STAT40720 Intro. to Data Analytics

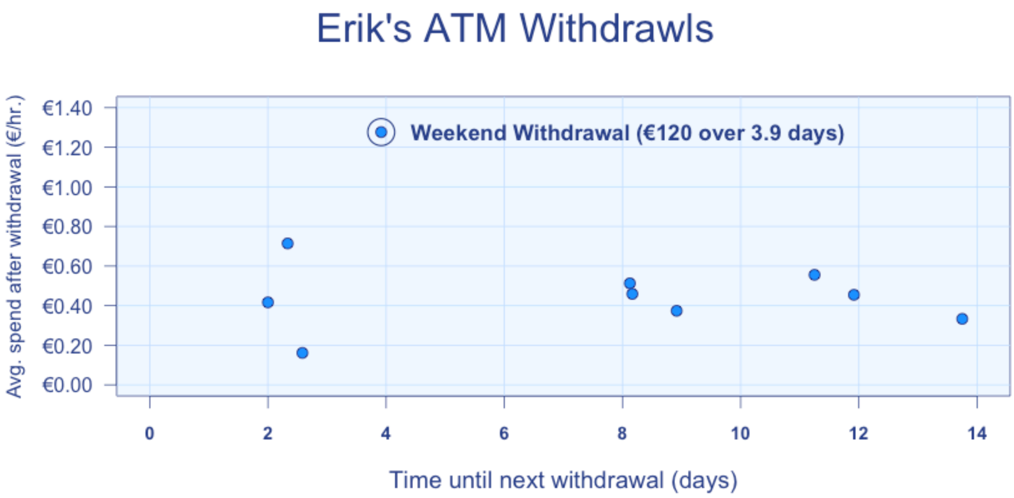
Assignment 4

Name	Kevin O'Connell
Student ID	########
Submission Date	16nd. November 2015

Question 1

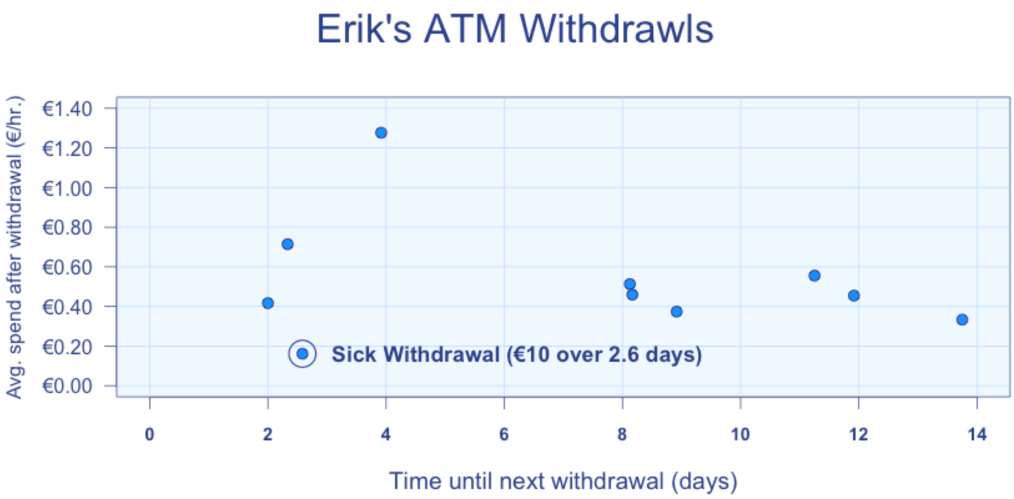
To complete the analysis, a plot of the average spend per hour versus time to next withdrawal was computed.

(a) Weekend Withdrawal



Erik’s withdrawal prior to the weekend away is likely to have the highest average spend over a period of 3 to 4 days. The most likely candidate is observation 5, where €120 was withdrawn and spent over 94 hours.

(b) Sick Withdrawal



Erik’s withdrawal prior to getting sick and confined to bed is likely to have the lowest average spend over a relatively short period of 2 to 3 days. The most likely candidate is observation 2, where €10 was withdrawn and spent over 62 hours.

