**** Target variable 이 normality 하지 않을 경우 일산 선형회귀분석 시의 문제점 고찰**

**** 최종적으로, 분류모델을 구축하기 위한 과정에서, 각 변수들 사이의 관계성을 고민해보고, 모델이 설명할 수 있는 영역을 명확히 하는 것은 중요하다. 단, 데이터가 설명하고 있는 데이터 단위를 줄이는 것 역시 실제 활용성 측면에서 중요**

문제 : 가령 현재 데이터 구조만 볼 시 다중 회귀모델을 구축해볼 수 있겠으나(+ Subset regression 을 통한 fitting),

Independent variable 의 몇가지 조건 특히, normality 하지 않을 경우 설명력은 줄어든다.

해결과정

- Skewness or Kurtosis 의 문제일 수 있다.
- Existence of a few outliers / extreme values

: 현재의 고려요건은 1 번 Cas 에 해당된다고 볼 수 있음

If there are also problems with heterogeneity of variance : 분산의 이질성에 대한 문제.

해결방책 나열 : ***1. marginal models***

: linear model fit with generalized least square estimation

(where you can define the correlation pattern)

: GEE and IPW(inverse probability weighting) with GLS, CR2, CR3

모델선택과정에서의 주된 고려요소

동분산성(homoscedasticity)

동분산성(homoscedasticity) 가정은 **잔차의 분산이 일정한 상수**임을 가정하는 것이다. 즉, 분산이 일정한 상수로 동일해야 한다. 동분산성이 어긋나면 이분산성(heteroscedasticity)이 나타난다. 그러나 **이분산성** 문제는 상대적으로 중요한 문제는 아니다. 왜냐하면 이분산성이 어긋나도 **추정계수는 불편추정치를 만족**하기 때문이다.

그러나 이 경우 효율성은 만족하지 않는다. 이분산성을 치료하는 것은 상대적으로 간단하다.

Generalized Least Square(GLS)가 그 방법이다.

Weighted Least Square(WLS)는 GLS 의 특수한 경우이다. 이는 오차항의 분산이 어떠한 함수를 갖는지를 알고서 그 함수를 회귀식의 양변에 나누는 방식이다. 그러나 여기서 주의할 점은 잔차와 오류항은 같지 않다는 것이다. 잔차는 어디까지나 추정식을 추정한 그 결과로 계산된 것인 반면, 오류항은 알 수 없는 미지의 것이기 때문이다. 때문에 오차항이 아니라 잔차를 이용하게 되고, 이 경우를

feasible Generalized Least Square(FGLS)라고 부른다. 그러나 오차항은 연구자가 알지 못하는 영역인데 알고 있다고 본다는 점에서 문제적일 수 있다.

If we assume the errors take a simple auto-regressive form such that each error is correlated with the prior:

$$\epsilon_{i+1} = \phi\epsilon_i + \delta_i$$

Where $\delta_i \sim N(0, \tau^2)$

We can estimate ϕ from the model.

And, below formula is for Sigma matrix

$$\Sigma_{ij} = \phi^{|i-j|}$$

Testcode1

```
library(faraway)
globwarm <- na.omit(globwarm)
colnames(globwarm)
```

```
##  [1] "nhtemp"     "wusa"       "jasper"     "westgreen"  "chesapeake"
##  [6] "tornetrask" "urals"      "mongolia"   "tasman"     "year"
```

@ ST1: Fit our ordinary least squares

```
lmod <- lm(nhtemp~ wusa + jasper + westgreen +
             chesapeake + tornetrask + urals +
             mongolia + tasman, data= globwarm)
summary(lmod)
```

```
##
## Call:
## lm(formula = nhtemp ~ wusa + jasper + westgreen + chesapeake +
##     tornetrask + urals + mongolia + tasman, data = globwarm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.43668 -0.11170  0.00698  0.10176  0.65352
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.242555   0.027011  -8.980 1.97e-15 ***
## wusa         0.077384   0.042927   1.803 0.073647 .
## jasper      -0.228795   0.078107  -2.929 0.003986 **
## westgreen    0.009584   0.041840   0.229 0.819168
## chesapeake  -0.032112   0.034052  -0.943 0.347346
## tornetrask   0.092668   0.045053   2.057 0.041611 *
## urals        0.185369   0.091428   2.027 0.044567 *
## mongolia     0.041973   0.045794   0.917 0.360996
## tasman       0.115453   0.030111   3.834 0.000192 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1758 on 136 degrees of freedom
## Multiple R-squared:  0.4764, Adjusted R-squared:  0.4456
## F-statistic: 15.47 on 8 and 136 DF,  p-value: 5.028e-16
```

@ ST2: The concern in that the errors are correlated aproxy for being able to determine how to errors behave is to look at the residual In this case residuals lagged by one year so, this looks at years one through n -1
assume the errors take a **simple auto-regressive form** such that each error is **correlated with the prior**

```
# error(i+1) = pi*errori + (pi times the previous error)
n <- length(residuals(lmod))
cor(residuals(lmod)[-1], residuals(lmod)[-n])

## [1] 0.583339
```

