MATH 564, 2020 Fall
Final Examination
(Total: 100 Points)

Instruction:

1. Problem 1-6 do not require any computing or coding to answer.
2. Problem 7-8 require coding to answer.
3. Please follow the format of the mid-exam solutions (my version) to prepare your solution. I prefer you to type out the solutions for Problem 1-8. You can use word, latex, or R-markdown. If you use R markdown, please knit the output to pdf file.

Problem 1 (10 pts=2+2+2+4) Accurately measuring a person's body fat percentage is difficult. An indirect method is to estimate the body fat percentage based on various body circumference measurements. Data were collected on the body fat percentage and several body measurements. The regression output of the data is given below (some entries in the output are deleted). Answer the following questions.

Summary of Fit

R^2	0.735011
$RMSE$	4.342724
$\bar{y}$	19.15079

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F-Ratio
Model	4	12920.754	3230.19	171.2787
Error	247	4658.236	18.86	Prob > F
C. Total		17578.990		$< .0001$

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> $\|t\|$
Intercept	-34.85407	7.245005	-4.81	$< .0001$
Weight	-0.135631	0.024748	-5.48	$< .0001$
Abs	0.9957513	0.056066	17.76	$< .0001$
Forearm	0.4729284		2.60	0.0098
Wrist	-1.505562	0.442666	-3.40	0.0008

(a) Is it possible for you to find out the number of observations in the data set? If so, do it.

(b) Compute the Adjusted R^2.

(c) Which body circumference measurement seems to be the most important in determining the body fat percentage?

(d) Compute a 95% confidence interval for the coefficient of Forearm ($t_{0.025,247} = 1.97$).

Problem 2 (10 pts=2+2+4+2 Data for 51 U.S. "states" (50 states, plus the District of Columbia) was used to examine the relationship between violent crime rate (violent crimes per 100,000 persons per year) and the predictor variables of urbanization (percentage of the population living in urban areas) and poverty rate. A predictor variable indicating whether or not a state is classified as a Southern state (1 = Southern, 0 = not) was also included. Some output for the analysis of this data is shown below (with some information intentionally left blank).

```
The regression equation is
Crime = -321.9 +4.69Urban +39.3Poverty -649.3South +12.1Urban*South -5.84Poverty*South

Predictor        Coef   SE Coef       T      P
Intercept     -321.90    148.20   -2.17  0.035
Urban           4.689     1.654    2.83  0.007
Poverty         39.34     13.52    2.91  0.006
South(S=1)    -649.30    266.96   -2.43  0.019
Urban*South     12.05     2.871    4.20  0.000
Poverty*South  -5.838    16.671   0.35  0.728

Analysis of Variance
Source           DF       SS      MS       F      P
Regression        5  2060459  412091   -----  0.000
Residual Error   --   882169  ------
Total            50  2942628
```

(a) For the Southern states, what is the fitted regression model for Crime with respect to urbanization and poverty.

(b) Predict the violent crime rate for a Southern state with an urbanization of 55.4 and a poverty rate of 13.7.

(c) Calculate the ANOVA F test statistic value, the DF corresponds to Residual Error, and the MSE. What should be the degree of freedom for the F statistic?

(d) Which predictors should probably be removed from the model to improve it? Why?

Problem 3 (10 pts: 3, 2, 5) faults were studied by changing the temperature from 40 °C to 80 °C in the spinning process. The following logistic regression was fitted to the data.
Parameter Estimates

Term	**Estimate**	**Std Error**	$p-$**value**
Intercept	-13.43198	0.4313758	0.0000
x	0.2067636	0.0068186	
$(x-60)^2$	0.0101906	0.0005469	$< .0001$

(a) Perform the two-sided hypothesis test on the coefficient corresponding to x, given $\alpha = 0.05$. (The normal quantile is $z_{0.975} = 1.96$.)

(a) Predict the percentage of faults when temperature is at 60 °C.

(b) Find the optimum value of of temperature to minimize the faults. (Hint: $e^z/(1+e^z)$ is a monotonic function in z).

Problem 4 (10 pts=3+2+2+3) Flu Shots. A local health clinic sent fliers to its clients to encourage everyone, but especially older persons at high risk of complications, to get a flu shot in time for protection against an expected flu epidemic. In a pilot follow-up study, 159 clients were randomly selected and asked whether they actually received a flu shot. A client who received a flu shot was coded $Y = 1$, and a client who did not receive a flu shot was coded $Y = 0$. In addition, data were collected on their age (X_1) and their health awareness. The latter data were combined into a health awareness index (X_2), for which higher values indicate great awareness. Also included in the data were client gender, when males were coded $X_3 = 1$ and females were coded $X_3 = 0$.

Multiple logistic regression model on $E(Y(\boldsymbol{x})) = \pi(\boldsymbol{x})$,

$$\log \frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3,$$

is fitted. Following are the output of the `R` function `glm`.

```
Call:
glm(formula = Y ~ ., family = binomial("logit"), data = flu)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4037  -0.5637  -0.3352  -0.1542   2.9394

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.17716    2.98242  -0.395  0.69307
X1           0.07279    0.03038   2.396  0.01658 *
X2          -0.09899    0.03348  -2.957  0.00311 **
X3           0.43397    0.52179   0.832  0.40558
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 134.94  on 158  degrees of freedom
Residual deviance: 105.09  on 155  degrees of freedom
AIC: 113.09
```

(a) Obtain $\exp(\hat{\beta}_2)$ (using the first `R` output), and interpret this number.

(b) What is the estimated probability that male clients aged 55 with a health awareness index of 60 will receive a flu shot? (Use the first `R` output.)