Question 2

(a) The equation of the regression line

$$
\begin{aligned}
b &= \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} \\
&= \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2} \\
&= \frac{371}{202} \\
&= 1.8366
\end{aligned}
$$

To find the value of *a*, the mean value of x and y must be found.

$$
\begin{aligned}
\bar{x} &= \frac{\sum x}{n} = \frac{112}{7} = 16 \\
\bar{y} &= \frac{\sum y}{n} = \frac{224}{7} = 32 \\
\therefore a &= \bar{y} - 1.8366\bar{x} \\
&= 32 - 1.8366(16) \\
&= 2.6144
\end{aligned}
$$

The above regression shows from the figure *b*, that for each minute after 8:08am the teacher leaves, the journey time increases by 1.84 minutes.

The value of *a* shows and extrapolation back to 8:00am from this data would yield a journey time of 2.6144 minutes. This is, of course, not practical, but it can be used to show the expected minimum journey time for an 8:08am departure is given by:

$$
\begin{aligned}
y_{x=8} &= 2.6144 + 1.8366(8) \\
&= 17.3
\end{aligned}
$$

Therefore, the expected minimum travel time is 17.3 minutes.

(b) 8:15am Departure Time

$$
\begin{aligned}
y_{x=15} &= 2.6144 + 1.8366(15) \\
&= 30.1634
\end{aligned}
$$

For an 8:15am departure, the expected journey duration is 30.2 minutes.

(c) 8:15am Departure Time 95% Confidence Interval

$$\begin{aligned}
\sum y^2 &= \sum(y-\bar{y})^2 + \frac{(\sum y)^2}{n} \\
&= 722 + \frac{224^2}{7} = 7890 \\
\sum xy &= \sum(x-\bar{x})(y-\bar{y}) + \frac{\sum x \sum y}{n} \\
&= 371 + \frac{112\times224}{7} = 3955 \\
SS_{resid} &= \sum y^2 - a\sum y - b\sum xy \\
&= 7890 - 2.6144\times224 - 1.8366\times3955 \\
&= 40.63 \\
95\%\ CI_{x=15} &= a + bx^* \pm t_{CRITICAL}\times s_{a+bx^*} \\
&= a + bx^* \pm t_{CRITICAL}\times s_e\sqrt{\frac{1}{n}+\frac{(x^*-\bar{x})^2}{S_{xx}}} \\
&= a + bx^* \pm t_{CRITICAL}\times\sqrt{\frac{SS_{resid}}{n-2}}\sqrt{\frac{1}{n}+\frac{(x^*-\bar{x})^2}{\sum(x-\bar{x})^2}} \\
&= 2.6144 + 1.8366\times15 \pm 2.571\times\sqrt{\frac{40.63}{5}}\sqrt{\frac{1}{7}+\frac{(15-16)^2}{202}} \\
&= 30.163 \pm 2.81766
\end{aligned}$$

(d) Interpret (c)

The value from (c) states that we can be 95% confident the mean population value of journey time for an 8:15am departure lies within 2.82 minutes of 30.16 minutes.

(e) 95% Prediction Interval

$$\begin{aligned}
95\%\ PI_{x=15} &= a + bx^* \pm t_{CRITICAL}\times\sqrt{{s_e}^2 + {s_{a+bx^*}}^2} \\
&= 2.6144 + 1.8366\times15 \pm 2.571\times \\
&\quad \sqrt{\left(\sqrt{\frac{40.63}{5}}\right)^2 + \left(\sqrt{\frac{40.63}{5}}\sqrt{\frac{1}{7}+\frac{(15-16)^2}{202}}\right)^2} \\
&= 30.163 \pm 2.571\times3.054 \\
&= 30.163 \pm 7.8519
\end{aligned}$$

(f) 8:30am Prediction Feasibility

Although a prediction of travel time for a departure time at 8:30am is physically possible, there is no evidence or observations to suggest this model holds for this departure. Therefore, this model should not be used to perform such a prediction.

Question 3

Case 1 – Removing case “+” should yield r = 0.285

Case 2 – Removing case “x” should yield r = 0.720

Case 3 – Removing case “+” and “x” should yield r = 0.925