Name Jerry

Student Number

STA 442/2101 f2016 Quiz 4

1. (3 points) For the general linear model with fixed effects and normal errors (see formmula sheet) find the distribution of the vector of residuals **e**. Show your work and **circle your final answer.**

No marks off for not saying this

$e = (I-H)y = (I-H)(X\beta + \epsilon)$.
Since ϵ is multivariate normal, e is multivariate normal by the formula sheet.

$E(e) = (I-H)X\beta = X\beta - X\underbrace{(X^TX)^{-1}X^TX}\beta$
$= X\beta - X\beta = 0$ and

$cov(e) = cov((I-H)y) = (I-H)\sigma^2 I_n (I-H)^T$
$= \sigma^2 (I-H)$, so

$e \sim N(0, \sigma^2(I-H))$

No marks for starting with $e = y - \hat{y}$ and treating y and $\hat{y}$ as independent.

Name Jerry

Student Number

STA 442/2101 f2016 Quiz Four

1. (3 points) For the general linear model with fixed effects and normal errors (see formmula sheet) find the distribution of the vector of predicted values $\hat{\mathbf{y}}$. Show your work and **circle your final answer.**

No marks off for not saying this

$\hat{y} = X\hat{\beta}$. Since $\hat{\beta}$ is normal by the formula sheet, $\hat{y}$ is normal as well.

$$E(\hat{y}) = E(X\hat{\beta}) = XE(\hat{\beta}) = X\beta$$

$$cov(\hat{y}) = cov(X\hat{\beta}) = X\,cov(\hat{\beta})X^T$$

$$= X\sigma^2(X^TX)^{-1}X^T$$

Or they could start with $\hat{y} = Hy$. In either case,

$$\hat{y} \sim N(X\beta, \sigma^2 X(X^TX)^{-1}X^T)$$

either is okay

H

2. In a study comparing the effectiveness of different exercise programmes, volunteers were randomly assigned to one of three exercise programmes (A, B, C) or put on a waiting list and told to work out on their own. Aerobic capacity is the body's ability to process oxygen. Aerobic capacity was measured before and after 6 months of participation in the program (or 6 months of being on the waiting list). The response variable was improvement in aerobic capacity. The explanatory variables were age (a covariate) and treatment group. A regression model for this problem is

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 d_{i,1} + \beta_3 d_{i,2} + \beta_4 d_{i,3} + \epsilon_i,$$

where x_1 is age, and d_1, d_2 and d_3 are dummy variables for exercise program.

(a) (2 points) In the table below, indicate how the dummy variables are defined. There is more than one right answer, but make Wait list the reference category. In the last column, give the conditional expected improvement in terms of β values. The *symbols* for the dummy variables should not appear in the last column.

Programme	d_1	d_2	d_3	$E(y\|\mathbf{x})$
A	1	0	0	$\beta_0+\beta_2 + \beta_1 x$
B	0	1	0	$\beta_0+\beta_3 + \beta_1 x$
C	0	0	1	$\beta_0+\beta_4 + \beta_1 x$
Wait List	0	0	0	$\beta_0 + \beta_1 x$

(b) (1 point) What null hypothesis would you test to find out whether, controlling for age, the three exercise programmes differ in their average effectiveness? Give your answer in terms of β values.

$H_0: \beta_2 = \beta_3 = \beta_4 \cancel{= 0}$

(c) (1 point) What null hypothesis would you test to find out whether, controlling for age, the exercise programmes B is better than being on the wait list? Give your answer in terms of β values. In spite of the phrasing, it's a 2-tailed test.

$H_0: \beta_3 = 0$

(d) (1 point) What null hypothesis would you test to find out whether, controlling for exercise programme, age is related to improvement in aerobic capacity? Give your answer in terms of β values.

$H_0: \beta_1 = 0$

(e) (1 point) What is the difference in average benefit between programmes A and C for a 27 year old participant? Give your answer in terms of β values.

$\beta_2 - \beta_4$

(f) (1 point) Is it safe to assume that age is independent of exercise program? Answer Yes or No and indicate why. No marks without saying why.

Yes because of random assignment.