UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

BIOSTATISTICS II

ASSIGNMENT 1

2022

STUDENT NAME: Nokwanda Themba

STUDENT NO.: THMNOK003

COURSE CODE: PPH7092S

DUE DATE: 12 August 2022

THIS ASSIGNMENT INCLUDES:

1. **Answers to questions**
2. **Figures and tables**

SUBMISSION OF THE ASSIGNMENT:

1. **Software used: R**
2. **Data Set: BiostatII_2022_Assign1_data.46.csv**
3. **THMNOK003_R Script Assignment 2**

LIST OF APPENDICES:

I. **R Script Code File**

Background:

This data is coming from a cohort of individuals living with HIV. These individuals had failed or were intolerant of zidovudine therapy (AZT).

Outcome: log10_vload – log base 10 viral load (copies/mL), which is one of the global responses to AIDS for improving treatment quality and health outcomes

Covariates of interest:

- patient - patient identifier
- age - measured in years
- sex - a factor with levels female and male
- AZT - a factor with levels intolerance and failure
- drug - antiretroviral drugs, coded as "didanosine (ddI)" and "zalcitabine (ddC)"
- prevOI - previous opportunistic infections coded as "AIDS" and "noAIDS""

Question 1 [27 marks]
a) Present a single table (Table 1) to describe your data. [10]

Table 1: Clinical Characteristics of study participants for a cohort study on individuals living with HIV who have failed or were intolerant to zidovudine therapy (overall sample and by AZT).

Variable	Total, N=693	Failure, N=143	Intolerance, N=550
Age (years)	31.00 (27.00, 35.00)	30.00 (27.00, 35.00)	31.00 (27.00, 35.00)
Median (IQR)			
v_load	9.18 (8.46, 9.71)	10.40 (9.90,10.62)	8.91 (8.29, 9.34)
Median (IQR)			
Educated [n(%)]	349 (50%)	68 (48%)	281 (51%)
Sex [n(%)]			
female	342 (49%)	54 (38%)	288 (52%)
male	351 (51%)	89 (62%)	262 (48%)
Drug [n (%)]			
ddC	330 (48%)	66 (46%)	264 (48%)
ddI	363 (52%)	77 (54%)	286 (52%)

1 Mean (SD); Median (IQR); n (%); s.d.,standard deviation; IQR, interquartile range
v_load, log10_vload – log base 10 viral load (copies/mL), which is one of the global responses to AIDS for improving treatment quality and health outcomes
drug ,antiretroviral drugs, coded as "didanosine (ddI)" and "zalcitabine (ddC)"
AZT, zidovudine therapy, a factor with levels of intolerance and failure

b) Present a single table (Table 2), showing results from the below models with coefficient estimates, 95% confidence intervals, p-values, and model fit statistics (only for the multiple linear regression model). [15]
i) Simple linear regression models using log10_vload as an outcome, for the covariates:
age, sex, drug, and zidovudine therapy.
ii) Multiple linear regression model using log10_vload as an outcome, including the covariates: age, gender, drug, and zidovudine therapy.

Table 2: Results from the multiple regression analysis conducted with log10_vload viral load as output and the covariates of interest as input, with empirically derived estimates and model statistics.

Model Summary Output:			
R-squared (R^2)	0.46		
Adjusted R Squared	0.46		
F-statistic	149.3		
Predictor variable	**Coefficient estimate**	**95% CI**	**P-value**
Intercept	9.80	[9.52, 10.07]	<0.001
Age	0.02	[0.01, 0.03]	<0.001
Sex -male	-0.14	[-0.24,-0.04]	0.007
Drug-ddI	-0.14	[-0.24,-0.04]	0.008
AZT-intolerance	-1.49	[-1.61,-1.37]	<0.001

Note: CI = Confidence interval
intercept = v_load, log10_vload – log base 10 viral load (copies/mL), which is one of the global responses to AIDS for improving treatment quality and health outcomes
Drug, antiretroviral drugs, coded as "didanosine (ddI)" and "zalcitabine (ddC)"
AZT, zidovudine therapy, a factor with levels of intolerance and failure
Dummy/indicator variables; drug-ddc, AZT-failure, sex -female as reference

c) Provide a conceptual AND fitted/predicted equation for each model in part (b). [2]
i) For the Simple Linear Regression:

Conceptual formula:	**Fitted/Predicted formula:**
log10_vload = β0 + β1 x Age + error	**$\hat{Y}$ = 8.58 + 0.02 x Age**
log10_vload= β0 + β1 x Sex + error	**$\hat{Y}$ = 9.10 + 0.02 x Sex**
log10_vload = β0 + β1 x Drug + error	**$\hat{Y}$ = 9.14 -0.07 x Drug**
log10_vload = β0 + β1 x AZT + error	**$\hat{Y}$= 10.27 -1.47 x AZT**

ii) For the Multiple linear regression:
Conceptual formula:

$$\text{log10\_vload\_i} = \beta 0 + \beta 1 \times \text{Age} + \beta 2 \times \text{Sex} + \beta 3 \times \text{Drug} + \beta 4 \times \text{AZT} + \text{error}$$

Fitted/Predicted formula:

$$\hat{Y} = 9.80 + 0.02 \times \text{Age} - 0.14 \times \text{Sex} - 0.14 \times \text{Drug} - 1.49 \times \text{AZT}$$

Question 2 [8 marks]
a) Describe in no more than 4 sentences the characteristics of the dataset based on your descriptive analysis. [4]
From Table 1 summarising the clinical characteristics of patients living with HIV in the sample cohort study, we can deduce that there were more educated people who were intolerant to AZT (51%) than those who had a failure to AZT(48%), and that females have a slightly higher intolerance to AZT than males with a percentage of 52% versus 48%%. With a contrast of 62 percent versus 38 percent, it also indicates that males have a much higher failure to respond to AZT than females do, and that both sexes have a higher proportion of intolerance to AZT than failure to respond to AZT. Additionally, it appears that more study participants who took the drug ddi (didanosine) were intolerant to AZT than those who took the drug DDC (zalcitabine), 52% versus 48%, and that there were more study participants who were taking the drug ddi (didanosine) had a failure to AZT than those who took the drug ddc (zalcitabine), 54% versus 46%. However, with a median of 10.40, log10 viral load was higher in study participants who had a failure to AZT than those who were intolerant to AZT.

b) Present figures that describe the relationship between the variables: age and outcome log10_vload, and zidovudine therapy and outcome log10_vload. Provide an appropriate caption for your figures (Figures 1 and 2), and a brief [no more than 4 sentences] explanation of what the results in the figures present. [4]

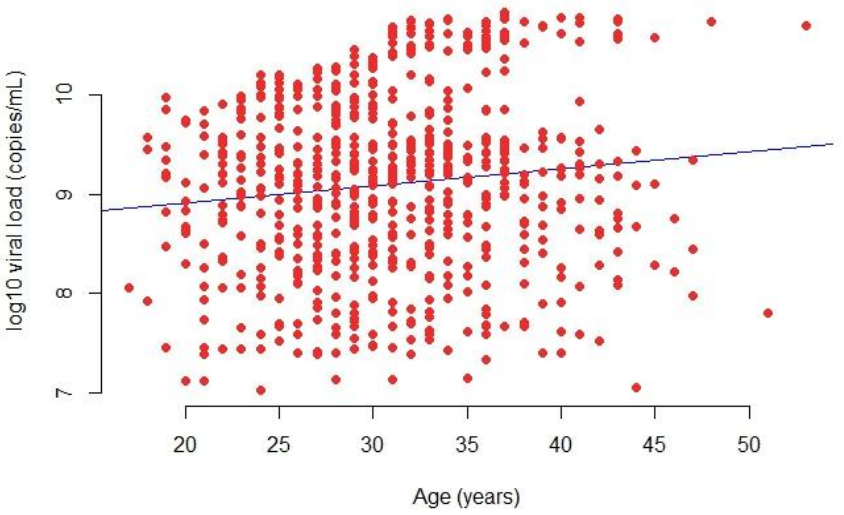

Figure 1 : Scatterplot illustrating the association between log base 10 viral load (copies/mL) and Age (years) of individuals in the cohort study sample.

