Michael Leibert
Math 651
Homework 1

1.5. When asked to state the simple linear regression model, a student wrote it as follows: $E[Y_i] = \beta_0 + \beta_1 X_i + \epsilon_i$. Do you agree?

No, the simple linear regression model is $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$. $E[Y_i]$ is equal to $\beta_0 + \beta_1 X_i$.

1.6. Consider the normal error regression model (1.24). Suppose that the parameter values are $\beta_0 = 200, \beta_1 = 5.0$, and $\sigma = 4$.

b. Explain the meaning of the parameters β_0 and β_1. Assume that the scope of the model includes $X = 0$.

The parameters β_0 and β_1 are regression coefficients. β_1 is the slope of the regression line. It indicates the change in the mean of the probability distribution of Y per unit increase in X. The parameter β_0 is the Y intercept of the regression line. When the scope of the model includes $X = 0$, β_0 gives the mean of the probability distribution of Y at $X = 0$.

1.19. 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 student's 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. Assume that first-order regression model (1.1) is appropriate.

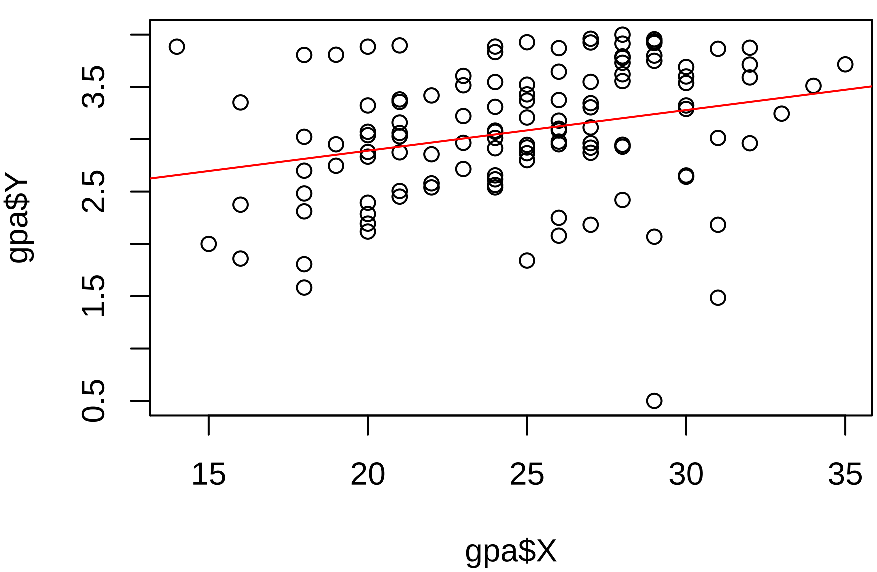
a. Obtain the least squares estimates of β_0 and β_1. and state the estimated regression function.

$$\beta_0 = 2.11405; \qquad \beta_1 = 0.03883; \qquad \hat{Y} = 2.11405 + 0.03883X$$

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

```
colnames(gpa)<-c("Y","X")
```

```
par(mar=c(5.1, 4.1, .6, 2.1))
plot(gpa$X,gpa$Y)
abline(lm( gpa$Y ~ gpa$X), col="red")
```



It is apparent that there is a relationship between ACT and GPA in the plot. I do not think the regression line fits the data well.

