625 Project

2023-11-02

```
library(haven)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.2      v purrr   0.3.4
## v tibble  3.2.1      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
birthweight=read_xpt("P_ECQ.XPT")
demo=read_xpt("P_DEMO.XPT")
activity=read_xpt("P_PAQY.XPT")
diet=read_xpt("P_DBQ.XPT")
foodsec=read_xpt("P_FSQ.XPT")
body=read_xpt("P_BMX.XPT")
```

In this proposed model, we assume a linear relationship between the outcome and predictors. This assumption can be checked by the residual vs fitted plot. Second, linear regression requires multivariate normality among the variables, and this assumption can be checked by the normal Q-Q plot. Another assumption is that residual errors have a distribution centered around 0 with constant variance and are independent with each other and to the predictors. This is checked by the residual plot and scale location plot.

Model 1:
$GrowthPercentile = \beta_0 + \beta_1 Sex + \beta_2 Race + \beta_3 BirthWeight + \beta_4 Smoke + \beta_5 FoodSecurity + \beta_6 FastFood$
$+\beta_7 Activity + \epsilon$

Model 2:
$GrowthPercentile = \beta_0 + \beta_1 Sex + \beta_2 Race + \beta_3 BirthWeight + \beta_4 Smoke + \beta_5 FoodSecurity + \beta_6 FastFood$
$+\beta_7 Activity + \beta_8 Sex * FoodSecurity + \beta_9 Sex * Activity + \epsilon$

```
data <- readRDS('clean_data.RDS')
data
```

```
## # A tibble: 3,970 x 10
##    seq_id smoke sex     age food_sec activity fastfood race  birthweight
##     <dbl> <fct> <fct> <dbl> <fct>    <fct>       <dbl> <fct>       <dbl>
##  1 109263 2     1         2 1        7               0 6            7.38
##  2 109264 2     2        13 1        4               0 1            6
##  3 109265 <NA>  1         2 1        7               2 3            6.25
##  4 109269 2     1         2 2        2               3 2            7.31
##  5 109270 2     2        11 2        7               1 4            6.25
##  6 109275 2     1         6 1        7               0 3            6.69
##  7 109277 2     2        12 2        7               2 1            8.5
##  8 109278 2     2         6 3        7               2 2            6.31
##  9 109280 2     2         2 4        2              NA 1            7.62
## 10 109287 2     2        11 3        3               1 3            7.88
## # i 3,960 more rows
## # i 1 more variable: growth_percentile <dbl>
```

```r
colnames(data)
```

```
##  [1] "seq_id"            "smoke"             "sex"
##  [4] "age"               "food_sec"          "activity"
##  [7] "fastfood"          "race"              "birthweight"
## [10] "growth_percentile"
```

```r
m1 <- lm(growth_percentile ~ sex+race+birthweight+smoke+food_sec+fastfood+activity, data=data)
summary(m1)
```

```
##
## Call:
## lm(formula = growth_percentile ~ sex + race + birthweight + smoke +
##     food_sec + fastfood + activity, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -77.358 -22.046   7.568  23.513  51.852
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  59.3123     4.4354  13.372  < 2e-16 ***
