3,4 due Monday 12th

# ST3451: Problem set 1

October, 2015

**Problems 3 and 4 due at the 5pm class on Monday 12th October. No late assignments will be accepted.**

1. Consider a population of $Y$ values

$$3.4, 10.6, 9.7, 5.4, 6.3, 12.1, 9.2, 4.7, 2.9, 8.5$$

   (a) Compute the mean and variance of $Y$ ($\mu$ and $\sigma^2$).

   (b) If we write $Y = \mu + \epsilon$

      i. What values does $\epsilon$ have?

      ii. Calculate the mean and variance of $\varepsilon$.

   (c) If we write $Y = 5 + \epsilon^*$

      i. What values does $\epsilon^*$ have?

      ii. Calculate the mean and variance of $\epsilon^*$.

2. Given the sample $Y_1, Y_2, \ldots, Y_n$ find the least squares estimator of $\alpha$ in the model

$$Y_i = \alpha + \epsilon_i, \qquad i = 1, \ldots, n$$

where $\mathrm{E}\{\epsilon_i\} = 0$, $\mathrm{var}\{\epsilon_i\} = \sigma^2$ and $\mathrm{corr}\{\epsilon_i, \epsilon_j\} = 0, i \neq j$.

3. Consider the simple linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \qquad i = 1, 2, \ldots, n$$

where $\mathrm{E}\{\epsilon_i\} = 0$, $\mathrm{var}\{\epsilon_i\} = \sigma^2$ and the $\epsilon_i$ are uncorrelated.

   (a) If $\beta_1$ is known, find the least squares estimator of $\beta_0$.

   (b) If $\beta_0$ is known, find the least squares estimator of $\beta_1$.

   (c) Do the answers to (a) and (b) seem reasonable? Why?

4. The reduction in blood pressure ($Y$) caused by an blood pressure drug was measured at each of a number of doses ($X$)

| Reduction $Y$ | Dose $X$ |
|---|---|
| 4.2 | 10 |
| 6.8 | 20 |
| 5.2 | 30 |
| 8.4 | 40 |
| 5.9 | 50 |
| 10.4 | 60 |
| 9.1 | 70 |
| 12.4 | 80 |
| 9.4 | 90 |
| 12.8 | 100 |
| 12.9 | 110 |
| 16.2 | 120 |
| 11.7 | 130 |
| 12.9 | 140 |
| 14.3 | 150 |

(a) Does the mean value of $Y$ depend on $X$?

(b) Fit a simple linear regression model to these data.

(c) Interpret the estimated parameters of the model in context.

(d) Find an estimate of the error about the regression line.

5. Show that the least squares estimators of $\widehat{\beta}_0, \widehat{\beta}_1$ can be written in the form

$$\widehat{\beta}_1 = \sum_{i=1}^{n} c_i Y_i \qquad \widehat{\beta}_0 = \sum_{i=1}^{n} d_i Y_i$$

giving the explicit $c_i, d_i$.

6. Write down a simple linear regression model forces the fitted line to go through the origin. Derive the least squares estimator for the parameter of your model. Use the extra sum of squares method to decompose the variance for this model.

7. (*) Suppose that the errors of the SLR model are assumed to be normally distributed. Find the maximum likelihood estimators of $\beta_0$ and $\beta_1$.

8. (*) Suppose that the heteroscedasticity assumption does not hold, so that $\text{var}\{Y_i\} = f_i\sigma^2$. Suggest a suitable least squares criterion in this case, and derive the form of the corresponding least squares estimators.

David Weisberger 1

15 14/10/15 ALSM1 PROBLEM SHEET 1

Q1 Population of y values

A. Mean of $y = \Sigma y_i/n = 7.28$   $\Sigma y = 728$   $\Sigma y_i^2 = 6206$

Variance $\sigma^2 = \frac{\Sigma(y_i-\mu)^2}{n} = \frac{\Sigma y_i^2 - n\mu^2}{n} = 9.068$

B. $y = \mu + \varepsilon$   $\varepsilon = y - \mu$

$\mu = \Sigma\varepsilon_i/n = 4.88\times10^{-4} \approx 0$

$\sigma_\varepsilon^2 = \frac{\Sigma\varepsilon_i^2 - n\mu_\varepsilon^2}{n} = 9.068$

$\sigma_\varepsilon^2 = \frac{1}{n}\left[\Sigma(y_i-\mu)^2 - n\left(\frac{\Sigma y_i - \mu}{n}\right)\right]$

