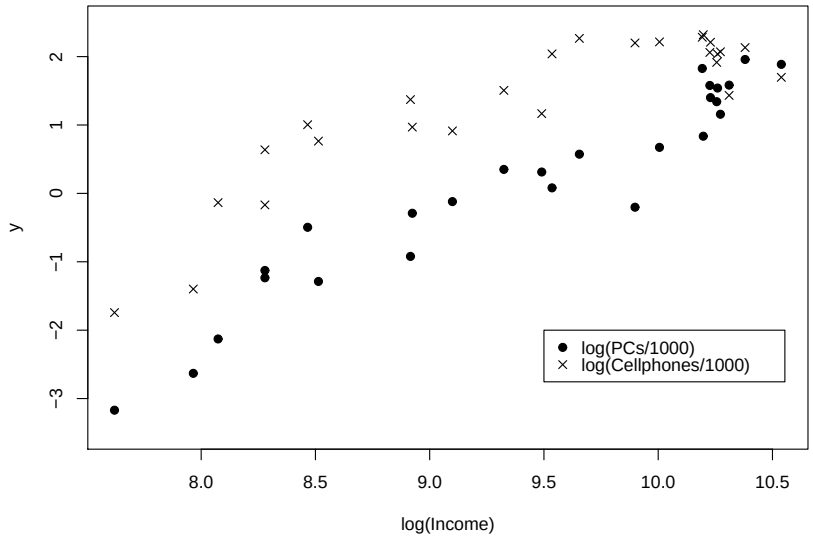
Q2 The following data reflect the purchasing power of individuals in 26 countries recorded in 2003. For each country the following variables were recorded in the data set

- `Cellphone` – the number of cellphones per 1000 individuals;
- `PCs` – the number of personal computers (PCs) per 1000 individuals;
- `Pcapincome` – the per capita income (in US Dollars)
- `ln.PCs` – the natural logarithm of the number of personal computers (PCs) per 1000 individuals (y_1);
- `ln.Cellphone` – the natural logarithm of number of cellphones per 1000 individuals (y_2);
- `ln.Income` – the natural logarithm of per capita income (x_1).



The objective of the analysis is to understand the role of per capita income (measured on the log scale), x_{i1}, in predicting the purchasing of these consumer electronics. An analysis in `R` is presented below:

```
> fit.PC<-lm(y1~x1);summary(fit.PC)    ##PCs
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -14.24314    0.87646  -16.25 1.87e-14 ***
x1            1.52633    0.09264   16.48 1.38e-14 ***
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1

Residual standard error: 0.418 on 24 degrees of freedom
Multiple R-squared:  0.9188,    Adjusted R-squared:  0.9154
F-statistic: 271.5 on 1 and 24 DF,  p-value: 1.383e-14

> fit.C<-lm(y2~x1);summary(fit.C)  ##Cellphones
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -8.9189     1.1211  -7.955 3.49e-08 ***
x1            1.0847     0.1185   9.154 2.68e-09 ***
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1

Residual standard error: 0.5347 on 24 degrees of freedom
Multiple R-squared:  0.7773,    Adjusted R-squared:  0.7681
F-statistic: 83.79 on 1 and 24 DF,  p-value: 2.684e-09
```

Question continued on the next page.

(a) Summarize the conclusions that can be made from the analysis concerning the relationship between per capita income and the amount of PCs and cellphones purchased per 1000 individuals. Make specific reference to line numbers when citing evidence to support you conclusions.

Is per capita income a better predictor of PC purchasing, or cellphone purchasing ? Justify your answer. 6 Marks

(b) Using the analyses above, predict the rate (per 1000 individuals) of PC and cellphone purchasing in a country whose per capita income is 8000 US dollars. 4 Marks

(c) Using the `predict` function in `R`, *confidence* and *prediction* intervals for the number of PCs (on the log scale) for a country whose per capita income is 2500 US dollars can be computed:

```
24 > newincome.data<-data.frame(x1=log(2500))
25 > predict(fit.PC,newdata=newincome.data,interval='confidence')
26        fit       lwr       upr
27 1 -2.301033 -2.649868 -1.952198
28 > predict(fit.PC,newdata=newincome.data,interval='predict')
29        fit       lwr       upr
30 1 -2.301033 -3.231621 -1.370445
```

For simple linear regression, the computation of the intervals is carried out using the following formulae

$$\widehat{y}^{\text{new}} \pm t_{\alpha/2,n-2} \times \widehat{\sigma}\sqrt{\left(\frac{1}{n} + \frac{(x^{\text{new}} - \overline{x})^2}{S_{xx}}\right)} \qquad \widehat{y}^{\text{new}} \pm t_{\alpha/2,n-2} \times \widehat{\sigma}\sqrt{\left(1 + \frac{1}{n} + \frac{(x^{\text{new}} - \overline{x})^2}{S_{xx}}\right)}$$

respectively. Explain the difference between these two intervals, in particular, state which interval is wider and explain why. 2 Marks

(d) At which value of log income would the lengths of the confidence and prediction intervals be at their shortest ? Justify your answer.

2 Marks

(e) Show that in a linear regression model fitted using least squares, the variance-covariance matrix of the residual random variable vector $\mathbf{Y} - \widehat{\mathbf{Y}}$ is the $(n \times n)$ matrix

$$\text{Var}_{\mathbf{Y}|\mathbf{X}}[\mathbf{Y} - \widehat{\mathbf{Y}}|\mathbf{X}] = \sigma^2(\mathbf{I}_n - \mathbf{H})$$

where, in standard notation, $\mathbf{H}$ is the 'hat matrix', and $\mathbf{I}_n$ is the $(n \times n)$ identity matrix.

