MATH 423/533 - ASSIGNMENT 2

To be handed in not later than 11:59pm, 24th October 2017.
Please submit your solutions with relevant R code snippets included as a pdf file via myCourses.
Please also upload a separate file containing your entire code as an .R script.

The following data gives data on average public teacher annual salary in dollars, recorded in the data frame `salary` as the variable `SALARY`, and spending (`SPENDING`) per pupil (in thousands of dollars) on public schools in 1985 in the 50 US states and the District of Columbia.

The objective of the analysis is to understand whether there is a relationship between teacher pay, y, and per-pupil spending, x. An analysis in `R` is presented below: some of the output has been deleted and replaced by `XXXXX`.

```
 1 > salary<-read.csv('salary.csv',header=TRUE)
 2 > x1<-salary$SPENDING/1000
 3 > y<-salary$SALARY
 4 > fit.Salary<-lm(y~x1);summary(fit.Salary)
 5
 6 Call:
 7 lm(formula = y ~ x1)
 8
 9 Residuals:
10     Min      1Q  Median      3Q     Max
11 -3848.0 -1844.6  -217.5  1660.0  5529.3
12
13 Coefficients:
14             Estimate Std. Error t value Pr(>|t|)
15 (Intercept)  12129.4      XXXXX   10.13 1.31e-13 ***
16 x1            3307.6      311.7   10.61 2.71e-14 ***
17 ---
18 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
19
20 Residual standard error: XXXXX on 49 degrees of freedom
21 Multiple R-squared:  0.6968,    Adjusted R-squared:  0.6906
22 F-statistic: XXXXX on 1 and 49 DF,  p-value: 2.707e-14
```

In answering the following questions, you may not use the `lm` function or its result on these data (or the functions `coef()`, `residuals()` etc.), but instead should use vector and matrix calculations.

(a) Write `R` code to verify the calculation of the entries in the `Estimate` column, and show that your code produces the correct results. 10 Marks

(b) Write `R` code to compute the value of the omitted entry for the `Residual standard error` on line 20. 10 Marks

(c) Compute the value of the entry in the `Std. Error` column on line 15 first using entries already given in the table, and then using the data directly. 10 Marks

(d) The entry for `Multiple R-squared` on line 21 is computed using the formula

$$R^2 = \frac{\mathrm{SS_R}}{\mathrm{SS_T}}$$

where $\mathrm{SS_R}$ is the 'regression sum-of-squares' and $\mathrm{SS_T}$ is the 'total sum of squares' as defined in lectures. Write `R` code to verify the calculation of R^2. 10 Marks

(e) Prove for a simple linear regression that, in the notation from lectures,

$$\text{SS}_\text{R} = \widehat{\beta}_1 S_{xy}$$

and show this result holds numerically for the salary data. 10 Marks

(f) The `F-statistic` on line 22 is computed using the sums-of-squares decomposition

$$\text{SS}_\text{T} = \text{SS}_\text{Res} + \text{SS}_\text{R}$$

and uses the formula

$$F = \frac{\text{SS}_\text{R}/(p-1)}{\text{SS}_\text{Res}/(n-p)}$$

where here $p = 2$ for simple linear regression. Write `R` code to compute the omitted value for F.

10 Marks

(g) In the notation from lectures, we have that the sums-of-squares decomposition can be written

$$\mathbf{y}^\top(\mathbf{I}_n - \mathbf{H}_1)\mathbf{y} = \mathbf{y}^\top(\mathbf{I}_n - \mathbf{H})\mathbf{y} + \mathbf{y}^\top(\mathbf{H} - \mathbf{H}_1)\mathbf{y}.$$

Show, mathematically and numerically, that

$$\text{trace}(\mathbf{I}_n - \mathbf{H}_1) = n - 1 \qquad \text{trace}(\mathbf{H} - \mathbf{H}_1) = p - 1$$

for this example, where $p = 2$ for simple linear regression.

20 Marks

(h) Using residual plots, assess the validity of the assumptions underlying the least squares analysis. Verify numerically the orthogonality results concerning the residuals, that is, in vector form

$$\mathbf{1}_n^\top \mathbf{e} = 0 \qquad \mathbf{X}^\top \mathbf{e} = \mathbf{0}_p \qquad \widehat{\mathbf{y}}^\top \mathbf{e} = 0.$$

(i) Using the fitted model, predict what the average public teacher annual salary would be in a state where the spending per pupil is \$4800.