(c) Based on the residual deviance, perform the Goodness-of-Fit test for $\alpha = 0.05$. Write down the H_0 and H_a for the Goodness-of-Fit test and make conclusion based on your test. (The chi-square quantile is $\chi^2(0.95, 155) = 185.0523$. Use the first `R` output.)

(d) Comment on the statistical significance of gender to the probability of getting a flu shot. The following output from `R` returned by `glm` function is for the logistic regression model

$$\log \frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})} = \beta_0 + \beta_1 X_1 + \beta_2 X_2.$$

```
Call:
glm(formula = Y ~ X1 + X2, family = binomial("logit"), data = flu)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4479  -0.5708  -0.3390  -0.1629   2.8430

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.45778    2.91534  -0.500  0.61705
X1           0.07787    0.02970   2.622  0.00873 **
X2          -0.09547    0.03241  -2.946  0.00322 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 134.94  on 158  degrees of freedom
Residual deviance: 105.80  on 156  degrees of freedom
AIC: 111.8
```

Use the likelihood ratio test (The chi-square quantile is $\chi^2(0.95, 1) = 3.84$.) to see if X_3 should be dropped from the first logistic regression model.

Problem 5 (10 pts=4+3+3) Researchers studied 41 male African elephants over a period of 8 years. The age (X) of the elephant at the beginning of the study and the number of successful mating (Y) during the 8 years were recorded. We assume the number of matings follows a Poisson distribution, where the mean depends on the age of the elephant in question. The probability distribution for the number of successful matings for the i-th elephant y_i is

$$\Pr(Y = y_i) = \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}.$$

(a) Show that the Poisson distribution is a member of the exponential family.

(b) The following output is returned by the `glm` function in `R`.

```
Call:
glm(formula = mating ~ age, family = poisson)
Coefficients:
            Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  -1.58201  0.54462    -2.905   0.00368 **
age           0.06869   0.01375     4.997   5.81e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 75.372  on 40  degrees of freedom
Residual deviance: 51.012  on 39  degrees of freedom
AIC: 156.46
```

Return the 95% confidence interval on the estimation of the coefficient for Age X. ($z_{0.975} = 1.96$)

(c) How much does the mean number of matings increases if age is increased by one year? Obtain a 95% confidence interval for the estimated increase.

Problem 6 (10 pts) Explain how you will use linear regression methods to approximately estimate the parameter in the following nonlinear model:

$$y = 1 - e^{-(\beta_0 + \beta_1 x)}.$$

Given the data (x_i, y_i) for $i = 1, \ldots, n$, write down the formula for estimating β_0 and β_1. (Hint: you need to use the parameter estimation formula for simple linear regression.)

Problem 7. Commercial Properties. 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, more attractive, and expensive for five specific geographic areas. The data contain the columns the age (X_1), operating expense and taxes (X_2), vacancy rates (X_3), total square footage (X_4) and rental rates (Y).

1. (2 pts) Obtain the scatter plot matrix of all the variables Y and $X_1 \sim X_4$. State and interpret your findings.

2. (1 pt) Fit regression model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$ (call this Model 1). State the estimated regression function.

3. (4 pts) Plot the residuals against $\hat{Y}$ (one plot in one picture), against 4 predictor variables (4 plots in one picture), against each two-factor interaction (6 plots in in one picture). Are the residuals look like i.i.d. normally distributed, i.e., the pattern is unsystematic random around zero?

4. (2 pts) Show the ANOVA table of the regression model. Is the F-ratio significant? ($\alpha = 0.05$).

5. (2 pts) What are the R^2 and R^2_{adj}?

6. (3 pts) Provide the point estimate of the mean values of the response at the following three new settings of $X_1 \sim X_4$. Provide the 95% confidence intervals and prediction intervals.

	1	**2**	**3**
X_1:	4.0	6.0	12.0
X_2:	10.0	11.5	12.5
X_3:	0.10	0	0.32
X_4:	80,000	120,000	340,000

7. (3 pts) Fit the regression model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_4 + \epsilon$ (call this Model 2). State the estimated regression function. Use partial F test to compare Model 1 and Model 2. What is your conclusion.

8. (2 pts) Plot Y against X_1, do you observe any curvature?

9. (4 pts) Fit the regression model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_4 + \beta_4 X_1^2 + \epsilon$ (call this Model 3). State the estimated regression function. Plot Y against the fitted $\hat{Y}$. Does Model 3 seem to be a good fit?

10. (2 pts) Use partial F test to compare Model 2 and Model 3. Can you conclude X_1^2 is a significant term?

Problem 8. Biopsy. A breast cancer database was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. He assessed biopsies of breast tumors for 699 patients up to 15 July 1992; each of nine attributes has been scored on a scale of 1 to 10, and the outcome is also known: benign ($Y = 0$) or malignant ($Y = 1$). The data is available in the `MASS` package of `R`. You can obtain it by call `library(MASS)` and then `data(biopsy)`. This data frame contains the following columns:

`V1` Clump thickness

`V2` Uniformity of cell size

`V3` Uniformity of cell shape

`V4` Marginal adhesion

`V5` Single epithelial cell size

`V6` Bare nuclei (16 values are missing

`V7` Bland chromatin

`V8` Normal nucleoli

`V9` Mitoses

`class` "benign" or "malignant"

1. (5 pts) Build a generalized linear model on the outcome (called `class` in the data) with respect to the variables `V1`, ..., `V9`. Write down the fitted model. Note that the data contain some `NA` values. Remove the rows containing `NA` to do all the questions.

2. (4 pts) From variables `V1`,..., `V9`, perform the forward stepwise regression with AIC criterion. What is the reduced model?

3. (3 pts) Use likelihood ratio test to compare the model in Question 1 and 2. Is there significant difference between the two?

4. (3 pts) Using the reduced model from Question 2, plot the deviance residual against the fitted value and perform the residual diagnostics.