Homework #2

Yunting Chiu

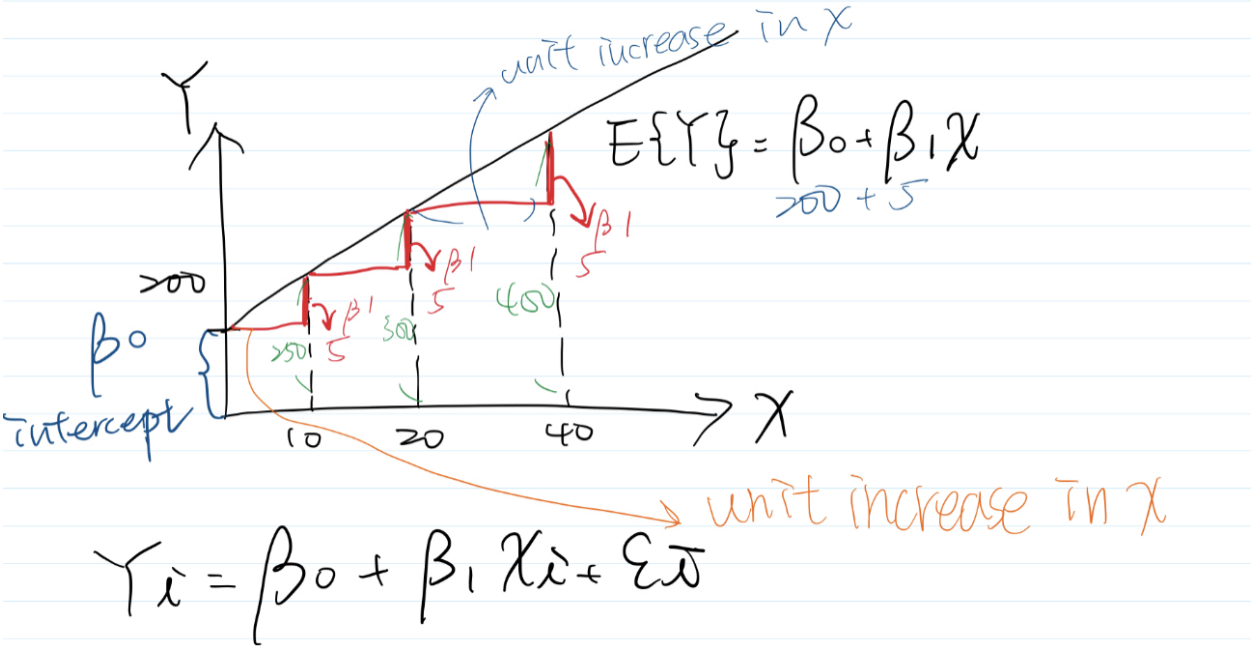
2021-02-05

1. (**1.2**) The members of a health spa pay annual membership dues of $$$300 plus a charge of $$$2 for each visit to the spa. Let Y denote the dollar cost for the year for a member and X the number of visits by the member during the year. Express the relation between X and Y mathematically. Is it a functional relation or a statistical relation (that is, is the relation deterministic or stochastic)?

- deterministic: the output of the model is entirely determined by the values of the parameters and the initial conditions; stochastic: random, unpredictable.
- The association between X and Y is: Y = 300 + 2X (dollars).
  This is a functional relationship because a effect in the value of the X will cause the corresponding change in the value of the Y. X and Y will not have a uncertain effect.

2. (**1.6**) Suppose the regression parameters are $\beta 0 = 200$ and $\beta 1 = 5.0$.

(a) Plot the regression equation.



(b) Predict the response for X = 10, 20, and 40.

