IDS 702 Midterm - KEY

Fall 2025

First name: _____ Last name: _____
Net ID: _____

I hereby state that I have not communicated with or gained information in any way from my classmates or unauthorized materials during this exam, and that all work is my own.

Signature: _____

Any potential violation of Duke's policy on academic integrity will be reported to the Office of Student Conduct & Community Standards. All work on this exam must be your own.

1. You have 75 minutes to complete the exam.
2. You are **not** allowed a cell phone (even if you intend to use it for checking the time), music device or headphones, notes, books, or other resources, or to communicate with anyone other than the professor or TAs during the exam.
3. Write clearly.

The exam has 15 questions: 12 multiple choice questions and 3 short answer questions. Question 14 has parts a-b and question 15 has parts a-e. There is also an optional bonus question. **Before you begin, make sure your exam has all questions.**

Formula Page

$Pr(A|B) = \frac{Pr(A \cap B)}{Pr(B)}$

$Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)$

Independence: $Pr(A|B) = Pr(A)$, $Pr(A \cap B) = Pr(A)Pr(B)$

Bayes Theorem: $Pr(B|A) = \frac{Pr(A|B)Pr(B)}{Pr(A)}$

Total probability: $Pr(A) = Pr(A|B)Pr(B) + Pr(A|B^c)Pr(B^c)$

General formula for confidence interval: $\bar{x} \pm z * \times \sigma / \sqrt{n}$

Multiple Choice (3 points each)

Circle the letter to mark your answer choice. Choose the single best answer for each question.

1. Which of the following is an accurate property of probability?

   a. Probability is always between -1 and 1, inclusive

   b. Marginal probability is the probability of an event given that another event has occurred

   c. Joint probability is the probability of an event regardless of another event

   d. **If knowing event A has occurred does not lead to any change in the probability of event B, then events A and B are independent**

2. A bank that offers small business loans wants to determine the probability that less than 5% of its 1400 new clients will default on their loan this year. Based on its 2024 loan metrics, the bank will assume that each loan has a default probability of 0.035. Which probability distribution fits this scenario and what are that distribution's parameters?

   a. Normal; $\mu = 1400$, $\sigma = 5$

   b. Normal; $\mu = 1400$, $\sigma = 0.035$

   c. **Binomial;** $n = 1400$**,** $p = 0.035$

   d. Binomial; $n = 1400$, $p = 0.05$

3. Which of the following does NOT affect the width of a theoretical confidence interval?

   a. Sample size

   b. Sample standard deviation

   c. **Sample statistic**

   d. Confidence level

4. When conducting a hypothesis test, the null and alternative hypotheses are always in terms of the _____.

   a. sample statistic

   b. **population parameter**

   c. probability distribution

   d. population mean

5. Which quantities are considered fixed (as opposed to random) in the theoretical representation of the simple linear regression model? **(Answer: B)**

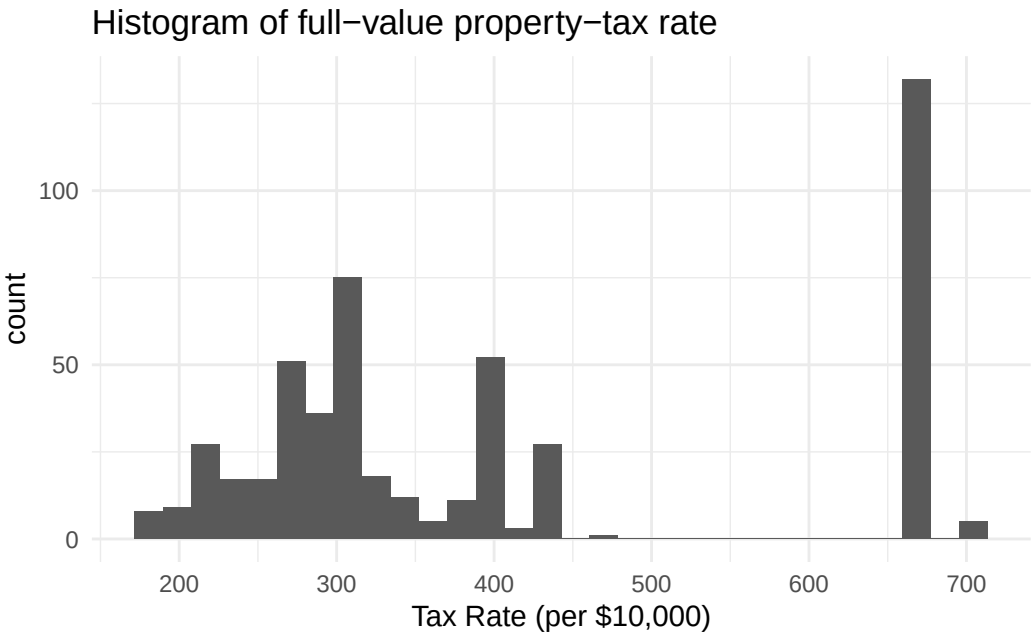
   a. Y, X, β_1, and ϵ

   b. **X, β_1, and σ^2**

   c. Y and ϵ

   d. β_1, ϵ, and σ^2

6. The dataset `Boston` contains housing values for 506 suburbs of Boston. Based on the histogram of the full-value property-tax rate per \$10,000 for each suburb, can you characterize the distribution of the sample mean of the property tax rate? Assume that each suburb is independent.



