Math 455 **Solutions to Midterm 1** **Spring 2025**

No books, no notes, no calculators.

1. The standard simple linear regression model assumes a linear relationship between X and Y:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

(a) (4 points) Are β_0, β_1 random variables? If not, what are they?

Solution: No, they are fixed, but unknown constants.

(b) (4 points) How are the "fitted values" $\hat{Y}_i$ defined? (Hint: Think of the $\hat{\beta}_i$. You need not give the definition of the $\hat{\beta}_i$.)

Solution: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

(c) (4 points) What is the definition of the residuals e_i?

Solution: $e_i = Y_i - \hat{Y}_i$

(d) (6 points) What is the difference, if any, between the residuals e_i and the errors ϵ_i?

Solution: The conceptual difference is that the ϵ_i are the differences $\epsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$ between the Y_i and the corresponding point on the line whose slope and intercept are the fixed but *unknown* parameters β_0, β_1, whereas the $e_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$ are the difference between the Y_i and the corresponding point on the *regression line*, whose slope and intercept are the *estimators* $\hat{\beta}_0, \hat{\beta}_1$.

Here is a list of more concrete differences: one is that the residuals e_i satisfy $\sum_{i=1}^n e_i = 0$ and the ϵ_i need not satisfy this condition. A second difference: the e_i are observable (computable from the data), while the ϵ_i are not. A third difference: the ϵ_i have a joint distribution which is iid normal with parameters $0, \sigma^2$, while the the e_i have a multivariate normal distribution with variance-covariance matrix $\sigma^2(I - H)$, where H is the hat matrix.

2. (4 points) You obtain a list of 4 samples y_1, y_2, y_3, y_4 from a random variable Y which has either the t-distribution or the F-distribution (degrees of freedom unknown).

This list turns out to be $0.34, 4.36, -0.93, 1.37$. Is the random variable Y more likely to have the t- or F-distribution? Why?

Solution: The t-distribution, because the F-distribution does not take on negative values.

3. (4 points) Fill in the blanks in the following definition of p-value:

   If W is a test statistic, the p-value, or attained significance level, is the smallest level of significance α for which the observed data indicate that the null hypothesis should be rejected.

4. Consider the following (very small) data frame and the regression output and answer the following questions.

```
> df
         Y X1 X2
1 3.142333  3  2
2 4.818786  4  4
3 5.147295  5  8
4 6.519227  6 16

> lmod1 <- lm(Y~X1, data = df)
> summary(lmod1)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.2003     0.8152   0.246   0.8289
X1            1.0459     0.1758   5.949   0.0271 *
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3931 on 2 degrees of freedom
Multiple R-squared:  0.9465,Adjusted R-squared:  0.9198
F-statistic: XXXXX on 1 and 2 DF,  p-value: 0.02711
```

   (a) (4 points) The p-value `0.0271` marked with `*` is defined by a hypothesis test. What are the null hypothesis and the alternative hypothesis of this test?

   Solution: With β_1 denoting the coefficient of `X1`, we have H_0: $\beta_1 = 0$ and H_a: $\beta_1 \neq 0$.

   (b) (6 points) The F-statistic is obscured by `XXXXX` in the output, but in fact it is determined by the information available to you in this question. How can the F-statistic be computed from the information that you are given here? (You need not give an exact answer—you only need to explain what computation you would do if you had a computer.)

   Solution: In this setting, with only one variable at issue, the F-statistic is the square of the t-statistic. So the F-statistic is $5.949^2 \approx 36$.

It is also possible to get the answer by using the connection between the F-statistic and the R^2 value; see Wackerly, et. al. 11.14.

(c) (8 points) With the linear regression model `lmod1` as above, consider the following sequence of steps which R has performed for you to produce the output:

1. Compute the estimate `1.0459`
2. Compute the 2nd column value `0.1758`
3. Compute the 3rd column value `5.949`
4. Compute the 4th column value `0.0271`
5. Declare the result statistically significant at the level $\alpha = 0.05$.

Which, if any, is the first of these steps to use the assumption that the errors ϵ_i are normally distributed, and how is the assumption used?

Solution: The fourth step is the first one to use the assumption that the errors are normally distributed. To compute the p-value we must know that the value `5.949` comes from a t-distribution. To make this conclusion we use the expression for the t-statistic and the remarks after Definition 7.2 on p. 360 in Wackerly, et al.

Note that the second step does not use the assumption. In the second step we use the computation of the variance of $\hat{\beta}_1$ and that $SSE/(n-2)$ is an unbiased estimator of σ^2. These steps only use that the ϵ_i are independent and have variance σ^2 (see Wackerly p. 582).

The third step just divides the first column value by the second column value.

(d) (5 points) If, after executing the code above, we typed into R

```
> model.matrix(lmod1)
```

what would the output be?

