~~~
---
title: "STAT5703 HW3 Exercise 1"
author: "Wen Fan(wf2255), Banruo Xie(bx2168), Hanjun Li(hl3339)"
output: pdf_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

## Exercise 1.
#### Question 1.
```{r message=FALSE, warning=FALSE, include=FALSE}
library(magrittr)
library(dplyr)
library(lubridate)
library(ggplot2)
library(forecast)
```
```{r}
df_milk <- read.table(file ="milk.txt",skip = 15,col.names=c("Month", "production"))
df_milk$Month<-ymd(df_milk$Month,truncated = 1)
df_milk <- df_milk %>%
  mutate(num_month=row_number())
df_milk %>% ggplot(aes(x=Month, y=production)) +
  geom_line() +
  geom_smooth(method = "lm", se = FALSE)
```

```{r}
linearMod <- lm(production ~ num_month, data=df_milk)
summary(linearMod)
```
From linear model, the production is 611 pounds per cow for the first month and increase 1.69
pounds per cow for each month.
```{r}
checkresiduals(linearMod)
```
The residuals are normally distributed, we can say it is a stationary time series

#### Question 2.
```{r}
ggAcf(df_milk$production)
```
~~~

From ACF plot above. the autocorrelation crosses the dashed blue line, it means that specific lag is significantly correlated with current series. The slow decrease in the ACF as the lags increase and due to the seasonality.

~~~
```{r}
ggPacf(df_milk$production)
```
~~~

After remove linear trends in a timeseries, we can say the plot indicates a seasonal AR(1) component because the Pacf Cuts off after lag 1.

Question 3.

~~~
```{r}
fitAR1 <- Arima(df_milk$production, order=c(1,0,0))
fitAR1
checkresiduals(fitAR1)
```


```{r}
fitAR2 <- Arima(df_milk$production, order=c(2,0,0))
fitAR2
checkresiduals(fitAR2)
```
```{r}
fitMA1 <- Arima(df_milk$production, order=c(0,0,1))
fitMA1
checkresiduals(fitMA1)
```
```{r}
fitMA2 <- Arima(df_milk$production, order=c(0,0,2))
fitMA2
checkresiduals(fitMA2)
```
~~~

In all cases, residals are normally distributed white noise, AR(1) is better with lower AICc.

Question 4.

~~~
```{r}
mol1 <- auto.arima(df_milk$production)
mol1
checkresiduals(mol1)
```
~~~

The moel with auto chosen gives us a lower AICc, which is better than AR(1) we get before. This model includes AR(1) and MA(4) with a first order difference.

~~~
```{r}
~~~

~~~
mol2 <- Arima(df_milk$production, order=c(1,1,4), seasonal = list(order = c(1, 1, 1), period = 12))
mol2
checkresiduals(mol2)
```
After manually changing the order, we get a better model ARIMA(1,1,4)(1,1,1)[12] has a much lower AICc.



---
title: "STAT5703 HW3 Ex2"
author: "Wen Fan(wf2255), Hanjun Li(hl3339), Banruo Xie(bx2168)"
output: pdf_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

## Exercise 2

#### Question 1

```{r}
library(mgcv)
cars$speed_sqr <- (cars$speed)^2
lr = lm(dist ~ speed + speed_sqr , data=cars)
summary(lr)
AIC(lr)
```

In this model, we can see from the summary table above that the model is significant and all features are significant as well. However, the R-squared value is not very high. Thus, we drop the relatively unsignificant variable 'speed' using 'stepwise' method. After dropping, we can see the AIC value of the model decreases, while both of the variables in the model become significant.

