STAT GR5205 Final Exam

Name:______________________________

UNI:______________________________

Please write your name and UNI

The Fall GR5205 final is closed notes and closed book. Calculators are allowed. Tablets, phones, computers and other equivalent forms of technology are strictly prohibited. Students are not allowed to communicate with anyone with the exception of the TA and the professor. If students violate these guidelines, they will receive a zero on this exam and potentially face more severe consequences. Students must include all relevant work in the handwritten problems to receive full credit.

Theory Component

Problem 1 [10 pts]

Part I (5 pts)

Let $\mathbf{X}$ be a full rank $n \times p$ design matrix and define the hat matrix as $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. Recall that the column space of $\mathbf{X}$, denoted $\mathcal{C}(\mathbf{X})$, is the set of all linear combinations of the columns in $\mathbf{X}$. Prove that if $\mathbf{v} \in \mathcal{C}(\mathbf{X})$, then $\mathbf{H}\mathbf{v} = \mathbf{v}$.

Part II (5 pts)

Let $\hat{\boldsymbol{\beta}}$ be the least squares estimator of $\boldsymbol{\beta}$. Use the result from Problem 1.I to prove that the sum of sample residuals is always zero, i.e., show $\sum_{i=1}^{n} e_i = 0$.

Problem 2 [25 pts]

Consider three models:

(1) $$Y_i = \beta_1 x_{i1} + \epsilon_i, \quad i = 1, \ldots, n, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2),$$

(2) $$Y_i = \beta_2 x_{i2} + \epsilon_i, \quad i = 1, \ldots, n, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2),$$

(3) $$Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, \quad i = 1, \ldots, n, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

Denote the respective data vectors and full design matrix by

$$\begin{aligned}
\mathbf{Y} &= \begin{pmatrix} Y_1 & Y_2 & \cdots & Y_n \end{pmatrix}^T, \\
\mathbf{x}_1 &= \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{n1} \end{pmatrix}^T, \\
\mathbf{x}_2 &= \begin{pmatrix} x_{12} & x_{22} & \cdots & x_{n2} \end{pmatrix}^T, \\
\mathbf{X} &= \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 \end{pmatrix}.
\end{aligned}$$

Further, let $\mathbf{H}_1$, $\mathbf{H}_2$, and $\mathbf{H}$ be the respective hat-matrices of models (1), (2) and (3).

Part I (10 pts)

Assuming that the vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ are perfectly uncorrelated (orthogonal), prove that

$$\mathbf{HX} = (\mathbf{H}_1 + \mathbf{H}_2)\mathbf{X}$$

Part II (10 pts)

Similarly, assuming that the vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ are perfectly uncorrelated (orthogonal), prove that

$$\mathbf{H}(\mathbf{x}_1 + \mathbf{x}_2) = (\mathbf{H}_1 + \mathbf{H}_2)(\mathbf{x}_1 + \mathbf{x}_2)$$

Part III (5 pts)

Note: the following exercise is not a proof. Assuming that the vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ are perfectly uncorrelated, state the (obvious) relationship between $\mathbf{H}_1$, $\mathbf{H}_2$, and $\mathbf{H}$.

Problem 3 [20 pts]

Consider the *heteroscedastic* regression through the origin model

$$Y_i = \beta x_i + \epsilon_i, \quad \imath = 1, \ldots, n, \quad \epsilon_i \overset{ind}{\sim} N(0, \sigma_i^2),$$

where $w_i = \frac{1}{\sigma_i^2}$ and σ_i^2 is known for $i = 1, \ldots, n$.

Part I (10 pts)

Derive the weighted least squares estimator $\hat{\beta}_w$ for the slope parameter β. You can use a matrix approach or a scalar approach for this problem. You cannot begin with the weighted least squares solution. You must solve the optimization problem from scratch.

Part II (5 pts)

Show that the weighted least squares estimator $\hat{\beta}_w$ is an unbiased estimator of β.

Part III (5 pts)

Derive an expression for the variance of $\hat{\beta}_w$. Simplify the expression completely.

Methods Component

Problem 4 [15 pts]

Patients who suffer from moderate to severe migraine headache took part in a double-blind clinical trial to assess an experimental surgery. A group of 79 patients were randomly assigned to receive either the real surgery in migraine trigger sites ($m = 53$) or a sham surgery ($n = 26$) in which an incision was made but no further procedure was performed. The surgeons hoped that patients would experience "a substantial reduction in migraine headaches," which we will label as "success." A substantial reduction means at least 50% reduction in migraine headache frequency, intensity, or duration when compared with baseline (presurgery) values.

