PART I: Modeling of mental health problems in adolescents with HIV positive parents. (45 pts)

For this part you will build a statistical model to understand what factors affect the mental health of adolescents with HIV positive parents. You will use the `BSI_overall` variable as your measure of poor mental health, where the higher the number the more severe the mental health problems.

The Brief Symptom Inventory (BSI) is a multidimensional symptom inventory designed to reflect psychological symptom patterns of psychiatric and medical patients. The symptom items are rated on a five-point scale to indicate the degree of distress within the last week. Scores can be calculated for symptom dimensions such as measures somatization, depression, and anxiety. The `BSI_overall` and subscales are already calculated for you and included in the dataset.

Your response should be written in a report / story format following the outline below.

- Start with a univariate analysis of the response variable. Create an appropriate plot (2 pts) and write a descriptive paragraph that contains appropriate summary statistics (3 pts).
- Explore the bivarate relationship between the response and *several* candidate explanatory variables (except `LIVWITH`). Your goal is to identify a few variables that may be associated with the response variable.
  - Choose these sensibly, there are some that will not make sense to include in a model.
  - At minimum you must explore 1 binary categorical, 1 categorical variable with more than 2 levels and 1 quantitative numeric variable. You may explore more if desired.
  - For **two** of the three (9 pts ea) explanatory variables chosen you must:
    - Clearly define your explanatory variable definition at the start of each section. (2 pts) Describe any data cleaning steps you took for each explanatory variable. If you didn't take any, you need to report what you did to decide that you didn't need to change anything. (2 pts)
    - Describe the relationship between your explanatory and response variable using a plot (2 pts) and a descriptive paragraph that contains appropriate summary statistics (3 pts)
- Create an appropriate regression model with at least the three predictors chosen above. (3 pts)
  - Check the model assumptions. Does the model fit well? Why or why not? (4 pts)
- Create a nicely formatted table to report/display the values that you are going to directly interpret in the next step (5 pts).
  - Interpret the regression coefficient for all predictors chosen using a clear English sentence that includes a point estimate, a CI and a p-value in the conclusion. (5 pts)
- Assess if whether or not living with both parents (versus living with one parent or someone else) moderates the relationship between BSI overall and your chosen *quantitative* explanatory variable.(10 pts)

Part I: Multivariable modeling of mental health.

Selected data for mental health analysis:

Primary response variable: adolescent' mental health problem (denoted BSI_overall)

Quantitative variable: Age started smoking (denoted AGESMOKE)

Binary variable: Gender (Female/Male)

Categorical variable: financial situation of household (denoted FINSIT)

```
1 = Very poor, struggling to survive
2 = Poor, barely paying the bills
3 = Have the necessities
4 = Comfortable
```

1. Univariate Analysis

Response variable 1 - adolescent' mental health problem (denoted BSI_overall)

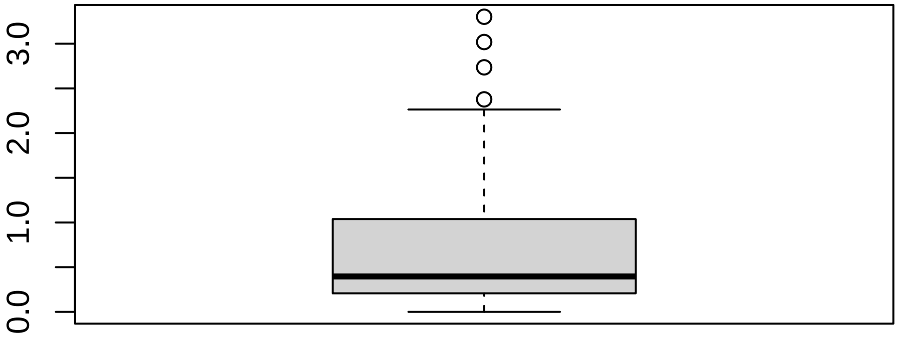
```
summary(hiv$BSI_overall)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
 0.0000  0.2075  0.3962  0.6517  1.0377  3.3019       3 
```

```
sd(hiv$BSI_overall, na.rm = TRUE)
```

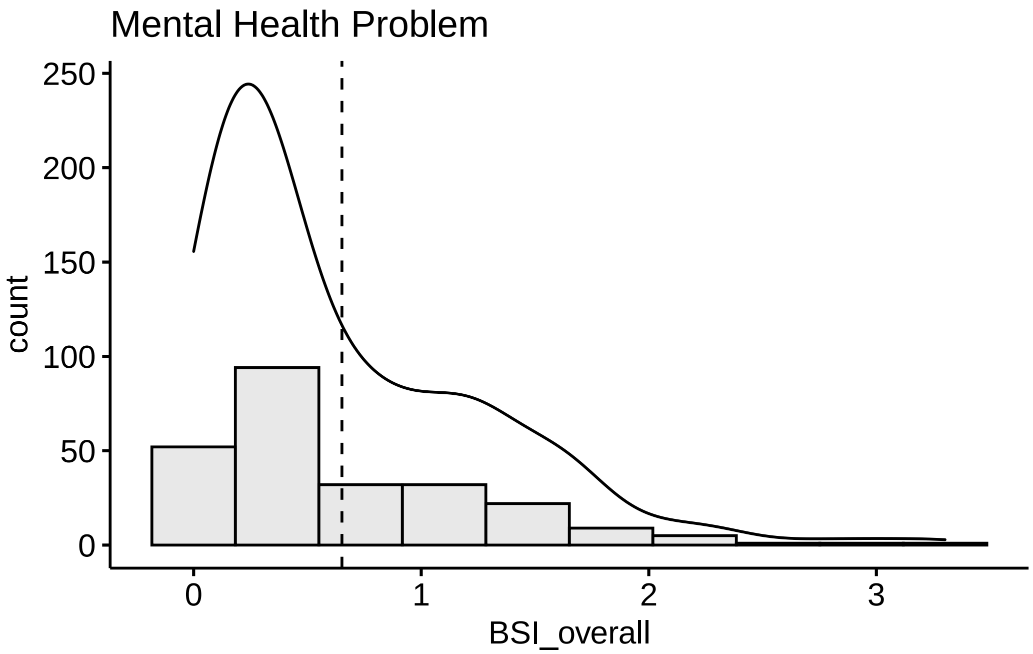
```
[1] 0.6117215
```

```
boxplot(hiv$BSI_overall)
```



The summary statistic and box plot analysis provide that 75% of sample values are below 1.04, and 25% of sample values are smaller than 0.21. The median is 0.40, and the standard deviation with 0.61. The box plot shows the upper potential outlines end of whiskers.

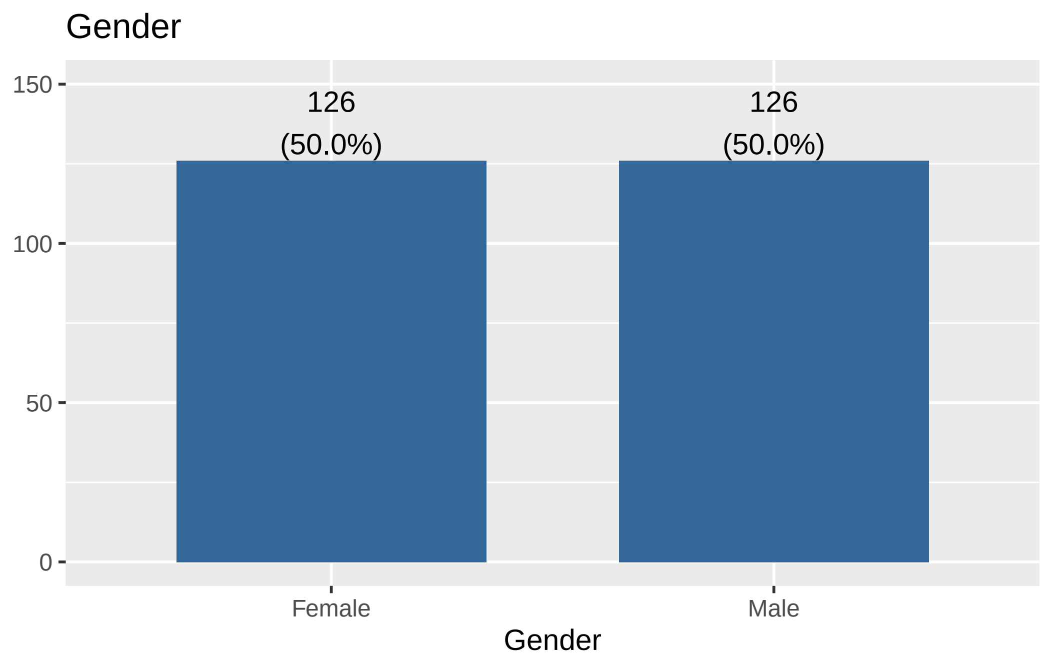
```
gghistogram(hiv, x = "BSI_overall", add = "mean",
  fill = "Lightgray",
  add_density = TRUE, bins = 10,
  xlab = "BSI_overall", title = "Mental Health Problem")
```



Mental health problem (BSI_overall) measured from 0 to 3.302 with a mean 0.65 which is higher than median of 0.40, and a standard deviation of 0.61. It shows a unimodal trend with right skewed. there have potential outlines in both histogram and the box plot.

