Solutions Autumn 2019

All questions are worth 10 points. Bullet points below each question give general grading instructions.

Exercise 1

a) We would conduct two tests for differences between proportions describing two populations. We could also conduct a chi square goodness-of-fit test with one population serving as the expected value, and the other two populations serving as the observed values.

Data tranformation: We would need to create a fraction of super happy people, by taking the number of people who answer 10 on the happiness question and dividing that by the number of observations. For example, for Norway this would be $\frac{262}{1541} = 0.17$

Then we would use the fraction, in order to compute the empirical standard deviation.

$$\sqrt{\frac{p(1-p)}{n}}) = \sqrt{\frac{0.17(1-0.17)}{1541}} \approx 0.01$$

- No test -5
- No data transformation - 5
- No Norway example -5
- Long argument with no clear intuition - 8

b) The fraction of cautious optimists for Norway are $\hat{p_1} = \frac{229+521+378}{1541} = \frac{1128}{1541} = 0.73$

The fraction of cautious optimists for Italy are $\hat{p_2} = \frac{655+732+299}{2591} = \frac{1686}{2591} = 0.65$

The pooled proportion estimate is $\hat{p} = \frac{1128+1686}{1541+2591} = \frac{2814}{4132} \approx 0.68$

The hypotheses of the test:

$$H_0 : (\hat{p_1} - \hat{p_2}) = 0$$

$$H_1 : (\hat{p_1} - \hat{p_2}) \neq 0$$

The test statistic:

$$z = \frac{(\hat{p_1} - \hat{p_2})}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}} = \frac{0.73 - 0.65}{\sqrt{0.68 \times 0.32 \times \frac{1541+2591}{1541 \times 2591}}} = \frac{0.08}{\sqrt{0.000225}} \approx 5.33$$

Answers may differ a bit depending on when you apply the rounding after the decimal comma.

$$z = 0.11 < z_{0.025} = 1.96$$

We reject the null hypothesis and find that Italians and Norwegians have different proportion of cautiously optimistic people

c)

$$H_0 : \frac{Var(Italy)}{Var(Norway)} = 1$$

$$H_1 : \frac{Var(Italy)}{Var(Norway)} \neq 1$$

The test statistic is:

$$F = \frac{1.66^2}{1.47^2} \approx 1.28$$

which is F-distributed with 2590,1541 degrees of freedom. The cutoff value is $F_{\infty,\infty,0.05} = 1$. We reject the null of equality of variances.

It is correct to use here also a two-tailed test with explicitly specified $H_1 : \frac{Var(Italy)}{Var(Norway)} \neq 1$. The cutoff value is the same.

Note: The larger variance goes in the numerator of the F-statistic. If the larger variance is in the denominator, then the student has to transform the cutoff value by taking the inverse and picking the cut-off value itself with flipped df - $\frac{1}{F_{\infty,\infty}} = 1$, which is the same.

We find that the variances of happiness between Norwegians and Italians are different. The happiness of Italians has a larger variance.

- Correct test, but with calculation mistake -5
- Puts the smaller variance on top, without mentioning the effect on the critical value -5
- Correct test, but wrong conclusion -7
- Correct test, but no conclusion -5
- Long argument with no clear intuition -8

d) This question should be solved with T-test for unequal variances:

$$H_0 : Mean(Italy) = Mean(Norway)$$

$$H_1 : Mean(Italy) \neq Mean(Norway)$$

The test statistic is:

$$T = \frac{7.18 - 8.12}{\sqrt{\frac{1.47^2}{1541} + \frac{1.66^2}{2591}}} = \frac{-0.94}{0.05} \approx -18.93$$

$$df = \infty$$

The cutoff value is $t_{\infty,0.025} = 1.96$, which means we reject the null hypothesis of no difference in means in favor of the alternative

The test indicates that Italians and Norwegians have different average happiness.

In fact, Norwegians are on average happier than Italians.

