STA363: Introduction to Statistical Modeling
Practice "Mini-Final" Key

TRUE OR FALSE

1. F In a one-way ANOVA, the *F*-test tests the null hypothesis H_0: $\mu_1 = \mu_1 = \cdots = \mu_k = 0$.

2. T R^2 is a poor means by which to compare the quality of fit of two models because R^2 will never decrease by adding predictors to the model.

3. F Consider two models for the same data. Model 1 has AIC = –32.9, and model 2 has AIC = –28.8. Model 2 is the better fitting model to the data.

5. F In a multiple regression model given by $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$, β_1 can be correctly interpreted as the mean change in Y given a one-unit increase in X_1.

6. F Cross-validation is a method used to determine what variables are significant in a statistical model.

7. T Logistic regression is a modeling tool for binary response variables. However, you can use either quantitative or qualitative predictors in a logistic regression model.

8. F Multicollinearity is a situation in a multiple regression where some of the predictors are related to the response variable.

9. F In ANCOVA models, we typically start by fitting a no-interaction model and then simplify the model if warranted.

10. T Poisson regression is a type of generalized linear model useful for data where the response Variable Y is a count.

MULTIPLE CHOICE

11. A A confidence interval for the mean response in a regression model is:
A] never wider than the corresponding prediction interval for the response.
B] always wider than the corresponding prediction interval for the response.
C] the same as a prediction interval for the response.

12. B Two variables are said to interact if:
A] they both have small *p*-values.
B] the effect that one of them has on the response depends on the value of the other.
C] they both have an effect on the response Y.

13. B Violations of the linearity assumption in a regression model may be addressed by:
A] transforming the response variable.
B] transforming one or more of the predictor variables.
C] running a cross-validation.

14. Suppose you intend to fit a multiple regression model to a set of data using four predictors X_1, X_2, X_3, and X_4. The model you fit will be of the form $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$. Suppose you wish to test the hypothesis H_0: $\beta_3 = \beta_4 = 0$. Describe the process of how to do this by using a reduced F-test. (Do not write R code; instead, describe the logic behind the process in your own words.)

1) Fit the full model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$.
2) Impose $H_0: \beta_3 = \beta_4 = 0$ onto this, resulting in the reduced model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$.
3) Run an F-test that compares explained variability in the response Y between these two models.

15. In an ADHD treatment study, a randomly selected group of 40 children were assigned to one of two experimental drugs (call them *A* and *B*) at two different dosages (10mg and 40mg). Each child received only one specific drug/dosage combination, and was administered that treatment over a one week period. Their response time to a stimulus was measured after administration of a treatment.

- What are the treatments in this study?

  4 of them: 1) A at 10mg; 2) A at 40mg; 3) B at 10mg, 4) B at 40mg.

- What are the experimental units in this study?

  Children.

16. A field study in ornithology (bird study) is conducted to determine how two different characteristics influence the likelihood of habitation of a particular species of bird on a given island. In this example, the response variable is called `occupied`: a value of 1 means that a given island was occupied by the species in question, and a 0 means it was not. The two predictor variables are the area of the island (`area`, in km^2) and the isolation of the island (`distmain`, distance from the mainland, in km). A logistic regression using these two predictors is performed below in R:

```
> model1 <- glm(occupied ~ distmain + area, data=d, family=binomial)
> summary(model1)
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   6.6417     2.9218   2.273  0.02302 *
distmain     -1.3719     0.4769  -2.877  0.00401 **
area          0.5807     0.2478   2.344  0.01909 *

> exp(model1$coefficients)
(Intercept)    distmain        area
766.3669575   0.2536142   1.7873322
```

Give a complete, clear, and correct interpretation of the output value 0.2536 in the context of this example.

For ea. additional km of distance from the mainland, we estimate that the odds of occupation of this species on an island are reduced by about 75% (assuming the area of the island is fixed).