Binary variable 1 - Gender (Female/ Males)

```
plot_frq(hiv$GENDER, title = "Gender",
axis.title = "Gender") + ylim (0,150)
```



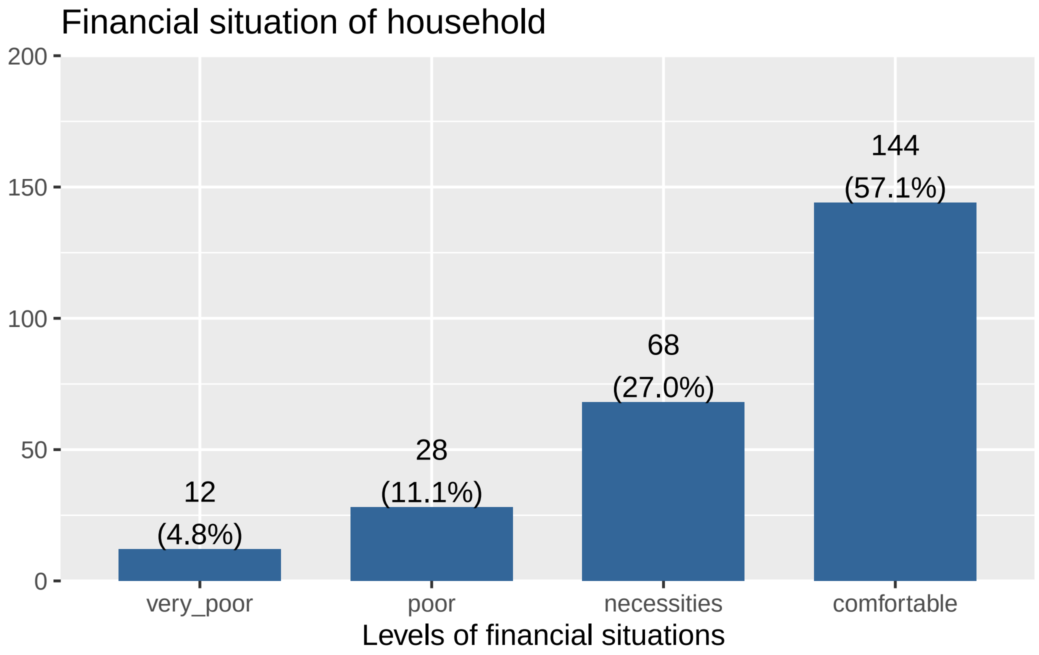
The surveyed adolescents most frequently record that both female and males adolescents are 50% equally attended in the mental health survey (n=252).

Categorical variable 1 - Financial situation of household (FINSIT)

```
1 = Very poor, struggling to survive
2 = Poor, barely paying the bills
3 = Have the necessities
4 = Comfortable
```

```
# relabeling for financial situation of household (FINSIT)
hiv$FINSIT <-factor(hiv$FINSIT,
                    labels = c("very_poor", "poor","necessities", "comfortable"))

plot_frq(hiv$FINSIT, title = "Financial situation of household",
axis.title = "Levels of financial situations ")
```



The financial situation of household participants most frequently report that 57.1% of people live comfortably (n=144), 27% of those need necessities(n=68), 11.1% of poor(n=28) living conditions, and 4.8% of those struggling to survive(n=12).

Quantitative Variable 1 - Age started smoking (AGESMOKE)

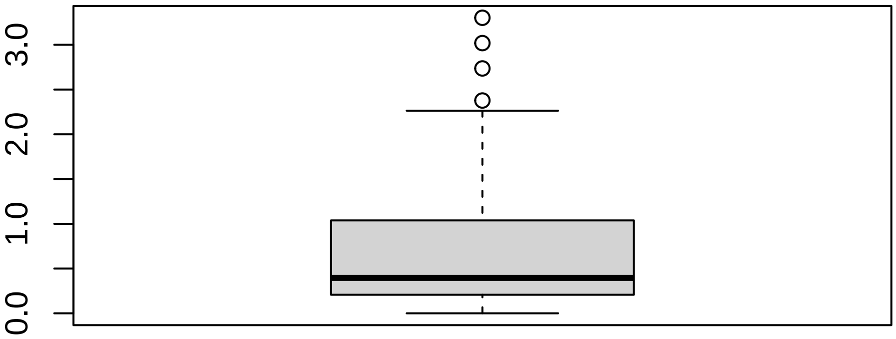
```
summary(hiv$AGESMOKE)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
   4.00   10.00   12.00   11.92   14.00   17.00     120 
```

```
sd(hiv$AGESMOKE, na.rm = TRUE)
```

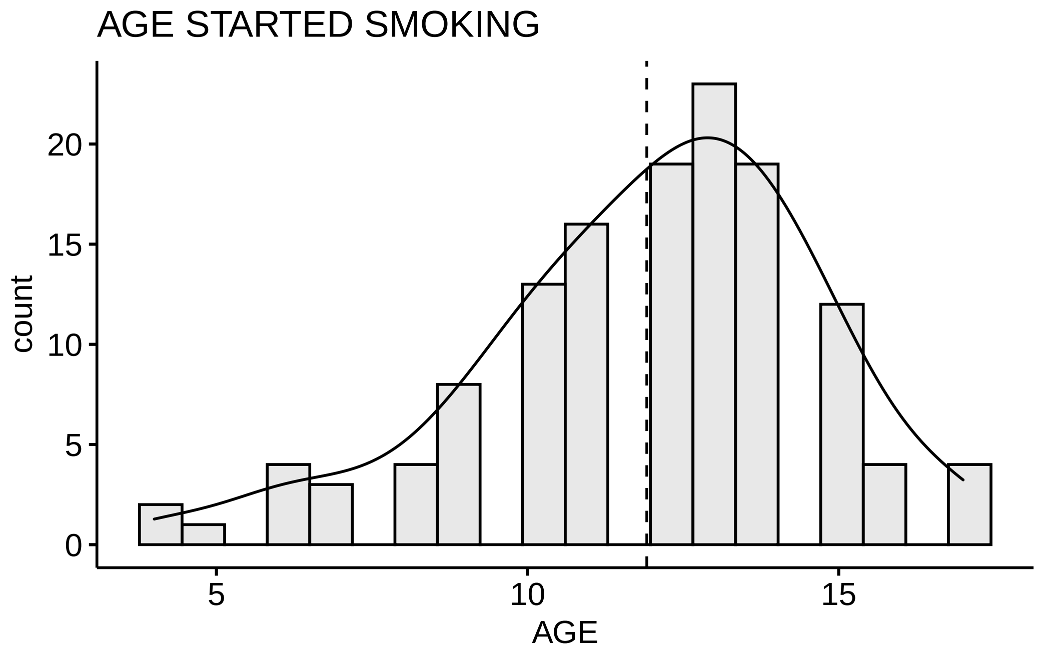
```
[1] 2.721253
```

```
boxplot(hiv$BSI_overall)
```



The summary statistic and box plot analysis shows that 75% of sample values are below 14, and 25% of sample values are smaller than 10. The median is 12, and the standard deviation with 2.72. The box plot shows the upper potential outlines end of whiskers.

```
gghistogram(hiv, x = "AGESMOKE", add = "mean",
  fill = "Lightgray",
  add_density = TRUE, bins = 20,
  xlab = "AGE", title = "AGE STARTED SMOKING")
```



Adolescents started smoking age ranged from 4 to 17 with a mean of 11.92, close to the median of 12, and a standard deviation of 2.72. It shows a unimodal trend with a slightly left skewed. Also, there are outlines in both the histogram and the box plot.

2.Covariante relationship

2.1 Quantitative response and categorical explanatory variable (Q ~ C).

Response variable is mental health problem of adolescents (denoted BSI_overall)

Categorical variable is Financial situation of household, which listed below (denoted FINSIT)

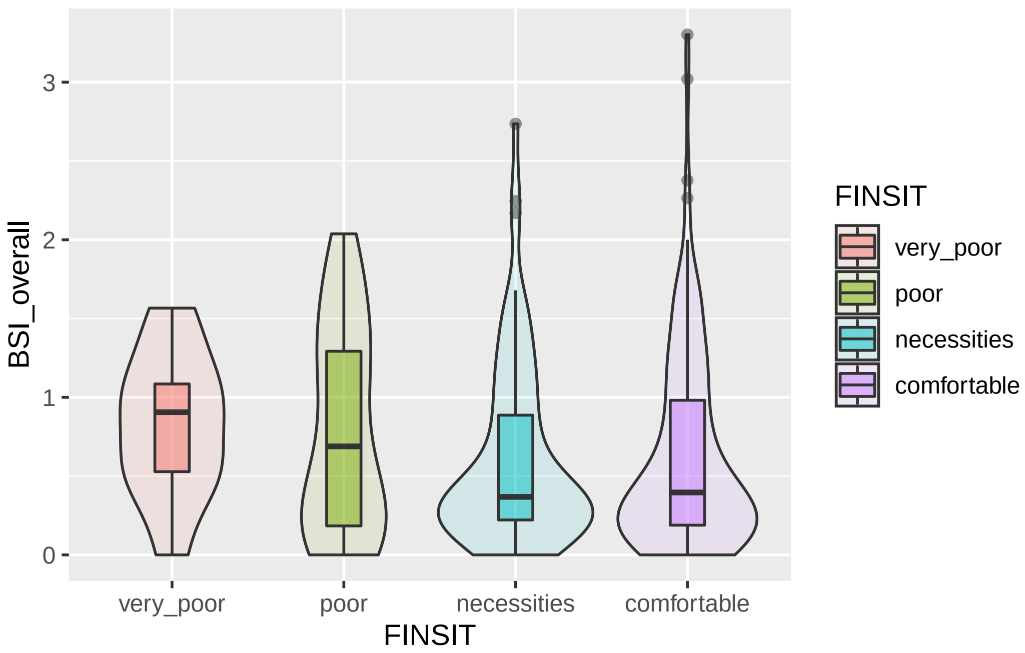
```
1 = Very poor, struggling to survive
2 = Poor, barely paying the bills
3 = Have the necessities
4 = Comfortable
```

