Week 5 Lab

Sophie Marion de Proce (edited by Christopher A Oldnall): Solutions

Package and Data Loading

In this lab, we will use the familiar NHAHES dataset to practice correlation and regression analysis, as well as checking assumptions for regression and t-tests. We will be using a significance level of 5% throughout. For this we will initially load in the data and packages necessary. This is the NHANES data set and also tidyverse.

```
library(tidyverse)
library(NHANES)
```

Exercise 1

From NHANES investigate the correlation between UrineFlow1 and UrineVol1 in adult males visually and numerically. We will consider that these variables are normally distributed for this exercise.

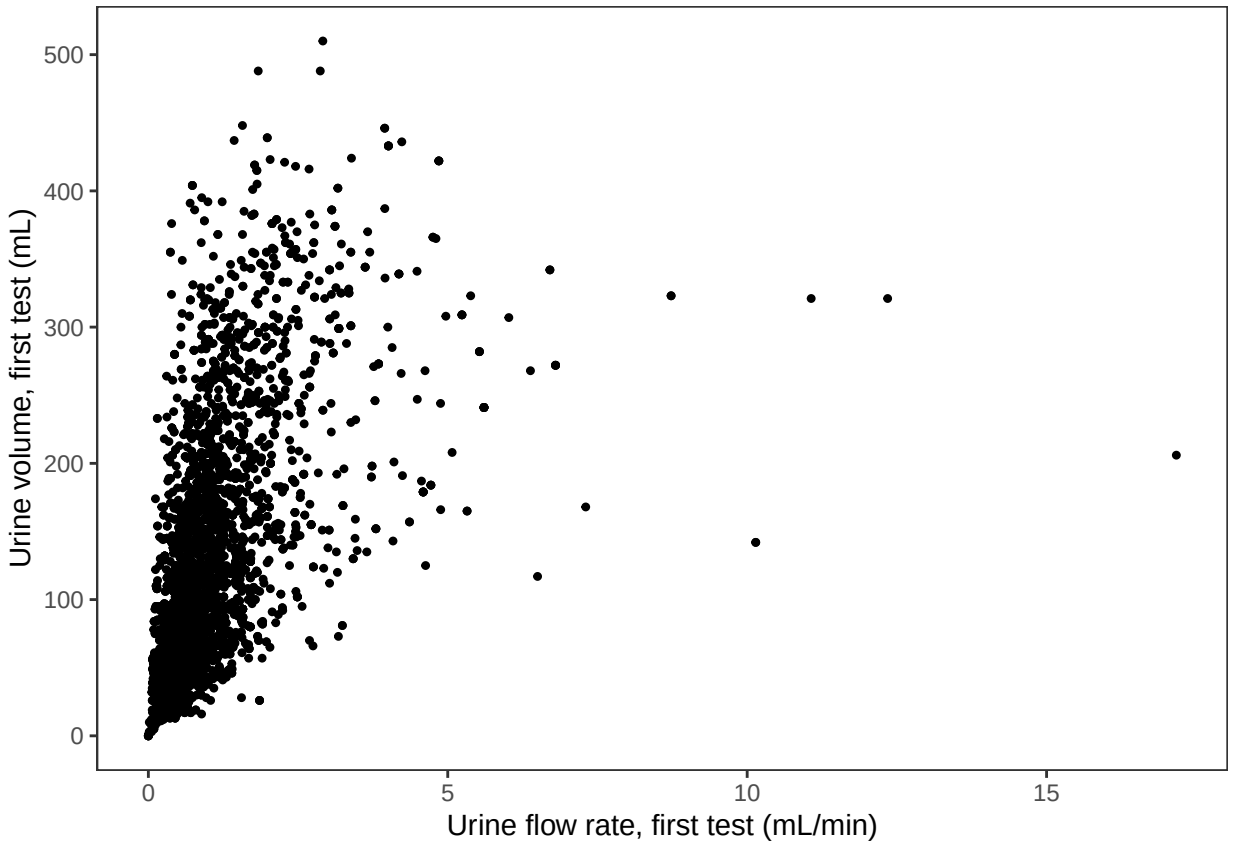
```
subsetNHANES <- NHANES %>%
  filter(Gender == "male", Age >= 18) %>%
  drop_na(UrineFlow1,UrineVol1) # remove rows where either UrineFlow1 or UrineVol1 is NA

dim(subsetNHANES)
```

```
## [1] 3446   76
```

