STATProject

Contents

Abstract **2**

Linear regression test **3**

T test **5**

χ^2 **test** **7**

References **8**

Abstract

This is the final demonstration of my skills and understanding in the statistical testing and writing in R. The three statistical tests that would be discussed are linear regression, t test and the χ^2 test. The package is separate to the codes in this report, but I have copied and pasted the code chunks from the package. In this way, the two parts of the project are quite closely linked.

Please enjoy!

Linear regression test

The first research question is whether there is a linear relationship between heights and weights of the sample population. The β is being tested, so the sampling distribution is: $\hat{\beta} \sim N(\beta, \frac{\sigma^2}{S_{xx}})$.

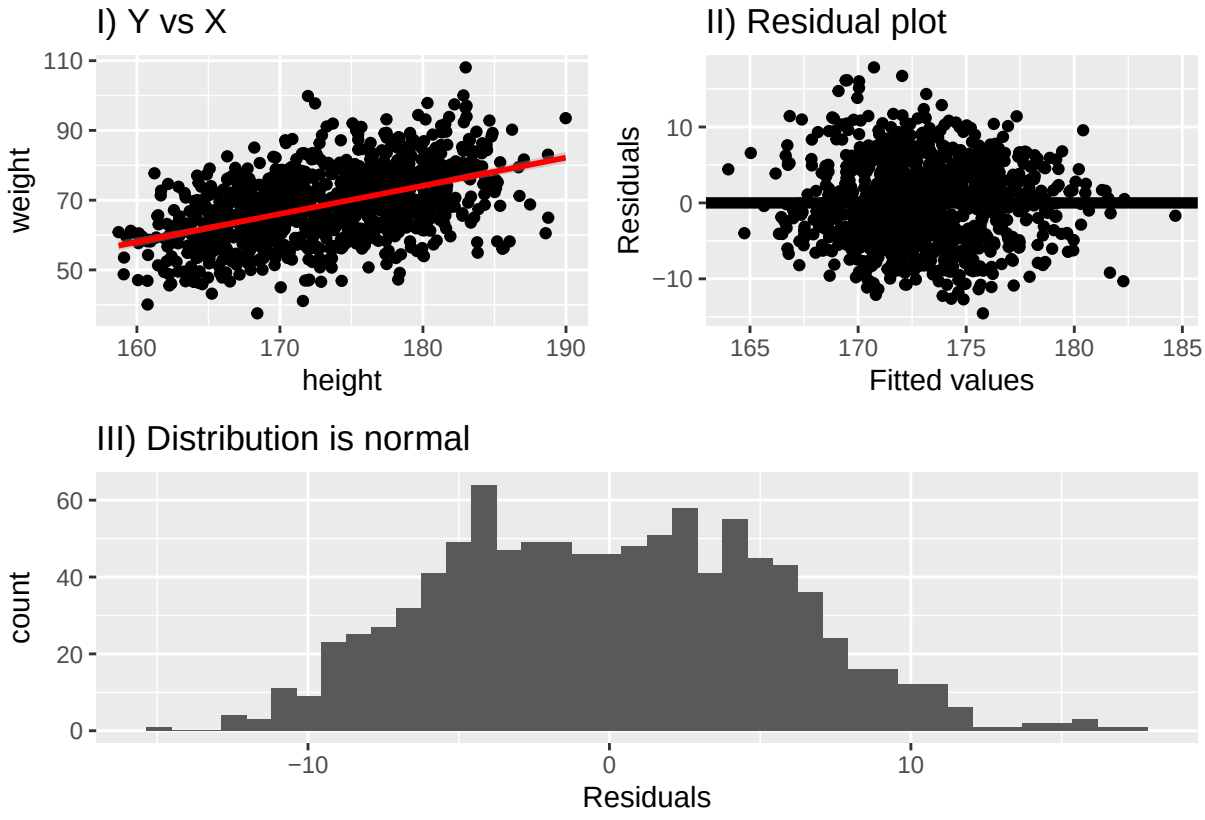
$H_0 : \beta = 0$ against $H_1 : \beta \neq 0$

We assume the linear regression model is appropriate: $Y_i = \alpha + \beta X_i + \epsilon_i$, where ϵ_i are independent and identically distributed $N(0, \sigma^2)$.

Test statistic: $\tau = \frac{\hat{\beta}}{s_{Y|X}/\sqrt{S_{xx}}} \sim t_{n-2}$ under H_0.

$\tau_{obs} = 132.3145054$

Making a proper linear regression decision requires the linearity between the two variables, constant variance and normality of the residuals. In fact, these are all assumed to hold already. The following graphs must show the linearity, evenly dispersed residuals and a bell shaped histogram of the residuals vs fitted values:



Once it is clear that these assumptions are met, the P value can then be computed directly using the r function lm().

The computed P value is $1.6279622 \times 10^{-60}$. With this P value, it is possible to make the conclusion about the test:

```
## REJECT H0:  1.627962e-60  < 0.05
##
## There is a relationship between height and weight: As the P-value is very small,
## we have very strong evidence to reject H0. I.E. very strong evidence that the
## slope parameter is significant and there is a relationship between the height
## and weight of the sample population.
```

