Midterm Exam
STA 207 Fall 2024
Prof. Priya Kohli

Name: P. Kohli

Signature: _____

Instructions:

1. Sign on this paper. By doing this you are agreeing to the HONOR CODE.

2. You have 75 minutes for this exam.

3. There are 7 questions in the exam and a BONUS problem. The exam is worth 25% of your final grade.

4. You can use calculator for this exam.

5. Show all steps to receive full credit for each problem.

6. Please write your answers in the exam paper and return it with all pages.

Good Luck!!!

Problem 1 [35 points]: First Year GPA data contains information from a sample of 219 first year students at a midwestern college that might be used to build a model to predict their first year GPA.

There are 10 variables in this data:

- GPA: First-year college GPA on a 0.0 to 4.0 scale
- HSGPA: High school GPA on a 0.0 to 4.0 scale
- SATV: Verbal/critical reading SAT score
- SATM: Math SAT score
- Male: 1= male, 0= female
- HU: Number of credit hours earned in humanities courses in high school
- SS: Number of credit hours earned in social science courses in high school
- FirstGen: 1= student is the first in her or his family to attend college, 0=otherwise
- White: 1= white students, 0= others
- CollegeBound: 1=attended a high school where >=50% students intended to go on to college, 0=otherwise

Using simple linear regression model to regress first year GPA on the high school GPA.

R-output

```
## lm(formula = GPA ~ HSGPA)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.17985    0.26194   4.504 1.09e-05 ***
## HSGPA        0.55501    0.07542   7.359 3.78e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4174 on 217 degrees of freedom
## Multiple R-squared:  0.1997, Adjusted R-squared:  0.196
## F-statistic: 54.15 on 1 and 217 DF,  p-value: 3.783e-12
```

```
confint(SLR,level = 0.95)
```

```
##                2.5 %    97.5 %
## (Intercept) 0.6635865 1.6961148
## HSGPA       0.4063587 0.7036663
```

Formulas:

1.) $t = \frac{\widehat{\beta}_i}{se_{\widehat{\beta}_i}}$ for i=0 and i=1. Here t is the t test statistic and se is the standard error.

2.) residual=$Y - \hat{Y}$

Plots:

Fig1: Scatterplot

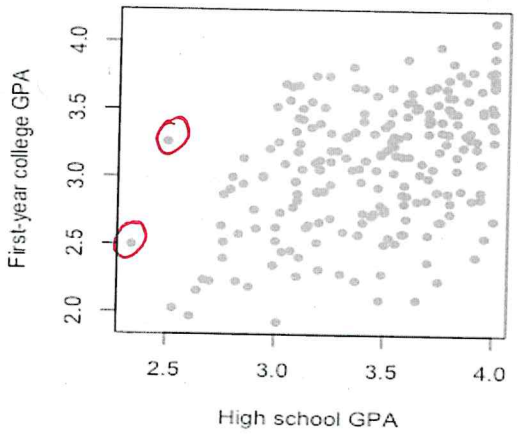


Fig2: Scatterplot with Fitted Model

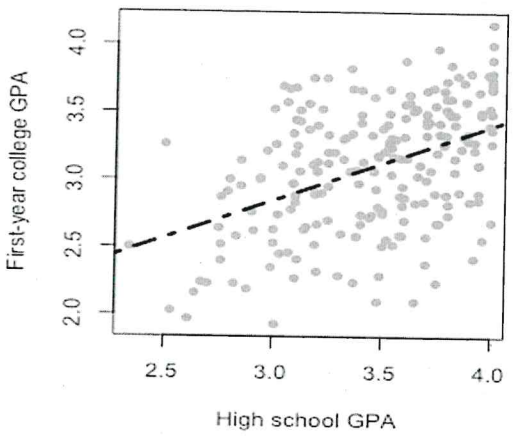


Fig3: Residual Plot

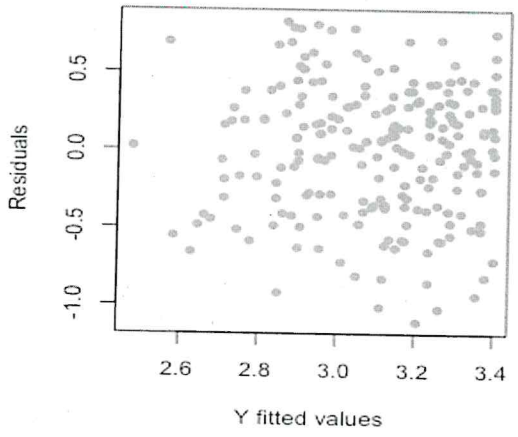
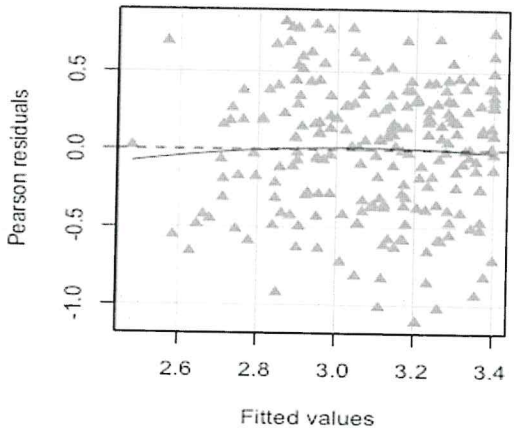


Fig4:Tukey's Curve



Tukey.test

Test	Pvalue
-0.4738946	63.55751e-02

Answer the following questions:

a. [5 points] Comment on the relationship between the First-year college GPA and High school GPA using their scatterplot.

There is a positive, linearish relationship between first year college GPA & High school GPA. The strength of the relation seems moderate. Two outliers are there (marked) as these students scored low in high school GPA but they did relatively as good as other students in the first year GPA. There is no obvious curve, thus linearish pattern.

b. [5 points] What is the correlation between the First-year college GPA and High school GPA (provide magnitude and sign).

$R = \sqrt{R^2}$ ✓

$= \sqrt{0.1997} = \pm 0.4469$

Since slope + relationship between the variables is positive

$\boxed{R = +0.4469}$

c. [5 Points] Report the fitted model equation for predicting First-year college GPA using High school GPA.

$\hat{Y} = 1.18 + 0.555X$, where

Y: first year college GPA

X: high school GPA

d. [5 points] Interpret the slope estimate and its standard error in above model in the context.

For a 1 point increase in HSGPA, we expected the first-year college GPA to increase by 0.555 points.

Sample to Sample variability in the above estimate is 0.555 points $\pm$ 0.075 points.

e. [5 points] Find the residual value for a randomly selected student whose High school GPA is 3.45. if their observed first-year score is 3 points.

$e = Y - \hat{Y}$

$\hat{Y} = 1.18 + 0.555(3.45)$

$= 1.18 + 1.91475$

$= 3.094$

$e = 3 - 3.094 = -0.094$

f. [5 points] Report the goodness of fit of this model. Interpret the measures in the context.

1.) $R^2 = 0.1997$

19.97% ~ 20% change in the first-year college GPA can be explained by the high school GPA.

2.) $\hat{\sigma} = 0.4174$

On an average, the variation between the observed & estimated first year college GPA is $\pm$ 0.4174 points.

3.) Using residual plot in Fig. 3, we see small to low relationship, if any between the residual & estimated Y values.

g. [5 points] Test the hypothesis that High school GPA is a useful predictor for the First-year GPA. Show all steps.

$H_0 : \beta_1 = 0$

$H_a : \beta_1 \neq 0$

$\alpha = 0.05$

$t = 7.359$

$P\text{-value} = 3.78 \times 10^{-12} = 0.000000000000378 \approx 0$

Since P-value $< \alpha$, we can reject H_0.