(b)

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

$Y = 200 + 5X_i + \varepsilon_i$ $\quad \varepsilon_i \sim N(0,1)$

$E\{Y_i\} = 200 + 5X_i + E\{\varepsilon_i\}$ ($E\{\varepsilon_i\} \to 0$)

$E\{Y_i\} = 200 + 5 \times 10 = 250$ # (if $X=10$)

$E\{Y_i\} = 200 + 5 \times 20 = 300$ # (if $X=20$)

$E\{Y_i\} = 200 + 5 \times 40 = 400$ # (if $X=40$)

(c) Explain the meaning of parameters $\beta 0$ and $\beta 1$.

- $\beta 0$ = Y intercept of regression line.
- $\beta 1$: one unit change in X, generates a $\beta 1$ unit change in Y.

3. (**1.10**) An analyst in a large corporation studied the relation between current annual salary (Y) and age (X) for the 46 computer programmers presently employed in the company. The analyst concluded that the relation is curvilinear, reaching a maximum at 47 years. Does this imply that the salary for a programmer increases until age 47 and then decreases? Explain.

- **curvilinear** only explain that X and Y are not linear relation. It is not true because it reaches its maximum at a point and then increasing at a decreasing rate meaning that wage first increases to a max at year 47 and then the increasing rate slows down. In reality, decreasing salary as people age in any company is not make sense, we also can find many examples in real world.

4. The time it takes to transmit a file always depends on the file size. Suppose you transmitted 30 files, with the average size of 126 Kbytes and the standard deviation of 35 Kbytes. The average transmittance time was 0.04 seconds with the standard deviation of 0.01 seconds. The correlation coefficient between the time and the size was 0.86. Based on this data, fit a linear regression model and predict the time it will take to transmit a 400 Kbyte file.

- Based on the model above, it will take 0.1085 sec to transmit a 400 Kbyte file.

30 samples { time = y average : 0.04sec sd : 0.01sec
file size : x average = 126K sd : 35K

correlation coefficient (r) = 0.86

Q: find predict lm take to transmit a 400kbyte file.

$$r = \frac{S_{xy}}{S_x S_y} \Rightarrow \beta_1 = \frac{S_{xy}}{S_x^2} = r\frac{S_y}{S_x}$$