- Correct test, but with calculation mistake -5
- Correct test, but wrong conclusion -7
- Correct test, but no conclusion -5

e) See Table 1 for solution.

Table 1: Chi Square Goodnes-of-fit Table

Observed	Expected	Difference	Summation Component
20	236	-216	197.69
53	236	-183	141.90
161	236	-75	23.83
194	236	-42	7.47
454	236	218	201.37
295	236	59	14.75
482	236	246	256.42
377	236	141	84.24
146	236	-90	34.32
182	236	-54	12.36
2364			974.37

$\chi^2 = 972.72 > \chi^2_{0.05,9} = 16.9$

Therefore, we reject the null hypothesis of equality between the observed and expected frequencies. The happiness in Russia is not uniformly distributed.

f) Solution in Table 2.

Table 2: Chi Squared Test Contigency Table

	Norway	Italy	Russia	
Very Happy	640	481	328	1449
Cautious Optimists	811	1743	1154	3708
Rest	90	367	882	1339
	1541	2591	2364	6496
Expected Frequencies				
Very Happy	344	578	527	
Cautious Optimists	880	1479	1349	
Rest	318	534	487	
Summation Components				
	255	16	75	
	5	47	28	
	163	52	320	963

The cutoff value is $\chi^2_{0.05,4} = 9.49 < 782$. With a test statistics of 782 we reject the null hypothesis and conclude that there is evidence of a relationship between happiness and country origin.

- Correct test, but with calculation mistake -5
- Correct test, but wrong conclusion -7

- Correct test, but no conclusion -5

g) The regression equation looks like this:

$$Happy_i = \alpha + \beta_1 I(country = Norway)_i + \beta_2 I(country = Italy)_i + \varepsilon_i$$

$I(country = Norway)_i$ is an indicator variables for respondents from Norway. When the respondent is from Norway the variable takes value 1 and 0 otherwise.

$I(country = Italy)_i$ is an indicator variables for respondents from Italy. As above.

The reference category are respondents from Russia.

The significance and sign of coefficient β_1 will show whether average happiness between Russians and Norwegians is significantly different. Same for β_2, but between Italians and Russians. A positive and significant β_1 will show that Norwegians are happier than Russians.

The data would look like in Table 3. Important variables for students to include are from columns 2, 3, 4 and 6. Depending on the way they have written the model above, they can have 4 and 5 instead of 4 and 6 for example.

Table 3: Example Dataset

1	2	3	4	5	6
Respondent ID	Happy	Country	I (Norway)	I(Russia)	I(Italy)
1	5	NO	1	0	0
2	2	RU	0	1	0
3	9	IT	0	0	1
4	10	NO	1	0	0
5	1	RU	0	1	0

- Correct equation type, +3
- Correct variable and coefficient description, +4
- Correct data table, +3

h) From the correlation matrix we can see that there are no variables with strong correlation, foe example above an arbitrary number of 0.8 in absolute terms. From Table 5 it is not obvious that this multicollinearity impacts the estimation of the standard errors.

Looking at the variable descriptions, it seems that social and social_active could be capturing the same source of variation - namely how sociable a person is. People who are more sociable with respect to their peer group are also more likely to meet other people several times a week.

crime and safety are two other variables that seem similar - if you have been the victim of a crime you would not feel safe, which implies that there could be a strong negative correlation between the two variables.

- Argues clearly about lack of mechanical multicollinearity, +5
- Discusses conceptual collinearity, +5
- Long argument with no clear intuition - 8

i) There can be no issues of autocorrelation because from the data description it is obvious that this is a cross-section, therefore no time structure and no autocorrelation.

