Stat 615/415 (Regression) HW #2 – Solutions

Regression Basics (chap. 1)

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)?

   Solution. *Y = 300+2X . This is a functional deterministic relation. There is no error term. The cost is completely determined by the number of visits, and there is no uncertainty about it.*

2. (**1.6**) Suppose the regression parameters are $\beta_0 = 200$ and $\beta_1 = 5.0$.

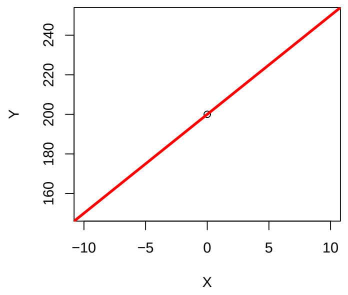
   (a) Plot the regression equation.

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

   (c) Explain the meaning of parameters β_0 and β_1.

   Solution.

   (a) *With these intercept and slope, the regression equation is* $y = 200 + 5x$.



   In R, you can write

```
X=0; Y=200; plot(X,Y,xlim=c(-10,10),ylim=c(150,250));
abline(200,5,col="red",lwd=3);
```

   A hand-made plot is certainly good too.

   (b) *Substitute* $X = 10, 20, 40$ *into the regression equation to get*

$$\begin{aligned} \hat{Y}(10) &= 200 + 5(10) = 250 \\ \hat{Y}(10) &= 200 + 5(20) = 300 \\ \hat{Y}(10) &= 200 + 5(40) = 400 \end{aligned}$$

   (c) *The intercept* $\beta_0 = 200$ *means that for all population units with independent variable* $X = 0$, *the expected response is* $\mathbf{E}(Y) = 200$. *The slope* $\beta_1 = 5$ *means that a 1-unit increment of* X *causes a 5-unit increase in the expected response.*

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.

   Solution. *It does not . The stated relation estimates trends in the population. One can use it to compare different age groups, but it says nothing about individual trends. To make conclusions about the change in salary for individual programmers, one has to track salaries of individual programmers in the sample, observing them over time.*

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.

   Solution. *We are given:* $n = 30$, $\bar{X} = 126$, $s_x = 35$, $\bar{Y} = 0.04$, $s_y = 0.01$, *and* $r = 0.86$.

   Compute the least squares estimates,

$$\begin{aligned} b_1 &= r\left(\frac{s_y}{s_x}\right) = (0.86)\left(\frac{0.01}{35}\right) = 0.000246 \\ b_0 &= \bar{Y} - b_1\bar{X} = 0.04 - (0.000246)(126) = 0.009. \end{aligned}$$

   The fitted regression line has an equation $y = 0.009 + 0.000246x$.

   The time it takes to transmit a 400 Kbyte file is predicted as

$$\hat{Y} = 0.009 + 0.000246X = 0.009 + (0.000246)(400) = \text{0.107 seconds} .$$

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.

	Sample mean	Standard deviation
Mileage	24,598	14,634
Miles per gallon	23.8	3.4

   The sample correlation coefficient is $r = -0.17$.

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

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

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

   Solution. *In this problem,* $n = 180$, $\bar{X} = 24,598$, $s_x = 14,634$, $\bar{Y} = 23.8$, $s_y = 3.4$, *and* $r = -0.17$.

   (a) *Parameters of the regression line predicting the number of miles per gallon based on the mileage are estimated as*

$$\begin{aligned} b_1 &= r\left(\frac{s_y}{s_x}\right) = (-0.17)\left(\frac{3.4}{14,634}\right) = -0.0000395 \\ b_0 &= \bar{Y} - b_1\bar{X} = 23.8 - (-0.0000395)(24,598) = 24.77. \end{aligned}$$

   The fitted regression line has an equation $y = 24.77 - 0.0000395x.$

   (b) *The negative slope means that cars with higher mileage (older cars) make a smaller number of miles per gallon. Namely, after a car is driven additional 100,000 miles, it is expected to make 3.95 less miles per gallon. According to the obtained intercept, brand new cars (with 0 mileage) on the average make 24.77 miles per gallon.*

   (c) *The number of miles per gallon is predicted by*

$$\hat{Y} = 24.77 - (0.0000395)(35,000) = \text{23.39 miles per gallon}$$

6. (**Stat-615 only**) Show that the sample intercept b_0 is a linear and unbiased estimator of the population intercept β_0.

