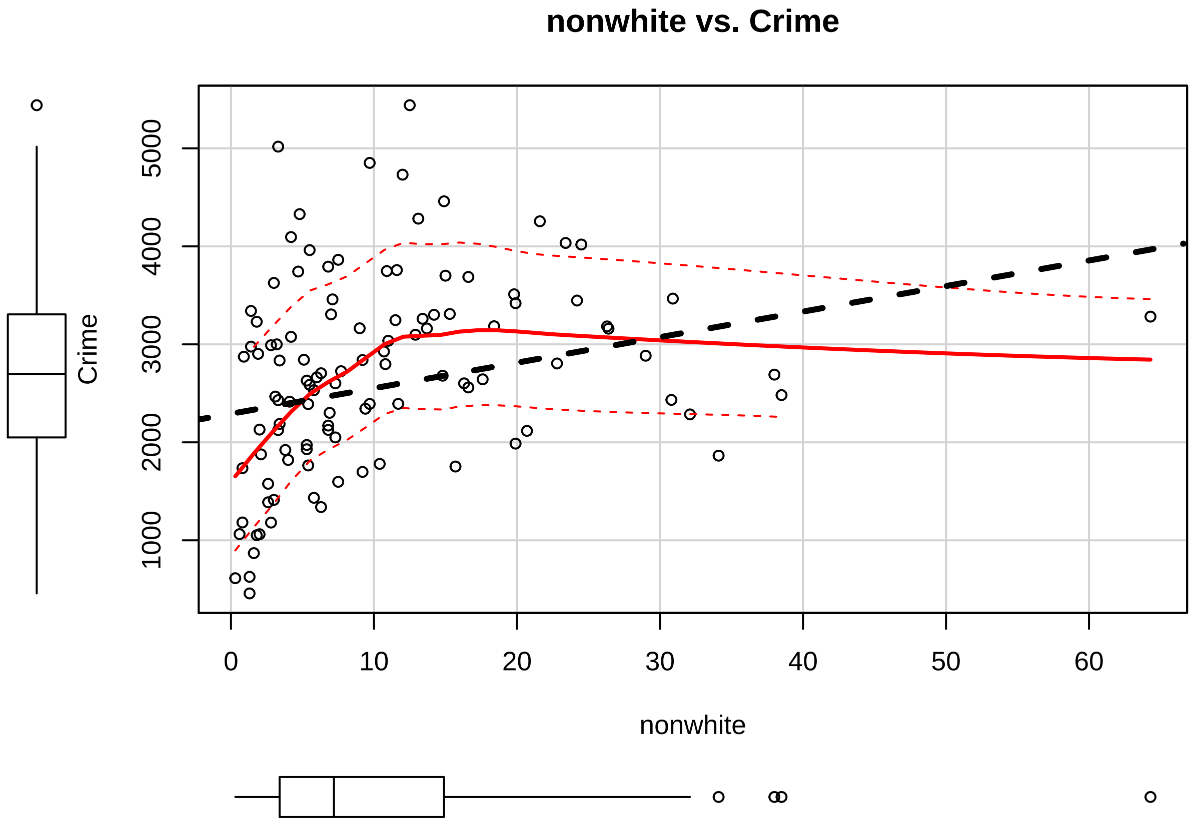
```r
upper_bound_A
```

```
## [1] 2672.731
```

For the linear regression model between population and crime, the 95% Confidence Interval for beta is [0.1335767,0.365431] and the 95% confidence interval for alpha is [2226.7411058,2672.7311946]

2. Simple regression on nonwhite and crime

```r
# plot the scatterplot for nonwhite and crime
with(Freedman, {
  scatterplot(crime ~ nonwhite, data = Freedman, main = "nonwhite vs. Crime", ylab = "Crime", xlab = "no
  abline(lm(crime ~ nonwhite), lwd=3, lty="dashed")
})
```




```r
# fitting the linear regression model for nonwhite and crime
coeff = lm(crime ~ nonwhite, data = Freedman)
summary = summary(coeff)
A = summary$coefficients["(Intercept)","Estimate"]
A
```

```
## [1] 2432.448
```

```r
B = summary$coefficients["nonwhite","Estimate"]
B
```

```
## [1] 26.07065
```

```r
r_squared = summary$r.squared
r = sqrt(r_squared)
r
```

```
## [1] 0.269784
```

```r
standard_error = summary$sigma
standard_error
```

```
## [1] 959.0505
```

```r
#standard errors of the intercept
SE_A = summary$coefficients["(Intercept)","Std. Error"]
SE_A
```

```
## [1] 133.108
```

```r
#standard errors of the slope
SE_B = summary$coefficients["nonwhite","Std. Error"]
SE_B
```

```
## [1] 8.953947
```

```r
# construct the 0.95 confidence interval for B
df = length(Freedman$crime) - 2
t_c = qt(0.975, df)
t_c
```

```
## [1] 1.982173
```

```r
lower_bound_B = B - t_c*SE_B
upper_bound_B = B + t_c*SE_B
lower_bound_B
```

```
## [1] 8.322377
```

```r
upper_bound_B
```

```
## [1] 43.81893
```

```r
# construct the 0.95 confidence interval for A
lower_bound_A = A - t_c*SE_A
upper_bound_A = A + t_c*SE_A
lower_bound_A
```

```
## [1] 2168.605
```

