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

No books, no notes, no calculators.

Name: _____

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?

   Solution: No, it does not. The values of X might be, for example, chosen deliberately by an experimenter. Also, X does not depend on ϵ, which by hypothesis is the defining random variable of the model.

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

   Solution: Yes, it does. The model specifies that ϵ is a random variable, and Y depends on ϵ.

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 $X_1 + 5X_2$.

   Solution:

   $V[X_1+5X_2] = V[X_1]+V[5X_2]+2\,\mathrm{Cov}[X_1, 5X_2] = 3+5^2 \cdot V[X_2]+2\cdot 5\cdot \mathrm{Cov}[X_1, X_2] = 3+5^2\cdot 2+2\cdot 5\cdot 1 = 63$

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 $E[\bar{X}]$.

   Solution:

$$E[\bar{X}] = E[\frac{1}{5}(X_1 + X_2 + X_3 + X_4 + X_5)] = \frac{1}{5}(E[X_1] + \cdots + E[X_5]) = 0$$

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

Solution:

$$V[\bar{X}] = V[\frac{1}{5}X_1 + \cdots + \frac{1}{5}X_5)] = \frac{1}{5^2} \cdot V[X_1] + \cdots + \frac{1}{5^2}V[X_5]) = \frac{1}{5^2} \cdot 5 = \frac{1}{5},$$

where the second equality uses the independence of the X_i.

(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 α and β, you should specify the numerical values of α and β.)

Solution: The distribution of $\bar{X}$ is normal, since $\bar{X}$ is a linear combination of independent normal random variables. The parameters are $\mu = 0$ and $\sigma^2 = \frac{1}{5}$, as computed in parts (a) and (b).

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

Solution: The distribution of W is chi-squared, with 4 degrees of freedom, by Theorem 7.3 in Wackerly.

(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.)

Solution: U has the t-distribution, with 4 degrees of freedom, because

$$U = \frac{\sqrt{5}\bar{X}}{\sqrt{W/4}}$$

where the numerator is normal with mean 0 and variance 1, and the distribution of W is chi-squared with 4 degrees of freedom by the previous part. Now we can apply the definition of the t-distribution from Wackerly.

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 <u>significance level</u> is the smallest level of significance α for which the observed data indicate that the <u>null hypothesis</u> should be <u>rejected</u>.

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?

Solution: The null hypothesis H_0 is $\beta_{\texttt{Nearest}} = \beta_{\texttt{Scruz}} = \beta_{\texttt{Adjacent}} = 0$, and the alternative hypothesis H_a is that at least one of $\beta_{\texttt{Nearest}}$, $\beta_{\texttt{Scruz}}$, and $\beta_{\texttt{Adjacent}}$ is nonzero.

(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`.

Solution: `A` is 3 and `B` is 26.

(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`?

Solution: $P(F > 0.4182)$. Let's check this with R:

```
> pf(0.4182,3,26, lower.tail=FALSE)
[1] 0.741428
```

(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?

Solution: The null hypothesis is $H_0 : \beta_{\texttt{Scruz}} = 0$ and the alternative hypothesis is $H_a : \beta_{\texttt{Scruz}} \neq 0$.

(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.

Solution:

$$\frac{1.292094 - 1}{1.987029}$$

(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.

Solution: It has the t distribution with 26 degrees of freedom.

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?

Solution:

$$7 + (-2) \cdot 10 + 5 \cdot 40 = 187 \text{ pounds per year}$$

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

Solution: Holding income constant, teenagers who defined one extra word correctly tended to spend 2 pounds per year less on gambling.

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 $E[Y]$.

Solution: The centroid of the disc is $(E[X], E[Y])$ because the joint distribution is uniform. Thus $E[X] = E[Y] = 0$.

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

Solution: We need to find $E[XY]$, which is just $\frac{1}{\pi}\int_{x^2+y^2\leq 1} xydxdy$. But xy is positive in the first and third quadrants, and negative in the second and fourth quadrants, so this is also zero.

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.

Solution:

$$\hat{\beta}_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y}))}{\sum (X_i - \bar{X})^2}$$

(b) (3 points) Rewrite this formula in terms of the correlation between X and Y and the standard deviations σ_X, σ_Y.

Solution:

$$\hat{\beta}_1 = \frac{\mathrm{Cov}(X, Y)}{\sigma_X^2} = \frac{\rho_{XY} \sigma_X \sigma_Y}{\sigma_X^2} = \rho_{XY} \frac{\sigma_Y}{\sigma_X}$$

(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?

Solution: Zero, by the previous part and 7 (b).

(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.

Solution: Here is a list of the assertions:

1. This data generating process has rotational symmetry about the origin.
2. We should expect $\hat{\beta}_0 \approx 0$.
3. Any line through the origin is equally likely (as the regression line).
4. If we redid the regression with the same data, we might get a different answer.

The first two are true, and the second two are false. The third assertion is false because the error measurement process does not have rotational symmetry: we measure errors only in Y. The fourth assertion is false because the regression output is the result of a deterministic procedure and will always yield the same result given the same data.

The first assertion is true because the unit disc is rotationally symmetric about the origin and the distribution is uniform. The second assertion is true because the regression line goes through $(\bar{X}, \bar{Y})$.

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?

   Solution: Best.

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

   Solution: Variance-minimizing among the class of estimators (of β_0, β_1 respectively) that are unbiased and linear in the Y_i.

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 chi-squared-distribution with ν_1, ν_2 degrees of freedom respectively. Then the random variable

    $$F = \frac{W_1/\nu_1}{W_2/\nu_2}$$

    is said to have the F-distribution with ν_1 numerator degrees of freedom,

    and ν_2 denominator 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: $H = X(X^T X)^{-1} X^T$, and is called the hat matrix because the fitted values $\hat{Y}$ may be expressed using H as $\hat{Y} = HY$.

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

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?

Solution: A vector, since we are taking the median of each column.

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

Solution: `x0` is the vector whose components are the medians of each predictor variable (including the constant 1) on the right-hand-side of the regression equation. Since the regression includes a constant, the first component is 1.

(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.

Solution: The period means "all other variables in the data frame except for `brozek`". The `-siri -density -adipos -free` removes `siri, density, adipos, free` from this list of variables. This specifies the regression equation, which includes a constant by default.

(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?

Solution: The `prediction` interval is a confidence interval for a single future observation of `brozek` for a man whose predictor values match those of `x0`. The `confidence` interval is a confidence interval for the mean of many future observations of `brozek` for a collection of men all of whose predictor values match those in `x0`.

The value of a single future observation is the dot product of `x0` and the vector of estimated $\hat{\beta}_i$s plus a random variable ϵ, so the variance of this observation must account for the variance of the $\hat{\beta}_i$s plus the variance of ϵ. The averaging (largely) eliminates the variance of ϵ, reducing the width of the second interval.

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.

Solution: False, neither model contains the other, and only nested models can be compared with an F-test.

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

Solution: True, Model I contains Model II, so Model I can always have SSE at least as small as Model II by setting $\beta_3 = \beta_4 = 0$ and setting β_1 and β_2 to the coefficients in Model II.

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

Solution: True, because Model III contains Model 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?

Solution: $0.64 = (-0.8)^2$.

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

Solution: $\hat{\beta}_1$ must be negative, by 8 (b).