STATS 3001 Statistical Modelling III
Assignment 1
Due: 4pm Monday 19th March (Week 4), 2018

`########################################## solutions ##########################################`

IMPORTANT In keeping with the university policy on plagiarism, you should read the University Policy Statement on Academic Honesty (plagiarism, collusion and related forms of cheating):

http://www.adelaide.edu.au/policies/230.

Assignments must be submitted with a signed Assessment Cover Sheet. These forms are available on MyUni/Canvas under `Modules`→`Assignment cover sheet`. Please note that assignment marks cannot be counted for your assessment unless a signed declaration is received.

Check off the following prior to submitting your assignment:

□ Sufficient working has been provided in each question to satisfactorily demonstrate to the marker that you understand the required concepts and steps in the question.

□ All R output and plots to support your answers are included where necessary.

□ A coversheet is attached to the submission that is completed and signed.

□ Answers are written on their own paper, not on the assignment handout.

□ The submission is neat and provides space for marker's comments.

□ The submission is stapled together.

□ {
- Submit your assignment into the Statistical Modelling III hand-in box on Level 6 (Ingkarni Wardli building).
- Late assignments will only be accepted by prior agreement with the Course Co-ordinator and relevant requests should usually be accompanied by a medical certificate.

.

---

(1) Let A and B be matrices as given in Tutorial 1, Question 4. Furthermore, assume C is a constant $n \times n$ matrix. Prove that

$$\text{tr}(ABC) = \text{tr}(BCA)$$

[**Hint**: use the result from Q4, T1.]

Solution:
We know from T1, Q4 that $\text{tr}(XY) = \text{tr}(YX)$. Let $X = AB$ and $Y = C$. Then

$$\text{tr}(ABC) = \text{tr}(CAB)$$

Now let $X = CA$ and $Y = B$. Then

$$\text{tr}(CAB) = \text{tr}(BCA),$$

as required.

[Total: 4]

(2) Let $\boldsymbol{Y}$ be a random $n \times 1$ vector with mean $\boldsymbol{\mu}$ and variance matrix $\boldsymbol{\Sigma}$, and let $\boldsymbol{A}$ be a constant $n \times n$ matrix. Show that

$$E(\boldsymbol{Y}^T \boldsymbol{A} \boldsymbol{Y}) = \operatorname{tr}(\boldsymbol{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^T \boldsymbol{A} \boldsymbol{\mu}.$$

Solution:

We know that the quadratic form $\boldsymbol{x}^T A \boldsymbol{x} = \sum_{i=1}^n \sum_{j=1}^n x_i a_{ij} x_j$, so that

$$\boldsymbol{Y}^T A \boldsymbol{Y} = \sum_{i=1}^n \sum_{j=1}^n Y_i a_{ij} Y_j$$

Also recall $\operatorname{cov}(Y_i, Y_j) = \sigma_{ij} = E(Y_i Y_j) - \mu_i \mu_j$. Then

$$E(Y_i Y_j) = \sigma_{ij} + \mu_i \mu_j$$

Then

$$\begin{aligned} E(\boldsymbol{Y}^T A \boldsymbol{Y}) &= \sum_{i=1}^n \sum_{j=1}^n a_{ij} E(Y_i Y_j) \\ &= \sum_{i=1}^n \sum_{j=1}^n a_{ij} (\sigma_{ij} + \mu_i \mu_j) \\ &= \sum_{i=1}^n \sum_{j=1}^n a_{ij} \sigma_{ji} + \sum_{i=1}^n \sum_{j=1}^n \mu_i \mu_j a_{ij} \\ &= \operatorname{tr}(A\boldsymbol{\Sigma}) + \boldsymbol{\mu}^T \boldsymbol{A} \boldsymbol{\mu} \end{aligned}$$

[Total: 8]

(3) Consider the multiple linear regression model

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\mathcal{E}}.$$

Prove that $X^T X$ is invertible if and only if the columns of X are linearly independent.

Solution:

Suppose the columns of X are linearly independent. In vector notation, this condition is expressed as $X\boldsymbol{\alpha} \neq \mathbf{0}$ for all vectors $\boldsymbol{\alpha} \neq \mathbf{0}$.