There can be issues of heteroscedasticity where the error terms for each country have different variance. Like we observe in the raw data, happiness in Italy has a larger variance than happiness in Norway. In Figure 1 we observe that the residuals are grouped and not centered around the 0. The variance (the range of the data points) is not constant around the 0, so heteroscedasticity could be a problem. From the histogram we observe that the residuals range from -6 to +4, so they are also not centered around the 0. However, the residuals do have a bell shape, so at least we have no problems with bimodal or other weird distributions.

Overall, the model does not seem to fit the data well.

- Correct intuition about autocorrelation, +3
- Correct intuition about heteroscedasticity, +3
- Good discussion of the plots, +4
- Overly long response with no clear idea, - 8

j) Easiest way to resolve this through a couple of point predictions:

Evaluate all variables at the their means except age:

$$X = 7.73 - 0.08 * 2.98 + 0.03 * 2.40 + 0.33 * 0.41 + 0.42 * 0.8 + 0.27 * 2.72 - 0.08 * 0.15 - 0.36 * 1.92 = 8.07$$

$$E(Happy|age = 15, X) = 8.07 - 0.14 \times log(15) = 7.91$$

$$E(Happy|age = 25, X) = 8.07 - 0.14 \times log(25) = 7.87$$

$$E(Happy|age = 50, X) = 8.07 - 0.14 \times log(50) = 7.83$$

$$E(Happy|age = 90, X) = 8.07 - 0.14 \times log(90) = 7.8$$

Overall, students should get a picture similar to Figure 1. It is important that the line is not straight, but slighly curved.

Interpretation: As people age their happiness decreases. If we increase age by 1 percent, we'd expect happiness to decrease by 0.0014 units (if happiness were a continuous variable). In addition, the non linearity implies that this negative effect is decreasing with age. The difference in happiness between ages 15 and 20 is larger than between 85 and 90.

- No intuition on the nonlinearity of the effect - 6
- No picture - 6
- Overly long response with no clear idea, - 8

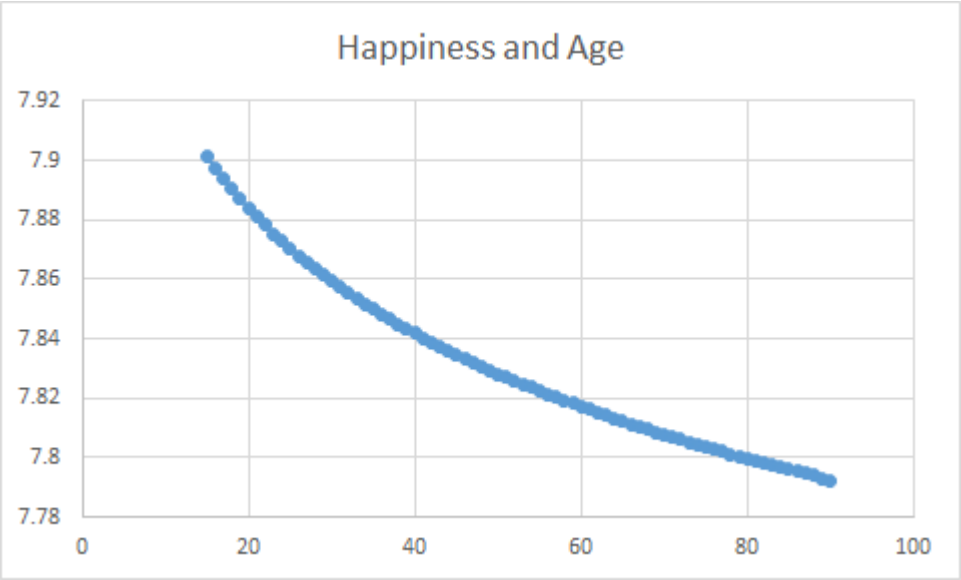

Figure 1

k) Here we require an interaction model:

$$Happiness_i = \alpha + \beta_1 difference\_income_i + \beta_2 fairness_i + \beta_3 female_i +$$

$$\beta_4 difference\_income_i \times female_i + \beta_5 fairness_i \times female_i + \varepsilon_i$$

Happiness is the dependent variable, difference_income, fairness and female are as described in Table 1.