There are 3446 rows in that data subset, so 3446 adult males have information for both UrineFlow1 and UrineVol1. Let's create a scatterplot for UrineFlow1 (x axis) and UrineVol1 (y axis)

```
subsetNHANES %>%
  ggplot(aes(x = UrineFlow1, y = UrineVol1)) +
  geom_point(pch=20) + # plot the data points and change the dot shape / size
  xlab("Urine flow rate, first test (mL/min)") +
  ylab("Urine volume, first test (mL)") +
  theme_bw() + ## remove gray background
  theme(panel.grid=element_blank()) ## remove grid
```



The scatterplot seems to show a linear positive relationship, we can investigate further with a correlation.As the task says to consider these variables as normally distributed (even though they are not), we will use the parametric Pearson's product-moment coefficient to calculate the correlation between them.Let's calculate Pearson's correlation coefficient between UrineFlow1 and UrineVol1 and display the output as a tibble (table) with columns for the coefficient estimate, t-statistic, and p-value.

```
subsetNHANES %>%
  summarize(Pearsonr = cor.test(UrineFlow1,UrineVol1,method="pearson")$estimate,
            tstat = cor.test(UrineFlow1,UrineVol1,method="pearson")$statistic,
            pval = cor.test(UrineFlow1,UrineVol1,method="pearson")$p.value)
```

```
## # A tibble: 1 x 3
##   Pearsonr tstat      pval
##      <dbl> <dbl>     <dbl>
## 1    0.531  36.8 4.57e-250
```

Pearson's product moment correlation coefficient is quite high and highly significant ($r = 0.53$, $t = 37.78$, $p < 0.005$). So there is a significant positive linear relationship between the Urine flow rate and Urine volume in adult males.

Exercise 2

Using the same dataset as in Exercise 1, build a simple linear regression model to predict UrineVol1 based on UrineFlow1 in adult males. Report its parameters, assess its quality and determine whether the model assumptions hold.

Let's build a linear regression model to predict UrineVol1 based on UrineFlow1 using the subsetNHANES dataset and save the output as an object called lmodel.

```
lmodel <- lm(UrineVol1 ~ UrineFlow1, data = subsetNHANES)
```

Now we can extract the model coefficients and their 95% confidence intervals. We can also display a summary of the linear model, including: the variables and dataset used; summary statistics for the distribution of the residuals; the coefficient estimates and their associated standard error, t-value and p-value; model measure of quality (e.g. residual standard error, adjusted R-squared and F statistic).

```
lmodel$coefficients
```

```
## (Intercept)  UrineFlow1
##    83.58769    48.19614
```

```
confint(lmodel)
```

```
##                 2.5 %   97.5 %
## (Intercept) 79.75244 87.42295
## UrineFlow1  45.62678 50.76550
```