SOLUTION. *The sample intercept is* $b_0 = \overline{Y} - b_1\overline{X} = \sum_{i=1}^{n} \frac{1}{n} Y_i - b_1\overline{X}$.

This is a linear function of $Y_1, \ldots, Y_n$. *Also, since* $\mathbf{E}(Y_i) = \beta_0 + \beta_1 X_i$, *and we already proved in class that* $\mathbf{E}(b_1) = \beta_1$, *we have*

$$\begin{aligned} \mathbf{E}(b_0) &= \sum_{i=1}^{n} \frac{1}{n} \mathbf{E}(Y_i) - \mathbf{E}(b_1)\overline{X} = \sum_{i=1}^{n} \frac{1}{n} (\beta_0 + \beta_1 X_i) - \beta_1 \overline{X} \\ &= \sum_{i=1}^{n} \frac{1}{n} \beta_0 + \beta_1 \sum_{i=1}^{n} \frac{1}{n} X_i - \beta_1 \overline{X} = \beta_0 + \beta_1 \overline{X} - \beta_1 \overline{X} = \beta_0, \end{aligned}$$

which proves that b_0 *is an unbiased estimator of* β_0.

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.

i	1	2	3	$\cdots$	118	119	120
X_i	21	14	28	$\cdots$	28	16	28
Y_i	3.897	3.885	3.778	$\cdots$	3.914	1.860	2.948

The full data set is available on our course blackboard site. To read text (ASCII) file, you can use an R command

```
read.table("CH01PR19.txt")
```

(a) Obtain the least squares estimates of β_0 and β_1 and state the estimated regression function.

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

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

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

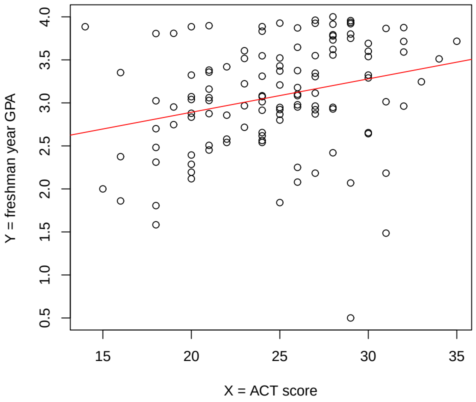
(e) Obtain the residuals e_i and the sum of the squared residuals $\sum e_i^2$.

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

SOLUTION. *Results are based on the R code and R output below.*

(a) $b_0 = 2.114$, $b_1 = 0.039$, *and the estimated regression function is* $y = 2.114 + 0.039x$.

(b) There is a lot of variation of responses around the regression line. A positive trend is apparent, and so, regression explains captures the trend, but it cannot produce an accurate prediction of the GPA. Points are relatively far from the regression line, so no, it does not fit the data well.



(c) $\hat{Y} =$ *3.279*

(d) *This is precisely the slope* $b_1 =$ *0.039* .

(e) $\sum e_i^2 =$ *45.82* .

(f) $\widehat{\sigma}^2 = s^2 =$ *0.388* , *measured in squared units of the GPA.*
$\widehat{\sigma} = s =$ *0.623* , *measured in the original units of the GPA.*

R code

```
> GPA = read.table("C:\\Data\\Book data\\Chapter  1 Data Sets\\CH01PR19.txt")
> attach(GPA)
> head(GPA)
     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
> X = V2; Y = V1;      # This matches the table of values given in the problem

> reg = lm(Y ~ X)
> reg
(Intercept)            X
    2.11405      0.03883

> plot(X,Y,xlab="X = ACT score",ylab="Y = freshman year GPA")
> abline(reg,col="red")

> predict(reg, data.frame(X=30))
3.278863

> e = Y - fitted.values(reg)      # Residuals
> sum(e^2)                        # Sum of squared residuals (Error Sum of Squares)
[1] 45.81761

> n = length(X)                   # Sample size
> var_est = sum(e^2)/(n-2)
> var_est                         # Estimated Var(Y)
[1] 0.3882848
> sqrt(var_est)                   # Estimated Std(Y)
[1] 0.623125

> summary(reg)                    # A direct way of estimating Std(Y)
Residual standard error: 0.6231 on 118 degrees of freedom

> anova(reg)                      # A direct way of estimating Var(Y)
Analysis of Variance Table        # is by computing the Mean Squared Error
           Df Sum Sq Mean Sq F value   Pr(>F)
X           1  3.588  3.5878  9.2402 0.002917 **
Residuals 118 45.818  0.3883
```