Solution:

```
> model.matrix(lmod1)
  (Intercept) X1
1           1  3
2           1  4
3           1  5
4           1  6
```

5. Consider the following regression summary output from a study examining factors affecting fuel efficiency in cars. The variables considered are:

- weight, the car's weight in thousands of pounds,
- horsepower, the engine's horsepower, and
- mpg, the miles per gallon (MPG) of the car.

```
Call:
lm(formula = mpg ~ weight + horsepower, data=cars)

Coefficients:
              Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)    50.000     5.500     9.091   3.2e-09 ***
weight         -6.000     1.200    -5.000   1.1e-05 ***
horsepower     -0.050     0.020    -2.500   0.018
```

(a) (5 points) What is the predicted fuel efficiency (MPG) for a car that weighs 3.5 thousand pounds and has 200 horsepower?

Solution: $50 - 6 \cdot 3.5 - 200 \cdot 0.05 = 19$mpg

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

Solution: Among cars with the same horsepower, a car that weighs 1000 pounds more tends to have a mileage rating 6 mpg lower.

6. Suppose X and Y are random variables whose joint distribution is the uniform distribution on the unit square in the first quadrant of the plane. That is, the square with side length 1 whose corners have coordinates $(0,0), (0,1), (1,1), (1,0)$. (Hint: this problem can be done using symmetry, without any laborious calculations.)

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

Solution: $E[X] = \frac{1}{2}$ and $E[Y] = \frac{1}{2}$, since $(\frac{1}{2}, \frac{1}{2})$ is the centroid of the unit square.

(b) (6 points) Find $\mathrm{Cov}(X, Y)$, and explain why your answer is correct.

Solution: $\mathrm{Cov}(X, Y) = 0$.
The probability density function $f(x, y)$ is a product $h(x)h(y)$, where $h(x)$ is the function that is 1 if $x \in [0, 1]$ and 0 elsewhere. Therefore X and Y are independent, by definition of independence. Thus the correlation and covariance are 0.

7. 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) (8 points) Suppose now that our data (X_i, Y_i) are independent and identically distributed samples from a uniform distribution on the unit the plane square in the first quadrant of the plane, as in the previous problem. If $n = 9999$, what approximate values do you expect to see in the regression output for $\hat{\beta}_0, \hat{\beta}_1$, and why?

Solution: Since we have a large number of points, we would guess that the empirical correlation is very close to the theoretical correlation, and that the empirical means $\bar{X}, \bar{Y}$ are close to the theoretical means. From the previous problem, we know the theoretical means are both 1/2 and that the theoretical correlation is 0.
Applying the formula $\hat{\beta}_1 = \rho_{XY}\frac{\sigma_Y}{\sigma_X}$, we obtain 0 for $\hat{\beta}_1$. Now use the fact that $(\bar{X}, \bar{Y})$ is on the regression line and we get 1/2 for $\hat{\beta}_0$.

(b) (12 points) Another student asserts, "The data generating process is symmetric about the line $y = x$, so the regression line is $y = x$. Thus we should expect an approximate result of 0 for $\hat{\beta}_0$ in the regression output, and an approximate result of 1 for $\hat{\beta}_1$." Comment on these four assertions. Your comment should address whether each assertion is true and why.

Solution:
First we enumerate the assertions:

1. The data generating process is symmetric about the line $y = x$.
2. The regression line is $y = x$.
3. We expect an approximate result of 0 for $\hat{\beta}_0$.
4. We expect an approximate result of 1 for $\hat{\beta}_1$.

The first assertion is true, but the remaining 3 are false.
The truth of the first assertion is clear: the unit square is symmetric about its diagonal; draw a picture if necessary.
The second assertion is false. In fact, the data generating process is symmetric about many lines, and not all of them can be the regression line. In addition, the regression line need not respect symmetries of the data generating process.
The third and fourth assertions are false, because we found the regression line in the previous problem, and these are not the coefficients.

8. Let X_1, X_2, X_3, X_4 be independent and identically distributed random variables which are all normal with mean 0 and variance 1. Let $\bar{X} = \frac{1}{4}(X_1 + X_2 + X_3 + X_4)$. Let $W = \sum_{i=1}^{4}(X_i - \bar{X})^2$.

(a) (4 points) What is the distribution of W? (Give the values of any parameters that are relevant.)

Solution: W has the χ^2 distribution with 3 degrees of freedom.

(b) (4 points) Notice that $\bar{X}$ appears in the formula defining W. Are $\bar{X}$ and W independent random variables? How do you know?

Solution: Yes, $\bar{X}$ and W independent random variables. This is a result from Math 447. See Theorem 7.2 in Wackerly, et. al.

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

   (a) (4 points) What do B, L, U, and E stand for?

   Solution: Best Linear Unbiased Estimators

   (b) (4 points) Explain the meaning of your answer for L above.

   Solution: L means "linear in the Y_i." Note that the "in the Y_i" part is important! The estimators are obviously *not* linear in the X_i.