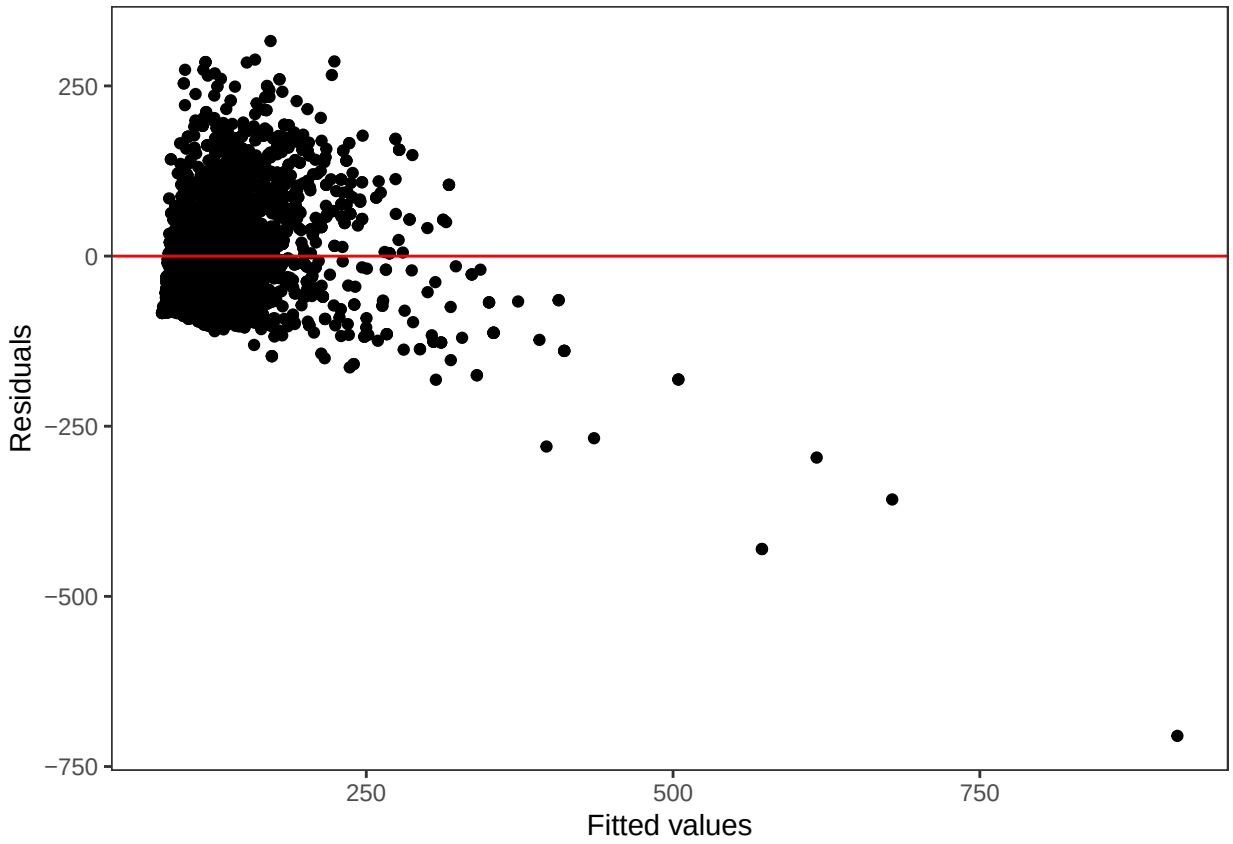
```
summary(lmodel)
```

```
##
## Call:
## lm(formula = UrineVol1 ~ UrineFlow1, data = subsetNHANES)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -704.97  -56.77  -17.60   42.67  315.97
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   83.588      1.956   42.73   <2e-16 ***
## UrineFlow1    48.196      1.310   36.78   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 77.72 on 3444 degrees of freedom
## Multiple R-squared:  0.282,  Adjusted R-squared:  0.2818
## F-statistic:  1353 on 1 and 3444 DF,  p-value: < 2.2e-16
```

The regression coefficient for UrineFlow1 is 48.2 [95% CI: 45.6-50.8] and the linear model intercept is 83.6 [95% CI: 79.8-87.4]. So an increase of 1 mL/min of urine flow rate is associated with an increase of 48.2 mL in urine volume. Both coefficients are significantly different from 0 (β_1: t = 36.78, p < 0.005; β_0: t = 42.73, p < 0.005). The model explains 28% of the variance in the response variable UrineVol1 ($R^2 = 0.28$). The residual standard error is a bit high (77.7), so there is some deviation of the response variable UrineVol1 from the regression line, but the F statistic is high (F = 1353), suggesting a significant relationship between UrineFlow1 and UrineVol1.

We can now check whether the assumptions of the linear regression model hold. We will first look at the assumption of constant variability of the residuals.

