STA302/1001 Autumn 2017 Homework #1 Solutions

With thanks to Alison Gibbs and Becky Lin

1. Either the left hand side should be Y_i or the e_i should not be on the right hand side.

2. (a) The model would go through the origin.
   (b) Minimize $\sum_{i=1}^{n}(y_i - \hat{\beta}_1 x_i)^2$ by differentiating with respect to $\hat{\beta}_1$ and setting the derivative equal to 0. This gives $\hat{\beta}_1 = \sum y_i x_i / \sum x_i^2$.
   (c) The model is a horizontal line. The fitted model would be $\hat{y} = \overline{y}$.
   (d) Minimize $\sum_{i=1}^{n}(y_i - \hat{\beta}_0)^2$ by differentiating with respect to $\hat{\beta}_0$ and setting the derivative equal to 0. This gives $\hat{\beta}_0 = \overline{y}$. This is unbiased for β_0 since $\mathrm{E}(\overline{Y}) = \frac{1}{n}\sum \mathrm{E}(Y_i) = \beta_0$ since $\mathrm{E}(e_i) = 0$.

3. (a)

$$\begin{aligned}\sum \hat{e}_i x_i &= \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)\, x_i \\ &= \sum (y_i - \overline{y} + \hat{\beta}_1 \overline{x} - \hat{\beta}_1 x_i)\, x_i \\ &= \sum x_i y_i - n\overline{x}\overline{y} - \hat{\beta}_1 SXX \\ &= 0\end{aligned}$$

   using $\hat{\beta}_1 = (\sum x_i y_i - n\overline{x}\overline{y})/SXX$.
   (b) $\sum \hat{e}_i \hat{y}_i = \sum (\hat{\beta}_0 + \hat{\beta}_1 x_i)\, \hat{e}_i = \hat{\beta}_0 \sum \hat{e}_i + \hat{\beta}_1 \sum x_i \hat{e}_i = 0$ using $\sum \hat{e}_i = 0$ and the result in part (a).

4. (a) When $x = 0$: $N(10, 4)$
   When $x = 5$: $N(35, 4)$
   (b) When $x = 2$, the conditional distribution of Y is $N(20, 4)$ and the probability that it is between 16 and 20 is $\Phi\left(\frac{20-20}{2}\right) - \Phi\left(\frac{16-20}{2}\right) = 0.477$ where Φ is the standard normal cumulative distribution function. (You should be able to approximate this (perhaps as 0.475) by knowing standard properties of the normal distribution.)

5. This could be an example of the regression effect where employees who did well before the training will tend to do worse, on average, the next time they are measured and employees who did poorly before the training will tend to do better, on average, the next time they are measured. However, the cut-off point (where the regression line crosses the line $y = x$) is at $x = 400$. But for this situation, x ranges from 40 to 100. Therefore, on average, employees did better after the training. The slope of 0.95 is not the whole story!

6. The slope is not statistically significantly different from 0. So the correct conclusions is that there is no evidence of a linear relationship between advertising expenditures and sales.

7. (a)

```r
fit = lm(eruption~waiting,data=q2data)
```

(b) The summary output from R:

```r
summary(fit)
```

```
##
## Call:
## lm(formula = eruption ~ waiting, data = q2data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.29917 -0.37689  0.03508  0.34909  1.19329
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.874016   0.160143  -11.70   <2e-16 ***
## waiting      0.075628   0.002219   34.09   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4965 on 270 degrees of freedom
## Multiple R-squared:  0.8115, Adjusted R-squared:  0.8108
## F-statistic:  1162 on 1 and 270 DF,  p-value: < 2.2e-16
```

(c) Estimates: $b_0 = -1.874016$ and $b_1 = 0.075628 waiting$.

8. (b)

$$\begin{aligned}\sum_{i}^{n}(X_i - \bar{X})^2 &= \sum_{i=1}^{n}(X_i^2 - 2X_i\bar{X} - \bar{X}^2) \\ &= \sum_{i=1}^{n} X_i^2 - 2\bar{X}\sum X_i - n\bar{X}^2 \\ &= \sum_{i=1}^{n} X_i^2 - 2n\bar{X}^2 - n\bar{X}^2 \\ &= \sum_{i=1}^{n} X_i^2 - n\bar{X}^2\end{aligned}$$

(c)

$$
\begin{aligned}
\sum_{i}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) &= \sum_{i=1}^{n}(X_i - \bar{X})Y_i - \bar{Y}\sum(X_i - \bar{X}) \\
&= \sum_{i=1}^{n}(X_iY_i - \bar{X}Y_i) - 0 \quad \text{by (a)} \\
&= \sum_{i=1}^{n} X_iY_i\bar{X}\sum_{i=1}^{n} Y_i \\
&= \sum_{i=1}^{n} X_iY_i - n\bar{X}\bar{Y}
\end{aligned}
$$