Now

$$\begin{aligned} X\boldsymbol{\alpha} \neq \mathbf{0} \quad &\Leftrightarrow \quad \|X\boldsymbol{\alpha}\|^2 > 0 \\ &\Leftrightarrow \quad \boldsymbol{\alpha}^T X^T X\boldsymbol{\alpha} > 0 \end{aligned}$$

Finally, observe that $\boldsymbol{\alpha}^T X^T X\boldsymbol{\alpha} > 0$ for all $\boldsymbol{\alpha} \neq \mathbf{0}$ if and only if all eigenvalues of $X^T X$ are positive. Since $|X^T X|$ is the product of its eigenvalues, the result is proved.

[Total: 5]

(4) The data file `fev.txt` contains data from an early study on the deleterous effects of smoking in children and young adults. Measures of lung function (FEV, forced expiratory volume in litres per second) were made in 654 healthy children seen for a routine check up in a paediatric clinic. The children participating in this study were asked whether they were current smokers. A higher FEV is usually associated with better respiratory function and it is well known that prolonged smoking diminishes FEV in adults.

The following variables were recorded: 654 subjects aged 3-19.

Variable	Description
`ID`	ID number
`Age`	years
`FEV`	litres
`Height`	inches
`Sex`	`Male` or `Female`
`Smoker`	`Non` = nonsmoker, `Current` = current smoker

The purpose of your analysis is to relate lung capacity (as measured by `FEV`) to the predictor variables: `Sex`, `Height`, `Smoker` and `Age` using multiple linear regression. We can do this by fitting a model of the form:

$$f\left(\texttt{FEV}_i\right) = \beta_0 + \beta_1 \times \texttt{Sex}_i + \beta_2 \times \texttt{Height}_i + \beta_3 \times \texttt{Smoker}_i + \beta_4 \times \texttt{Age}_i + \mathcal{E}_i,$$

where f is some suitable transformation and $\mathcal{E}_i \overset{iid}{\sim} N(0, \sigma^2)$; $i = 1, 2, \ldots, 654$.

(a) In R, create an appropriate design matrix X using the function `model.matrix()`. Call this matrix `X` and provide the first 10 rows.

Solution:

```
X<-model.matrix( ~  Sex + Height + Smoker + Age, data=fev )
X[1:10,]

##    (Intercept) SexMale Height SmokerNon Age
## 1            1       0   57.0         1   9
## 2            1       0   67.5         1   8
## 3            1       0   54.5         1   7
## 4            1       1   53.0         1   9
## 5            1       1   57.0         1   9
## 6            1       0   61.0         1   8
## 7            1       0   58.0         1   6
## 8            1       0   56.0         1   6
## 9            1       0   58.5         1   8
## 10           1       0   60.0         1   9
```

(b) In your matrix `X`, what are the values for the *categorical variables*? What levels of the variables do these values correspond to?

Solution:

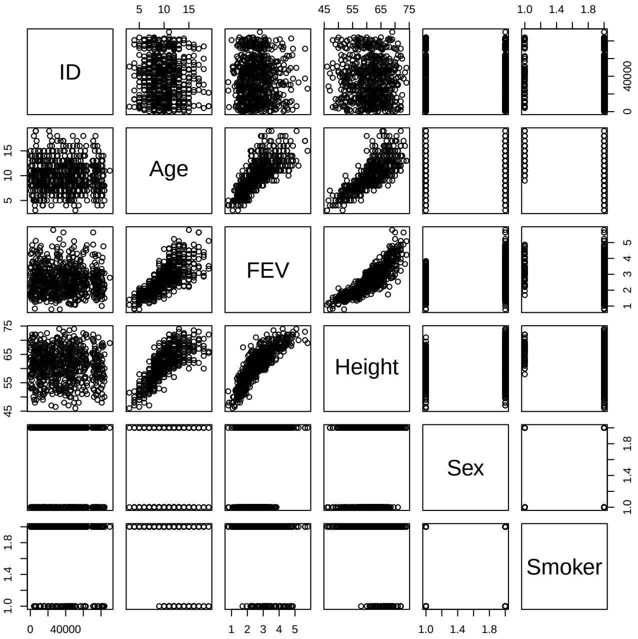
- Values in the categorical columns are 1s and 0s.

- `Sex` has been coded as `1=Male` and `0=Female`.
- `Smoker` has been coded as `1=Non` and `0=Current`.

(c)
- Using the `pairs` command in R, create a scatterplot matrix for all the variables.
- Comment on the relationship between `FEV` and the continuous predictor variables.
- Repeat the scatterplot matrix but now using `log(FEV)` as the response variable. Comment again on the relationship between `FEV` and the continuous predictor variables.

