STAT GU4205/GR5205 Assignment 4 [40 pts]
Due 11:59pm Friday, November 22nd

Please scan the written part (if you do not want to use Word or LaTeX), and submit your assignment on **Gradescope**. For coding problems, your `R` code is also required.

Note: for coding problems, even the code is submitted through Canvas, please make sure in your Gradescope submission, you include all output, plots, and necessary explanations for the output.

Problem 1, Properties of multivariate normal distribution, 4 points

Consider the model

$$Y_i = \mu + \epsilon_i \qquad i = 1, 2, \ldots, n \qquad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

The sample mean and sample variance are defined respectively as

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

and

$$S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2.$$

Use Theorem 3.6 on Page 93 to prove that the sample mean $\bar{Y}$ and sample variance S_Y^2 are independent random variables.

Problem 2, Inference for slope parameters, 10 points

A commercial real estate company evaluates vacancy rates, square footage, rental rates, and operating expenses for commercial properties in a large metropolitan area in order to provide clients with quantitative information upon which to make rental decisions. The data below are taken from 81 suburban commercial properties that are the newest, best located, most attractive, and expensive for five specific geographic areas. The data consists of variables age (X_1), operating expenses and taxes (X_2), vacancy rates (X_3), total square footage (X_4), and rental rates (Y). For this data set, we skip residual diagnostics but in practice, that should be included in the analysis. The data set `HW4Problem2.txt` is posted on Canvas. Use `R` to perform the following tasks:

i. Regress the rental rates (Y) against all of the covariates; age (X_1), operating expenses and taxes (X_2), vacancy rates (X_3), total square footage (X_4). Write down the estimated linear model.

ii. What percentage of variation in rental rates is explained by this model?

iii. Are there any marginal relationships between the response variable and covariates? (Run t-tests on all slope parameters.)

iv. Run a F-test to see if there is an *overall relationship* between the rental rates and all of the covariates.

v. Run a F-test to simultaneously test the slopes for age (X_1) and vacancy rates (X_3).

vi. Run a F-test to see if vacancy rates (X_3) is a significant predictor after holding all other variables constant. To perform this test, use the full and reduced models. How does this test relate to the summary output from part (ii)?

vii. The researcher wishes to obtain 95% interval estimates of the mean rental rates for four typical properties specified as follows. Find the four confidence intervals using the Bonferroni procedure.

	1	2	3	4
x_1	5.0	6.0	14.0	12.0
x_2	8.25	8.50	11.50	10.25
x_3	0	0.23	0.11	0
x_4	250,000	270,000	300,000	310,000

Problem 3, "General" linear models, 8 points

i. Recall the multiple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i. \tag{1}$$

For each of the following regression models, indicate whether it can be expressed in the form of (1) by a suitable transformation. To receive full credit, describe the transformation if it exists.

a. $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 \log(x_{i2}) + \beta_3 x_{i1}^2 + \epsilon_i$

b. $Y_i = \epsilon_i \exp\{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2\}$

c. $Y_i = \log(\beta_1 x_{i1}) + \beta_2 x_{i2} + \epsilon_i$

d. $Y_i = \beta_0 \exp\{\beta_1 x_{i1}\} + \epsilon_i$

e. $Y_i = [1 + \exp\{\beta_0 + \beta_1 x_{i1} + \epsilon_i\}]^{-1}$

ii. Consider the toy data set:

The data set is provided in the file `HW4Problem3.txt` on canvas. Use multiple linear regression techniques to fit a polynomial to the above data set. To receive full credit,

y	2.44	8.36	98.33	115.06	128.91	123.46	148.30	138.10	153.10	119.08
	87.66	134.88	91.71	126.81	40.41	54.94	33.03	35.74	14.99	-1.18
	2.44	8.36	28.33	45.06	48.91	43.46	118.30	108.10	233.10	199.08
	337.66	384.88								
x	0.00	0.00	1.00	1.00	2.00	2.00	3.00	3.00	4.00	4.00
	5.00	5.00	6.00	6.00	7.00	7.00	8.00	8.00	9.00	9.00
	10.00	10.00	11.00	11.00	12.00	12.00	13.00	13.00	14.00	14.00
	15.00	15.00								

write down the estimated model and create a scatter plot with the estimated curve overlaid on the plot.

Problem 4, Model diagnostics, 8 points The Tri-City Office Equipment Corporation sells an imported copier on a franchise basis and performs preventive maintenance and repair service on this copier. The data have been collected from 45 recent calls on users to perform routine preventive maintenance service; for each call, x is the number of copiers serviced and Y is the total number of minutes spent by the service person. The data set `HW4Problem4.txt` is posted on canvas.

Use `R` to perform the following tasks:

i. Obtain the estimated regression function.

ii. Create a scatterplot of the data set with the line of best fit overlaid on the graph. Create a QQ plot of the studentized deleted residuals, histogram of the studentized deleted residuals, line plot of the studentized deleted residuals, plot the studentized deleted residuals verses predicted values $\hat{y}$, and studentized residuals verses predictor variable x. Based on the plots, discuss whether any of the regression assumptions have been violated. In your descriptions, relate your explanations to the relevant plots.

iii. Perform a Box-Cox procedure on the data set. What is the estimated value of λ? Is it necessary to perform this transformation on the response variable? Briefly explain your reasoning.

Problem 5, Heteroskedasticity, 10 points

Observations on Y are to be taken when $x = 10, 20, 30, 40$, and 50, respectively. The true regression function is $E[Y] = 20 + 10x$. The error terms are independent and normally distributed with $E[\epsilon_i] = 0$ and $Var[\epsilon_i] = .8x$.

i. Generate a random Y observation for each x level and calculate both the ordinary and weighted least squares estimates of the regression coefficient β_1 in the simple linear

regression function.

ii. Repeat part (a) 10,000 times, generating new random numbers each time.

iii. Calculate the mean and variance of the 10,000 ordinary least squares estimates of β_1 and do the same for the 10,000 weighted least squares estimates.

iv. Do both the ordinary least squares and weighted least squares estimators appear to be unbiased? Explain. Which estimator appears to be more precise here? Comment.