Interpretation: Through visual inspection, it appears that there is a very weak positive to no correlation between log10 viral load and age in years. There is an overall very weak positive relationship between log base 10 viral load and participant age (correlation=0.12) as data points are less clustered together and spread out more, although we can see that higher viral load is also slightly associated with older age in participants (>30 years old have a log base 10 viral load higher than 10). This ultimately means there is no trend to the data, thus, there is no correlation.

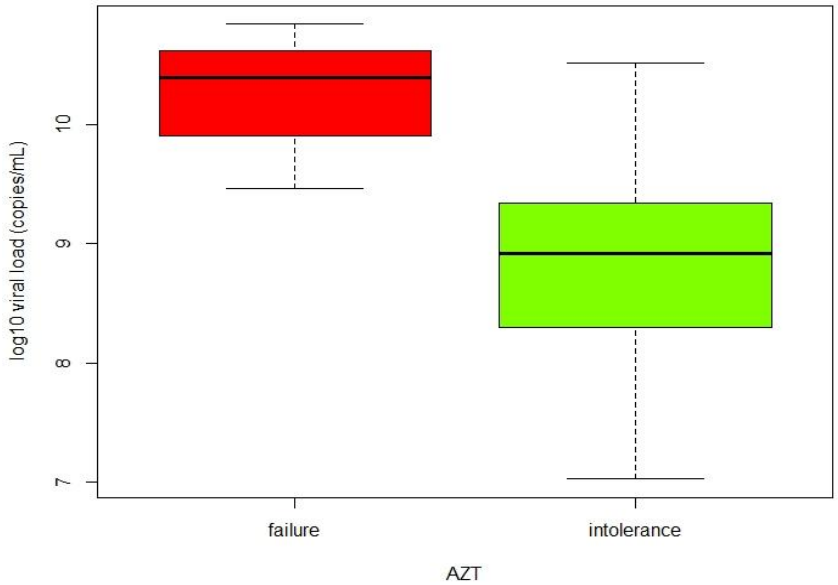

Figure 2: Boxplot showing the distribution of viral load at different AZT (zidovudine therapy) levels intolerance and failure in the cohort study sample.

Interpretation: It appears that participants who show an intolerance to AZT have lower log base 10 viral load levels than participants who show a failure to AZT (with lower median,interquartile range and maximum log base 10 viral load levels). The graph also reveals that participants who show a failure to AZT have higher distribution of log base 10 viral load levels in the sample, and that the log base 10 viral load distribution of failure to AZT is higher than the intolerance to AZT. We therefore conclude that in general, participants who have an intolerance to AZT have lower log base 10 viral load levels than those with AZT failure. Additionally, both data are screwed to the left (negatively skewed) and do not exhibit normal distribution, with AZT intolerance showing more dispersion of data as compared to AZT failure.

Question 3 [15 marks]

a) Completely interpret your regression findings and justify the use of multiple linear regression. [15]

How well does our model fit our data?

R squared: From our R squared (R^2) value of 0.46 we can see that our independent predictor variables explain 46% of the variability of our dependent variable, log base 10 viral load. An R^2 of 46% reveals that 46% of the variability observed in the dependent variable is explained by the regression model. For our cohort study sample data, the R^2 value indicates that the model provides a somewhat "okay" fit to the data. Ultimately, the model isn't fitting the data very well (R^2 <0.5), but if we take into account other aspects, a low R^2 value can also make a good predictive model. Even with a low R^2 value, we do have statistically significant independent variables in our model, which means we can still make predictions about the relationships between the variables.

F-statistic: The F-statistic equals 149.3 producing a p-value of <0.001, which is highly significant. Therefore, our independent predictor variables statistically significantly predict the outcome, log base 10 viral load. The regression model seems to be a good fit for our cohort study sample data.

Confidence interval: Looking at the confidence interval, we can say that there is approximately a 95% chance that the interval [9.52, 10.07] will contain the true value of intercept b0, which is narrow, making it highly precise.

