Modelling income per state

There are many classical data sets available in `R`, and one of them can be found by typing `state.x77`. This is an object in the "datasets" package of `R`, and it contains data about the stetes of the US from the year 1977. We are interested in making a model for the income per capita of each state, where the covariates will be the total population of the state, its area, the illiteracy of the state and its murder rate. Define the objects `Income, Pop, Area, Illit` and `Murder`, for example by

```
Income = state.x77[,"Income"]
```

We start by considering the model

$$\texttt{Income}_i = \beta_0 + \beta_1 \texttt{Pop}_i + \beta_2 \texttt{Area}_i + \beta_3 \texttt{Illit}_i + \beta_4 \texttt{Murder}_i + U_i,$$

where i represents the state and the U_i's are all independent and have a $N(0, \sigma^2)$ distribution.

(a) Fit the model, writing a script using the command

```
fit=lm(Income~Pop+Area+Illit+Murder)
summary(fit)
```

Give an estimate of σ^2.

(b) Plot the residuals, which can be found in `fit$residuals`, against `Area`. Furthermore, make a histogram of the residuals. Proof that the mean of the residuals equals 0 (remember that the vector of residuals in the linear model is given by $Y - X\hat{\beta}$). Argue why it may be a good idea to take the logarithm of the Area as a covariate.

Define the object `logArea` as the logarithm of `Area`. Refit the model using `logArea` instead of `Area`.

(c) In this new model, test the hypothesis that `logArea` and `Murder` are not significant (so their corresponding variables are 0), using a significance level of 5%.

Now fit the model

$$\texttt{Income}_i = \beta_0 + \beta_1 \texttt{Illit}_i + \beta_2 \texttt{Pop}_i + U_i. \tag{1}$$

(d) Mike wonders if it would be a good idea to add another covariate, namely $\texttt{Illit2} = \texttt{Illit}^2$ (the square of the illiteracy). Test an appropriate hypothesis to see if Mike is right.

(e) We are still using model (1). A few years after the data was collected, researchers found that based on new data, a good value for (β_1, β_2) would be $(-500, 0.05)$. Test whether the value of the parameters in 1977 differed significantly from these new values, using a significance of 5%.