Improving High School Math Education in Portugal in GLM Framework

Heqiao Ruan
Department of Statistics, University of California, Davis

Background

- Portugal, a country located in southwestern Europe
- Statistics keeps high school education quality in Portugal at tail end in Europe
- Especially serious problems in students' **failure rate and dropout rate**
- Officials want to improve math education
- Our goal: identify important factors influencing students' math grade
- Does they match common sense?
- Propose corresponding advice to help government to overcome difficulties

Dataset

- **Student Performance Dataset**
- Depicts students' achievement collected using school reports and questionnaires
- 395 observations,33 features
- Response: G1(First Stage),G2(2rd Stage), G3(Final Grade)
- 30 Predictors: Most Categorical
- Distribution of Grade Level(G3):

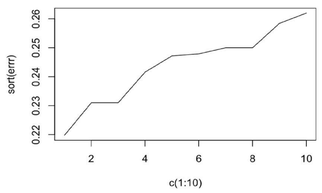
A(G3≥16)	$B(13 < G3 < 16)$	$C(11 < G3 < 14)$	$D(9 < G3 < 12)$	$F(G3 < 10)$
40	60	62	103	130

A-D Pass, F Fail

Method

- **Logistic Regression Model**: Fail(0,F)~Pass(1), 2-category classification /Probability of Passing on G3
- Prediction Error, 10-fold Cross Validation
- Model Selection by AIC

Predic/Level	Fail	Pass
Fail	60	25
Pass	70	240



Prediction Error 24.0% CV Error: 25.6%

Final Model after stepAIC:

$$Logit(E[Y|X]) \sim \beta_0+\beta_1 I(sexM)+\beta_2 age+\beta_3 I(Mjobhealth)+\beta_4 I(Mjobother)+\beta_5 I(Mjobservice)+\beta_5 I(Mjobteacher)+\beta_6 failures+\beta_7 I(schoolsupYes)+\beta_8 I(famsupYes)+\beta_9 I(higherYes)+\beta_{10} goout+\beta_{11} health+\beta_{12} failures * I(schoolsupYes)$$

Why add interaction? Schoolsupport effect contradict common sense

Method

Multinomial Model with 5-category: Proportional Odds Model, Baseline Odds Model on G3

Prediction Error:52.4% for Baseline Odds Model, 56.2% for Proportional OddsModel

Merge Categories

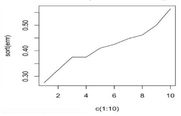
A:High, BCD:Medium, F:Low 3-category

Low(G3<10)	Medium($10 \leq G3 < 16$)	High($G3 \geq 16$)
130	225	40

Baseline Odds Model: Severe Lack of Fit. **Choose Proportional Odds Model**

Prediction Error:36.2% CV Error(10-fold):40.2% Performance Measurements:

Predic/Level	Low	Medium	High
Low	57	32	0
Medium	73	193	38
High	0	0	2



Accuracy	Precision	Recall	F-Score	F-Score($\beta = 0.5$)
0.638	0.649	0.703	0.676	0.683

Final Model after model selection :

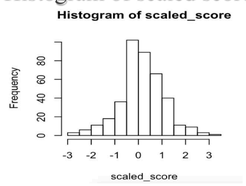
$$logit(P(Y \leq k)) = \beta_{0,k}+\beta_1 I(sexM)+\beta_2 age+\beta_3 I(PstatusT)+\beta_4 I(Mjobhealth)+\beta_5 I(Mjobother)+\beta_6 I(Mjobservice)+\beta_7 I(Mjobteacher)+\beta_8 studytime+\beta_9 failures+\beta_{10} I(schoolsupYes)+\beta_{11} I(famsupYes)+\beta_{12} I(higheryes)+\beta_{13} freetime+\beta_{14} goout+\beta_{15} health+\beta_{16} failures * I(schoolsupYes)$$

Transformed Model: Want to extract information from G1,G2 as well as G3

Perform PCA on cbind(G1,G2,G3)->Extract first principal component Y-> Scale Y to percent(0-100%) ->logit transformation on Y->Scaled Score

Scale-Score=logit(PERCENT(0.4629G1+0.5614G2+0.6859G3))

Histogram of scaled score: Approximately Normal, do **ordinary lm** on Scale-Score!



