Multiple Choice

1. Which plot do we use to check the normality condition for simple linear regression?

   (a) Normal QQ-plot of residuals

2. Which plot do we use to check the linearity condition of multiple linear regression?

   (a) Residuals versus fitted values plot

3. Suppose we fit a linear regression model with two predictors, and its R^2 is 0.68. Then we fit another model that includes those two predictors and their interaction. Which of the following could be the larger, three-predictor model's R^2?

   (b) 0.70

4. Consider the simple linear regression model: $Y = \beta_0 + \beta_1 X + \epsilon$. Which of the following is correct?

   (b) β_1 is a parameter

5. Consider a categorical predictor, with three categories, and a quantitative response variable. When we conduct a one-way ANOVA test, what are the hypotheses?

   (c) H_0 : all group means are equal
   H_A : at least one group's mean does not equal the other group means

6. If we want to draw cause-effect conclusions from a study, what must be true about the study?

   (c) Treatments are randomly assigned to subjects.

7. Use the following fitted linear regression model to calculate the residual for an observation with $X_1 = 0$, $X_2 = 5$, and $Y = -17$.

   $$\widehat{Y} = -1 + 2X_1 - 3X_2$$

   (b) -1

8. Suppose with a certain treatment the probability of recovery from a disease is 0.3. What are the odds of recovery with this treatment?

   (d) 0.43

9. Which of the following is **true**?

(a) Being 90% confident that an interval captures μ means that if we were to repeatedly take samples and construct the corresponding intervals, in the long run 90% of them would capture μ.

(c) If every predictor in a multiple linear regression model has a VIF > 5, we should not use that model for predicting the response.

10. Use the following fitted model to predict Y when $X = 2$. Log indicates natural log.

$$log(Y) = 5 - X^2$$

(b) e^1

11. What is the distribution of the error term in a multiple linear regression model with two predictors?

(e) Normal$(0, \sigma^2)$

12. What method is used to estimate logistic regression model coefficients?

(b) Maximum likelihood

Short Answer Analysis Questions

13. *(20 points)* Consider a linear regression model that predicts a quantitative response Y from a quantitative predictor X and the season of the year. Seasons include spring, summer, fall, and winter. The model is:

$$Y = \beta_0 + \beta_1 X + \beta_2 Summer + \beta_3 Fall + \beta_4 Winter + \epsilon,$$

where $Summer$, $Fall$, and $Winter$ are indicator variables ($= 1$ if that's the season; $= 0$ otherwise).

(a) Which season is the reference category?

Answer: The reference category here is Spring season because it is the only level in the categorical variable Season missing in in the about multiple linear regression.

(b) Briefly interpret what it means if the coefficient of $Summer$ is positive.

Answer: The positive coefficient for Summer suggests that for a fixed X, the response Y will be higher for Summer compare to Spring by the magnitude of β_2 on average.

(c) Briefly interpret what it means if the coefficient of X is negative.

Answer: Foe any given season, the response Y will decrease by about the magnitude of β_1 for every unit increase in X on average.

(d) State the null and alternative hypotheses to test whether, holding X fixed, the average response differs between fall and spring.

Answer:

$$H_0 : \beta_3 = 0 \quad versus \quad Ha : \beta_3 \neq 0$$

(e) Suppose this is the (incomplete) ANOVA table for this model.

Source	DF	Sum of Squares	Mean Squares	F-statistic
Model	4	14872	3718	89.24769
Error	91	3791	41.65934	—
Total	95	18663	—	—

i. Calculate R^2 for this model.

Answer:

$$R^2 = \frac{SSModel}{SSTotal} = \frac{14872}{18663} = 0.7968708$$

ii. Briefly interpret the R^2 value you calculated.

Answer: About 79.68708% of variability in the response Y can be explained by the model.

iii. Calculate the test statistic for the overall ANOVA F-test.

