OQC

Oxford Q-Step Centre

# LAB SESSION 3

rose.degeus@politics.ox.ac.uk

OQC
Oxford
Q-Step
Centre

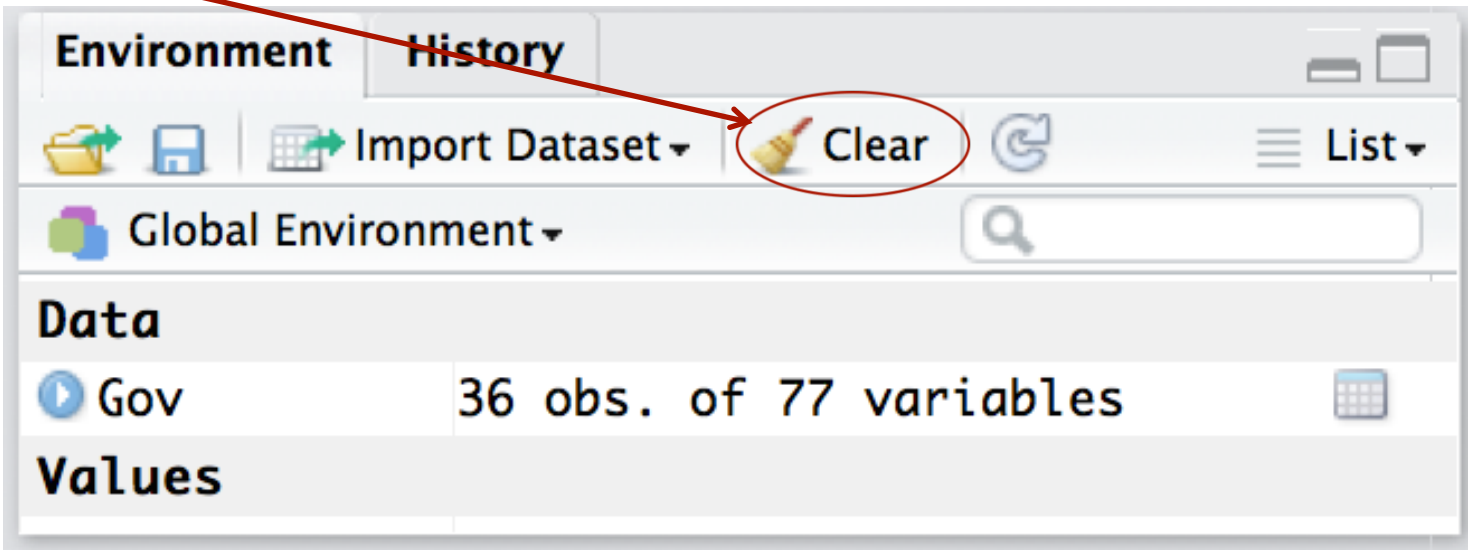
# Log-in details

| **Login:** | **Password:** |
| --- | --- |
| q-step-01 | q-step-01 |
| q-step-02 | q-step-02 |
| …… | …… |
| …… | …… |
| q-step-10 | q-step-10 |
| q-step-11 | q-step-11 |

OQC
Oxford
Q-Step
Centre

# First: clean your workspace!

- **Command:** CTRL + L
- **Or:**



OQC
Oxford
Q-Step
Centre

# Today

- Recap lab 2
- Homework Assignment
- Learning objectives lab 3
- Exercises:
  - Comparing samples (continuous/categorical)
  - Bivariate regression
  - Multivariate regression

OQC
Oxford
Q-Step
Centre

# 1. Recap lab 2

- Calculating the mean/median/standard deviation
- Percentiles + Ordering Data
- Missing Data
- Scatterplot

# 1. Recap Lab 2

OQC
Oxford
Q-Step
Centre

**Descriptive Statistics**

- Mean; median ; variance and standard deviation
- Percentiles and quantiles
- Creating scatterplots and boxplots

**Commands**

- Mean(); median (); sd(); percentile()
- Ordering data → data[order(Data$unemployment), c ("country", "unemployment")]

**Important**

- Dealing with missing data → na.rm = TRUE

# Homework Assignment

OQC
Oxford
Q-Step
Centre

- Create a scatterplot

```
plot(data$hdi_2010,data
$womens_parl_representation_2010, xlab="HDI
index",ylab="Women in parliament 2010") # gives axis
labels

text(data$hdi_2010,data
$womens_parl_representation_2010, labels = data
$country, cex = 0.7) # provides country labels
```