Summary of PCA: Table 10:PCA of G1,G2,G3:

Factor	PC1	PC2	PC3
Proportion of Variance	0.9095	0.06162	0.02892
Cumulative Proportion	0.9095	0.9711	1
G1	0.4629	0.8024	-0.3764
G2	0.5614	0.0632	0.8251
G3	0.6859	-0.5933	-0.4212

Final model after model selection:

$$E[Y|X] = E[ScaledScore|X] = \beta_0+\beta_1 sexM+\beta_1 I(Mjobhealth)+\beta_2 I(Mjobother)+\beta_3 I(Mjobservice)+\beta_4 I(Mjobteacher)+\beta_5 studytime+\beta_6 failures+\beta_7 I(schoolsupYes)+\beta_8 I(famsupYes)+\beta_9 I(higherYes)+\beta_{10} goout+\beta_{11} failures * schoolsupyes+\beta_{12} freetime+\beta_{13} health$$

Result(Important Predictors)

3.Logit Transformed Model after model selection

Predictors	Coefficient	Std.Error	t value	pvalue
sexM	0.251	0.092	2.735	6.53e-3
studytime	0.147	0.053	2.784	5.64e-3
failures	-0.418	0.0643	-6.506	2.5e-10
schoolsupyes	-0.465	0.137	-3.401	7.45e-4
famsupyes	-0.212	0.0872	-2.433	0.0154
romanticyes	-0.218	0.0890	-2.446	0.015
goout	-0.119	0.039	-3.098	0.002
health	-0.063	0.0299	-2.087	0.038
higher	0.750	0.632	1.186	0.236
failures:schoolsupyes	0.403	0.167	2.408	0.017

Result

1.Logistic Regression Model:

Predictors	Coefficient	Standard Error	z value	P Value
age	-0.217	0.108	-2.002	0.045
sexM	0.569	0.268	2.126	0.033
failures	-1.233	0.226	-5.460	4.76e-8
schoolsupyes	-1.334	0.385	-3.462	5.36e-4
goout	-0.346	0.114	-3.039	2.37e-3
higheryes	0.965	0.588	1.641	0.100
failures:schoolsupyes	1.412	0.475	2.982	2.87e-3

2.Proportional Odds Type Model after model selection

Predictors	Coefficient	Standard Error	z value	PVALUE
sexM	0.562	0.239	2.345	0.0195
age	-0.196	0.095	-2.065	0.396
failures	-1.278	0.221	-5.77	1.62e-8
schoolsupyes	-1.404	0.361	-3.888	1.19e-4
goout	-0.346	0.107	-2.047	0.041
higheryes	0.904	0.572	1.58	0.057
failures:schoolsupyes	1.478	0.447	3.309	1.027e-3

2(I):Intercepts

Prediction	Value	Std.Error	t value	pvalue
1|2	-4.706	1.863	-2.526	0.012
2|3	-1.163	1.842	-0.633	0.5270

Conclusion and Discussion

- Pay more attention to students who have fail before and provide extra support
- Pay more attention to Girls' study
- Arouse students' motivation to pursue higher degree, potentially help reduce dropout rate
- Extra Family support on math impose negative effect on students' grade. Avoid parents intervening
- Study more time, don't party too much

REFERENCES

[1].Paulo Cortez and Alice Silva. Using Data Mining to Predict Secondary School Student Performance. University of Minho Guimaraes, Portugal
[2].Hans-Georg Mueller. Generalized Linear Models Lecture Notes. UC Davis Winter 2018

Acknowledgement

I would like to express my thankfulness to Prof Hans Georg Mueller and TA Yaqing.

STA 232B - APPLIED STATISTICS

FINAL PROJECT

Analyses of the Iowa Crops Data

***Authors**:*
Alphabetical ordering

Rui HU
Heqiao RUAN
Tesi XIAO
Zitong ZHANG
Yejiong ZHU

***Instructor**:*
Dr. Jiming JIANG

March 19, 2019

Contents

1 Reading

1.1 Paper Summary

This paper considers the problem of transforming satellite information into good estimates of crop areas at the individual pixel and segment levels. The authors analyzed the data about corn and soybeans from both farm-level survey in 1978 June and land observatory satellites (LANDSAT) during the 1978 growing season of 12 Iowa counties. A linear regression model is specified for the relationship between the reported hectares of corn and soybeans in the survey and the corresponding satellite determination of them. The correlation structure within the counties is given by a nested-error model, i.e. the mean hectares of the crop per segment in a county is defined as the conditional mean of reported hectares given the satellite data and the realized (random) county effect. Based on the proposed model, the authors defined and estimated the variance-component and obtained the generalized least-squares estimators. Predictions of mean hectares of corn and soybeans and their standard errors were obtained as well.

The components-of-variance model considered in the paper is

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + u_{ij}$$

where $i = 1, \cdots, 12$ denotes the county; $j = 1, \cdots, n_i$ denotes the segment, n_i is the number of segments in the ith county; y_{ij} is the number of hectares of crops in the jth segment of the ith county as reported in the June Survey; x_{1ij} and x_{2ij} are the number of pixels classified as corn and soybeans, respectively, in the j-th segment of the i-th county. The random error u_{ij} can be written as

$$u_{ij} = v_i + e_{ij}$$

where $v_i \sim N\left(0, \sigma_v^2\right)$ is the i-th county effect and $e_{ij} \sim N\left(0, \sigma_e^2\right)$ is the random effect associated with the j-th sample segment within the i-th county. Thus, the model expressed in matrix notation is $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$.