Conclusion: At 5% significance, we can conclude that the high school GPA is a useful (statistically significant) predictor for the first-year college GPA.

h. [5 points] Report and interpret the 95% confidence interval for slope in above model.

We are 95% confident that for a 1 unit increase in high school GPA, the expected increase in the first year college GPA is between 0.406 to 0.704 points.

i. [5 points] If you were to construct a 99% confidence interval for slope, would this be wider or narrower than the 95% confidence interval. Why?

Since we are increasing the confidence, the interval estimate will be wider.

The reason is that the t critical value is higher when confidence is higher.

j. [5 points] The results from Tukey's test of a model are given in Fig 4. Show all steps of hypothesis testing and share the conclusion.

H_0: Linearity b/w first year college GPA & high school GPA holds.

H_a: linearity fails

$\alpha = 0.05$

test stat $= -0.474$

P-value $= 0.635$

P-value $> 0.05 \Rightarrow$ Do not reject H_0

Conclusion: At 5% significance, we have enough evidence that there is a linear relation between the first-year college GPA & the high school GPA.

Problem 2 [15 points]: To test the LINE assumptions for a given model, six plots are given. Using these plots what can you tell about the assumptions for the model I fitted. Explain all steps.

Fig.1(a)

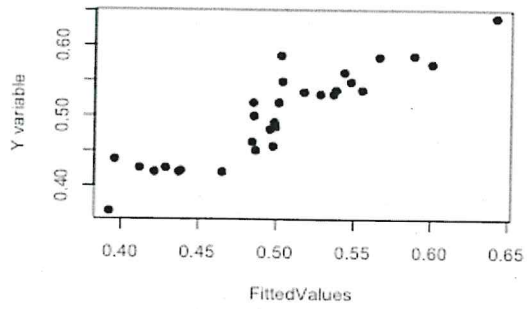


Fig1(b)

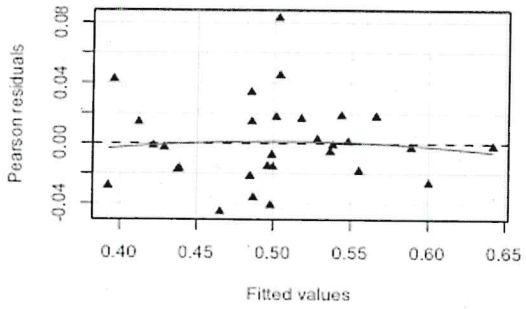


Fig.1(c)

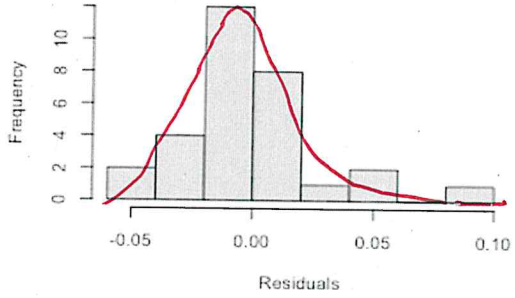


Fig.1(d)

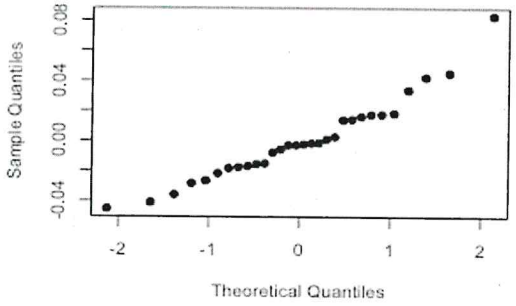


Fig.1(e)