- What is the correlation?

```
cor.test(data$womens_parl_representation_2010,
data$hdi_2010)
```

# Homework

OQC
Oxford
Q-Step
Centre

- Interpretation of results
- Correlation Coefficient = ~0.39
  - (1) This is **positive**
  - (2) correlation ranges between -1 and 1; 0.39 can be considered moderate association
- P-value = 0.019
  - This is smaller than the cut-off point 0.05
  - We say 'statistically significant at 5 percent level'
  - This means → it is unlikely we would find this outcome (coefficient 0.39) if the **null-hypothesis were true [i.e. we would only in 5% of cases]**
  - **Null-Hypothesis =** no association
  - **Thus** – we are confident that the positive and moderate association between human development index and women's representation in parliament is reflective of the real world [statistical inference from sample to population]

**Tip:** when writing your essay; look carefully at other articles + books to see how they interpret / what language they use to describe significance and size of effects

**Conducting the analysis is step 1; writing an intelligible interpretation step 2!**

# 2. Learning objectives

OQC
Oxford
Q-Step
Centre

- 1. Significance testing
- 2. The means difference test (t-test) and the chi-square test of independence
- 3. Regression analysis
  - Bivariate
  - Multivariate

# 1. Significance testing

OQC
Oxford
Q-Step
Centre

- When? E.g. regression analysis, correlation test, means difference test (t-test)
- Always work with **hypotheses:**
  - **Null Hypothesis** = the hypothesis of no effect/no relationship
  - **Alternative Hypothesis** = there *is* a statistically significant relationship
- Two-tailed/one-tailed test:
  - There **is a difference** = two-tailed test
  - The level will be **higher** [lower] in new[old] democracies = one-tailed test

# P-Value

OQC
Oxford
Q-Step
Centre

- **P-value** = probability that quantifies the strength of the evidence against the null hypothesis in favour of the alternative
- We have the value of **the test-statistic**, and the **null distribution** of the test-statistic, and we want to see if the test-statistic is in the middle of the distribution, or in the tail
- Think about the p-value in terms of the following question: *if we would repeat the study (using the same sampling procedure), what is the probability that the new value we get is further out in the tail (for a one-tailed test)?*

# 2.1 Difference in mean scores: two-sample T-test

OQC
Oxford
Q-Step
Centre

- Statistical hypothesis test
- Means difference test → are two sets of data statistically significantly different from each other?
- NULL Hypothesis = there is no significant difference in the mean score of corruption perception between old and new democracies
- ALTERNATIVE Hypothesis = there **is a significant difference** in corruption perception between old and new democracies
  - There **is a difference** = two-tailed test
  - The level will be **higher** [lower] in new[old] democracies = one-tailed test

# The t-test

OQC
Oxford Q-Step Centre

```
> t.test(data$womens_parl_representation_1990,data$womens_parl_representation_2010)

        Welch Two Sample t-test

data:  data$womens_parl_representation_1990 and data$womens_parl_representation_2010
t = -5, df = 70, p-value = 2e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -17.9  -8.0
sample estimates:
mean of x mean of y
     12.0      24.9
```

Test statistic – the "T" statistic. The difference between means / difference standard errors

Measure of significance
Smaller P-value → the more strongly data contradict H0 [and thus we reject this]

Alternative hypothesis = there is a difference in means

Means for representation of women in parliament in 1990 and 2010 respectively

# 2.2 The chi-squared test of independence (i)

OQC
Oxford
Q-Step
Centre

- Remember → for continuous data we did a Pearson's Test Coefficient [correlation]
- Now → categorical data we use the **Chi-Squared Test of independence**
- H0 = var x and y are independent
- Ha = var x and y are interdependent / associated

# 2.2 The chi-squared test of independence (ii)

OQC
Oxford
Q-Step
Centre

- We compare the observed values with the expected values
- Expected values → the values we would expect to see if the null hypothesis is true (i.e. no association)
- Observed values → your data!
- Sum of the squared differences between the observed and the expected data / divided by the expected data
- → "Goodness of fit" how well does your data fit the data as expected from the null hypothesis?