5 Mark

(j) The prediction at an arbitrary new x value, x_1^{new} can be written in terms of the estimates $\widehat{\beta}$ as

$$\widehat{y}^{\text{new}} = \mathbf{x}_1^{\text{new}}\widehat{\beta} = [1 \;\; x_1^{\text{new}}]\,\widehat{\beta} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1^{\text{new}}$$

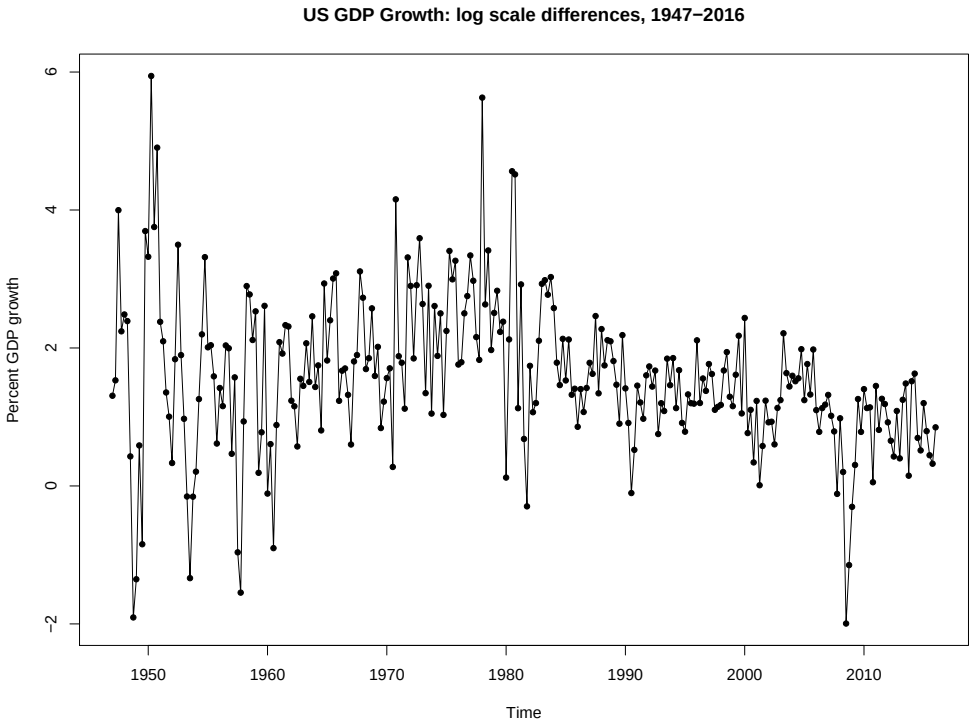
with $\widehat{\beta}$ the least squares **estimate**. Compute the *estimated standard prediction error* for $\widehat{y}^{\text{new}}$, that is, the square root of the estimated variance of the corresponding random variable

$$\widehat{Y}^{\text{new}} = \mathbf{x}_1^{\text{new}}\widehat{\beta} = [1 \;\; x_1^{\text{new}}]\,\widehat{\beta}$$

now with $\widehat{\beta}$ the least squares **estimator**, if x_1^{new} is \$4800. 15 Marks

EXTRA QUESTION FOR STUDENTS IN MATH 533

The figure below plots the percent differences on the log scale between successive recorded quarterly Gross Domestic Product (GDP) values in the US between the first quarter of 1947 and the first quarter of 2016 (277 data points).



The data may be read in from the file `US-GDP.txt` as follows. For regression purposes, we define the predictor x_1 by considering time (in quarters) since Q1, 1947.

```
1 y0<-scan('US-GDP.txt')
2 y<-100*log(y0[-1]/y0[-278])
3 x1<-c(1:277)
```

Is there any statistical evidence that there is a 'changepoint' in the GDP series at the year 1980 (when $x_1 = 133$), that is, that the relationship between y and x_1 prior to Q1 1980 is different from the relationship after that time ? Investigate this possibility using straight line regression modelling (**not** a single simple linear regression), and report the result of an appropriate hypothesis test.

25 Marks