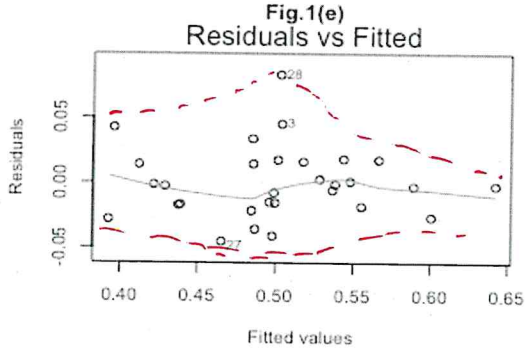
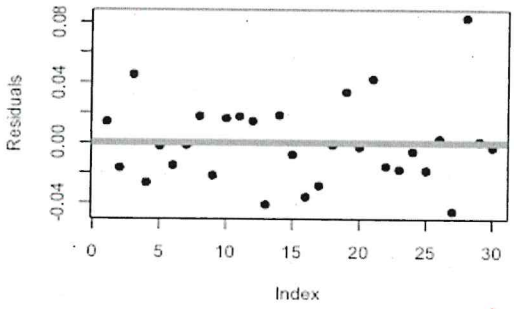


Fig.1(f)



Linearity: Using Fig1(a), we see a linear (upward) relation, no obvious outliers but there is a slight curve in the middle. Fig1(b) shows that the possible curve is nearly flat. Using the two plots linearity seems to hold.

Independence of Errors: Using Plot 1(f), the index plot of residuals, there is no obvious

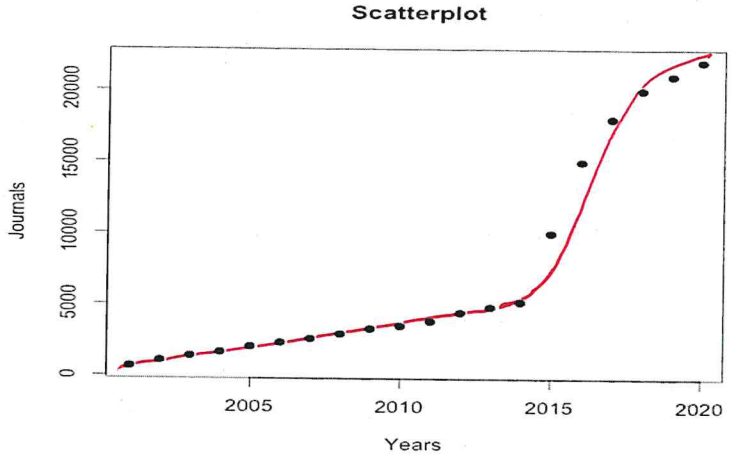
Normality of Errors: Fig 1(c) shows histogram of errors which has a bell-shaped curve distribution. There is an outlier but overall it still looks symmetric & bell shaped.

Fig. 1(d), the qq plot shows an agreement between the sample & normal quantiles with some deviation in the middle. Overall, the normality seems to hold.

Errors are homoscedastic: The residual plot in Fig. 1(e) shows a slight change in range/variability. It is not so obvious if homoscedastic is met using this plot. BP test would be helpful.

* be appropriate as the trend is non-linear.

Problem 3 [7.5 points]: Data from academic journals published on the internet over a 20-year period showed following relationship



(a) Comment on the trend in journals published online. Do you think a linear model would work well in this case, why or why not?

There is an exponential like curve with # of academic journals published on the internet increasing linearly for first 15 years or so after which it increases drastically. A linear model might not *

(b) Regressing log(journals) on years, following regression summary was obtained:
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	87.690e-02	1.4270e-01	6.145.	0.001659 ***
time	3.4555e-01	3.091e-01	10.829	0.000017 ***

Using this output, report the fitted model in terms of journals.

