sample midterm

Instructions

Answer as many questions as you can. Be brief but include enough details to let me follow your reasoning.

- Notes, textbook, and calculator are allowed; computer is only allowed for Q2 and course materials.
- Each problem is 20 pts. Total points = 40. Time = 1 hr

1 Estimate a missing value.

Do the following by hand…

A sample of size $n = 6$ contains two independent variables X_1, X_2 and one categorical response variable Y. However, the response value for the 3$^{\text{rd}}$ sampling unit is missing.

	1	**2**	**3**	**4**	**5**	**6**
X_1	-2	-1	0	1	2	3
X_2	2	-2	0	0	3	2
Y	A	A	???	B	B	A

Predict the missing $Y_{i=3}$ by the following classification methods.

a. Predict $Y_{i=3}$ by KNN method with $k = 1$.

b. Predict $Y_{i=3}$ by KNN method with $k = 3$.

Obs.	X_1	X_2	Y	Dist to $\vec{X} = (0,0)$	Neigh for $k = 1$	Neigh for $k = 3$
1	-2	2	A	$\sqrt{8}$		A
2	-1	-2	A	$\sqrt{5}$		A
3	0	0				
4	1	0	B	$\sqrt{1}$	B	B
5	2	3	B	$\sqrt{13}$		
6	3	2	A	$\sqrt{13}$		
$\hat{Y}$					B	A

c. Logistic regression, without the 3$^{\text{rd}}$ sampling unit, produced the following results.

```
> z = 1*(y=="A")
> lreg = glm( z ~ x1 + x2, family=binomial)

> summary(lreg)

Coefficients: Estimate

Intercept 0.8373

x1 -0.4494

x2 -0.0778
```

- Estimate the probability $Y_{i=3} = A$.

The log odds is $\frac{p}{1-p} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 = -.8373 + 0 + 0.$

Thus $p(Y_{i=3} = A) = p = \frac{e^{.8373}}{1+e^{.8373}} = .6979.$

Predict A based on higher probability.

d. Suppose X_1 and X_2 are independent Normal random variables with the same variance $\sigma^2 = 1$. Assuming equal prior probabilities $P(A) = P(B) = 0.5$, predict $Y_{i=3}$ using linear discriminant analysis.

Given both are normal with the same variance and equal priors, the prediction is based on which group mean is the closest to the point of interest.

The discriminant function $\delta_k(x)$ can be written as:

$$\delta_k(X) = \ln(p_k) - \frac{1}{2\sigma^2}(X - \mu_k)^2 = \ln(.5) - \frac{1}{2}(X - \mu_k)^2$$

Where $X = (0, 0)$ and the mean μ_k is estimated by the sample mean for each group, $k \in (A, B)$.

We calculate the sample mean for a group by calculating the mean of the X_{1_i} for the points in the group and then the mean of the X_{2_i} for the points in the group. Group A has points 1, 2, and 6 and Group B has points 4 and 5. The group mean is thus $(\bar{X}_1, \bar{X}_2)$

- Group A: $\hat{\mu}_A = \left(\frac{-2-1+3}{3}, \frac{2-2+2}{3}\right) = (0, \frac{2}{3})$.
- Group B: $\hat{\mu}_B = \left(\frac{1+2}{2}, \frac{0+3}{2}\right) = (\frac{3}{2}, \frac{3}{2})$.

To maximize the discriminant function, we minimize the distance between the point of interest $X = (0, 0)$ and the group means. Comparing two groups under these conditions simplifies the comparison to whether

$$\ln(.5) - \frac{1}{2}(X - \hat{\mu}_A)^2 > \ln(.5) - \frac{1}{2}(X - \hat{\mu}_B)^2$$
$$(X - \mu_A)^2 > (X - \mu_B)^2$$

We will predict the Group whose mean is closest to $X = (0, 0)$ using Euclidean distance.

