UNIVERSITEIT VAN DIE VRYSTAAT
UNIVERSITY OF THE FREE STATE

STSM 3744

WISKUNDIGE STATISTIEK & AKTUARIËLE WETENSKAP/
MATHEMATICAL STATISTICS & ACTUARIAL SCIENCE

Test 1 — 7 September 2018

MEMORANDUM

TYD/TIME: 140 Minutes + 20 min. computer time PUNTE/MARKS: 50

INSTRUCTIONS:

1. **DO NOT put your details in the document header or anywhere else in the document.** *This is different to what you are used to — please take careful note. The inside of your document must be anonymous.* Your details should only be in the file name.
2. Answer all questions in a single Word document but convert the final version to **.pdf** before submitting (backup the Word document and all working files separately).
3. Label questions clearly, as it is done on the question paper.
4. **Show all code, calculations, formulas, hypotheses and reasoning.**
5. Use a significance level of 5% for all hypothesis tests unless stated otherwise.
6. No USB drives or data transfer devices may be used during the course of the test.
7. No attempt at communication of any kind is allowed with other students. No email or messaging service on any device may be accessed.
8. You may access any notes, textbooks, or other papers that you personally bring into the venue. You may not share notes with any other student.
9. You may use a laptop running Windows or Linux during this exam, but NO CELL-PHONE, tablet, music player, or similar devices.
10. You may use headphones, but only with UFS computers, not your own devices.
11. You may access learn.ufs.ac.za as well as any search engine and any reference site of statistical information.
12. You may not access any site (or non-browser software) that allows communication with other students. Do not attempt to communicate with another student in any way, do not attempt to use any credentials but your own at any time.
13. Make backups of your work at least once per hour, on both the D drive of the computer and on learn.ufs.ac.za.

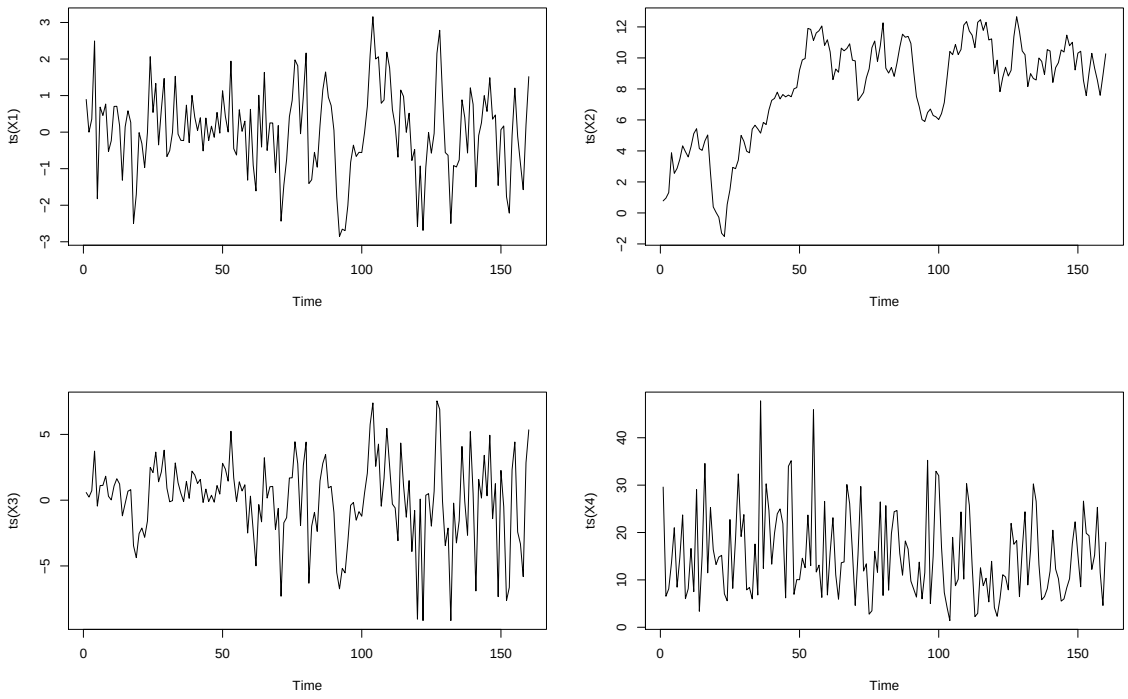
This style text refers to marking guidance. It is there to help you decide how to allocate marks. If it is not clear enough then use your own discretion.