The part we will explore a relationship between adolescents' mental health problem and their financial situation of household.

```
hiv %>%
  select(FINSIT, BSI_overall) %>%
  na.omit()%>%
  group_by(FINSIT)%>%
  summarize(mean = mean(BSI_overall),
  sd = sd(BSI_overall),
  IQR = IQR(BSI_overall),
  n = n()) %>% kable(digits = 2)
```

FINSIT	mean	sd	IQR	n
very_poor	0.82	0.46	0.56	11
poor	0.78	0.65	1.11	28
necessities	0.62	0.58	0.67	68
comfortable	0.63	0.63	0.79	142

```
hiv %>%
  select(BSI_overall, FINSIT)%>%
  na.omit()%>%
  ggplot(aes(x=FINSIT, y=BSI_overall, fill=FINSIT)) +
  geom_violin(alpha=.1) +
  geom_boxplot(alpha=.5, width=.2)
```



Summary table & violin graph

On average, adolescents' mental health problem scores were different at each level of household living conditions. The highest score of 0.82 was found in very poor living conditions with a standard deviation of 0.46 and IQR of 0.56. Poor living had also a mean of 0.78 with a standard deviation of 0.65 and IQR of 1.11. While there is slight differentiation between basic needs and comfortable living conditions at 0.01 with standard deviations of 0.58 and 0.63 respectively. There is upper potential outliners in necessities and comfortable groups.

2.2 Quantitative response and categorical explanatory variable (Q ~ B).

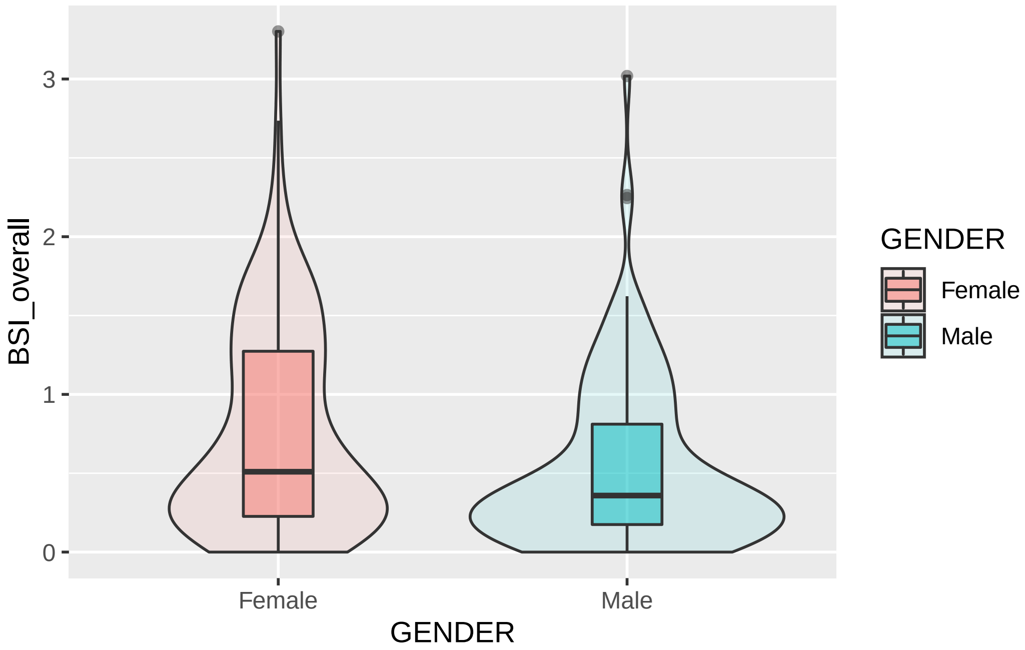
Response variable is mental health problem of adolescents (denoted BSI_overall)

Binary variable is Gender (female/Male)

```
hiv %>%
  select(GENDER, BSI_overall) %>%
  na.omit()%>%
  group_by(GENDER)%>%
  summarize(mean = mean(BSI_overall),
  sd = sd(BSI_overall),
  IQR = IQR(BSI_overall),
    n = n()) %>% kable(digits = 2)
```

GENDER	mean	sd	IQR	n
Female	0.78	0.66	1.05	123
Male	0.53	0.53	0.64	126

```
hiv %>%
  select(BSI_overall, GENDER)%>%
  na.omit()%>%
  ggplot(aes(x=GENDER, y=BSI_overall, fill=GENDER)) +
  geom_violin(alpha=.1) +
  geom_boxplot(alpha=.5, width=.2)
```



The violin boxplot indicates a comparison of mental health problem scores between females and males among adolescents. Female adolescents' average mental health score is 0.78 with a standard deviation of 0.66 and IQR of 1.05. While Male adolescents' average mental health score is 0.53 with a standard deviation of 0.53 and IQR of 0.64. Also, there are upper outliners, and both gender females and males skewed right.

2.3 Quantitative response and quantitative explanatory variable. (Q ~ Q)

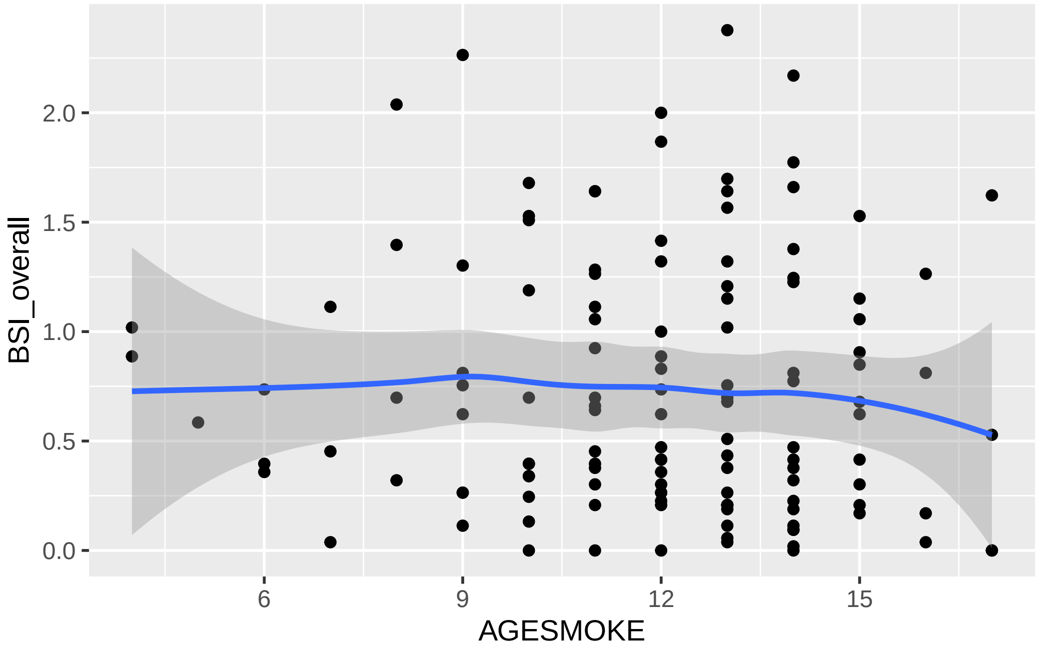
Response variable is mental health problem of adolescents (denoted BSI_overall)

Explanatory variable - Age started smoking (denoted AGESMOKE)

```
cor(hiv$BSI_overall, hiv$AGESMOKE, use = "pairwise.complete.obs")
```

```
[1] -0.05688903
```

```
hiv %>%
  select(BSI_overall, AGESMOKE)%>%
  na.omit()%>%
  ggplot(aes(x=AGESMOKE, y=BSI_overall)) + geom_point()+ geom_smooth()
```



There is a weak negative correlation between adolescents' mental health scores and adolescents' age started smoking. The correlation coefficient is r = -0.005, and no appear linear correlation between them. Thus we could not use this explanatory variable of smoking into analysis.

4.Covariante Analysis

4.1 t-Test analysis

The t-Test was used to see if adolescents' average mental health problem score was the difference between both gender females and males.

