Stat 208 Final Exam

Jordan Berninger

5/29/2020

Question 1

(a) This model includes a location parameter, so we know that $\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i) = 0$. Estimated residuals are defined as $\hat{\epsilon_i} = Y_i - \hat{Y}_i$.

Accordingly, $\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i) = \frac{1}{n}\sum_{i=1}^{n}(\hat{\epsilon_i}) = 0$.

Therefore, $\sum_{i=1}^{n}(\hat{\epsilon_i}) = 0$.

(b) We have two cases to solve here, when n is even and when n is odd. We first consider the case where n is even. Note that we can factor the sum into pairs $\sum_{i=1}^{n} y_i = (y_1 + y_n) + (y_2 + y_{n-1}) + \ldots$. Accordingly, if we show that each pair sums to zero, then the whole series sums to zero. To do this, we will make use of the symmetry of ϕ^{-1}, which is plotted below. Because of symmetry, we know that $\phi^{-1}(1/2 + c) = -\phi^{-1}(1/2 - c) \forall c \in (0, 1/2)$.

We now consider the general pair in the summation, $\phi^{-1}(i - 1/2)/n + \phi^{-1}(n - i + 1 - 1/2)/n$. Now,

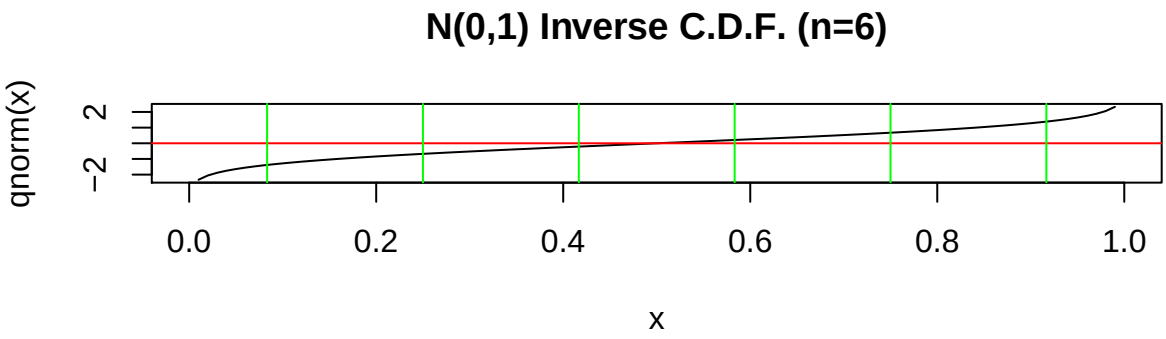
$$\begin{aligned}
\phi^{-1}(\frac{i-1/2}{n}) &= \phi^{-1}(\frac{1}{2} + \frac{i}{n} - \frac{1}{2n} - \frac{1}{2}) \\
&= \phi^{-1}(\frac{1}{2} + \frac{2i-1-n}{2n}) \\
\phi^{-1}(\frac{n-i+1-1/2}{n}) &= \phi^{-1}(\frac{1}{2} + \frac{n}{n} - \frac{i}{n} + \frac{1}{n} - \frac{1}{2n} - \frac{1}{2}) \\
&= \phi^{-1}(\frac{1}{2} + \frac{2n-2i+1-n}{2n}) \\
&= \phi^{-1}(\frac{1}{2} + \frac{n-2i+1}{2n})
\end{aligned}$$