   a. Yes, the sample mean would have a bimodal distribution

   b. **Yes, the sample mean would have a normal distribution**

   c. Yes, the sample mean would have a distribution centered around 300 but with many outliers at ~675.

   d. No, we would need to know the population distribution of the tax rate to characterize the distribution of the sample mean

7. Which type of question can NOT be answered with a linear model?

   a. On average, how much does Y change per increase in the variable X?

   b. What is the predicted value of Y for a given value of X?

   c. What is the average value of Y when $X = 0$?

   d. **On average, per increase in Y, how much does X change?**

8. What distinguishes R^2 and adjusted R^2?

    a. R^2 is used to evaluate multiple linear regression models, and adjusted R^2 is used to evaluate simple linear regression models.
    b. **R^2 is the proportion of variability explained, and adjusted R^2 adds a penalty term for the number of predictors in a multiple linear regression model.**
    c. R^2 can be used to evaluate any linear regression model and adjusted R^2 is only used to evaluate models with interaction terms.
    d. R^2 can be negative in some cases, but adjusted R^2 is always between 0 and 1.

9. In linear regression, what is a residual?

    a. **The difference between the observed value and the estimated line**
    b. The difference between the true population value and the estimated line
    c. The difference between the true population value and the true population line
    d. The difference between the observed value and the true population line

10. A researcher fits a model regressing house listing price on number of bedrooms, number of bathrooms, distance to nearest grocery store (stored categorically as 0-1 mile, 1-5 miles, or more than 5 miles), and presence or absence of a homeowner's association (stored categorically as yes/no). The researcher has data on 1500 homes. What is the dimension of the design matrix for this model? **(Answer: C)**

    a. 1500×4
    b. 1500×5
    c. 1500×6
    d. 1500×8

11. What is the difference between interpreting coefficient estimates for a numeric variable and a categorical variable in a regression model?

    a. **Unlike estimates for numeric variables, estimates for categorical variables are interpreted in comparison to a reference level**
    b. Unlike estimates for categorical variables, estimates for numeric variables are interpreted in comparison to a reference level
    c. Unlike estimates for numeric variables, estimates for categorical variables are interpreted as the increase in the outcome per unit increase in the predictor
    d. There is no difference in the interpretations of estimates for numeric and categorical variables

12. Which of the following metrics evaluates the influence of individual observations on the regression as a whole?

    a. Leverage

    b. **Cook's Distance**

    c. Variance Inflation Factor

    d. R^2

Short Answer

Question 13 (9 points)

For each term/symbol listed below, decide whether it is most closely associated with the **population**, the **sample/data**, or the **sampling distribution**. Write each term under the correct heading.

1. Parameter
2. Observed statistic
3. Central Limit Theorem
4. Standard error
5. σ
6. $\bar{x}$
7. Sample size n
8. Null distribution
9. Null hypothesis

Population	Sample/Data	Sampling Distribution
Parameter	Observed statistic	CLT
σ	$\bar{x}$	Standard error
Null hypothesis	Sample size n	Null distribution

Question 14 (10 points)

Let $X_1, X_2, ..., X_n \overset{iid}{\sim} Geometric(p)$, where $Pr(X = x) = (1-p)^{x-1}p$ for $x = 1, 2, 3, ...$

a. Show that the likelihood function is $L(p) = p^n(1-p)^{(\sum_{i=1}^n x_i)-n}$

Joint distribution: $\prod_{i=1}^n (1-p)^{x_i-1}p$

Simplifying, we have $L(p) = p^n \prod_{i=1}^n (1-p)^{x_i-1} = p^n(1-p)^{\sum_{i=1}^n (x_i-1)} = p^n(1-p)^{\sum_{i=1}^n (x_i)-n}$

b. Show that the maximum likelihood estimate is $\hat{p} = \dfrac{1}{\bar{X}}$

Taking the log of the likelihood, we have $l(p) = n\log(p) + (\sum_{i=1}^n x_i - n)\log(1-p)$

Then we take the derivative of the log-likelihood: $\frac{dl}{dp} = \frac{n}{p} - \frac{\sum_{i=1}^n x_i - n}{1-p}$

Set to 0 and solve for p: $\frac{n(1-p)-p(\sum_{i=1}^n x_i - n)}{p(1-p)} = 0$

$n - np - p\sum_{i=1}^n x_i + np = 0 \implies n = p\sum_{i=1}^n x_i \implies \hat{p} = \frac{n}{\sum_{i=1}^n x_i} \implies \hat{p} = \dfrac{1}{\bar{X}}$

Question 15 (24 points)

The Carseats dataset contains information on sales of child carseats at 400 stores. You want to understand how certain factors are related to sales.