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

$$E[Y_{30}] = 2.11405 + 0.03883 \cdot 30 = 3.278863$$

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

$$\beta_1 = 0.03882713$$

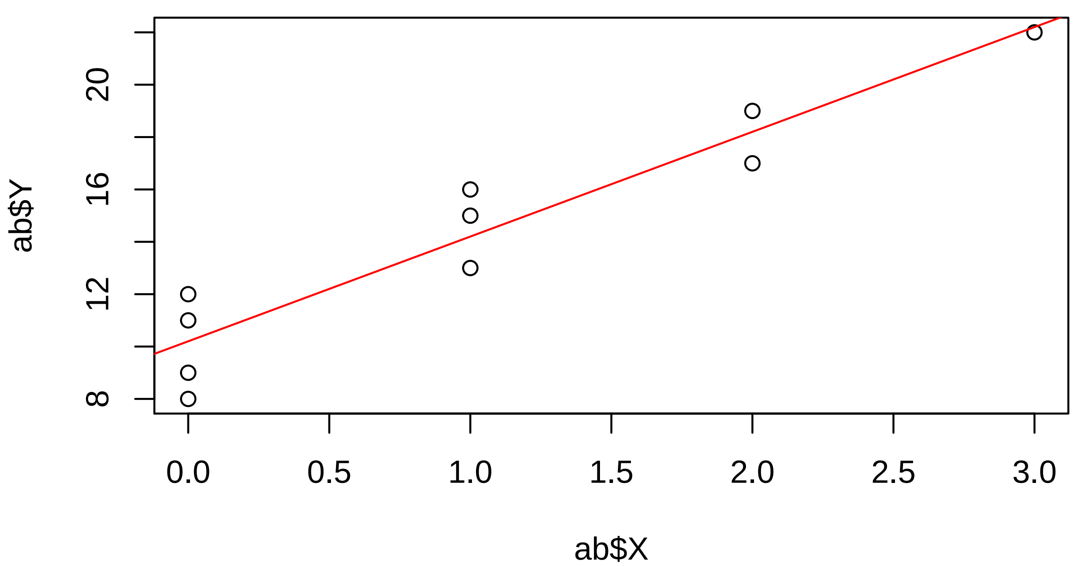
1.21. Airfreight breakage. A substance used in biological and medical research is shipped by airfreight to users in cartons of 1,000 ampules. The data below, involving 10 shipments, were collected on the number of times the carton was transferred from one aircraft to another over the shipment route (X) and the number of ampules found to be broken upon arrival (Y). Assume that first-order regression model (1.1) is appropriate.

a. Obtain the estimated regression function. Plot the estimated regression function and the data. Does a linear regression function appear to give a good fit here?

$$\beta_1 = 4; \qquad \beta_2 = 10.2; \qquad \hat{Y} = 10.2 + 4X$$

```
colnames(ab)<-c("Y","X")
```

```
par(mar=c(5.1, 4.1, .6, 2.1))
plot(ab$X,ab$Y)
abline(lm(ab$Y ~ ab$X), col="red")
```



The linear regression function appears to be a good fit.

b. Obtain a point estimate of the expected number of broken ampules when $X = 1$ transfer is made.

$$\hat{Y_h} = 10.2 + 4(1) = 14.2$$

c. Estimate the increase in the expected number of ampules broken when there are 2 transfers as compared to 1 transfer.

$$\big(10.2 + 4(2)\big) - \big(10.2 + 4(1)\big) = 4$$

d. Verify that your fitted regression line goes through the point $(\bar{X}, \bar{Y})$.

$$\bar{X} = 1; \qquad \bar{Y} = 14.2; \qquad 10.2 + 4(1) = 14.2$$

1.23. Refer to Grade point average Problem 1.19.

a. Obtain the residuals e_i;. Do they sum to zero in accord with (1.17)?

```
options(scipen=999)
b0<-as.numeric(lm( gpa$Y ~ gpa$X)[[1]][1])
b1<-as.numeric(lm( gpa$Y ~ gpa$X)[[1]][2])
gpa$ei<- (gpa$Y - (b1* gpa$X + b0 ) );gpa$ei
```

```
##   [1]  0.96758105  1.22737094  0.57679116 -0.42824608  0.09858105
##   [6]  0.54730978 -0.39451735  0.79861829 -2.74003597  0.05444541
##  [11]  0.26409967  0.25913691  0.03709967 -0.03290033 -0.15034448
##  [16] -0.19938171  0.43727254 -0.30469022 -0.13772746 -0.77259183
##  [21] -0.48290033  0.42758105  0.52979116  0.76261829  0.35479116
##  [26] -0.02255459 -0.78120884 -0.38924608  0.74744541  0.13058105
##  [31]  0.84227254 -0.36028332 -0.27220884  0.25144541 -0.11124608
##  [36]  0.02609967  0.45158105  0.01113691  0.38661829  0.52244541
##  [41] -0.14555459 -0.62486309 -0.50590033 -0.87355459 -1.17103597
##  [46] -0.42890033 -1.13469022 -0.69645619  0.10023530  0.99306243
##  [51] -0.29138171  0.61671668  0.14261829 -0.17155459  0.50109967
##  [56]  0.41213691  0.23058105 -0.69659183  0.04413691  0.69596403
##  [61] -0.16272746 -0.29107321  0.28527254  0.59892679 -0.63686309
##  [66] -0.47741895 -0.39090033  0.35748265 -1.00693757  0.50892679
##  [71]  0.14840817 -0.04107321 -0.33093757 -0.11293757  0.67996403
##  [76] -0.05659183  0.21492679 -0.03955459  0.79879116  0.07682840
##  [81]  0.43240817  0.18140817 -1.04455459  0.51848265  0.12327254
##  [86] -0.24238171  0.18261829  0.71596403  0.95623530 -0.42341895
##  [91]  0.84009967 -0.97938171  0.34427254  0.21106243  0.50996403
##  [96]  0.78709967 -0.04938171 -0.05441895 -0.10476470 -0.50193757
## [101] -1.24372746 -1.22993757 -0.01159183  0.23448265 -0.13190033
## [106]  0.24300127 -0.28472746  0.41979116  0.59079116 -0.21772746
## [111]  0.45075392  0.32113691 -0.49659183 -0.60459183 -1.83169022
## [116]  0.99440817  0.55996403  0.71279116 -0.87528332 -0.25320884
```

```
sum(gpa$ei)
```

```
## [1] 0.0000000000001474376
```