$$r = 0.86 \Rightarrow \beta_1 = 0.86\,\frac{0.01}{35} = 0.00025$$

$$\beta_0 = \bar{Y} - \beta_1 \bar{X} \Rightarrow 0.04 - 0.00025 \times 126$$

$$= 0.04 - 0.0315 = 0.0085$$

Model

$$y_i = \beta_0 + \beta_1 x_i \cdot \varepsilon_i$$

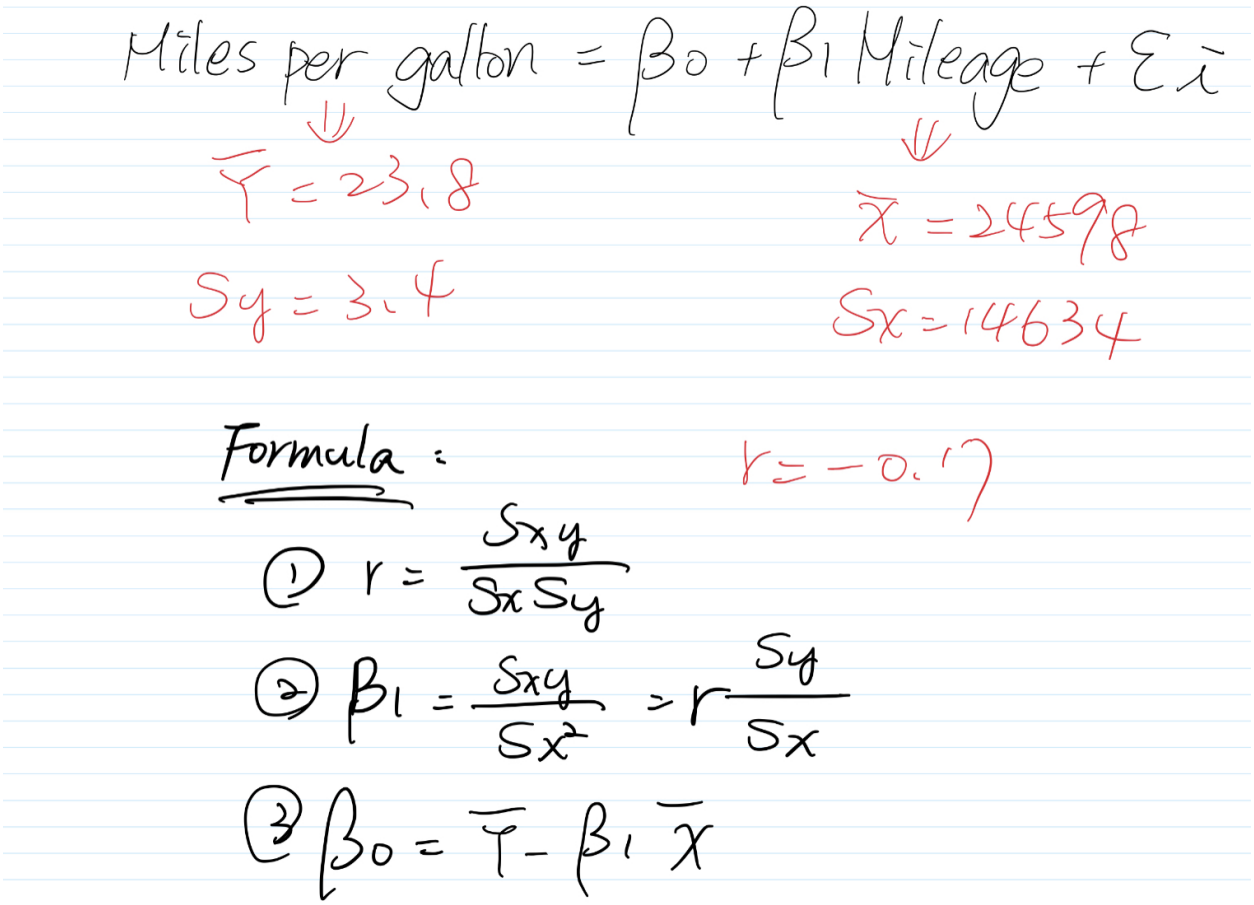
$$E\{Y_i\} = 0.0085 + E\{0.00025 \times 400\}$$

$$= 0.0085 + E\{0.1\}$$

$$= 0.1085 \text{ (sec)} \#$$

5. At a gas station, 180 drivers were asked to record the mileage of their cars and the number of miles per gallon. The results are summarized in the table.

(a) Compute the least squares regression line which describes how the number of miles per gallon depends on the mileage.



- Reference: https://youtu.be/yttN024P-Gg

```
# slope(b1) = r*sd(y)/sd(x)
slope <- ((-0.17*3.4)/14634)
slope
```

```
## [1] -3.949706e-05
```

```
# y intercept(b0) = sample mean of y - slope* sample mean of x
yIntercept <- 23.8-(slope*24598)
yIntercept
```

```
## [1] 24.77155
```

(b) What do the obtained slope and intercept mean in this situation?

- X is Mileage, Y is Miles per gallon. The regression line indicates that every increment in mileage will decrease miles per gallon by 0.00003.949706.
- Intercept means that when mileage equals to zero, the miles per gallon would be 24.77155. However, mileage would be zero but miles per gallon would be 24.77155. In reality, it's doesn't have any practical meaning.

(c) You purchase a used car with 35,000 miles on it. Predict the number of miles per gallon.

