Exploratory Data analysis R

Pwani R Workshop

12th July 2024

Session objectives

1. Compute and interpret Pearson's r between two variables.
2. Set-up a simple linear regression model.
3. Conduct and interpret a chi-square test.

Data

- We will use data from the World Happiness Report.
- The csv file (WHR2018.csv) has been provided.

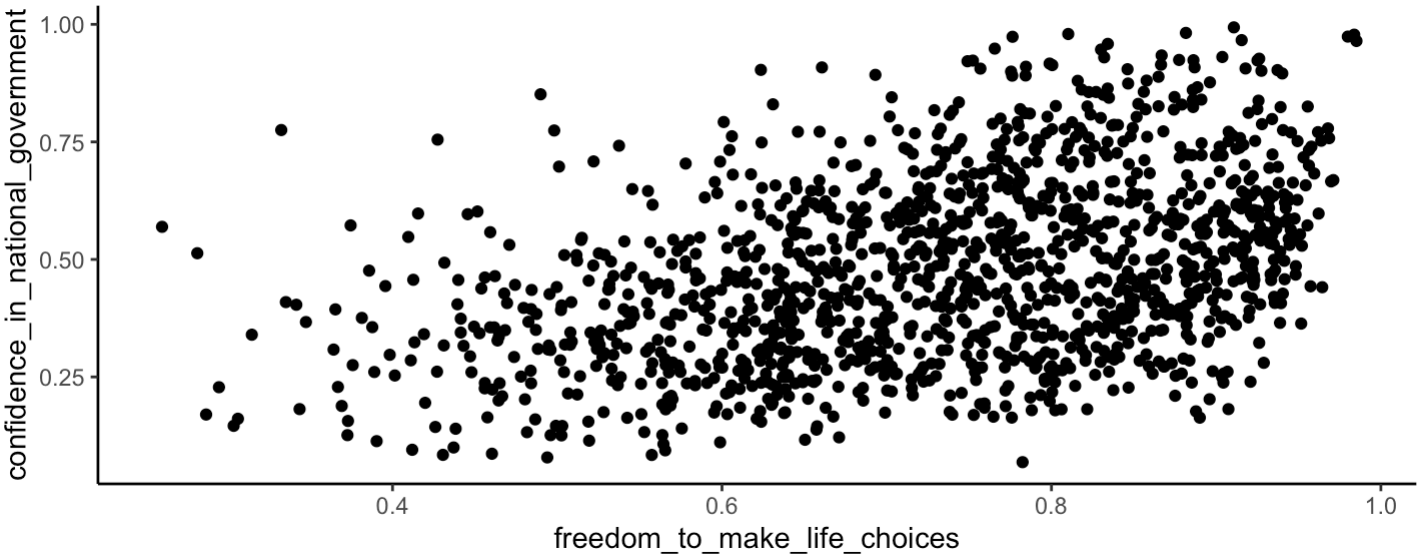
Linear Regression

Scatter with *happiness!*

What is the relationship between *Confidence in national government* and *Freedom to make life choices?*

