

Solutions to Exam—style question (2)

(a) The general form of the full regression model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon, \quad \epsilon \sim N(0, \sigma^2).$$

X_4 is the indicator variable.

(b) (i) The full regression equation is

$$Y = 1.1 + 0.443X_1 - 1.39X_2 - 0.258X_3 - 1.38X_4 + \epsilon,$$

where $\epsilon \sim N(0, 1.159^2)$.

(ii) Recall that

$$T = \frac{\hat{\beta}_i}{\text{s.e.}(\hat{\beta}_i)}.$$

Thus,

$$A = \frac{0.4433}{0.2869} = 1.545$$

$$B = \frac{-1.3936}{-2.03} = 0.6865$$

$$C = \frac{-0.2584}{0.7883} = -0.3278$$

and

$$D = \frac{-1.381}{-1.10} = 1.255.$$

(iii) Variable X_3 (Morale) should be removed, since this has the largest p -value. In fact, we retain both

$$H_0 : \beta_3 = 0 \quad \text{and} \quad (1)$$

$$H_0 : \beta_4 = 0, \quad (2)$$

though the p -value for (1) shows that β_3 is the *least* significant and so we remove X_3 first.

(iv) $R^2 = 95.3\%$, i.e. 95.3% of the variation in Y is explained by the full regression model.

- (c) (i) In the *final model* we are left with just X_1 (overtime hours) and X_2 (sick days). We discussed in part (b) that X_3 (Morale) should be removed from the model because β_3 had a large p -value in the initial fit. After re-fitting without X_3 , β_4 must have then had the largest p -value above 0.05, and so X_4 (“shut-down indicator”) was also removed, leaving us with the *final model* including only X_1 and X_2 . Both X_1 and X_2 in the *final model* have p -values less than 0.05, meaning we would reject the hypotheses

$$\begin{aligned} H_0 &: \beta_1 = 0 & \text{and} \\ H_0 &: \beta_2 = 0. \end{aligned}$$

Thus, only overtime hours and sick days are important indicators of productivity.

- (ii) We now have $R^2 = 94.2\%$ – this has reduced slightly from that under the *full model*, but only slightly. The simpler the model the better, and we should be prepared to accept a small reduction in R^2 if it means we can remove two predictor variables!
- (iii) The histogram of residuals shows that the assumption of Normality here is reasonable – the plot peaks at about zero and then slopes away fairly symmetrically around this. However, some might say that this plot suggest the residuals are positively skewed, which could indicate a problem. These graphs are open to interpretation, and should only be used as a guide.
- (iv) If $X_1 = 36$ and $X_2 = 3$, we have

$$\begin{aligned} Y &= -1.33 + 0.458 \times 36 - 1.53 \times 3 \\ &= 10.568, \end{aligned}$$

i.e. 10,568 units.