Final exam

total points: 100

- There are 3 exam problems, each worth 50 points. Pick **only 2 out of 3** that you would like to solve. Look through all the problems beforehand in order to select the ones you are most comfortable with. Your total score cannot be higher than 100, e. g. if you chose problems 1 and 3, you are not allowed to cherry-pick items from problem 2. State clearly at the beginning of your solution which problems you picked.
- Read each question carefully in order to avoid surprises: it can have multiple sub-questions.
- You are allowed to look up the session codes and use on-line search engines to find specific answers. However, keep in mind that even popular answers to some problems might be different from what is asked in the exam or even wrong (e. g. some people forget to make their standard errors robust).
- You are **not allowed to communicate** with other students or copy others' answers, i. e. **plagiarise**. In case virtually identical answers or suspiciously similar mistakes are found in the solutions of two or more students, they will be asked to clarify their answers on an individual basis. Failure to do so will automatically result in legal action carried out by the Ethics Committee of the university for the entire group of students.
- Your submission must be a single `.R` file with your last name, first name, and `exam`, all separated by hyphens. Five points will be deducted for wrong file naming.
  Example: `ivanov-ivan-exam.R`
  Solutions submitted in other formats (`.docx` files, screenshots, text in email body etc.) **will not be graded and will be awarded zero points!**
- If a question is theoretical, write your answer as a comment (starting with a pound sign `#`). If it contains code, write it simply as a code line to be run. If you are using any libraries, make sure you load all of them; there should be no '`could not find function "ggplot"`'-like errors. Delimit separate questions clearly.
  Example:

```
###### Problem 1-1
# The coefficient on beer tax is equal to 0.28 (0.21 rob. SE) and is statistically insignificant (at 5% level).

###### Problem 1-3
library(data.table)
psid <- fread("psid1990.csv")
```

- Do not hard-code your answers. Your solution should rely on the functions you learned during the course.

Hint. When interpreting the coefficients in regressions, keep in mind that depending on the presence of logarithms in dependent/explanatory variables, the interpretation changes (assuming X is exogenous):

- $Y = \alpha + \beta X + U$, `lm(Y ~ X)`: a change of X by 1 unit causes Y to change by β units;
- $\log Y = \alpha + \beta X + U$, `lm(I(log(Y)) ~ X)`: a change of X by 1 unit causes Y to change **approximately** by $(\mathrm{e}^{\beta} - 1) \cdot 100\%$.
- $Y = \alpha + \beta \log X + U$, `lm(Y ~ I(log(X)))`: a change of X by $p\%$ causes Y to change by $\beta \cdot \log(1 + p/100)$ units;
- $\log Y = \alpha + \beta \log X + U$, `lm(I(log(Y)) ~ I(log(X)))`: a change of X by 1% causes Y to change by $\beta\%$ (elasticity).

Problem 1 (50 points)

Use the data from the file `demography.csv` in order to make a demographic forecast of the global population structure. It contains the data for the year 2000 for women in 5-year age cohorts (`AgeCohort`).

- `Females`: the total number of females (in thousands) in that cohort,
- `Fertility`: the average number of girls born by a female in that cohort,
- `Mortality`: the probability of death for a female in that age cohort.

1. *(6 points)* We know that the average age in a cohort is the lower bound plus 2.5. We assume that the fertilities for every age within a cohort are equal. Compute the overall mean age of mother at birth, a: multiply the average age in each cohort by the probability of birth in that cohort (fertility rate divided by the sum of fertility rates) and add them up.
2. *(4 points)* Is the population with these fertility rates expected to increase or decrease (assuming constant fertility rates)? Answer this question in two steps:
   - At first, compute the average number of daughters per woman, d,
   - Then compute the average population percentage change per year, $d^{1/a}$.

   In how many years will it change twofold?
3. *(6 points)* Compute and plot in one graph the following series (with age along the horizontal axis):
   - The probability of survival at a certain age (one minus mortality),
   - The probability of survival up to a certain age (the cumulative product of survival probabilities you have just computed),
   - Fertility (given in the table),
   - Mortality (given in the table).