```{r}
lr2 = step(lm(dist ~ speed_sqr+speed , data=cars))
summary(lr2)
AIC(lr2)
```
According to above results, we would choose speed-squared only model as the best model.
#### Question 2
~~~

As we drop the variable 'speed' in point 1, we could only get value of 'reaction time' in the residual of the formulation.
$$time=\frac{dist-\hat{\beta_0}-\hat{\beta_1}*speed}{speed}$$
By getting its value, we can estimate its distribution.
Question 3
```{r}
lr_1 <- function(y, X)
  {
  qrx <- qr(X)
  Q <- qr.Q(qrx,complete=TRUE)
  QR <- qr.R(qrx)
  return(backsolve(QR, (t(Q) %*% y)))
  }
```
Question 4

```{r}
newM2 <- model.matrix(dist ~speed + I(speed^2),cars)
lr_1(cars$dist,newM2)
```
Therefore, the function gives the right result for the coffecient for the model.
Question 5
```{r}
lr_2 <- function (X, y) {
  qrx <- qr(X) ## returns a QR decomposition object
  Q <- qr.Q(qrx,complete=TRUE) ## extract Q
  R <- qr.R(qrx) ## extract R
  f <- t(Q)%*%y
  f <- f[1:ncol(X),]
  beta <- solve(R)%*%f
  residual <- y-X%*%beta
  sigma <- as.vector(t(residual)%*%residual/(nrow(X)-ncol(X)))
  variance <- solve(R)%*%t(solve(R))*sigma
  list(coefficient=beta,std_error=sqrt(as.matrix(diag(variance),ncol=ncol(X))),
       residual_variance=sigma)
}
newM2 <- model.matrix(dist ~speed + I(speed^2),cars)
lr_2(newM2,cars$dist)
```
Therefore, the function gives the right result for the model.
Question 6
```{r}
lr_3 <- function (X, y) {
```

~~~
  qrx <- qr(X) ## returns a QR decomposition object
  Q <- qr.Q(qrx,complete=TRUE) ## extract Q
  R <- qr.R(qrx) ## extract R
  f <- t(Q)%*%y
  f <- f[1:ncol(X),]
  beta <- solve(R)%*%f
  residual <- y-X%*%beta
  sigma <- as.vector(t(residual)%*%residual/(nrow(X)-ncol(X)))
  variance <- solve(R)%*%t(solve(R))*sigma
  vrr <- solve(t(X)%*%X)
  dia <- as.matrix(diag(vrr))
  pvalue <- 2*pt(-abs(beta)/sqrt((sigma*dia)),df=nrow(X)-ncol(X))
  list(coefficient=beta,std_error=sqrt(as.matrix(diag(variance),ncol=ncol(X))),
      pvalue=pvalue,residual_variance=sigma)
}
newM2 <- model.matrix(dist ~speed + I(speed^2),cars)
lr_3(newM2,cars$dist)
```
~~~

Therefore, the function gives the right result for the model.

~~~
---
title: "HW3 Exercise 3"
author: "Wen Fan(wf2255), Banruo Xie(bx2168), Hanjun Li(hl3339)"
date: "4/13/2020"
output: pdf_document
---

```{r, warning=FALSE}
library(readr)
library(MASS)
CpsWages <- read_table2("~/Documents/Columbia/STAT W5703/HW/HW3/CpsWages.txt")
names <- c('sex', 'race', 'marr', 'occupation', 'sector', 'south', 'union')
CpsWages[, names] <- as.data.frame(sapply(CpsWages[, names], as.factor)) # make some of the
variables as vector type
```
~~~

Problem 1

We could use a multiple linear regression to examine this dataset. That is, use wage as our target variables and all else to be the predictors. It is not a good idea to include age, education, and experience at the same time because those variables are highly correlated. i.e. as a person ages, one tend to have higher education and more experience; thus, it might lead to issues of collinearity.

Problem 2

~~~
```{r, warning=FALSE}
m1 <- lm(wage ~ ., data = CpsWages)
summary(m1)

par(mfrow=c(2,2))
plot(m1)
```
~~~

From the plot of residual v.s. fitted value, it seems that there's a cone-shaped pattern as the fitted values get larger, so it doesn't agree with the hypothesis of homoscedasticity. From the normal QQ-plot, although there are some points depart from the QQ-line, we see that most of the points are aligned with the normal QQ-line, so we can say that the hypothesis of normality is generally met.

Problem 3

Using $\alpha=0.05$, it appears that only sex, union, and occupation are statistically significant according to the associated p-values from the model. To test if sector is significant, we can again use the associated p-value calculated by t-test from the model. Using $\alpha=0.05$, we fail to reject the null hypothesis as the p-value for sector level 1 and 2 are both greater than 0.05.

Problem 4

Since we see that only sex, union, and occupation are statistically significant from the model, we can fit a simpler model using only those three variables as predictors. We can also implement methods like AIC or BIC to reduce our model as some varaibles that do not appear to be significant might be significant in the reduced model.

Problem 5