- Group A: $\sqrt{(0-0)^2 + (\frac{2}{3} - 0)^2} = 2/3.$

- Group B: $\sqrt{(\frac{3}{2}-0)^2+(\frac{3}{2}-0)^2}=\frac{3}{2}\sqrt{2}$.
- $2/3 < \frac{3}{2}\sqrt{2}$ so predict Group A.

e. Estimate the training error rate of the $k = 3$ KNN algorithm used in b.

Obs.	X_1	X_2	Y	**Dist**	**Neighbors**	$\hat{Y}$	**Error**
1	-2	2	A	0|17|13|17|25	1,4, 2 or 5	A/B	½
2	-1	-2	A	17|0|8|34|32	2,4,1	A	0
3	0	0	NA	NA	NA	NA	
4	1	0	B	13|8|0|10|8	4,2,6	A	1
5	2	3	B	17|34|10|0|2	5,6,4	B	0
6	3	2	A	15|32|8|1|0	6,5,4	B	1
							2.5/5 = .5

f. (Stat-627 only) Estimate the testing error rate of the algorithm used in b using leave-one-out cross-validation.

Obs.	X_1	X_2	Y	**Dist**	**Neighbors**	$\hat{Y}$	**Error**
1	-2	2	A	0|17|13|17|25	4, 2 , 5	B	1
2	-1	-2	A	17|0|8|34|32	4,1,6	A	0
3	0	0	NA	NA	NA	NA	
4	1	0	B	13|8|0|10|8	2,6,5	A	1
5	2	3	B	17|34|10|0|2	6,4,1	A	1
6	3	2	A	15|32|8|1|0	5,4, A	B	1
							4/5 = .8

2 Toothgrowth

An experiment was conducted to evaluate the effect of vitamin C on tooth growth. Sixty guinea pigs received various doses of vitamin C by one of two delivery methods, orange juice or ascorbic acid.

- Results of this experiment are in dataset `ToothGrowth` which is already loaded in R.
- You can look at it with commands `names(ToothGrowth)`, `summary(ToothGrowth)`, `dplyr::gimpse(ToothGrowth`, and `?ToothGrowth`.

```
library(tidyverse)
glimpse(ToothGrowth)
```

```
Rows: 60
Columns: 3
$ len  <dbl> 4.2, 11.5, 7.3, 5.8, 6.4, 10.0, 11.2, 11.2, 5.2, 7.0, 16.5, 16.5,…
$ supp <fct> VC, VC, VC, VC, VC, VC, VC, VC, VC, VC, VC, VC, VC, VC, VC, VC, V…
$ dose <dbl> 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 1.0, 1.0, 1.0, …
```

a. Fit a linear regression model to predict tooth length based on the predictors `dose` and the supply delivery method of vitamin C (`supp` where Orange Juice (OJ) = 0 and Ascorbic Acid (VC) = 1) including all interactions.

1. Is delivery method significant?
2. Is there a significant interaction between the `dose` and delivery method?
3. Write two regression equations explicitly, one equation for each delivery method.
4. What percent of the total variation of the tooth length is explained by this regression?

```
reg <- lm(len ~ dose*supp, data = ToothGrowth)
summary(reg)
```

```
Call:
lm(formula = len ~ dose * supp, data = ToothGrowth)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.2264 -2.8462  0.0504  2.2893  7.9386 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   11.550      1.581   7.304 1.09e-09 ***
dose           7.811      1.195   6.534 2.03e-08 ***
suppVC        -8.255      2.236  -3.691 0.000507 ***
dose:suppVC    3.904      1.691   2.309 0.024631 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.083 on 56 degrees of freedom
Multiple R-squared:  0.7296,	Adjusted R-squared:  0.7151 
F-statistic: 50.36 on 3 and 56 DF,  p-value: 6.521e-16
```

- a.1 There is substantial evidence (p = 0.000507) to reject the null hypotheses of no relationship between length and supply method.

- a.2 There is strong evidence (p = 0.024631) to reject the null hypotheses of no interactions between dose and supply method.

