Homework 2

Alyssa Sharma

2023-09-12

R Markdown

Question 1.19

Part a:

The least squares estimate of β_1 is 0.03883 according to the summary function. The least squares estimate of β_0 is 2.11405 according to the summary function.

The estimated regression function is $\hat{Y}_i = 2.11405 + 0.03883(X_i) + \epsilon$

```
adf <- read.table("http://www.cnachtsheim-text.csom.umn.edu/Kutner/Chapter%20%201%20Data%20Sets/CH01PR19

head(adf)
```

```
##      V1 V2
## 1 3.897 21
## 2 3.885 14
## 3 3.778 28
## 4 2.540 22
## 5 3.028 21
## 6 3.865 31
```

```
colnames(adf) <- c("gpa","act")

a_LS_model <- lm(gpa ~ act, data = adf)
summary_a_LS_model <- summary(a_LS_model)
```

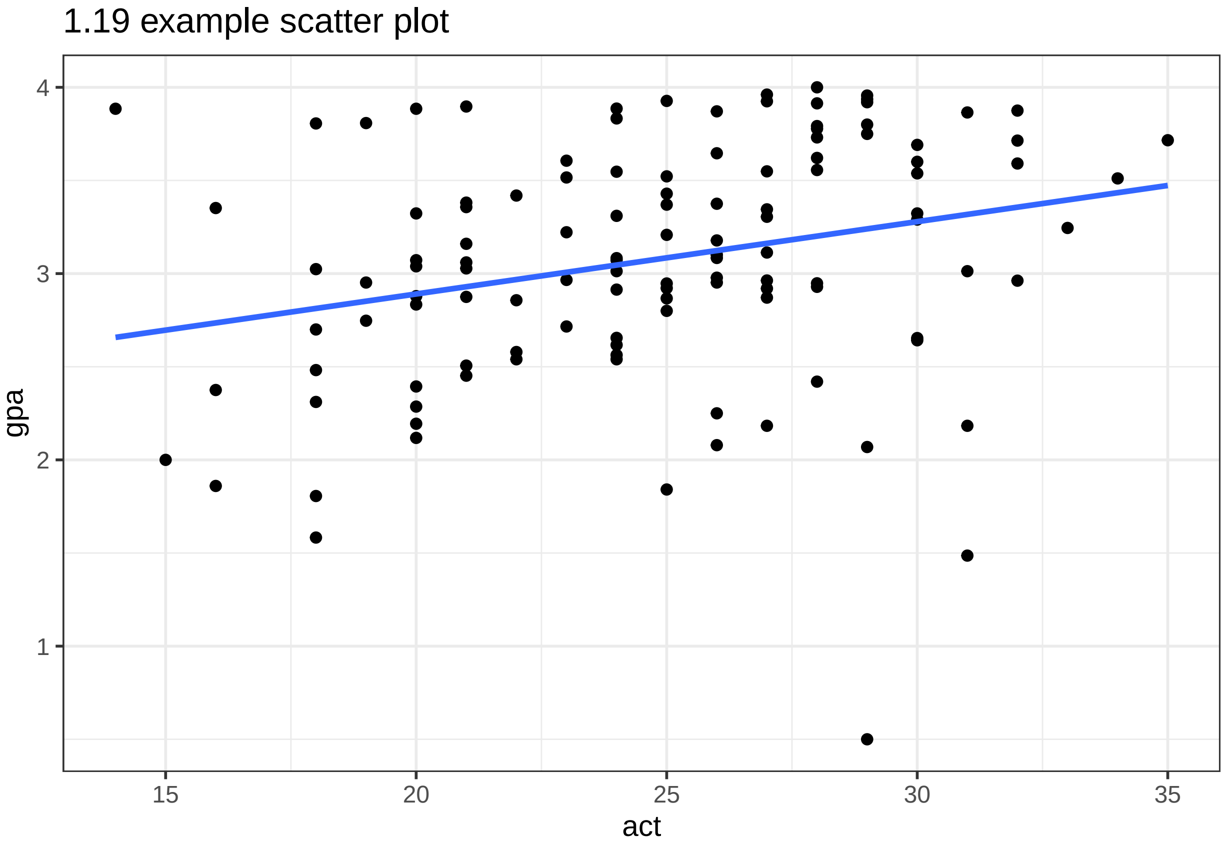
```
##
## Call:
## lm(formula = gpa ~ act, data = adf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.74004 -0.33827  0.04062  0.44064  1.22737
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.11405    0.32089   6.588  1.3e-09 ***
## act          0.03883    0.01277   3.040  0.00292 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6231 on 118 degrees of freedom
## Multiple R-squared:  0.07262,    Adjusted R-squared:  0.06476
## F-statistic:  9.24 on 1 and 118 DF,  p-value: 0.002917
```