   Make sure the lines for the four series are visually distinct and the legend is legible.
4. *(16 points)* In this sub-question, you will compute 5-year demographic projections of the female population structure. Demographers regularly provides such forecasts for age pyramids: if the present dynamics are preserved, how many people of every age will live in a country in 100 years?
   Make a 21 $\times$ 21 data frame. Its columns should correspond to years—name the columns `y2000`, `y2005` ..., `y2100`. The rows should correspond to age cohorts—name them `age0_4`, ..., `age100_104`. The cells will contain numerical values showing the population of a certain age group in a certain year. Copy the `Females` column from the source data frame into the first column (`y2000`). Fill this data frame using the following algorithm:
   - The number of newborn girls in the first cohort (`age0_4`) in the next 5-year time period is the expected number of girls produced by all females who were alive at the beginning of the previous period (with given fertilities).
   - The number of females in each subsequent age cohort (i. e. the next row) is equal to the number of females in the previous cohort in the previous period multiplied by the probability of survival at that age.
5. *(6 points)* Plot these projections as coloured lines ostensively showing the change of the population structure throughout the years (age groups denoted by 0, 5, ..., 100 on the x axis, population on the y axis). Give a short summary and an interpretation of the changes that are going to happen by the year 2100 compared to 2000 in this model.
6. *(12 points)* Consider the following alternative scenario.
   - The fertility rates for cohorts '15–19', ..., '45–49' all go up by 0.03, i. e. become equal to (0.049 0.165 0.345 0.318 0.158 0.058 0.031).

   Compare the population structure in the year 2100 from (5) with the structure from this scenario and make a conclusion.

Problem 2 (50 points)

Many governments try to implement measures to reduce spirits consumption in order to improve some social indicators, such as vehicle fatality rate. The data at your disposal (`fatality.csv`) contain observations from 1982 to 1988 on 48 American states: vehicle fatality rates per capita (`mrall`), spirits consumption in gallons per capita (`spircons`), presence of a breath test law (`breath`), and other variables.

1. *(3 points)* Load the data from `fatality.csv` using the `fread` function from the `data.table` package. How many observations does it contain?
2. *(6 points)* Create a histogram with bin width 0.1 and a kernel density plot (smooth histogram) with bandwidth 0.037 with a rug for the variable `spircons`. Label the axes properly, add a title. Use any plotting theme that does not have background colour but has subtle axis lines instead.

3. *(7 points)* Using the `dplyr` package, select only the columns `spircons`, `year`, and `state`, and columns with names beginning with '`mr`'. Using the `summarise` function, compute the number of observations and the average mortality rate per state. Does every state have the same number of observations? Then, using the same function, compute the number of observations and the average mortality rate per year. Does every year have the same number of observations?
4. *(3 points)* Generate a new variable in the data frame, `fatalityrate`, based on `mrall`, equal to fatality rate per 10,000 people, and provide its summary statistics, including the standard deviation.
5. *(6 points)* Plot the two variables, `spircons` (x axis) and `fatalityrate` (y axis) in a scatter plot. Make it colourful: colour the points by the variable `state` (observations from the same state should have the same colour). Do not forget to label the axes and to add a title and a legend to the plot.
6. *(5 points)* Calculate the correlation matrix (a set of pairwise correlations) for variables `fatalityrate`, `spircons`, and `breath`. Give a brief interpretation of the results.
7. *(6 points)* Estimate a linear regression where `fatalityrate` is the dependent variable and `spircons` and `breath` are the explanatory variables. Report the coefficients and heteroskedasticity-robust standard errors. Is the effect of spirits consumption on fatality rate significant, ceteris paribus? Is the breath test law efficient? Do your findings agree with general logic?
8. *(3 points)* Test the joint significance of `spircons` and `breath`. Do not forget to supply the `vcov. = vcovHC` argument!
9. *(7 points)* Now, add `year` and `state` factor variables into the model in order to control for unobserved heterogeneity across time and states. Report heteroskedasticity-robust standard errors. What can you say about the effect of the spirits consumption and breath test law now? Are they individually and jointly significant after controlling for time and state effects? Are the time and state effects jointly significant?
10. *(4 points)* Write a short summary of your findings about the impact of spirits consumption on vehicle fatality rate and provide an explanation for the differences between estimates in these models. Which one, in your opinion, is the best one for policymakers?

