CSCI E-106: Data Modeling

Assignment 2

Due: February, 11 2019 at 7:19 pm EST

Instructions: Students should submit their reports on Canvas. The report needs to clearly state what question is being solved, step-by-step walk-through solutions, and final answers clearly indicated. Please solve by hand where appropriate.

Please submit either scanned hand-written solution or typed solutions and two files: (1) a R Markdown file (.Rmd extension) and (2) a PDF document generated using knitr for .Rmd file submitted in (1) where appropriate. Please, use RStudio Cloud for your solutions.

All questions are coming from Kutner, M. *et al*: Applied Linear Statistical Models, Fifth Edition.

Problem 1.22 done on R, shown on Knitr. Some calculations of 2.8 done on R, also on Knitr.

1. (1.22)
2. (2.8)
3. (2.16)
4. (2.51)
5. (2.53)

Yinan Kang

2.16

a) 98% CI for mean hardness when elapsed time = 30 hr.

Confidence Interval: $\hat{Y}_h \pm t(\frac{1-\alpha}{2}; n-2)\, s\{\hat{Y}_h\}$ Eqn. 2.33

$\hat{Y}_{30} \pm t(.98; 14)\, s\{\hat{Y}_h\}$

$s^2\{\hat{Y}_h\} = MSE\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right]$ Eqn. 2.30 $X_h = 30$

MSE = 9.15

$\bar{X} = 28$

$\sum(X_i - \bar{X})^2 = 1280$

→ calculated on R, included in Knitr PDF

$$s^2\{\hat{Y}_{h=30}\} = 9.15\left[\frac{1}{16} + \frac{(30-28)^2}{1280}\right] = 0.60$$

$$s\{\hat{Y}_{h=30}\} = 0.77$$

$$t(0.98; 14) = 2.264$$

$\hat{Y}_{h=30} \pm 2.264(0.77)$ ... $\hat{Y}_{h=30} = 2.034(30) + 168.6 = 229.62$ ← from 1.22

CI: 229.62 ± 1.74

$$\boxed{227.88 \le E\{Y_{30}\} \le 231.36}$$

→ We conclude with 98% confidence the mean hardness after 30 elapsed hours is between 227.88 and 231.36 Brinell units.

b) Prediction Interval: $\hat{Y}_h \pm t(\frac{1-\alpha}{2}; n-2) \times s\{pred\}$

$\hat{Y}_{30} = 229.62$

$s^2\{pred\} = MSE\left[1 + \frac{1}{n} + \frac{(X_{30} - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}\right]$

$= 9.15\left[1 + \frac{1}{16} + \frac{(30-28)^2}{1280}\right] = 9.75$

$s\{pred\} = 3.12$

PI: $229.62 \pm 2.264(\cancel{9.75}\ 3.12)$

$\boxed{PI: 222.56 \le Y_{h(new)} \le 236.68}$ @ 98% confidence hardness of new item is in this range.

c) $s^2\{predmean\} = MSE\left[\frac{1}{m} + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}\right]$ 2.396

$\hat{Y}_h \pm t(\frac{1-\alpha}{2}, n-2)\, s\{predmean\}$

$s^2\{predmean\} = 9.15\left[\frac{1}{10} + \frac{1}{16} + \frac{4}{1280}\right] = \cancel{0.5941}$ 1.515

$s\{predmean\} = 1.231$

$229.62 \pm 2.264(1.231)$

$\boxed{PI: 226.833 \le \bar{Y}_{h(new)} \le 232.407}$ 98% confident mean of 10 newly molded items have mean hardness in this range.

d) Interval in (c) is narrower than (b), as (c) predicts a mean which has lesser variance than a one-time prediction.

(e) Confidence band: $\hat{Y}_h \pm W s\{\hat{Y}_h\}$

$W^2 = 2F(1-\alpha; 2, n-2) = 2F(.98; 2, 14)$
$= 10.48$

$W = 3.24$, $s\{\hat{Y}_h\} = 0.77$ (from (a))

$\boxed{CB: 229.62 \pm 3.24(0.77) = 227.13 \le \beta_0 + \beta_1 X_h \le 232.11}$

Yes, this is wider than CI in (a). This should be, as confidence band applies to entire regression, not just one X-value.

