***Name:* ______________________________ *NetID:* __________________________**

***Signature:* ____________________________**

STAT 420 – Midterm Exam 1B

Sharing or copying any part of the exam is an infraction of the University's rules on Academic Integrity and will be disciplined accordingly pursuant to the Student Code of Conduct.

Students are allowed to use a standard or graphing calculator, but no other form of electronic device. This is a "closed notes" exam, however students are allowed one 8.5" x 11" sheet containing anything they like on both sides.

No credit will be given without supporting work. Partial credit is available for incorrect answres, but only with accompanying work.

Page	Possible Points	Earned
1	12	
2	12	
3	12	
4	16	
5	6	
6	17	
7	6	
8	11	
9	8	
Total		

1. The number of points a student earns on an exam is often thought to be determined by how prepared the student is. For $n = 10$ students, the following values have been recorded.

y = Final Exam points,
x_1 = number of absences,
x_2 = average number of hours spent studying per week

Consider the model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon,$$

where ε's are i.i.d. $N(0, \sigma^2)$.

x_1	x_2	y
1	3	40
1	5	40
2	9	37
2	3	17
3	12	79
3	10	79
4	16	76
4	12	66
5	14	63
5	16	93

Then $$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 10 & 30 & 100 \\ 30 & 110 & 360 \\ 100 & 360 & 1220 \end{bmatrix}; \quad (\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} 0.575 & -0.075 & -0.025 \\ -0.075 & 0.275 & -0.075 \\ -0.025 & -0.075 & 0.025 \end{bmatrix};$$

$$\mathbf{X}^T\mathbf{Y} = \begin{bmatrix} 590 \\ 2{,}010 \\ 6{,}820 \end{bmatrix}, \quad \text{and} \quad \hat{\boldsymbol{\beta}} = \begin{bmatrix} 18 \\ -3 \\ 5 \end{bmatrix}, \qquad \sum (y_i - \hat{y}_i)^2 = 1{,}400,$$

and $\sum (y_i - \bar{y})^2 = 5{,}280.$

a) (12) Perform the significance of the regression test at the 5% level of significance.

$H_0 : \beta_1 = \beta_2 = 0$ vs. H_1 : at least one $\beta_j \neq 0$

Completing the ANOVA table,

Source	SS	df	MS	F
Regression	3880	$p - 1 = 2$	1940	**9.70**
Error	1400	$n - p = 7$	200	
Total	5280	$n - 1 = 9$		

The critical region is $F > F_\alpha(2,7) = F_{0.05}(2,7) =$ **4.74**.
With a calculator, you get p-value = **0.0096**.
As a result, **we reject H_0**; the model is a significant model for predicting Final Exam score.

1. (continued)

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 18 \\ -3 \\ 5 \end{bmatrix}, \quad (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} = \begin{bmatrix} 0.575 & -0.075 & -0.025 \\ -0.075 & 0.275 & -0.075 \\ -0.025 & -0.075 & 0.025 \end{bmatrix}.$$

b) (7) Test $\mathrm{H}_0 : \beta_2 = 0$ vs. $\mathrm{H}_1 : \beta_2 \neq 0$ at the 10% level of significance.

Use MSE = 200 from part a as the estimate of the variance of the residuals.

$$\hat{\mathrm{V}}\mathrm{ar}\left[\hat{\beta}_2\right] = \hat{\sigma}^2 \cdot C_{33} = (200)(0.025) = 5$$

Calculate the test statistic.

$$t = \frac{\hat{\beta}_2 - \beta_{20}}{\sqrt{\hat{\mathrm{V}}\mathrm{ar}\left[\hat{\beta}_2\right]}} = \frac{5-0}{\sqrt{5}} = \mathbf{2.236}$$

There are $n - p = 7$ degrees of freedom. The critical region is $|t| > t_{\alpha/2}(n-p) = t_{0.05}(7) =$ **1.895**.

With a calculator, you get p-value = **0.0604**.

As a result, **we reject *H*₀**; β_2 is a significant predictor in the model.

c) (5) Construct a 95% confidence interval for β_1.