Problem 3 (50 points)

The goal of this analysis is to estimate the classical Mincer wage equation. The data set `psid1990-exam.csv` contains observations on Americans for the year 1990. The variables present are:

- `age`: age in years;
- `male`: 1 for males, 0 for females.
- `empl`: current employment status equal to 1 for working now, 2 for temporarily laid off, 3 for unemployed looking for work, 4 for retired, 5 for permanently disabled, 6 for housewives, 7 for students, 8 for other.
- `educ`: total years of schooling.
- `postgrad`: 1 for holders of a post-graduate degree, 0 otherwise.
- `inc`: taxable money income of an individual (censored at 999,999 from above).
- `tran`: income from social transfers (censored at 99,999 from above).
- `hours`: work hours.
- `bwnorm`: 1 if birth weight was above 2.5 kilograms, 0 otherwise (a proxy for health of a respondent at birth).
- `children`: number of children as of 1990.
- `marr`: total number of marriages in one's life.
- `currmarr`: current marital status equal to 1 for married, 2 for never married, 3 for widowed, 4 for divorced, 5 for separated.

1. *(5 points)* Which share of the population (in %) are receiving social transfers? Compute the descriptive statistics for two sub-samples of people: for those who receive social transfers (`tran > 0`) and for those who don't. Are these two sub-samples different in terms of education and hours worked? Is there anything strange about the data values?

2. *(7 points)* Compute the total income, `totinc`, equal to money income plus social transfers. How many observations have negative or zero values of total income? Select and keep only the observations with strictly positive total income in the data frame. Also, select and keep only the observations where the variables `marr`, `currmarr`, or `children` do not contain missing values. Finally, generate the variable `ltotinc` equal to the logarithm of total income.
3. *(5 points)* Produce histograms of `totinc` and `ltotinc` in one plot.
4. *(6 points)* Using the `dplyr` package, sort the observations by age and compute the average `totinc` and `ltotinc` by age, as well as the standard deviation of these variables in each group. Plot these averages (on the y axis) versus age (on the x axis) in two line plots without any background colour.
5. *(4 points)* Compute the correlation between the number of children and income. Is there any correlation between the number of children and hours worked? Give an explanation.
6. *(4 points)* Calculate the correlation between income and birth weight separately for males and females. Give a brief explanation.
7. *(7 points)* Estimate a linear regression with `ltotinc` as the dependent variable, and the number of years of schooling as the explanatory variable, including the following control variables: logarithm of work hours, age, age squared, gender, total number of marriages, number of children, presence of a post-graduate degree, employment status (factor), marital status (factor), and healthy birth weight indicator. Report robust standard errors. What is the effect of education on income, ceteris paribus? Is it significant? Does its sign agree with the theory?
8. *(2 points)* Are males or non-males better paid, according to this model, ceteris paribus? Is the difference in wages between males and non-males significant?
9. *(6 points)* Now we suspect that the effect of the number of kids is non-linear. Create a new factor variable `kidcat` equal to `"0"` if `children` takes the value 0, `"1"` if `children` is equal to 1, `"2_3"` if `children` is 2 or `3`, and `"4+"` if `children` is greater or equal to 4. Replace the `children` variable in the model from (7) with `kidcat` and estimate a new model. Test the hypothesis that having 1, 2, or 3 kids does not have an effect on the total income.
10. *(5 points)* Output a simple text table comparing the models from questions 7 and 9 using the `stargazer` package and the `stargazer` command with arguments `type = "text"` and `se` equal to the list of heteroskedasticity-robust standard errors. Copy and paste it into your script file as a comment (to comment out a block of text in RStudio, use the hot key Ctrl + Shift + C). Keep it as is, do not spend time making it any prettier.