$\log(\hat{Y}) = 0.877 + 0.345X$, where Y: # of journals published online

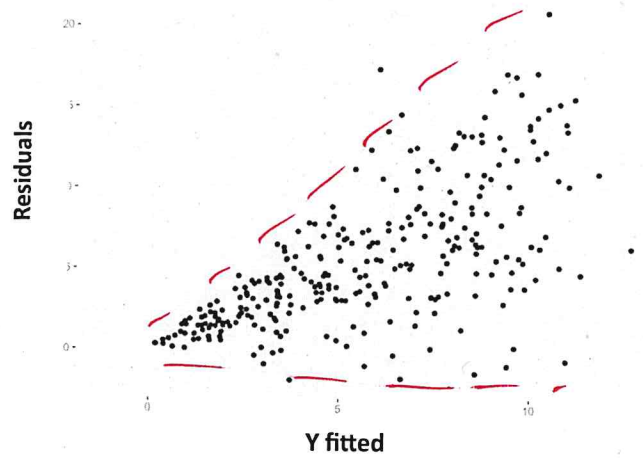
X: time in years

$\hat{Y} = \exp(0.877 + 0.345X)$

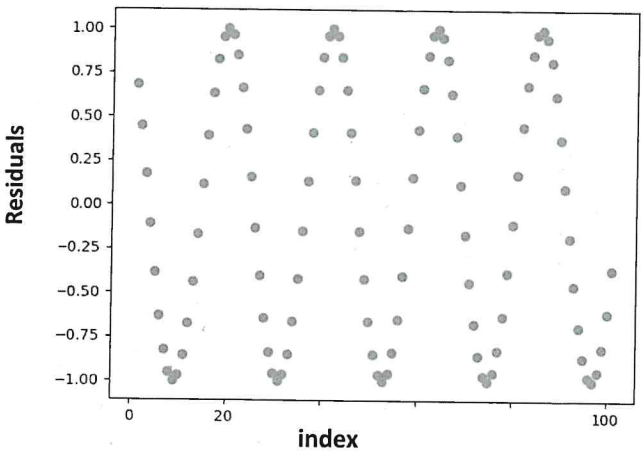
$\hat{Y} = \exp(0.877)\exp(0.345X)$.

Problem 5 [7.5 points]: For each plot given below, state which regression assumptions can you test and what conclusions can you make about those assumptions, explain.

Plot (a)



Plot (b)



Plot (a) shows an upward ~~trend~~ trend, hence using this pattern in a residual plot, we can tell that the linearity assumption would fail.
Plot (a) also shows fanning out, hence homoscedasticity would also fail.

Plot (b) shows the index plot of residuals. There is a clear cyclic pattern, hence independence of errors assumption ~~will~~ will fail.

Problem 6 [15 points]: In each of the following set of variables,

- identify which of the variables can be regarded as a response variable and which can be used as predictors
- classify each variable as numerical (discrete or continuous) or categorical (ordinal or nominal)
- state which type of regression can be used in the analysis.

a.) [5 points] Whether or not an applicant is accepted admission using their grade point average, SAT score, and gender.

Y: whether an applicant is accepted for admission or not (categorical nominal)

X_1: GPA numeric, continuous

X_2: SAT score numeric, continuous

X_3: Gender

Model: logistic Regression

b.) [5 points] The predict the time it takes to finish a race using the distance of a race and the weather conditions.

Y: time to finish a race (numeric, continuous)

X_1: distance of race (numeric, continuous)

X_2: weather conditions (categorical, nominal)

Model: Multiple Linear Regression (MLR)

c.) [5 points] To predict whether or not the person has lung cancer, the weight of a person, and whether or not the person is a smoker

Y: whether or not person has lung cancer (categorical, nominal)

X_1: weight (numeric, continuous)

X_2: Smoker or not (categorical, nominal)

Model: logistic Regression

Wrong as correlation doesn't imply causation.

Problem 7 [5 points]: Some studies suggest that students who eat breakfast regularly perform better academically and conclude that eating breakfast causes higher grades! What is wrong with this conclusion and is there any justification for this reasoning?

It might be because there is a confounding variable like the overall nutritional status of the students, as those who eat breakfast regularly may also follow healthier eating habits that boost their academic performance.

BONUS Problem (5 Points): For the following scatter plot and the fitted regression line answer the following questions:

a.) What are the possible **signs** of intercept and slope?

both positive

b.) Show in the graph the residuals for **any two points**.