STAT340 Final exam

Points:

MC1-6 (/12)	MC7-9 (/6)	SA1,2 (/8)	SA3 (/4)	SA4,5 (/8)	SA6 (/4)	Total (/42)

First (given) name:

Write here: _____

Last (family) name:

Write here: _____

Rules:

- You must show work for all computations (unless otherwise specified) to receive full credit.
- You do NOT need to simplify any expressions you write down.

Multiple choice 2pts each

MC1,2

Let X and Y be independent random variables such that $E(X) = -1$, $E(Y) = 3$, $Var(X) = 1$, $Var(Y) = 2$. Which of the following values is ***closest*** to $E(3X - 2Y)$?

a. -10
b. -5
c. 0
d. 5
e. 10

Which of the following values is ***closest*** to $Var(3X - 2Y)$?

a. -10
b. -5
c. 0
d. 5
e. 10

MC3,4

Let $X_1, X_2, \ldots, X_n$ be an independent and identically distributed sample such that X_i has mean μ and variance σ^2. As n increases, how does your sample variance (i.e. $\frac{1}{n-1}\sum(X_i - \bar{X})^2$) tend to change?

a. Increases proportional to n
b. Increases proportional to $\sqrt{n}$
c. Does not tend to change
d. Decreases proportional to $\sqrt{n}$
e. Decreases proportional to n

As n increases, how does the variance of the sample mean $\bar{X}$ tend to change?

a. Increases proportional to n
b. Increases proportional to $\sqrt{n}$
c. Does not tend to change
d. Decreases proportional to $\sqrt{n}$
e. Decreases proportional to n

MC5

All else held equal, in a simple linear regression context, which of the following is **true** as σ^2 increases? ***Choose ALL that apply!*** (i.e. the number of right choices is ≥1)

a. $\hat{\beta}_1$ tends to increase
b. $\mathrm{SE}(\hat{\beta}_1)$ tends to increase
c. RSE (residual standard error) tends to increase
d. R^2 tends to increase
e. Df (degree of freedom) tends to increase

MC6

Which of the following is NOT an assumption of a linear regression model?

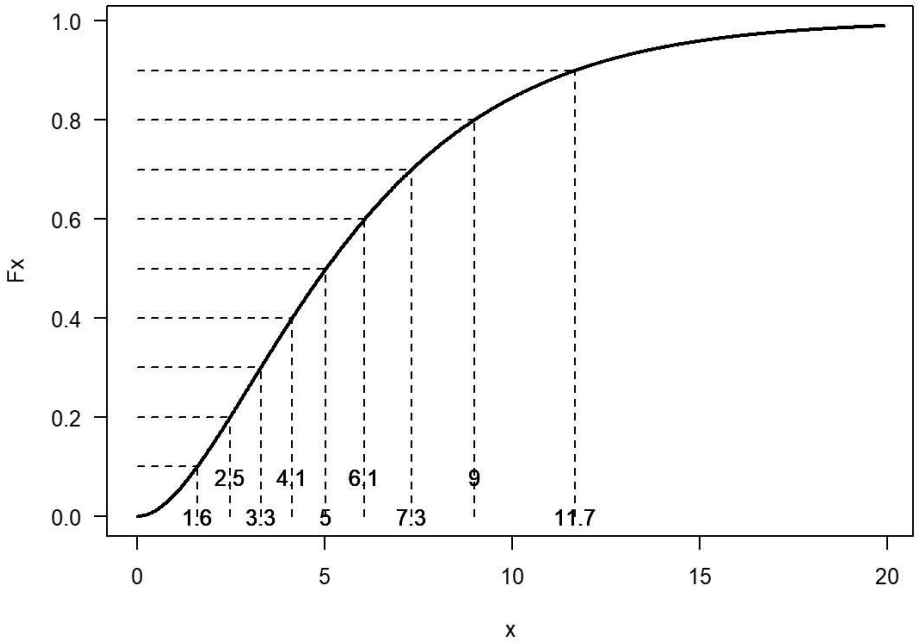
a. The response variable has a linear relationship with the predictor variables.
b. The predictor variables are normally distributed.
c. The errors are normally distributed.
d. The errors have constant variance.
e. The errors are independent.