Use MSE = 200 from part a as the estimate of the variance of the residuals.

$$\hat{\mathrm{V}}\mathrm{ar}\left[\hat{\beta}_1\right] = \hat{\sigma}^2 \cdot C_{22} = (200)(0.275) = 55$$

So, the 95% confidence interval for β_1 is

$$\hat{\beta}_1 \pm t_{\alpha/2}(n-p) \cdot \sqrt{\hat{\mathrm{V}}\mathrm{ar}\left[\hat{\beta}_1\right]} = -3 \pm t_{0.025}(7) \cdot \sqrt{55} = -3 \pm 2.365 \cdot 7.416$$

$$= \mathbf{-3 \pm 17.54}$$

$$= \mathbf{(-20.54,\ 14.54)}$$

1. (continued) $\hat{\boldsymbol{\beta}} = \begin{bmatrix} 18 \\ -3 \\ 5 \end{bmatrix}$, $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} = \begin{bmatrix} 0.575 & -0.075 & -0.025 \\ -0.075 & 0.275 & -0.075 \\ -0.025 & -0.075 & 0.025 \end{bmatrix}$.

d) (6) Construct a 90% prediction interval for the final exam score of a student who missed 2 days of class and studied an average of 14 hours per week.

The vector representing these predictors is $\boldsymbol{x}_0 = [\;1 \quad 2 \quad 14\;]$. The estimate for the average wait time is

$$\hat{y} = [\;1 \quad 2 \quad 14\;]\begin{bmatrix} 18 \\ -3 \\ 5 \end{bmatrix} = 18 - 3(2) + 5(14) = 82 .$$

To calculate the estimate for the variance of the estimate, we need

$$\boldsymbol{x}_0'(X'X)^{-1}\boldsymbol{x}_0 = [\;1 \quad 2 \quad 14\;]\begin{bmatrix} 0.575 & -0.075 & -0.025 \\ -0.075 & 0.275 & -0.075 \\ -0.025 & -0.075 & 0.025 \end{bmatrix}\begin{bmatrix} 1 \\ 2 \\ 14 \end{bmatrix} = [\;1 \quad 2 \quad 14\;]\begin{bmatrix} 0.075 \\ -0.575 \\ 0.175 \end{bmatrix} = 1.375$$

So, the 90% prediction interval is

$$\hat{y} \pm t_{\alpha/2}(n-p)\cdot\sqrt{\hat{\mathrm{V}}\mathrm{ar}[Y \mid x]} = 82 \pm t_{0.05}(7)\cdot\sqrt{\hat{\sigma}^2\cdot(1+1.375)} = 82 \pm 1.895\cdot\sqrt{200\cdot 2.375}$$

$$= \mathbf{82 \pm 41.30}$$

$$= \mathbf{(40.7,\ 123.3)}$$

e) (3) Interpret β_0 in the context of the problem.

β_0 represents the average final exam score of students who did not miss any classes but who also did not do any studying.

f) (3) Interpret β_2 in the context of the problem.

β_1 represents the average change in the final exam score for each additional absence from class (while holding study hours constant)

2. (16) Suppose a complete second-order model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \varepsilon$$

was fit to $n = 27$ data points.

```
> sum( lm( y ~ 1 )$residuals^2 )
[1]  400

> summary( lm( y ~ x2 + x4 + x5 + x6 ) )$r.squared
[1]  0.60

> summary( lm( y ~ x1 + x3 ) )$r.squared
[1]  0.70

> sum( lm( y ~ x1 + x2 + x3 + x4 + x5 + x6 )$residuals^2 )
[1]  90
```

Test $H_0 : \beta_1 = \beta_3 = 0$ at a 5% level of significance.
State the alternative hypothesis, the value of the test statistic, the critical value(s), and a decision.

$H_0 : \beta_1 = \beta_3 = 0$ is also represented by the Null Model of

$$Y = \beta_0 + \qquad + \beta_2 x_2 + \qquad + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \varepsilon$$

H_1 : at least one $\beta_j \neq 0$ is also represented by the Full Model of

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \varepsilon$$

For the Full Model, $df = n - p = 27 - 7 = 20$ and $\text{SSE}_{\text{Full}} = 90$.

For the Null Model, $df = n - q = 27 - 5 = 22$ but SSE_{Null} is not given. However, $R^2 = 0.60$ for that model, so $R^2 = 1 - \frac{\text{SSE}_{\text{Null}}}{SYY} = 1 - \frac{\text{SSE}_{\text{Null}}}{400} = 0.60$. Thus, $\text{SSE}_{\text{Null}} = 160$.

