STAT 420 – Midterm Exam 2A

1. In order to compare the average GPA for the students of three distinct departments at a university, eight students were randomly chosen from each of the three departments (Astronomy, Biology, Communication), the students' GPA (y) and the average time spent studying per week (x) in hours was recorded. Consider the model

$$Y = \beta_0 + \beta_1 v_1 + \beta_2 v_2 + \beta_3 x + \varepsilon,$$

where $v_1 = 1$ if a student is from Astronomy, 0 otherwise;
$v_2 = 1$ if a student is from Biology, 0 otherwise.

To answer parts a-c, consider…

Astronomy:	$Y = \beta_0 + \beta_1$		$+ \beta_3 x + \varepsilon$		
Biology:	$Y = \beta_0$	$+ \beta_2$	$+ \beta_3 x + \varepsilon$		
Communication:	$Y = \beta_0$		$+ \beta_3 x + \varepsilon$	$\leftarrow$	base category

a) (3) Give an interpretation (in the context of the problem) to the regression coefficient β_0.

The average GPA of Communication majors who do not study ($x = 0$).

b) (3) Give an interpretation (in the context of the problem) to the regression coefficient β_1.

The difference between the average GPA of Astronomy majors and the average GPA of Communication majors who spend the same amount of time studying.

c) (3) Give an interpretation (in the context of the problem) to the regression coefficient β_3.

The change in the average GPA for one additional hour of studying per week (regardless of major).

1. (continued)

Examine the following sum of squares calculations.

```
> sum( lm( y ~ v1 + v2 + x )$residuals^2 )
[1] 8.0
> sum( lm( y ~ v1 + v2 )$residuals^2)
[1] 14.0
> sum( lm( y ~ x + 0 )$residuals^2)
[1] 18.0
> sum( lm( y ~ x )$residuals^2)
[1] 11.0
> sum( lm( y ~ 1 )$residuals^2)
[1] 20.0
```

d) (14) We wish to test if the relationship between GPA and time spent studying is the same for all three departments. Perform the appropriate test at $\alpha = 0.10$. Based on the result of this test, which model (full or null) is preferred?

$H_0 : \beta_1 = \beta_2 = 0$ is also represented by the Null Model of $Y = \beta_0 + \beta_3 x + \varepsilon$

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

$$Y = \beta_0 + \beta_1 v_1 + \beta_2 v_2 + \beta_3 x + \varepsilon$$

For the Full Model, $df = n - p = 24 - 4 = 20$ and $SSE_{Full} = 8$.

For the Null Model, $df = n - q = 24 - 2 = 22$ and $SSE_{Null} = 11$.

	SS	df	MS	*F*
Difference	3	$p - q = 2$	1.5	**3.75**
Full Model	8	$n - p = 20$	0.4	
Null Model	11	$n - q = 22$		

The critical region is $F > F_\alpha(2,7) = F_{0.10}(2,20) =$ **2.59**.
With a calculator, you get *p*-value = **0.0414**.
As a result, **we reject H_0**; the Full Model is better.

1. (continued)

e) (9) Compute the AIC values for the full and the null models from part d. If the AIC model selection criteria is used, which model (full or null) is preferred?

Using the actual formula,…

$$\text{AIC}_{\text{Full}} = 24 + 24\ln(2\pi) + 24\ln\left({}^{8.0}\!/_{24}\right) + 2\times 4 = \mathbf{49.74}$$

$$\text{AIC}_{\text{Null}} = 24 + 24\ln(2\pi) + 24\ln\left({}^{11.0}\!/_{24}\right) + 2\times 2 = \mathbf{53.39}$$

Or using the reduced formula, since $n + n\ln(2\pi)$ is constant for all models,…

$$\text{AIC}_{\text{Full}} = 24\ln\left({}^{8.0}\!/_{24}\right) + 2\times 4 = \mathbf{-18.37}$$

$$\text{AIC}_{\text{Null}} = 24\ln\left({}^{11.0}\!/_{24}\right) + 2\times 2 = \mathbf{-14.72}$$

The lower AIC is preferred, so in either case, we choose **the Full Model**.

f) (9) Compute the Adjusted R-squared values for the full and the null models from part d. If the Adjusted R-squared model selection criteria is used, which model (full or null) is preferred?