Note: the Chi2 does not give any information about the **strength** of the association

# Chisq.test(representation.hdi)

OQC
Oxford
Q-Step
Centre

Pearson's Chi-squared test with Yates' continuity correction

Note we use the **table** that we created

data: representation.hdi

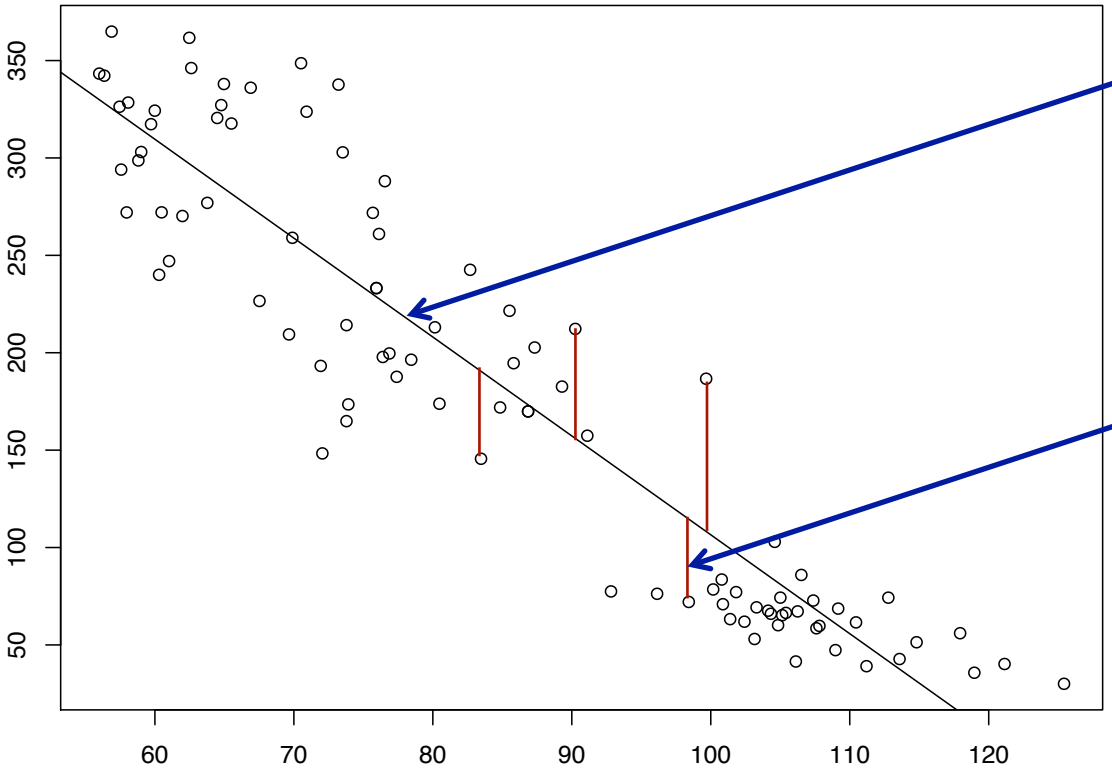
X-squared = 2.7778, df = 1, p-value = 0.09558

The test statistic; the X-Squared statistic

P-value = measure of significance.
How likely is it to receive the test – statistic (X-squared) under null hypothesis? If p = small then we reject the H0.

# 3. Regression (i)

OQC
Oxford
Q-Step
Centre




**Regression** = finding the "line of best fit" between the dependent variable "y" (or: response variable) and the independent variable

**Residual** = distance between line and observation (e.g. **red lines**)

# 3. Regression (ii)

OQC
Oxford
Q-Step
Centre

**S.E.** = estimate of the standard deviation of the sample mean (based on the population mean)

```
Call:
lm(formula = L$enviro_performance_index_2010 ~ L$exec_parties_1945_2010)

Residuals:
     Min       1Q   Median       3Q      Max
-23.5465  -5.4566  -0.5074   8.2585  22.1650

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                 69.076      1.877  36.810   <2e-16 ***
L$exec_parties_1945_2010     4.262      1.961   2.174   0.0372 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.9 on 32 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.1287,    Adjusted R-squared:  0.1014
F-statistic: 4.725 on 1 and 32 DF,  p-value: 0.03724
```

**T-value** = coefficient/s.e. → for t-test: compare t-value with critical value on t-distribution with degrees of freedom (n-1)

**P-value**: here - probability of observing a value larger than the t-value here, given the t-distribution with d.f. (n-1) (general: the probability of obtaining an effect that *is at least as extreme* as the estimate in your sample data, *assuming the null hypothesis is true*. I.e. A P-value evaluates how well the sample data support the idea that the null hypothesis is true.

**Constant/intercept** = expected mean value of DV when all IV's are zero

**F-test for the model:** null hypotheses = all coefficients are zero;
Alternative hypothesis: at least one of the coefficients is not zero

OQC
Oxford
Q-Step
Centre

# SEE YOU IN WEEK 8

☺