~~~
```{r}
m2 <- lm(wage ~ sex+union+occupation, data = CpsWages)
summary(m2)

par(mfrow=c(2,2))
plot(m2, which=c(1,2))
boxcox(m2)
```
~~~

From the box-cox plot, we can see that $\lambda = 0$ lies inside the confidence interval, so we could apply log-transformation to transform the target variable wage into log(wage), and the resulting model is

```{r}
m3 <- lm(log(wage) ~ sex+union+occupation, data = CpsWages)
summary(m3)

par(mfrow=c(1,2))
plot(m3, which=c(1,2))
```

We can see that the cone-shaped pattern does not appear in the residual vs fitted plot, meaning homoscedasticity is met; also, there are only 2 points not aligned with the QQ-line, meaning the normality condition is met. In addition, all variables are statistically significant from the model summary; hence, this simplified model is appropriate.

Problem 6
As we can see from the QQ-plot in problem 5, point 171 and 200 appear to be outliers. Removing those two points could somehow improve our model, but it would not alter our conclusion. As shown below, sex, union, and occupation are still significant.

```{r}
summary(lm(log(wage) ~ sex+union+occupation, data = CpsWages[-c(171,200),]))
```

---
title: "STAT5703 HW3 Exercise 4"
author: "Wen Fan(wf2255), Banruo Xie(bx2168), Hanjun Li(hl3339)"
output: pdf_document

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

Exercise 4.
Question 1.
$$var(\hat\epsilon_i) = var[Y_i-\hat Y_i] = E[(Y_i-\hat Y_i)^2]- E([Y_i-\hat Y_i])^2$$
$$var(\delta_i) = var[\hat Y_i-E[Y_i]] = E[(\hat Y_i-E[Y_i])^2]- E([\hat Y_i-E[Y_i]])^2$$
$$var(\delta_i)-var(\hat\epsilon_i) = E[(\hat Y_i-(E[Y_i])^2]- E[(Y_i-\hat Y_i)^2] =\\ E[(\hat Y_i-E[Y_i])^2]- E[RSS(\hat Y_i)]
$$

$$E[RSS(\hat Y_i] + \sum var(\delta_i)-\sum var(\hat\epsilon_i) =\sum E[(\hat Y_i-(E[Y_i])^2]
=\\ E[\sum (\hat Y_i-E[Y_i])^2] = \sigma^2\Gamma
$$

Question 2.