Surgery	No success	Success
Real	12	41
Sham	11	15

Consider the following logistic regression output related to the above dataset.

```
> Success <- c(rep(1,41+15),rep(0,12+11))
> Real <- c(rep(1,41),rep(0,15),rep(1,12),rep(0,11))
> summary(glm(Success~Real,family=binomial(link="logit")))

Call:
glm(formula = Success ~ Real, family = binomial(link = "logit"))

Coefficients:
            Estimate Std.  Error  z-value   Pr(>|z|)
(Intercept)   0.3102      0.3970   0.781      0.4346
Real          Missing     0.5151   Missing    0.0745 .
```

Part I (10 pts)

Compute the two values missing from the above R output.

Part II (5 pts)

Run the appropriate test to see if successful reduction of migraine headache was more common among patients who received the real surgery than among those who received the sham surgery.

Problem 5 [10 pts]

A university medical center urology group was interested in the association between prostate-specific antigen (PSA) and a number of prognostic clinical measurements in men with advanced prostate cancer. Data were collected on 97 men who were about to undergo radical prostatectomies. The 8 variables are:

Variable	Variable Name	Description
X_1	PSA level	Serum prostate-specific antigen level (mg/ml)
X_2	Cancer volume	Estimate of prostate cancer volume (cc)
X_3	Weight	Prostate weight (gm)
X_4	Age	Age of patient (years)
X_5	Benign prostatic hyperplasia	Amount of benign prostatic hyperplasia (cm2)
X_6	Seminal vesicle invasion	Presence or absence of seminal vesicle invasion
X_7	Capsular penetration	Degree of capsular penetration (cm)
Y	Gleason score	Pathologically determined grade of disease

In our setting we create a new binary response variable Y, called high-grade cancer by letting $Y = 1$ if Gleason score equals 8, and $Y = 0$ otherwise (i.e., if Gleason score equals 6 or 7). The goal of this exercise is to carry out a logistic regression analysis and decide what the "best" cut-off value is for classifying whether or not a respondent belongs to the high-grade cancer group.

Note: If you are taking STAT 5206 with me, this is the same dataset from Lab 6.

Consider three different cutoff values: $\hat{p} = .23$, $\hat{p} = .50$, $\hat{p} = .73$. Based on the three two-by-two contingency tables displayed below, which of the three cutoff values gives the most reliable predictions for classifying whether or not a respondent belongs to the high-grade cancer group? To receive full credit, please show the relevant calculations.

$\hat{p} = .23$	$\hat{Y} = 1$	$\hat{Y} = 0$	$\hat{p} = .50$	$\hat{Y} = 1$	$\hat{Y} = 0$	$\hat{p} = .73$	$\hat{Y} = 1$	$\hat{Y} = 0$
$Y = 1$	17	4	$Y = 1$	13	8	$Y = 1$	7	14
$Y = 0$	9	67	$Y = 0$	4	72	$Y = 0$	1	75

Problem 5 [20 pts]

Fifteen stores were selected for the study, and a completely randomized experimental design was utilized. Each store was randomly assigned one of the promotion types, with five stores assigned to each type of promotion. Other relevant conditions under the control of the company, such as price and advertising, were kept the same for all stores in the study. Data on the number of cases of the product sold during the promotional period, denoted by Y, are presented in Table 1, as are also data on the sale of the product in the preceding period, denoted by X. Indicator variables are defined below:

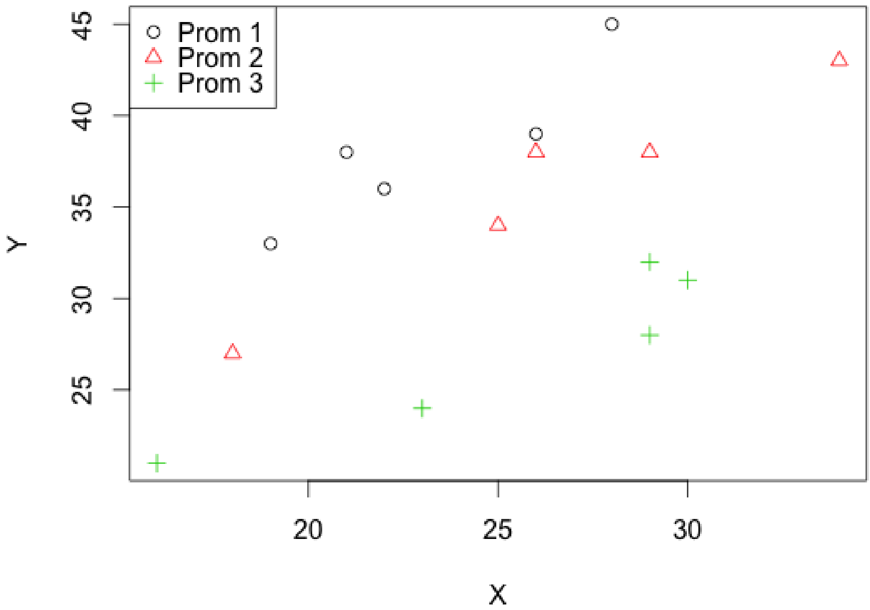
Table 1

	Store 1		Store 2		Store 3		Store 4		Store 5	
	Y	X	Y	X	Y	X	Y	X	Y	X
Promotion 1	38	21	39	26	36	22	45	28	33	19
Promotion 2	43	34	38	26	38	29	27	18	34	25
Promotion 3	24	23	32	29	31	30	21	16	28	29

$$P_1 = \begin{cases} 1 & \text{promotion 2} \\ 0 & \text{otherwise} \end{cases} \qquad P_2 = \begin{cases} 1 & \text{promotion 3} \\ 0 & \text{otherwise} \end{cases}$$

$$I_1 = \begin{cases} 1 & \text{store 2} \\ 0 & \text{otherwise} \end{cases} \qquad I_2 = \begin{cases} 1 & \text{store 3} \\ 0 & \text{otherwise} \end{cases} \qquad I_3 = \begin{cases} 1 & \text{store 4} \\ 0 & \text{otherwise} \end{cases} \qquad I_4 = \begin{cases} 1 & \text{store 5} \\ 0 & \text{otherwise} \end{cases}$$

Figure 1



Part I (5 pts)

Do you believe that an interaction should be included between the sale of the product in the preceding period with promotion? Justify your answer based on Figure 1.

Part II (5 pts)

Write down the theoretical linear model relating the number of cases of the product sold during the promotional period versus the promotion type, store index, and the sale of the product in the preceding period.

Part IV (5 pts)

Run a testing procedure to see if average sales differ per promotion group after controlling for the variation in other covariates in the model. To receive full credit, show all relevant steps of the testing procedure.

Part V (5 pts)

Run a testing procedure to see if promotion type 2 has the same impact as promotion type 3 on sales. Run this test after controlling for the variation of all other covariates in the model. To receive full credit, show all relevant steps of the testing procedure including computing the correct test statistic and degrees of freedom. Note the correct P-value for this test is `1-pf(f.calc,?,?)=0.0003`.

R code

```
#-------------------------- Model 1


model.1 <- lm(Y~Promotion)
summary(model.1)
anova(model.1)


#-------------------------- Model 2


model.2 <- lm(Y~Store+X+Promotion)
summary(model.2)
anova(model.2)


#-------------------------- Model 3


Int1 <- P1*X
Int2 <- P2*X
model.3 <- lm(Y~Store+X+Promotion+Int1+Int2)
summary(model.3)
anova(model.3)


#-------------------------- Model 4


Prom.combine <- I(Promotion=="Prom 2")+I(Promotion=="Prom 3")
model.4 <- lm(Y~Store+X+Prom.combine)
summary(model.4)
anova(model.4)
```