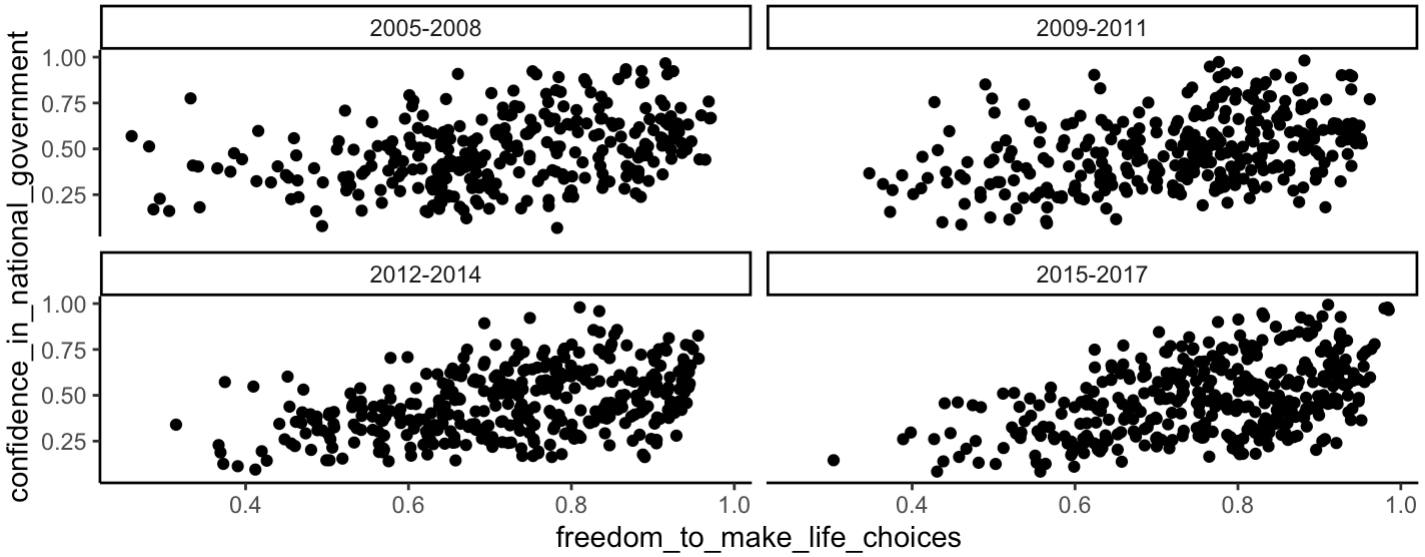
```
# import the happy dataset
happy <- import("../data/WHR2018.csv") %>%
  clean_names()

# plot the scatter plot
ggplot(happy, aes(x=freedom_to_make_life_choices,
                  y=confidence_in_national_government))+
  geom_point()+
  theme_classic()
```



Scatter with *happiness!* - multiple facets

```
# plot the scatter plot
ggplot(happy, aes(x=freedom_to_make_life_choices,
                  y=confidence_in_national_government)) +
  geom_point()+
  theme_classic() +
  facet_wrap(~year_grp)
```



cor for correlation

Correlation between *Confidence in national government* and *Freedom to make life choices?*

```
  cor(happy$freedom_to_make_life_choices,
      happy$confidence_in_national_government)
```

```
## [1] NA
```

This returns `NA` - why?

cor for correlation

Correlation between *Confidence in national government* and *Freedom to make life choices?*

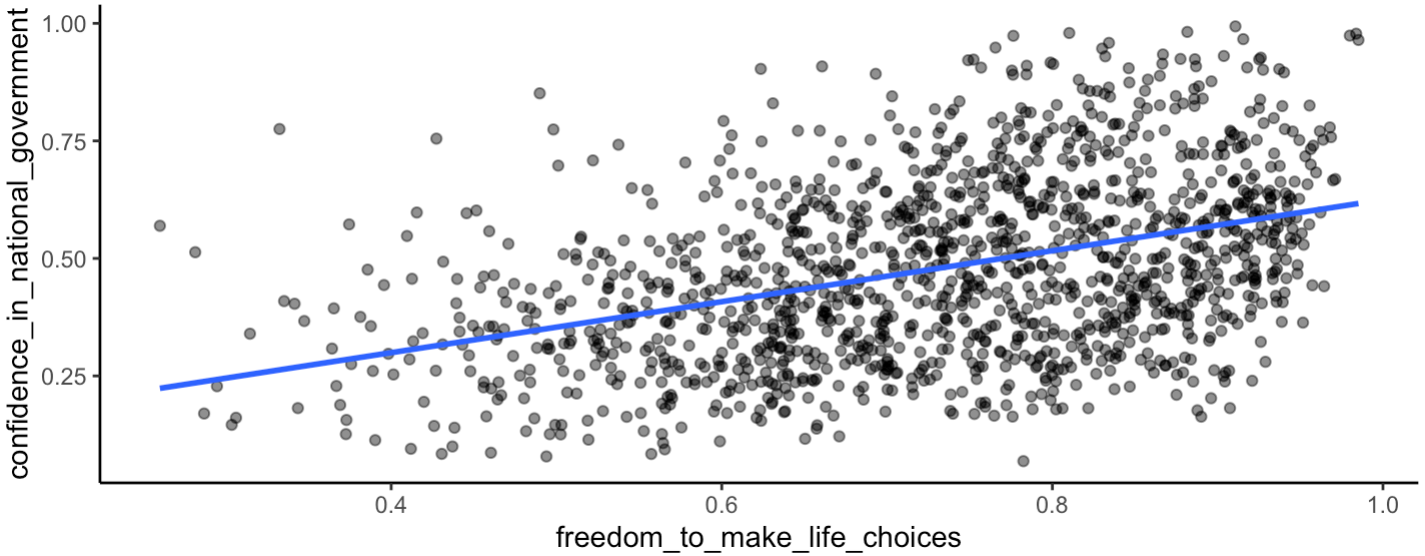
```
happy <- happy %>%
  filter(freedom_to_make_life_choices!="",confidence_in_national_government!="" )

  cor(happy$freedom_to_make_life_choices,
      happy$confidence_in_national_government)
```

```
## [1] 0.4080963
```