to prove $E[\frac{1}{\sigma^2}RSS(\hat Y)+2tr(S)-n] = \Gamma$, we only need to prove
$\sigma^2E[RSS(\hat Y)+2tr(S)-n] = \sigma^2\Gamma$, $\sigma^2(2tr(S)-n)=\sum var(\delta_i)-
\sum var(\hat\epsilon_i)\\ = E[(\delta-\delta_\mu)^T(\delta-\delta_\mu)]-E[(\hat \epsilon-\hat
\epsilon_\mu)^T(\hat \epsilon-\hat \epsilon_\mu)]$\\
$$\hat \epsilon = Y-\hat Y = (I-S)Y$$
$$\hat \epsilon_\mu=(I-S)E[Y] = (I-S)\mu$$
$$\delta = \hat Y-E[Y] = \hat Y - \mu = SY-\mu$$
$$\delta_\mu = SE{Y}-\mu=(S-I)\mu$$
Therefore, $$\hat \epsilon - \hat \epsilon_\mu = (I-S)Y - (I-S)\mu = (I-S)(Y-\mu)$$
$$\delta - \delta_\mu = S(Y-\mu)$$
Now, we can go back to $E[(\delta-\delta_\mu)^T(\delta-\delta_\mu)]-E[(\hat \epsilon-\hat
\epsilon_\mu)^T(\hat \epsilon-\hat \epsilon_\mu)]\\=E[(Y-\mu)^TS^TS(Y-\mu)] - E[(Y-\mu)^T(I-
S)^T(I-S)(Y-\mu)]\\=E[(Y-\mu)^T(-I+S+S^T)(Y-\mu)]\\=tr[(-I+S+S^T)cov(Y-\mu)]+(E[Y-\mu])^T(-
I+S+S^T)E[Y-\mu]\\=\sigma^2tr[-I+S+S^T] = \sigma^2(-n+2tr[s])$
proved.

Question 3.
$re(S) = tr(X^TX(X^TX)^{-1}) = tr(I_P) = p$
therefore $C_p = \frac{1}{\sigma^2}RSS(\hat Y)+2p-n$
Note that AIC = 2p-2l and $l = -\frac{1}{2}(n\log\sigma^2+\frac{1}{\sigma^2}RSS(\hat Y)+C$
So $AIC = \frac{1}{\sigma^2}RSS(\hat Y)+2p-n\log\sigma^2+C$ is similar to C_P if we treat
σ^2, n to constant.

Question 4.
We have $AIC(\hat \beta_{q+1})-AIC(\hat \beta) = 2-2(l_{q+1}-l_q)$ where $l_q = -
\frac{1}{2}(n\log(RSS(\hat\beta_q)+n-nlogn)$
Therefore $P(AIC(\hat \beta_{q+1})-AIC(\hat \beta))<0) \\=
P(2+nlog\frac{RSS(\hat\beta_{q+1})}{RSS(\hat\beta_q)}<0)\\=P(nlog(1-X_1^2/n)<-2)$
where n$\to \infty$
$P(nlog(1-X_1^2/n)<-2) = P(X_1^2>2)>0$

Question 5.
$BIC = -2l(\hat \theta)+plogn$, So $P(BIC(\hat\beta_{q+1})-BIC(\hat\beta_q)<0) =
\\P(logn+nlog\frac{RSS(\hat\beta_{q+1})}{RSS(\hat\beta_q)}<0)$
where n$\to \infty$
$P(logn+nlog\frac{RSS(\hat\beta_{q+1})}{RSS(\hat\beta_q)}<0) = P(logn-X_1^2<0)
\\P(X_1^2>logn)=0$