P value: The Statistical Hypotheses are as follows:

H_0: $\beta 1 = 0$ (the coefficients are equal to zero i.e., no relationship between x and y)

H_A: $\beta 1 \neq 0$ (the coefficients are not equal to zero (i.e., there is a relationship between x and y)

In our model, for all predictor values, the p value is less than the level of significance (p<0.05), which indicates that we have at least one coefficient in our model that isn't zero, additionally, all our predictor variables are statistically significant. The cohort study sample data provides us with enough statistical evidence to reject the null hypothesis. There is therefore strong evidence that a relationship does exist between our predictor variables and our outcome, log base 10 viral load. Changes in the predictor variables are associated with changes in outcome.

Interpretation of coefficients: For *dummy variables:* Sex is coded as Male=1,Female=0 (reference) *i.e. b0 + b1 if person is male; bo if person is female*, AZT (zidovudine therapy) Failure=0 (reference),Intolerance=1; *i.e. b0 + b1 if a person has AZT intolerance; bo if a person has AZT failure* and Drug (antiretroviral drugs) coded as didanosine (ddI)=1,zalcitabine (ddC)=0 (reference) *i.e. b0 + b1 if a person is taking ddi; bo if a person is taking ddc.*

From the coefficient estimates in the table, the multiple regression model is:

$$\hat{Y} = 9.80 + 0.02(\text{Age}) - 0.14(\text{Sex}) - 0.14(\text{Drug}) - 1.49\ (\text{AZT})$$

This tells us that, holding all other predictor variables constant, for every unit increase in Age, there is an associated 0.02 increase in log base 10 viral load, for every Male participant there is a 0.14 decrease in log base 10 viral load and no decrease or increase in log base 10 viral load for a Female participant. For every participant using ddi as a drug there is a 0.14 decrease in log base 10 viral load and no decrease for participants using ddc as a Drug. For every participant intolerant to AZT there is a 1.49 decrease in log base 10 viral load.

Age (coeff: 0.02) The average age is statistically significant at the 0.05 level (p<0.001). There seems to be a relationship between age and log base 10 viral load. The positive coefficient indicates that a unit increase in age is associated with an increase of 0.02 in the estimated log base 10 viral load, keeping all other covariates constant.

Sex (coeff: -0.14) The average sex for males is statistically significant at the 0.05 level (p=0.007). The negative coefficient indicates that on average, being a male participant is associated with a log base 10 viral load that is 0.14 lower than those who are female participants. Every male participant is associated with a 0.14 unit decrease in log base 10 viral load as compared to female participants, holding age, drug and AZT constant.

Drug (coeff: -0.14) The average drug ddi is statistically significant at the 0.05 level (p=0.008). The negative coefficient indicates that on average, taking ddi as a drug is associated with a log base 10 viral load that is 0.14 lower than those who were taking ddc as a drug. Participant who used ddi as a drug, were associated with a unit decrease of 0.14 log base 10 viral load, as compared to participants who used ddc as a drug of use, holding age, sex and AZT constant.

AZT (coeff: -1.49) The average AZT intolerance is statistically significant at the 0.05 level (p<0.001). The negative coefficient indicates that on average, having AZT-intolerance is associated with a viral load that is 1.49 lower than those who had AZT failure. Every participant who shows an intolerance to AZT is associated with a 1.49 unit decrease in log base 10 viral load, as compared to participants who showed a failure to AZT, holding all other covariates constant.

Conclusion: In our cohort study of individuals living with HIV, we found significant relationships between age, sex, drug and AZT and the outcome, log base 10 viral load (p < 0.001 for each).

Question 4 [4 marks]

a) What is the predicted value of log base 10 viral load for an individual using ddC, with intolerance AZT, who is a female at median age? Interpret your findings briefly. [4]

From our fitted/predicted formula, we deduce:

log10_vload = 9.80 + 0.02 (Age) + (-0.14 Sex-Female) + (-0.14 Drug-ddc) + (-1.49 AZT-intolerance)

log10_vload = 9.80 + 0.02 (31) -0.14 (0) -0.14 (0) -1.49 (1) =8.93

On average, the predicted log base 10 viral load of a 31 year old female with intolerance to AZT, using ddC as a drug is 8.93. The association is statistically significant (p-value <0.05)