- a.3

  - Ascorbic Acid: $\hat{Y} = 11.5 + 7.811 * dose - 8.255(1) + 3.904 * dose = 3.245 + 11.715 * dose$
  - Orange Juice: $\hat{Y} = 11.5 + 7.811 * dose$

- a.4 $R^2 = 0.7296$, or 73% of the total variation is explained by this model.

b. Conduct a lack-of-fit test to decide whether the relation between the dose and the tooth length is linear or their might be a better non-linear relationship.

- State the test statistic, the p-value, and your conclusion.

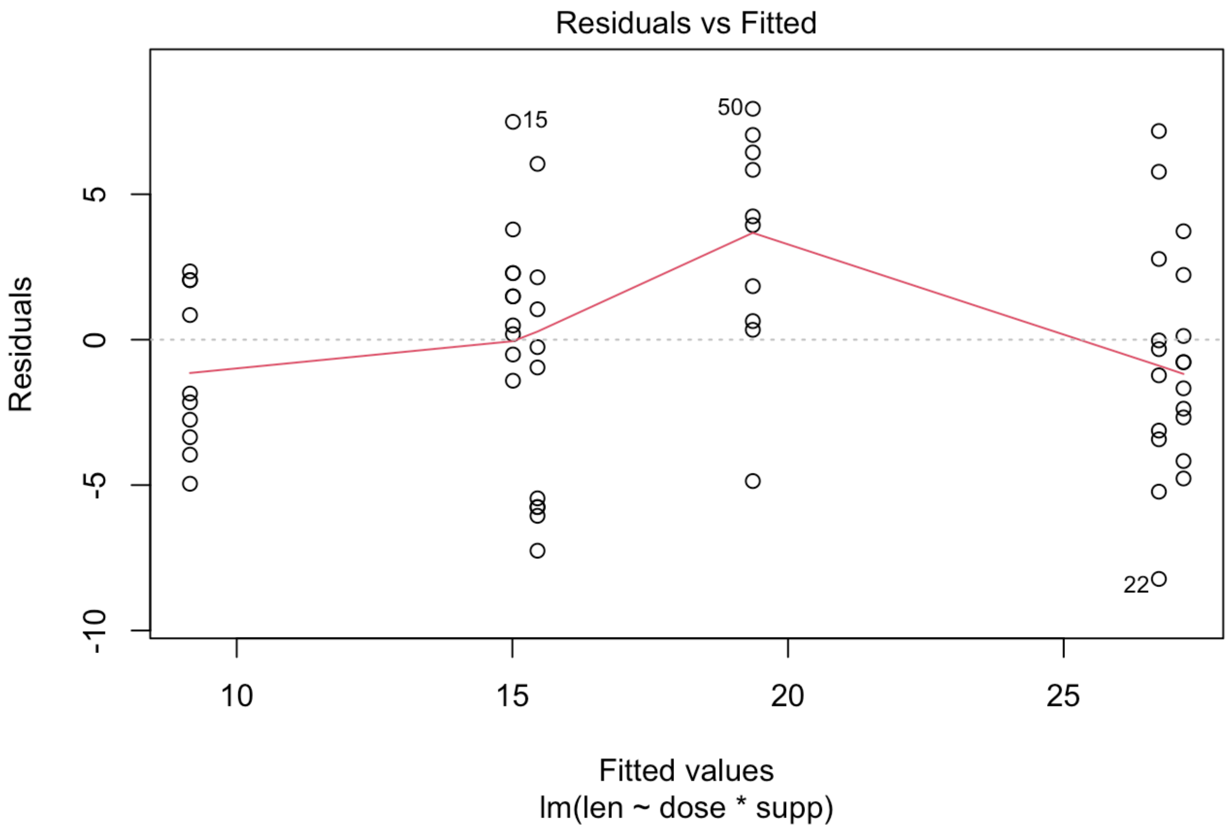
```
reg_saturated <- lm(len ~ as.factor(dose) * supp, data = ToothGrowth)
anova(reg, reg_saturated)
```