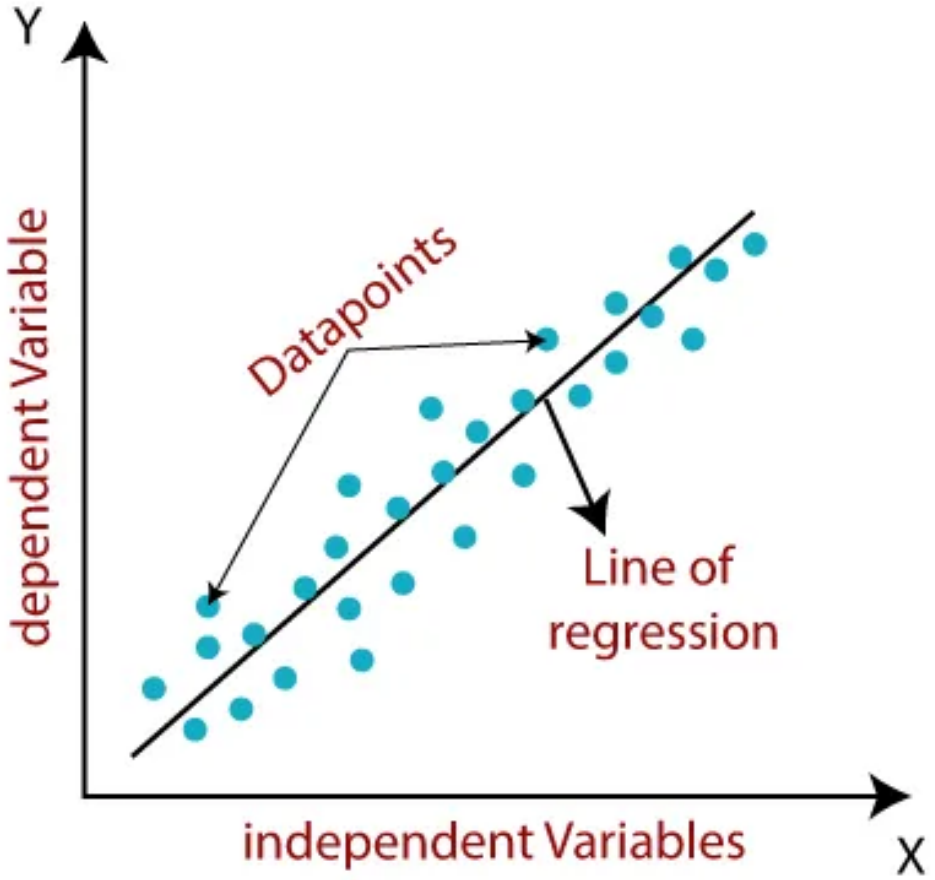
plot the line of best fit

```
ggplot(happy, aes(x=freedom_to_make_life_choices,
                      y=confidence_in_national_government)) +
  geom_point(alpha=.5) +
  geom_smooth(method=lm, se=FALSE) +
  theme_classic()
```



