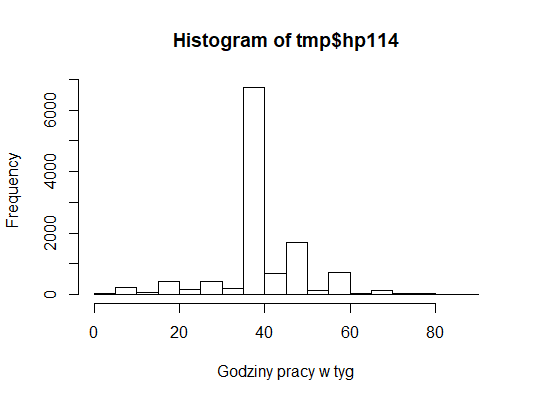
Statistics project – Description

Exercise 1:

I decided to go with the question: “ILE PRZECIETNIE GODZIN W TYGODNIU PRACUJESZ?”.



Here are some main calculation that I got:

* Mean is equal to 41.3 hour per week,
* Standard deviation is equal to 10.7,
* Skewness is equal to -0.29,
* Kurtosis is equal to 2.77.

Quartiles & percentiles:

* 5% - 20 h/w,
* 25% - 40 h/w,
* 50% - 40 h/w,
* 75% - 45 h/w,
* 95% - 60 h/w.

As we can see, majority of people work about 40 hours per week.

As we can see, majority of people work about 40 hours per week.

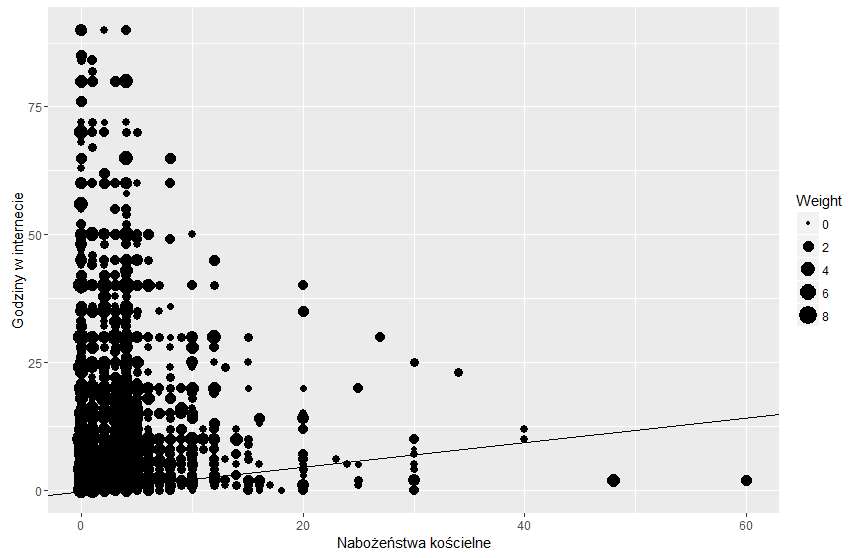
Skewness is almost zero, which means that mean is very close to the mode.

Also this means that data is not very asymmetric. Standard deviation tells us about spread of data, which is equal to almost 11.

Exercise 2:

For this exercise I choose two entirely other questions:

* “Ile godzin w ostatnim tygodniu korzystała Pani z internetu?”
* “Jak często przeciętnie w ciągu miesiąca bierze Pani udział w nabożeństwach lub innych spotkaniach o charakterze religijnym?”



We can see that most of people who uses internet more than 50 hours per week are not likely to go to the church very often. On the other hand, people who go more than 20 times to church, don’t use internet that often.

* Correlation is equal to 0.09689039

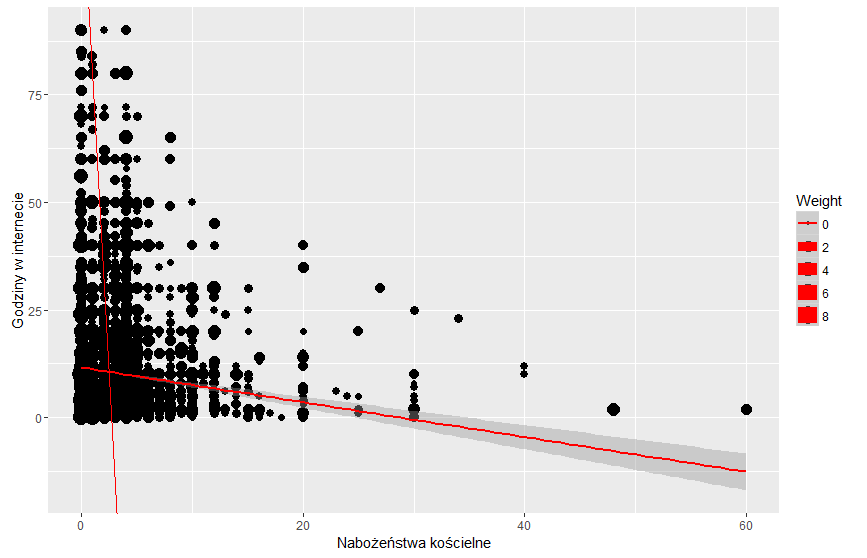
Small correlation tell us that these two questions don’t really have anything In common. They are for sure quite opposite questions and data. This set of data shows that correlation between chuch gatherings and using internet is

negligible.

