**Math 455** **Midterm 1** **Spring 2024**

No books, no notes, no calculators.

Name: ________________________________________

| Question | Points | Score |
|---|---|---|
| 1 | 8 | |
| 2 | 8 | |
| 3 | 15 | |
| 4 | 4 | |
| 5 | 24 | |
| 6 | 13 | |
| 7 | 8 | |
| 8 | 21 | |
| 9 | 6 | |
| 10 | 6 | |
| 11 | 6 | |
| 12 | 20 | |
| 13 | 12 | |
| 14 | 8 | |
| Total: | 159 | |

1. Consider the simple (one predictor variable) linear regression model, $Y = \beta_0 + \beta_1 X + \epsilon$.

   (a) (4 points) Does this model specify that $X$ is a random variable? Why or why not?

   (b) (4 points) Does this model specify that $Y$ is a random variable? Why or why not?

2. (8 points) Let $X_1, X_2$ be random variables whose joint distribution has the variance-covariance matrix

$$\begin{bmatrix} 3 & 1 \\ 1 & 2 \end{bmatrix}$$

   Find the variance of $X_1 + 5X_2$.

3. Let $X_1, X_2, X_3, X_4, X_5$ be independent and identically distributed random variables which are all normal with mean 0 and variance 1. Let $\bar{X} = \frac{1}{5}(X_1 + X_2 + X_3 + X_4 + X_5)$. Let $W = \sum_{i=1}^{5}(X_i - \bar{X})$.

   (a) (1 point) Find $E[\bar{X}]$.

   (b) (3 points) Find $V[\bar{X}]$.

(c) (3 points) What is the distribution of $\bar{X}$? (Give the values of any parameters that are relevant. This means that if you say [for example] that $X$ has the beta distribution with parameters $\alpha$ and $\beta$, you should specify the numerical values of $\alpha$ and $\beta$.)

(d) (3 points) What is the distribution of $W$? (Give the values of any parameters that are relevant.)

(e) (5 points) What is the distribution of $U = 2\sqrt{5}\frac{\bar{X}}{\sqrt{W}}$? (Give the values of any parameters that are relevant.)

4. (4 points) Fill in the blanks in the following definition of $p$-value:

If $W$ is a test statistic, the $p$-value, or attained ______________________________, is the smallest level of significance $\alpha$ for which the observed data indicate that the ________________________ should be ______________________________.

5. The “Galapagos” data set has 30 rows of data and 7 variables. Consider the regression summary below and answer the following questions.

```
> lmod <- lm(Species ~ Nearest + Scruz + Adjacent, data=gala)
> summary(lmod)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 96.421923  29.636973   3.253  0.00315 **
Nearest      1.292094   1.987029   0.650  0.52123
Scruz       -0.460349   0.414648  -1.110  0.27707
Adjacent     0.007821   0.025895   0.302  0.76503
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 118.2 on B degrees of freedom
Multiple R-squared:  0.04603,Adjusted R-squared:  -0.06404
F-statistic: 0.4182 on A and B DF,  p-value: 0.7414
```

(a) (4 points) The $p$-value `0.7414` above arises from a statistical test involving the $F$-statistic. What are the null hypothesis and alternative hypothesis of that statistical test?

(b) (4 points) The $F$-statistic (numerator and denominator) degrees of freedom normally printed in the regression summary output have been replaced by `A` and `B`. From the information above, find `A` and `B`.

(c) (4 points) If $F$ is a random variable having the $F$-distribution with `A` and `B` numerator and denominator degrees of freedom, what probability involving $F$ is the $p$-value `0.7414`?

(d) (4 points) The $p$-value `0.27707` above arises from a statistical test. What are the null hypothesis and alternative hypothesis of that statistical test?

(e) (4 points) Suppose you wished to test the null hypothesis $\beta_{\texttt{Nearest}} = 1$ against the alternative hypothesis $\beta_{\texttt{Nearest}} \neq 1$. What test statistic would you compute? You need not do any calculuation, just write the expression for the relevant test statistic.

(f) (4 points) Assuming the null hypothesis in part (e), what is the distribution of your test statistic? Give the values of any relevant parameters.

6. Consider the following regression summary output from a study examining teenage gambling in Britain. The variables considered are `income`, measured in pounds per week, `verbal`, a verbal score in words out of 12 correctly defined, and `gamble`, expenditure on gambling in pounds per year.

```
Call:
lm(formula = gamble ~ income + verbal, data=teengamb)

Coefficients:
               Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)     7.000     14.583    0.480    0.634
verbal         -2.000      2.075   -0.964    0.340
income          5.000      0.985    5.074 7.56e-06 ***
```

(a) (5 points) What is the predicted expenditure on gambling *in pounds per year* for a teenager with an income of 40 pounds per week and a verbal score of 10?

(b) (8 points) Write a sentence interpreting the coefficient of `verbal` in this regression, comprehensible to the non-statistician.

7. Suppose $X$ and $Y$ are random variables whose joint distribution is the uniform distribution on the unit disc in the plane ($x^2 + y^2 \leq 1$). (Hint: this problem can be done using symmetry, without any laborious calculations.)