Yes the residuals do sum to zero.

b. Estimate σ^2 and σ . In what units is σ expressed?

$$\hat{\sigma}^2 = s^2 = MSE = \frac{SSE}{n-2} = \frac{\sum e_i^2}{n-2} = \frac{45.81761}{118} = 0.3882848$$

$$s = \sqrt{0.3882848} = 0.623125. \qquad \sigma \text{ is expressed in GPA.}$$

2.15. Refer to Airfreight breakage Problem 1.21.

b. Compute $\sum e_i^2$ and MSE. What is estimated by MSE?

$$\begin{aligned}\sum e_i^2 &= (16-14.2)^2+(9-10.2)^2+(17-18.2)^2+(12-10.2)^2+(22-22.2)^2+(13-14.2)^2+\\ &\qquad (8-10.2)^2+(15-14.2)^2+(19-18.2)^2+(11-10.2)^2\\ &= 3.24+1.44+1.44+3.24+0.04+1.44+4.84+0.64+0.64+0.64\\ &= 17.6\end{aligned}$$

$$MSE = \frac{\sum e_i^2}{n-2} = \frac{17.6}{10-2} = 2.2 \qquad \text{MSE is an unbiased estimator for } \sigma^2$$

2.15 Refer to the regression model $Y_i = \beta_0 + \epsilon_i$ in Exercise 1.30. Derive the least squares estimator of β_0 for this model.

$$\begin{aligned}Q &= \sum_{i=1}^{n}(Y_i-\beta_0-\beta_1 X_i)^2\\ &= \sum_{i=1}^{n}(Y_i-\beta_0)^2 \qquad \text{where } \beta_1 = 0\end{aligned}$$

$$\frac{\partial Q}{\partial \beta_0} = -2(Y_i-\beta_0)$$

Set equal to 0

$$\begin{aligned}0 &= -2\sum(Y_i-\beta_0)\\ 0 &= \sum Y_i - \sum \beta_0\\ n\beta_0 &= \sum Y_i\\ \beta_0 &= \bar{Y}\end{aligned}$$

1.36 Prove the result in (1.20) - that the sum of the residuals weighted by the fitted values is zero.

$$\begin{aligned}0 &= \sum_{i=1}^{n}\hat{Y}_i e_i\\ &= \sum_{i=1}^{n}(\beta_1 X_i+\beta_0)\,e_i\\ &= \sum_{i=1}^{n}\beta_1 X_i e_i + \beta_0 e_i\\ &= \underbrace{\beta_1\sum_{i=1}^{n}X_i e_i}_{0 \text{ from } (1.19)} + \underbrace{\beta_0\sum_{i=1}^{n}e_i}_{0 \text{ from } (1.17)}\\ &= 0\end{aligned}$$

1.25. Refer to Airfreight breakage Problem 1.21.

a. Because of changes in airline routes, shipments may have to be transferred more frequently than in the past. Estimate the mean breakage for the following numbers of transfers: $X = 2, 4$. Use separate 99 percent confidence intervals. Interpret your results.

$$X_h = 2; \qquad \hat{Y}_h = 2 \cdot 4 + 10.2 = 18.2$$

$$s^2\left(\hat{Y}_h\right) = 2.2\left[\frac{1}{10} + \frac{(2-1)^2}{\sum(X_i - 1)^2}\right] = 0.44; \qquad s\left(\hat{Y}_h\right) = \sqrt{.44}$$

99 percent Confidence Interval:

$$18.2 - 3.3553 \cdot \sqrt{.44} \leq E\left[Y_h\right] \leq 18.2 + 3.3553 \cdot \sqrt{.44}$$

$$15.97454 \leq E\left[Y_h\right] \leq 20.42546$$

The amount of ampules to be broken upon arrival if there is 2 shipments is expected to be between 21 and 31.

$$X_h = 4; \qquad \hat{Y}_h = 4 \cdot 4 + 10.2 = 26.2$$

$$s^2\left(\hat{Y}_h\right) = 2.2\left[\frac{1}{10} + \frac{(4-1)^2}{\sum(X_i - 1)^2} = 2.2\right]; \qquad s\left(\hat{Y}_h\right) = \sqrt{2.2}$$