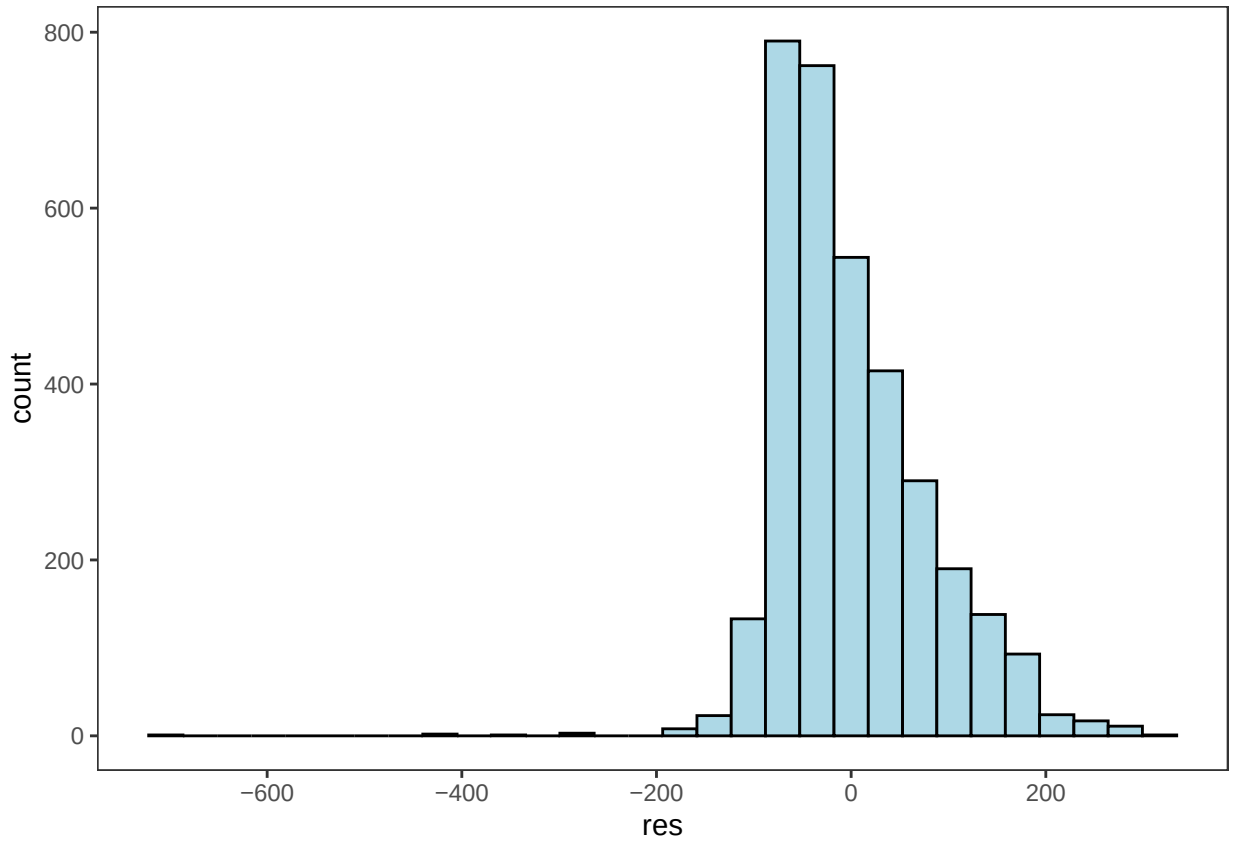
```
# Produce a residual vs. fitted plot
ggplot(lmodel, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(aes(yintercept=0),color="red") +
  labs(y='Residuals',x='Fitted values') +
  theme_bw() + ## remove gray background
  theme(panel.grid=element_blank()) ## remove grid
```