Question 1

For the four series in 'STSM3744Test1of2018Q1data.csv', determine graphically whether each series is or is not stationary, and explain your reasoning. You may use hypothesis tests to check your reasoning but these are not required and will not receive marks.

```
mydata <- read.csv('STSM3744Test1of2018Q1data.csv',row.names=1)
attach(mydata)
windows(12,8)
par(mfrow=c(2,2))
plot(ts(X1)); plot(ts(X2)); plot(ts(X3)); plot(ts(X4))
detach(mydata)
```

✔**for code reading in data,** ✔**for code to draw plots.**



✔ ✔**for giving 4 plots in total.**

For X_1 we note a change in the correlation pattern near the middle of the series (the first autocorrelation goes from negative to positive). ✔Thus X_1 is not stationary. ✔

For X_2 we note a change in the mean early on in the series (the first quarter is much lower than the rest). ✔Thus X_2 is not stationary as a whole. ✔

For X_3 we note a gradual change in the variance (increasing variance throughout the series). ✔Thus X_3 is not stationary. ✔

For X_4 we note stable mean, stable variance, and stable autocorrelation patterns on the graph. ✔Thus we continue to assume X_4 is stationary. ✔

Total for Question 1: 12

Question 2

Consider the data set given in 'STSM3744Test1of2018Q2data.csv'. The three variables are as follows:

Y = Years of education, **and is assumed to follow a Poisson distribution.**

X_1 = Household income (in Fairyland Dollars per month)

X_2 = Log of social network follower count

(a) In general, when is it appropriate to do a Poisson regression and when is it not appropriate? [3]

This model is appropriate when it is clear that the dependent variable follows a Poisson distribution. ✔This mostly occurs with count data ✔(*e.g.* number of accidents in a period). It is not appropriate when the mean and variance of the dependent variable differ greatly ✔as this contradicts the Poisson assumption.

(b) Fit, give, and store the following five GLMs using the given data: [7]

$$
\begin{aligned}
GLM1: &\ \log E[Y_i] = \beta_0 \\
GLM2: &\ \log E[Y_i] = \beta_0 + \beta_1 X_{1i} \\
GLM3: &\ \log E[Y_i] = \beta_0 + \beta_2 X_{2i} \\
GLM4: &\ \log E[Y_i] = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} \\
GLM5: &\ \log E[Y_i] = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i}
\end{aligned}
$$

NB: This question was wrong on the question paper handed out. The question paper did not have the x terms.

```
q2 <- read.csv('STSM3744Test1of2018Q2data.csv')
summary(m1 <- glm(y~1,family=poisson))
summary(m2 <- glm(y~x1,family=poisson))
summary(m3 <- glm(y~x2,family=poisson))
summary(m4 <- glm(y~x1+x2,family=poisson))
summary(m5 <- glm(y~x1*x2,family=poisson))
```

✔for reading in data. ✔ ✔ ✔for saying 'glm' and 'family=poisson' in all cases (give 3 or 0).

```
glm(formula = y ~ 1, family = poisson)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.43186    0.01711   142.1   <2e-16 ***
---
    Null deviance: 702.81  on 299  degrees of freedom
Residual deviance: 702.81  on 299  degrees of freedom
AIC: 1957.4

glm(formula = y ~ x1, family = poisson)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 1.108e+00  7.203e-02   15.38   <2e-16 ***
x1          8.427e-04  4.275e-05   19.71   <2e-16 ***
---
    Null deviance: 702.81  on 299  degrees of freedom
Residual deviance: 323.22  on 298  degrees of freedom
AIC: 1579.8

glm(formula = y ~ x2, family = poisson)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.480149   0.067472  36.758   <2e-16 ***
x2          -0.009706   0.013139  -0.739     0.46
---
    Null deviance: 702.81  on 299  degrees of freedom
Residual deviance: 702.26  on 298  degrees of freedom
AIC: 1958.9

glm(formula = y ~ x1 + x2, family = poisson)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.143e+00  9.899e-02  11.543   <2e-16 ***
x1           8.420e-04  4.274e-05  19.698   <2e-16 ***
x2          -6.728e-03  1.322e-02  -0.509    0.611
---
    Null deviance: 702.81  on 299  degrees of freedom
Residual deviance: 322.96  on 297  degrees of freedom
AIC: 1581.6

glm(formula = y ~ x1 * x2, family = poisson)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.469e+00  2.849e-01   5.157 2.51e-07 ***
x1           6.401e-04  1.708e-04   3.749 0.000178 ***
x2          -7.495e-02  5.743e-02  -1.305 0.191848
x1:x2        4.221e-05  3.455e-05   1.222 0.221761
---
    Null deviance: 702.81  on 299  degrees of freedom
Residual deviance: 321.47  on 296  degrees of freedom
AIC: 1582.1
```

✔ ✔ ✔for giving 5 correct model outputs. Give 2 marks if they gave 5 outputs but they are not for glm or not poisson. Give 1 if multiple mistakes but still attempted. 0 otherwise.