$$R^2_{\text{Full}} = 1 - \frac{\text{SSResidual}}{\text{SYY}} = 1 - \frac{8.0}{20.0} = 0.60.$$

$$\text{Adjusted } R^2_{\text{Full}} = 1 - \frac{n-1}{n-p}\cdot\left(1 - R^2\right) = 1 - \frac{23}{20}\cdot(1 - 0.60) = \mathbf{0.54}.$$

$$R^2_{\text{Null}} = 1 - \frac{\text{SSResidual}}{\text{SYY}} = 1 - \frac{11.0}{20.0} = 0.45.$$

$$\text{Adjusted } R^2_{\text{Null}} = 1 - \frac{n-1}{n-p}\cdot\left(1 - R^2\right) = 1 - \frac{23}{22}\cdot(1 - 0.45) = \mathbf{0.425}.$$

The higher Adjusted R^2 is preferred, so we choose **the Full Model**.

1. (continued)

Consider a new variable. Let $v_3 = 1$ if a student is from Communication, 0 otherwise.

g) (3) Find the Residual Sum of Squares for the model

$$Y = \beta_1 v_1 + \beta_2 v_2 + \beta_3 v_3 + \varepsilon.$$

This new model is the same as $Y = \beta_1 v_1 + \beta_2 v_2 + \varepsilon$ under the original setup. Therefore, the two models

$$Y = \beta_1 v_1 + \beta_2 v_2 + \varepsilon \qquad \text{and} \qquad Y = \beta_1 v_1 + \beta_2 v_2 + \beta_3 v_3 + \varepsilon.$$

would have the same predicted/fitted values, the same residuals, and the same Residual Sum of Squares.

```
> sum( lm( y ~ v1 + v2 )$residuals^2)
[1] 14.0
```

h) (4) Suppose we suspect that the rate (the slope) of the relationship between GPA and time spent studying may be different for the different departments. Suggest an appropriate model.

If you want different slopes, you need interaction terms.

$Y = \beta_0 + \beta_1 v_1 + \beta_2 v_2 + \beta_3 x + \beta_4 v_1 x + \beta_5 v_2 x + \varepsilon,$
Then,

Astronomy:	$Y = \beta_0 + \beta_1 + (\beta_3 + \beta_4)x + \varepsilon$
Biology:	$Y = \beta_0 + \beta_2 + (\beta_3 + \beta_5)x + \varepsilon$
Communication:	$Y = \beta_0 + \beta_3 x + \varepsilon$

OR…
$Y = \mu_1 v_1 + \mu_2 v_2 + \mu_3 v_3 + \gamma_1 v_1 x + \gamma_2 v_2 x + \gamma_3 v_3 x + \varepsilon,$
Then,

Astronomy:	$Y = \mu_1 + \gamma_1 x + \varepsilon$
Biology:	$Y = \mu_2 + \gamma_2 x + \varepsilon$
Communication:	$Y = \mu_3 + \gamma_3 x + \varepsilon$

i) (2) For your model in part h, we wish to test if the rate (the slope) of the relationship between GPA and time spent studying is the same for the three departments. Specify the null hypothesis H_0 using the notations of your part h model.

$H_0 : \beta_4 = \beta_5 = 0$ OR… $H_0 : \gamma_1 = \gamma_2 = \gamma_3$

---

2. A vaccine is shipped by airfreight to medical facilities in cartons, each containing 1,000 vials. Consider the model: $Y = \beta_0 + \beta_1 x + \varepsilon$,

where y = number of broken vials at final destination;
x = number of times the carton was transferred from one aircraft to another.

