TEST1

Started: 23 Mar at 14:39

Quiz instructions

- Once started you have 60 minutes to complete this test.
- It is worth 10% of your final grade.
- A single attempt is allowed for each question.

Question 1 — 4 pts

Consider the multiple linear regression model

$$\mathbf{Y} = X\boldsymbol{\beta} + \mathbf{E},$$

where

$$\mathbf{E} \sim N_3(\mathbf{0}, \sigma^2 I),$$

and in this case

$$X = \begin{bmatrix} 1 & a/\sqrt{2} \\ 1 & 0 \\ 1 & -a/\sqrt{2} \end{bmatrix}$$

Let

$$\hat{\boldsymbol{\gamma}} = \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix} = A\hat{\boldsymbol{\beta}} + \mathbf{b},$$

where

$$A = \begin{bmatrix} b & 0 \\ 0 & c \end{bmatrix} \text{ and} \mathbf{b} = \begin{bmatrix} d \\ e \end{bmatrix}$$

Calculate $Var(\hat{\boldsymbol{\gamma}})$ and choose which of the following are true.

This question may have more than one correct answer. You must select all of the correct responses.

Marking: canvas marks as follows:

If there are k correct answers, and you select X answers of which

- X_1 are correct, and
- X_2 are incorrect, then

Your mark is

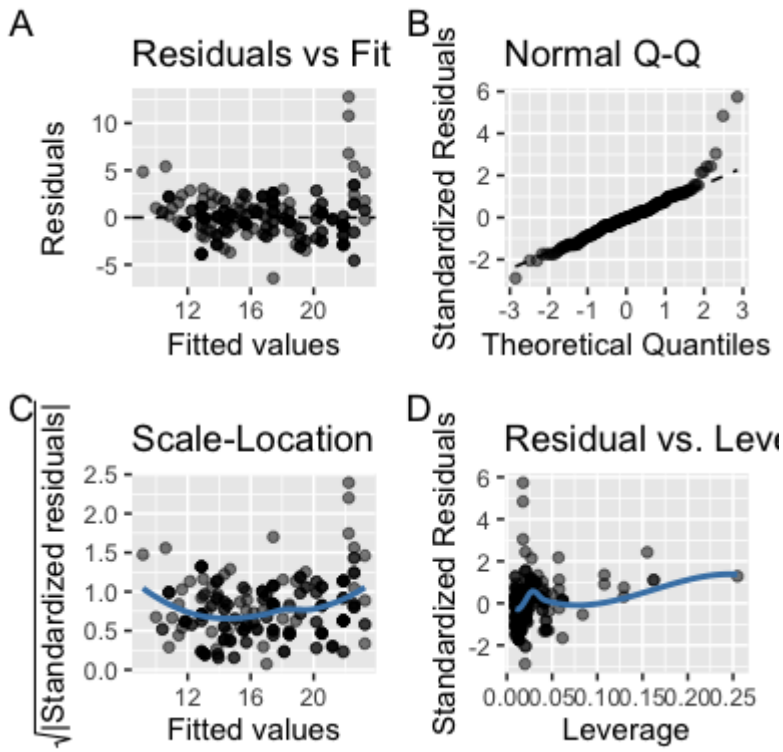
$$\frac{X_1}{k} - \frac{X_2}{k}.$$

- [ ] $var(\widehat{\gamma}_2) = c^2\sigma^2 + e^2$
- [ ] $var(\widehat{\gamma}_2) = \frac{c^2\sigma^2}{a^2}$
- [ ] $cov(\widehat{\gamma}_1, \widehat{\gamma}_2) = b^2 + c^2 + 2de$
- [ ] $var(\widehat{\gamma}_1) = \frac{b^2\sigma^2}{3} + d^2$

Question 2

4 pts

The following assumption plots have been produced for a linear model.



Assumption plots

Which of the following statements are correct?

This question may have more than one correct answer. You must select all of the correct responses.

Marking: canvas marks as follows:

If there are k correct answers, and you select X answers of which

- X_1 are correct, and
- X_2 are incorrect, then

Your mark is

$$\frac{X_1}{k} - \frac{X_2}{k}.$$

- [ ] Influential points are more likely to appear on the top right or bottom right of plot D
- [ ] The best plot to check normality of the residuals is B
- [ ] Influential points are more likely to appear close to the origin of plot D
- [ ] The best plot to check homoscedascity is D

Question 3

4 pts

The following design matrix X has been created.

```
##   (Intercept) TxB TxC x TxB:x TxC:x
## 1           1   0   0 1     0     0
## 2           1   1   0 1     1     0
## 3           1   0   1 1     0     1
## 4           1   0   0 2     0     0
## 5           1   1   0 2     2     0
## 6           1   0   1 2     0     2
## 7           1   0   0 3     0     0
## 8           1   1   0 3     3     0
## 9           1   0   1 3     0     3
## attr(,"assign")
## [1] 0 1 1 2 3 3
## attr(,"contrasts")
## attr(,"contrasts")$Tx
## [1] "contr.treatment"
```

Which of the following statements are correct?

This question may have more than one correct answer. You must select all of the correct responses.