APPENDIX 1

R SCRIPT- BIOSTAT II ASSIGNMENT 1

```
install.packages("dplyr")
install.packages("ggplot2")
install.packages("skimr")
install.packages("readr")
install.packages("lattice")
install.packages("tidyr")
install.packages("here")
install.packages("tidyverse")
install.packages("ggubr")
install.packages("describer")
install.packages("psych")
install.packages("gtsummary")
install.packages("janitor")
install.packages("stringr")
install.packages("ggpubr")
install.packages("gmodels")
install.packages("summarytools")
install.packages("epiDisplay")
install.packages("epiR")
install.packages("survival")
install.packages("tab")
install.packages("gtools")
install.packages("gtable")
install.packages("stringi")
install.packages("MASS")
install.packages("glue")
install.packages("broom")
install.packages("palmerpenguins")
install.packages("Hmisc")
install.packages("car")
install.packages("caret")
install.packages("ROCR")
install.packages("labelled")
library("here")
library("dplyr")
library("skimr")
library("ggplot2")
library("readr")
library("lattice")
library("tidyr")
library("tidyverse")
library("ggubr")
library("describe")
library("psych")
library("gtsummary")
library("janitor")
library("stringr")
library("gmodels")
library("summarytools")
library("epiDisplay")
library("epiR")
library("survival")
library("tab")
library("gtools")
```

```
library("gtable")
library("stringi")
library("MASS")
library("glue")
library("broom")
library("palmerpenguins")
library("Hmisc")
library("car")
library("caret")
library("ROCR")
library("labelled")
##-----------------------------------------------------------------------------------------------
##Question 1A:

##Summary statistics
View(BiostatII_2022_Assign1_data.46)
str(BiostatII_2022_Assign1_data.46)
head(BiostatII_2022_Assign1_data.46)
names(BiostatII_2022_Assign1_data.46)
summary(BiostatII_2022_Assign1_data.46)
skim(BiostatII_2022_Assign1_data.46)
summary(BiostatII_2022_Assign1_data.46$age)
summary(BiostatII_2022_Assign1_data.46$v_load)
describe(BiostatII_2022_Assign1_data.46)
##Check Variable characters:
is.factor(BiostatII_2022_Assign1_data.46$educated)
as.factor(BiostatII_2022_Assign1_data.46$educated)
BiostatII_2022_Assign1_data.46$educated<-as.factor(BiostatII_2022_Assign1_data.46$educated)
is.factor(BiostatII_2022_Assign1_data.46$educated)
class(BiostatII_2022_Assign1_data.46$educated)
is.factor(BiostatII_2022_Assign1_data.46$sex)
as.factor(BiostatII_2022_Assign1_data.46$sex)
BiostatII_2022_Assign1_data.46$sex<-as.factor(BiostatII_2022_Assign1_data.46$sex)
is.factor(BiostatII_2022_Assign1_data.46$sex)
class(BiostatII_2022_Assign1_data.46$sex)
is.factor(BiostatII_2022_Assign1_data.46$drug)
as.factor(BiostatII_2022_Assign1_data.46$drug)
BiostatII_2022_Assign1_data.46$drug<-as.factor(BiostatII_2022_Assign1_data.46$drug)
is.factor(BiostatII_2022_Assign1_data.46$drug)
class(BiostatII_2022_Assign1_data.46$drug)
is.factor(BiostatII_2022_Assign1_data.46$AZT)
as.factor(BiostatII_2022_Assign1_data.46$AZT)
BiostatII_2022_Assign1_data.46$AZT<-as.factor(BiostatII_2022_Assign1_data.46$AZT)
is.factor(BiostatII_2022_Assign1_data.46$AZT)
class(BiostatII_2022_Assign1_data.46$AZT)
is.factor(BiostatII_2022_Assign1_data.46$v_load)
is.numeric(BiostatII_2022_Assign1_data.46$v_load)