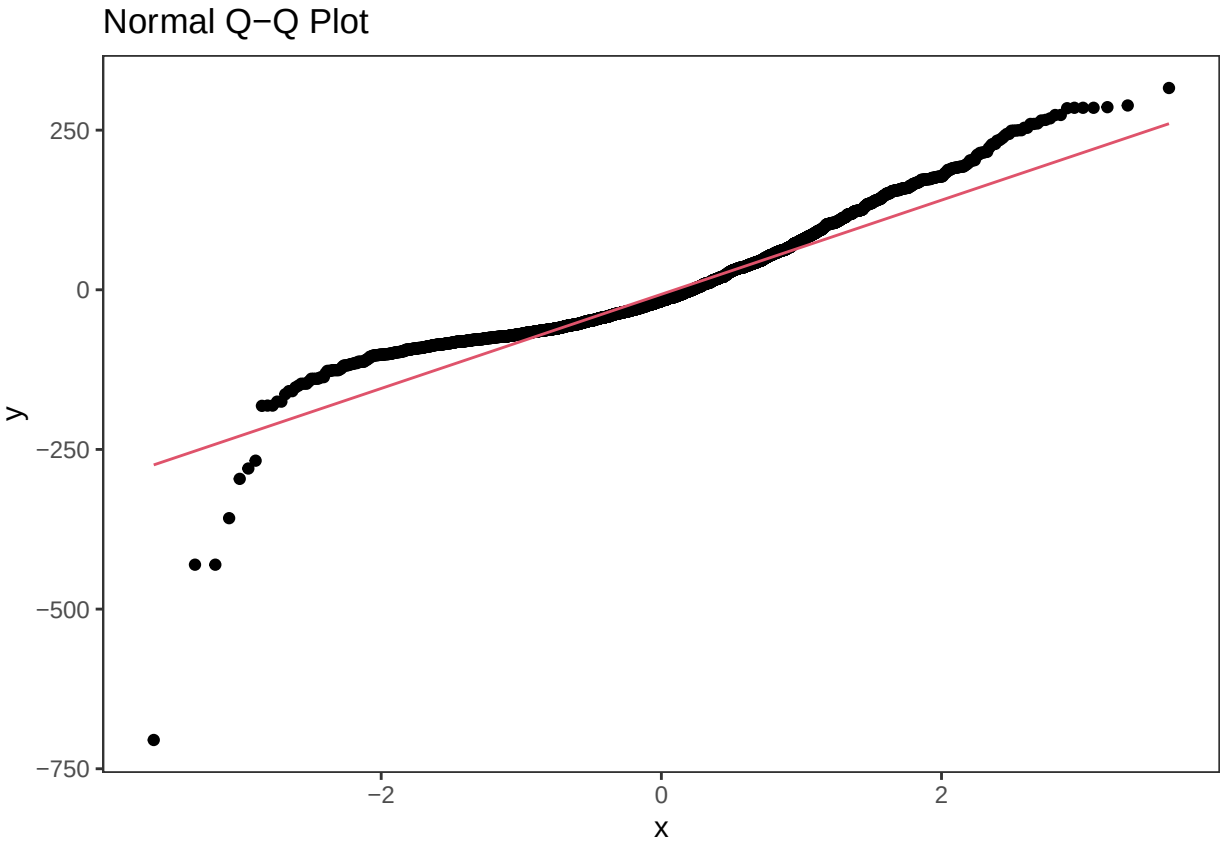
This plot suggests some non-random patterns of distribution of the residuals. It is likely that this assumption is not met. Then we can check the assumption of normal distribution of the residuals.

```
# Get the list of residuals
res <- resid(lmodel)

# Plot a histogram for the residuals
res %>%
  as_tibble() %>%
  ggplot(aes(x = res)) +
  geom_histogram(col="black",fill="lightblue",bins = 30 ) +
  theme_bw() + ## remove gray background
  theme(panel.grid=element_blank()) ## remove grid
```




```
# Create a Normal Q-Q plot for the residuals
res %>%
  as_tibble() %>%
  ggplot(aes(sample=res)) +
  stat_qq() +
  stat_qq_line(color=2) +
  labs(title="Normal Q-Q Plot") +
  theme_bw() +
  theme(panel.grid=element_blank()) ## remove grid
```




```
# Perform the Shapiro-Wilk test on the residuals
shapiro.test(res)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res
## W = 0.93036, p-value < 2.2e-16
```

The histogram suggests a right skew, and the Q-Q plot show that there is some deviation from the normal distribution. The Shapiro-Wilk test also has a very low p-value ($W = 0.93$, $p < 0.005$). It seems like the assumptions of the regression model don't hold, so it is probably necessary to transform the variable.