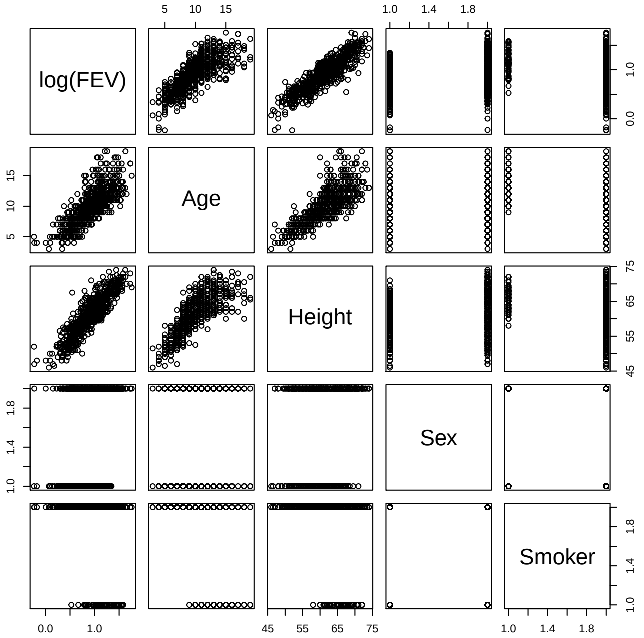
Solution:

-
  ```
  pairs(fev)
  ```



- There is a nonlinear, possibly quadratic relationship between `FEV` and `Height`. There is a possibly nonlinear relationship between `FEV` and `Age` and the variance of `FEV` increases with increasing `Age`.

-
  ```
  pairs(log(FEV)~ Age + Height + Sex + Smoker, data=fev)
  ```



The relationship between `log(FEV)` and `Height` is now linear with constant variance. The relationship between `log(FEV)` and `Age` also now looks linear and the constant variance assumption is approximately true.

(d) Obtain a set of diagnostic plots (and neatly include them in your answers) for each of the models

$$\texttt{log(FEV)} \sim X + 0$$

and

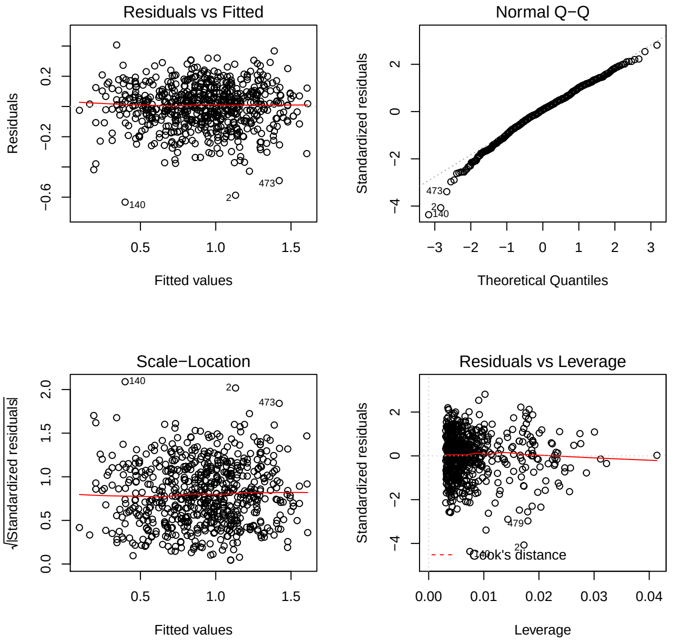
$$\texttt{FEV} \sim X + 0.$$

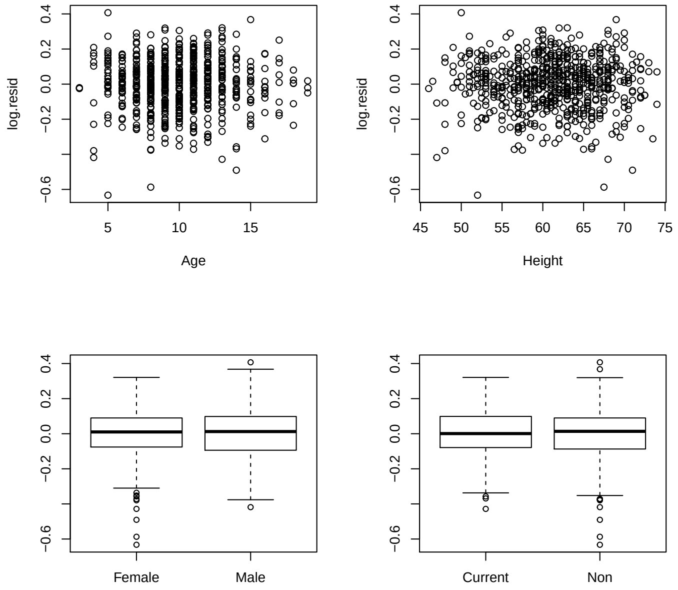
(Note the +0 in the formula ensures R does not automatically include an intercept term; the intercept is already accounted for in your design matrix.)