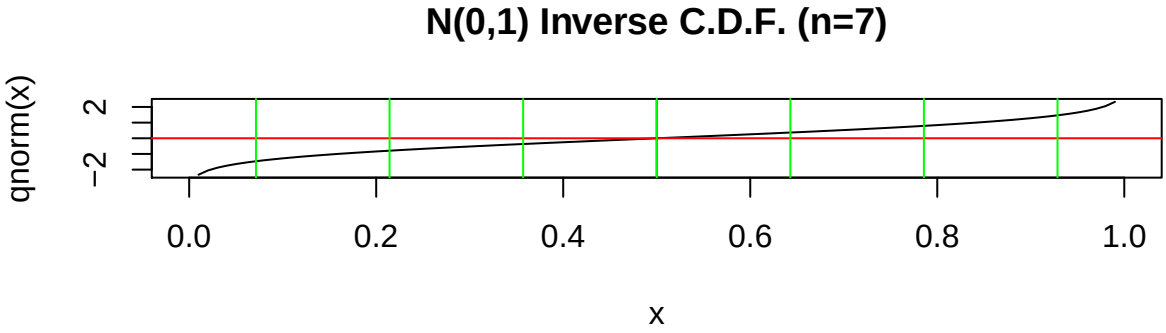
We can see the general pair is in the defired form: $\phi^{-1}(1/2 \pm c)$, and thus each pair in the sum where n is even evaluates to zero. Therefore we have shown $\sum_{i=1}^{n} \phi^{-1}(i - 1/2)/n = 0$ for even n.

To prove the odd case, we first consider the visual difference between an arbitrary even case (n=6) and an arbitrary odd case (n=7). Note that in the odd case, the middle line appears to fall at $x = 0.5$ and all the other lines appear to have a symmetric pair across the origin, as in the even case.

```
n=6
j=c(1:n)
par(mfrow=c(2,1))
x=seq(from=0,to=1,0.01)
plot(x,qnorm(x),type="l",main="N(0,1) Inverse C.D.F. (n=6)")
abline(h=0,col="red")
for(i in 1:n){
  abline(v=(j[i]-0.5)/n,col="green")
}
n=7
j=c(1:n)
```

```r
plot(x,qnorm(x),type="l",main="N(0,1) Inverse C.D.F. (n=7)")
abline(h=0,col="red")
abline(v=0.5,col="blue")
for(i in 1:n){
  abline(v=(j[i]-0.5)/n,col="green")
}
```





Now consider, when n is odd, our summation can be factored as, $\sum_{i=1}^{n} y_i = (y_1+y_n)+(y_2+y_{n_1})+\cdots+y_{(n+1)/2}$. The same math from the even pairs applies for the pairs in the odd case, so we know each pair in this summation evaluates to zero. Since n is odd, clearly, the $\frac{n+1}{2}$ element cannot find a match in this ordering. We note that $\phi^{-1}((n+1)/2-1/2)/n) = \phi^{-1}(1/2) = 0$, and therefore the entire series sums to zero when n is odd.

(c) First we note the the trigonometric identity: $cos(2\pi t/T) = \frac{1}{2sin(2\pi/T)}(sin(2\pi(t+1)/T) - sin(2\pi(t-1)/T))$. We plug this into the summation and see that all the interior terms cancel out, it follows that:

$$\begin{aligned}
\sum_{t=1}^{T} cos(2\pi t/T) &= \frac{1}{2sin(2\pi/T)}\Big(sin(2\pi(n+1)/T)) - \\
&\quad sin(2\pi n/T) - sin(2\pi/T) - sin(2\pi(1-1)/T)\Big) \\
&= \frac{1}{2sin(2\pi/T)}\Big(sin(2\pi(n+1)/T)) - sin(2\pi/T)\Big) \\
&= \frac{1}{2sin(2\pi/T)}\Big(sin(2\pi n/T)) - sin(2\pi/T)\Big) \\
&= 0
\end{aligned}$$

Problem 2

(a) We are given that $M\vec{v} = \vec{v} \forall$ n-dimensional vectors $\vec{v}$, and this is a very strong statement. This tells us that $\lambda = 1$ is the eignevalue for all n-dimensional vectors. We know that the sum of the eigenvalues $\sum^n \lambda_i = n = trace(A)$ and $\prod^n \lambda_i = 1 = det(A)$. We also know that $0 = det(A - \lambda \mathbf{I})$. We can factor $A = P\Lambda P^{-1}$, where Λ is a diagonal matrix constructed of the eigenvalues of A, and P contains the eigenvectors as columns. Therefore $P = P^{-1} = \lambda = \mathbf{I}$. It follows that A is the identity matrix.

