Stat 021 correction — Sumin Byun

- **Question 1** -- (b), (d)

(a)

(b) ✓

(c)

(d) ✓

- **Question 3** -- (b)

(a)

(b) ✓

(c)

- **Question 6**

1. Since we suppose I am interested in exploring a linear relationship between the variables, we can choose linear regression model because we can explore the linear relationship between my credit card account payments (predictor) and APR at the time of my payment (response). The simple linear regression model (the equation) is $Y = \beta_0 + \beta_1 x + \epsilon$ and the estimated equation is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.
2. Most of the work in this step can be done in R, importing the sample dataset of card payments and APR, making the scatter plot with regression line, having summary() function to find the estimated slope (β_1) and β_0 (intercept), standard error, and so on, then we can explore how this linear regression model fits the data.
   ↳ residual: observed value of APR − predicted value of APR (in the context)
3. We can now check the residual vs fitted values plot to see if the linearity assumption and constant variance assumption are met. Also, we can examine the residuals if the measurement errors follow a normal distribution with mean zero. For the normality assumption, we can see the normal quantile plot to check if the normality assumption is met; and, we should check if the errors are independent from one another (independence assumption). For the inference, we should make sure if the unseen errors (ϵ) in the model follow a normal distribution and if the data are obtained using a random process.
   ↳ data of the payments made on the credit card account and the APR at the time of the payment.

4. There are two things that we can do: estimation / prediction of SLR and SLR inference. In the estimation, we can estimate the parameters and predict the APR with the certain value of my card payment. In the inference, we can run hypothesis test or confidence interval to infer the true slope and draw practical conclusions.

- **Question 8.1**

1. Linearity
2. Zero mean of ϵ (random variable) : we assume that the distribution of ϵ is centered at 0.
3. uniform spread

Based on the residual plot, we can conclude that linearity assumption and the uniform spread assumption are both met.

4. Independence : we can assume that the samples are independent.
5. Normality : Based on the normal quantile plot, it is not a perfect good normal distribution, but still can say the errors roughly follow the normal distribution.
6. Randomness : In the setting, the sample of Singaporean diamonds is the simple random sample, so the randomness assumption is met.

- **Question 9.3**

We can calculate prediction interval of $(1-\alpha)$%. in order to determine if our runner's rate of burning calories is different from the overall average for all people in the age group. If the interval contains our runner's rate of burning calories, we cannot say our runner's rate is different from the average for all people in the age group. However, if the interval does not contain it, you could say that you are $(1-\alpha)$%. confident that the runner has a statistically significantly different rate of burning calories from the average for all people in the age group.

- **Question 9.4**

1. p-value of ANOVA F-test ($= 0.002$). If we check the p-value of test for β_1, where H_0 (null hypothesis) is $\beta_1 = 0$ which means the slope $= 0$; H_A (alternate hypothesis) is $\beta_1 \neq 0$ which means the slope $\neq 0$, we can reject the H_0 - because we had a small p-value ($0.002 < 0.005$ with $\alpha = 0.05$) - in favor of H_A that the slope is not 0. (β_1 here is the slope of the predictor, which is the speed of the run (in mph))
2. $R^2 = 0.413$ because R^2 shows the proportion of variability of y explained by the model. $R^2 = 0.413$ shows that the model is not a terribly bad fit for the data, but also not an extraordinarily good fit for the data.