Exercise 3:

Still using the same data:

Scatter plot with two regression lines:



As we can see two lines are In opposite directions, they don’t even converge, which means that data is not connected at All.

Residuals:

Ile godzin w ostatnim tygodniu korzystała Pani z internetu?

Min 1Q Median 3Q Max

-11.636 -7.636 -3.636 3.364 79.979

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 11.63648 0.14335 81.17 <2e-16 \*\*\*

tmp2$ep39 -0.40392 0.03805 -10.62 <2e-16 \*\*\*

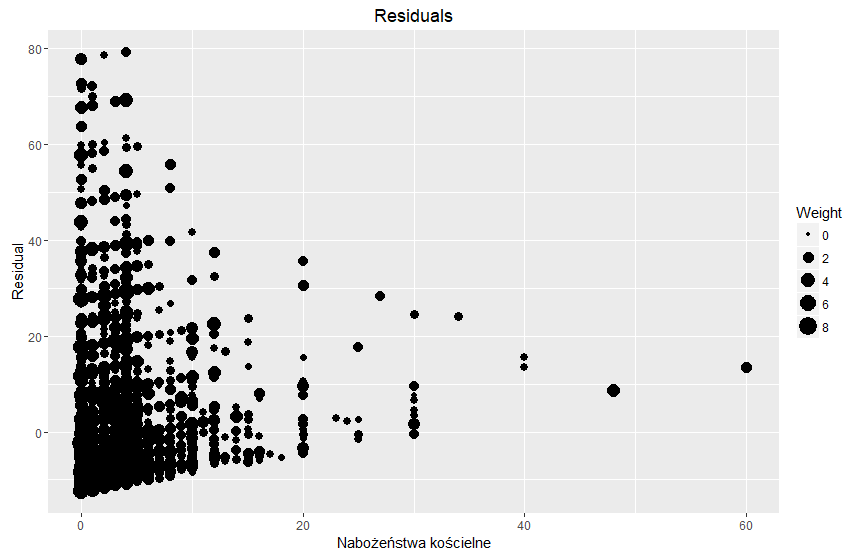
---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.69 on 11893 degrees of freedom

Multiple R-squared: 0.009388, Adjusted R-squared: 0.009304

F-statistic: 112.7 on 1 and 11893 DF, p-value: < 2.2e-16

Chart of the residuals:

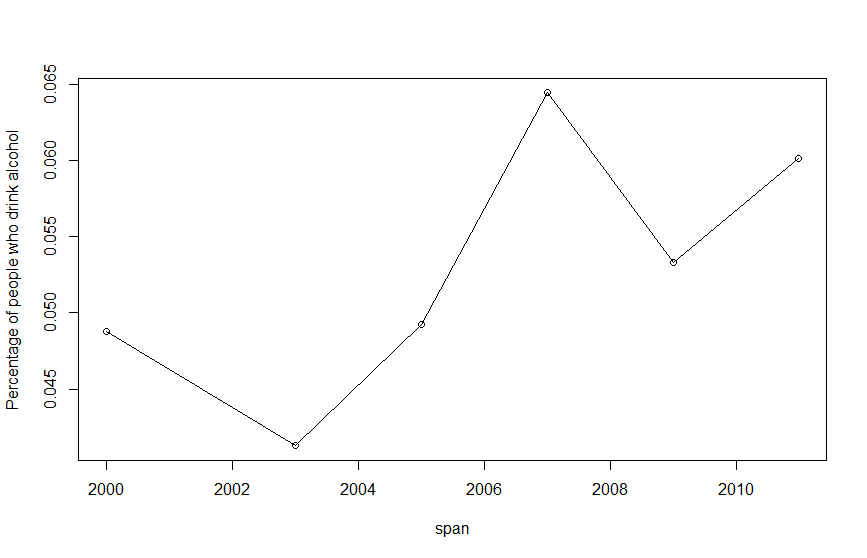
Values predicted by linear regression model have significant residuals. Based on the chart of residuals linear reggression model is worse choice compared to non-linear reggression model. This is also shown by R-squared equal to 0.009388

which proves that both data sets are independent from one another.