Answer:

$$F = \frac{MSModel}{MSE} = \frac{3718}{41.65934} = 89.24769$$

iv. State the null and alternative hypotheses that would be tested by the overall ANOVA F-test.

Answer:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 \quad versus \quad H_a : \text{At least one} \quad \beta_i \neq 0$$

(f) How would we change this model to allow each season to have a different slope of X? You may answer in words or by stating the new model.

Answer: By adding the interaction term between the quantitative predictor X and the categorical predictor Season in the model.

14. *(8 points)* Consider the fitted logistic regression model stated below, where π represents the probability of success and X is a quantitative predictor ranging from 0 to 10. Use the model to answer the questions that follow.

$$log\left(\frac{\widehat{\pi}}{1-\widehat{\pi}}\right) = -3 + 0.4X$$

(a) What is the estimated probability of success for $X = 5$?

Answer:

$$\widehat{\pi} = \frac{\exp(-3 + 0.4 \times 5)}{1 + \exp(-3 + 0.4 \times 5)} = 0.2689414$$

(b) What is the estimated probability of failure for $X = 5$?

Answer: Let's denote by $\widehat{\pi}_{failure}$ the probability of failure. Thus,

$$\widehat{\pi}_{failure} = 1 - \widehat{\pi} = 1 - 0.2689414 = 0.7310586$$

(c) What is the estimated odds of success for $X = 5$?

Answer:

$$Odds_{\widehat{\pi}} = \frac{\widehat{\pi}}{1-\widehat{\pi}} = \frac{0.2689414}{0.7310586} = 0.3678794$$

(d) Calculate the predictor value associated with a 0.5 probability of success.

Answer:

$$X = -\frac{\beta_0}{\beta_1} = -\frac{-3}{0.4} = 7.5$$

15. *(16 points)* Two students at Grinnell College took a simple random sample of students who were US citizens and conducted phone interviews to investigate patterns of political involvement. Data is in **Political** in the `Stat2Data` package. We will look at the variables *Participate* (= 1 if voted; = 0 if eligible to vote but didn't) and *Edit* (= 1 if read editorial page; = 0 if don't).

(a) Fit a logistic regression model predicting *Participate* from *Edit*. State the fitted model in logit form. I've started it for you.

Answer: $log\left(\frac{\widehat{\pi}}{1-\widehat{\pi}}\right) = 0.5108 - 0.1854 Edit$

(b) The estimated slope on "Edit" is negative. Briefly, what does that tell us?

Answer: The negative coefficient of of Edit suggests that students who read editorial pages are less likely to vote compare to students who don't read editorial pages.

(c) Use the slope from the fitted model to estimate the odds ratio of voting for those that read the editorial page versus those who don't. **Show your calculation.**

Answer:

$$OR = \exp(\widehat{\beta}_1) = \exp(-0.1854) = 0.8307719$$

(d) Interpret the odds ratio in context.

Answer: The odds of a student voting is 0.8307719 times lower for students reading editorial pages compare to student who don't.

(e) Report a 95% confidence interval for the odds ratio.

Answer: The 95% confidence interval for β_1 is $(-1.277, 0.9063)$. Thus a 95% confidence interval for the odds ratio is

$$(e^{-1.277}, e^{0.9063}) = (0.2788727, 2.475148)$$

(f) We want to know if the relationship between reading the editorial page and the odds of voting is statistically significant. Use the model to test whether $\beta_1 = 0$ at a significance level of 0.05.

i. Report the Z-test p-value to at least three decimal places.

Answer: The Z-test $= -0.3329$ and the p-value$= 0.7392$

ii. Report the likelihood ratio test p-value to at least three decimal places.

Answer: The likelihood ratio $= 0.11$ and the p-value$= 0.7392$

iii. These two p-values should lead you to the same conclusion. Circle it.

B. We do not see evidence that this relationship is significant.

16. *(16 points)* Consider the **Pulse** dataset in the `Stat2Data` package. After loading the data, run `?Pulse` and read the variable definitions. Our goal is to build a model to predict resting heart rate.