MC7

You decide to buy a lottery ticket every day until you win the jackpot. Assume you have an infinite line of credit (i.e. your credit card lets you buy an unlimited number of lottery tickets). Which of the following is the best random variable to use to model the number of tickets you end up buying?

a. Normal
b. Binomial
c. Poisson
d. Geometric
e. Exponential

MC8



We generate observations from the random variable defined by the CDF above F. We will generate a sequence or random variables $U_1, U_2, \ldots, U_n \sim Unif(0,1)$ (iid), and then let $X_i = F^{-1}(U_i)$. Which of the following intervals do you expect to see the most number of observations X_i fall in?

a. $(-\infty, 2.5)$
b. $(2.5, 4.1)$
c. $(5, 9)$
d. $(9, \infty)$
e. Unable to determine

MC9

A statistics instructor gives each of the 160 students in a class a different random data set. Each data set is obtained by doing `rnorm(n=20, mean=5, sd=4.5)`. The students are told that $\sigma = 4.5$, but they do NOT know $\mu = 5$. The students are tasked to use a Monte Carlo test to test the hypotheses $H_0 : \mu = 5$ vs $H_1 : \mu \neq 5$ at $\alpha = 0.05$. Which of the following is true?

a. Using as many Monte Carlo replications as computationally possible will lower the standard error of the point estimate.
b. Using as many Monte Carlo replications as computationally possible will improve the power of the test.
c. Use a t-statistic $\frac{\bar{x}-5}{s/\sqrt{20}}$ instead of a z-statistic $\frac{\bar{x}-5}{4.5/\sqrt{20}}$ increases the power since s is based on the data.
d. We expect to see on average 8 students committing a type I error.
e. None of the above are true.

Short answer 4pts each

SA1

A **fair coin** is flipped 4 times. Assume each flip has no influence on any other flip. Let A and B denote the following two events:

- A: There are 2 heads in total out of 4 flips.
- B: The first flip is a head.

Answer each of the following, **showing all work for full points**. If it helps you, here's a list of every possible outcome:

```
HHHH, HHHT, HHTH, HHTT, HTHH, HTHT, HTTH, HTTT,
THHH, THHT, THTH, THTT, TTHH, TTHT, TTTH, TTTT.
```

a. What is $P(A)$?
b. What is $P(B)$?
c. What is $P(A \& B)$? (i.e. what is the probability of both events occurring simultaneously?)
d. Are the events dependent or independent? Explain.

SA2

Each statement below may or may not be correct. For each statement, identify if it's correct or incorrect and explain why. If it is incorrect, rewrite the statement to be correct.

a. 0.01 is always a better value to use for α than 0.05 because it gives a lower rate of false positives.
b. For a computed 95% confidence interval for μ, there is a 95% chance that μ is contained in the interval.
c. You can decide whether or not to include an interaction term in a model by checking if the two predictors are correlated in the data.
d. If two events are independent, then they are also mutually exclusive.

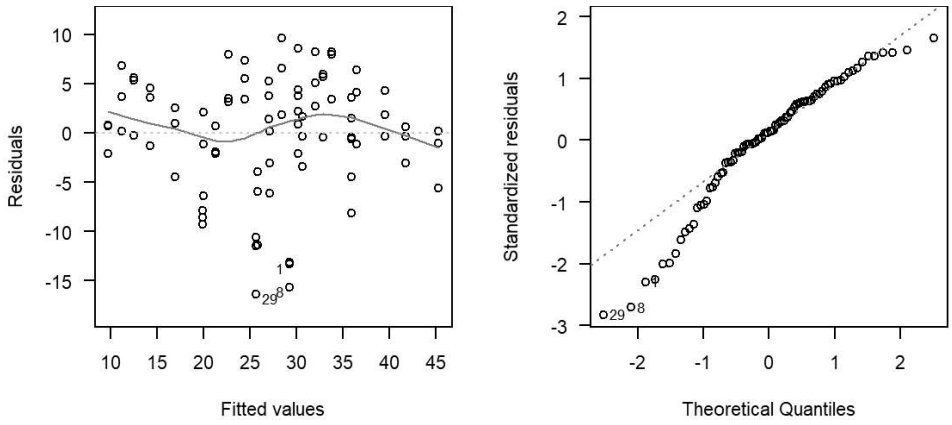
SA3

We have a dataset of CO2 uptake levels in a certain grass species under different temperature conditions. Here are the variables:

- `uptake` : numeric response measuring amount of CO2 uptake of each sample
- `Type` : categorical predictor (levels: “Quebec”, “Mississippi”) denoting location where the sample was originally from
- `Treatment` : categorical predictor (levels: “nonchilled”, “chilled”) denoting temperature treatment applied to the sample
- `conc` : numeric predictor denoting level of ambient CO2 the sample was kept in.