T test

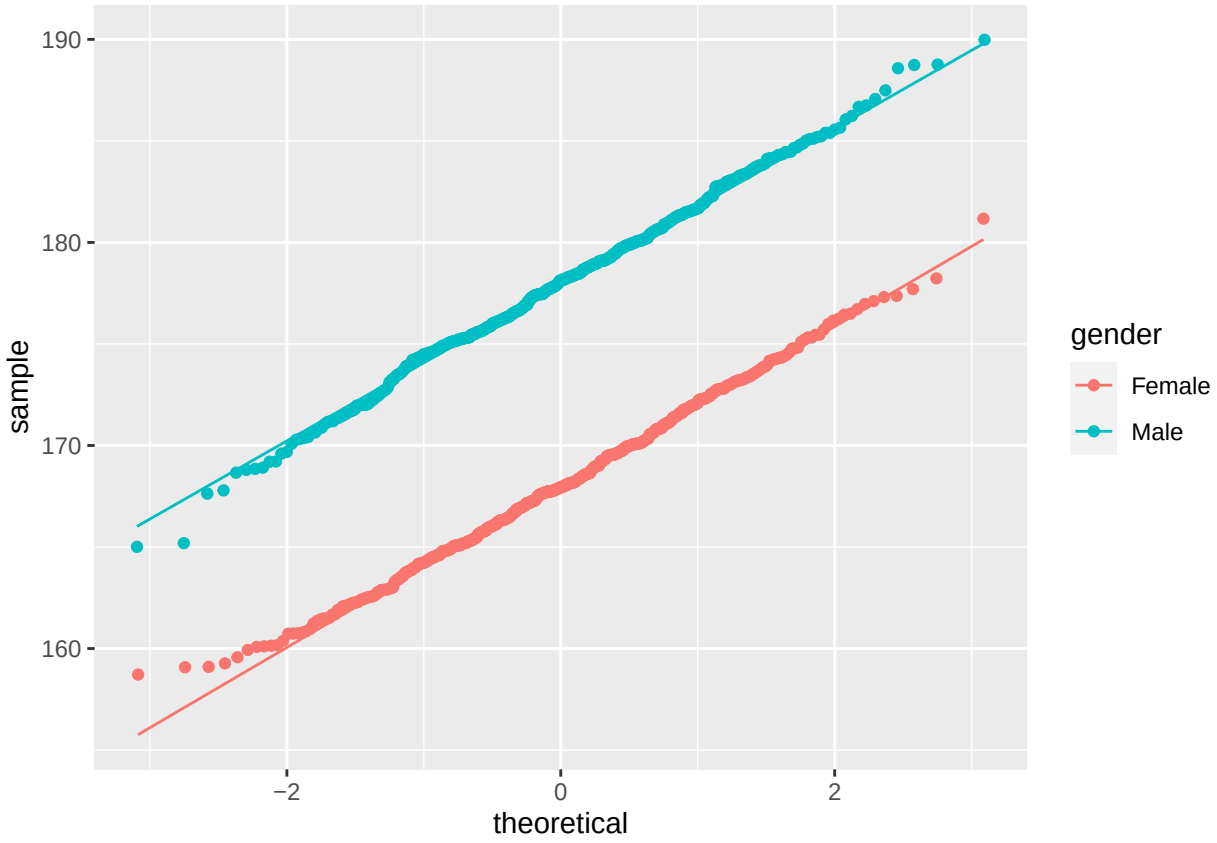
The next test is about the mean heights of male and female. It is to test if they are same or not. The t test is based on equal variances assumption. Assuming the null hypothesis $H_0 : \mu_1 = \mu_2$, the resulting sampling distribution is $\frac{\overline{X_1}-\overline{X_2}}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim N(0,1)$.

$H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$

Test statistic: $\tau = \frac{\overline{X_1}-\overline{X_2}}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2}$, if H_0 is true.

$\tau_{obs} = 39.9471643$.

2 conditions must be met to carry out this test. The normal distribution and equal variance must hold. Firstly, the QQ plot can be examined to confirm the normality assumption:



Since the equal variance is assumed, the larger standard deviation divided by the smaller must not exceed 2. The larger is 3.9596539 and the smaller is 3.8744566. The division gives 1.0219895.

Once all assumptions are confirmed to hold, the t.test() function directly outputs the P value ($3.3213865 \times 10^{-209}$) and the conclusion can be made:

```
## REJECT H0:  3.321387e-209  < 0.05
##
## The mean height of male and female are NOT the same: As the P-value is very
```

```
## small, we have very strong evidence to reject H0. I.E. very strong evidence that
## the mean height of male is not the same as the mean height of female.
```

χ^2 test

The last test is to see if male and female have different amount of physical activity. In other words, if gender affects the amount of physical activity. In terms of statistics, this is equivalent to saying there is association between the two variables gender and physical activity. A χ^2 distribution is constructed by squaring a single standard normal distribution: $Q \sim \chi_i^2$ where Q is an example of a χ^2 distribution. Then $Q = Z^2$ where $Z \sim N(0, 1)$.

H_0 : the two variables are independent against each other. H_1 : not H_0.

The Pearson's χ^2 test-statistic (without continuity correction) for the test of independence is: $\tau = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(r-1)(c-1)}$, under H_0.
$\tau_{obs} = 1.0267993$.

Since the test is based on the normal approximation, all entries (O_{ij} and E_{ij}) in the table below must be at least 5:

```
## # A tibble: 3 x 3
##   `Physical activity`  Male Female
##   <chr>               <int>  <int>
## 1 None                  127    116
## 2 Moderate              255    242
## 3 Intense               125    135
```

Assuming they are all greater than or equal to 5, the P value can then be generated with the chisq.test() function. The P value is 0.5984576. Based on this P value, the conclusion to the test can be made:

```
## DO NOT REJECT H0:  0.5984576  > 0.05
##
## Gender does NOT affect the amount of physical activity: As the P-value is large,
## we have no evidence to reject H0. I.E. no evidence that the two variables are
## dependent against each other. The two variables are independent against each
## other and there is no association between gender and the amount of physical
## acitivity.
```

References

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.