1) a + c

2) b + d

6) step 1: chose

- identify response & explainatory variables: APR = explainatory & payment = response
- both variables are continuous & quantitative, so try to plot a scatter plot to see whether the relationship looks linear or not.

$$Y_{(payment)} = \beta_0 + \beta_1 x_{(APR)} + \epsilon \ ; \ \epsilon \sim N(0, \sigma^2)$$

step 2: fit

- calculate the estimated value for all parameters ($\hat{\beta}_0, \hat{\beta}_1$) and intepret them
- this can be done either by hand or by other software.

step 3: assess

- figure out how well the model fit the observation.
- looking at residuals plots and q-q plot to see whether the residual are constant, linear & normal or not; noted that normality of residuals is only needed for inference.

step 4: use

- use the model to predict or answer the question we want an answer to For example, use CI or p-value for slope (β_1) to see whether APR has a significant effet on users' payment.

8.1 (8.a)

To do a SLR model, the sample has to be independent & random. The question says that the sample is random so we assume that we can trust it that the sample is random. Also, it is safe to assume that 47 ($n \equiv$ sample size = df + 1) is not more than 10% of all Singaporean diamonds (population), so we can assume that the sample is also independent.

Now, we move to assessing whether this model fit the data well or not. From residuals plot, we see that the residuals are uniform: linear and having constant variance. There is no expansion in the size of residual as fitted values grow, so we can do estimation on the model.

Moving to see whether we can do inference or not, we take a look at the q-q plot. The observed residuals almost lie atop of the theoretical Line, so we see that the residual is normal. Thus, we can do inferencing too.

9.1) $\hat{\sigma}_\epsilon = 30.84$

Which me that on we estimate the standard deviation of calories burned by a runner at some average speed x to be 30.84.

0 becaus constant 0 because constant

$Var(y_{at\ x}) = Var(\cancelto{0}{\beta_0}) + Var(\cancelto{0}{\beta_1 x}) + Var(\epsilon)$

9.3) We can calculate CI of some specific significan level (popular on are 90, 95, 99% CI). Then the CI for that specific α is $80.82 \pm t_{(\frac{\alpha}{2})} \times \sigma_{\beta_1} \approx 80.82 \pm 2 \times 22.51$ which will include 100 cal per increase in average speed. Therefore, we cannot reject the null hypothesis that say our runners' rate is different from the average people in the same age group.

$\star$ $t_{\alpha/2}$ is the critical t-value of $n-2 = 16$ degrees of freedom that accounted for $(1-\alpha) \times 100\%$.

9.4)
- $R^2 = 0.4313$ which mean that our model can explain 43.13% of variability in the number of calories burned which is not particularly good. Because if our model can describe the data well, the R^2 should be close to 1.

- p-value from F-test $= 0.002255$ which means that the chance that we get $\hat{\beta}_1$ as extreme as 80.82 assuming the true β_1 is 0 is $\approx 0.2\%$ which is extremely low. Therefore, we could assume that reject the null hypothesis that there is no relationship between calories burned and running speed.