STAT 420 – Midterm Exam 1A

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

x_1	x_2	y
1	1	20
1	3	29
2	7	43
2	1	6
3	10	75
3	8	66
4	14	89
4	10	66
5	12	71
5	14	95

Consider the model

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

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

Then $$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 10 & 30 & 80 \\ 30 & 110 & 300 \\ 80 & 300 & 860 \end{bmatrix}, \qquad (\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} 0.575 & -0.225 & 0.025 \\ -0.225 & 0.275 & -0.075 \\ 0.025 & -0.075 & 0.025 \end{bmatrix},$$

$$\mathbf{X}^T\mathbf{Y} = \begin{bmatrix} 560 \\ 2{,}020 \\ 5{,}780 \end{bmatrix}, \quad \text{and} \quad \hat{\boldsymbol{\beta}} = \begin{bmatrix} 12 \\ -4 \\ 7 \end{bmatrix},$$

$$\sum (y_i - \hat{y}_i)^2 = 350, \quad \text{and} \quad \sum (y_i - \bar{y})^2 = 8{,}090.$$

a) (12) Perform the significance of the regression test at the 10% 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	7740	$p - 1 = 2$	3870	**77.40**
Error	350	$n - p = 7$	50	
Total	8090	$n - 1 = 9$		

The critical region is $F > F_{\alpha}(2,7) = F_{0.10}(2,7) =$ **3.26**.
With a calculator, you get p-value = **1.68E-05**.
As a result, **we reject H_0**; the model is a significant model for predicting Final Exam score.

1. (continued) $\hat{\boldsymbol{\beta}} = \begin{bmatrix} 12 \\ -4 \\ 7 \end{bmatrix}$, $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} = \begin{bmatrix} 0.575 & -0.225 & 0.025 \\ -0.225 & 0.275 & -0.075 \\ 0.025 & -0.075 & 0.025 \end{bmatrix}$.

b) (7) Test $\mathrm{H}_0 : \beta_2 = 0$ vs. $\mathrm{H}_1 : \beta_2 \neq 0$ at the 5% level of significance.

Use MSE = 50 from part a as the estimate of the variance of the residuals.

$$\hat{\mathrm{V}}\mathrm{ar}\left[\hat{\beta}_2\right] = \hat{\sigma}^2 \cdot C_{33} = (50)(0.025) = 1.25$$

Calculate the test statistic.

$$t = \frac{\hat{\beta}_2 - \beta_{20}}{\sqrt{\hat{\mathrm{V}}\mathrm{ar}\left[\hat{\beta}_2\right]}} = \frac{7-0}{\sqrt{1.25}} = \mathbf{6.261}$$

There are $n - p = 7$ degrees of freedom. The critical region is $|t| > t_{\alpha/2}(n-p) = t_{0.025}(7) =$ **2.365**.

With a calculator, you get p-value = **0.00042**.

As a result, **we reject H_0**; β_2 is a significant predictor in the model.

c) (5) Construct a 90% confidence interval for β_1.

Use MSE = 50 from part a as the estimate of the variance of the residuals.

$$\hat{\mathrm{V}}\mathrm{ar}\left[\hat{\beta}_1\right] = \hat{\sigma}^2 \cdot C_{22} = (50)(0.275) = 13.75$$

So, the 90% confidence interval for β_1 is

$$\hat{\beta}_1 \pm t_{\alpha/2}(n-p) \cdot \sqrt{\hat{\mathrm{V}}\mathrm{ar}\left[\hat{\beta}_1\right]} = -4 \pm t_{0.05}(7) \cdot \sqrt{13.75} = -4 \pm 1.895 \cdot 3.708$$

$$= \mathbf{-4 \pm 7.03}$$

$$= \mathbf{(-11.03,\ 3.03)}$$

1. (continued) $\hat{\boldsymbol{\beta}} = \begin{bmatrix} 12 \\ -4 \\ 7 \end{bmatrix}$, $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} = \begin{bmatrix} 0.575 & -0.225 & 0.025 \\ -0.225 & 0.275 & -0.075 \\ 0.025 & -0.075 & 0.025 \end{bmatrix}$.

d) (6) Construct a 95% prediction interval for the final exam score of a student who missed 3 days of class and studied an average of 12 hours per week.