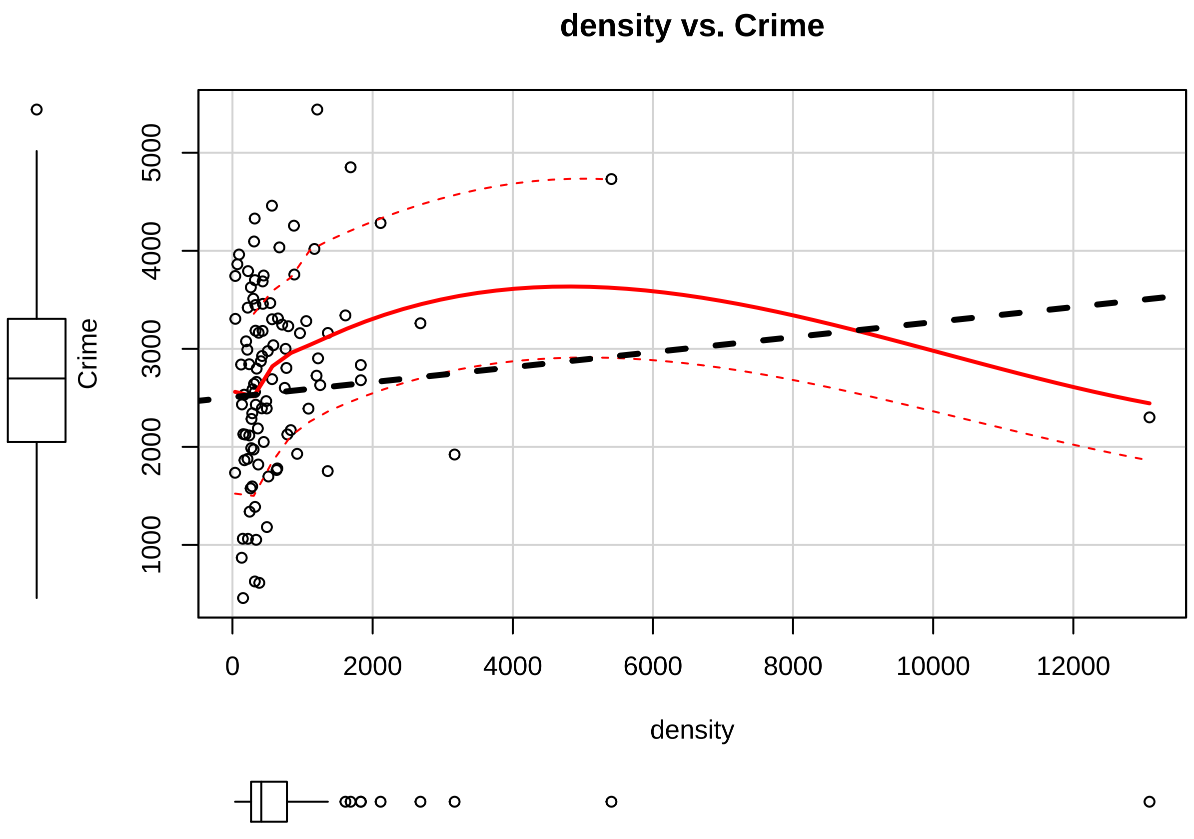
```r
upper_bound_A
```

```
## [1] 2696.291
```

For the linear regression model between nonwhite and crime, the 95% Confidence Interval for beta is [8.3223766,43.8189291] and the 95% confidence interval for alpha is [2168.6045322,2696.2907996]

3. Simple regression on density and crime

```r
# plot the scatterplot for density and crime
with(Freedman, {
  scatterplot(crime ~ density, data = Freedman, main = "density vs. Crime", ylab = "Crime", xlab = "dens
  abline(lm(crime ~ density), lwd=3, lty="dashed")
})
```




```
# fitting the linear regression model for density and crime
coeff = lm(crime ~ density, data = Freedman)
summary = summary(coeff)
A = summary$coefficients["(Intercept)","Estimate"]
A
```

```
## [1] 2674.556
```

```
B = summary$coefficients["density","Estimate"]
B
```

```
## [1] 0.07655275
```

```
r_squared = summary$r.squared
r = sqrt(r_squared)
r
```

```
## [1] 0.1122283
```

```
standard_error = summary$sigma
standard_error
```

```
## [1] 982.3378
```

```
#standard errors of the intercept
SE_A = summary$coefficients["(Intercept)","Std. Error"]
SE_A
```

```
## [1] 111.3472
```

```
#standard errors of the slope
SE_B = summary$coefficients["density","Std. Error"]
```

```
SE_B
```

```
## [1] 0.06846885
```

```
# construct the 0.95 confidence interval for B
df = length(Freedman$crime) - 2
t_c = qt(0.975, df)
t_c
```

```
## [1] 1.982173
```

```
lower_bound_B = B - t_c*SE_B
upper_bound_B = B + t_c*SE_B
lower_bound_B
```

```
## [1] -0.0591644
```

```
upper_bound_B
```

```
## [1] 0.2122699
```

```
# construct the 0.95 confidence interval for A
lower_bound_A = A - t_c*SE_A
upper_bound_A = A + t_c*SE_A
lower_bound_A
```

```
## [1] 2453.846
```

```
upper_bound_A
```

```
## [1] 2895.265
```

For the linear regression model between density and crime, the 95% Confidence Interval for beta is [-0.0591644,0.2122699] and the 95% confidence interval for alpha is [2453.8463075,2895.2654053]