## sex2          1.3540     1.1219   1.207 0.227607
## race2         1.1027     2.3719   0.465 0.642041
## race3        -5.3088     1.8113  -2.931 0.003407 **
## race4        -3.0733     1.8455  -1.665 0.095971 .
## race6       -16.4338     2.5937  -6.336 2.76e-10 ***
## race7        -6.3799     2.3363  -2.731 0.006360 **
## birthweight   3.0953     0.4214   7.345 2.73e-13 ***
## smoke2       -5.7413     1.7304  -3.318 0.000919 ***
## food_sec2     1.6975     1.5999   1.061 0.288764
## food_sec3     2.1916     1.5298   1.433 0.152085
## food_sec4     4.2223     2.1391   1.974 0.048493 *
## fastfood      0.5489     0.3537   1.552 0.120815
## activity1   -14.7757     3.8281  -3.860 0.000116 ***
## activity2    -4.3283     3.3628  -1.287 0.198168
## activity3    -4.2554     3.1410  -1.355 0.175601
## activity4    -4.1332     3.1805  -1.300 0.193864
## activity5    -3.6476     2.8750  -1.269 0.204647
## activity6    -6.3743     3.3566  -1.899 0.057664 .
## activity7   -13.5898     2.6685  -5.093 3.78e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.87 on 2678 degrees of freedom
##   (1272 observations deleted due to missingness)
## Multiple R-squared:  0.07702,    Adjusted R-squared:  0.07047 
## F-statistic: 11.76 on 19 and 2678 DF,  p-value: < 2.2e-16
```

```
m2 <- lm(growth_percentile ~ sex+race+birthweight+smoke+food_sec+fastfood+activity+sex*food_sec+sex*acti
summary(m2)
```

```