Codebook:

`Sales` Unit sales (in thousands) at each location

`Advertising` Local advertising budget for company at each location (in thousands of dollars)

`Price` Price company charges for car seats at each site

`ShelveLoc` A factor with levels `Bad`, `Good` and `Medium` indicating the quality of the shelving location for the car seats at each site

`Age` Average age of the local population

You fit a model regressing unit sales on advertising budget, average age, and an interaction term for price and shelving location.

a. What research question does the interaction term assess in this model?

Does the relationship between price and sales depend on shelving location?

b. Write the theoretical model. Be sure to define the terms in the model.

$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_3 x_4 + \beta_7 x_3 x_5 + \epsilon, \epsilon \sim N(0, \sigma^2)$

Y = sales; x_1 = Advertising budget; x_2 = Average age; x_3 = price

$x_4 = 1$ if ShelveLoc=Good, 0 otherwise

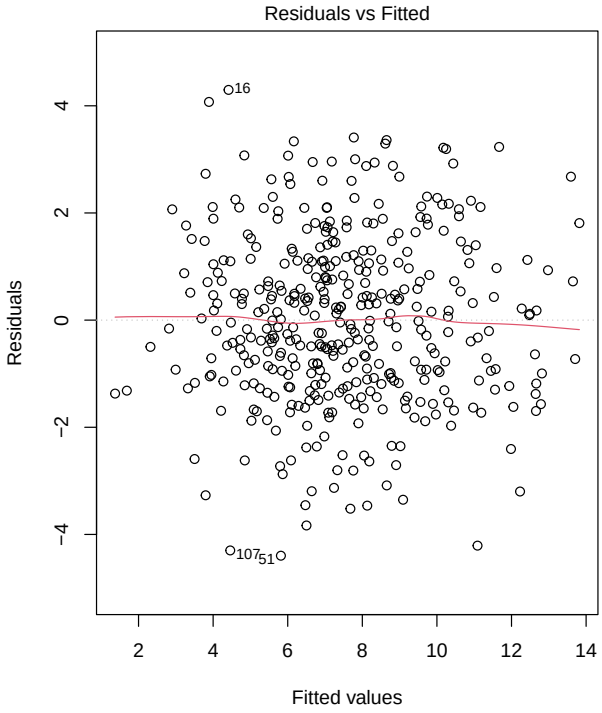
$x_5 = 1$ if ShelveLoc=Medium, 0 otherwise

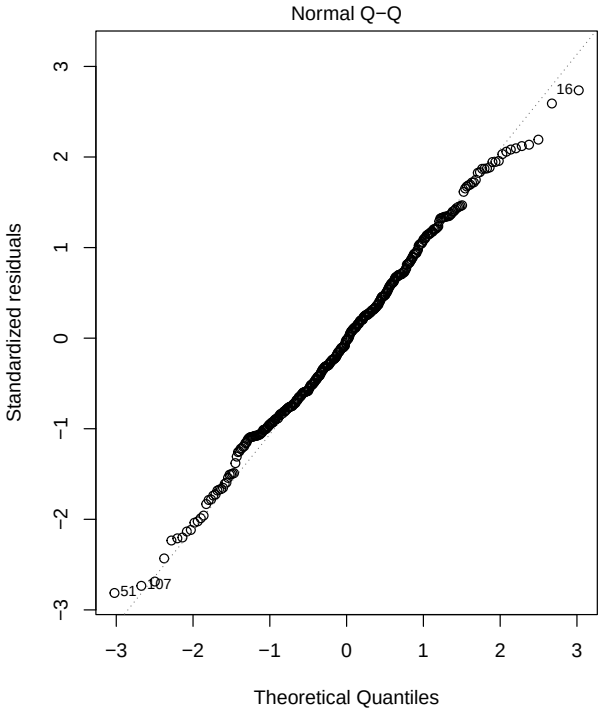
c. Using the coefficient table provided below, write an interpretation for the coefficient estimate of Price.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	15.201271541	0.887804715	17.1223145	1.799024e-49
Advertising	0.108820503	0.011963374	9.0961384	4.765309e-18
Age	-0.050521545	0.004969894	-10.1655181	1.058384e-21
Price	-0.067605447	0.006923915	-9.7640493	2.657378e-20
ShelveLocGood	4.505449776	1.157584694	3.8921124	1.167616e-04
ShelveLocMedium	0.748110900	0.972244803	0.7694676	4.420792e-01
Price:ShelveLocGood	0.002838802	0.009762804	0.2907773	7.713754e-01
Price:ShelveLocMedium	0.010379444	0.008316370	1.2480738	2.127490e-01