99 percent Confidence Interval:

$$26.2 - 3.3553 \cdot \sqrt{2.2} \leq E\left[Y_h\right] \leq 26.2 + 3.3553 \cdot \sqrt{2.2}$$

$$21.22316 \leq E\left[Y_h\right] \leq 31.17684$$

The amount of ampules to be broken upon arrival if there is 4 shipments is expected to be between 21 and 31.

b. The next shipment will entail two transfers. Obtain a 99 percent prediction interval for the number of broken ampules for this shipment. Interpret your prediction interval.

$$X_h = 2; \qquad \hat{Y}_{h(\text{new})} = 2 \cdot 4 + 10.2 = 18.2$$

$$s^2\left(\text{pred}\right) = 2.2\left[1 + \frac{1}{10} + \frac{(2-1)^2}{\sum(X_i - 1)^2}\right] = 2.64; \qquad s\left(\hat{Y}_{h(\text{new})}\right) = \sqrt{2.64}$$

99 percent Prediction Interval:

$$18.2 - 3.3553 \cdot \sqrt{2.64} \leq Y_{h(\text{new})} \leq 18.2 + 3.3553 \cdot \sqrt{2.64}$$

$$12.74814 \leq Y_{h(\text{new})} \leq 23.65186$$

c. In the next several days, three independent shipments will be made, each entailing two transfers. Obtain a 99 percent prediction interval for the mean number of ampules broken in the three shipments. Convert this interval into a 99 percent prediction interval for the total number of ampules broken in the three shipments.

$$X_h = 2; \qquad m = 3; \qquad \hat{Y}_{h(\text{new})} = 2 \cdot 4 + 10.2 = 18.2$$

$$s^2\left(\text{predmean}\right) = 2.2\left[\frac{1}{3} + \frac{1}{10} + \frac{(2-1)^2}{\sum(X_i - 1)^2}\right] = 1.173333; \qquad s\left(\hat{Y}_{h(\text{new})}\right) = 1.083205$$

99 percent Multiple Prediction Interval:

$$18.2 - 3.3553 \cdot 1.083205 \leq \bar{Y}_{h(\text{new})} \leq 18.2 + 3.3553 \cdot 1.083205$$

$$14.56543 \leq \bar{Y}_{h(\text{new})} \leq 21.83457$$

99 percent Total Prediction Interval:

$$14.56543 \cdot 3 \leq T_{h(\text{new})} \leq 21.83457 \cdot 3$$

$$43.701 \leq T_{h(\text{new})} \leq 65.50371$$

e. Determine the boundary values of the 99 percent confidence band for the regression line when $X_1 = 2$ and when $X_1 = 4$. Is your confidence band wider at these two points than the corresponding confidence intervals in part (a)? Should it be?

$$X_h = 2; \qquad \hat{Y}_h = 2 \cdot 4 + 10.2 = 18.2$$

$$s^2\left(\hat{Y}_h\right) = 2.2\left[\frac{1}{10} + \frac{(2-1)^2}{\sum(X_i - 1)^2}\right] = 0.44; \qquad s\left(\hat{Y}_h\right) = \sqrt{.44}$$

$$W^2 = 2F(.99, 2, 8) = 2 \cdot 8.649111 = 17.29822; \qquad W = 4.159113$$

Confidence band for the regression line:

$$18.2 - 4.159113 * \sqrt{.44} \leq \beta_0 + \beta_1 X_h \leq 18.2 + 4.159113 * \sqrt{.44}$$

$$15.44116 \leq \beta_0 + \beta_1 X_h \leq 20.95884$$

$$X_h = 4; \qquad \hat{Y}_h = 4 \cdot 4 + 10.2 = .2$$

$$s^2\left(\hat{Y}_h\right) = 2.2\left[\frac{1}{10} + \frac{(2-1)^2}{\sum(X_i - 1)^2}\right] = 2.2; \qquad s\left(\hat{Y}_h\right) = \sqrt{2.2}$$

$$W^2 = 2F(.99, 2, 8) = 2 \cdot 8.649111 = 17.29822; \qquad W = 4.159113$$

Confidence band for the regression line:

$$26.2 - 4.159113 * \sqrt{2.2} \leq \beta_0 + \beta_1 X_h \leq 26.2 + 4.159113 * \sqrt{2.2}$$

$$20.03104 \leq \beta_0 + \beta_1 X_h \leq 32.36896$$

YES IT IS WIDER...