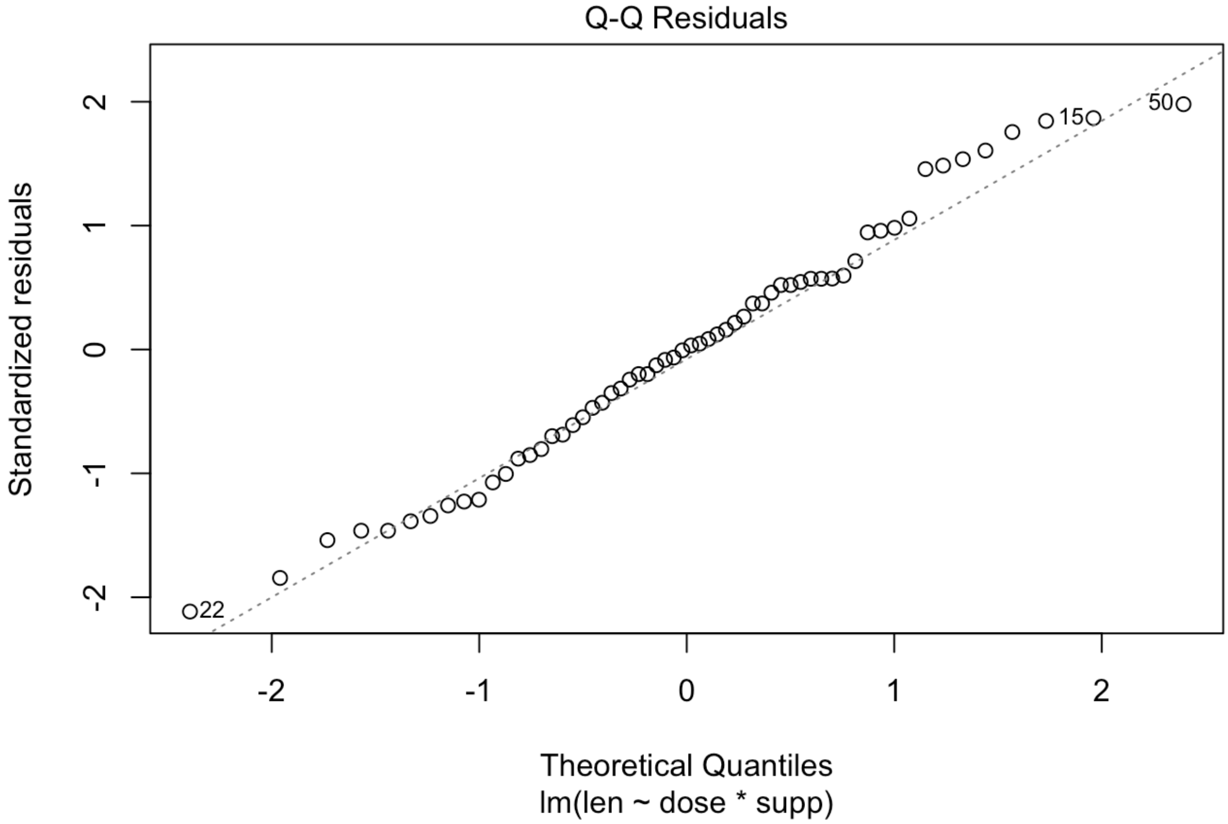
```
Analysis of Variance Table

Model 1: len ~ dose * supp
Model 2: len ~ as.factor(dose) * supp
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)    
1     56 933.63                                  
2     54 712.11  2    221.53 8.3994 0.0006667 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

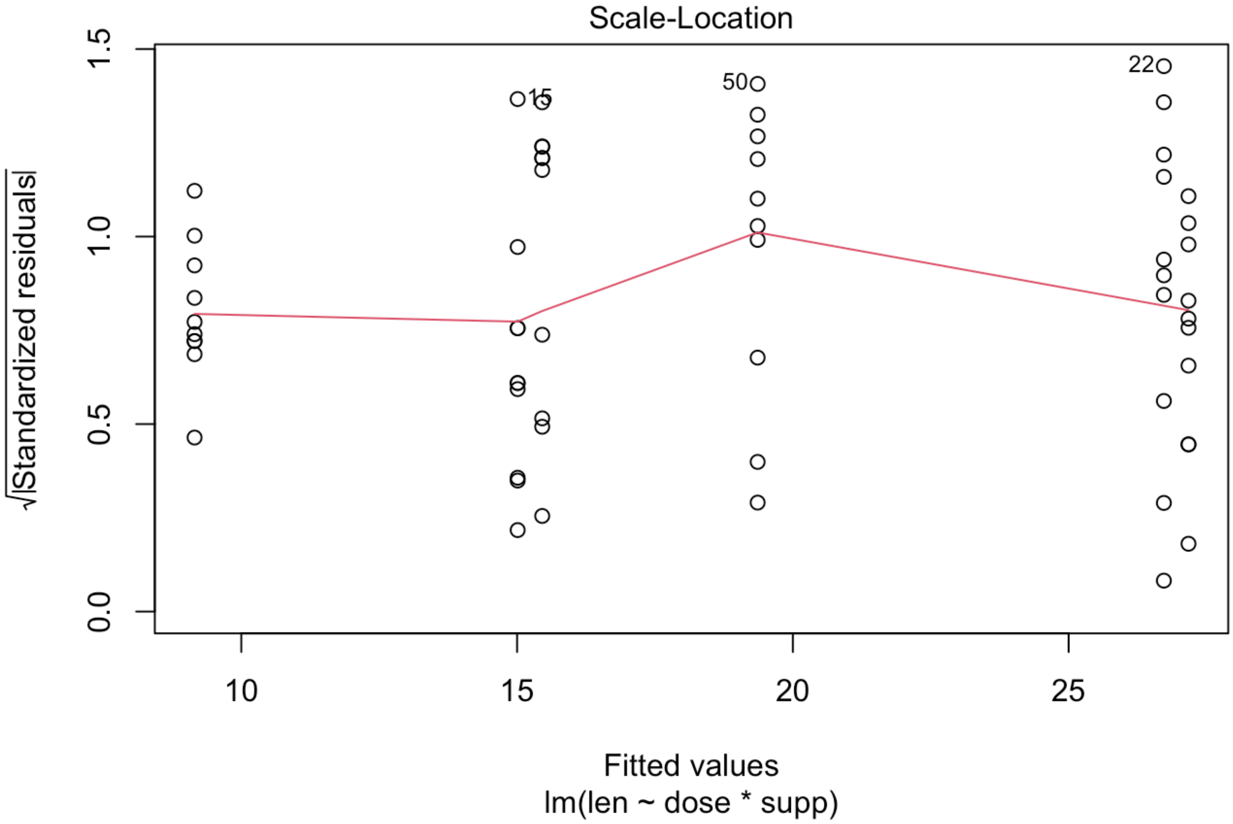
- The test statistic is the F statistic for a partial F-test which here is 8.3994.
- The p-value is 0.0006667.