## 
## Call:
## lm(formula = growth_percentile ~ sex + race + birthweight + smoke + 
##     food_sec + fastfood + activity + sex * food_sec + sex * activity, 
##     data = data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -78.449 -21.734   7.463  23.318  59.962 
## 
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      62.4490     5.1622  12.097  < 2e-16 ***
## sex2             -4.5873     5.1661  -0.888 0.374634    
## race2             1.3990     2.3706   0.590 0.555132    
## race3            -5.3340     1.8142  -2.940 0.003310 ** 
## race4            -3.3074     1.8490  -1.789 0.073765 .  
## race6           -16.5226     2.5905  -6.378 2.10e-10 ***
## race7            -6.5286     2.3363  -2.794 0.005237 ** 
## birthweight       3.1054     0.4208   7.379 2.12e-13 ***
## smoke2           -5.7115     1.7271  -3.307 0.000955 ***
## food_sec2        -0.1113     2.1977  -0.051 0.959598    
## food_sec3         0.2950     2.1062   0.140 0.888612    
## food_sec4        -2.4959     3.1001  -0.805 0.420829    
## fastfood          0.5456     0.3530   1.546 0.122341    
## activity1       -29.2531     5.9306  -4.933 8.61e-07 ***
## activity2        -5.5038     4.8739  -1.129 0.258896    
## activity3        -9.4306     4.6967  -2.008 0.044753 *  
## activity4        -4.5821     4.5008  -1.018 0.308742    
## activity5        -4.8435     4.1037  -1.180 0.237996    
## activity6        -6.4340     4.8161  -1.336 0.181685    
## activity7       -15.4552     3.8306  -4.035 5.62e-05 ***
## sex2:food_sec2    3.5840     3.1277   1.146 0.251946    
## sex2:food_sec3    3.3128     2.9693   1.116 0.264659    
## sex2:food_sec4   12.5134     4.1602   3.008 0.002655 ** 
## sex2:activity1   23.8673     7.7864   3.065 0.002197 ** 
## sex2:activity2    2.4321     6.7136   0.362 0.717181    
## sex2:activity3    8.8261     6.3172   1.397 0.162479    
## sex2:activity4    0.3078     6.3254   0.049 0.961190    
## sex2:activity5    2.2803     5.7058   0.400 0.689444    
## sex2:activity6    0.4726     6.6623   0.071 0.943450    
## sex2:activity7    3.3614     5.2843   0.636 0.524761    
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 28.78 on 2668 degrees of freedom
##   (1272 observations deleted due to missingness)
## Multiple R-squared:  0.08604,    Adjusted R-squared:  0.0761
## F-statistic:  8.66 on 29 and 2668 DF,  p-value: < 2.2e-16
```

```
#sex*food_sec
#sex*activity
```

```
AIC(m1)
```

```
## [1] 25824.05
```

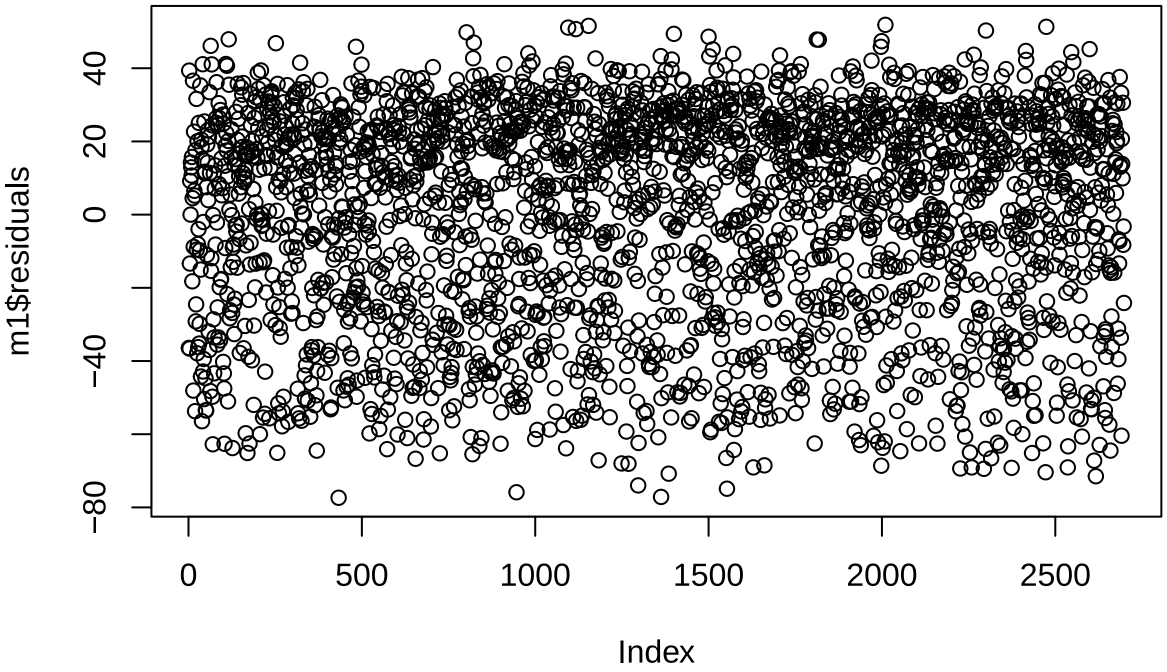
```
AIC(m2)
```

```
## [1] 25817.57
```

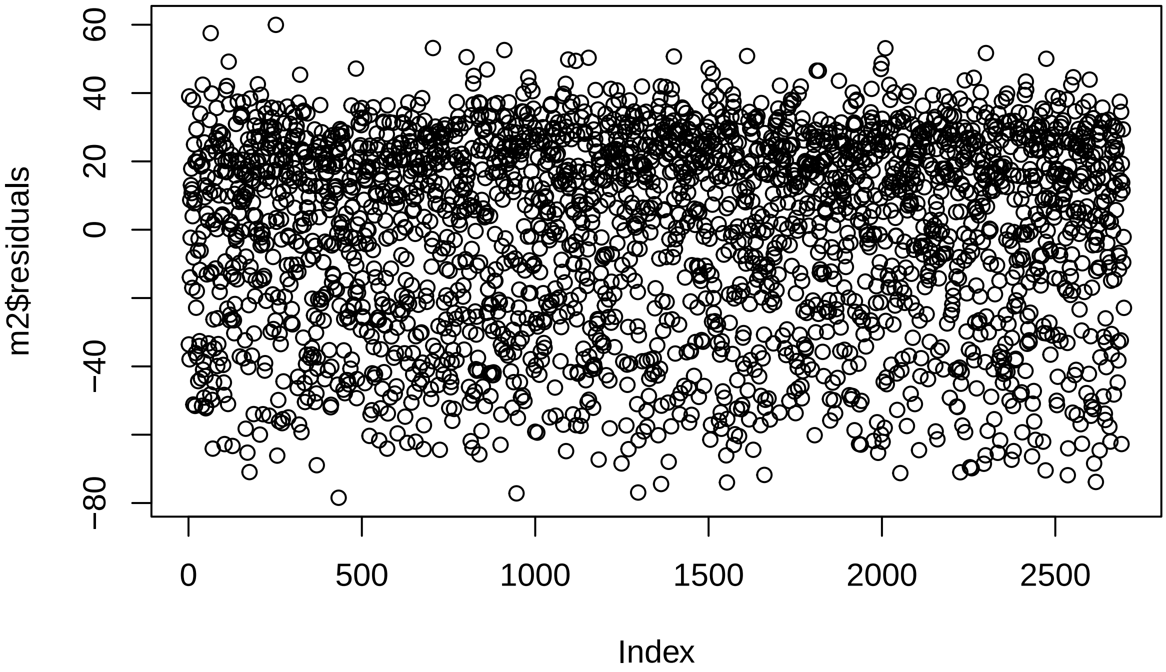
```
anova(m1,m2)
```

```
## Analysis of Variance Table
##
## Model 1: growth_percentile ~ sex + race + birthweight + smoke + food_sec +
##     fastfood + activity
## Model 2: growth_percentile ~ sex + race + birthweight + smoke + food_sec +
##     fastfood + activity + sex * food_sec + sex * activity
##   Res.Df     RSS Df Sum of Sq      F   Pr(>F)
## 1   2678 2231927
## 2   2668 2210125 10     21802 2.6318 0.003445 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(m1$residuals)
```

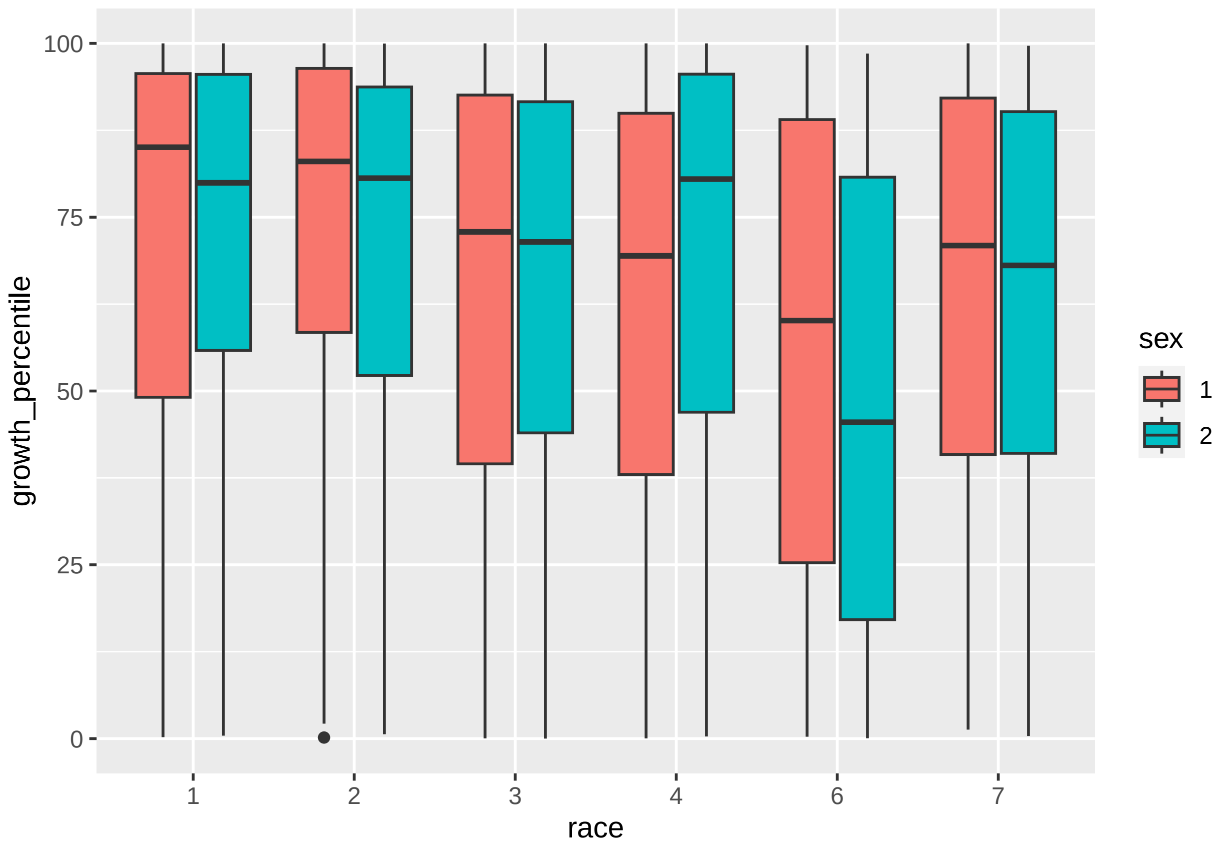


```
plot(m2$residuals)
```



```
ggplot(data, aes(x=race, y=growth_percentile, fill=sex)) +
    geom_boxplot()
```