Identify response and explanatory variables

The response variable is adolescents' mental health problem score, which numerical variable. (denoted BSI_overall). The explanatory variable is gender (female/male).

Research hypothesises

Null Hypothesis: there is no difference in average mental health problem scores between female and male adolescents.

Alternate Hypothesis: There is a difference in average mental health problem scores between female and male adolescents.

Assumption for t-Test

1.Mutually exclusive and independent. It is valid because a person could not be female and Male.

2.Differences are normally distributed. it is valid both groups are normally distributed, thus differences are normally distributed too.

3.Variances are similar for both groups. It is valid because the standard deviation is close to both gender females and males (0.66 vs 0.53).

Let μ_1 be the mean of female adolescents' mental health problem

Let μ_2 be the mean of male adolescents' mental health problem

$H_0 : \mu_1 - \mu_2 = 0$
$H_A : \mu_1 - \mu_2 \neq 0$

```
t.test(hiv$BSI_overall ~ hiv$GENDER)
```

```
    Welch Two Sample t-test

data:  hiv$BSI_overall by hiv$GENDER
t = 3.2293, df = 233.2, p-value = 0.001419
alternative hypothesis: true difference in means between group Female and group Male is not
95 percent confidence interval:
 0.09607302 0.39672094
sample estimates:
mean in group Female   mean in group Male
           0.7763461            0.5299491
```

As a result of the t-test, female adolescents have on average 0.25 (95% CI 0.10, 0.40) higher mental health problem scores compared to male adolescents. This is a significant difference (p=0.0014).

4.2 ANOVA Test analysis

ANOVA test is used to see if adolescents' mental health problem scores change across levels of their financial situation of household.

Identify response and explanatory variables

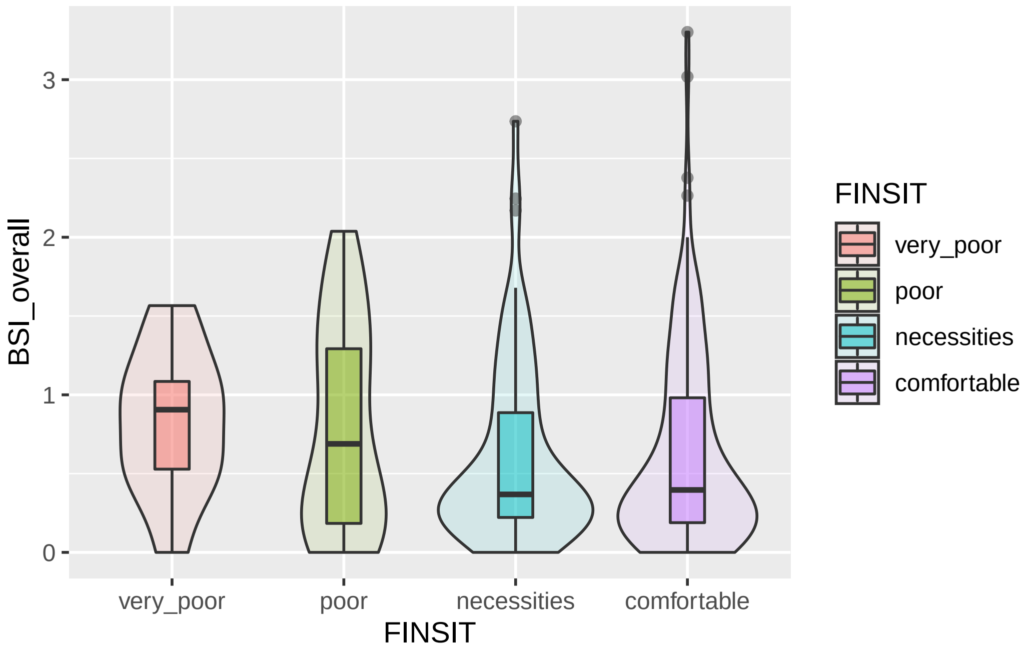
The response variable is adolescents' mental health problem score, which numerical variable. (denoted BSI_overall).

Categorical variable is Financial situation of household, which listed below (denoted FINSIT)

```
        1 = Very poor, struggling to survive
        2 = Poor, barely paying the bills
        3 = Have the necessities
        4 = Comfortable
```

Visualize and summarise bivariate relationship

```
hiv %>%
  select(BSI_overall, FINSIT)%>%
  na.omit()%>%
  ggplot(aes(x=FINSIT, y=BSI_overall, fill=FINSIT)) +
  geom_violin(alpha=.1) +
  geom_boxplot(alpha=.5, width=.2)
```



The distribution of average adolescents' mental health problem scores varies across their levels of the financial situation of the household. The highest score of 0.82 was found in very poor living conditions with a standard deviation of 0.46 and IQR of 0.56. Poor living had also a mean of 0.78 with a standard deviation of 0.65 and IQR of 1.11. While there is slight differentiation between basic needs and comfortable living conditions at 0.01 with standard deviations of 0.58 and 0.63 respectively.

Research hypothesises

Null Hypothesis: There is no association between adolescents' mental health problem scores and the financial situation of the household.

Alternate Hypothesis: There is an association between adolescents' mental health problem scores and the financial situation of the household.

Assumption of ANOVA test

1.Independence: assuming adolescents, who have a different financial situation of households are sampled independently, all groups have a large enough sample size.

2.Normality: the distribution of mental health score within each group are fairly normal

3.Equal variances: standard deviation and IQR are similar across all groups.

Let μ_1 be the true mean adolescents' mental health problem score with very poor group

Let μ_2 be the true mean adolescents' mental health problem score with poor group

Let μ_3 be the true mean adolescents' mental health problem score with have the necessities group

Let μ_4 be the true mean adolescents' mental health problem score with comfortable

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$

H_A : At least one group mean is different.

```
aov(hiv$BSI_overall~hiv$FINSIT) |> summary()
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
hiv$FINSIT    3   0.96  0.3192   0.852  0.467
Residuals   245  91.84  0.3749
3 observations deleted due to missingness
```

The p-value (0.467) is large enough, thus there is no evidence to reject the null hypothesis: average mental problem health score is no difference in levels of the financial situation of the household among adolescents.

4.3 Linear Regression Analysis (Q~Q)

Identify response and explanatory variables

The response variable is adolescents' mental health problem score, which numerical variable. (denoted BSI_overall).

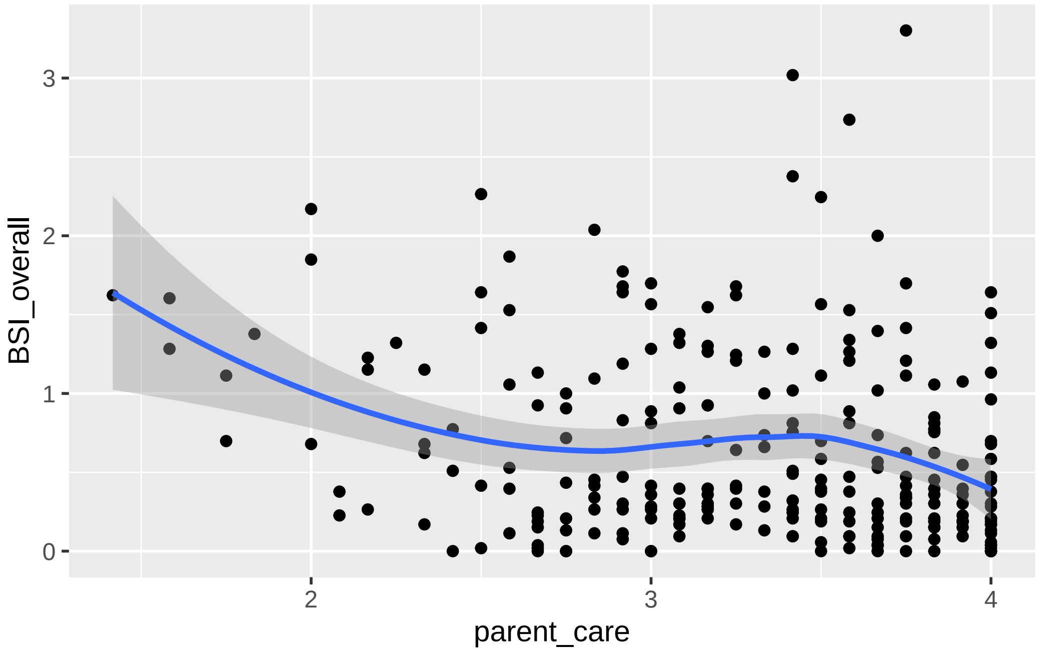
Explanatory variable - Parent care for adolescents (denoted parent_care) numerical variable.

Visualize and summarise bivariate relationship

```
cor(hiv$BSI_overall, hiv$parent_care, use = "pairwise.complete.obs")
```

```
[1] -0.2116093
```

```
hiv %>%
  select(BSI_overall, parent_care)%>%
  na.omit()%>%
  ggplot(aes(x=parent_care, y=BSI_overall)) + geom_point()+ geom_smooth()
```



There is a negligible linear correlation between adolescents' mental health problem score and their parental care (r=-0.21) with a small negative slope.