(b) We are given that $A\vec{v} = B\vec{v} \forall \vec{v} \in R^n$ and that either matrix is invertible. Another strong statement. Consider the case where A is non-singular, without loss of generality. Now,

$$\begin{aligned} A\vec{v} &= B\vec{v} \\ A^{-1}A\vec{v} &= A^{-1}B\vec{v} \\ \mathbf{I}\vec{v} &= A^{-1}B\vec{v} \end{aligned}$$

This implies that $A_{-1}B = \mathbf{I}$, which means that $B^{-1} = A^{-1}$. Therefore, $A = B$ by the uniquness of matrix inverses.

Problem 3

We can assume that $Q' = Q$, therefore $Y'AY = (Y'AY)' = Y'A'Y$, which implies that $A' = A$. The Fisher Cochran Facts tell us that if A is a projection matrix, we know that $Y'AY \sim \chi^2(m)$, where $m = rank(A)$ if and only if $A\Sigma A = A$, where Σ is the covariance matrix of Y.

Problem 4

We know that Γ_0 is symmetric and therefore, there exists a diagonal matrix with all the eigenvalues, D, and orthogonal matrix with all the eigenvectors P, such that $P\Gamma_o P' = D$. It follows that $\Gamma_0^{-1} = P'D^{-1}P$ and thus $\Gamma_0^{-1/2} = (P'D^{-1}P)^{1/2}$.

This could also be accomplished with a Cholesky Decomposition. Since Γ_0 is symmetric, we know there exists a lower triangular matrix, L such that $\Gamma_0 = LL'$. It follows that $\Gamma_0^{-1/2} = L^{-1}$. A Cholesky defines L as $l_{i,i}\sqrt{a_{i,i} - \sum_{j=1}^{k-1} l_{k,j}^2}$ and $l_{k,j} = a_{k,i} - \sum_{j=1}^{i-1} a_{k,j}$. Lower triangular matrices are not the hardest to invert.

Problem 5

```
# S12 mauna loa model
{
  d=60#2018-1959+1
  T=12
  N=d*T
  p2=14 # updated this

  # Define Observations and Times
  # Initialize time (tax) and observation (Y) arrays
  tax=1:N
  tyr=1959+tax/12

  Y=loa$v3
```

```r
  #Plot the data
  # Seasonal effect matrix
  S=matrix(0,12,11)
  diag(S)=rep(1,11)
  S[12,]=rep(-1,11)
  # Define Design Matrix
  D2=matrix(0,N,p2)
  for(i in 1:N){
    D2[i,1]=1
    D2[i,2]=i
    D2[i,3]=i^2
    if(mod(i-1,12) == 0) {
      D2[i:(i+11),4:14] = S
    }
  }

  # Compute the OLS estimators
  H1.2=t(D2) %*% D2
  H2.2=solve(H1.2)
  H3.2= H2.2 %*% t(D2)
  theta_hat2= H3.2 %*% Y

  # Compute Estimated Y
  Yhat2= D2 %*% theta_hat2

  # Compute Residuals
  Resid2=Y-Yhat2

  # Compute Parameter Standard Errors
  se2=matrix(0,p2,1)
  SSE2=t(Resid2) %*% Resid2
  sighat2=SSE2/(N-p2)
  for(i in 1:p2){
    se2[i]=sighat2^(1/2)*H2.2[i,i]^(1/2)
  }

  R.sq2 <- 1-SSE2/sum((Y-mean(Y))^2)
}
# sample variance of the residuals as \hat{\sigma^2}
sigma.sq.hat<-var(Resid2)
# sample 1 lag correlation as the estimate of \hat{\rho}
rho.hat<-acf(Resid2,plot=F)$acf[2]
gamma.hat<-diag(N)
# create the \Gamma_0 matrix
for(i in 1:nrow(gamma.hat)){
  for(j in 1:ncol(gamma.hat)){
    gamma.hat[i,j]=rho.hat^(abs(i-j))
  }
}
gamma.hat.inv<-solve(gamma.hat)
# GLM Diagnostics Slide 17
beta.hat.gls<-solve(t(D2)%*%gamma.hat.inv%*%D2)%*%t(D2)%*%gamma.hat.inv%*%Y
## report all parameter estimators and R^2.
```

```
## Is the quadratic coefficient significantly positive? Report a p-value
Yhat3=D2%*%beta.hat.gls
Resid3=Y-Yhat3
se3=matrix(0,p2,1)
SSE3=t(Resid3) %*% Resid3
sighat3=SSE3/(N-p2)
for(i in 1:p2){
  se3[i]=sighat3^(1/2)*H2.2[i,i]^(1/2)
}

R.sq3 <- 1-SSE3/sum((Y-mean(Y))^2)
# parameter estimates
beta.hat.gls
```

```
##                [,1]
##  [1,]  3.150532e+02
##  [2,]  6.660800e-02
##  [3,]  8.866352e-05
##  [4,]  6.811523e-02
##  [5,]  6.945073e-01
##  [6,]  1.450437e+00
##  [7,]  2.584057e+00
##  [8,]  3.021856e+00
##  [9,]  2.311993e+00
## [10,]  6.857907e-01
## [11,] -1.477755e+00
## [12,] -3.172484e+00
## [13,] -3.253731e+00
## [14,] -2.043665e+00
```