	SS	df	MS	F
Difference	70	$p - q = 2$	35	**7.78**
Full Model	90	$n - p = 20$	4.5	
Null Model	160	$n - q = 22$		

The critical region is $F > F_\alpha(2,20) = F_{0.05}(2,20) =$ **3.49**.
With a calculator, you get p-value = **0.0032**.
As a result, **we reject H_0**; the Full Model is better.

3. A vaccine is shipped by airfreight to medical facilities in cartons, each containing 1,000 vials. The data presented here concerns 10 such shipments.

y = number of broken vials at final destination

x = number of times the carton was transferred from one aircraft to another

x	y
1	14
0	7
2	15
0	10
3	20
1	11
0	6
1	13
2	17
0	9

Consider the model

$$Y = \beta_0 + \beta_1 x + \varepsilon,$$

where ε's are i.i.d. $N(0, \sigma^2)$.

$\sum x = 10$; $\sum y = 122$; $\sum x^2 = 20$; $\sum y^2 = 1{,}666$; $\sum x\,y = 162$;

$\sum (x - \overline{x})^2 = 10$; $\sum (y - \overline{y})^2 = 177.6$; $\sum (x - \overline{x})(y - \overline{y}) = 40.$

a) (6) Find the equation of the least-squares regression line.

$$\overline{x} = \frac{\sum x}{n} = \frac{10}{10} = 1 \qquad \overline{y} = \frac{\sum y}{n} = \frac{122}{10} = 12.2$$

$$\hat{\beta}_1 = \frac{\text{SXY}}{\text{SXX}} = \frac{40}{10} = 4$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \cdot \overline{x} = 12.2 - 4\,(1) = 8.2.$$

Least-squares regression line: $\boldsymbol{\hat{y} = 8.2 + 4\,x}$

3. (continued)

b) (12) Perform the significance of the regression test at the 1% level of significance.

$H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$

We have to have SSE, so

$$\text{SSReg} = \hat{\beta}_1^2 \cdot SXX = (4)^2 \cdot 10 = 160$$

SSE = SYY – SSReg = 177.6 – 160 = 17.6

Solution A:

Completing the ANOVA table,

Source	SS	df	MS	F
Regression	160	$p - 1 = 1$	160	**72.73**
Error	17.6	$n - p = 8$	2.2	
Total	177.6	$n - 1 = 9$		

The critical region is $F > F_\alpha(1,8) = F_{0.01}(1,8) =$ **11.26**.

With a calculator, you get p-value = **2.74E-05**.

As a result, **we reject H_0**; the model is a significant model for predicting the number of broken vials.

Solution B:

The variance of the residuals is $s_e^2 = \frac{SSE}{n-2} = \frac{17.6}{8} = 2.2$.

Calculate the t-test statistic.

$$t = \frac{\hat{\beta}_1 - \beta_{10}}{s_e / \sqrt{SXX}} = \frac{4-0}{\sqrt{2.2}/\sqrt{10}} = \mathbf{8.528}$$

The critical region is $|t| > t_{\alpha/2}(8) = t_{0.005}(8) =$ **3.355**.

With a calculator, you get p-value = **2.74E-05**.

As a result, **we reject H_0**; the model is a significant model for predicting the number of broken vials.

c) (5) Construct a 90% prediction interval for the number of broken vials after a shipment that had 3 aircraft transfers.

$$\hat{y} \pm t_{\alpha/2} \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{SXX}} = (8.2 + 4 \cdot 3) \pm t_{0.05}(8) \cdot 1.48 \cdot \sqrt{1 + \frac{1}{10} + \frac{(3-1)^2}{10}}$$

$$= 20.2 \pm 1.860 \cdot 1.48 \cdot \sqrt{1 + \frac{1}{10} + \frac{4}{10}}$$

$$= \mathbf{20.2 \pm 3.37}$$

$$= \mathbf{(16.83, 23.57)}$$

4. (6) Consider the following data set:

x_1	x_2	y
1	1	6
2	1	9
3	0	10
4	0	11
2	1	15
4	0	17
3	1	18
5	0	18

Consider the model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i.,$$

$$i = 1, \ldots, 8.$$

where ε_i's are i.i.d. $N(0, \sigma_e^2)$.

$$\mathbf{X}^{\mathrm{T}}\mathbf{X} = \begin{bmatrix} 8 & 24 & 4 \\ 24 & 84 & 8 \\ 4 & 8 & 4 \end{bmatrix}; \quad (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} = \begin{bmatrix} 4.25 & -1 & -2.25 \\ -1 & 0.25 & 0.5 \\ -2.25 & 0.5 & 1.5 \end{bmatrix}; \quad \mathbf{X}^{\mathrm{T}}\mathbf{Y} = \begin{bmatrix} 104 \\ 340 \\ 48 \end{bmatrix}$$

.