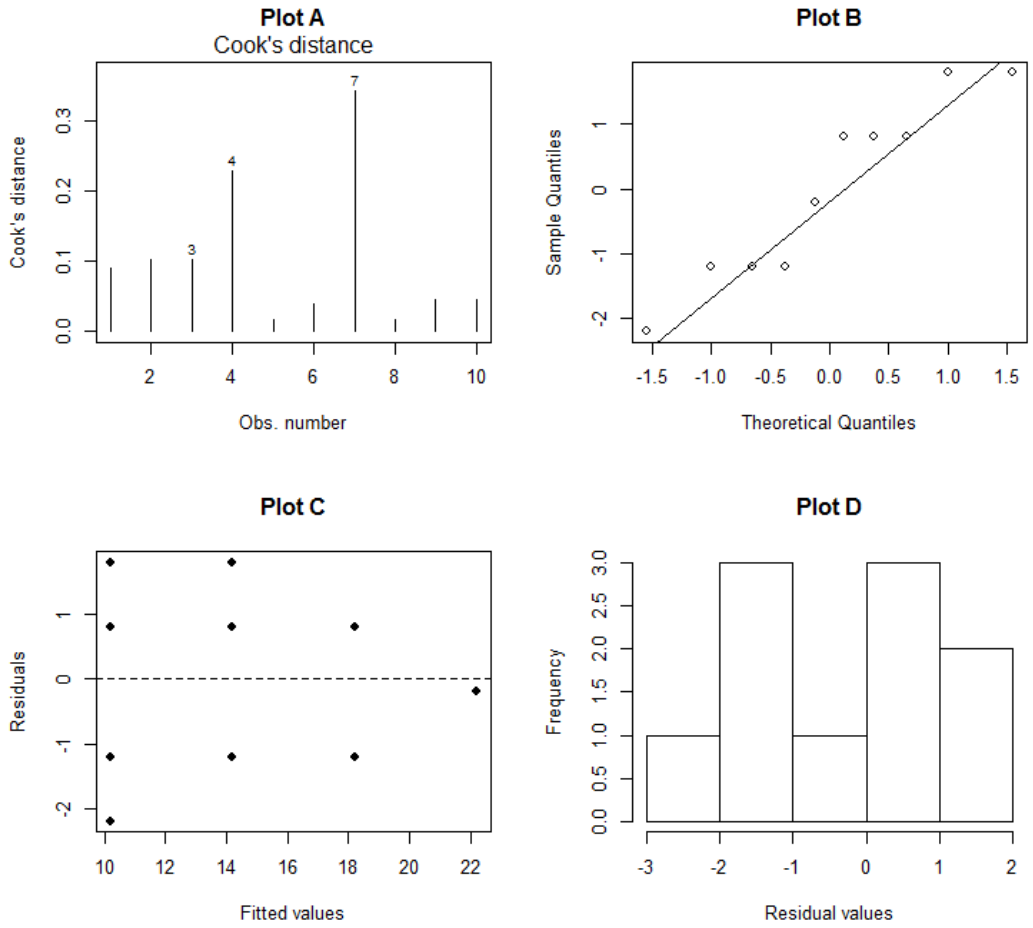
a) (8) List the assumptions about the distribution of the error terms (ε) that must hold in order for the model to be considered valid.

We assume that the error terms are

- independent,
- (identically) Normally distributed with
- mean 0 and
- constant variance, σ^2.

2. (continued)

b) (16) For each diagnostic plot below, (i) identify what model issues it is used to check and (ii) <u>briefly</u> explain whether this model is okay for each issue listed in (i) or if there's a problem.



Plot A: The Cook's Distance plot checks for any influential points. As long as all observations have Cook's Distance less than $4/n$ (which is $4/10 = 0.4$ here; or some use 0.5), then we're okay. This model does not seem to have any overly influential observations.

Plot B: The Normal Probability Plot (aka QQ Plot) checks the Normality assumption. The points generally follow the line, but we can't get a strong gauge on this because there's only 10 points.

Plot C: The residual plot is used to check the randomness of the error terms and constant variance. The points are generally scattered above and below the mean of 0, but the funnel pattern suggests a decrease in variance as the response values increase.

Plot D: The histogram of the residuals is used to check the Normality assumption. This plot doesn't appear to have much of a bell-shape, but with only 10 observations, it's hard to solidify behavior in the population.

2. (continued)

c) (6) There is something of particular concern regarding one of the model assumptions as seem in the plots of part b. Which of the following tests address that assumption? Do you still have a concern?

```
Shapiro-Wilk test

W = 0.9073, p-value = 0.2632
```

```
Durbin-Watson test

DW = 1.875, p-value = 0.4968
```

```
studentized Breusch-Pagan test

BP = 3.0628, df = 1, p-value = 0.0801
```

```
Levene's Test
      Df F value Pr(>F)
group  3  1.2667 0.3671
       6
```

```
Kolmogorov-Smirnov test

D = 0.2164, p-value = 0.7373
```

```
Bartlett test

Bartlett's K-squared = 0.4765, df = 1, p-value = 0.49
```

There are concerns about the constant variance assumption. The Breusch-Pagan test has a relatively low *p*-value which suggests that its null hypothesis (constant variance) does not hold.