This did not appear to change the model significantly.

```
# R^2
R.sq3
```

```
##          [,1]
## [1,] 0.999233
```

The R^2 is, again, extremely high.

```
# confidence interval for quadratic coefficient
alpha <- 0.05
beta_hat_ci_upp <- beta.hat.gls + qt(1-alpha/2, N-p2)*sqrt(se3)
beta_hat_ci_low <- beta.hat.gls - qt(1-alpha/2, N-p2)*sqrt(se3)
beta_hat_ci_low[3];beta_hat_ci_upp[3] # insignificant
```

```
## [1] -0.00159912
```

```
## [1] 0.001776447
```

The 95% confidence interval for the quadratic coefficient includes zero, indicating insignificance.

```r
# p-value for quadratic term
dt(beta.hat.gls[3],df=N-p2)
```

```
## [1] 0.398801
```

The p-value provides weak evidence to reject the null hypothesis that the quadratic term = 0.

Question 6

It is easier to write this one out by hand, see the attached image.

Question 7

We can use F-Tests with nested model for all of these tests. Following a homework problem, we first test the hypothesis that $\mu_1 = \mu_2 = \mu_3$ and $a_1 = a_2 = a_3$. Accordingly, our null model will only have 1 set of regression parameters, and the alternative model will have a set for each group.

```r
y1<-c(8.42,14.68,21.42,25.45,27.14,30.53,34.51,34.54,33.24,39.63,43.98,47.77)
y2<-c(9.86,9.54,11.96,12.46,11.38,14.69,16.48,20.11)
y3<-c(6.52,5.11,7.75,6.84,7.65,9.49,7.03,9.41,12.01)
x1<-c(1,3,5,6,7,8,9,9,10,11,12,14)
x2<-c(3,3,4,5,6,8,9,12)
x3<-c(2,5,7,8,10,15,16,18,20)

n1=length(y1)
n2=length(y2)
n3=length(y3)

mu1<-matrix(c(rep(1,n1),rep(0,n2),rep(0,n3)),ncol=1)
mu2<-matrix(c(rep(0,n1),rep(1,n2),rep(0,n3)),ncol=1)
mu3<-matrix(c(rep(0,n1),rep(0,n2),rep(1,n3)),ncol=1)
a1<-matrix(c(x1,rep(0,n2),rep(0,n3)),ncol=1)
a2<-matrix(c(rep(0,n1),x2,rep(0,n3)),ncol=1)
a3<-matrix(c(rep(0,n1),rep(0,n2),x3),ncol=1)

D<-cbind(mu1,mu2,mu3,a1,a2,a3)

n<-n1+n2+n3
y<-matrix(c(y1,y2,y3))
xs<-matrix(c(x1,x2,x3))
r<-3

# complete null model, just an intercept term and the x values
x.o<-matrix(c(rep(1,n),xs),ncol=2)
m.o<-x.o%*%solve(t(x.o)%*%x.o)%*%t(x.o)
m.a<-D%*%solve(t(D)%*%D)%*%t(D)
F.stat<-((t(y)%*%(m.a-m.o)%*%y)/(r-1))*((n-r)/(t(y)%*%(diag(n)-m.a)%*%y))
F.stat
```

```
##          [,1]
## [1,] 1578.608
```