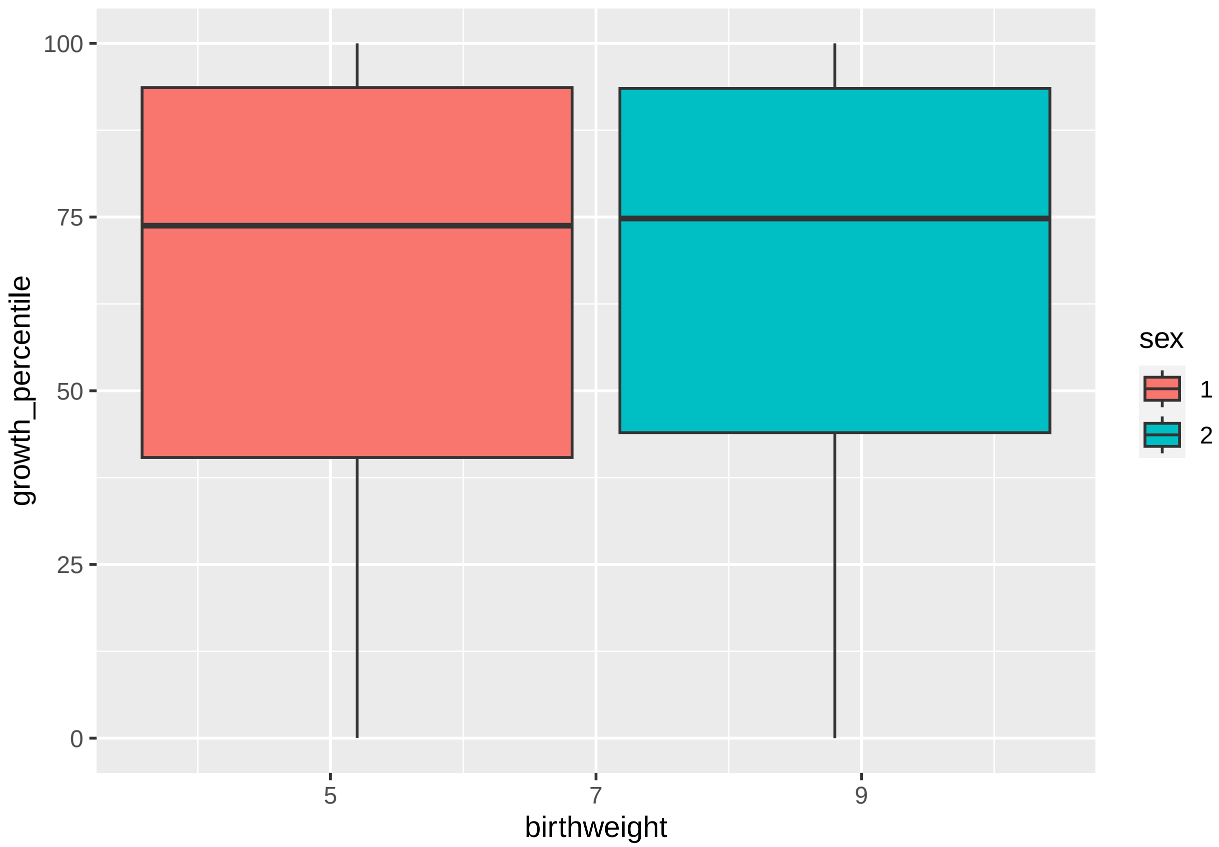
```
## Warning: Removed 97 rows containing non-finite values (‘stat_boxplot()‘).
```



```
ggplot(data, aes(x=birthweight, y=growth_percentile, fill=sex)) +
    geom_boxplot()
```

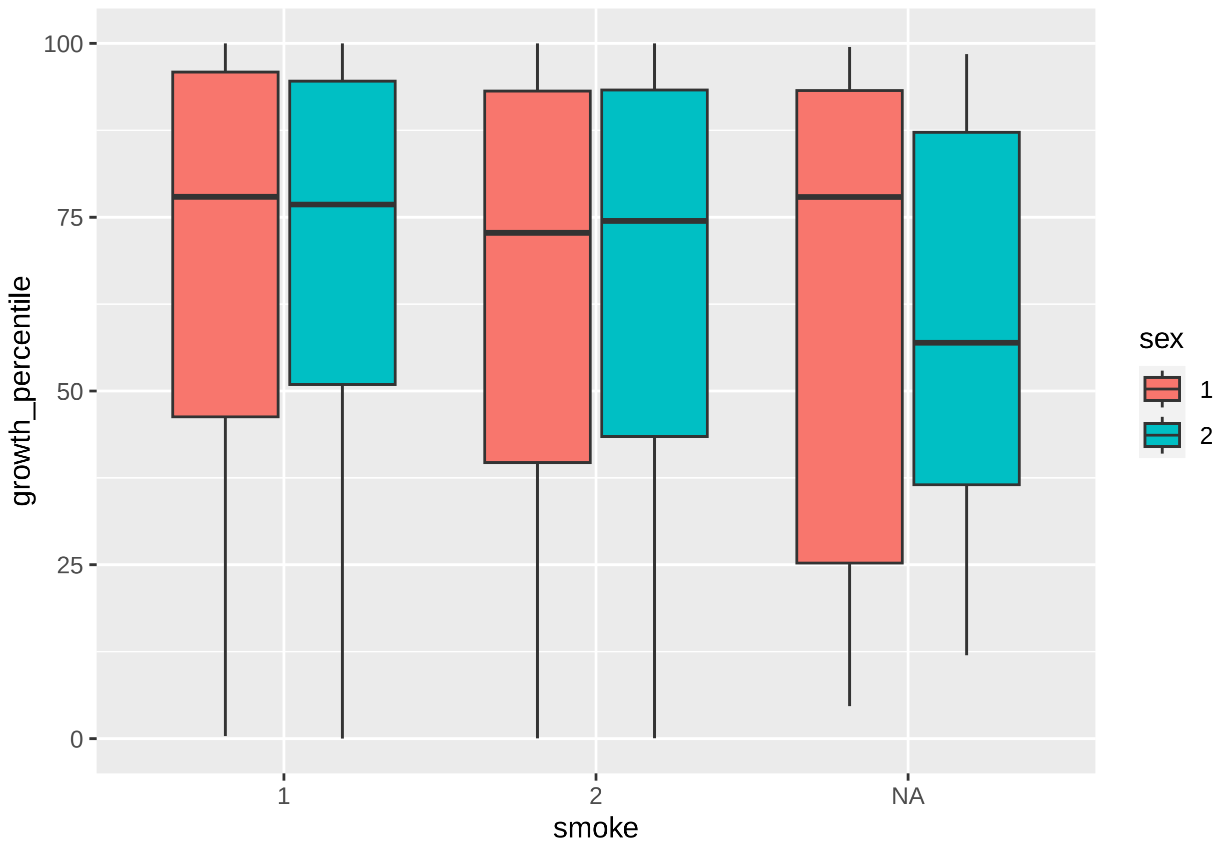
```
## Warning: Removed 137 rows containing missing values (‘stat_boxplot()‘).
```