If you were concerned about the normality test, look at the Shapiro-Wilk test. The *p*-value is rather large which suggests that the distribution can be explained with the Normal distribution.

[Bonus Knowledge: Levene's Test and the Bartlett Test also both check for constant variance. The Kolmogorov-Smirnov test checks for Normality. These each have such high *p*-values that we'd say the assumptions are met. However, the small sample size likely means that none of these tests are very reliable.

3. Apex Corporation, which makes corrugated paper products, is currently working to improve its cost control program. The company is analyzing its manufacturing costs to understand more fully the important influences on its costs. Monthly data has been assembled on a group of variables over the course of $n = 27$ months, and regression analysis is to be used to assess how these variables are related to total manufacturing costs. The variables are:

Y = total manufacturing cost per month in thousands of dollars (Cost)
X_1 = total production of paper per month in tons (Paper)
X_2 = total machine hours used per month (Machine)
X_3 = total overhead costs per month in thousands of dollars (Overhead).

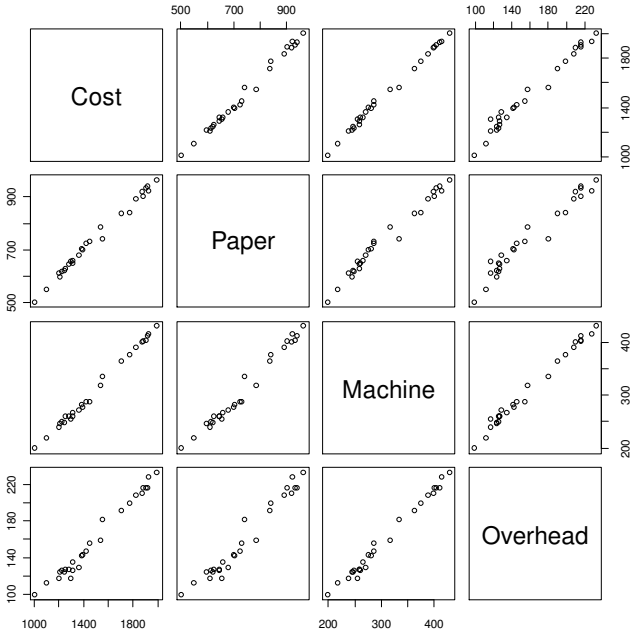
a) (4) Clearly state the first-order, multiple linear regression model for relating Cost to the predictors under consideration.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

b) (6) Consider the following scatterplot matrix and correlation matrix for the predictors and response variable. Comment on the relationship between the predictors and response, and among the predictors

	Cost	Paper	Machine	Overhead
Cost	1.000	0.996	0.997	0.989
Paper	0.996	1.000	0.989	0.978
Machine	0.997	0.989	1.000	0.994
Overhead	0.989	0.978	0.994	1.000

Each of the individual predictors is highly correlated with Cost suggesting that any one of them might be good for simple linear regression. However, the predictors are also highly correlated with one another suggesting collinearity, so it's unlikely that we'll find a significant multiple regression model.



3. (continued)

c) (7) Consider the following output and fill in the missing values.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	60.35437	20.84177	2.896	0.0082
Paper	0.95630	0.12133	7.882	0.0000
Machine	2.32128	0.45561	5.095	0.0000
Overhead	0.08603	0.53067	0.162	0.8726

Residual standard error: 11.213
Multiple R-squared: 0.9987 Adjusted R-squared: 0.9986

Analysis of Variance Table

	Sum Sq	Df	Mean Sq	F value	Pr(>F)
Regression	2255666	3	751888.7	5979.75	0.0000
Residuals	2892	23	125.739		
Total	2258558	26			

$$\text{SE(error)} = \sqrt{\text{MSE}} = \sqrt{125.739} = 11.213$$

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{2892}{2258558} = 0.9987$$

$$\text{Adjusted } R^2 = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE}{SST} = 1 - \left(\frac{26}{23}\right)\frac{2892}{2258558} = 0.9986$$

d) (3) Given only the information and analysis contained thus far, clearly state what you feel is the best linear regression model for predicting Cost.

The explanatory variable of Overhead (x_3) is not significant in the full model based on the individual *t*-test. That coupled with the likelihood of collinearity means that we're justified to remove it from the full model.

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