class(BiostatII_2022_Assign1_data.46$v_load)
is.factor(BiostatII_2022_Assign1_data.46$age)
is.numeric(BiostatII_2022_Assign1_data.46$age)
class(BiostatII_2022_Assign1_data.46$age)
is.factor(BiostatII_2022_Assign1_data.46$pid)
is.numeric(BiostatII_2022_Assign1_data.46$pid)
BiostatII_2022_Assign1_data.46$pid <- as.character(BiostatII_2022_Assign1_data.46$pid)
is.character(BiostatII_2022_Assign1_data.46$pid)
class(BiostatII_2022_Assign1_data.46$pid)
is.factor(BiostatII_2022_Assign1_data.46$X)
is.numeric(BiostatII_2022_Assign1_data.46$X)
```

```
class(BiostatII_2022_Assign1_data.46$X)
##-----------------------------------------------------------------------------------------------
##Summary statistics table:
tbl_summary(BiostatII_2022_Assign1_data.46,by = AZT, statistic = list(c(1,7,8) ~ "{mean} ({sd})",c(2,3) ~ "{median}
({p25}, {p75})",all_categorical() ~ "{n} ({p}%)"),digits = all_continuous() ~ 2,missing = "no") %>%
  add_overall() %>%
  modify_caption("Characteristics of the sample (overall and by AZT)",) %>%
  bold_labels()
##-----------------------------------------------------------------------------------------------
##Check for Normality

##Age
class(BiostatII_2022_Assign1_data.46$age)
hist(BiostatII_2022_Assign1_data.46$age,main= "", xlab="Age (years)")
summary(BiostatII_2022_Assign1_data.46$age)
shapiro.test(BiostatII_2022_Assign1_data.46$age)
qqnorm(BiostatII_2022_Assign1_data.46$age, ylab = "Age (years) ", col = "dark green", pch = 1, frame = FALSE)
qqline(BiostatII_2022_Assign1_data.46$age, col = "steelblue", lwd = 2)
qqPlot(BiostatII_2022_Assign1_data.46$age)
boxplot(BiostatII_2022_Assign1_data.46$age,xlab = "Age (years)")

##Viral load
hist(BiostatII_2022_Assign1_data.46$v_load,main = "",xlab = "log10_vload")
summary(BiostatII_2022_Assign1_data.46$v_load)
shapiro.test(BiostatII_2022_Assign1_data.46$v_load)
qqnorm(BiostatII_2022_Assign1_data.46$v_load, ylab = "log10_vload ", col = "dark green", pch = 1, frame = FALSE)
qqline(BiostatII_2022_Assign1_data.46$v_load, col = "steelblue", lwd = 2)
qqPlot(BiostatII_2022_Assign1_data.46$v_load)
boxplot(BiostatII_2022_Assign1_data.46$v_load,xlab = "log10_vload")

##Summary statistics continued
View(BiostatII_2022_Assign1_data.46)
str(BiostatII_2022_Assign1_data.46)
summary(BiostatII_2022_Assign1_data.46)
table(BiostatII_2022_Assign1_data.46$sex)
tabyl(BiostatII_2022_Assign1_data.46$sex)
summary(BiostatII_2022_Assign1_data.46)
nrow(BiostatII_2022_Assign1_data.46)
skim(BiostatII_2022_Assign1_data.46)
quantile(BiostatII_2022_Assign1_data.46$age)
quantile(BiostatII_2022_Assign1_data.46$v_load)
mean(BiostatII_2022_Assign1_data.46$v_load)
mean(BiostatII_2022_Assign1_data.46$age)
sd(BiostatII_2022_Assign1_data.46$v_load)
sd(BiostatII_2022_Assign1_data.46$age)
var(BiostatII_2022_Assign1_data.46)
var(BiostatII_2022_Assign1_data.46$v_load)
var(BiostatII_2022_Assign1_data.46$age)
cor(BiostatII_2022_Assign1_data.46$v_load,BiostatII_2022_Assign1_data.46$age)
sd(BiostatII_2022_Assign1_data.46$v_load) / mean(BiostatII_2022_Assign1_data.46$v_load)
sd(BiostatII_2022_Assign1_data.46$age) / mean(BiostatII_2022_Assign1_data.46$age)
median(BiostatII_2022_Assign1_data.46$v_load)
median(BiostatII_2022_Assign1_data.46$age)
IQR(BiostatII_2022_Assign1_data.46$v_load)