Linear regression model

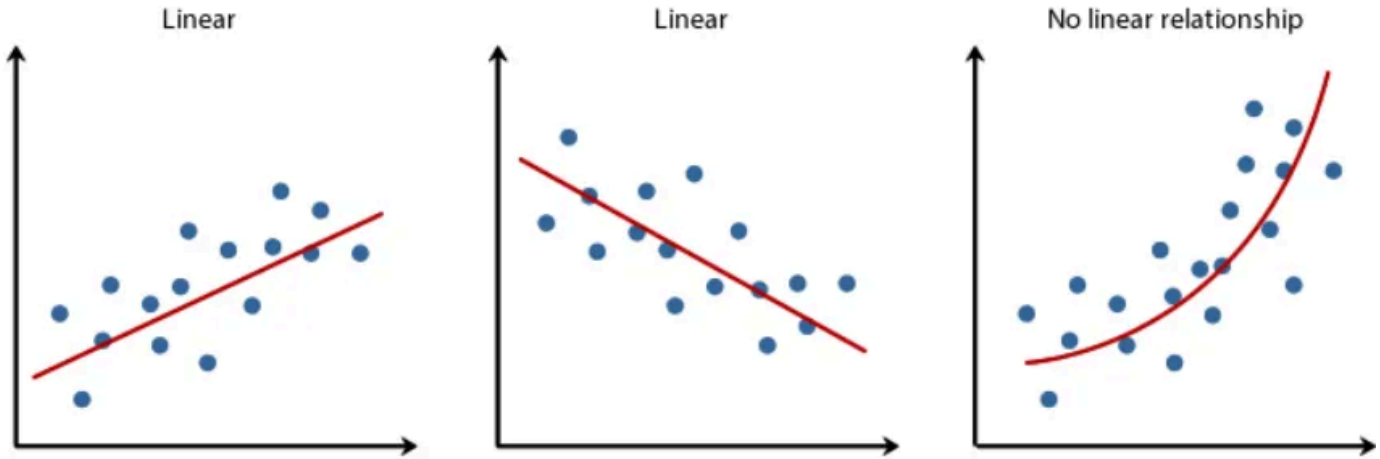
A **linear regression model** uses a straight line to describe the relationship between variables.



Linear regression model

Step 1: Make sure your data meets the assumptions

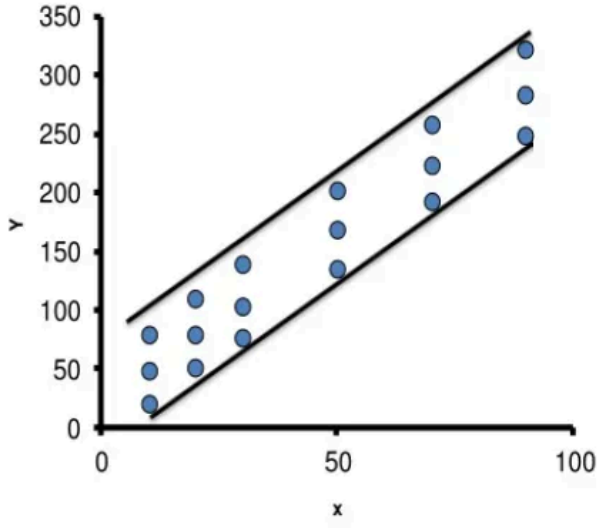
Linearity: the relationship betwen the outcome and explanatory variable must be linear. Test visually scatter plot.



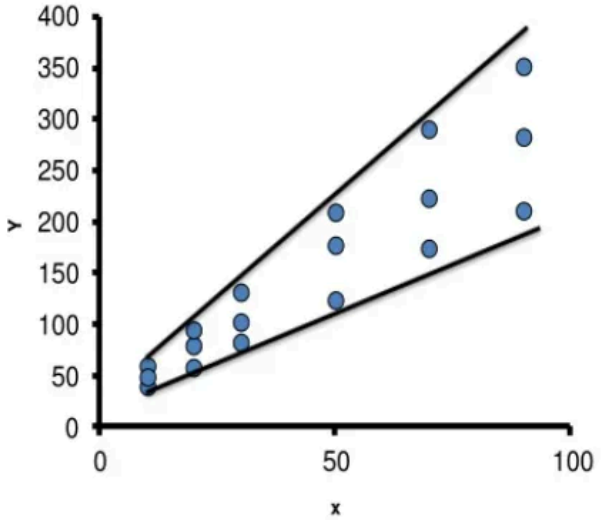
Linear regression model

Step 1: Make sure your data meets the assumptions

Homoscedasticity: variance of the residuals is constant across all levels of the independent variables.