4 Marks

Hence deduce the form of the estimated variance of the ith residual, $E_i = Y_i - \widehat{Y}_i$.

2 Marks

Answer to Q2:

(a) Income is a reasonable/good predictor of purchasing in both consumer products. Lines 10 and 22 indicate that the R^2 measures are very high (0.9188 for PCs) and quite high (0.7773 for cellphones) respectively. In both models, the coefficient β_1 is estimated to be significantly positive (line 5 for PCs, 17 for cellphones), indicating that as income increases in a country, purchasing power for the two consumables is also increased. The estimated slope coefficients are significantly different, with a higher slope for PCs.

Going by the R^2 statistics (lines 10 and 22), income is a better predictor of PC purchasing.

6 Marks

(b) We have

$$y_1 : \widehat{y}_1 = -14.24314 + 1.52633 \times \ln(8000) = -0.5256773$$

$$y_2 : \widehat{y}_2 = -8.9189 + 1.0847 \times \ln(8000) = 0.8298659$$

Check:

```
 1 > newincome.data<-data.frame(x1=log(8000))
 2 > predict(fit.PC,newdata=newincome.data)
 3          1
 4 -0.5256773
 5 > predict(fit.C,newdata=newincome.data)
 6         1
 7 0.8298659
```

4 Marks

(c) Using the `predict` function in `R`, *confidence* and *prediction* intervals for the number of PCs (on the log scale) for a country whose per capita income is 2500 US dollars can be computed:

```
 8 > newincome.data<-data.frame(x1=log(2500))
 9 > predict(fit.PC,newdata=newincome.data,interval='confidence')
10        fit       lwr       upr
11 1 -2.301033 -2.649868 -1.952198
12 > predict(fit.PC,newdata=newincome.data,interval='predict')
13        fit       lwr       upr
14 1 -2.301033 -3.231621 -1.370445
```

The second (prediction) interval is wider as it factors in the variability due to the random residual error that would be present in any future observed value. The width of the prediction interval is determined by the variance of the fitted value plus the residual error variance.

2 Marks

(d) They are shortest at the mean of the log predictor values; this is because the prediction variance is an increasing function at $(x^* - \overline{x}_1)^2$

2 Marks

(e) We have that

$$\mathbf{Y} - \widehat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\widehat{\beta} = \mathbf{Y} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y} = (\mathbf{I}_n - \mathbf{H})\mathbf{Y}$$

so the variance of this random variable is, by the general result for linear transformed random variables,

$$(\mathbf{I}_n - \mathbf{H})\sigma^2\mathbf{I}_n(\mathbf{I}_n - \mathbf{H}) = \sigma^2(\mathbf{I}_n - \mathbf{H})$$

as $(\mathbf{I}_n - \mathbf{H})$ is idempotent. Hence the form of the estimated variance of the ith residual is the diagonal element

$$\widehat{\sigma}^2(1 - h_{ii})$$

6 Marks

Q1 (a) Describe the least squares procedure for fitting a simple linear regression model to a sample of data $\{(x_{i1}, y_i), i = 1, \ldots, n\}$. 5 Marks

(b) Show that the least squares criterion for estimating β in simple linear regression is equivalent to requiring *orthogonality* between the columns of design matrix $\mathbf{X}$ and the residual vector $\mathbf{e} = \mathbf{y} - \mathbf{X}\widehat{\beta}$.
5 Marks

(c) Show that the least squares estimates $\widehat{\beta}$ and the *fitted values*, $\widehat{y}_i$, can each be written as linear combinations of the original response data $\mathbf{y} = (y_1, \ldots, y_n)^\top$.
4 Marks

(d) The following `R` output records the analysis of a small sample of data using simple linear regression. Some output has been removed and replaced by `XXX`.

```
 1 > summary(lm(y~x))
 2 Coefficients:
 3             Estimate Std. Error t value Pr(>|t|)
 4 (Intercept)      XXX      1.7799   0.839   0.4293
 5 x             0.3635      0.1388    XXX     XXX
 6 ---
 7 Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1
 8
 9 Residual standard error: 1.704 on 7 degrees of freedom
10 Multiple R-squared:  0.4948,    Adjusted R-squared:  0.4227
11 F-statistic: 6.857 on 1 and 7 DF,  p-value: 0.03448
```

From the output identify

(i) the sample size n; 1 Mark

(ii) the `Estimate` omitted from line 4; 1 Mark

(iii) the conclusion of the test of the null hypothesis

$$\mathrm{H}_0 \ : \ \beta_1 = 0$$

(against the usual two-sided alternative hypothesis) which is omitted from line 5 – this test is carried out at the $\alpha = 0.05$ level, and the corresponding t-test critical value is $t_{0.025,7} = 2.3646$.
1 Mark

(iv) whether `x` is a useful predictor of `y`; 1 Mark

(v) the three terms in the sums of squares decomposition

$$\mathrm{SS}_\mathrm{T} = \mathrm{SS}_\mathrm{Res} + \mathrm{SS}_\mathrm{R}.$$

2 Marks

State clearly which pieces of output you use when giving answers.