(c) Since model $GLM4$ is nested in model $GLM1$ we can compare them with a χ^2 test in order to determine the global fit of model $GLM4$. Perform this test and explain the result. [4]

```
anova(m1,m4,test='Chisq')
```

✔**for giving code that looks like it uses the correct models based on earlier code. The calculations can be done manually, in which case give marks if the numbers match the output below.**

```
 Analysis of Deviance Table
Model 1:  y   1
Model 2:  y   x1 + x2
Resid.  Df Resid.  Dev Df Deviance Pr(>Chi)
1 299 702.81
2 297 322.96 2 379.85 < 2.2e-16 ***
```

✔**for giving output.**

Since the p-value $< \alpha$ we reject H_0: the added terms as a group are not significant in the nested model. ✔**for giving the null hypothesis. Any correct form is fine.** Thus, the combination of terms included in $GLM4$ is meaningful in predicting y. ✔**for rejecting the null hypothesis in words. Do NOT work with any mistake here.**

(d) Explain what is meant by the term 'parsimonious', and then explain which of the models can be said to be the most parsimonious. [5]

```
BIC(m1,m2,m3,m4,m5)
```

✔**for giving code that references the models from earlier.**

A parsimonious model is one that is both simple and works well. ✔ ✔**for mentioning both simplicity and accuracy in some way.**

The model with only term X_1 ✔is the most parsimonious because it has the lowest value of BIC. ✔

(e) Using the most parsimonious model you chose only, interpret the meaning of β_1 in detail. [3]

$\hat{\beta}_1 \approx 0.00084271$, implying that a unit increase in X_1 multiplies ✔$E[Y]$ by $\exp(0.00084271) = 1.000843$. ✔**for exp.** In other words, according to the model, an increase of 1 Fairyland Dollar per month in Household income is associated with the expected value of Years of education increasing by about 0.08430652% of its previous value. ✔**for well worded explanation.**

(f) Briefly explain the steps you would follow to obtain a prediction interval for a single new observation in either $GLM2$ or $GLM3$, whichever you prefer. [4]

1. Simulate K vectors from a $N_2\left(\hat{\boldsymbol{\beta}},\ \hat{\Sigma}_{\hat{\boldsymbol{\beta}}}\right)$. ✔
2. For each vector, calculate $\exp\left(X^{(new)}\boldsymbol{\beta}_k\right)$ where $X^{(new)}$ is the new observation in row form. ✔
3. For each vector, simulate a $Poisson\left(\exp\left(X^{(new)}\boldsymbol{\beta}_k\right)\right)$ value. ✔
4. Arrange the resulting Poisson values in vector form and obtain lower and upper quantiles ✔, *e.g.* 2.5% and 97.5%.

Student numbers do not need to match my numbers, or even have to be numbered, but order of steps must be correct. Verbal explanations in place of formulas are also acceptable as long as they are detailed enough. Defining more terms and using them is also good, like $\lambda_k^{(new)} = \exp\left(X^{(new)}\boldsymbol{\beta}_k\right)$.

(g) Remember to remove all identifying information from your document, and make sure you have a backup made of your work in the first two hours of the test. [2]

✔**if you can't figure out whose test you are marking.**

✔**if the filename has a 'b' at the end indicating that this person made a backup in a reasonable time.**

Total for Question 2: 28

Question 3

Consider the time series given in 'STSM3744Test1of2018Q3data.csv'. Assume the series has constant variance and autocovariance patterns.