Which assumptions look better for the model fit on the log-transformed data compared to the raw data? In each case, justify your answer.

Solution:

```
# Log transformed fev as outcome
lm.log <- lm(log(FEV) ~ X + 0, data=fev)
par(mfrow=c(2,2))
plot(lm.log)
log.resid <- residuals(lm.log)
plot(Age, log.resid)
plot(Height, log.resid)
boxplot(log.resid~Sex)
boxplot(log.resid~Smoker)
```

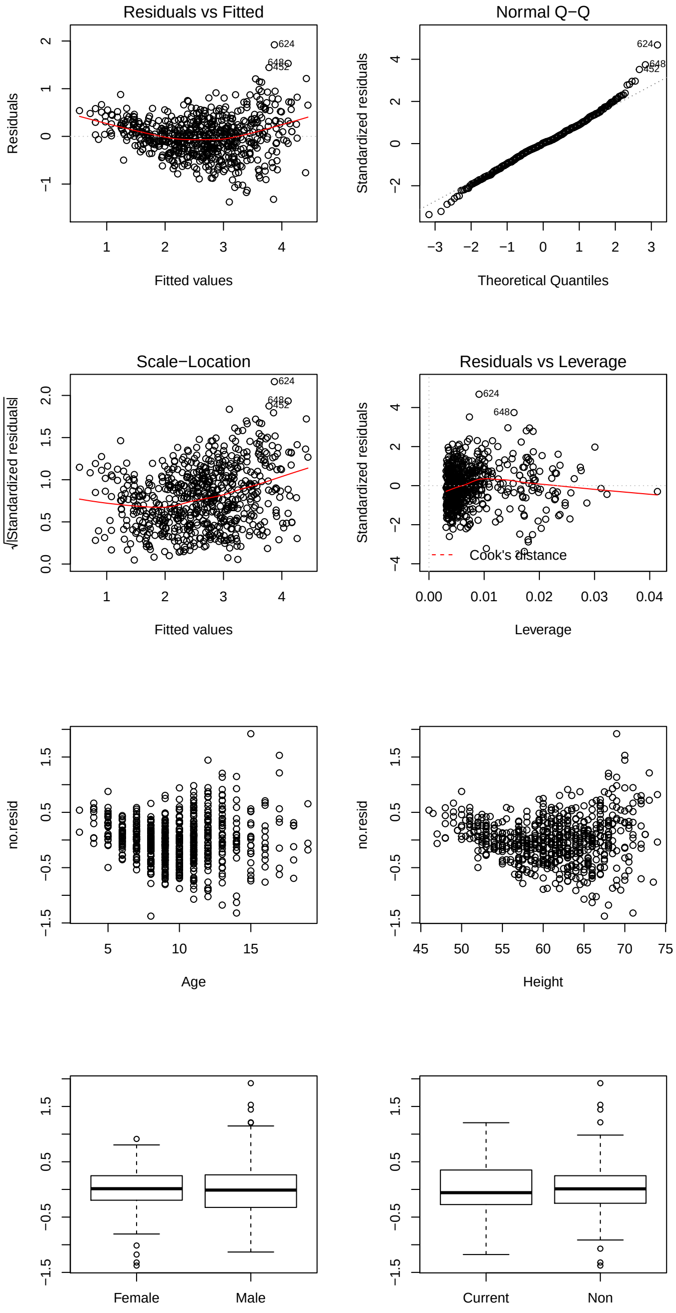




Examination of the Residuals versus Fitted plot, Scale-Location plot and the Residuals versus predictor variable plots for the raw (untransformed) data show clear evidence of curvature (non-linearity) and heteroscedasticity (increasing variance). There is also evidence of non-constant variance in the boxplots of the Residuals for the categorical variables Sex and Smoker. On the other hand, the model fit to the log transformed data shows no evidence of the non-linearity or heteroscedasticity seen on the raw data. There is perhaps some minor curvature in the Residuals versus Age plot.

```r
# Raw fev as outcome
lm.nolog <- lm(FEV ~ X + 0, data=fev)
par(mfrow=c(2,2))
```

```r
plot(lm.nolog)
no.resid <- residuals(lm.nolog)
plot(Age, no.resid)
plot(Height, no.resid)
boxplot(no.resid~Sex)
boxplot(no.resid~Smoker)
```



