# Multiple Choice

1. Which plot do we use to check the normality condition for simple linear regression?
   (a) Normal QQ-plot of residuals
   (b) Normal QQ-plot of the predictor
   (c) Normal QQ-plot of the response
   (d) None of the above; this is one of the conditions we can't check with a plot.

2. Which plot do we use to check the linearity condition of multiple linear regression?
   (a) Residuals versus fitted values plot
   (b) Empirical logit plot
   (c) Scatterplots of each predictor versus $Y$
   (d) None of the above; this condition depends on how the data was collected.

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 $R^2$?
   (a) 0.66
   (b) 0.70
   (c) Either of these
   (d) Neither of these

4. Consider the simple linear regression model: $Y = \beta_0 + \beta_1 X + \epsilon$. Which of the following is correct?
   (a) $Y$ is a parameter
   (b) $\beta_1$ is a parameter
   (c) $\epsilon$ is a parameter
   (d) All of these are correct.

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?
   (a) $H_0$ : all group variances are equal
   $H_A$ : at least one group's variance does not equal the other group variances
   (b) $H_0$ : all group variances are equal
   $H_A$ : none of the three groups have the same variance
   (c) $H_0$ : all group means are equal
   $H_A$ : at least one group's mean does not equal the other group means
   (d) $H_0$ : all group means are equal
   $H_A$ : none of the three groups have the same mean

6. If we want to draw cause-effect conclusions from a study, what must be true about the study?
   (a) There are an equal number of subjects in each treatment group.
   (b) The response variable is quantitative.
   (c) Treatments are randomly assigned to subjects.

(d) Researchers never directly contact the subjects.
(e) None of the above are necessary; as long as the results are statistically significant, we can conclude a cause-effect relationship.

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

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

(a) -2
(b) -1
(c) 0
(d) 1
(e) 2
(f) We need $\widehat{\sigma}_\epsilon$.

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?

(a) 2.33 (b) 1.35 (c) 0.7 (d) 0.43 (e) 0.3

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

(a) Being 90% confident that an interval captures $\mu$ means that if we were to repeatedly take samples and construct the corresponding intervals, in the long run 90% of them would capture $\mu$.
(b) A p-value is the probability that the null hypothesis is true.
(c) If every predictor in a multiple linear regression model has a VIF $> 5$, we should not use that model for predicting the response.
(d) If constructed from the same data, a 90% confidence interval for $\mu$ will be wider than a 99% confidence interval for $\mu$.

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

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

(a) 1 (b) $e^1$ (c) 3 (d) $e^3$ (e) 9 (f) $e^9$

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

(a) t($df = n - 1$)
(b) t($df = n - 2$)
(c) t($df = n - 3$)
(d) t($df = n - 4$)
(e) Normal$(0, \sigma^2)$
(f) Normal$(0, 1)$

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

(a) Least squares
(b) Maximum likelihood
(c) Method of moments
(d) Bonferroni estimation

# 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?

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

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

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

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

| Source | DF | Sum of Squares | Mean Squares | F-statistic |
|---|---|---|---|---|
| Model | 4 | 14872 | | |
| Error | | 3791 | | — |
| Total | 95 | | — | — |

i. Calculate $R^2$ for this model.

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

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

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

(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.

14. *(8 points)* Consider the fitted logistic regression model stated below, where $\pi$ 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 $X = 5$?

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

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

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

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.

$log\left(\frac{\widehat{\pi}}{1-\widehat{\pi}}\right) =$

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

(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.**

(d) Interpret the odds ratio in context.

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

(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.

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

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

A. We see significant evidence that odds of voting is related to whether a person reads the editorial page.

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.

    (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.

    (c) Interpret your part (b) interval.

    (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.

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

    (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.

    (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?

    (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.

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.

    (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.*