   (a) (2 points) Find $E[X]$ and $E[Y]$.

   (b) (6 points) Find $\text{Cov}(X, Y)$.

8. Suppose we have a simple (one predictor variable) regression model $Y = \beta_0 + \beta_1 X + \epsilon$ and data $(X_i, Y_i)$, $i = 1, \ldots, n$.

   (a) (3 points) Write a formula for $\hat{\beta}_1$ in terms of the $X_i$ and $Y_i$.

   (b) (3 points) Rewrite this formula in terms of the correlation between $X$ and $Y$ and the standard deviations $\sigma_X, \sigma_Y$.

(c) (5 points) Suppose now that our data $(X_i, Y_i)$ are independent and identically distributed samples from a uniform distribution on the unit disc in the plane ($x^2+y^2 \leq 1$). If $n = 9999$, what approximate value do you expect to see in the regression output for $\hat{\beta}_1$, and why?

(d) (10 points) Another student observes, "This data generating process has rotational symmetry about the origin. Therefore we should expect an approximate result of 0 for $\hat{\beta}_0$ in the regression output, but any line through the origin is equally likely because of the rotational symmetry. In fact, even if we redid the regression with the same data, we might get a different answer for $\hat{\beta}_1$." Comment on these observations. There are many assertions, and your comment should address whether each assertion is true and why.

9. The Gauss-Markov Theorem states that the OLS estimators $\hat{\beta}_0, \hat{\beta}_1$ are BLUE.

(a) (2 points) What does the B stand for?

(b) (4 points) What is the mathematical meaning of your answer to part (a)?

10. (6 points) Fill in the blanks in the following definition of the $F$-distribution.

Let $W_1, W_2$ be a independent random variables, having the ____________-distribution with $\nu_1, \nu_2$ degrees of freedom respectively. Then the random variable

$$F = \underline{\hspace{3cm}}$$

is said to have the $F$-distribution with ______ __________________ degrees of freedom,

and ______ __________________ degrees of freedom.

11. (6 points) Fill in the blanks in the following statements about the "hat matrix" in the context of the multiple regression model $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$. (Here with $p$ predictor variables and $n$ rows of data.)

The hat matrix $H$ can be expressed in terms of the model matrix $X$ by the formula: $H =$ ____________,

and is called the hat matrix because the fitted values $\hat{Y}$ may be expressed using $H$ as $\hat{Y} =$ ____________.

The expression for $H$ in terms of the model matrix comes from linear algebra. It is the matrix that projects $\mathbb{R}^n$ onto ________________________.

12. In Chapter 4, the prediction problem was examined in the context of the body fat data set `fat`. Answer the following questions about the R code and output below.

```
> data(fat,package="faraway")
> lmod <- lm(brozek ~ . -siri -density -adipos -free, data = fat)
> x <- model.matrix(lmod)
> x0 <- apply(x,2,median)
> predict(lmod,new=data.frame(t(x0)),interval="prediction")
       fit     lwr      upr
1 17.49322 9.61783 25.36861
> predict(lmod,new=data.frame(t(x0)),interval="confidence")
       fit      lwr      upr
1 17.49322 16.94426 18.04219
```

(a) (2 points) After the code `x0 <- apply(x,2,median)` is executed, is `x0` a number, a vector, or a matrix?

(b) (2 points) How would you characterize the value of `x0`? (Give a one-sentence description using only words, not mathematical symbols.)

(c) (6 points) In the second line of code, what does `brozek ~. -siri -density -adipos -free` mean? Explain specifically the period and the minus signs.

(d) (10 points) The outputs of the two calls to the `predict` function are very different. What, conceptually, is the difference between the `prediction` interval and the `confidence` interval? How does this explain the difference in output?

13. Television advertising would ideally be aimed at exactly the audience that observes the ads. A study was conducted to determine the amount of time that individuals spend watching TV during evening prime-time hours. Twenty individuals were observed for a 1-week period, and the average time spent watching TV per evening, Y , was recorded for each. Four other bits of information were also recorded for each individual: $x_1$ = age, $x_2$ = education level, $x_3$ = disposable income, and $x_4$ = IQ. Consider the three models given below:

$$\begin{aligned} \text{Model I:} \quad & Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon \\ \text{Model II:} \quad & Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \\ \text{Model III:} \quad & Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon \end{aligned}$$

Indicate whether each of the following statements are true or false and give short reason for each answer:

(a) (4 points) Models I and III can be compared using an $F$-test.

(b) (4 points) After fitting models I and II, and computing the SSEs, $\text{SSE}_{\text{I}} \leq \text{SSE}_{\text{II}}$.

(c) (4 points) After fitting models II and III, and computing the $R^2$s, $R^2_{\text{III}} \geq R^2_{\text{II}}$.

14. Suppose you have a simple (one predictor variable) linear regression model for which you have computed the fitted values $\hat{Y}$ and the correlation $\text{Cor}(Y, \hat{Y})$ is $-0.8$.

(a) (4 points) What is the coefficient of determination $R^2$ for this model?

(b) (4 points) What can be deduced about $\hat{\beta}_1$ from this information?