Exercise 3

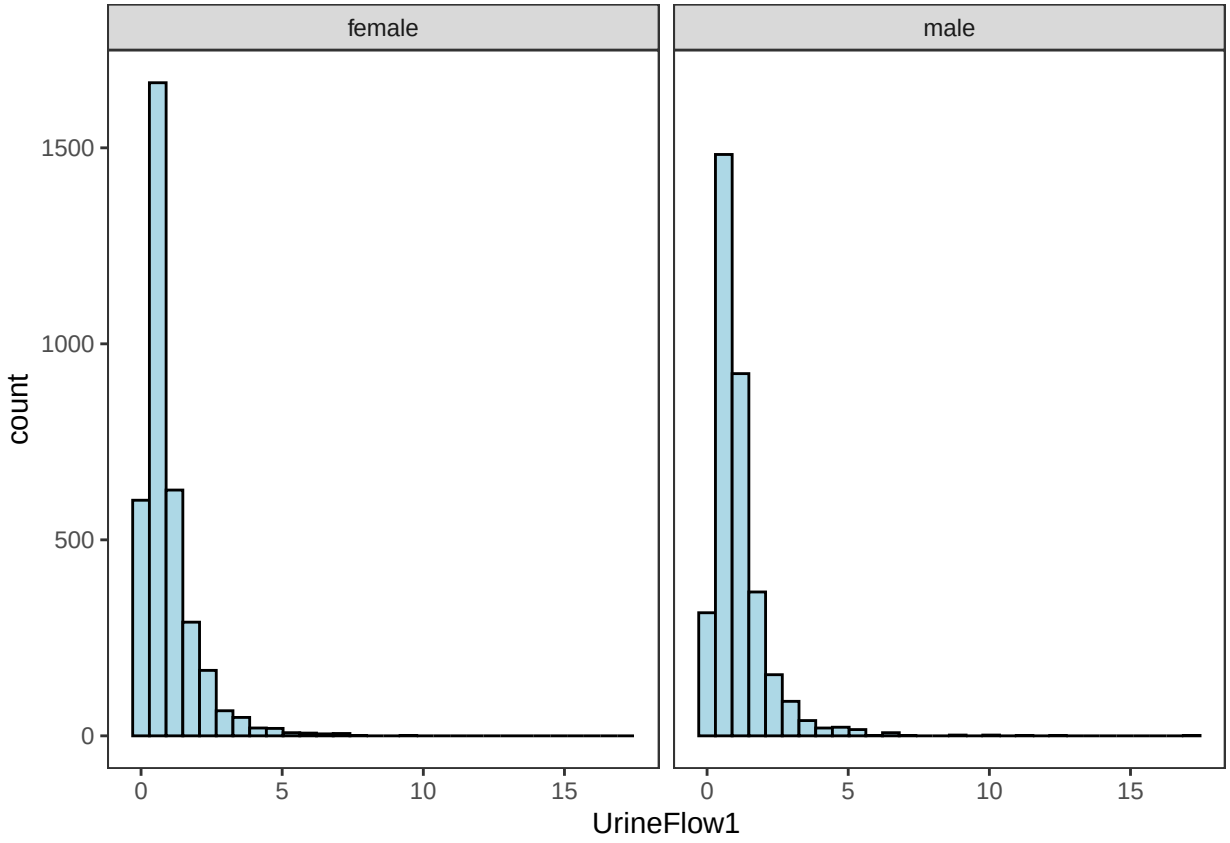
If you were to investigate the difference in UrineFlow1 between adult males and females, check whether the assumptions for parametric tests hold.

Let's create a new subset of NHANES with only adults with Gender and UrineFlow1 information.