R output

```
#-------------------------- Model 1

> model.1 <- lm(Y~Promotion)
> summary(model.1)

Call:
lm(formula = Y ~ Promotion)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       38.200      2.264  16.871 1.01e-09 ***
PromotionProm 2   -2.200      3.202  -0.687  0.50511
PromotionProm 3  -11.000      3.202  -3.435  0.00494 **

Residual standard error: 5.063 on 12 degrees of freedom
Multiple R-squared:  0.5241,Adjusted R-squared:  0.4448
F-statistic: 6.609 on 2 and 12 DF,  p-value: 0.01161

> anova(model.1)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value  Pr(>F)
Promotion  2  338.8 169.400  6.6086 0.01161 *
Residuals 12  307.6  25.633
```

```
#------------------------- Model 2

> model.2 <- lm(Y~Store+X+Promotion)
> summary(model.2)

Call:
lm(formula = Y ~ Store + X + Promotion)


Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      16.7385     3.1495   5.315 0.001106 **
StoreStore 2      0.3969     1.5351   0.259 0.803419
StoreStore 3     -0.9364     1.5351  -0.610 0.561130
StoreStore 4      0.9943     1.6567   0.600 0.567310
StoreStore 5     -1.7726     1.5433  -1.149 0.288448
X                 0.9364     0.1189   7.876 0.000101 ***
PromotionProm 2  -5.1966     1.2451  -4.174 0.004170 **
PromotionProm 3 -13.0601     1.2140 -10.758 1.32e-05 ***


Residual standard error: 1.874 on 7 degrees of freedom
Multiple R-squared:  0.9619,Adjusted R-squared:  0.9239
F-statistic: 25.28 on 7 and 7 DF,  p-value: 0.0001857

> anova(model.2)

Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value   Pr(>F)
Store      4  65.07  16.267  4.6296 0.038270 *
X          1 138.58 138.578 39.4399 0.000412 ***
Promotion  2 418.16 209.080 59.5051 4.04e-05 ***
Residuals  7  24.60   3.514
```

```
#------------------------- Model 3

> model.3 <- lm(Y~Store+X+Promotion+Int1+Int2)
> summary(model.3)

Call:
lm(formula = Y ~ Store + X + Promotion + Int1 + Int2)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      22.4116     8.3187   2.694  0.04309 *
StoreStore 2      1.4809     1.8526   0.799  0.46034
StoreStore 3     -0.4533     1.6874  -0.269  0.79892
StoreStore 4      2.8667     2.6197   1.094  0.32372
StoreStore 5     -1.2619     1.7408  -0.725  0.50103
X                 0.9337     0.2264   4.125  0.00913 **
PromotionProm 2 -17.7415    12.5649  -1.412  0.21705
PromotionProm 3 -19.4550    12.1809  -1.597  0.17112
Int1             -0.2759     0.5063  -0.545  0.60922
Int2              0.2331     0.2589   0.900  0.40922


Residual standard error: 1.961 on 5 degrees of freedom
Multiple R-squared:  0.9703,Adjusted R-squared:  0.9167
F-statistic: 18.13 on 9 and 5 DF,  p-value: 0.002635

> anova(model.3)

Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value   Pr(>F)
Store      4  65.07  16.267  4.2310 0.072772 .
X          1 138.58 138.578 36.0447 0.001841 **
Promotion  2 418.16 209.080 54.3825 0.000405 ***
Int1       1   2.26   2.256  0.5868 0.478229
Int2       1   3.12   3.116  0.8106 0.409220
Residuals  5  19.22   3.845
```

```
#------------------------- Model 4

> Prom.combine <- I(Promotion=="Prom 2")+I(Promotion=="Prom 3")
> model.4 <- lm(Y~Store+X+Prom.combine)
> summary(model.4)

Call:
lm(formula = Y ~ Store + X + Prom.combine)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   14.8433     7.8846   1.883  0.09652 .
StoreStore 2   0.3186     3.8590   0.083  0.93623
StoreStore 3  -1.0147     3.8590  -0.263  0.79923
StoreStore 4   1.4119     4.1617   0.339  0.74314
StoreStore 5  -1.6421     3.8794  -0.423  0.68323
X              1.0147     0.2974   3.412  0.00920 **
Prom.combine  -9.3398     2.7031  -3.455  0.00863 **


Residual standard error: 4.712 on 8 degrees of freedom
Multiple R-squared:  0.7252,Adjusted R-squared:  0.5191
F-statistic: 3.518 on 6 and 8 DF,  p-value: 0.05225

> anova(model.4)

Analysis of Variance Table

Response: Y
             Df  Sum Sq Mean Sq F value   Pr(>F)
Store         4  65.067  16.267  0.7325 0.594715
X             1 138.578 138.578  6.2406 0.037050 *
Prom.combine  1 265.110 265.110 11.9388 0.008628 **
Residuals     8 177.645  22.206
```