Instructions:

- Your R code shall be written in a single script file.
- Name the script file `lab2_[your_id].R`. For example, `lab2_5123700044.R`.
- Separate your answers/code into sections according to `task_id` and `part_id`.

Task 1 (16 points)

This task is about running linear regression in R. The dataset `debonding.txt` is about cathodic debonding of elastomeric metal bonds. The variables measured were:

`debonding`	the amount of debonding measured in centimetres
`time`	time used measured in minutes
`voltage`	voltage used measured in volts
`ph`	pH at time of bonding
`temp`	temperature measured in Fahrenheit

(a) (1 point) Use R function `read.table` to load the contents of the data file `debonding.txt` into a R data frame called `deb.df`.

(b) (1 point) Use R function `lm` to construct the full model, name it as `deb.full.LM`,

$$\texttt{debonding} = \beta_0 + \beta_1\texttt{time} + \beta_2\texttt{voltage} + \beta_3\texttt{ph} + \beta_4\texttt{temp} + \varepsilon$$

(c) (1 point) What is the proportion of variability in `debonding` explained by our model?

(d) (1 point) Use R to plot the standardised residual $\hat{e}'_i$ against the fitted value $\hat{y}_i$.

(e) (1 point) Use R to plot the residual $\hat{e}_i$ against the previous residual $\hat{e}_{i-1}$.

(f) (1 point) Use R to plot the estimated autocorrealtion function of $\hat{e}_i$, i.e. the acf plot.

(g) (1 point) Use R to plot the sample quantiles of $\hat{e}_i$ against the theoretical quantiles of

$$\text{Nornal}(0, 1)$$

(h) (1 point) Use R to plot the sample quantiles of $\hat{e}'_i$ against the theoretical quantiles of

$$\text{Nornal}(0, 1)$$

(i) (1 point) Perform a Box-Cox transformation on the response by choosing an appropriate λ from its 95% confidence interval, name the transformed response as `debonding.bc`.

(j) (1 point) Construct the following model, name it as `deb.full.bc.LM`,

$$\texttt{debonding.bc} = \beta^*_0 + \beta^*_1\texttt{time} + \beta^*_2\texttt{voltage} + \beta^*_3\texttt{ph} + \beta^*_4\texttt{temp} + \varepsilon$$

Is there any evidence that working with the transformed response is better?

(k) (1 point) Suppose assumptions are satisfied, what does F-test say about the models.

(l) (1 point) Suppose assumptions are satisfied, comment on the significance of `ph` in the presence of the other three regressors.

(m) (1 point) Use R to find the 95% confidence interval of β_1.

(n) (1 point) Use R to find the 95% confidence interval of the mean `debonding` when

$$\texttt{time} = 30.00, \quad \texttt{voltage} = 1350, \quad \texttt{ph} = 4.00, \quad \texttt{temp} = 300$$

(o) (1 point) Use R to find the 95% prediction interval of `debonding` when

$$\texttt{time} = 30.00, \quad \texttt{voltage} = 1350, \quad \texttt{ph} = 4.00, \quad \texttt{temp} = 300$$

(p) (1 point) Construct the following two, name them as `deb.no.v.LM` and `deb.no.v.p.LM`

$$\texttt{debonding} = \beta_0^{**} + \beta_1^{**}\texttt{time} + \beta_3^{**}\texttt{ph} + \beta_4^{**}\texttt{temp} + \varepsilon$$
$$\texttt{debonding} = \beta_0^{***} + \beta_1^{***}\texttt{time} + \beta_4^{***}\texttt{temp} + \varepsilon$$

Suppose the residual analysis has provided no evidence that model assumptions are violated by the two models above. Is there anything in the output of `summary(deb.no.v.LM)` and `summary(deb.no.v.p.LM)` that can be used as the basis for deciding which of the two models is better? If so, how to use it. And can we use it to compare amongst all four models?

Task 2 (4 points)
This task is about categorical regressors and partial F-test.

(a) (1 point) Load `sim_lab2.csv.bz2` into a R data frame, name it as `sim.df`, then run

```
> sim.LM = lm(y~.+x1*x3+x1*x4+x2*x3+x2*x4+x3*x4, data = sim.df)
> summary(sim.LM)
```

Explain why there are `NA`s in the summary output.

(b) (1 point) Suppose there is no evidence against model assumptions being satisfied. According to our model, is the conditional mean of Y significantly different for `x4=B` and `x4=C` given other regressors take the same value. Justify your answer.

(c) (1 point) Read the help documentation on `anova.lm`, then comment on the meaning of having a large p-value in the following output

```
> anova(sim.LM)
```

```
Analysis of Variance Table

Response: y
            Df    Sum Sq   Mean Sq    F value Pr(>F)
x1           1 388292353 388292353 60154.2397 <2e-16 ***
x2           1  36089461  36089461  5590.9782 <2e-16 ***
x3           3     16007      5336     0.8266 0.4790
x4           2  20440965  10220482  1583.3568 <2e-16 ***
x1:x3        3     30274     10091     1.5633 0.1961
x1:x4        2   6467655   3233828   500.9845 <2e-16 ***
x2:x3        3     17517      5839     0.9046 0.4380
x2:x4        2     25772     12886     1.9963 0.1360
x3:x4        4      6405      1601     0.2481 0.9109
Residuals 3978  25677774      6455
```

(d) (1 point) What can we learn from the following?

```
> source("~/Desktop/lab2_func.R")
> f.vec = replicate(1e4, sim.p.func(n=200)); sim.plot()
```