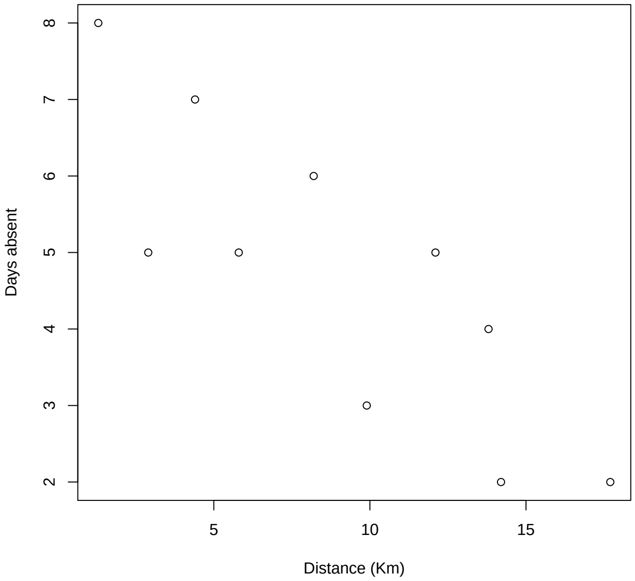
SMM637 Answers - Simple Linear Regression

1. (a) The distance from work is being used to predict number of days absent. Therefore ‘distance from work’ is the x variable and ‘days absent’ is the y variable.

   (b) Scatter plot:



   (c)

```
hosp.lm <- lm(absent ˜ distance, data=hospital)
summary(hosp.lm)

Call:
lm(formula = absent ˜ distance, data = hospital)

Residuals:
    Min      1Q  Median      3Q     Max
-1.5829 -1.0070  0.3641  0.9137  1.2430

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.47360    0.75328   9.921    9e-06 ***
distance    -0.30715    0.07248  -4.238  0.00285 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.179 on 8 degrees of freedom
Multiple R-squared:  0.6918,    Adjusted R-squared:  0.6533
F-statistic: 17.96 on 1 and 8 DF,  p-value: 0.002846
```

   So the fitted regression line is given by the equation

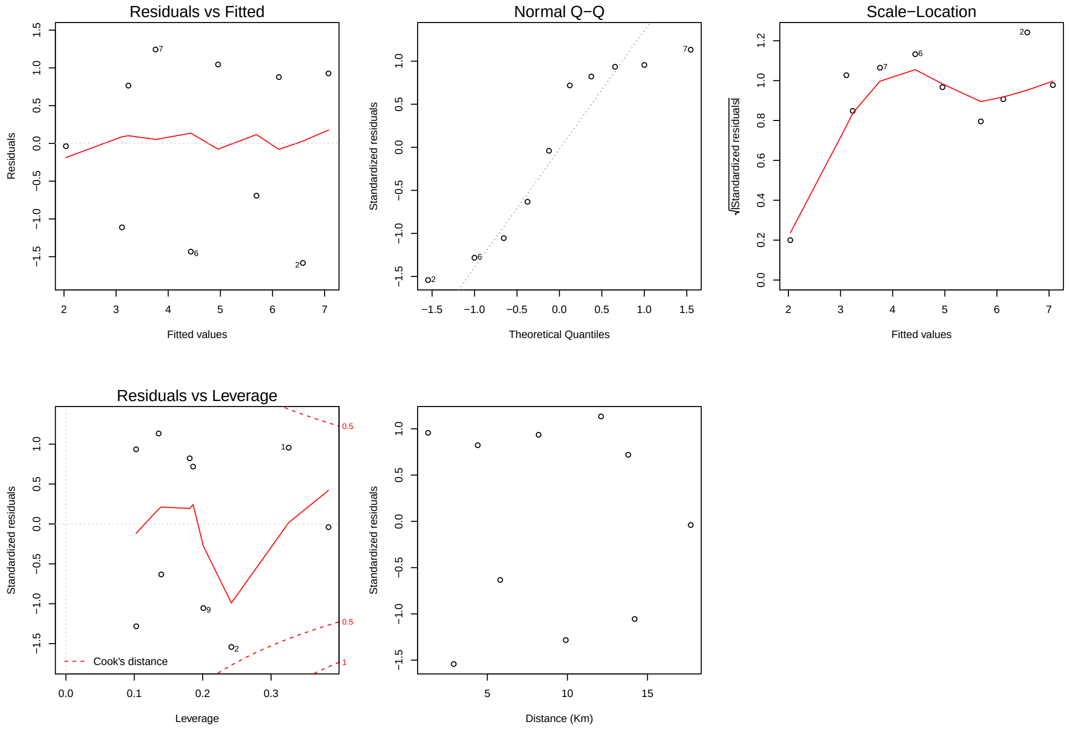
$$\hat{y} = 7.474 - 0.3072x$$

(d) On average, people living close to the hospital take more days off than people living some distance from the hospital.

(e) $\hat{y} = 7.4736 - 0.3072 \times 10 = 4.4$. Thus, 4-5 days.

(f) The same slope -0.3072 cannot continue for arbitrarily large distances. It is only valid for the range of the distance variable $(1.3, 17.7)$. If it continuous, then someone living 25 or more kilometres from the hospital will have a negative number of days absent. The relationship might cease to be linear.

(g) Residuals plots:



The top left plot of Residuals vs Fitted, does not show any indication of increasing variance with mean, which means that the constant variance assumption holds here. The other feature to look for is a pattern in the average value of the residuals as the fitted values change. The solid curve shows a running average of the residuals to help judging this: there no pattern here. The top right plot shows the square root of the absolute value of the standardized residuals against fitted value (again with a running average curve). If all is well the points should be evenly spread with respect the vertical axis here, with no trend in their average value. A trend in average value is indicative of a problem with the constant variance assumption, and, although there is a slight pattern, it is not a concern in this case. The top middle plots the ordered standardized residuals against quantiles of a standard normal: a systematic deviation from a straight line is slightly present here indicating a slight departure from normality in the residuals. However there are few data to get a full picture. The lower left plot is looking at leverage and influence of residuals, by plotting standardized residuals against a measure of leverage. Cook distances over 0.5 are considered borderline problematic, so no points warrant investigation. Finally, the last plot shows the standardized residuals against then explanatory variable. There are no obvious patterns against it, so that the adequacy of the linear model is not questioned.