IQR(BiostatII_2022_Assign1_data.46$age)
by(BiostatII_2022_Assign1_data.46, BiostatII_2022_Assign1_data.46$sex, summary)
fivenum(BiostatII_2022_Assign1_data.46$v_load)
fivenum(BiostatII_2022_Assign1_data.46$age)
```

```
table(BiostatII_2022_Assign1_data.46$educated)
prop.table(table(BiostatII_2022_Assign1_data.46$educated))
100*prop.table(table(BiostatII_2022_Assign1_data.46$educated))
ggplot(data = BiostatII_2022_Assign1_data.46, aes(educated, age)) +
  geom_boxplot()+theme_light()
table(BiostatII_2022_Assign1_data.46$sex)
prop.table(table(BiostatII_2022_Assign1_data.46$sex))
100*prop.table(table(BiostatII_2022_Assign1_data.46$sex))
ggplot(data = BiostatII_2022_Assign1_data.46, aes(sex, age)) +
  geom_boxplot()+theme_light()
table(BiostatII_2022_Assign1_data.46$drug)
prop.table(table(BiostatII_2022_Assign1_data.46$drug))
100*prop.table(table(BiostatII_2022_Assign1_data.46$drug))
ggplot(data = BiostatII_2022_Assign1_data.46, aes(drug, age)) +
  geom_boxplot()+theme_light()
table(BiostatII_2022_Assign1_data.46$AZT)
prop.table(table(BiostatII_2022_Assign1_data.46$AZT))
100*prop.table(table(BiostatII_2022_Assign1_data.46$AZT))
ggplot(data = BiostatII_2022_Assign1_data.46, aes(AZT, age)) +
  geom_boxplot()+theme_light()
table(BiostatII_2022_Assign1_data.46$AZT)
prop.table(table(BiostatII_2022_Assign1_data.46$AZT))
100*prop.table(table(BiostatII_2022_Assign1_data.46$AZT))
ggplot(data = BiostatII_2022_Assign1_data.46, aes(AZT,v_load)) +
  geom_boxplot()+theme_light()
table(BiostatII_2022_Assign1_data.46$age,BiostatII_2022_Assign1_data.46$educated)
table(BiostatII_2022_Assign1_data.46$sex,BiostatII_2022_Assign1_data.46$educated)
table(BiostatII_2022_Assign1_data.46$age,BiostatII_2022_Assign1_data.46$sex)
table(BiostatII_2022_Assign1_data.46$age,BiostatII_2022_Assign1_data.46$AZT)
table(BiostatII_2022_Assign1_data.46$sex,BiostatII_2022_Assign1_data.46$AZT)
##-------------------------------------------------------------------------------------------------------------------------------------------------------
-
##Question 1b)

##MULTIPLE REGRESSION
lm(v_load ~ age + sex + drug + AZT , data = BiostatII_2022_Assign1_data.46)
fit <- lm(v_load ~ age + sex + drug + AZT, data = BiostatII_2022_Assign1_data.46)
summary(fit)
summary(fit)$coefficients
coef(summary(fit))
plot(fit)
coef(fit)
fit$coef
confint(fit)
confint(fit, conf.level=0.95)
tidy(fit)
glance(fit)
##OR
glm(v_load ~ age + drug + AZT + sex,data = BiostatII_2022_Assign1_data.46, family = gaussian(link = "identity" ))
fit_1 <- glm(v_load ~ age + drug + AZT + sex,data = BiostatII_2022_Assign1_data.46, family = gaussian(link = "identity"
))
summary(fit_1)
summary(fit_1)$coefficients
coef(summary(fit_1))
plot(fit_1)
coef(fit_1)
fit_1$coef
confint(fit_1)
```

```
confint(fit_1, conf.level=0.95)
tidy(fit_1)
glance(fit_1)

###SIMPLE LINEAR REGRESSION
##Age variable
mod1_age = lm(v_load ~ age, data = BiostatII_2022_Assign1_data.46)
summary(mod1_age)
summary(mod1_age)$coefficients
##OR
lm(v_load ~ age, data = BiostatII_2022_Assign1_data.46)
mod1_age <- lm(v_load ~ age, data = BiostatII_2022_Assign1_data.46)
summary(mod1_age)
summary(mod1_age)$coefficients

##SEX Variable
mymodel<-glm(v_load ~ sex,
        data = BiostatII_2022_Assign1_data.46,
        family = 'binomial')
summary(mymodel)