Obtain the least-squares estimates $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$.

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} = (X'X)^{-1}(X'Y) = \begin{bmatrix} 4.25 & -1 & -2.25 \\ -1 & 0.25 & 0.5 \\ -2.25 & 0.5 & 1.5 \end{bmatrix}\begin{bmatrix} 104 \\ 340 \\ 48 \end{bmatrix} = \begin{bmatrix} \mathbf{-6} \\ \mathbf{5} \\ \mathbf{8} \end{bmatrix}$$

5. For this problem we will use a random sample of 35 vehicles from a data set provided by the Environmental Protection Agency regarding fuel economy in cars.

- y = mileage (in miles per gallon)
- x_1 = engine horsepower
- x_2 = top speed (in miles per hour)
- x_3 = vehicle weight (in hundreds of lbs.)

Consider the model $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$, where $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$, and the following output from R.

```
> summary(fit.mpg)

Call:
lm(formula = y ~ x1 + x2 + x3)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 215.2455    32.6848   6.585 3.24e-07 ***
x1            0.4724     0.1118   4.227 0.000215 ***
x2           -1.5165     0.3377  -4.491 0.000104 ***
x3           -2.1171     0.2704  -7.830 1.23e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.51 on [redacted] degrees of freedom
Multiple R-squared:  0.9195,     Adjusted R-squared:  0.9112
F-statistic: 110.5 on [redacted] and [redacted] DF,  p-value: [redacted]
```

a) (5) Construct a 90% confidence interval for β_2.

The *df* for the model is $n - p = 35 - 4 = 31$, so the 90% confidence interval for β_2 is

$$\hat{\beta}_2 \pm t_{\alpha/2}(n-p)\cdot \mathrm{SE}\left[\hat{\beta}_2\right] = -1.5165 \pm t_{0.05}(31)\cdot 0.3377 = -1.5165 \pm 1.645\cdot 0.3377$$

$$= \mathbf{-1.5165 \pm 0.5555}$$

$$= \mathbf{(-2.07, -0.96)}$$

You could also opt to use the more conservative $t_{0.05}(30) = 1.697$ in which case the margin of error would be 0.5731 and the CI would be (–2.09, –0.94).

b) (6) Perform the significance of the regression test at the 5% level of significance.

The test statistic of F = **110.5** is given.
The critical region is $F > F_\alpha(p-1, n-p) = F_{0.05}(3,31)$ = **2.91**.
With a calculator, you get p-value = **1.11E-16**.
As a result, **we reject H_0**; the model is a significant model for predicting mileage.

6. (8) Suppose the number of number of times a carton is transferred from one aircraft to another (X) and the number of broken vials upon delivery (Y) follow a bivariate normal distribution with

$$\mu_X = 2.0, \quad \sigma_X = 1.5, \quad \mu_Y = 12, \quad \sigma_Y = 5, \quad \rho = 0.70.$$

Suppose that a recent carton shipment makes 4 aircraft transfers. What is the probability that the carton contains no more than 15 broken vials?

We want $P(Y \le 15 \mid X = 4)$. Given that $X = 4$, Y is Normal with

$$\mathrm{E}[Y \mid X] = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X) = 12 + 0.70 \cdot \frac{5}{1.5}(4-2) = 16.67,$$

$$\mathrm{Var}[Y \mid X] = (1-\rho^2)\sigma_Y^2 = (1-0.70^2)5^2 = 12.75, \text{ and}$$

$$\mathrm{SD}[Y \mid X] = 3.57.$$

Thus,

$$P(Y \le 15 \mid X = 4) = P\left(Z \le \frac{15-16.67}{3.57}\right) = P(Z \le -0.47) = \mathbf{0.3192}.$$