Below is the output of a multiple linear regression fit and the diagnostic plots. You may reference this output as justification in your answers below, but PLEASE clearly state which numbers you are referring to. Please **show all work for full points**.

```
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     27.620528   1.627945  16.966  < 2e-16 ***
## TypeMississippi                 -9.380952   1.851185  -5.068 2.59e-06 ***
## Treatmentchilled                -3.580952   1.851185  -1.934   0.0566 .
## conc                             0.017731   0.002225   7.969 1.00e-11 ***
## TypeMississippi:Treatmentchilled -6.557143   2.617972  -2.505   0.0143 *
##
## Residual standard error: 5.999 on 79 degrees of freedom
## Multiple R-squared:  0.7072, Adjusted R-squared:  0.6923
## F-statistic: 47.69 on 4 and 79 DF,  p-value: < 2.2e-16
```



a. Construct a 95% confidence interval for the interaction term and **interpret it**. You may use the normal approximation to the t-distribution due to large sample size (i.e. use 1.96 as your critical value).
b. What response would you expect to observe on average for a new sample of Mississippi type, nonchilled treatment, and with concentration 500? Write an expression for the answer (you do NOT need to simplify the expression to a single number!)
c. Approximately what proportion of the change in response is explained by the change in predictors?
d. Looking at the diagnostic plots, is there evidence of model assumption violations? Explain.

SA4

You are a scientist working for big pharma developing a test for a disease. Define the following variables:

- Let p denote the prevalence of the disease in the population of interest, i.e. $P(\text{disease})$
- Let α denote the false positive rate of the disease, i.e. $P(\text{positive test} \mid \text{no disease})$ (this is also equivalent to 1–specificity)
- Let s denote the sensitivity of the test, i.e. $P(\text{positive test} \mid \text{disease})$ (this is also 1–false negative rate)

Suppose you gather a sample of n subjects for a study. Assume the sample is very representative of the overall population of interest. On average, how many subjects would you expect to be in each category below? (Write an expression giving the **expected average COUNT of subjects in each category** out of n total subjects in the sample.)

a. True positives
b. True negatives
c. False positives
d. False negatives

SA5

This question is based on the manifest (i.e. list of passengers) of the Titanic. Here are the relevant columns:

- `Survived` This is a categorical response, 1 indicating the passenger survived, 0 indicating the passenger died
- `Pclass` This is a categorical predictor indicating the class of the passenger's ticket (i.e. 1st class, 2nd class, 3rd class)
- `Sex` This is also a categorical predictor indicating the sex of the passenger
- `Age` This is a numerical predictor

```
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.777013   0.401123   9.416  < 2e-16 ***
## Pclass2     -1.309799   0.278066  -4.710 2.47e-06 ***
## Pclass3     -2.580625   0.281442  -9.169  < 2e-16 ***
## Sexmale     -2.522781   0.207391 -12.164  < 2e-16 ***
## Age         -0.036985   0.007656  -4.831 1.36e-06 ***
```

a. Which of the predictors appear to be the most significant in this model? Give an interpretation of one of these coefficients.
b. Give a 95% confidence interval for the male coefficient and interpret the interval.
c. Suppose you are a 20-year old male passenger in 1st class. Write an expression for your predicted log-odds of survival.
d. Convert the log-odds of survival from part c to a probability of survival.

SA6

Suppose you are given a vector `data` with $n = 100$ observations X_i of a poisson process (i.e. each observation is a count of the number of occurrences of some event in a fixed interval). Write a sequence of instructions using **pseudo-code** (i.e. a mix of some R code as well as some English descriptions is allowed) that will **compute a 95% confidence for λ interval using a Monte-Carlo based method**.

Your pseudo-code should be specific enough so that someone who has a BASIC understanding of the R programming language and its functions but has NO formal training in statistics should be able to follow your instructions and produce the correct computation. Also, **every step/command/instruction should at LEAST mention what specific R function to use**. Any R expressions you write do NOT have to perfectly evaluate in an R console to receive full credit, but **your response SHOULD show both a clear understanding of the statistical methodology as well as at least a BASIC understanding of R functions and syntax**.

As an example, saying something like “compute a point estimate of lambda” would be considered too vague, but saying something like “use `mean()` to find the mean of `data` and store as `lambda_hat` ” is acceptable. Also note: There are multiple possible solutions to this problem, and your number of steps may differ from someone else’s number of steps even if you use the same method. This is completely ok, as long as both responses are clear, complete, and specific enough. You may also add additional annotation/commentary in each step to help clarify your thought process to the graders, such as in step 1 below which has already been done for you to help get you started.

1. Use `mean()` to find the mean of `data` and store as `lambda_hat` . This is our sample estimate of $E(X) = \lambda$.

…