Marking: canvas marks as follows:

If there are k correct answers, and you select X answers of which

- X_1 are correct, and
- X_2 are incorrect, then

Your mark is

$$\frac{X_1}{k} - \frac{X_2}{k}.$$

- [ ] `Tx` is a categorical variable with 2 levels
- [ ] The model used is `Tx / x`
- [ ] The model used is `Tx * x`
- [ ] The data set has 9 observations

Question 4

4 pts

A random sample of the `mpg` dataset has been taken and the following linear model fitted.

```
mpg_lm <- lm(cty ~ displ * drv, data = mpg)
```

The following output has been obtained:

```
##
## Call:
## lm(formula = cty ~ displ * drv, data = mpg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7034 -1.2850  0.0461  1.2936 10.6584
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  23.3059     1.2722  18.320  < 2e-16 ***
## displ        -2.2552     0.3166  -7.124 1.11e-10 ***
## drvf          5.9091     1.8402   3.211  0.00173 **
## drvr         -3.1252     4.5451  -0.688  0.49315
## displ:drvf   -1.3624     0.6053  -2.251  0.02636 *
```

```
## displ:drvr    1.0803     0.8894   1.215  0.22705
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.249 on 111 degrees of freedom
## Multiple R-squared:  0.7227, Adjusted R-squared:  0.7102
## F-statistic: 57.87 on 5 and 111 DF,  p-value: < 2.2e-16
```

Which of the following statements are correct?

This question may have more than one correct answer. You must select all of the correct responses.

Marking: canvas marks as follows:

If there are k correct answers, and you select X answers of which

- X_1 are correct, and
- X_2 are incorrect, then

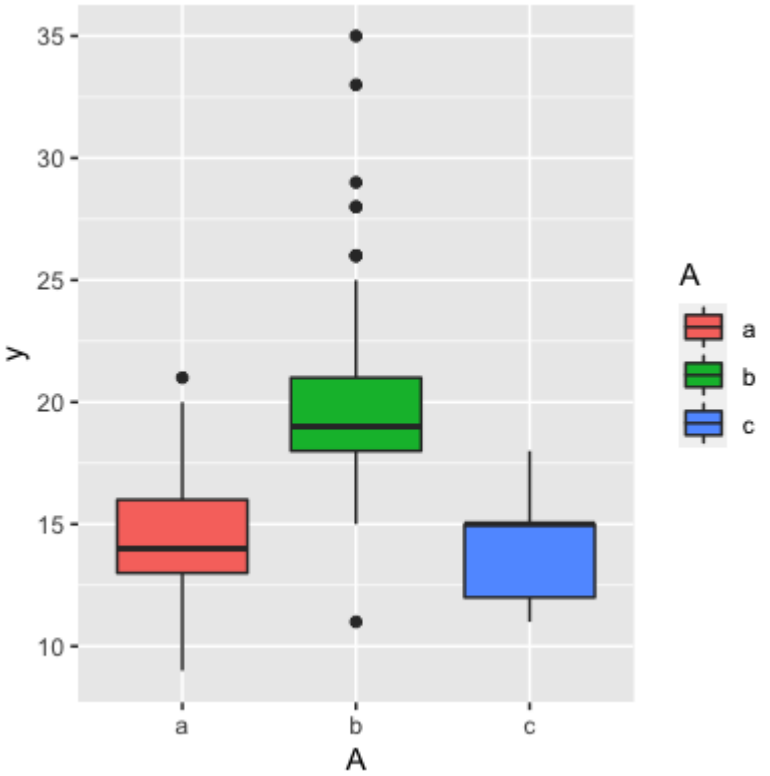
Your mark is

$$\frac{X_1}{k} - \frac{X_2}{k}.$$

- [ ] The reference level `drv` is f
- [ ] The statistic for the ANOVA test for comparing the full model to the null model is 57.87 on 5 and 111 degrees of freedom
- [ ] The estimated slope for front-wheel drive is -3.6176
- [ ] The reference level `drv` is 4

Question 5 1 pts

Consider the following output



Plot for question

Which one of the following R commands was used to make it?

- `df %>% group_by(A) %>% summarise(across(where(is.numeric), mean))`
- `df %>% ggplot(aes(A, y, fill = A)) + geom_boxplot()`
- `df %>% group_by(A) %>% summarise(m = mean(y)) %>% mutate(diff = m - mean(m))`
- `df %>% ggplot(aes(x, y, col = A)) + geom_point() + geom_smooth(method = lm)`

Question 6

4 pts

A data set has a response variable `y` and a categorical predictor `Tx`. The categorical variable has three levels. The following linear model is fitted:

```
lm1 <- lm(y ~ Tx, data = df, contrasts = list(Tx = "contr.sum"))
```

The following output is obtained:

```
##
## Call:
## lm(formula = y ~ Tx, data = df, contrasts = list(Tx = "contr.sum"))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9613 -0.4870 -0.1018  0.7271  1.6622
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.00830    0.18279  54.754   <2e-16 ***
## Tx1         -4.79023    0.25850 -18.531   <2e-16 ***
## Tx2         -0.03931    0.25850  -0.152     0.88
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.001 on 27 degrees of freedom
## Multiple R-squared:  0.9447, Adjusted R-squared:  0.9407
## F-statistic: 230.8 on 2 and 27 DF,  p-value: < 2.2e-16
```

Which of the following statements are correct?

This question may have more than one correct answer. You must select all of the correct responses.

Marking: canvas marks as follows:

If there are k correct answers, and you select X answers of which

- X_1 are correct, and
- X_2 are incorrect, then

Your mark is

$$\frac{X_1}{k} - \frac{X_2}{k}.$$

- [ ] The overall mean of `y` is 10.0083
- [ ] The group mean of Level 1 is 5.22
- [ ] The group mean of Level 3 is 14.84
- [ ] The overall mean of `y` is 5.1788

No new data to save. Last checked at 14:41 Submit quiz