```
## Warning: Removed 94 rows containing non-finite values (‘stat_boxplot()‘).
```



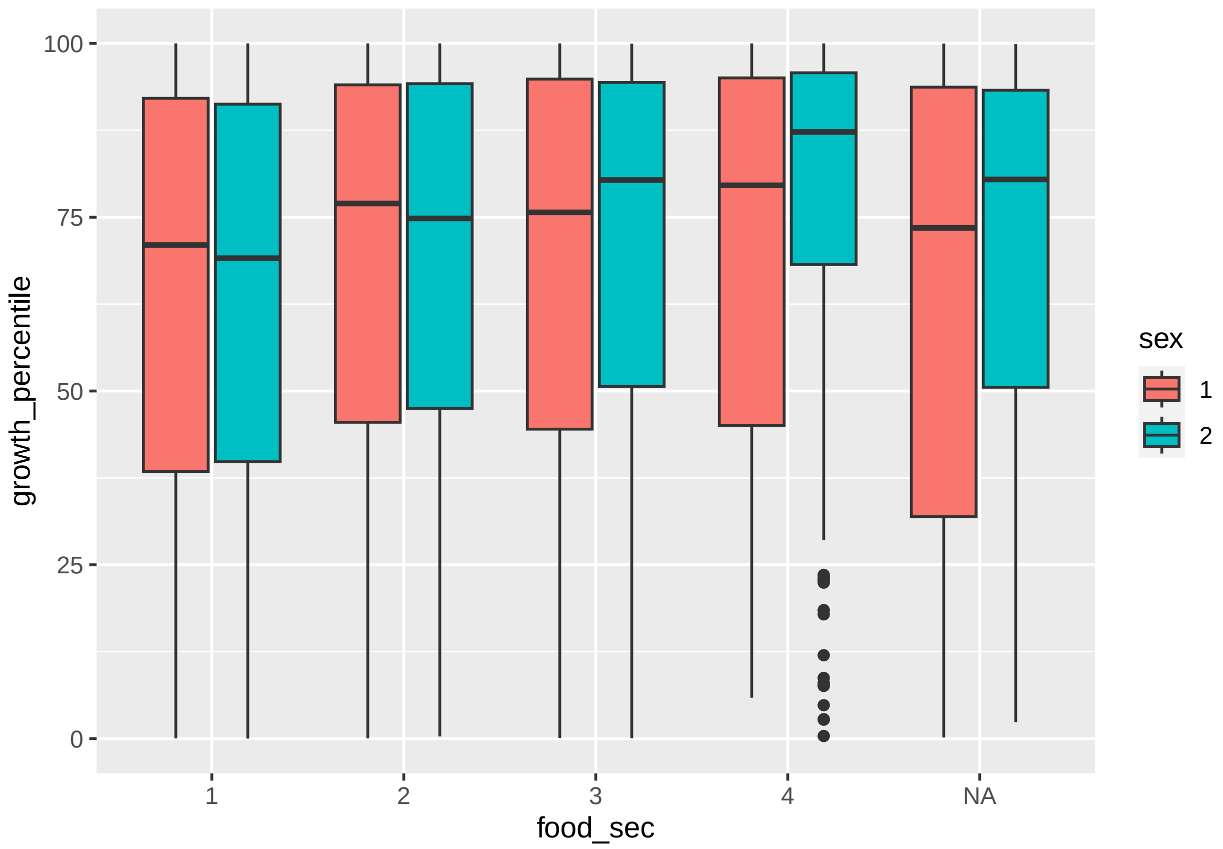
```
ggplot(data, aes(x=smoke, y=growth_percentile, fill=sex)) +
    geom_boxplot()
```

```
## Warning: Removed 97 rows containing non-finite values (‘stat_boxplot()‘).
```



```
ggplot(data, aes(x=food_sec, y=growth_percentile, fill=sex)) +
    geom_boxplot()
```

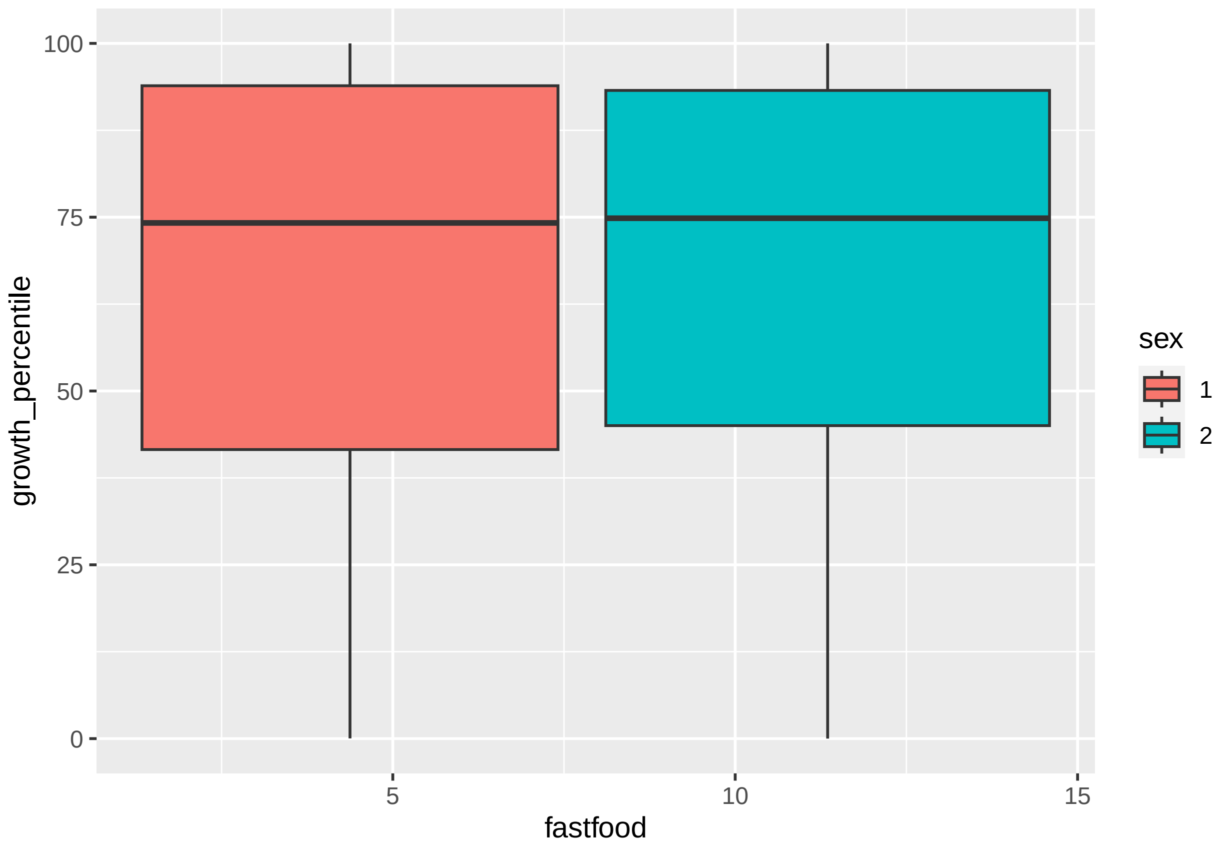
```
## Warning: Removed 97 rows containing non-finite values (‘stat_boxplot()‘).
```



```
ggplot(data, aes(x=fastfood, y=growth_percentile, fill=sex)) +
    geom_boxplot()
```