$\Sigma(y_i-\mu) = \Sigma y_i - n\mu = n\mu - n\mu = 0$

C $y = 5 + \varepsilon^*$   $\varepsilon^* = y - 5$

$\mu_{\varepsilon^*} = \frac{\Sigma\varepsilon^*}{n} = 2.28$

$\sigma_{\varepsilon^*}^2 = \frac{\Sigma(\varepsilon_i^* - \mu_{\varepsilon^*})^2}{n} = 9.068$

Q2 Done in class

Q3 $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$   $E[\varepsilon_i] = 0$, $Var[\varepsilon_i] = \sigma^2$   $Cov[\varepsilon_i, \varepsilon_j] = 0 \quad i \neq j$

A. If $\beta_1$ is known, what is $\hat{\beta}_0$?

$\hat{y}_i = \hat{\beta}_0 + \beta_1 x_i$

LS: $Q(\hat{\beta}_0) = \sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \Sigma(y_i - \hat{y}_i)^2 = \Sigma(y_i - \hat{\beta}_0 - \beta_1 x_i)^2$

$\frac{dQ}{d\beta_0} = -2\Sigma(y_i - \hat{\beta}_0 - \beta_1 x_i) = 0$

$\Sigma y_i - n\hat{\beta}_0 - \beta_1\Sigma x_i = 0$

$\hat{\beta}_0 = \bar{y} - \beta_1\bar{x}$

Compare with calculation for $\hat{\beta}_0, \hat{\beta}_1$ $\Rightarrow$ the same $\Rightarrow$ reasonable

B If $\beta_0$ is known, what is $\hat{\beta}_1$? $\qquad \hat{y}_i = \beta_0 + \hat{\beta}_1 x_i$

$$\frac{dQ}{d\beta_1} = 2 \sum (y_i - \beta_0 - \beta_1 x_i)(-(x_i)) = 0$$

$$\sum x_i y_i - \beta_0 \sum x_i - \hat{\beta}_1 \sum x_i = 0$$

$$\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2} - \frac{\beta_0 \sum x_i}{\sum x_i^2}$$

Reasonable? Usually $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$ $\Rightarrow$ by knowing $\beta_0$ we knowing $\beta_0$ we know $y|x=0$

replaces sample mean of $y$ - seems a reasonable estimate

$\uparrow$

- $Var[\hat{\beta}_1]_F = \sigma^2 / S_{xx} \geq Var[\hat{\beta}_1]$
- We'd expect a smaller variance of the least squares estimator than when both $\beta_0$ and $\beta_1$ are estimated
- Unbiased if expectation $= \beta_1$

Unbiased? $\mathbb{E}[\hat{\beta}_1] = \mathbb{E}\left[\frac{\sum x_i y_i}{\sum x_i^2} - \frac{\beta_0 \sum x_i}{\sum x_i^2}\right]$ $\qquad \beta_0 = \bar{y} - \beta_1 \bar{x}$

$$= \mathbb{E}\left[\frac{\sum x_i y_i}{\sum x_i^2} - \frac{(\bar{y} - \beta_1 \bar{x}) \sum x_i}{\sum x_i^2}\right]$$

$$= \frac{1}{\sum x_i^2} \mathbb{E}\left[\sum x_i y_i - \beta_0 \sum x_i\right]$$

$$= \frac{1}{\sum x_i^2} \left[\mathbb{E}\left[\sum x_i y_i\right] - \beta_0 \mathbb{E}\left[\sum x_i\right]\right]$$

$$= 1/\sum x_i^2 \left[\sum x_i (\beta_0 + \beta_1 x_i) - \beta_0 \sum x_i\right]$$

$$= 1/\sum x_i^2 \left[\beta_0 \sum x_i + \beta_1 \sum x_i^2 - \beta_0 \sum x_i\right]$$

$$= \beta_1 \Rightarrow \text{unbiased estimator}$$

$$Var[\beta_1] = Var\left[\frac{\sum x_i y_i - \beta_0 \sum x_i}{\sum x_i^2}\right]$$

$$= \left(\frac{1}{\sum x_i}\right)^2 Var\left[\sum x_i y_i - \beta_0 \sum x_i\right]$$

