Assignment 1, Statistical Modelling III

Andrew Martin

March 19, 2018

1. Let A and B be matrices as given in T1Q4. Furthermore assume C is a constant $n \times n$ matrix. Prove that

$$tr(ABC) = tr(BCA)$$

(Use the result from T1Q4)

Solution Since B and C are 'compatible', let $D = BC$.

$$tr(AD) = tr(DA), \implies tr(ABC) = tr(BCA) \quad \text{from tute 1 question 4}$$

As Required

2. Let Y be a random $n \times 1$ vector with mean μ and variance matrix Σ, and let A be a constant $n \times n$ matrix. Show that

$$E(Y^T AY) = tr(A\Sigma) + \mu^T A\mu$$

Solution Using the covariance formula:

$$cov(\mathbf{X}, \mathbf{Y}) = E[\mathbf{X}\mathbf{Y}^\mathbf{T}] - E[\mathbf{X}]E[\mathbf{Y}]^T$$

$$\implies E[\mathbf{X}\mathbf{Y}^\mathbf{T}] = E[\mathbf{X}]E[\mathbf{Y}]^T + cov(\mathbf{X}, \mathbf{Y})$$

By grouping AY this can be applied: (In the formula above using $X = Y^T$ and $Y^T = AY$)

$$
\begin{aligned}
E(Y^TAY) &= E[\mathbf{Y^T}]E[(\mathbf{AY})^T]^T + cov(\mathbf{Y^T}, (\mathbf{AY})^\mathbf{T}) \\
&= E[\mathbf{Y}^T]E[\mathbf{Y^T A^T}]^T + E\left[(\mathbf{Y}^T - E(\mathbf{Y}^T))(\mathbf{Y^T A^T} - E(\mathbf{AY}))^T\right] \\
&= \mu^T(\mu^T A^T)^T + E\left[(\mathbf{Y}^T - \mu^T)((\mathbf{AY})^T - (A\mu)^T)^T\right] \\
&= \mu^T(A\mu) + E\left[(\mathbf{Y} - \mu)^T(\mathbf{AY} - A\mu)\right] \\
&= \mu^T(A\mu) + E\left[(\mathbf{Y} - \mu)^T\mathbf{A}(\mathbf{Y} - \mu)\right] \\
&= \mu^T(A\mu) + E\left[tr[(\mathbf{Y} - \mu)^T\mathbf{A}(\mathbf{Y} - \mu)]\right] \ \textit{the inside is a scalar.} \\
&= \mu^T(A\mu) + E\left[tr[\mathbf{A}(\mathbf{Y} - \mu)(\mathbf{Y} - \mu)^T]\right] \ \textit{using Q1} \\
&= \mu^T(A\mu) + tr\left[E[\mathbf{A}(\mathbf{Y} - \mu)(\mathbf{Y} - \mu)^T]\right] \\
&= \mu^T(A\mu) + tr\left[\mathbf{A}E[(\mathbf{Y} - \mu)(\mathbf{Y} - \mu)^T]\right] \\
&= \mu^T(A\mu) + tr\left[\mathbf{A}\Sigma\right] \ \textit{definition for the variance} \\
&= tr(A\Sigma) + \mu^T A\mu
\end{aligned}
$$

As Required

3. Consider the multiple linear regression model

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

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

Solution $\Leftarrow$ Given the columns of X are linearly independent, prove that X^TX is invertible.
Using the definition for linear independence,

$$Xv = 0 \implies v_i = 0 \ \forall i$$

Assume X^TX has linear dependency (meaning it isn't invertible) I.e. $X^TXv = 0$

$$
\begin{aligned}
&\implies v^TX^TXv = 0 \\
(Xv)^TXv = v^TX^TXv &= 0 \\
&\implies (Xv)\cdot(Xv) = 0 \\
&||Xv||^2 = 0
\end{aligned}
$$

Which is a contradiction as X has linearly independent columns. This implies the columns of X^TX are linearly independent, which since X^TX is square, implies it is invertible (matrix-inverse theorem) **As Required**

4. The data file `fev.txt` contains data from an early study on the deleterious 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 variables were recorded as given: 654 subjects aged 3-19.
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:

$$f(FEV_i) = \beta_0 + \beta_1 \times Sex_i + \beta_2 \times Height_i + \beta_3 \times Smoker_i + \beta_4 \times Age_i + \mathcal{E}_i$$

Where f is a suitable transformation and $\mathcal{E} \sim N(0, \sigma^2) \quad \forall i$

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

Solution The first 10 rows are as below:

```
> head(X,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
```

As Required

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

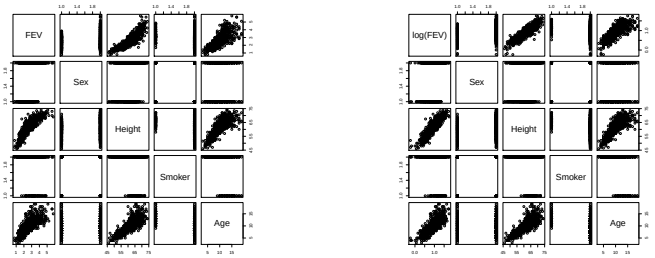
Solution The categorical variables are represented with integer values corresponding to their state, since each of the categorical levels in this example only has 2 factors, they are represented as a 0 or a 1.
For Sex, male is denoted with a 1 and female with a 0.
For Smoker, non-smoker is denoted with a 1 and smoker with a 0. I.e.

```
> contrasts(fev$Sex)
       Male
Female    0
Male      1
> contrasts(fev$Smoker)
        Non
Current   0
Non       1
```

As Required

(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, instead using `log(FEV)` as the response variable. Comment again on the relationship between `FEV` and the continuous predictor variables.