   (c) (4 points) Explain the meaning of your answer for B above.

   Solution: Best means "lowest variance among unbiased estimators of β_0, β_1 which are linear in the Y_i". Note again the "in the Y_i" part: the statement would not be true without it!

Multiple Choice: Circle the letter corresponding to the best response.

10. (4 points) In a least-squares regression model with an intercept, expressed in matrix notation as $Y = X\beta + \epsilon$, the residual vector is always orthogonal to which of the following?

    (A) The response vector Y

    (B) The fitted values vector $\hat{Y}$

    (C) The column space of the design matrix X

    (D) The null space of $X^T X$

Solution: (C)

11. (4 points) The hat matrix H in a regression is defined as

$$H = X(X^T X)^{-1} X^T.$$

    What are the properties of H?

    (A) It is symmetric and idempotent.

    (B) It is diagonal.

    (C) It is always invertible.

    (D) It has eigenvalues strictly greater than zero.

Solution: (A)

12. (4 points) The ordinary least squares estimator is unbiased under which of the following conditions?

(A) The residuals are normally distributed.

(B) The independent variables are uncorrelated.

(C) The expectation of the error term is zero, given the independent variables.

(D) The independent variables have equal variance.

Solution: (C)

13. (4 points) In a multiple regression setting, which of the following correctly describes the variance of an estimated coefficient?

(A) It decreases when more predictors are added, regardless of relevance.

(B) It is independent of the correlations among predictors.

(C) It depends on the sample size and the variance of the predictor.

(D) It is equal to the standard error of the residuals.

Solution: (C)

14. (4 points) In an ordinary least squares (OLS) regression, what are the fitted values?

(A) The values of the response variable Y_i that minimize the sum of squared residuals

(B) The predicted values $\hat{Y}_i$ obtained from the estimated regression model

(C) The residuals after removing the influence of the predictors

(D) The true values of the response variable in the population

Solution: (B)

15. (4 points) How is the vector of fitted values $\hat{Y}$ computed in an OLS regression model?

(A) By adding the residuals to the observed response values

(B) By multiplying the design matrix X by the estimated coefficient vector $\hat{\beta}$

(C) By taking the inverse of the covariance matrix of the predictors

(D) By subtracting the estimated residuals from the response variable

Solution: (B)

16. (4 points) The difference between a confidence interval for the mean response and a prediction interval for a new observation is that:

(A) A confidence interval is always wider than a prediction interval.

(B) A prediction interval accounts for both the uncertainty in estimating the mean response and the variability of future observations.

(C) A confidence interval is used for predicting new observations, while a prediction interval is used for estimating the population mean.

(D) A prediction interval depends only on the sample size, while a confidence interval depends on both the sample size and variability of residuals.

Solution: (B)

17. (4 points) Which of the following best explains why a prediction interval is wider than a confidence interval?

(A) A prediction interval accounts for both the uncertainty in the estimated regression function and the natural variability in future observations.

(B) A prediction interval uses a smaller significance level than a confidence interval.

(C) A confidence interval assumes that future observations will be close to the mean response.

(D) A prediction interval is based on a different test statistic than a confidence interval.

Solution: (A)

18. (4 points) What is the primary source of additional uncertainty in a prediction interval compared to a confidence interval?

(A) Sampling variability in estimating $\hat{\beta}$

(B) The residual variance, which accounts for random variation in individual observations

(C) The presence of collinearity in the predictors

(D) The need to estimate the variance of the independent variables

Solution: (B)

19. (4 points) Which of the following Hill criteria suggests that the relationship between an exposure and outcome should be seen in multiple studies and populations?

A) Plausibility

B) Specificity

C) Consistency

D) Analogy

Solution: (C)

20. (4 points) The Hill criterion *gradient* refers to:

A) The effect disappearing when the exposure is removed

B) A dose-response relationship, where higher exposure leads to a greater effect

C) A plausible explanation for why the effect occurs

D) A strong correlation coefficient

Solution: (B)

21. (4 points) Which Hill criterion is concerned with whether the observed relationship aligns with existing biological or theoretical knowledge?

A) Strength of association

B) Plausibility

C) Specificity

D) Experiment

Solution: (B)

22. (4 points) In a simple linear regression analysis with one explanatory variable and n data points (X_i, Y_i), which one of the following quantities is the same for all $i = 1, \ldots, n$?

A) The i-th diagonal entry of the hat matrix

B) The standard deviation of Y_i

C) The standard deviation of the residual e_i

D) The standard deviation of $\hat{Y}_i$

Solution: (B)

23. (4 points) In a linear regression analysis with n rows of data, with the usual assumptions, which of the following quantities can never be zero?

A) The estimated intercept, $\hat{\beta}_0$

B) The first residual e_1

C) The last fitted value, $\hat{Y}_n$

D) All of the above quantities can be zero.

Solution: (D)