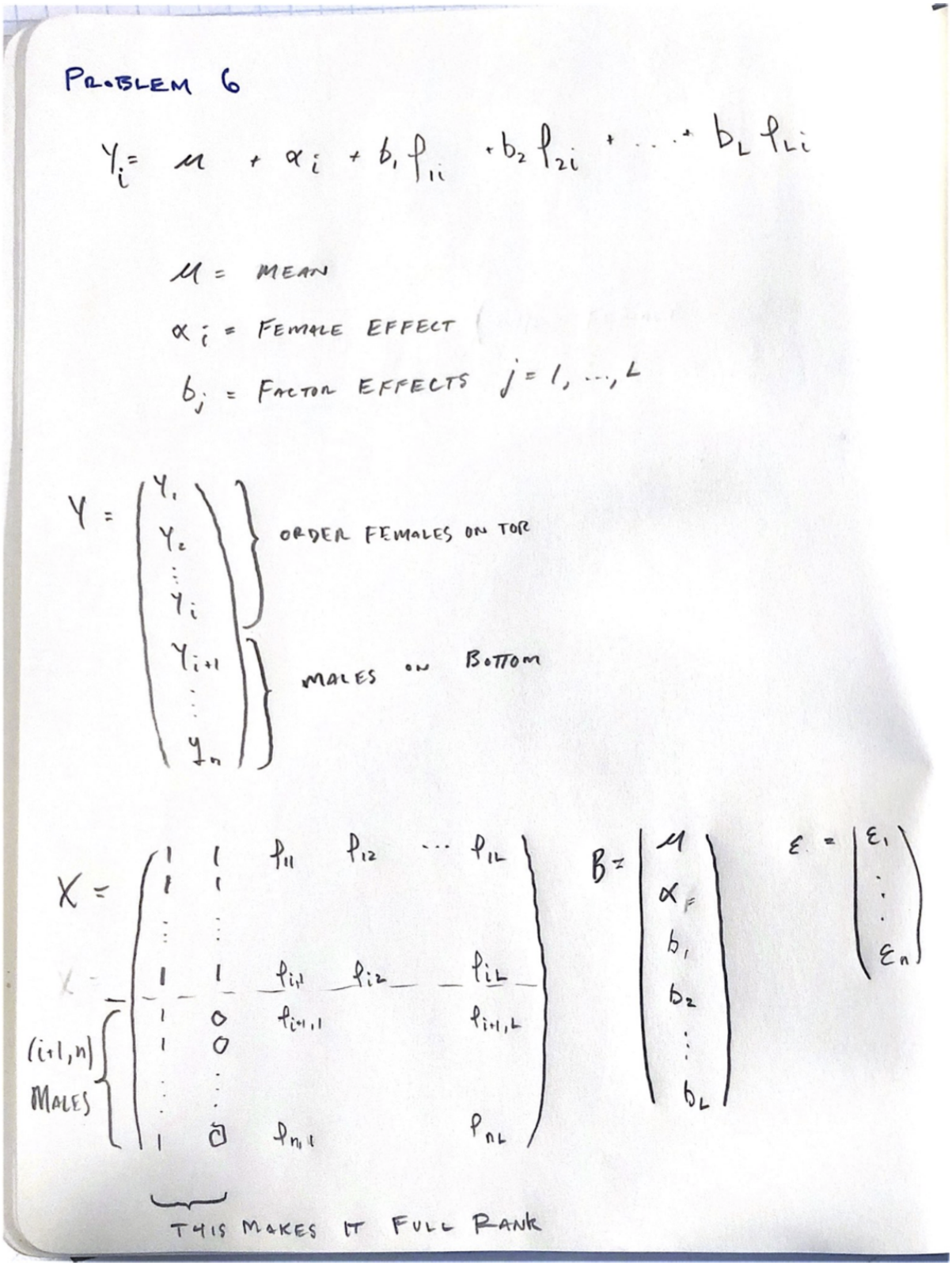


Figure 1: Problem 6

```r
qf(p=0.95,df1=r-1,df2=n-r)
```

```
## [1] 3.369016
```

```r
F.stat > qf(p=0.95,df1=r-1,df2=n-r) # If true, reject Ho
```

```
##      [,1]
## [1,] TRUE
```

This F-Test rejects the null hypothesis and indicates that fitting different regression terms for each group results in a superior model.