Part b:

The estimated regression function does not seem to fit the data well.

```
library(ggplot2)
ggplot(adf, aes(x = act, y = gpa)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "act", y = "gpa", title = "1.19 example scatter plot") +
  theme_bw()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Part c:

The point estimate of GPA when the ACT test score is 30 is $\hat{Y_{12}} = 3.279$.

```
library(moderndive)
Fittedandresiduals <-get_regression_points(a_LS_model)
Fittedandresiduals
```

```
## # A tibble: 120 x 5
##        ID   gpa   act gpa_hat residual
##     <int> <dbl> <int>   <dbl>    <dbl>
##  1      1  3.90    21    2.93    0.968
##  2      2  3.88    14    2.66    1.23
##  3      3  3.78    28    3.20    0.577
##  4      4  2.54    22    2.97   -0.428
##   5      5  3.03    21    2.93    0.099
##   6      6  3.86    31    3.32    0.547
##   7      7  2.96    32    3.36   -0.395
##   8      8  3.96    27    3.16    0.799
##   9      9  0.5     29    3.24   -2.74
## 10     10  3.18    26    3.12    0.054
## # i 110 more rows
```

```
#Method 2
#Simply use the summary_toluca_LS_model object to get residuals and fitted values

#Here is 1.23:
sumris <- sum(Fittedandresiduals$residual)
sse <- sum((Fittedandresiduals$residual)^2)
sumris
```

```
## [1] -0.005
```

```
sse
```

```
## [1] 45.82027
```

```
#part b
mse <- sse/(nrow(adf) - 2)
mse
```

```
## [1] 0.3883074
```

```
sqrt(mse)
```

```
## [1] 0.6231431
```

Part d: The point estimate of the change in the mean response when the entrance test score increases by one point is β_1, which is equal to 0.03883.

Question 1.20

Part a:

The estimated regression function is $\hat{Y_i} = 0.254192 + 0.063683(X_i) + \epsilon$

Here is the code for 1.20:

```
copyvstime <- read.table("http://www.cnachtsheim-text.csom.umn.edu/Kutner/Chapter%20%201%20Data%20Sets/

head(copyvstime)
```

```
##    V1 V2
## 1  20  2
## 2  60  4
## 3  46  3
## 4  41  2
## 5  12  1
## 6 137 10
```

```
colnames(copyvstime) <- c("copy","time1")