Controlling for age and advertising, per unit increase in price, on a bad shelving location, sales decrease by 0.068, on average.

d. Describe how you would use the diagnostic plots below to assess three linear regression assumptions. First name the assumption, then name the most relevant plot for that assumption, describe what you look for to assess a violation, and state whether or not you see strong evidence that the assumption is violated in this case.



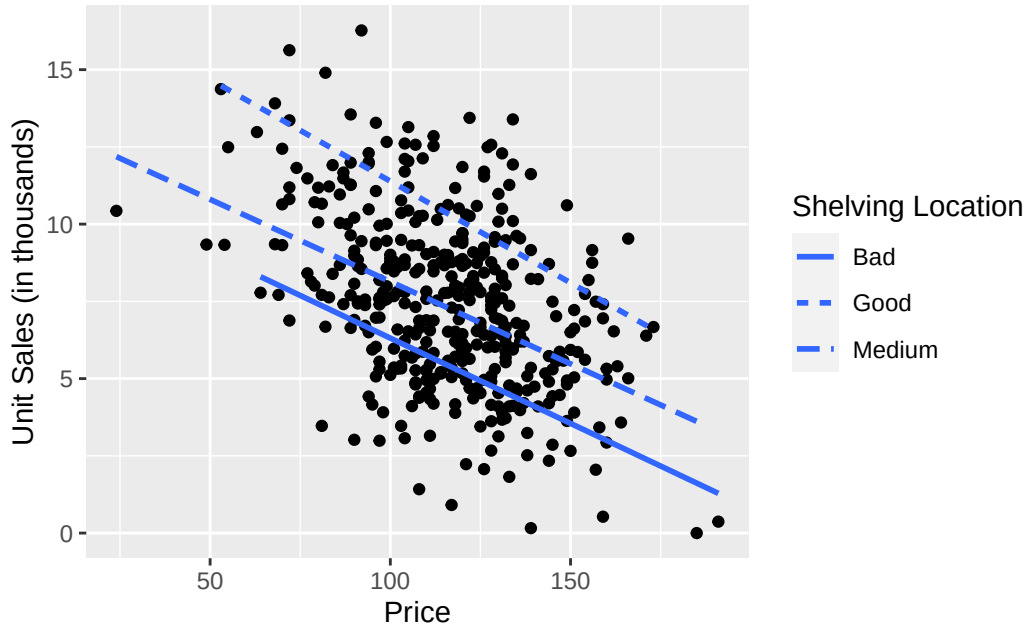


i. Linearity: Residuals vs fitted plot, looking for cloud of points, no clear pattern, straight fitted line across. No strong evidence of violation here

ii. Equal variance/homoscedasticity: Residuals vs fitted plot, looking for unequal spread moving left to right (i.e., "fanning"). No strong evidence of violation here

iii. Normality of errors: QQ plot, looking to see if points fall along the 45 degree line. No strong evidence of violation here

e. Based on the plot and the model metrics provided below, are there any changes you would make to this model? Explain your answer.




```
summary(lm(Sales ~ Advertising+Age+Price*ShelveLoc, data=Carseats))$adj.r.squared
```

```
[1] 0.6855111
```

```
summary(lm(Sales ~ Advertising+Age+Price+ShelveLoc, data=Carseats))$adj.r.squared
```

```
[1] 0.6856247
```

I would consider dropping the interaction term. The lines in the plot are nearly parallel, so price doesn't seem to be interacting with shelving location very much. The adjusted R^2 values for the models with and without the interaction term are also nearly identical, indicating no substantial improvement in model performance with the interaction term. That said, if two variables are known to interact *a priori,* we may choose to leave it in the model.

Bonus (5 points)

Show that $Var(X) = \mathbf{E}[X^2] - \mathbf{E}[X]^2$

$Var(X) = \mathbf{E}[(X - \mathbf{E}[X])^2] = \mathbf{E}[X^2 - 2X\mathbf{E}[X] + \mathbf{E}[X]^2] = \mathbf{E}[X^2] - 2\mathbf{E}[X]^2 + \mathbf{E}[X]^2 = \mathbf{E}[X^2] - \mathbf{E}[X]^2$

Blank Page - End of exam