STAT 21 Midterm I Corrections

Stat 21 Directory

Quick note: I lost 0.5 points on question 6 for "Step 1. Mentions which is a predictor and which is a response."

- I think I did this. I wrote: "We want pay(APR)," which is shorthand for "We want payments as a function of APR," so clearly payments is the response and APR is the explanatory. I use this shorthand throughout the exam, for example on question 8 where I write "price(size)" for "price of the diamond is a function of size of the diamond."
- Also, I wrote that I would fit a model: `lm(payments ~ APR...)`, so clearly I knew which is which.

Question 4

Which of the following conditions for inference in regression does a fitted-values vs. residual plot not aid in assessing?

Answer: (c) Independence and (d) Normality

Reasoning: We need to use a quantile plot to check for normality.

Question 5

In which of the following situations do we not need to assess the normality and randomness conditions for inference in regression?

Answer: (a) A scientist studying a random sample... (d) A used car salesperson wants to determine if an increase in the number of miles on a car has a statistically significant impact on the sale price of the vehicle.

Reasoning: For (a), we have a random sample which means we get these conditions for free. For (b), we're not trying to do inference, e.g. finding the average effect of an increase of n miles. Instead, we're just checking if there is indeed a relationship.

Question 6

Describe what you would do at each of the four modeling steps above in no more than 1-3 sentences per step. Make sure to provide some justification/explanation for each activity.

Quick note: I lost 0.5 points on question 6 for "Step 1. Mentions which is a predictor and which is a response."

- I think I did this. I wrote: "We want pay(APR)," which is shorthand for "We want payments as a function of APR," so clearly payments is the response and APR is the explanatory. I use this shorthand throughout the exam, for example on question 8 where I write "price(size)" for "price of the diamond is a function of size of the diamond."
- Also, I wrote that I would fit a model: `lm(payments ~ APR...)`, so clearly I knew which is which.

Answer: (I'm just covering the points I missed.)

Step 3. We also need to check for normality of residuals. To do so, I'd use a normal quantile plot and see if there is any noticeable deviation from the quantile line.

Question 8a:

Based on these plots, what conclusions can we make about the conditions necessary for inference with a SLR model?

Answer: We want to make sure $\epsilon \sim i.i.d. N(0, \sigma_\epsilon)$, with σ_ϵ constant.

Based on the plots, things look good for inference. This is a simple random sample, so we get the randomness condition. There are no visible trends in the residuals plot (neither non-linearity nor non-constant variance), and the errors are centered at 0. It's also important that the errors be normally distributed. Though not totally perfect (some of the residuals on the high- and low-ends are bit higher than the theoretical quantiles would predict), the normal quantile plot looks mostly solid.

Question 9a.

What is the estimate for the standard deviation of the number of calories burned based on this linear model?

Answer: This is the standard error of regression, which is: 30.84 calories.

Question 9c.

Suppose, on average... every mph increase in running speed corresponds to 100 additional calories burnt. Describe a procedure to determine if our runner's rate of burning calories is different...

Answer: Here, we could do a one sample *t*-test. We want to check our null hypothesis $H_0 : \beta_1 = 100$, where $\hat{\beta}_1$ is the slope for our runner, estimated at 80.82 additional calories burnt per mph increase. Our alternative hypothesis is $H_a : \beta_1 < 100$. We'll use $\alpha = 0.05$, corresponding to a 95% confidence level.

We can just compute this in R. We'd get back a *p*-value, which would give the probability that we'd observe data this extreme if there is no difference between our runner's calories burnt per mph and 100 calories burnt per mph. If that value is below our $\alpha = 0.05$, then we reject the null hypothesis in favor of the alternative hypothesis, which is that fewer than 100 calories burnt are associated with each mph increase for our runner.