- There is substantial evidence to reject the Null hypothesis that a non-liner model cannot do better than a linear model. We don’t know which form it would take.

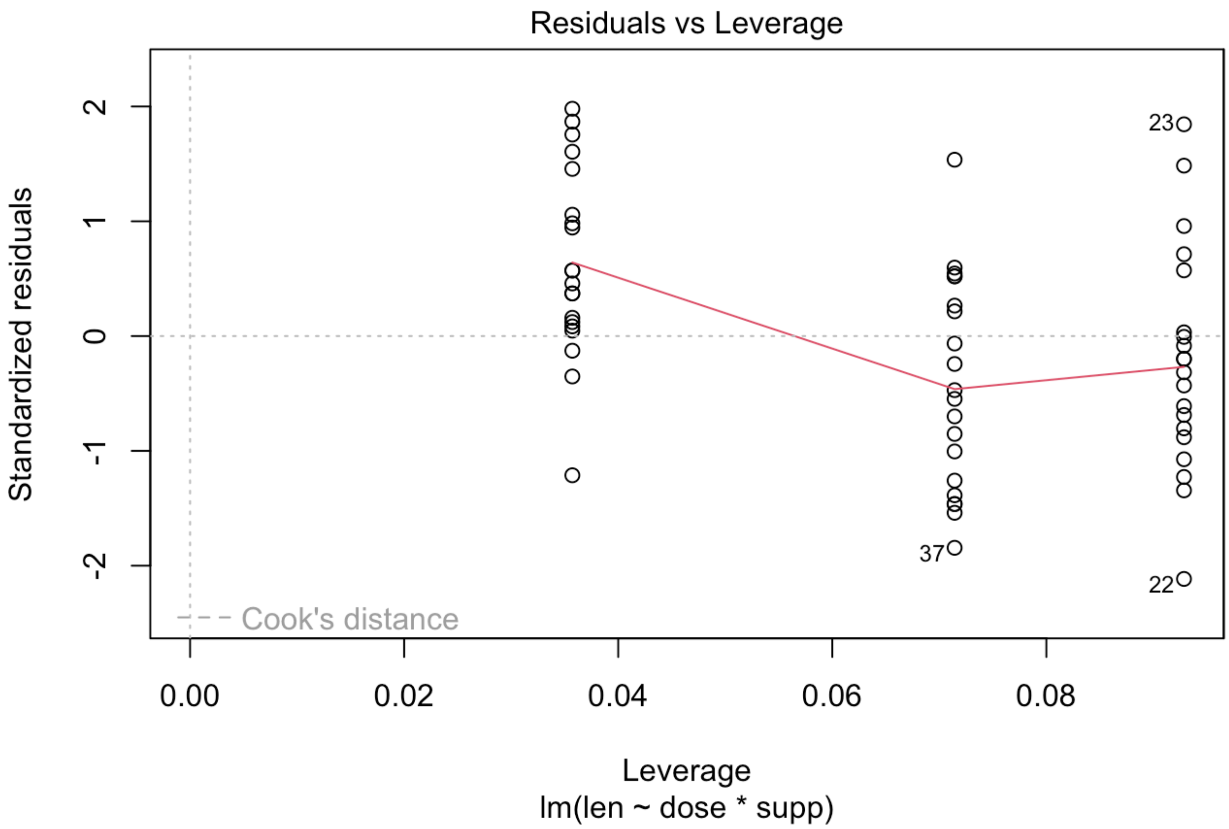
c. Use `plot()` to examine the studentized residuals. Are there any obvious extreme values? Explain your reasoning.

```
plot(reg)
```




Q-Q Residuals
Standardized residuals
2
1
0
-1
-2
50
15
22
-2
-1
0
1
2
Theoretical Quantiles
lm(len ~ dose * supp)


Scale-Location
√|Standardized residuals|
1.5
1.0
0.5
0.0
15
50
22
10
15
20
25
Fitted values
lm(len ~ dose * supp)



None of the values are greater than 3 standard deviations from 0.

d. Create training and testing subsets and estimate the prediction mean squared error. Use a seed of 1234 and a 60% split.

```
set.seed(1234)
Z <- sample(nrow(ToothGrowth), .6 * nrow(ToothGrowth))
reg <- lm( len ~ dose * supp, data = ToothGrowth[Z ,])
len_predicted <- predict(reg, data.frame(ToothGrowth[-Z,]))
MSE <-  mean((len_predicted - ToothGrowth$len[-Z])^2 )
MSE
```

```
[1] 17.1946
```