$$= \frac{1}{(\sum x_i)^2} Var\left[\sum x_i y_i\right]$$

$$= \frac{1}{(\sum x_i^2)^2} \left[\sum Var[x_i y_i] + 2 \sum_{i=1}^{n-1} \sum_{j>i} Cov(x_i y_i, x_j y_j)\right]$$

$$= 1/(\sum x_i^2)^2 \left[\sum x_i^2 Var[y_i]\right]$$

$$= \sigma^2 / \sum x_i^2$$

Comparison: $S_{xx} = (x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2 \Rightarrow \sum x_i^2 = S_{xx} + n\bar{x}^2$

$$\frac{\sigma^2}{\sum x_i^2} \leq \frac{\sigma^2}{S_{xx}}$$

DAVID WETHERELL 18/20
12200464

1

14/10/15 ALSM PROBLEM SHEET 1

Q5 Show $\hat{\beta}_0, \hat{\beta}_1$ can be written $\hat{\beta}_1 = \Sigma c_i y_i$ $\hat{\beta}_0 = \Sigma d_i y_i$
Linear estimators → linear combinations of the data

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{1}{S_{xx}} \Sigma (x_i - \bar{x}) y_i = \frac{\Sigma(x_i - \bar{x})}{S_{xx}} y_i = \Sigma c_i y_i$$

So $c_i = \frac{x_i - \bar{x}}{S_{xx}}$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \bar{y} = \frac{1}{n} \Sigma y_i$$

$$= \frac{1}{n} \Sigma y_i - (\Sigma c_i y_i) \bar{x}$$

$$= \Sigma \left(\frac{1}{n} - c_i \bar{x}\right) y_i = \Sigma (d_i) y_i$$

$$d_i = \Sigma \left(\frac{1}{n} - c_i \bar{x}\right)$$

$$= \frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{S_{xx}}$$

Q4 Reduction in blood pressure ($y$) for dosage ($x$) $n = 15$

A. Scatterplot ⇒ possible dependency

B $\Sigma x_i = 1200$ $\Sigma x_i^2 = 124\,000$ $\Sigma y_i = 152.6$ $\Sigma y_i^2 = 1724.86$ $\Sigma x_i y_i = 14179$

SLR: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i$

$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$ $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

$S_{xy} = 1971$ $S_{xx} = 28\,000$ $\hat{\beta}_1 = 0.07039$ $\hat{\beta}_0 = 4.542$

C Interpretation: The estimate mean reduction in BP at dose 0 is 4.54 (not meaningful) The estimated mean reduction in BP for every extra unit dose of the drug is 0.07039

D Estimate of error

$SSE = \Sigma (y_i - \hat{y}_i)^2$

$\hat{\sigma}^2 = MSE = SSE/n-2$

$SSE = S_{yy} - \hat{\beta}_1 S_{xy}$ SS(uncorrected) - SS(Reg)

$= 177 - 0.70(1974) = 38.67$ $\hat{\sigma}^2 = 38.67/15-2 = 2.97$

DANO WETBRECHT 1
12300644

18/20

ST3451 Problem Set 1

1 (a) mean = $\frac{3.4 + 10.6 + 9.7 + 5.4 + 6.3 + 12.1 + 9.2 + 4.7 + 2.9 + 8.5}{10} = \frac{72.8}{10}$

7.28.

$\frac{(3.4-7.28)^2 + (10.6-7.28)^2 \quad \ldots \quad + (8.5-7.28)^2}{10}$

$\frac{15.0544 + 11.0224 + 5.8564 + 3.5344 + 0.9604 + 23.2324 + 3.8044 + 6.6564 + 19.1844 + 1.4884}{10}$