Solution



Pairs plots for the variables of `FEV`. Left uses the standard values, while right uses a logarithmically transformed FEV variable (denoted `log(FEV)`

The continuous predictors are height and age. For FEV with height, there is an exponential trend (easily observed by comparing the FEV/Height and Height/FEV plots). For FEV with age, the trends are much more linear, but the spread of the data points seems to be significantly worse for the higher values.

For `log(FEV)` the exponential trend with height is cancelled out, giving what appears to be a linear relationship. However, for `log(FEV)` with age, this introduces a slight exponential relationship.

As Required

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

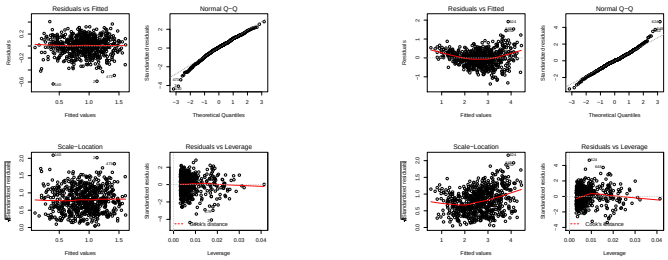
$$log(FEV) \sim X + 0$$

and

$$FEV \sim X + 0$$

Note: the `+0` is necessary so that `R` does not include an intercept term. Which assumptions look better for the model fit on the log-transformed data compared to the raw data. In each case, justify the answer.

Solution



Diagnostic plots. Left is the model with `log(FEV)`, while right is the model with `FEV`.

For the plots regarding `FEV` (on the right), the first plot shows that there is a dip in the residual error values around the centre of the graph (following the red line), which shows that the data may not be linear. The normal q-q shows that the data is normally distributed, i.e. the data (mostly) follows the straight line. However, at the top-right of the graph, it can be seen that the data all spreads off the line (and this can be observed slightly on the bottom left too) - which implies it may not quite be normally distributed. Homoscedasticity is observed in the Residuals vs fitted plot (top left) - the variance of the points does appear to change a little bit for the `FEV` plots.

For `log(FEV)` (shown in the plots on the left), linearity can be observed from the residuals vs fitted plot, the red line indicates that the data is linear. Normality is shown in the Normal q-q plot - and it can be seen that the data mostly follows the dotted line - so the data can be assumed to be normal. Homoscedasticity can be seen in the residuals vs fitted plot; the points seem to have equal variance from the centre throughout, implying homoscedasticity.

For both `FEV` and `log(FEV)`, independence is a design assumption.

The assumptions of homoscedasticity, linearity and normality all look better on the `log(FEV)` plots.

As Required

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

Solution Using `summary(fevmodel)` yields:

> **summary**(fevmodel)

Call:
lm(**formula** = FEV ~ X + 0, **data** = fev)

Residuals:

Min	1**Q**	Median	3**Q**	Max

−1.37656 −0.25033 0.00894 0.25588 1.92047

Coefficients :

	Estimate	Std. Error	**t** value	Pr(>|**t**|)	
X(Intercept)	−4.544220	0.232046	−19.583	< 2e−16	***
XSexMale	0.157103	0.033207	4.731	2.74e−06	***
XHeight	0.104199	0.004758	21.901	< 2e−16	***
XSmokerNon	0.087246	0.059254	1.472	0.141	
XAge	0.065509	0.009489	6.904	1.21e−11	***

−−−

Signif. **codes**: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 0.4122 **on** 649 degrees of freedom
Multiple **R**−squared: 0.9781, Adjusted **R**−squared: 0.9779
F−statistic: 5800 **on** 5 and 649 DF, p−value: < 2.2e−16

The coefficient for Sex is the value 0.157103. This means that a male will expect to have an FEV 15.7103% higher than a female.

As Required

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

Solution We can generate the prediction interval as it is simply a confidence interval for the estimable function containing the new data points. When generating a prediction interval on R, it generates 654 intervals. **As Required**

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

Solution We cannot perform a confidence interval for this, as we can only produce a confidence interval for the predictor variables in this case. **As Required**

The R code used for this assignment is below:

```
library(tidyverse)
library(ggplot2)
setwd("~/Uni")
fev =read.table(file='fev-1.txt',header=TRUE)
##a
X=model.matrix(~Sex + Height + Smoker + Age,data=fev)
head(X,10)

#b
contrasts(fev$Sex)
contrasts(fev$Smoker)

#c
pdf( file="PairsPlot.pdf")
pairs(FEV~Sex + Height + Smoker + Age,data=fev)
pairs(log(FEV) ~ Sex + Height + Smoker + Age, data=fev)

#d
logfevmodel=lm(log(FEV)~X+0,data=fev)
fevmodel=lm(FEV~X+0,data=fev)

tmp = par(mfrow = c(2,2))
plot(logfevmodel)
plot(fevmodel)
par(tmp)

#e
summary(fevmodel)

#f
summary(fevmodel)
summary(logfevmodel)
predict(fevmodel,interval="prediction")
predict(logfevmodel,interval="prediction")
dev.off()
```