Research hypothesises

Null Hypothesis: There is no linear relationship between adolescents' mental health problem score and their parental care.

Alternate Hypothesis: there is a linear relationship between adolescents' mental health problem score and their parental care.

Let β_1 be slope parameter which express the change in adolesecnt meantal health problem score as parental care decreases.

b.

$H_0 : \beta_1 = 0$
$H_A : \beta_1 \neq 0$

Assumptions

Adolescents' mental health problem score and their parental care are continuous numerical variables.

```
slm_model<- lm(BSI_overall ~ parent_care, data=hiv)
summary(slm_model)
```

```

Call:
lm(formula = BSI_overall ~ parent_care, data = hiv)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.8408 -0.4099 -0.2123  0.3115  2.7594 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.38136    0.21937   6.297 1.39e-09 ***
parent_care -0.22369    0.06587  -3.396 0.000798 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.597 on 246 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.04478,	Adjusted R-squared:  0.0409 
F-statistic: 11.53 on 1 and 246 DF,  p-value: 0.0007975
```

```
tbl_regression(slm_model, intercept = TRUE) %>%
add_glance_table(include = c(adj.r.squared))
```

Characteristic	**Beta**	**95% CI**	**p-value**
(Intercept)	1.4	0.95, 1.8	<0.001
parent_care	-0.22	-0.35, -0.09	<0.001
Adjusted R^2	0.041		

There is significant (<.0001) evidence to believe that there is a linear relationship between adolescents' mental health problem score and their parental care. Parental care for adolescents explains 4% of the variation in adolescents' mental health problem scores.

```
slm_model |> confint()
```

```
                 2.5 %      97.5 %
(Intercept)  0.9492634  1.81344862
parent_care -0.3534351 -0.09394664
```

Each point increases in parental care for adolescents is associated with a significant decreases of 0.22 (-0.35, -0.09) points in adolescents' mental health problems score (p<.0001).

5.Model building - Multiple Linear Regression with a Categorical Predictor

Identify response and explanatory variables

Outcome: Adolescent's mental health problem score, who have HIV positive parents (denoted BSI_overall), numerical variable.

Predictor: Parental care for adolescent (denoted patent_care), numerical variable.

Predictor: Financial situation of household, categorical variable includes following groups:

```
1 = Very poor, struggling to survive
2 = Poor, barely paying the bills
3 = Have the necessities
4 = Comfortable
```

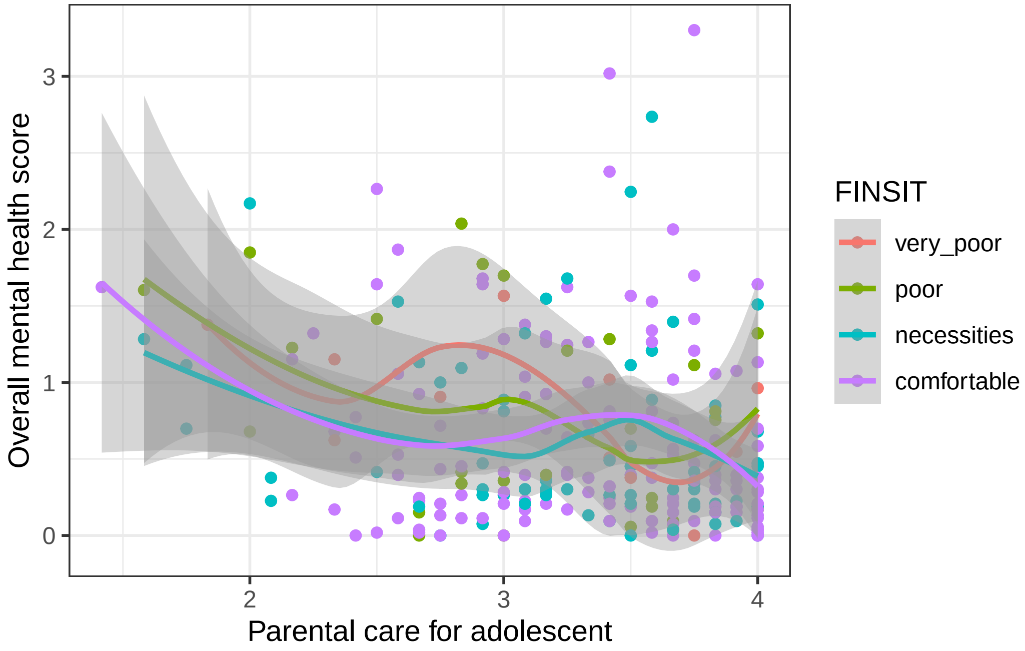
State hypothesis here

Null Hypothesis: There is no relationship between parental care and the adolescent's mental health problem score.

Alternative: there is a relationship between parental care and the adolescent's mental health problem score.

Visualize and describe the relationships

```
hiv%>%
  select(FINSIT, BSI_overall, parent_care) %>%
  na.omit()%>%
  ggplot(aes(x=parent_care, y = BSI_overall, color = FINSIT)) +
  geom_point() + geom_smooth() + theme_bw() +
  labs(x="Parental care for adolescent", y="Overall mental health score")
```



The relationship is negative, weak, and possibly linear correlation between adolescent mental health problem scores and parental care. The trends fluctuate at the end of the scatter plot between adolescents' mental health problem scores and parental care. The adolescents' mental health problem scores with the necessities and comfortable groups appear similar trend. The linear model could fit this model, but there is a skewness. I assume there to be a significant relationship between adolescent mental health problems score and parental care, but I do not expect there to be a difference in adolescent mental health problems score across levels of the financial situation of the household.

Fit simple linear regression model

```
lm.model_1 <- lm(BSI_overall ~ parent_care, data=hiv)
tidy(lm.model_1)
```

```
# A tibble: 2 x 5
  term        estimate std.error statistic       p.value
  <chr>          <dbl>     <dbl>     <dbl>         <dbl>
1 (Intercept)    1.38     0.219       6.30 0.00000000139
2 parent_care   -0.224    0.0659     -3.40 0.000798
```

The estimate of the simple regression model coefficient for parental care has a small negative slope but it is not significant (b1=-0.22, p =.0008). Thus, parental care might be associated with adolescents' mental health problems score.

Multilinear regression models

Let y be adolescents' mental health problem score (BSI_overall)

Let x1 be parental care (parent_care)

Let x2 = 1 when Financial situation of household = "Poor", otherwise 0

Let x3 = 1 when Financial situation of household = "Necessities", otherwise 0

Let x4 = 1 when Financial situation of household = "Comfortable", otherwise 0

The reference group is Very poor living conditions

The mathematical model would look like:

$$Y \sim \beta_0 + \beta_1 * x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

```r
lm_model.2 <- lm(BSI_overall ~ parent_care + FINSIT, data=hiv)
tidy(lm_model.2)
```

```
# A tibble: 5 x 5
  term                estimate std.error statistic     p.value
  <chr>                  <dbl>     <dbl>     <dbl>       <dbl>
1 (Intercept)           1.49      0.275      5.40  0.000000156
2 parent_care          -0.214     0.0668    -3.20  0.00154
3 FINSITpoor           -0.0419    0.213     -0.197 0.844
4 FINSITnecessities    -0.174     0.195     -0.893 0.373
5 FINSITcomfortable    -0.151     0.188     -0.804 0.422
```

```r
tbl_regression(lm_model.2, intercept = TRUE) %>%
   add_glance_table(include = c(adj.r.squared))
```

Characteristic	Beta	95% CI	p-value
(Intercept)	1.5	0.95, 2.0	<0.001
parent_care	-0.21	-0.35, -0.08	0.002
FINSIT			

Characteristic	Beta	95% CI	p-value
very_poor	—	—	
poor	-0.04	-0.46, 0.38	0.8
necessities	-0.17	-0.56, 0.21	0.4
comfortable	-0.15	-0.52, 0.22	0.4
Adjusted R^2	0.035		

Interpret the regression coefficients

b0:The predicted adolescent mental health problem score with no parental care and very poor financial situation is 1.5 (0.95, 2.0, p-value <.0001)

b1: After controlling for the financial situation of the household, for each unit decreases in parental care, adolescents' mental health problem scores increased by -0.21 (-0.35, -0.08, p-value = 0.002)