We create the interaction variables $difference\_income_i \times female_i$ and $fairness_i \times female_i$.

A clear alternative is to estimate the regression model $Happiness_i = \alpha + \beta_1 difference\_income_i + \beta_2 fairness_i + \varepsilon_i$ for each gender. Then we would get estimates that are a function of the estimated coefficients in the interaction model.

l) 1st stage:

$$E(difference\_income|Female) = 2.9 + 0.14 \times 1 = 3.04$$

$$E(difference\_income|Male) = 2.9 + 0.14 \times 0 = 2.9$$

$$E(fairness|Female) = 2.42 - 0.05 \times 1 = 2.37$$

$$E(fairness|Male) = 2.42 - 0.05 \times 0 = 2.42$$

2nd stage:

$$E(happiness|Female) = 7.67 - 0.09 * 3.04 + 0.05 * 2.37 + 0.03 * 1 = 7.54$$

$$E(happiness|Male) = 7.67 - 0.09 * 2.9 + 0.05 * 2.42 + 0.03 * 0 = 7.53$$

Females seem to be more happy, although the difference is probably not significant.

- No predictions - 6
- No answer to the female vs male happiness question -4
- Overly long response with no clear idea, - 8

m) For Norway there could be two possible answers: 1) none of the lags seem to be significant and lag 6 could be just noise, so white noise process. Basically, happiness is not dependent on history, and each period is a new draw.

2) Happiness depends on happiness 6 years ago. This could be an AR(6) process with only the 6th lag significant.

For Italy: the first 4 lags are significant, it could be an AR(1) process with a high autocorrelation coefficient.

For Russia: only the first lag is significant. This could be an AR(1) process with low autocorrelation coefficient, or an MA(1) process. It could be also white noise. There is little dependence on history, similar to Norway.

- Overly long response with no clear idea, - 8

n) In this R output we are fitting 3 ARIMA models. The best one - the one with the lowest AIC- is Arima(France, order = c(0, 1, 2))). ARIMA(0,1,2):

$$\Delta Happiness_t = -0.6452\Delta\varepsilon_{t-1} - 0.0992\Delta\varepsilon_{t-2}$$

where Δ denotes the first differences.

Looking at the coefficients, the second MA term is not significant, so it probably can be dropped. Similarly examining the third model, one would also think that all the other terms expect the MA(1) are irrelevant because they are not significant.

$$\Delta Happiness_t = \phi\Delta\varepsilon_{t-1}$$

- Recognizing the best ARIMA, +2
- Writing down the equation, +3
- Figuring out that MA(1) is the best description of French happiness, +5
- Overly long response with no clear idea, - 8

o) The point of this question is to get an intuition about fixed effects.

From this assignment we learn that happiness is different in different countries (point f)). It seems like Norwegians are happier than Italians, who are happier than Russians, on average. However, once we look into the distributions we find that there are equal amounts of cautious optimists - people with above average happiness, but not super happy - between Norway and Italy. Happiness is more variable between the two countries, with larger variance for Italy. We observe within a lifetime, happiness is a different process for each country. Whereas Norway seems to be more of a white noise, it seems like for Russia and Italy happiness is dependent on history, for Italy and France even more so.

Therefore, looking at Table 8, it seems obvious that the fixed effect model would be a better way to model happiness. There are notable differences in happiness between

countries, and likely these difference reflect also fixed factors that we are not aware of or that are not measured. For e.g. air quality, urbanization, population concentration, informal social norms, etc. By taking fixed effects, we explain more of the variation in happiness. If there was an R^2 it would be increasing.

Another argument to make would be that obviously Russia and Norway have a different intercept for happiness (given that there is different average happiness), therefore in order to explain better happiness we need to take these different starting points into account.

- Overly long response with no clear idea, - 8