Homework 11

Stat061-F23

Prof Amanda Luby

---

1. In some regression situations, it is appropriate to assume that the relationship being approximated should pass through the origin. If so, the resulting equation has the form $y = bx$.

   (a) Use the least squares criterion to find the formula for the slope in this case.
   (b) Luxury suites, many costing more than $100,000 to rent, have become big-budget status symbols in new sports arenas. The `arena_data` table below (data also created in .qmd document) gives the number of luxury suites and their projected revenues (in millions of USD) for nine US sports facilities. Explain why a no-intercept model may be appropriate for this setting, and find the equation for the least squares line $y = bx$.

Arena	Suites	Projected_Revenue
Palace (Detroit)	180	11.0
Orlando Arena	26	1.4
Bradley Center (Milwaukee)	68	3.0
America West (Phoenix)	88	6.0
Charlotte Coliseum	12	0.9
Target Center (Minneapolis)	67	4.0
Salt Lake City Arena	56	3.5
Miami Arena	18	1.4
ARCO Arena (Sacramento)	30	2.7

2. Show that the MLEs for β_0 and β_1 are unbiased

3. Using the simple linear regression estimates for $\hat{\beta}_0$ and $\hat{\beta}_1$, show each of the following:

   (a) The least squares line must pass through the point $(\bar{x}, \bar{y})$
   (b) $\sum(y_i - \hat{y}_i) = 0$
   (c) $\sum(\hat{y}_i - \bar{y})^2 = \hat{\beta}_1^2 \sum(x_i - \bar{x})^2$

4. A simple exponential decay model says that the concentration $C_{(t)}$ of a pesticide (for example) remaining after time t is $C_{(t)} = C_0 e^{-\gamma t}$ for $t > 0$, where C_0 is the initial concentration and γ is a constant that determines the rate of decay.

   (a) Show that this model can be transformed into a linear model for $Y = \ln(C_{(t)})$ on t.
   (b) Using the terms in the original exponential decay model, what are the slope and intercept for the linear model in (a)?
   (c) If you have data on concentrations at n different times t_i, you could estimate γ by fitting a simple linear regression of Y_i on t_i. This implicitly assumes an error term, ϵ_i, that is approximately normally distributed. Write out the implied model for $C_{(t)}$ and describes how error enters the model. What is the distribution of the error term in the model for $C_{(t)}$?

5. Each year, the Scottish Hill Runners Association publishes a list of hill races in Scotland for the year. The code below reads in the results for 68 races and includes `distance`: the distance of the

course in km; `climb`: the elevation of the course in km; `timeM`: the record time to complete the course (for registrants in the men's category); and `timeW`: the record time to complete the course (for registrants in the women's category). Your submitted work for this problem should include both (1) a nicely formatted output that is easy for the grading team to read, and (2) your code to support your answers.

```
races = tibble(read.table("http://stat4ds.rwth-aachen.de/data/ScotsRaces.dat",
                    header = TRUE))
races
```

```
# A tibble: 68 x 5
   race              distance climb timeM timeW
   <chr>                <dbl> <dbl> <dbl> <dbl>
 1 AnTeallach            10.6 1.06   74.7  89.7
 2 ArrocharAlps          25   2.4   187.  222.
 3 BaddinsgillRound      16.4 0.65   87.2 102.
 4 BeinnLee              10.2 0.26   41.6  52.5
 5 BeinnRatha            12   0.24   47.8  58.8
 6 BeinnResipol          12   0.845  68.2  81.4
 7 BenGullipen           13.5 0.35   50.0  56.7
 8 BenLedi               10   0.75   50.1  63.6
 9 BenLomond             12   0.97   62.3  72.0
10 BenNevis              14   1.34   85.6 103.
# i 58 more rows
```

(a) Make a scatterplot of `timeW` against `distance`. Does a linear model appear to be appropriate?
(b) Fit the least-squares regression and check the residual plots. Do you have any concerns about the fit of the model? If so, rectify it before moving to part (c).
(c) Find 90% intervals for:

i. The slope of the regression line
ii. The average winning time for a 60km race
iii. The predicted winning time for a 60km race