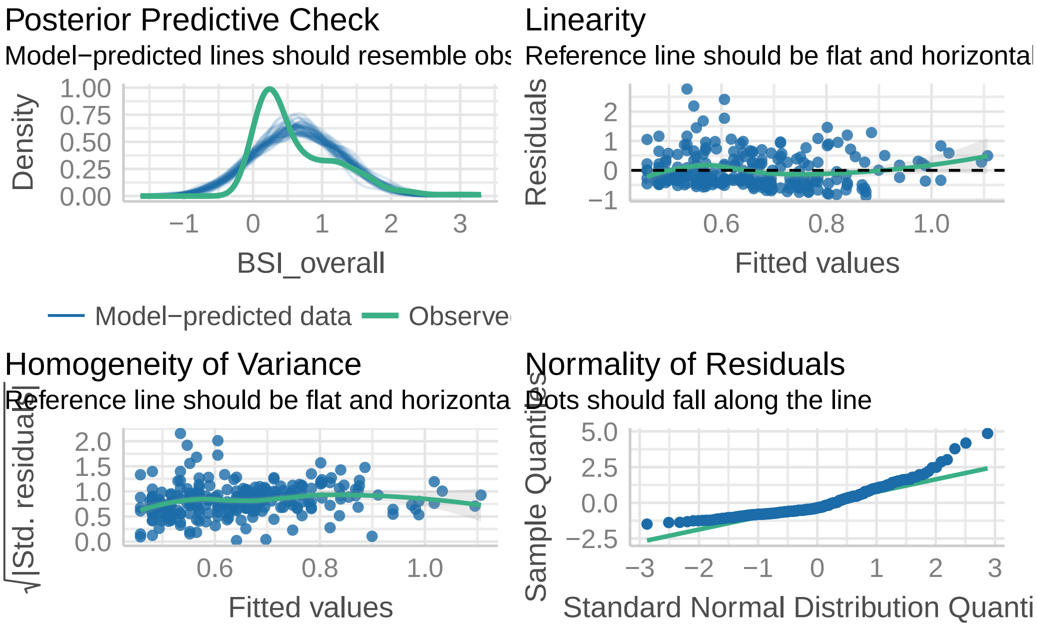
b2: those reported financially poor situations of households have -0.04 (-0.46, 0.38, p-value=0.8) mental health problems score lower compared to the very poor household group, and this is not a significant relationship.

b3: Those reported financial need necessities of households have -0.17(-0.56, 0.21, p-value=0.4) mental health problems score lower compared to the very poor household group, and this is not a significant relationship.

b4: Those reported financially comfortable situations of household have -0.15(-0.52, 0.22, p-value=0.4) mental health problems score lower compared to the very poor household group, and this is not a significant relationship.

Assess model fit

```r
check_model(lm_model.2,
  check = c("qq", "linearity", "homogeneity", "pp_check"))
```



The residual models appear not well fit each model. The residuals are somewhere not constant, with predicted adolescents' mental health scores higher than observed.

Conclusion

```
r2(lm_model.2)
```

```
# R2 for Linear Regression
       R2: 0.051
  adj. R2: 0.035
```

After controlling the financial situation of households, parental care is not significantly (p-value=0.002) associated with adolescents' mental health problem scores. The financial situation of households was not significantly associated with an adolescent mental health problem score because the p-values were large enough. The financial situation of households and parental care combined explain 3.5% in adolescents' mental health scores.

Additional variable “Live with both partents and without”

Outcome: Adolescent’s mental health score, who have HIV positive parents (denoted BSI_overall), numerical variable.

Predictor: Parental care for adolescent (denoted patent_care), numerical variable.

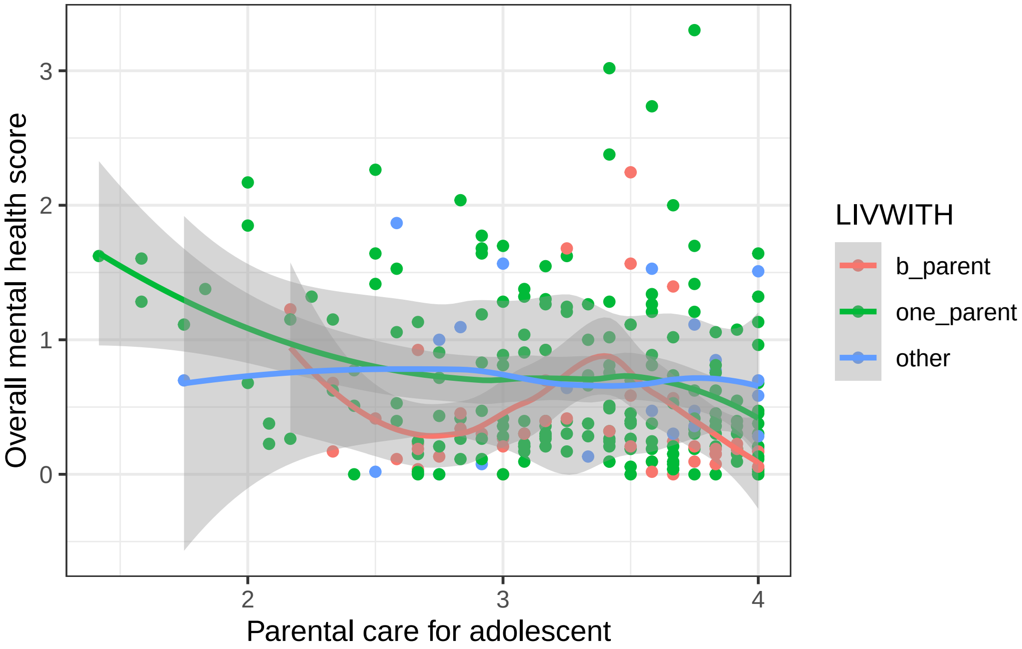
Moderator variable: Currently living with? (denoted LIVWITH)

```
1 = Both parents
2 = One parent
3 = Other
```

Visualize and describe the relationships

```
hiv$LIVWITH <- factor(hiv$LIVWITH,
                      labels = c("b_parent", "one_parent", "other"))

hiv%>%
  select(LIVWITH, BSI_overall, parent_care) %>%
  na.omit()%>%
  ggplot(aes(x=parent_care, y = BSI_overall, color = LIVWITH)) +
  geom_point() + geom_smooth() + theme_bw() +
  labs(x="Parental care for adolescent", y="Overall mental health score")
```



The relationship is negative, weak, and possibly linear correlation between adolescent mental health problem scores and parental care. Trends are moderate variance between mental health problem scores and parental care, but those living with others are barely flat and have no linear correlation with mental health problem scores. There is an upper skewness.

Multilinear regression models

```
Let y be adolescents' mental health problem (BSI_overall)
Let x1 be parental care (parent_care)
Let x2 = 1 when Living with = "One parent", otherwise 0
Let x3 = 1 when Living with = "other", otherwise 0
The reference group is living with both parents
```

The mathematical model would look like:

$$Y \sim \beta_0 + \beta_1 * x_1 + \beta_2 x_2 + \beta_3 x_3$$

```r
lm_model.3 <- lm(BSI_overall ~ parent_care + LIVWITH, data=hiv)
tidy(lm_model.3)
```

```
# A tibble: 4 x 5
  term                estimate std.error statistic    p.value
  <chr>                  <dbl>     <dbl>     <dbl>      <dbl>
1 (Intercept)            1.15     0.232       4.97 0.00000127
2 parent_care           -0.221    0.0652     -3.39 0.000812
3 LIVWITHone_parent      0.265    0.0976      2.72 0.00700
4 LIVWITHother           0.296    0.151       1.96 0.0509
```

```
tbl_regression(lm_model.3, intercept = TRUE) %>%
   add_glance_table(include = c(adj.r.squared))
```

Characteristic	Beta	95% CI	p-value
(Intercept)	1.2	0.70, 1.6	<0.001
parent_care	-0.22	-0.35, -0.09	<0.001
LIVWITH			
b_parent	—	—	
one_parent	0.27	0.07, 0.46	0.007
other	0.30	0.00, 0.59	0.051
Adjusted R^2	0.063		

Interpret the regression coefficients

b0: The predicted adolescent mental health score with no parental care and very poor financial situation is 1.2 (0.70, 1.60, p-value <.0001)

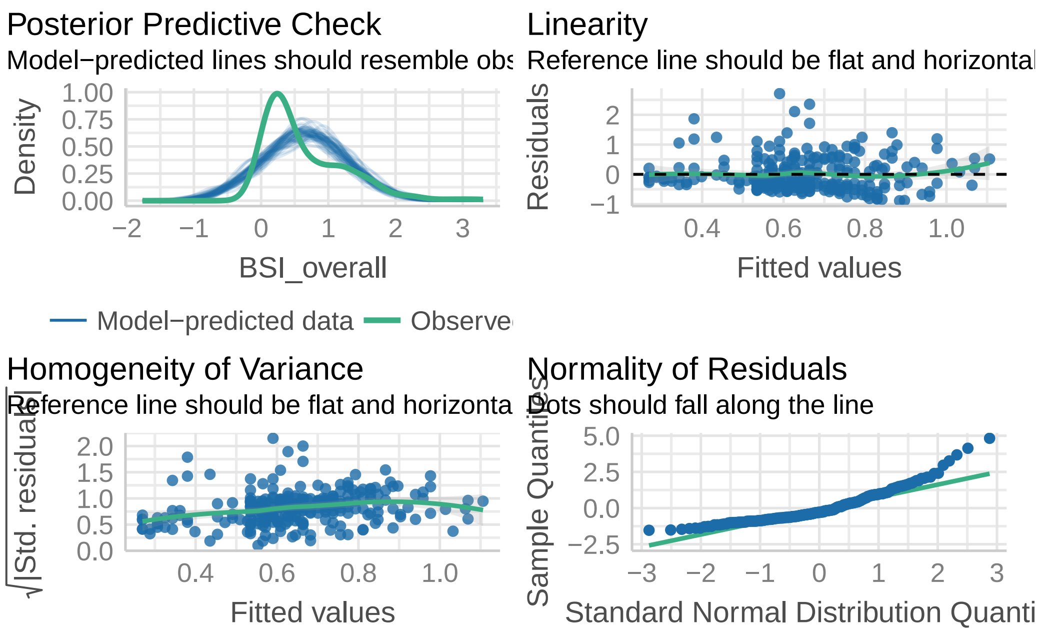
b1: After controlling for living with family, for each unit increase in parental care, adolescents' mental health scores decreased by -0.22 (-0.35, -0.09, p-value < .0001), and this is significant relationship.

b2 those living with one parent have 0.27 (0.07, 0.46, p-value=0.007) higher mental health problem scores compared to those with both parents, and this is not a significant relationship.

b3 Those living with others have 0.30 (0.00, 0.59, p-value=0.051) higher mental health problem scores compared to those with both parents, and this is not a significant relationship.