Exercise 4:

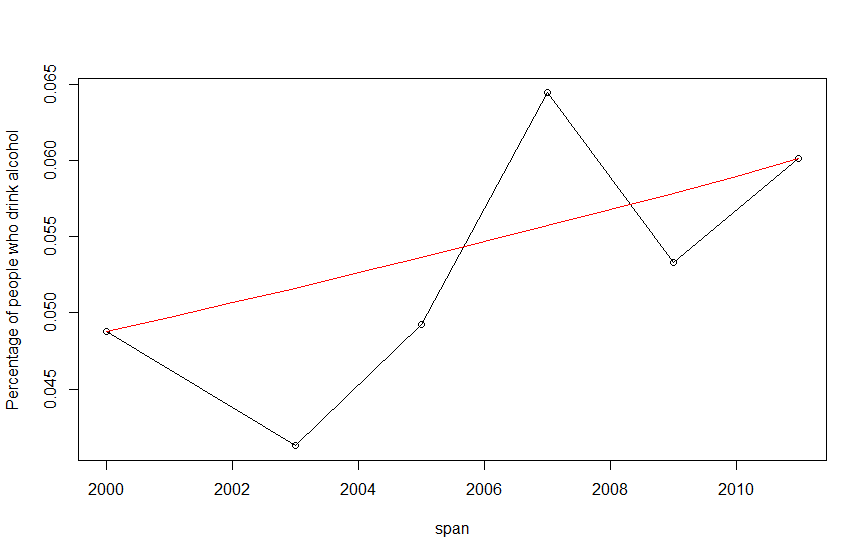
In this exercise I used question: “Czy uważasz, że pijesz za dużo alkoholu?”

I gathered 6 data points, years: 2000, 2003, 2005, 2007, 2009, 2011.



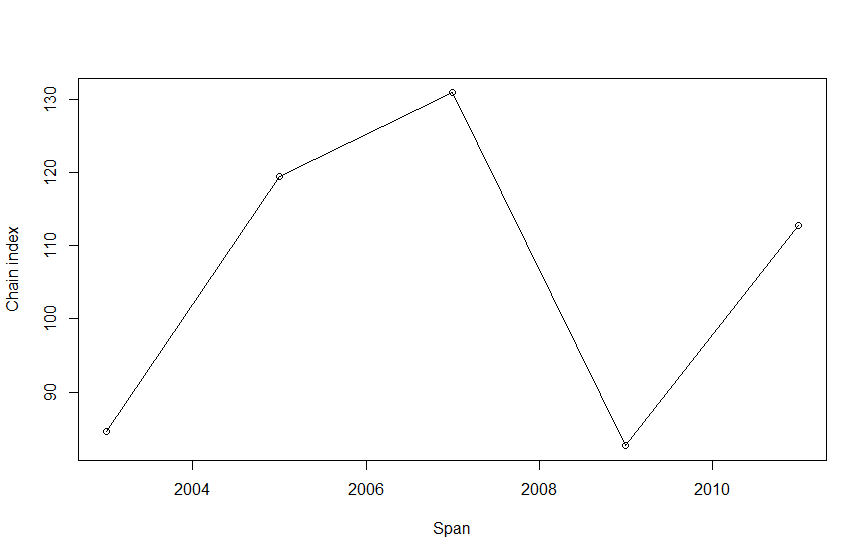
Unfortunately there is more and more alcoholics every year ☹.

CAGR plot:

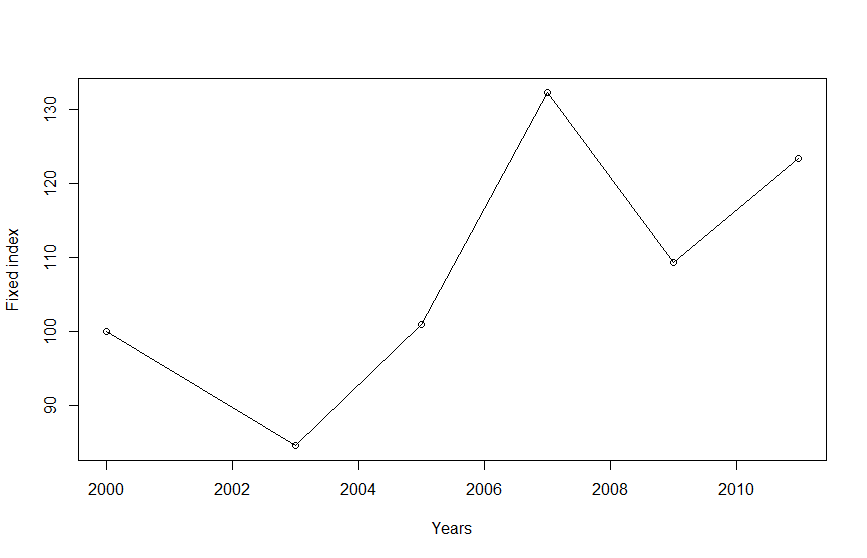


CARG = about 2%

Chain index:



Fixed index:



A chain index is an [index number](https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Index_number) in which the value of any given period is related to the value of its immediately preceding period and on that graph, we can see that after fluctuations in each year, finally the alcoholics are almost at the preceding period which is 100.

Group of people who drinks alcohol has increased by almost 25% since 2000.

The biggest growth is from 2003 to 2007, most probably because there was the biggest unemployment in Poland.