~~~
---
title: "HW3 Exercise 5"
author: "Wen Fan(wf2255), Banruo Xie(bx2168), Hanjun Li(hl3339)"
date: "4/15/2020"
output: pdf_document
---

```{r, warning=FALSE}
library(SMPracticals)
library(MASS)
data(pollution)
```

### Problem 1
```{r}
pairs(pollution)
```

Since the plots are too dense, we can not distinguish any patterns from it.

```{r}
pairs(pollution[,c(1:3,15:16)]) # association of mortality with weather
```

We do not see any patterns among weathers and mortality in these plots as the points are
scattered across the plots; we could not recognize if there is outlier either.

```{r}
pairs(pollution[,c(4:11,16)])   # and social factors
```

We don't see any significant covariates from the plots and we could not recognize if there is
outlier.
```{r}
pairs(pollution[,c(12:14,16)])  # and pollution measures
```

We still could not see any significant covariates from the plots, but we see that there might be
possbile outliers in **hc** and **noc** since one or two points are apart from the rest of the
points, which appear to form a cluster. We should consider transforming the dataset as most of
the features don't seem to correlate with mortality. The issue of outliers and collinearity among
the variables might arise in accounting for the effect of air pollution on mortality.

### Problem 2
~~~

```{r}
fit <- step(glm(mort~.-hc-nox-so,data=pollution))
```

The model associated with the lowest AIC contains variables prec, jant, jult, popn, educ, dens, and nonw, so we want to run a reduced model using those 7 variables

```{r}
summary(lm(mort ~ prec + jant + jult + popn + educ + dens + nonw, data = pollution))
```

The reduced model has an adjusted R-squared of 0.68, meaning 68% of the variance could be explained by those 7 variables. Keeping all else constant:

* an increase in average annual precipitation would increase the mortality and this change is statistically significant. It might sound plausible as high precipitation could cause flood.

* a decrease in average January temperature would increase the mortality and this change is statistically significant. It sounds plausible as the elderly are especially susceptible to cold weather.

* a decrease in average July temperature would increase the mortality and this change is statistically significant. It sounds plausible due to the same reason as the previous one.

* household size is not a significant variable, which is reasonable as household size has nothing to do with mortality.

* a decrease in median school years completed by those over 22 would increase the mortality and this change is statistically significant. It sounds plausible since the level of education might affect one's knowledge in nutrition and healthy diet.

* Population per square mile in urbanized areas in 1960 is not a significant variable.

* an increase in percentage non-white population in urbanized areas in 1960 would increase the mortality and this change is statistically significant, which is not reasonable.

```{r}
boxcox(fit)
```

The box-cox plot suggests that a log transformation might be appropriate as $\lambda=0$ falls inside the 95% confidence interval.

```{r}
plot.glm.diag(fit) # model adequate?
```

As we can see, homoscedasticity and normality condition are generally met as we don't see obvious pattern in the residual v.s. fitted value plot, nor do we see many points apart from the normal QQ-line. By checking the Cook's distance, we can see there are indeed some outliers.
```{r}
fit2 <- update(fit,log(mort)~.) # try log transform of response plot.glm.diag(fit) # model adequate?
summary(fit2)
```
We can see the conclusion is the same as the reduced model with no log-transformation applied.
```{r}
plot.glm.diag(fit) # model adequate?
```

Again, homoscedasticity and normality conditions are satisfied, so the log transform version of the reduced model is adequate as well.

Hence, using the reduced model resulted from step function and apply log transformation as suggested by the box-cox plot would be the chosen model.

Problem 3
```{r}
pairs(resid(lm(cbind(log(mort),hc,nox,so)~.,data=pollution)))
```

We see that **hc** and **nox** are not appropriate to be fitted using a linear model as there is obvious pattern in the residual plots, only **so** seems to have a significant linear relationship with the other variables and there seem to be outliers in all three pollution variables.
```{r}
fit3 <- lm(log(mort) ~ prec + jant + jult + popn + educ + dens + nonw + hc + nox + so, data = pollution)
summary(fit3)
```
After adding the three pollution variables to the reduced model, the adjusted r-squared has improved by around 3%. From the previous conclusion, we want to log-transform **hc** and **nox** to eliminate the patterns in the residual plot.
```{r}
fit4 <- lm(log(mort) ~ prec + jant + jult + popn + educ + dens + nonw + log(hc) + log(nox) + so, data = pollution)
summary(fit4)
```

~~~~
```
We can see the adjusted r-squared is further improved. Hence, the reduced model with added
variables **so**, **log(hc)**, and **log(nox)** is a better model.

### Problem 4
```{r}
rfit <- lm.ridge(log(mort) ~ prec + jant + jult + popn + educ + dens + nonw + log(hc) + log(nox) +
so, data=pollution,lambda=seq(0,20,0.01))
plot(rfit)
```

As we can see, as the penalty term $\lambda$ increases, the coefficient of the varialbes tend to
approache 0, and two of the variables are especially sensitive to the value of $\lambda$ as their
curvature are large.
```{r}
select(rfit)
```
Those three values are estimates of the penalty term using different methods.


### Problem 5
```{r}
lqs_fit <- lqs(log(mort) ~ prec + jant + jult + popn + educ + dens + nonw + log(hc) + log(nox) + so,
data=pollution)
lqs_fit
```
Using least trim square regression, we see that popn appear to be statistically more import than
prec, which doesn't agree with what we have previously.

```{r}
rlm_fit <-rlm(log(mort) ~ prec + jant + jult + popn + educ + dens + nonw + log(hc) + log(nox) +
so, data=pollution)
summary(rlm_fit)
```
From the t-values, we can see that using the robust M-estimation generally results in the same
conclusion as what we have previously.
~~~~