Elaine Kim (Student ID : 902212991)

STAT 021, Test 1 Corrections

1. $\hat{p}_{state} = \frac{31.6}{100000}$

90% CI for p_{rural} = [31.2, 40.1]

(a) The confidence interval indicates that we are 90% confident that the true average prevalence (per 100,000 individuals) of lung cancer lies between 31.2 and 40.1. ⇒ valid

The CI lies above 31.6, so it is likely much higher than the state-wide average.

(b) Chance is the occurrence of events in the absence of any obvious intention or cause. It is, thus, the probability of something happening; when the chance is defined in mathematics it is called probability. The confidence interval simply quantifies our certainty that a certain value will be contained within an interval - it is not the probability. Therefore, the usage of "chance" is incorrect in this situation. ⇒ not valid

(c) A 90% CI can be interpreted as, if you were to repeatedly perform the study with new samples, the resulting confidence intervals would contain the true value 90% of the time. ⇒ valid

(d) As in the explanation for (b), the word "chance" is not appropriate in this situation. ⇒ not valid

Answer) b, d

3.

(a) A confidence interval is used to estimate the mean response, and a prediction interval is used to estimate an individual response. If you look at the mathematical equations for each of the standard errors:

$$SE_{\hat{\mu}} = \hat{\sigma}_\varepsilon \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\Sigma (x - \bar{x})^2}}$$

(for conf. int)

$$SE_{\hat{y}} = \hat{\sigma}_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\Sigma (x - \bar{x})^2}}$$

(for predn. int)

The only difference between the two standard errors is the extra 1 in the standard error for prediction. This extra 1 reflects the fact that an individual response will vary from the mean response μ_Y with a standard deviation of σ_ε. This also makes the prediction interval wider, if all other values are the same ⇒ true

(b) $SS_{Total} = SS_{Model} + SSE$

$SS_{Model} = \Sigma(\hat{y} - \bar{y})^2$

$SSE = \Sigma(y - \hat{y})^2$

The total sum of squared errors is the sum of the sum of squared errors due to the model and the sum of squared errors due to the model.

It is not necessary that the sum of squared errors is smaller than the sum of squared errors due to the model.

If the model did a terrible job at explaining the variability within the data, then SSE could be larger than SS_{Mod}.

⇒ false

(c) Correlation between x_1 and $Y = r_1$

Correlation between x_2 and $Y = r_2$

$|r_1| > |r_2|$,

Therefore $r_1^2 > r_2^2$ (the coefficient of determination for regressing Y on X_1 is greater than the coefficient of determination for regressing Y on X_2. ⇒ true

Answer) b

5. We assess normality and randomness conditions when we want to make an inference from the regression. To make an "inference" means to use the data we can see to infer facts about numbers or relationships we can't see (definition from textbook).

(a) ⇒ this is simply trying to model the data, with no inference involved.

⇒ Don't need to assess.

(b) To "predict" something means to get the fitted value given by the model we have chosen to use. This also requires no inference.

⇒ Don't need to assess

(c) To obtain an upper and lower bound on the size of the increase (= an interval), this requires an inference.

⇒ Need to assess

(d) Determining if something is statistically significant (or, that it is too extreme to be just due to random variation from one value to the next), this also requires inference. ⇒ Need to assess

Answer) a, b

6. 1) I would gather a data set of sufficient size (at least, approximately 30) of payments made on my credit card and the APR at the time of my payment. I would then use a scatter plot to visually determine if there was a linear relationship.

2) In order to fit my data, I could compute β_0 and β_1 through R or by hand. In this case, $\hat{y}$ would be the average APR (response), and x would be the payment made on my credit card (predictor).

3) To assess the fit of my plot, I would use a residual plot (residual vs $\hat{y}$) and a normal quantile plot. The regression plot would show if there was a linear relationship and if the variances were constant. The normal quantile plot would tell if the residuals followed a normal distribution.

4) If 1-3 are complete, I can now answer statistical questions. For example, I could use the CI of the slope to determine the interval of the slope for a certain confidence level

8. a)

Linearity - The residual plot shows no particular pattern (ex. fanning, curvature), and the data points appear to be randomly distributed, so we can say that the linearity condition has been met.

Constant variance - The residual plot shows fairly even distribution above and below 0, so the variance can be predicted to be fairly constant.

Normality - The residuals follow a relatively unimodal plot (studentized resids.), and the Normal probability plot shows that the residuals roughly follow a normal distribution. So, the "normality" condition is met.

Zero mean - The R output shows the model parameters for the estimated regression equation, and by default the linear part of the model, $\beta_0 + \beta_1 x$, shows that the error distribution is centered at 0.

Randomness - We do not have information about how the data was collected; therefore, we do not know if the data was obtained using a random process.

Independence - From the residual plots, we do not know if the errors are independent from one another.