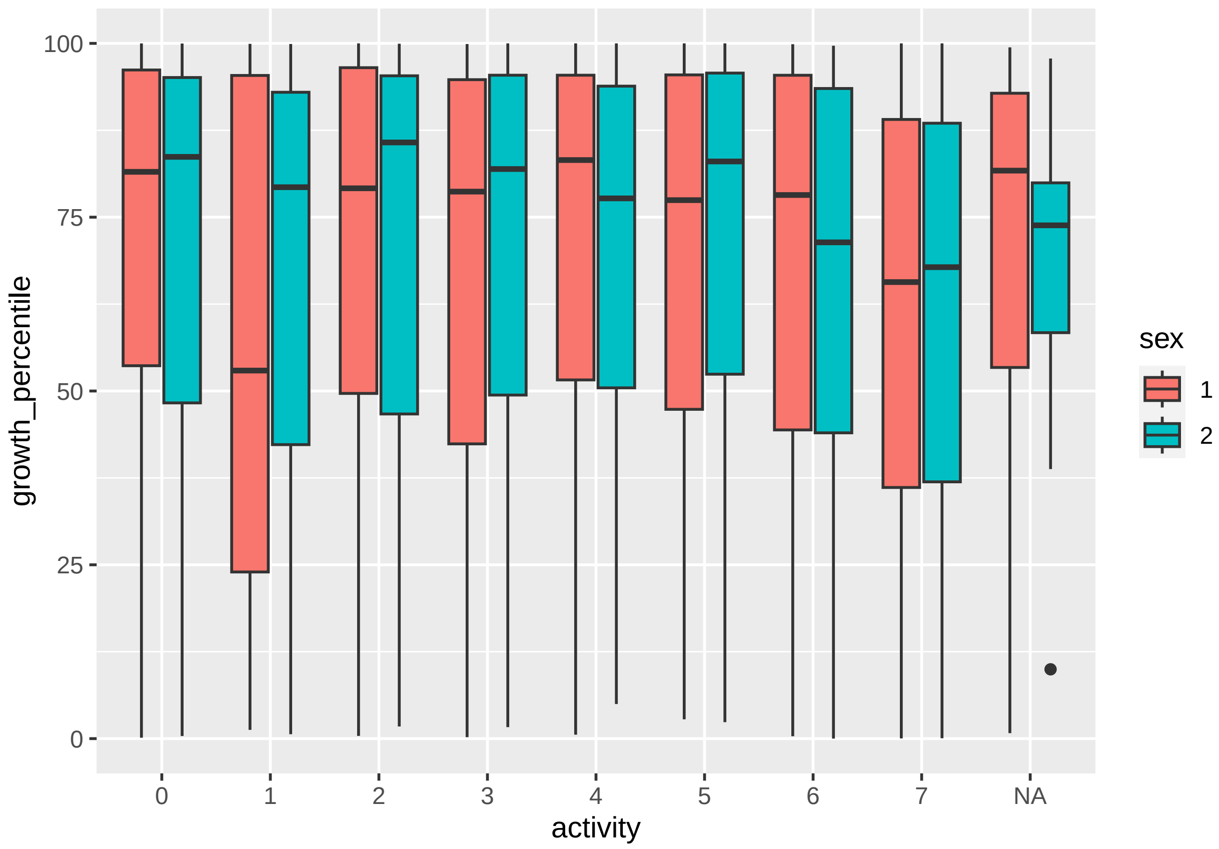
```
## Warning: Removed 951 rows containing missing values (‘stat_boxplot()‘).
```

```
## Warning: Removed 66 rows containing non-finite values (‘stat_boxplot()‘).
```



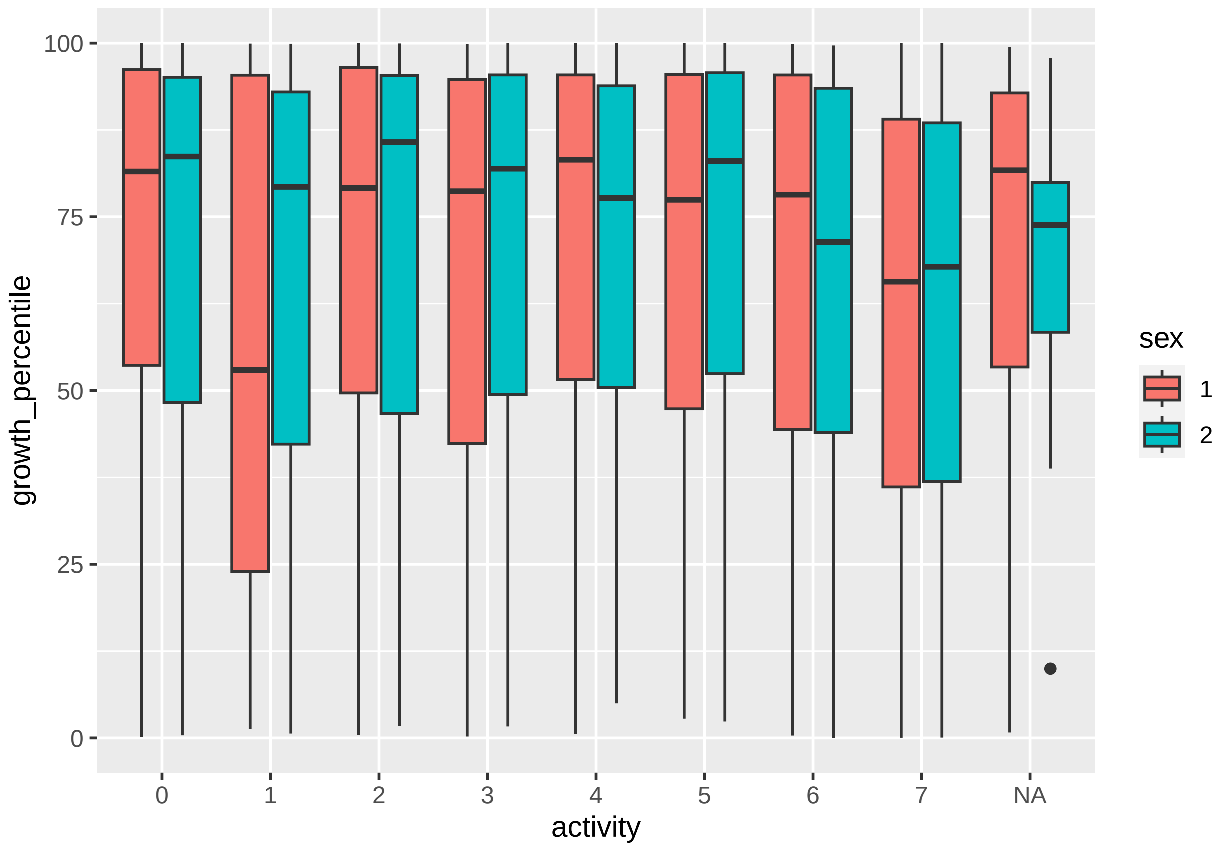
```
ggplot(data, aes(x=activity, y=growth_percentile, fill=sex)) +
    geom_boxplot()
```

```
## Warning: Removed 97 rows containing non-finite values (‘stat_boxplot()‘).
```



```
ggplot(data, aes(x=activity, y=growth_percentile, fill=sex)) +
    geom_boxplot()
```

```
## Warning: Removed 97 rows containing non-finite values (‘stat_boxplot()‘).
```


100
75
50
25
0
growth_percentile
0
1
2
3
4
5
6
7
NA
activity
sex
1
2