Scanned by CamScanner

2.51 $b_0 = \bar{Y} - b_1 \bar{X}$

$$E[b_0] = E[\bar{Y} - b_1 \bar{X}]$$

$$= E(\bar{Y}) - \bar{X} E(b_1)$$

$$E(\bar{Y}) = E\left[\frac{1}{n} \Sigma Y_i\right]$$

$$E(b_1) = \beta_1$$ ← from Lecture + Notes

$\left(\Sigma = \sum_{i=1}^{n}\right)$

$$= \frac{1}{n} \Sigma E(y_i)$$

$$= \frac{1}{n} \Sigma(\beta_0 + \beta_1 x_i)$$

$$= \beta_0 + \beta_1 \bar{x}$$

$$E[b_0] = E(\bar{Y}) - \bar{X} E(b_1)$$

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

$$\rightarrow E[b_0] = \beta_0$$

2.53

a) $L(\beta_0, \beta_1, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2\right]$

where X_i are independent random variables.

b) The estimators β_0 and β_1 of the MLE are not the same as those in 1.27 as the conditions require that $g(X_i)$ does not involve β_0 or β_1 (or σ^2).

probability distribution

(Not sure how to derive the MLEs)

Assignment2

Yinan Kang

2/10/2019

R Markdown

Problem 1.22

(a)

```
batches.df <- data.frame(X = c(16,16,16,16,24,24,24,24,32,32,32,32,40,40,40,40),
                 Y = c(199,205,196,200,218,220,215,223,237,234,235,230,250,248,253,246))
head(batches.df,3)
```

```
##    X   Y
## 1 16 199
## 2 16 205
## 3 16 196
```

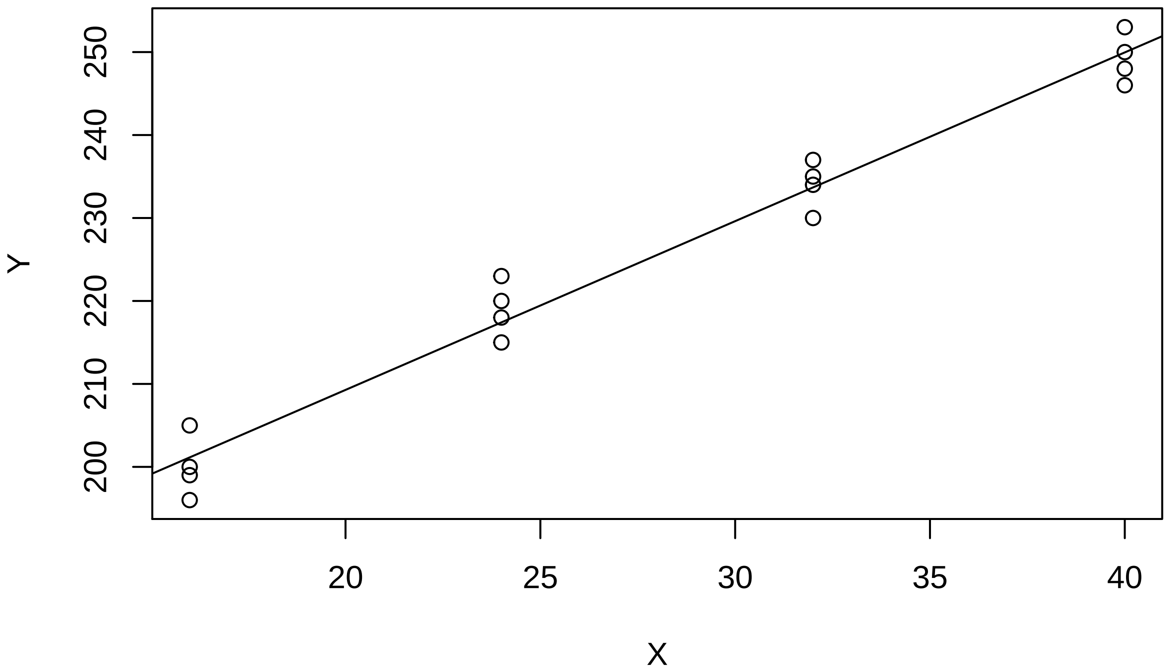
```
set.seed(258)

batches.lm <- lm(batches.df$Y ~ batches.df$X)
batches.lm
```

```
##
## Call:
## lm(formula = batches.df$Y ~ batches.df$X)
##
## Coefficients:
##  (Intercept)  batches.df$X
##      168.600         2.034
```

The estimated linear regression function is Y = 2.034*X + 168.6.

```
plot(batches.df)
abline(168.6, 2.034)
```



Yes, the regression function looks like a good fit here as the values tend to be clustered near the line.

(b)

```
#Using X = 40:
Y.40 <- 2.034*40 + 168.6
print(Y.40)
```

```
## [1] 249.96
```

Point estimate at X = 40 is 249.96

(c)

Point estimate of change in mean hardness when X increases by 1 is equal to the regression coefficient, 2.034 Brinell units

Problem 2.16

```
mean((batches.lm$residuals)^2)
```

```
## [1] 9.151562
```

```
mean(batches.df$X)
```

```
## [1] 28
```

```
sum.xtot <- 0
for (i in 1:nrow(batches.df)) {
  sum.xind <- (batches.df[i,1] - mean(batches.df$X))^2
  sum.xtot <- sum.xtot + sum.xind
```

```
}
```

MSE of linear regression in Problem 1.22 = 9.15
X-bar = 28 hr
sum((Xi - X-bar)^2) = 1280