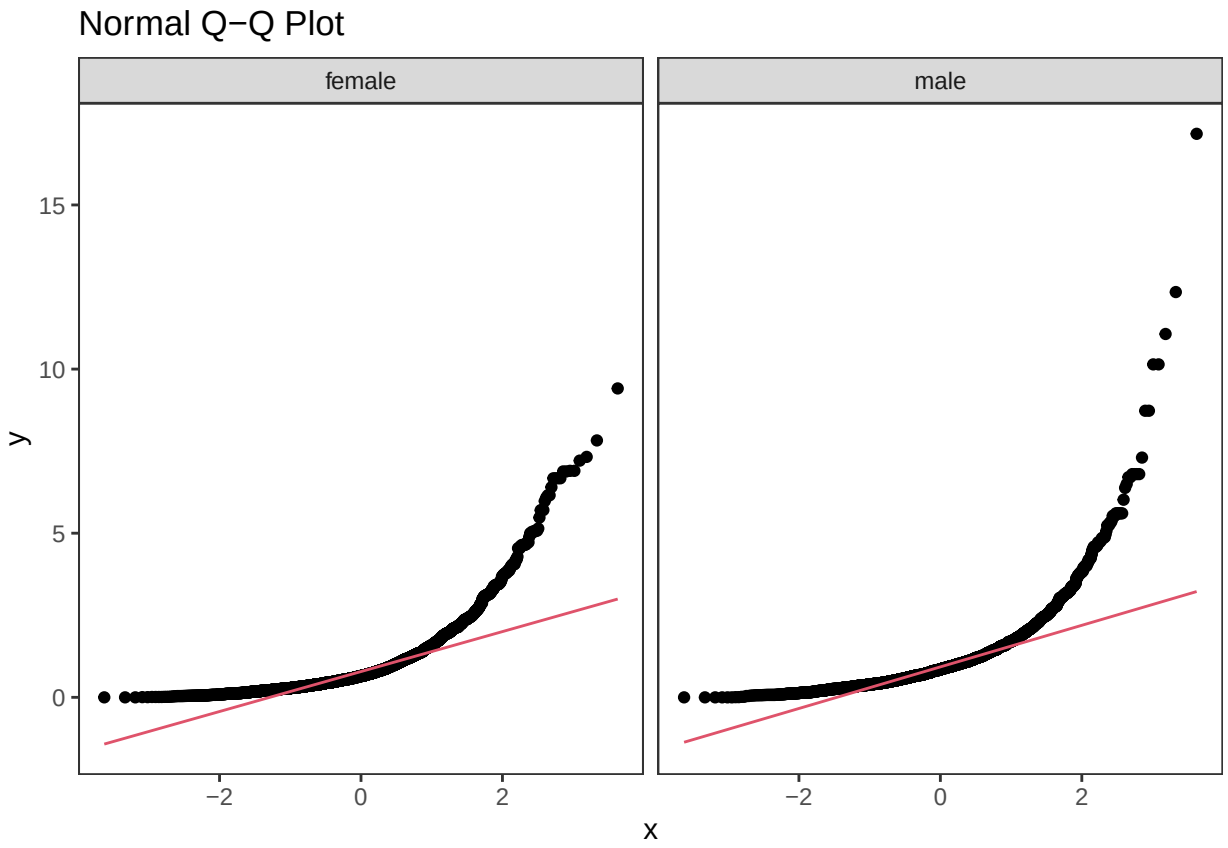
```
subsetNHANES2 <- NHANES %>%
  filter(Age >= 18) %>% # keep only adults
  drop_na(UrineFlow1, Gender) # remove individuals with either missing Gender or UrineFlow
```

First we will test the assumption of normality. When comparing the means of two different independent groups (such as male vs female heights), both sets of data are assumed to be normal, and both should be tested either individually. Let's create a histogram of UrineFlow1 for each Gender. We will also create a Normal Q-Q plot for UrineFlow1 for each Gender and perform the Shapiro-Wilk test on each gender.

```
subsetNHANES2 %>%
  ggplot(aes(x = UrineFlow1)) +
  facet_wrap(~Gender) +
  geom_histogram(col="black",fill="lightblue",bins = 30 ) +
  theme_bw() + ## remove gray background
  theme(panel.grid=element_blank()) ## remove grid
```




```
subsetNHANES2 %>%
  ggplot(aes(sample=UrineFlow1)) +
  stat_qq() +
  stat_qq_line(color=2) +
  facet_wrap(~Gender) +
  labs(title="Normal Q-Q Plot") +
  theme_bw() +
  theme(panel.grid=element_blank()) ## remove grid
```




```
subsetNHANES2 %>%
  filter(Gender == "male") %>%
  pull(UrineFlow1) %>%
  shapiro.test()
```

```
##
##  Shapiro-Wilk normality test
##
## data:  .
## W = 0.70025, p-value < 2.2e-16
```