- 23.38915 Miles per gallon

```
# y = b0+b1*X
predict35000 <- yIntercept + slope * 35000
predict35000
```

```
## [1] 23.38915
```

6. (**Stat-615 only**) Show that the sample intercept b0 is a linear and unbiased estimator of the population intercept $\beta 0$.

$$E(b_0) = \beta_0 \; ?$$

$$\bar{Y}_i = b_0 + b_1 \bar{X}_i \Rightarrow b_0 = \bar{Y}_i - b_1 X_i$$

$$\begin{aligned}
E(b_0) &= E(\bar{Y} - b_1 \bar{X}) \\
&= E(\bar{Y}) - E(b_1 \bar{X}) \\
&= E(\bar{Y}) - E(b_1)\bar{X} \\
&= E(\bar{Y}) - \beta_1 \bar{X} \\
&= \frac{1}{n}\sum Y_i - \beta_1 \bar{X} \qquad \left(Y_1 + Y_2 + \cdots Y_n = \sum Y_i\right) \\
&= \frac{1}{n}\sum(\beta_0 + \beta_1 X_i) - \beta_1 \bar{X} \qquad = \beta_1 X_1 + \beta_1 X_2 + \beta_1 X_3 + \cdots \beta_1 X_n \\
&= \frac{1}{n}\left(\sum \beta_0 + \sum \beta_1 X_i\right) - \beta_1 \bar{X} \\
&= \frac{1}{n}\left(n \cdot \beta_0 + \beta_1 n \cdot \bar{X}\right) - \beta_1 \bar{X} \\
&= \beta_0 + \beta_1 \bar{X} - \beta_1 \bar{X} \\
&= \beta_0
\end{aligned}$$

$$\therefore E(b_0) = \beta_0$$

- We know that E(b1) = $\beta 1$, because of the proof is from class book and https://people.stat.sc.edu/hansont/stat704/lecture4.pdf

7. (Computer project - 1.19, 1.24). Grade point average. The director of admissions of a small college selected 120 students at random from the new freshman class in a study to determine whether a students grade point average (GPA) at the end of the freshman year (Y) can be predicted from the ACT test score (X). The results of the study follow.

- X is ACT test score, and Y is students GPA.

(a) Obtain the least squares estimates of $\beta 0$ and $\beta 1$ and state the estimated regression function.

- The $\beta 0$ is 2.11405 and the $\beta 1$ is 0.03883. Based on this regression model below, if ACT score increase by one, GPA will increase by 0.03883.

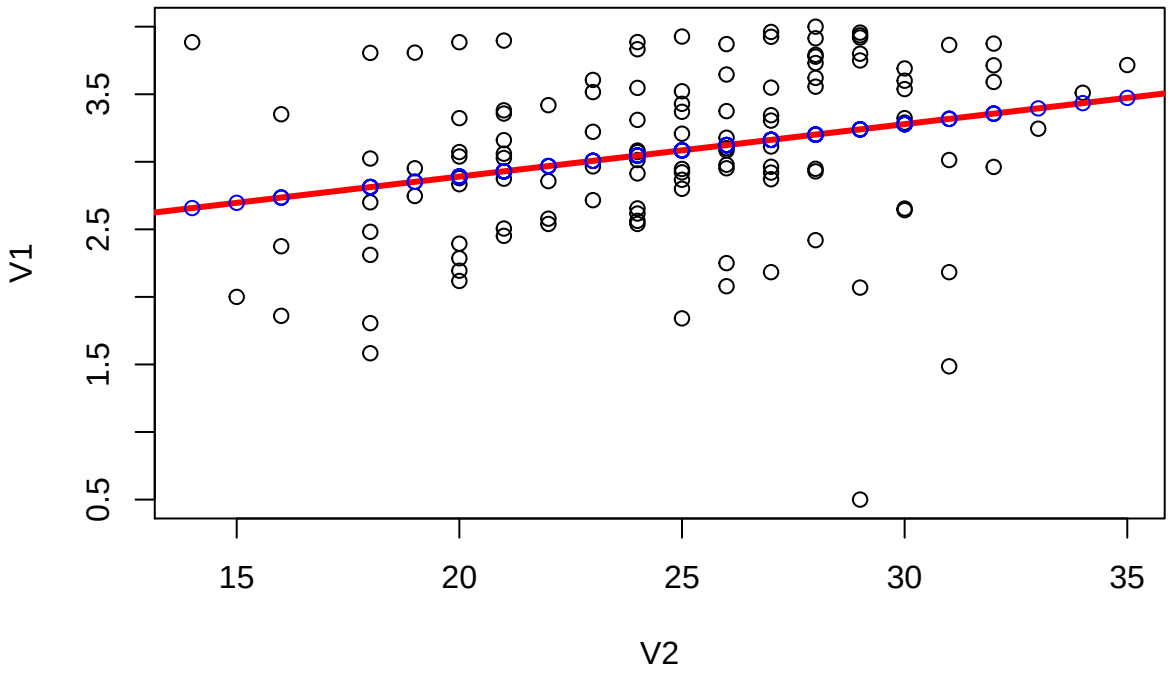
```r
asc <- read.table("./data/CH01PR19.txt")

reg <- lm(V1 ~ V2, data = asc)
summary(reg)
```

```
##
## Call:
## lm(formula = V1 ~ V2, data = asc)
##
## 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 ***
## V2           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
```