Accurancy

```
check_model(lm_model.3,
  check = c("qq", "linearity", "homogeneity", "pp_check"))
```



Posterior Predictive Check

Model−predicted lines should resemble obs

Linearity

Reference line should be flat and horizontal

Homogeneity of Variance

Reference line should be flat and horizonta

Normality of Residuals

Dots should fall along the line

The residual models appear not well fit each model. The residuals are somewhere not constant, with predicted adolescents' mental health scores higher than observed.

Conclusion

```r
r2(lm_model.3)
```

```
# R2 for Linear Regression
       R2: 0.074
  adj. R2: 0.063
```

After controlling living with family, parental care is significantly (p-value<.0001) associated with adolescents' mental health scores. But those living with one parent were significantly associated with an adolescent mental health score compared to those living with both parents and others. Living with family situations and parental care combined explains 6.3% of adolescents' mental health scores.

PART II: Modeling the probability of smoking (30 pts)

Build a regression model to model the probability a student will have started smoking prior to the study. Before you start, make sure you check the response variable data type, and identify the appropriate regression model to use.

Your model must include at least one quantitative variable from the BSI subscales and one binary predictor. You may include other variables as you desire.

- Describe the distribution of smoking status using a plot (2 pt) and a descriptive paragraph. (3 pts)
- Describe the relationship between smoking status and one of your predictors using an appropriate bivariate plot (2 pt) and description. (3 pts)
- Create an appropriate regression model your chosen predictors. (5 pts)
- Create a nicely formatted table to report/display the values that you are going to directly interpret in the next step (2 pts).
- Interpret the effect of at least one quantitative, and one binary predictor on the response. You must include a point estimate, confidence interval, and p-value in your response. (8 pts)
- Report the accuracy of your model. (5 pts)

PART II: Modeling the probability of smoking

Identify response and explanatory variables

Outcome: Age started smoking, continuous numerical variable (denoted AGESMOKED)

Predictor: Depression BSI Depression subscale (denoted BSI_depress)

Binary: Gender (female/Male)

Relationship between outcome and binary predictor variable

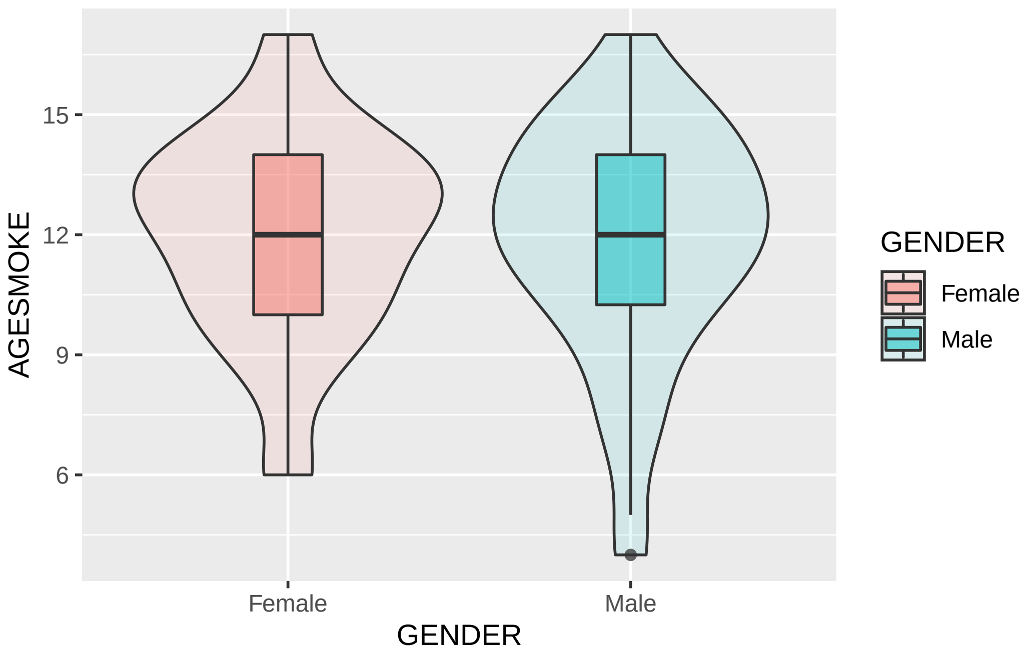
Outcome: age started smoking

Binary: Gender

```
hiv %>%
  select(GENDER, AGESMOKE) %>%
  na.omit()%>%
  group_by(GENDER)%>%
  summarize(mean = mean(AGESMOKE),
  sd = sd(AGESMOKE),
  IQR = IQR(AGESMOKE),
  n = n()) %>% kable(digits = 2)
```

GENDER	mean	sd	IQR	n
Female	11.99	2.57	4.00	70
Male	11.84	2.90	3.75	62

```
hiv %>%
  select(AGESMOKE, GENDER)%>%
  na.omit()%>%
  ggplot(aes(x=GENDER, y=AGESMOKE, fill=GENDER)) +
  geom_violin(alpha=.1) +
  geom_boxplot(alpha=.5, width=.2)
```



Summary table & violin graph

The violin box plot indicates a comparison of age started smoking between females and males. Female students' average started smoking have a mean of 11.99 with a standard deviation of 2.57 and an IQR of 4. While, males have a mean of 11.84 with a standard deviation of 2.90, and an IQR of 3.75. Both gender females and males are normally distributed, except left skew for males.

t-Test analysis for AGESOMKE ~ GENDER

The t-Test was used to see if there's a difference between the average male and female students who started smoking.

Identify response and explanatory variables

The response variable is age started smoking (AGESMOKE)

The explanatory variable is gender (female/male).

Research hypothesises

Null Hypothesis: there is no difference in average male and female students who started smoking.

Alternate Hypothesis: There is a difference in average mental male and female students who started smoking.

Assumption for t-Test

1.Mutually exclusive and independent. It is valid because a person could not be female and Male. 2.Differences are normally distributed. it is valid both groups are normally distributed, thus differences are normally distributed too. 3.Variances are similar for both groups. It is valid because the standard deviation is close to both gender females and males (2.57 vs 2.90).

```
t.test(hiv$AGESMOKE ~ hiv$GENDER)
```

```

    Welch Two Sample t-test

data:  hiv$AGESMOKE by hiv$GENDER
t = 0.30639, df = 122.72, p-value = 0.7598
alternative hypothesis: true difference in means between group Female and group Male is not
95 percent confidence interval:
 -0.8027543  1.0967635
sample estimates:
mean in group Female   mean in group Male 
            11.98571             11.83871 
```

As a result of the t-test, female students, who started smoking have on average 0.14 (95% CI -0.80, 1.10) higher age smoking compared to a male students. This is not a significant difference (p=0.76).

Quantitative response and quantitative explanatory variable. (Q ~ Q)

Identify response and explanatory variables

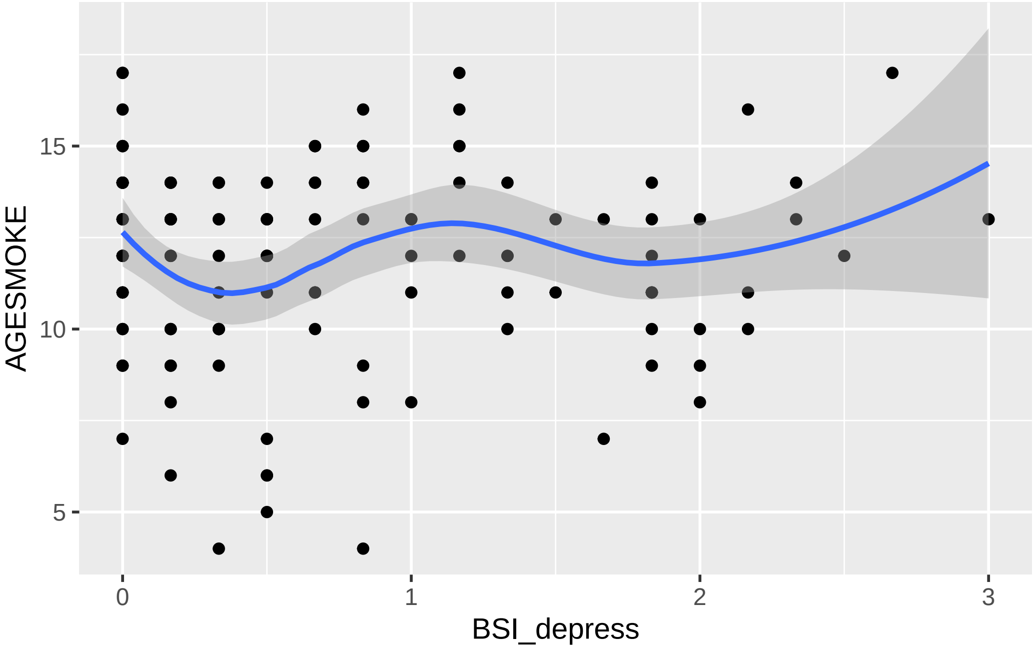
Response variable is mental health of adolescents (denoted BSI_overall), numerical variable.