summary(mymodel)$coefficients

##Drug variable
model2 <- lm(v_load ~ drug,data = BiostatII_2022_Assign1_data.46)
summary(model2)
summary(model2)$coefficients

##AZT Variable
model3 <- lm(v_load ~ AZT,data = BiostatII_2022_Assign1_data.46)
summary(model3)
summary(model3)$coefficients

##ANOVA
anova(mod1_age)
anova(fit_1)
anova(mod1_age,fit)
anova(mod1_age,fit_1)

##TABULATING MLR STATISTICS
tbl_regression(fit, intercept = TRUE) %>% bold_labels()

##Residual Diagnostics
##CHECK LINEARITY ASSUMPTIONS

##1.Normality (QQ-plot)
plot(fit)
plot(fit_1)
par(mfrow=c(2,2))

##QQ plot-Normality met
##plot does not deviate significantly from line, normality  met
##There is a pattern in the residual plot. This suggests that we can assume linear relationship between the predictors and the
outcome variables.

##2.Multi-Collinearity
vif(fit)
vif(fit_1)
##Less than 5, independent variables are not highly correlated
```

```
##3.Outliers,leverage,influence)
plot(cooks.distance(fit_1), pch = 16, col = "blue")
##4.HeterosCedasticity (Constant variance assumption)
##5.Independence

##PREDICTED VALUES
head(augment(fit_1))
head(augment(fit))
head(augment(fit_1),n=20)
head(augment(fit),n=20)
tail(augment(fit_1))
tail(augment(fit))

contrasts(BiostatII_2022_Assign1_data.46$sex)
##female=0,male=1
contrasts(BiostatII_2022_Assign1_data.46$drug)
##ddC=0,ddI=1
contrasts(BiostatII_2022_Assign1_data.46$AZT)
##failure=0,intolerance=1
##-------------------------------------------------------------------------------------------------
##Question 2b:
##Fig1 v_load vs Age
lm(BiostatII_2022_Assign1_data.46$age~ BiostatII_2022_Assign1_data.46$v_load)
abline(23.6539, 0.8009)
abline(lm(BiostatII_2022_Assign1_data.46$age ~ BiostatII_2022_Assign1_data.46$v_load),col = "blue")
plot(BiostatII_2022_Assign1_data.46$age,BiostatII_2022_Assign1_data.46$v_load,main = "v load vs age",xlab ="Age
(years)",ylab = "log10 viral load (copies/mL)",pch = 16,frame = FALSE,col =
"firebrick2",abline(lm(BiostatII_2022_Assign1_data.46$v_load ~ BiostatII_2022_Assign1_data.46$age),col = "blue"))
##OR
plot(v_load ~ age, data = BiostatII_2022_Assign1_data.46)
abline(lm(BiostatII_2022_Assign1_data.46$v_load ~ BiostatII_2022_Assign1_data.46$age),col = "red")
##OR
scatter.smooth(BiostatII_2022_Assign1_data.46$age,BiostatII_2022_Assign1_data.46$v_load,main="v load vs age",xlab
="Age (years)", ylab = "log10 viral load" )
##OR
ggplot(BiostatII_2022_Assign1_data.46, aes(x = age, y = v_load)) + geom_point() + stat_smooth() + theme_minimal()
ggplot(BiostatII_2022_Assign1_data.46, aes(age, v_load)) + geom_point() + stat_smooth(method = lm)
##
cor (BiostatII_2022_Assign1_data.46$v_load,BiostatII_2022_Assign1_data.46$age,method = "pearson")
cor(BiostatII_2022_Assign1_data.46$v_load,BiostatII_2022_Assign1_data.46$age, use = "complete.obs")
cor (BiostatII_2022_Assign1_data.46$age, BiostatII_2022_Assign1_data.46$v_load)
##-------------------------------------------------------------------------------------------------
##Fig2 v_load vs AZT
boxplot(BiostatII_2022_Assign1_data.46$v_load ~ BiostatII_2022_Assign1_data.46$AZT, col=rainbow(4),xlab =
"AZT",ylab = "log10 viral load (copies/mL)")
##-------------------------------------------------------------------------------------------------
```