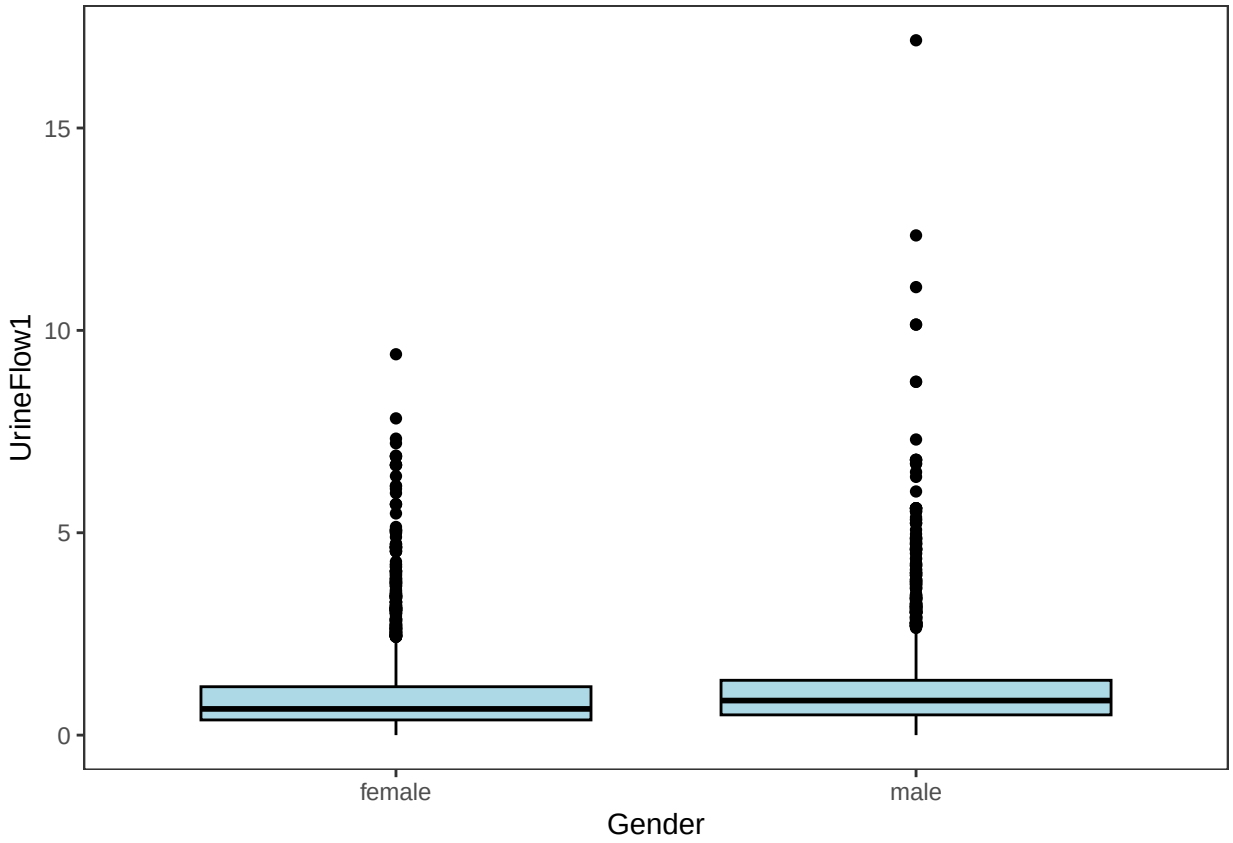
```
subsetNHANES2 %>%
  filter(Gender == "female") %>%
  pull(UrineFlow1) %>%
  shapiro.test()
```

```
##
##  Shapiro-Wilk normality test
##
## data:  .
## W = 0.74158, p-value < 2.2e-16
```

The histograms, Q-Q plots and Shapiro-Wilk tests ($p < 0.005$) all suggest that UrineFlow1 is not normally distributed for either Gender, therefore we would perform non-parametric tests (or transform the data). We will now look at the assumption of homogeneity of variances. We can first create a boxplot for UrineFlow1 for both genders.

```
subsetNHANES2 %>%
  ggplot(aes(x = Gender, y = UrineFlow1)) +
```

```
  geom_boxplot(col="black",fill="lightblue") +
  theme_bw() + ## remove gray background
  theme(panel.grid=element_blank()) ## remove grid
```



We will also display the standard deviations of UrineFlow1 for both groups.

```
subsetNHANES2 %>%
  group_by(Gender) %>%
  summarise(sd = sd(UrineFlow1))
```

```
## # A tibble: 2 x 2
##   Gender    sd
##   <fct>  <dbl>
## 1 female 0.931
## 2 male   1.01
```

Finally, we use the F-test of homogeneity of variances

```
subsetNHANES2 %>%
  var.test(UrineFlow1 ~ Gender, data = .)
```

```
##
##  F test to compare two variances
##
## data:  UrineFlow1 by Gender
## F = 0.84838, num df = 3528, denom df = 3445, p-value = 1.218e-06
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.7938663 0.9066171
## sample estimates:
```

```
## ratio of variances
##          0.8483819
```

The boxplot, standard deviations, and F-test (F=0.85, df=3528, p=1.22x10-6) all suggest that the variances are equal between the two groups. Overall, since the data is not normally distributed, it would be best to transform the data, or use non- parametric tests.