(a) Fit a linear regression model to predict resting heart rate from weight only. We'll call this model **Model 1**. State the fitted model.

Answer:

$$\widehat{Rest} = 77.43 - 0.05749Wgt$$

(b) Use Model 1 to provide an interval that you are 90% confident captures the resting heart rate of one specific person who weighs 180 pounds.

Answer: The 90% prediction interval of the resting heart rate of one specific person who weighs 180 pounds is $(47.71, 86.45)$.

(c) Interpret your part (b) interval.

Answer: 90% of all individual with weight 180 pounds will have there predicted resting rate to be between 47.71 and 86.45.

(d) Now fit a model predicting resting heart rate from exercise, height, and weight. Use linear terms only, no interactions or transformations. We'll call this model **Model 2**. State the fitted model.

Answer:

$$\widehat{Rest} = 114.2 - 6.949Exercise - 0.4658Hgt + 0.01035Wgt$$

(e) Which of the two models seems to better meet the conditions for linear regression? Briefly justify your answer.

Answer: Model 2 meets the condition better as the constant variance assumption is better in model 2 compare to model 1.

(f) Which of the two models would you prefer to use to predict resting heart rate, based on the "Residual standard error" values reported in the `summary()` outputs. Briefly justify your answer.

Answer: Based on the "Residual standard error", model 2 seems to be better as the 'Residual standard error" for model 2 is lower than the 'Residual standard error" for model 1.

(g) What's another value reported in the `summary()` output that we might use to compare these models? Which model do you prefer based on it?

Answer: The adjusted coefficient of (multiple) determination R^2_{adj} could also be used to compare the two model. We observed that as we added the two new predictors in model 1 the adjusted coefficient of (multiple) determination R^2_{adj} went up. So, we prefer model 2 compare to model 1. However, a significant test needs to be perform here to statistically see if model 2 is better than model 1.

(h) Conduct a Nested F-test to compare the models. Report the **test statistic** and state **which model** appears to be "better" based on this test. Use a reasonable significance level.

Answer:

$$H_0 : \beta_1 = \beta_2 = 0 \quad versus \quad H_a : \beta_1 \neq 0 \quad or \quad \beta_2 \neq 0$$

Decision: The resulting p-value= 8.17×10^{-18} which less than the significance level $\alpha = 0.05$. So we reject the null hypothesis in favor of the alternative hypothesis.

Conclusion: We do see significant evidence that model 2 is statistically significantly better than model 1 at the significant level of $\alpha = 0.05$.

17. *(10 points)* An extensive survey was conducted by the Center for Disease Control to study health-related risky behavior of "youths". Some of the data is stored in **YouthRisk2009** in the `Stat2Data` package. We are interested in answering the question: What is the relationship – or is there one at all – between age and the odds of ever having smoked marijuana? Use `?YouthRisk2009` to see the data documentation and find the relevant variables.

   (a) Very briefly explain why we should use a logistic regression model, as opposed to a linear regression model, to analyze this data.

   Answer: Because we are interested in a relationship between age (quantitative predictor) and the odds of ever having smoked marijuana (binary response), a logistic regression will be more appropriate for this relationship compare to a simple linear regression model.

   (b) Use the data and a logistic regression model to thoroughly answer the research question. *You **do not** need to fill this page. This is just how the spacing worked out. **Do** write enough to demonstrate your understanding of the topic.*

   Answer:

   - Both the logit and the probability form of the logistic regression.
     - Logit form

$$\log\left(\frac{\widehat{\pi}}{1-\widehat{\pi}}\right) = -5.761 + 0.3216Age$$

     - Probability form

$$\widehat{\pi} = \frac{\exp(-5.761 + 0.3216Age)}{1 + \exp(-5.761 + 0.3216Age)}$$

   - Then use this model to talk about the relationship indeed by applying some of the things from problem 15.