The assumption of normality looks marginally better satified for the log transformed data, although there are some departures in the lower tail corresponding to the large negative residuals observed in the boxplots for non-smokers and females.

The assumption of independence is not affected by the log transform. We are not given information about the design of the study, so it may be reasonable to assume the study population behaves like a random sample. However, if there were sibling groups represented in the sample for example, the independence assumption may not be satisfied.

(e) Provide a careful interpretation of the coefficient for the predictor variable of `Sex` on the *orginal scale*.

Solution:

```
# Log transformed model
summary(lm.log)

##
## Call:
## lm(formula = log(FEV) ~ X + 0, data = fev)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.63278 -0.08657  0.01146  0.09540  0.40701
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## X(Intercept) -1.990066   0.081888 -24.302  < 2e-16 ***
## XSexMale      0.029319   0.011719   2.502   0.0126 *
## XHeight       0.042796   0.001679  25.489  < 2e-16 ***
## XSmokerNon    0.046068   0.020910   2.203   0.0279 *
## XAge          0.023387   0.003348   6.984  7.1e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1455 on 649 degrees of freedom
## Multiple R-squared:  0.9779,Adjusted R-squared:  0.9777
## F-statistic:  5736 on 5 and 649 DF,  p-value: < 2.2e-16

summary(lm.log)$coefficients

##                Estimate  Std. Error    t value      Pr(>|t|)
## X(Intercept) -1.99006571 0.081887729 -24.302368  2.876539e-93
## XSexMale      0.02931936 0.011718565   2.501958  1.259586e-02
## XHeight       0.04279579 0.001678968  25.489334 7.664775e-100
## XSmokerNon    0.04606754 0.020910198   2.203113  2.793739e-02
## XAge          0.02338721 0.003348451   6.984488  7.096410e-12

exp(summary(lm.log)$coefficients[,1])

## X(Intercept)     XSexMale      XHeight   XSmokerNon         XAge
##    0.1366864    1.0297534    1.0437247    1.0471451    1.0236628
```

Based on the regression output, Sex, Height, Smoking Status and Age are all significant predictors of `log(FEV)`.

Recall that in the context of the multiple regression, the interpretation of an individual coefficient corresponds to the effect of a unit change in that variable if all other predictors are kept fixed.

For fixed Height, Smoking Status and Age, the mean `log(FEV)` for males is higher than for females by 0.029. On the original FEV scale, this equates to an approximate percentage increase of 3 for males relative to females.

(f) (i) Explain why it is possible to obtain a $100(1-\alpha)\%$ prediction interval for `FEV`. *In your answer, give the form of the interval on each of the transformed and untransformed scales.*

Solution:

A $100(1-\alpha)\%$ prediction interval for `FEV` can be obtained by exponentiation of the endpoints of the interval for `log(FEV)`, because a prediction interval refers to a single random variable Y_0.
Since

$$I_L \leq Y_0 \leq I_U \Leftrightarrow e^{I_L} \leq e^{Y_0} \leq e^{I_U},$$

it follows the latter is a $100(1-\alpha)\%$ prediction interval for e^{Y_0}, where $Y_0 =$ `log(FEV)` and $e^{Y_0} =$ `FEV`.

(ii) Can you perform a similar calculation for a confidence interval for `FEV`? Justify your answer.

Solution:

No. A confidence interval is for an expected value.
If (L, U) is a $100(1-\alpha)\%$ confidence interval for $E(Y_0)$, we cannot use the same argument to deduce that (e^L, e^U) is a $100(1-\alpha)\%$ confidence interval for $E(e^{Y_0})$ because

$$E(e^{Y_0}) \neq e^{E(Y_0)}.$$

In other words, (e^L, e^U) is a confidence interval for $e^{E(Y_0)}$, where $Y_0 =$ `log(FEV)`, but you want a confidence interval for $E(\text{FEV}) = E(e^{Y_0})$.

[Total: 25]