ct_LS_model <- lm(time1 ~ copy, data = copyvstime)
summary_ct_LS_model <- summary(ct_LS_model)
summary_ct_LS_model
```

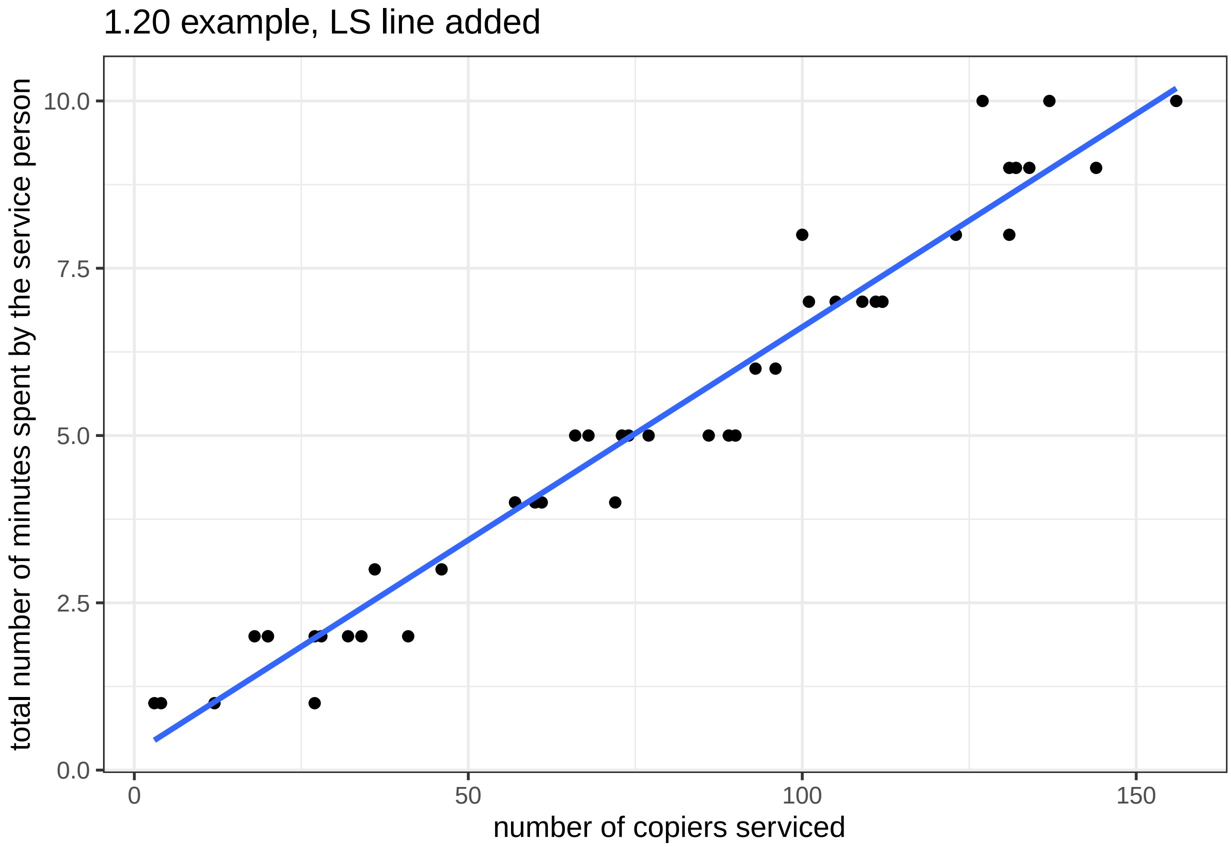
```
##
## Call:
## lm(formula = time1 ~ copy, data = copyvstime)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.98570 -0.36780 -0.03733  0.40328  1.65802
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.254192   0.178413   1.425    0.161
## copy        0.063683   0.002046  31.123   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5801 on 43 degrees of freedom
## Multiple R-squared:  0.9575, Adjusted R-squared:  0.9565
## F-statistic: 968.7 on 1 and 43 DF,  p-value: < 2.2e-16
```

Part b:

The estimated regression function fits the data very well. The residuals are quite small as compared to the previous problem. There is a strong positive correlation in both the estimated regression function and in the original data.

```
ggplot(copyvstime, aes(x = copy, y = time1)) +
  geom_point() +
  labs(x = " number of copiers serviced", y = "total number of minutes spent by the service person", ti
  geom_smooth(method = "lm", se = FALSE) +
  theme_bw()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Part c: β_0 has no meaning in this problem. If 0 copiers are serviced, then in the real world, the number of minutes spent servicing is also going to be 0, not 0.254192.