The vector representing these predictors is $\boldsymbol{x}_0 = \begin{bmatrix} 1 & 3 & 12 \end{bmatrix}$. The estimate for the average wait time is

$$\hat{y} = \begin{bmatrix} 1 & 3 & 12 \end{bmatrix} \begin{bmatrix} 12 \\ -4 \\ 7 \end{bmatrix} = 12 - 4(3) + 7(12) = 84 .$$

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

$$\boldsymbol{x}_0'(X'X)^{-1}\boldsymbol{x}_0 = \begin{bmatrix} 1 & 3 & 12 \end{bmatrix} \begin{bmatrix} 0.575 & -0.225 & 0.025 \\ -0.225 & 0.275 & -0.075 \\ 0.025 & -0.075 & 0.025 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \\ 12 \end{bmatrix} = \begin{bmatrix} 1 & 3 & 12 \end{bmatrix} \begin{bmatrix} 0.2 \\ -0.3 \\ 0.1 \end{bmatrix} = 0.5$$

So, the 95% prediction interval is

$$\hat{y} \pm t_{\alpha/2}(n-p) \cdot \sqrt{\hat{\mathrm{V}}\mathrm{ar}[Y \mid x]} = 84 \pm t_{0.025}(7) \cdot \sqrt{\hat{\sigma}^2 \cdot (1 + 0.5)} = 84 \pm 2.365 \cdot \sqrt{50 \cdot 1.5}$$

= 84 ± 20.48

= (63.52, 104.48)

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_1 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

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \varepsilon$$

was fit to $n = 32$ data points.

```
> sum( lm( y ~ 1 )$residuals^2 )
[1]  600

> summary( lm( y ~ x2 + x4 + x6 ) )$r.squared
[1]  0.65

> summary( lm( y ~ x1 + x3 + x5 + x7 ) )$r.squared
[1]  0.72

> sum( lm( y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7)$residuals^2 )
[1]  150
```

Test $H_0 : \beta_2 = \beta_4 = \beta_6 = 0$ at a 10% level of significance.
State the alternative hypothesis, the value of the test statistic, the critical value(s), and a decision.

$H_0 : \beta_2 = \beta_4 = \beta_6 = 0$ is also represented by the Null Model of

$$Y = \beta_0 + \beta_1 x_1 \qquad + \beta_3 x_3 \qquad + \beta_5 x_5 \qquad + \beta_7 x_7 + \varepsilon$$

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

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \varepsilon$$

For the Full Model, $df = n - p = 32 - 8 = 24$ and $\text{SSE}_{\text{Full}} = 150$.

For the Null Model, $df = n - q = 32 - 5 = 27$ but SSE_{Null} is not given. However, $R^2 = 0.72$ for that model, so $R^2 = 1 - \frac{\text{SSE}_{\text{Null}}}{SYY} = 1 - \frac{\text{SSE}_{\text{Null}}}{600} = 0.72$. Thus, $\text{SSE}_{\text{Null}} = 168$.

	SS	df	MS	F
Difference	18	$p - q = 3$	6	**0.96**
Full Model	150	$n - p = 24$	6.25	
Null Model	168	$n - q = 27$		

The critical region is $F > F_\alpha(2,7) = F_{0.10}(2,7) = \mathbf{2.33}$.
With a calculator, you get p-value = **0.4276**.
As a result, **we fail to reject H_0**; the Null 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	16
0	9
2	17
0	12
3	22
1	13
0	8
1	15
2	19
0	11

Consider the model

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

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

$\sum x = 10$; $\sum y = 142$; $\sum x^2 = 20$; $\sum y^2 = 2{,}194$; $\sum x y = 182$;

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

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

$$\bar{x} = \frac{\sum x}{n} = \frac{10}{10} = 1 \qquad \bar{y} = \frac{\sum y}{n} = \frac{142}{10} = 14.2$$

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

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x} = 14.2 - 4\,(1) = 10.2.$$

Least-squares regression line: $\boldsymbol{\hat{y} = 10.2 + 4\,x}$

3. (continued)

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

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

We have to have SSE, so

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

$$\text{SSE} = \text{SYY} - \text{SSReg} = 177.6 - 160 = 17.6$$

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 $F > F_{\alpha}(1,8) = F_{0.05}(1,8) =$ **5.32**.