Explanatory variable - Depression (denoted BSI_depress) numerical variable.

```r
cor(hiv$BSI_depress, hiv$AGESMOKE, use = "pairwise.complete.obs")
```

```
[1] 0.02442096
```

```r
hiv %>%
  select(BSI_depress, AGESMOKE)%>%
  na.omit()%>%
  ggplot(aes(x=BSI_depress, y=AGESMOKE)) + geom_point()+ geom_smooth()
```



There is a weak linear correlation between age-start smoking and students' depression level (r=0.24) with a small positive slope.

Model building for student smoking

Identify variables

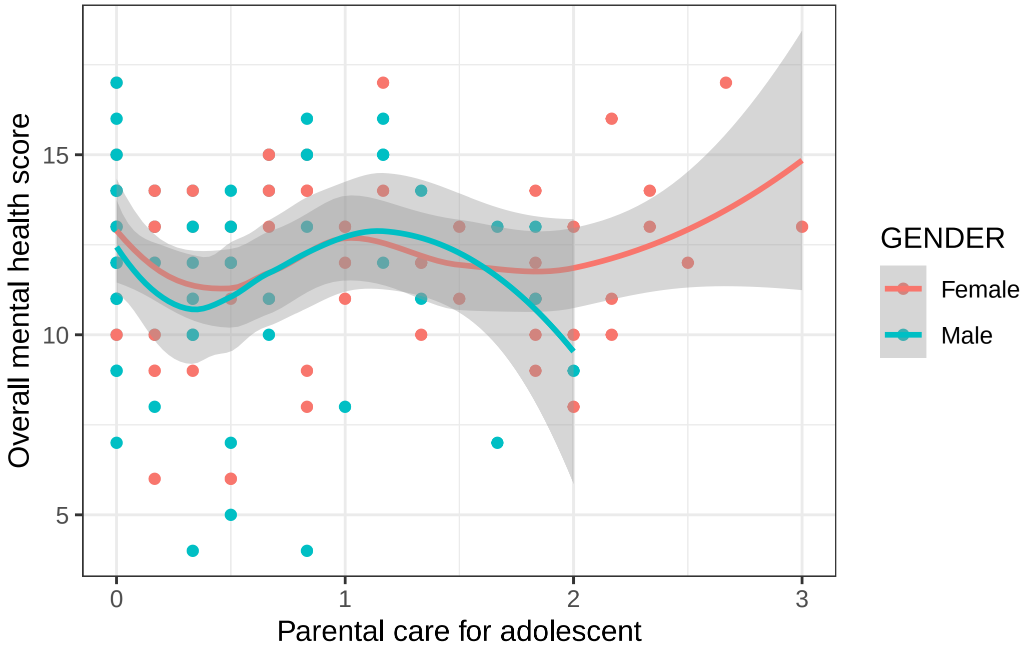
Outcome: Age started smoking, continuous numerical variable (denoted AGESMOKED)

Predictor: Depression BSI Depression subscale (denoted BSI_depress)

Binary: Gender (female/Male)

Visualize and describe the relationships

```
hiv%>%
  select(GENDER, BSI_depress, AGESMOKE) %>%
  na.omit()%>%
  ggplot(aes(x=BSI_depress, y = AGESMOKE, color = GENDER)) +
  geom_point() + geom_smooth() + theme_bw() +
  labs(x="Parental care for adolescent", y="Overall mental health score")
```



The relationship is weak, positive slope, and possibly a linear correlation between students' age started smoking and their depression level. There is upper and lower skewness.

Multilinear regression models

```
Let y be student past smoking status - Age started smoking (denoted AGESMOKE)
Let x1 be student' depression level - BSI depression (denoted BSI_depress)
Let x2 = 1 when GENDER = "Female", otherwise 0
The reference group is female
```

The mathematical model would look like:

$$Y \sim \beta_0 + \beta_1 * x_1 + \beta_2 x_2$$

```
lm.mod_smoke_dep <- lm(AGESMOKE ~ BSI_depress, data=hiv)
tidy(lm.mod_smoke_dep)
```

```
# A tibble: 2 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)  11.9        0.339    35.1   3.44e-67
2 BSI_depress   0.0899     0.326     0.275 7.84e- 1
```

The estimation of the regression coefficient for depression is not significant (b1=0.09, p-value=0.78). Thus, depression might not be associated with age-started smoking.

```
lm_model.4 <- lm(AGESMOKE ~ BSI_depress + GENDER, data=hiv)
tidy(lm_model.4)
```

```
# A tibble: 3 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)  12.0        0.462    26.0   1.90e-52
2 BSI_depress   0.0498     0.342     0.145 8.85e- 1
3 GENDERMale   -0.200      0.498    -0.402 6.89e- 1
```

```
tbl_regression(lm_model.4, intercept = TRUE) %>%
   add_glance_table(include = c(adj.r.squared))
```

Characteristic	**Beta**	**95% CI**	**p-value**
(Intercept)	12	11, 13	<0.001
BSI_depress	0.05	-0.63, 0.73	0.9
GENDER			
Female	—	—	
Male	-0.20	-1.2, 0.79	0.7
Adjusted R^2	-0.014		

Interpret the regression coefficients

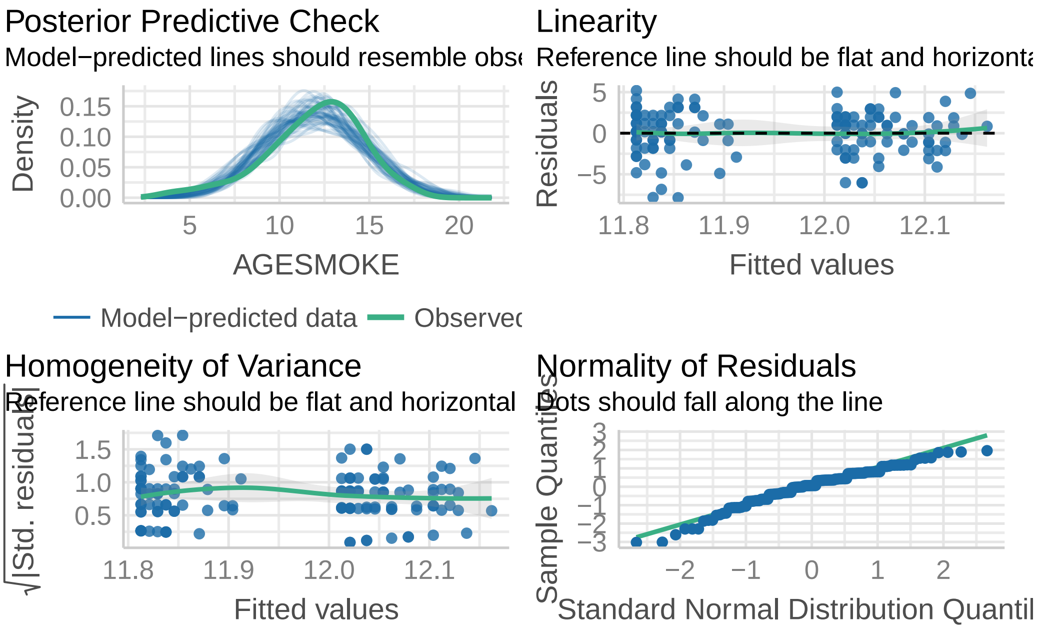
b0: The predicted age-started smoking with no depression and gender is 12 (95%CI 11, 13, p-value <.0001)

b1: After controlling gender, each unit increased in depression while age-started smoking increased by 0.05 (95%CI -0.63, 0.73, p-value=0.9). This is not a significant relationship.

b2: After controlling the depression, the predicted average age-started smoking for females is -0.20 (95% CI -1.2, 0.79, p-value=0.7) lower than for males. This is not a significant relationship.

Accurancy

```
check_model(lm_model.4,
  check = c("qq", "linearity", "homogeneity", "pp_check"))
```



The residual models appear not well fit each model. The residuals are somewhere not constant, with predicted age started smoking higher than observed.

Conclusion

```
r2(lm_model.4)
```

```
# R2 for Linear Regression
       R2: 0.002
  adj. R2: -0.014
```

After controlling gender, student depression is not significant (p-value = 0.9) associated with age-started smoking. Gender and student depression combined explain 1.4% of age-started smoking.