Part d: The point estimate of the mean service time when 5 copiers are serviced is 0.572607. This number is calculated below.

```
library(moderndive)
Fittedandresiduals <-get_regression_points(ct_LS_model)
Fittedandresiduals
```

```
## # A tibble: 45 x 5
##       ID time1  copy time1_hat residual
##    <int> <int> <int>     <dbl>    <dbl>
##  1     1     2    20     1.53     0.472
##  2     2     4    60     4.08    -0.075
##  3     3     3    46     3.18    -0.184
##  4     4     2    41     2.86    -0.865
##  5     5     1    12     1.02    -0.018
##  6     6    10   137     8.98     1.02
##  7     7     5    68     4.58     0.415
##  8     8     5    89     5.92    -0.922
##  9     9     1     4     0.509    0.491
## 10    10     2    32     2.29    -0.292
## # i 35 more rows
```

```
a <- .254192 + (.063683*5)
a
```

```
## [1] 0.572607
```

Question 1.23

Part a: The residuals sum to -0.005, which is approximately 0. This number may have been produced due to a floating point error. Part b: The estimate of σ^2 is 0.3883074. The estimate of σ is 0.6231431.

```
library(moderndive)
Fittedandresiduals <-get_regression_points(a_LS_model)
Fittedandresiduals
```

```
## # A tibble: 120 x 5
##       ID   gpa   act gpa_hat residual
##    <int> <dbl> <int>   <dbl>    <dbl>
##  1     1  3.90    21    2.93    0.968
##  2     2  3.88    14    2.66    1.23
##  3     3  3.78    28    3.20    0.577
##  4     4  2.54    22    2.97   -0.428
##  5     5  3.03    21    2.93    0.099
##  6     6  3.86    31    3.32    0.547
##  7     7  2.96    32    3.36   -0.395
##  8     8  3.96    27    3.16    0.799
##  9     9  0.5     29    3.24   -2.74
## 10    10  3.18    26    3.12    0.054
## # i 110 more rows
```

```
#Method 2
#Simply use the summary_toluca_LS_model object to get residuals and fitted values

#Here is 1.23:
sumris <- sum(Fittedandresiduals$residual)
sse <- sum((Fittedandresiduals$residual)^2)
sumris
```

```
## [1] -0.005
```

```
sse
```

```
## [1] 45.82027
```

```
#part b
mse <- sse/(nrow(adf) - 2)
mse
```

```
## [1] 0.3883074
```

```
sqrt(mse)
```

```
## [1] 0.6231431
```

Question 1.24

Part a:

The residuals e_i are computed with the get_regression_points formula below. The sum of the residuals Σe_i is -0.002 which is approximately zero. The sum of the squared residuals $\Sigma {e_i}^2$ is 14.47071. The sum of squared residuals is much larger than the sum of the residuals.

Part b: The estimate of σ^2 is 0.3365282. The estimate of σ is 0.5801105. These values are expressed in minutes serviced and squared minutes serviced.

```
#here is the problem 1.24 code
library(moderndive)
Fittedandresiduals <-get_regression_points(ct_LS_model)
Fittedandresiduals
```

```
## # A tibble: 45 x 5
##       ID time1  copy time1_hat residual
##    <int> <int> <int>     <dbl>    <dbl>
##  1     1     2    20     1.53     0.472
##  2     2     4    60     4.08    -0.075
##  3     3     3    46     3.18    -0.184
##  4     4     2    41     2.86    -0.865
##  5     5     1    12     1.02    -0.018
##  6     6    10   137     8.98     1.02
##  7     7     5    68     4.58     0.415
##  8     8     5    89     5.92    -0.922
##  9     9     1     4     0.509    0.491
## 10    10     2    32     2.29    -0.292
## # i 35 more rows
```

```
#a <- .254192 + (.063683*5)
#a

sumris <- sum(Fittedandresiduals$residual)
sumris
```

```
## [1] -0.002
```

```
sumrissq <-  sum((Fittedandresiduals$residual)^2)
sumrissq
```

```
## [1] 14.47071
```

```
b <- sumrissq/(nrow(copyvstime) - 2)
b
```

```
## [1] 0.3365282
```

```
sqrt(b)
```

```
## [1] 0.5801105
```