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 $s_e^2 = \dfrac{SSE}{n-2} = \dfrac{17.6}{8} = 2.2$.

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 $|t| > t_{\alpha/2}(8) = t_{0.025}(8) =$ **2.306**.

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 95% prediction interval for the number of broken vials after a shipment that had 2 aircraft transfers.

$$\hat{y} \pm t_{\alpha/2} \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{SXX}} = (10.2 + 4 \cdot 2) \pm t_{0.025}(8) \cdot 1.48 \cdot \sqrt{1 + \frac{1}{10} + \frac{(2-1)^2}{10}}$$

$$= 18.2 \pm 2.306 \cdot 1.48 \cdot \sqrt{1 + \frac{1}{10} + \frac{1}{10}}$$

$$= \mathbf{18.2 \pm 3.74}$$

$$= \mathbf{(14.46, 21.94)}$$

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

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

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

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) 172.7589    41.0736   4.206 0.000205 ***
x1            0.3459     0.1406   2.461 0.019633 *
x2           -1.1219     0.4210  -2.665 0.012112 *
x3           -1.7274     0.3504  -4.930 2.62e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.142 on [redacted] degrees of freedom
Multiple R-squared:  0.8436,     Adjusted R-squared:  0.8285
F-statistic: 55.75 on [redacted] and [redacted] DF,  p-value: [redacted]
```

a) (5) Construct a 95% confidence interval for β_2.

The df for the model is $n - p = 35 - 4 = 31$, so the 95% confidence interval for β_2 is

$$\hat{\beta}_2 \pm t_{\alpha/2}(n-p)\cdot \mathrm{SE}\left[\hat{\beta}_2\right] = -1.1219 \pm t_{0.025}(31)\cdot 0.4210 = -1.1219 \pm 1.96 \cdot 0.4210$$

$$= \mathbf{-1.1219 \pm 0.8252}$$

$$= \mathbf{(-1.95, -0.30)}$$

You could also opt to use the more conservative $t_{0.025}(30) = 2.042$ in which case the margin of error would be 0.8597 and the CI would be (–1.98, –0.26).

b) (6) Perform the significance of the regression test at the 10% level of significance.

The test statistic of F = **55.75** is given.
The critical region is $F > F_\alpha(p-1, n-p) = F_{0.10}(3,31) =$ **2.27**.
With a calculator, you get p-value = **1.36E-12**.
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

$$\mu_X = 1.5, \quad \sigma_X = 1, \quad \mu_Y = 15, \quad \sigma_Y = 4, \quad \rho = 0.60.$$

Suppose that a recent carton shipment makes 3 aircraft transfers. What is the probability that the carton contains no more than 20 broken vials?

We want $P(Y \le 20 \mid X = 3)$. Given that $X = 3$, Y is Normal with

$$\mathrm{E}[Y \mid X] = \mu_Y + \rho \frac{\sigma_Y}{\sigma_X}(x - \mu_X) = 15 + 0.60 \cdot \frac{4}{1}(3 - 1.5) = 18.6,$$

$$\mathrm{Var}[Y \mid X] = (1 - \rho^2)\sigma_Y^2 = (1 - 0.60^2)4^2 = 10.24, \text{ and}$$

$$\mathrm{SD}[Y \mid X] = 3.2.$$

Thus,

$$P(Y \le 20 \mid X = 3) = P\left(Z \le \frac{20 - 18.6}{3.2}\right) = P(Z \le 0.44) = \mathbf{0.6700}.$$