@0.583339 : very strong correlation @ ST3 fit it__0.583339 using GLS ? Fit it by basically estimating sigma >> Incorporately that correlation appropriately

```
# Sigma(ij) = pi^|i-j| : Sigma Matrix using this exact equation
X <- model.matrix(lmod)
Sigma <- diag(n) # creat empty matrix with n
Sigma <- 0.5833^abs(row(Sigma) - col(Sigma)) # = pi^|i-j|

y <- globwarm$nhtemp
Sigma_inv <- solve(Sigma)                                  #__ inverse of sigma
XTX_inv <- solve(t(X) %*% Sigma_inv %*% X)
betahat <- XTX_inv %*% t(X) %*% Sigma_inv %*% y      #__ This X, transposed X
inverse that include that sigma and
                                                           # sandwiched in between
the Xs
betahat

##                      [,1]
## (Intercept) -0.234134783
## wusa         0.068425906
## jasper      -0.218438446
## westgreen    0.003880871
## chesapeake  -0.014952072
## tornetrask   0.057691347
## urals        0.222078555
## mongolia     0.055247801
## tasman       0.122999856
```

coefficients for this mdoel

```
# (Intercept) -0.234134783
# wusa         0.068425906
# jasper      -0.218438446
# westgreen    0.003880871
# chesapeake  -0.014952072
# tornetrask   0.057691347
# urals        0.222078555
# mongolia     0.055247801
# tasman       0.122999856
```

@ ST4 Check what the correlation is

```
res <- y - X %*% betahat
cor(res[-1], res[-n])

## [1] 0.5887776

# 0.5887776
```

@--

```
# another way to do this
S <- chol(Sigma)     #__square root to get pi
S_inv <- solve(t(S))
SX <- S_inv %*% X
Sy <- S_inv %*% y
lm(Sy ~ SX - 1)

##
## Call:
## lm(formula = Sy ~ SX - 1)
##
## Coefficients:
## SX(Intercept)         SXwusa       SXjasper    SXwestgreen   SXchesapeake
##     -0.234135       0.068426      -0.218438       0.003881      -0.014952
##  SXtornetrask        SXurals     SXmongolia       SXtasman
##      0.057691       0.222079       0.055248       0.123000

matrix(lm(Sy ~ SX -1)$coef)

##               [,1]
##  [1,] -0.234134783
##  [2,]  0.068425906
##  [3,] -0.218438446
##  [4,]  0.003880871
##  [5,] -0.014952072
##  [6,]  0.057691347
##  [7,]  0.222078555
##  [8,]  0.055247801
##  [9,]  0.122999856
```

```r
library(nlme)
glmod <- gls(nhtemp ~ wusa + jasper + westgreen +
               chesapeake + tornetrask + urals +
               mongolia + tasman,
             correlation =corAR1(form = ~year),
             data= globwarm)
summary(glmod)
```

```
## Generalized least squares fit by REML
##   Model: nhtemp ~ wusa + jasper + westgreen + chesapeake + tornetrask +
urals + mongolia + tasman
##   Data: globwarm
##         AIC       BIC   logLik
##   -108.2074 -76.16822 65.10371
##
## Correlation Structure: AR(1)
##  Formula: ~year
##  Parameter estimate(s):
##       Phi
## 0.7109922
##
## Coefficients:
##                   Value  Std.Error   t-value p-value
## (Intercept) -0.23010624 0.06702406 -3.433188  0.0008
## wusa         0.06673819 0.09877211  0.675678  0.5004
## jasper      -0.20244335 0.18802773 -1.076668  0.2835
## westgreen   -0.00440299 0.08985321 -0.049002  0.9610
## chesapeake  -0.00735289 0.07349791 -0.100042  0.9205
## tornetrask   0.03835169 0.09482515  0.404446  0.6865
## urals        0.24142199 0.22871028  1.055580  0.2930
## mongolia     0.05694978 0.10489786  0.542907  0.5881
## tasman       0.12034918 0.07456983  1.613913  0.1089
##
##  Correlation:
##             (Intr) wusa   jasper wstgrn chespk trntrs urals  mongol
## wusa        -0.517
## jasper      -0.058 -0.299
## westgreen    0.330 -0.533  0.121
## chesapeake   0.090 -0.314  0.230  0.147
## tornetrask -0.430  0.499 -0.197 -0.328 -0.441
## urals       -0.110 -0.142 -0.265  0.075 -0.064 -0.346
## mongolia     0.459 -0.437 -0.205  0.217  0.449 -0.343 -0.371
## tasman       0.037 -0.322  0.065  0.134  0.116 -0.434  0.416 -0.017
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -2.31122523 -0.53484054  0.02342908  0.50015642  2.97224724
##
```

```
## Residual standard error: 0.204572
## Degrees of freedom: 145 total; 136 residual

intervals(glmod, which = "var-cov")

## Approximate 95% confidence intervals
##
##  Correlation structure:
##         lower      est.     upper
## Phi 0.5099744 0.7109922 0.8383752
## attr(,"label")
## [1] "Correlation structure:"
##
##  Residual standard error:
##     lower      est.     upper
## 0.1540709 0.2045720 0.2716263

# Correlation structure:
#   lower      est.     upper
# Phi 0.5099744 0.7109922 0.8383752
```