homoscedasticity

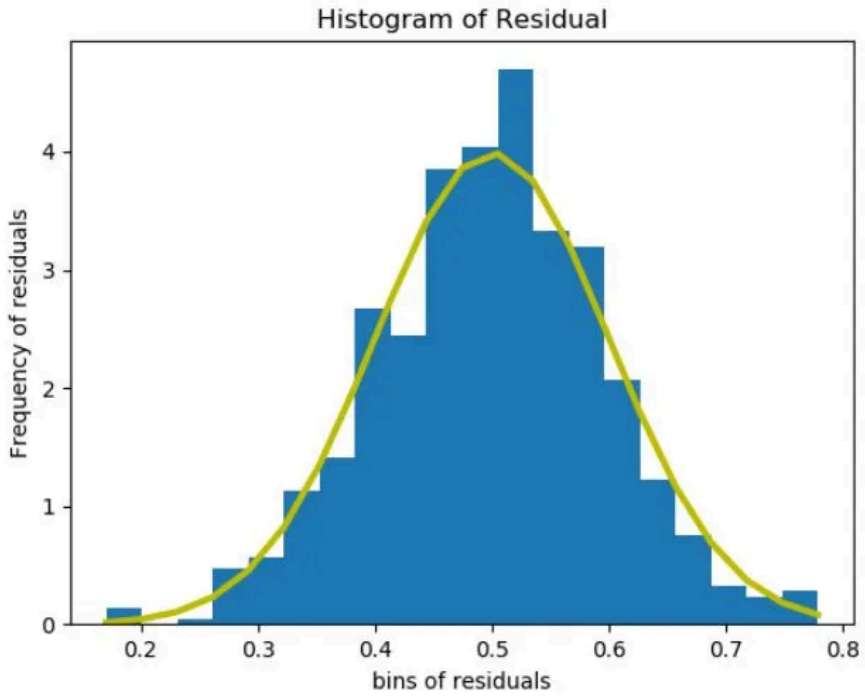


heteroscedasticity

Linear regression model

Step 1: Make sure your data meets the assumptions

Normality: The residuals, when plotted, follow a normal distribution.



Linear regression model

Step 1: Make sure your data meets the assumptions

Multicollinearity: this is when the independent variables are highly correlated with each other. Multicollinearity can make it difficult to analyze the individual effects of the features in predicting the target

Linear regression model

Step 2: Perform a linear regression analysis

- Two lines of code are needed to perform a linear regression analysis.

- The first line of code makes the linear model, and the second line prints out the summary of the model.

```
model_1 <- lm(freedom_to_make_life_choices ~ confidence_in_national_government, data = happy)

summary(model_1)
```

Linear regression model

```
model_1 <- lm(freedom_to_make_life_choices ~ confidence_in_national_government, data = happy)

summary(model_1)
```

```
##
## Call:
## lm(formula = freedom_to_make_life_choices ~ confidence_in_national_government,
##     data = happy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.49912 -0.08191  0.00183  0.10498  0.26737
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     0.584291   0.009495   61.54   <2e-16 ***
## confidence_in_national_government 0.307186   0.018412   16.68   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1309 on 1393 degrees of freedom
## Multiple R-squared:  0.1665, Adjusted R-squared:  0.1659
## F-statistic: 278.4 on 1 and 1393 DF,  p-value: < 2.2e-16
```

Chi-square

Chi-square

A Chi-square is used to examine whether there is an association between the row variable and the column variable.

Example 17.1

Table 17.1 shows the data from the influenza vaccination trial described in the last chapter (see Example 16.1). Since the exposure is vaccination (the row variable), the table includes row percentages. We now wish to assess the strength of the evidence that vaccination affected the probability of contracting influenza.

Table 17.1 2 × 2 table showing results from an influenza vaccine trial.

(a) Observed numbers.

	Influenza		
	Yes	No	Total
Vaccine	20 (8.3%)	220 (91.7%)	240
Placebo	80 (36.4%)	140 (63.6%)	220
Total	100 (21.7%)	360 (78.3%)	460

Source: Essentials of Medical Statistics (Betty R. Kirkwood)

Chi-square test in R

Step 1:Load dataset

In this example we have created a fictional dataset

```
# Sample data frame (replace this with your actual data)

data_summary <- data.frame(
  cough = sample(c("Yes", "No"), 100, replace = TRUE),
  gender = sample(c("Male", "Female"), 100, replace = TRUE)
)

head(data_summary)
```

```
##   cough gender
## 1    No Female
## 2    No Female
## 3    No   Male
## 4    No   Male
## 5    No Female
## 6   Yes Female
```

Chi-square test in R

Step 2: Create a contingency table

```
# Create the 2 by 2 table
contingency_table <- table(data_summary$cough, data_summary$gender)

contingency_table
```

```
##
##       Female Male
##   No      25   22
##   Yes     26   27
```

Chi-square test in R

Step 3: Perform the chi-square test!

```
contingency_table
```

```
##
##       Female Male
##   No      25   22
##   Yes     26   27
```

```
chisq.test(contingency_table)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  contingency_table
## X-squared = 0.045124, df = 1, p-value = 0.8318
```

What does this mean?