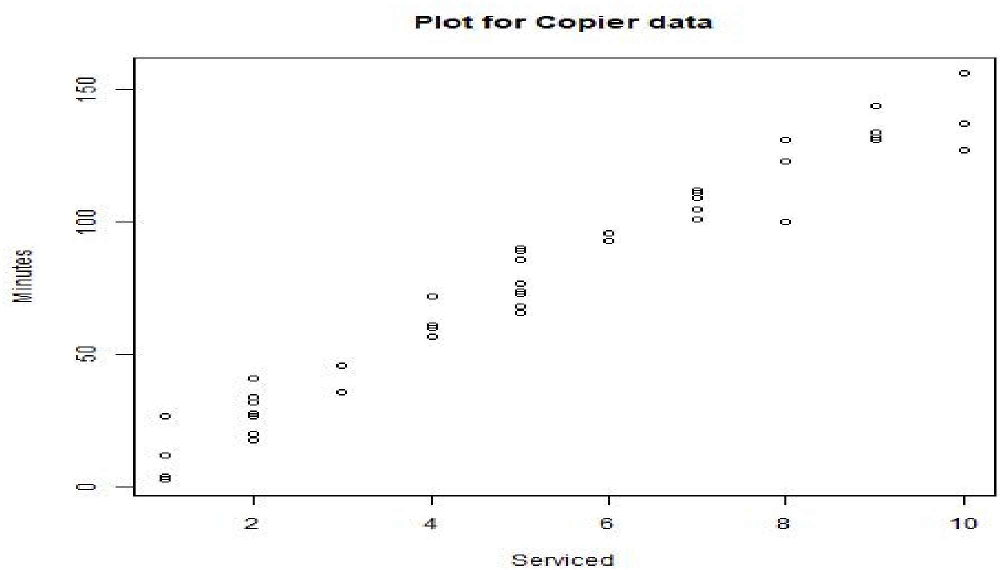
Stat 6021: Homework Set 3 Solutions

1. (a) The response variable is *Minutes*, the total time taken by the service person, and the predictor is *Serviced*, the number of copiers serviced.

   (b) The scatterplot is shown below. We can see there is a strong positive linear association between the total time taken by the service person and the number of copiers serviced.



   (c) The values are

   - $\hat{\beta}_1 = 15.0352$
   - $\hat{\beta}_0 = -0.5802$
   - $R^2 = 0.9575$
   - $\hat{\sigma}^2 = 8.914^2 = 79.4594$

   (d) For each additional copier serviced, the predicted service time increases by 15.0352 minutes .

   When the number of copiers serviced is 0, the predicted service time is -0.5802 minutes. The intercept makes no sense in context because service time cannot be negative. (This is a by product of extrapolation)

(e) The ANOVA F statistic is 968.66. The hypotheses are $H_0 : \beta_1 = 0, H_a : \beta_1 \neq 0$. Since the p-value is small, we reject the null hypothesis. The data supports the claim that there is a linear association between the total service time and the number of copiers serviced.

Alternatively, using the critical value approach. Critical value is 4.07 (using qf(0.95,1,43) in R). Since the F-stat is greater than the critical value, we reject the null hypothesis. The data supports the claim that there is a linear association between the total service time and the number of copiers serviced.

2. (a) The table is displayed below.

x_i	70	75	80	80	85	90
y_i	75	82	80	86	90	91
$\hat{y}_i$	76	80	84	84	88	92
e_i	-1	2	-4	2	2	-1

(b)

	DF	SS	MS	F-stat	p-value
Regression	1	160	160	21.3333	0.0099
Residual	4	30	7.5	***	***
Total	5	190	***	***	***

$SS_{res} = \sum_i e_i^2 = (-1)^2 + 2^2 + (-4)^2 + 2^2 + 2^2 + (-1)^2 = 30.$

$$\begin{aligned} SS_T &= \sum_i (y_i - \bar{y})^2 \\ &= (75-84)^2 + (82-84)^2 + (80-84)^2 + (86-84)^2 + (90-84)^2 + (91-84)^2 \\ &= 190. \end{aligned}$$

(c) $\hat{\sigma^2} = \frac{SS_{res}}{n-2} = \frac{30}{4} = 7.5.$

(d) $R^2 = \frac{SS_{res}}{SS_T} = \frac{160}{190} = 0.842.$

(e) $H_0 : \beta_1 = 0, H_a : \beta_1 \neq 0$. Since the p-value is less than 0.05, we reject the null hypothesis. Our data supports the claim that there is a linear relationship between scores on the second quiz and scores on the first quiz.

Critica value approach: Critical value is 7.71 (using qf(0.95,1,4) in R). Since the F-stat is greater than the critical value, we reject the null hypothesis. Our data supports the claim that there is a linear relationship between scores on the second quiz and scores on the first quiz.

3. We want to find $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize

$$Q = SS_{res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Taking partial derivatives with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$ and setting them to 0, we obtain

$$
\begin{aligned}
\frac{\partial Q}{\partial \hat{\beta}_0} &= \sum_i 2\left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right)(-1) = 0, && (1)\\
\frac{\partial Q}{\partial \hat{\beta}_1} &= \sum_i 2\left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right)(-x_i) = 0. && (2)
\end{aligned}
$$

$$
\begin{aligned}
\sum_i e_i &= \sum_i y_i - \hat{y}_i\\
&= \sum_i y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\\
&= \sum_i y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_i x_i\\
&= n\bar{y} - n\bar{y} + n\hat{\beta}_1\bar{x} - n\hat{\beta}_1\bar{x} = 0 && (3)
\end{aligned}
$$

Alternatively, can use partial derivative (1) after the second line.

Sum / average of residuals is 0.

$$
\begin{aligned}
\sum_i \hat{y}_i &= \sum_i \hat{\beta}_0 + \hat{\beta}_1 x_i\\
&= n\hat{\beta}_0 + \hat{\beta}_1 n\bar{x}\\
&= n\left(\hat{\beta}_0 + \hat{\beta}_1\bar{x}\right)\\
&= n\bar{y} = \sum_i y_i && (4)
\end{aligned}
$$

Sum of fitted values is equal to sum of observed responses.

$$
\begin{aligned}
\sum_i x_i e_i &= \sum_i x_i (y_i - \hat{y}_i)\\
&= \sum_i x_i \left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right] = 0, \text{using equation (2).} && (5)
\end{aligned}
$$

Sum of residuals weighted by x_i is 0.

Common mistake in (5) will be to write $\sum_i x_i e_i = x_i \sum_i e_i = 0$ using (3). But cannot pull x_i out of the summation since it's not a constant.

$$
\begin{aligned}
\sum_i \hat{y}_i e_i &= \sum_i (\hat{\beta}_0 + \hat{\beta}_1 x_i) e_i\\
&= \hat{\beta}_0 \sum_i e_i + \hat{\beta}_1 \sum x_i e_i\\
&= 0, \text{using equations (3) and (5).} && (6)
\end{aligned}
$$

Sum of residuals weighted by fitted responses is 0.

Common mistake in (6) will be to write $\sum_i \hat{y}_i e_i = \hat{y}_i \sum_i e_i = 0$ using (3). But cannot pull $\hat{y}_i$ out of the summation since it's not a constant.