(b) Plot the estimated regression function and the data. Does the estimated regression function appear to fit the data well?

- According to the plot below, there are a few blue data points fitting the line, meaning that the majority of points does not fit the model.

```r
attach(asc)
plot(V2, V1)
reg <- lm(V1 ~ V2)
abline(reg, col = "red", lwd = 3)
Yhat = predict(reg, x = V2)
points(V2, Yhat, col = "blue")
```




```
# summary(reg)
```

(c) Obtain a point estimate of the mean freshman GPA for students with ACT test score X = 30.

- The estimated mean of freshman GPA is 3.278863

```
predict(reg, data.frame(V2 = 30))
```

```
##        1
## 3.278863
```

(d) What is the point estimate of the change in the mean response when the entrance test score increases by one point?

- The slope and the intercept are both significant. According to the regression results, the change in the mean of GPA will increase in 0.03883.

```
summary(reg)
```

```
##
## Call:
## lm(formula = V1 ~ V2)
##
## 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 ***
## V2           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
```

(e) Obtain the residuals ei and the sum of the squared residuals.

- Ans: residual is -7.175e-05, and sum of the squared residuals is 45.818 with 118 degrees of freedom.

```
# formula = smaple mean of Y - b0 + b1* sample mean of x + residual(ei)
meanGPA <- mean(asc$V1)
meanGPA
```

```
## [1] 3.07405
```

```
meanACT <- mean(asc$V2)
meanACT
```

```
## [1] 24.725
```

```
residual <- meanGPA - 2.11405 - 0.03883 * meanACT
residual
```

```
## [1] -7.175e-05
```

```
anova(reg)
```

```
## Analysis of Variance Table
##
## Response: V1
##            Df Sum Sq Mean Sq F value   Pr(>F)
## V2          1  3.588  3.5878  9.2402 0.002917 **
## Residuals 118 45.818  0.3883
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(f) Obtain point estimates of σ^2 and σ. In what units is each of them expressed?

- The mean square of the error $= s^2 = 0.3883$. Since s^2 is an unbiased estimator for σ^2, the point estimate for σ^2 is 0.3883. Similarly s = 0.6231372 and it is an unbiased estimator for σ, so the point estimate for σ is 0.6231372. The unit of each is GPA score.

```
anova(reg)
```

```
## Analysis of Variance Table
##
## Response: V1
##            Df Sum Sq Mean Sq F value   Pr(>F)
## V2          1  3.588  3.5878  9.2402 0.002917 **
## Residuals 118 45.818  0.3883
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
sigma <- sqrt(0.3883)
sigma
```

```
## [1] 0.6231372
```