$= \frac{90.670}{10} = 9.0676 = \sigma^2$

1 (b) $y = \mu + \varepsilon \qquad \varepsilon = y - \mu$

i = $-3.88, 3.32, 2.42, -1.88, -0.98, 4.82, 1.92, -2.38, -4.38, 1.22$

ii. mean = $\Sigma \uparrow / 10 = 2.220$

variance = 9.0676

c (c). $\varepsilon^* = y - 5$

$-1.6, 5.6, 4.7, 0.4, 1.3$
$7.1, 4.2, -0.3, -2.1, 4.5$

mean = $\Sigma \uparrow / 10 = 2.38$

varian = $\frac{\Sigma (y_i - \mu)^2}{n} = \frac{94.016}{10} = 9.4016$

2. $y_i = \alpha + \varepsilon_i$ $\qquad \varepsilon_i = y_i - \alpha$

Minimise $\sum \varepsilon_i^2$

minimise $\sum (y_i - \alpha)^2$

$$\frac{d\left(\sum (y_i - \alpha)^2\right)}{d(\alpha)}$$

$$-2 \sum (y_i - \alpha) = 0$$

$$\sum (y_i - \alpha) = 0$$

$$\sum y_i - \sum \alpha = 0$$

$$\sum y_i - n\alpha = 0$$

$$n\alpha = \sum y_i$$

$$\alpha = \frac{\sum y_i}{n}$$

$$\alpha = \bar{y}$$

3(a) $\beta_1$ is known Find $\beta_0$

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ $\qquad \varepsilon_i = y_i - \beta_0 + \beta_1 x_i$

Minimise $\sum \varepsilon_i^2$

$$\frac{d \sum (y_i - \beta_0 - \beta_1 x_i)^2}{d(\beta_0)}$$

$$-2 \sum (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\sum (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\sum y_i - n\beta_0 - \beta_1 \sum x_i = 0$$

$$n\beta_0 = \sum y_i - \beta_1 \sum x_i$$

$$\beta_0 = \frac{\sum y_i}{n} - \beta_1 \frac{\sum x_i}{n}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

APPLIED LINE
ST3451
3(b) If $\beta_0$ is k
Minimise $\sum \varepsilon$
10

DAVID WEMBRECHT
12300644


3(b) If $\beta_0$ is known, find $\beta_1$

minimise $\sum \varepsilon_i^2$

min. $\sum \varepsilon_i^2$

$$\frac{d\left(\sum (y_i - \beta_0 - \beta_1 x_i)^2\right)}{d(\beta_1)}$$

$$-2 \sum (y_i - \beta_0 - \beta_1 x_i)(x_i) = 0.$$

$$\sum (y_i x_i - \beta_0 x_i - \beta_1 x_i x_i) = 0$$

$$\sum x_i y_i - \beta_0 \sum x_i - \beta_1 \sum x_i^2 = 0$$

$$\beta_1 \sum x_i^2 = \sum x_i y_i - \beta_0 \sum x_i$$

$$\beta_1 = \frac{\sum x_i y_i}{\sum x_i^2} - \beta_0 \frac{\sum x_i}{\sum x_i^2}$$

10

(c) yes they are as expected, they are the exact same answers as if we didn't know $\beta_1$ ~~or~~ $\beta_0$ etc
Answers seem reasonable

4(a) yes the mean value of y does depend on X. By plotting the data, ~~therefore~~ it is a reasonable assumption that the data is correlated, i.e that the mean value of y depend on X.

b. $\sum x_i = 1200$ $\quad \sum x_i^2 = 124\,000$

$\sum y_i = 152.6$ $\quad \sum y_i^2 = 1729.86$

$\bar{x} = 80$ $\quad \sum x_i y_i = \cancel{2764300}$ 14174

$\bar{y} = 10.173$

$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$ $S_{xy} = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n}$

$14179 - \frac{(1200)(152.6)}{15}$ ~~2764352~~

$=$ ~~2757400~~ $1971$

$S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$ $124000 - \frac{(1200)^2}{15} = 28000$

$\hat{\beta}_1 = \frac{\cancel{2757400}\ 1971}{28000} = \cancel{98.49640} = \cancel{0.07034}\ -33.0247$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

$80 - \cancel{0.07034}\ (-33.0247)(10.1733) = \cancel{79.28}\ 11.109$

$y_i = \cancel{79.28 + 0.07039 x_i} = -33.025 + 11.1 x_i$

c When the dose in 0 the reduction is ~~79.28~~ $-33.020$ which has no meaning in this situation

For each unit increase in the dose, the mean value of the reduction will increase by 11.1.

D Calculate the SSE given by: $\sum(y_i - \hat{y})^2 = \sum(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$

$= S_{yy} - \frac{S_{xy}^2}{S_{xx}}$

8

$S_{yy} = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$ $124000 - \frac{(1200)^2}{15} = 28000.$

$S_{xx} = \sum x^2 - \frac{(\sum x_i)^2}{n} = 1729.86 - \frac{(152.6)^2}{15} = 177.409$

$S_{xy}$. $\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}$ $14179 - \frac{(152.6)(1200)}{15} = 1971$

$= 28000 - \frac{1971}{177.409} = 27988.89$