Final Question

This dataset is a time series of 139 data points. The units are unknown. We denote the index x and the variable of interest y. This analysis focuses on determing the relationship between y and x. We consider whether a linear or quadratic relationship is more appropriate. We also consider whether seasonal effects modeled sinusiodal, and periodic factor allow the model to explain more of the response data. The scatterplot shows that y increases with the index.

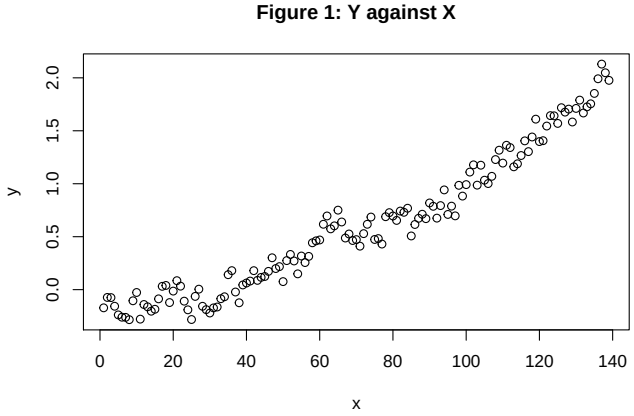
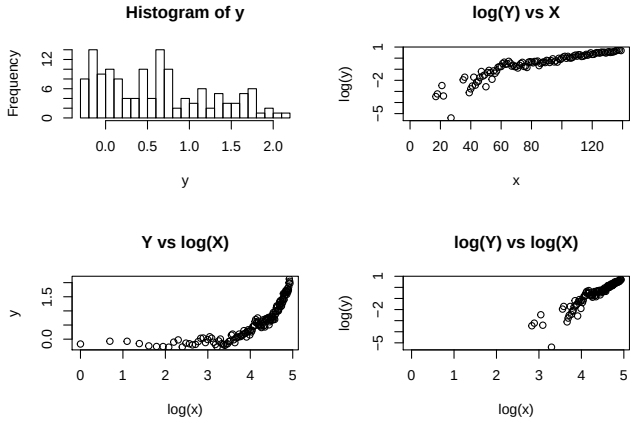


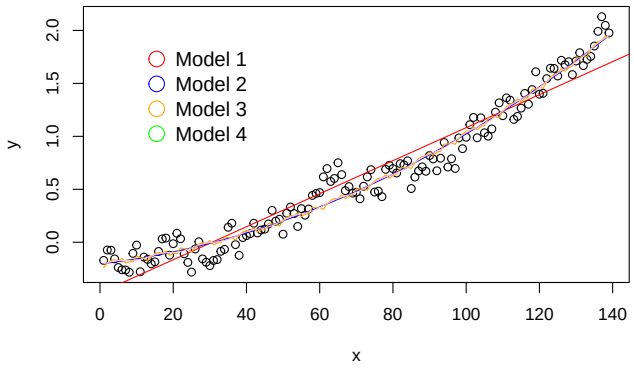
Figure 1 provides strong evidence to fit a model on the untransformed variables, however, our exploratory data analysis should not end there. The following histogram shows that y has a multimodal, asymmetric distiburion, and the additional scatterplots hint at a strong relationshoip between the log-transformed variables.



The models under consideration in this analysis are: Model 1: $y = \mu + ax$, Model 2: $y = \mu + ax + bx^2$, Model 3: $y = \mu + ax + bx^+ccos(2\pi x/T) + dsin(2\pi x/T)$ (where $T = 12$), Model 4: $y = \mu + ax + bx^2 + cS_1 + \cdots + nS_{12}$. These models were all studied this quarter. It is worth noting that I also fit Model 3 with different values of T, and the following results apply to all models of this family.

