Exam question 1.

You have been asked to study the correlation between studying years and salary. To help you we have gathered information from 3010 individuals in USA year 1976 from which you are going to pick out a random sample of 500 individuals. You read though the data and choose your sample with these commandos:

```
rm(list = ls())
set.seed(personalnumber)
library(Ecdat)
data("Schooling")
df <- Schooling df <- df[sample(nrow(df), 500) , ]
```

To help you we have these variables:
wage76: The individuals salary in cent(1/100 dollar).
ed76: Studying time of the individual in years.
daded: total studying years for dad
momed: total studying years for mom

let X be ed76 and Y be wage76

1. (6p) Solve $SS_{XX} = \sum_{i=1}^{500} (X_i - \overline{X})^2$, $SS_{YY} = \sum_{i=1}^{500} (Y_i - \overline{Y})^2$ and $SS_{XY} = \sum_{i=1}^{500} (X_i - \overline{X})(Y_i - \overline{Y})$
2. (4p) With the answers from the previous question solve the Pearson correlation coefficient
3. (4p) Estimate the following linear regression model: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$. Interpret the estimated coefficient $\widehat{\beta_1}$
4. (6p) Solve a hypothesis test with the null hypothesis of $\beta_1 = 25$ with 5% significance level. it's not necessary to state any presumptions because they are assumed to be met.
5. (6p) One potential extradited variable in the model for equation 1 is the environment at home. It would be reasonable to think that individuals that have highly educated parents i) would be more prone to get high education and ii) have easier to get higher salary with help of their parents network of contacts. In such a case the regression above would lead to "omitted variable bias". Explain and motivate in which direction such a bias would go and answer if the result of 1.3 is a probable overestimation or underestimation of β_1
6. (3p) Estimate a new model where both mom and dads study level is included. What happens with the estimation of β_1 . Did the change go in the direction you thought it would?
7. (8p) The model of equation 1 above implies that one extra year of studying is associated with an absolute increase of salary per hour. It might be more fair to think that an increase in study years will increase the salary with a certain percentage. Formulate an alternative model that would enable such an interpretation. Estimate this model and interpret the estimated slope of the coefficient.

Exam question 2.

To solve this question we will use quarterly data from the UK regarding income. You read the data and create necessary data with the following commandos:

```
rm(list = ls())
 set.seed(personnummer)
library(Ecdat)
data("IncomeUK")
df <- data.frame(IncomeUK)[1:56, ]
df$t <- seq(1:dim(df)[1])
df$qtr <- rep(1:4, length=nrow(df))
```

The following variables is available:
income: Aggregated income in millions of pounds
t: A variable that gets the number 1 for first quarter 1971, 2 for second quarter 1972….. until it reaches 56 for the fourth quarter of 1984.
qtr: A variable that gets number 1 for first quarter, 2 for second quarter, 3 for third quarter and 4 for fourth quarter.

A potential regression model: Y = $\beta_0 + \beta_1 q1 + \beta_2 q2 + \beta_3 q3 + \beta_4 t + \varepsilon,$

In the model q1 is a dummy variable that indicate quarter 1(same with q2 and q3)

1. (7p) Create the variables q1, q2, q3 and estimate the regression model above. With regards to the point estimation, create a new forecast for income for first quarter 1985(t=57) and second quarter 1985(t=58).

2. (6p) The actual values for the first and second quarter were 56 727 and 59 790 million pound. Use these values to calculate root mean square error (RMSE) and the mean absolute error (MAD) for the prognosis.

Exam question 3

There was a survey 1980 regarding household spending in Spain. Read this code to get the information.

```
rm(list = ls())
set.seed(personnummer)
library(Ecdat)
data("BudgetFood")
df <- BudgetFood df <- df[sample(nrow(df), 5000) , ]
```

In the code the following variables are available:

- wfood - percentage of the households total expenses that goes to groceries
- totexp - TOTAL expenses for the household per month(pesetas)
- age - age of the contact person in the household
- size - size of the household
- town - the size of the town divided into 5 categories (1 smallest and 5 largest)
- sex - gender of contact person in household

The survey is a stratified sample survey done with proportional allocation. The stratum variable is town. Assume that the selection frame is based on a register of households and independent random selection without return has been done from each stratum. You can also assume that the population size is very large relative to the sample selection and ignore in this question the finality correction

1. (2p) Create a variable foodexp, that indicates household spending on groceries per month in pesetas.
2. (4p) Create a figure that illustrates household spending on groceries per month divided depending on town size.
3. (4p) Create a table with sample size, sample mean, medians and standard deviation for the sample.
4. (15p) Calculate with 5% significance level if the size of the towns has a connection with the total spending on groceries for a household per month.
5. (2p) Indicate two aspects that affects how well a stratification works in terms of efficiency
6. (5p) Estimate the average household spending per month that goes to groceries in pesetas.
7. (10p) Do an interval estimation with 95% confidence level on the average household spending on groceries per month. Interpret the interval.

Exam question 4.

Assume that you have a sample of N = 20000 individuals and you want to estimate the average value for the variable X. You know that the variable is normally distributed. Though reading the code down below you get the population sample deviation, Sigma 2 for earlier years, which you can use to calculate the sample deviation.

```
rm(list = ls())
set.seed(personnummer)
sigma2 <- round(100*runif(1)) + 1
```

1. (8p) calculate the necessary sample size if the 95% confidence interval can at maximum be 5 values wide.