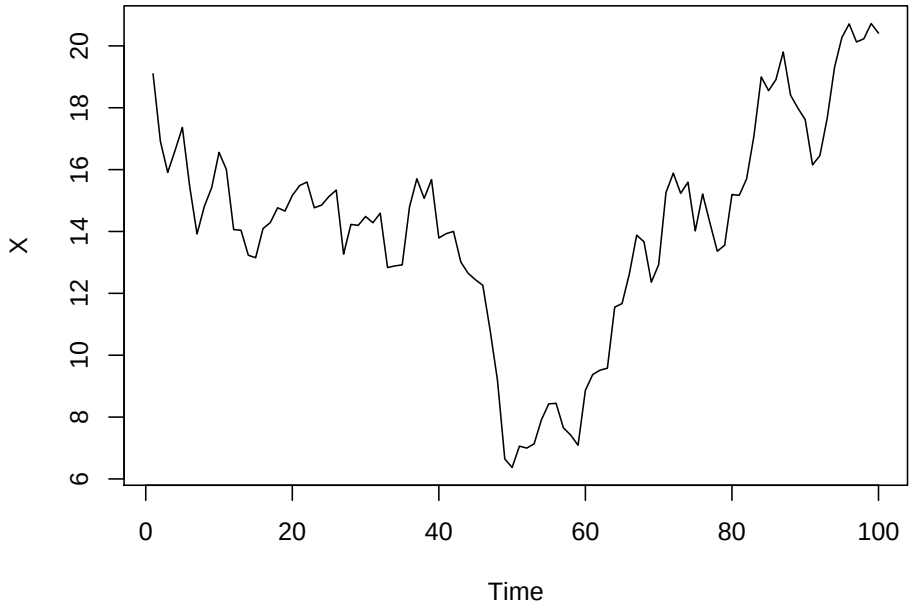
Using a plot and a unit root test, determine whether the series could be stationary. If not, difference the series.

Repeat the above until stationarity is achieved.

Total for Question 3: 10

```
mydata <- read.csv('STSM3744Test1of2018Q3data.csv',row.names=1)
X <- ts(mydata$x)
windows(6,4); par(mar=c(4,4,0.1,0.1))
plot(X)
PP.test(X)
DX <- diff(X)
windows(6,4); par(mar=c(4,4,0.1,0.1))
plot(DX)
PP.test(DX)
```

✔**for sensible looking code. Does not need to match memo, but must look like it can run.**



✔**for giving graph.**

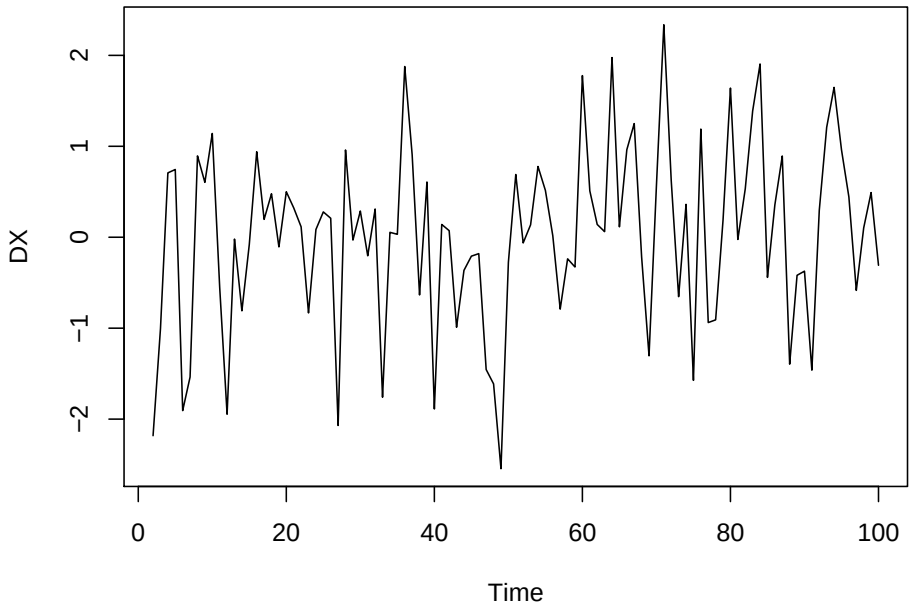
On the graph we see that the mean is changing ✔gradually over time and that each value seems strongly connected to the previous value. It appears to not be stationary.

H_0 : Differencing is appropriate ✔**(saying 'unit root is present' is also acceptable)** *vs* H_1 : differencing is not appropriate

```
 Phillips-Perron Unit Root Test
data:  X
Dickey-Fuller = -1.7096, Truncation lag parameter = 3,
p-value = 0.697
```

✔

Since the p-value is larger than α we fail to reject the null hypothesis and continue to assume that differencing is appropriate for this series. ✔**if the student somehow had a small p-value and rejects here then give this mark but not the next one.**



✔

We note stable mean, stable variance, and stable autocorrelation patterns on the graph. Thus we continue to assume DX is stationary. ✔ **no reason = no mark. If the student did not difference then they lose this mark and the two that follow.**

```
 Phillips-Perron Unit Root Test
data:  DX
Dickey-Fuller = -8.638, Truncation lag parameter = 3,
p-value = 0.01
```

✔

Since the p-value is smaller than α we reject the null hypothesis and conclude that differencing is not appropriate for this series. ✔

Total half marks on memo = 100 vs. 100 = Double total margin points (=50).