The 4 models are fit and Figure 3 shows that they all appear to fit the data well. We apply several methods to select the best model.

Figure 3: Data and Models 1−4



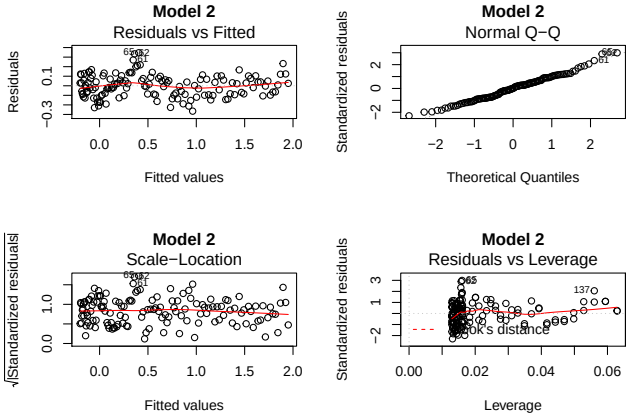
Since the 4 models under consideration qualify as nested models, it is appropriate to apply an F-Test for model selection. This is done with the *anova*() function, which tells us that model 2 models significantly more response variance than the other models.

```
anova(m1,m2,m3,m4)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ 1 + x
## Model 2: y ~ 1 + x + I(x^2)
## Model 3: y ~ 1 + x + I(x^2) + factor(mod(x - 1, 12))
## Model 4: y ~ 1 + x + I(x^2) + cos((2 * pi * (x))/(12)) + sin((2 * pi *
##     (x))/(12))
##   Res.Df    RSS Df Sum of Sq        F Pr(>F)
## 1    137 3.8181
## 2    136 1.8430  1   1.97510 138.5383 <2e-16 ***
## 3    125 1.7821 11   0.06089   0.3883 0.9586
## 4    134 1.8344 -9  -0.05232   0.4077 0.9290
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The *step*() function optimizes a model based on AIC, and model 2 is also deemed the best by this evaluation.

The residual plots for all 4 models were evaluated, but for the sake of brevity, only the residual plots of model 2 are inspected here. The residual plots show that the residuals are approximately centered at zero across the range of fitted values. The residuals also appear homoskedastic. The residuals closely match the theoretical quantiles in the QQ plot.



In Question 5, we developed a Generalized Least Squares estimate for a similar model by takings the varriance of the fitted residuals. We follow the same process here to and evaluate the resultant model.

The GLS model coefficients are very similar to those of model 2. However, the 95% confidence interval for the quadratic term includes zero, and the p-value indicates that the quadratic term is not-significant.

```
alpha <- 0.05
beta_hat_ci_upp <- beta.hat.gls + qt(1-alpha/2, N-p2)*sqrt(se3)
beta_hat_ci_low <- beta.hat.gls - qt(1-alpha/2, N-p2)*sqrt(se3)
c(beta_hat_ci_low[3],beta_hat_ci_upp[3])
```

```
## [1] -0.005095142  0.005262029
```

```
dt(beta.hat.gls[3],df=N-p2)
```

```
## [1] 0.3982096
```

I conclude that model 2 is the most appropriate, and that y is increasing quadratically with respect to x. Hopefully this is a time series for a World Peace and Happiness metric, and not something like temperature. The F-tests between models indicated that this model accounts for the most response variance, and the AIC selection determined each term (intercept, linear, quadratic) is statistically significant. While the GLS model determined that the quadratic term is insignificant, this model used insight from model 2's residuals. The best models's metrics and coefficients are printed below.

```
summary(m2)
```

```
##
## Call:
## lm(formula = y ~ 1 + x + I(x^2))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.26552 -0.08953  0.00266  0.07624  0.34549
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.046e-01  3.005e-02  -6.809 2.88e-10 ***
## x            4.000e-03  9.910e-04   4.036 9.02e-05 ***
## I(x^2)       8.279e-05  6.857e-06  12.073  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1164 on 136 degrees of freedom
## Multiple R-squared:  0.9683, Adjusted R-squared:  0.9679
## F-statistic:  2080 on 2 and 136 DF,  p-value: < 2.2e-16
```