The authors also considered several other circumstances: (i) other correlation structures considering geographical distance between segments; (ii) a multivariate framework that considers the correlation between reported areas of crop; (iii) the model including quadratic terms of the numbers of pixels of the crops. But none of these circumstances improve the precision of the estimation or have statistically significant results.

The sample mean of the reported hectares of the two crops $\overline{y}_{i.} = n_i^{-1} \sum_{j=1}^{n_i} y_{ij}$ can be expressed as

$$\overline{y}_{i.} = \beta_0 + \beta_1 \overline{x}_{1i.} + \beta_2 \overline{x}_{2i.} + v_i + \overline{e}_{i.}$$

where $\overline{x}_{1i.} \equiv n_i^{-1} \sum_{j=1}^{n_i} x_{1ij}$ and $\overline{x}_{2i.} \equiv n_i^{-1} \sum_{j=1}^{n_i} x_{2ij}$ are the sample mean numbers of pixels of two crops, respectively, within county i, and $\overline{e}_{i.} \equiv n_i^{-1} \sum_{j=1}^{n_i} e_{ij}$ is the sample mean of the within county effects in the i-th county.

In addition, the population mean hectares of two crops in the i-th county (y_i) can be defined as the conditional mean given the realized county effect v_i and the values of the satellite data:

$$y_i \equiv \beta_0 + \beta_1 \overline{x}_{1i(p)} + \beta_2 \overline{x}_{2i(p)} + v_i$$

where $\overline{x}_{1i(p)} \equiv N_i^{-1} \sum_{j=1}^{N_i} x_{1ij}$ and $\overline{x}_{2i(p)} \equiv N_i^{-1} \sum_{j=1}^{N_i} x_{2ij}$ are the population mean numbers of pixels classified as two crops, respectively, in the i-th county, which are known. The focus of this paper is to predict the mean crop hectares per segment.

The best predictor of v_i is the conditional expectation of v_i given the sample mean $\overline{u}_{i.}$. Suppose the variance σ_v^2 and σ_e^2 are known, then the generalized least-squares estimator of β is $\tilde{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}$, where $\mathbf{V} = E(\mathbf{u}\mathbf{u}')$. Then a possible predictor of the i-th county effect v_i is

$$\tilde{v}_i = \tilde{u}_{i.} g_i$$

where $\tilde{u}_{i.} = n_i^{-1} \sum_{j=1}^{n_i} \tilde{u}_{ij}, \tilde{u}_{ij} = y_{ij} - \mathbf{x}_{ij}\tilde{\boldsymbol{\beta}}$, and $g_i = \frac{\sigma_v^2}{\sigma_v^2 + n_i^{-1}\sigma_e^2}$. Then the corresponding predictor $\tilde{y}_i$ is

$$\tilde{y}_i = \overline{\mathbf{x}}_{i(p)}\tilde{\boldsymbol{\beta}} + \tilde{v}_i$$

which is the best linear unbiased predictor (BLUP) of y_i and have the variance of the error as

$$\mathrm{E}\left\{\left(\tilde{y}_i - y_i\right)^2\right\} = \sigma_v^2\left(1 - g_i\right) + \mathbf{c}_i\mathbf{V}(\tilde{\boldsymbol{\beta}})\mathbf{c}_i'$$

where $\mathbf{V}(\tilde{\boldsymbol{\beta}}) = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}$, $\mathbf{c}_i = \overline{\mathbf{x}}_{i(p)} - g_i\overline{\mathbf{x}}_i$ and $\overline{\mathbf{x}}_\mathbf{i} = n_i^{-1}\sum_{j=1}^{n_i}\mathbf{x}_{ij}$

Consider a class of predictors of the county mean crop area y_i as

$$N_i^{-1}\left[\sum_{j=1}^{n_i} y_{ij} + \sum_{j=n_i+1}^{N_i}\left(\mathbf{x}_{ij}\tilde{\boldsymbol{\beta}} + \tilde{v}_i\right)\right]$$

- When $\delta_i = g_i$ and $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}$, this is the BLUP above.
- When $\delta_i = 0$, this predictor is called the regression synthetic predictor.
- When $\delta_i = 1$, this predictor is called the survey regression predictor.

Since the variances σ_v^2 and σ_e^2 in the nested-error model are unknown, they are estimated by the residual mean square for the regression model. One choice for $\hat{\sigma_e^2}$ is

$$\hat{\sigma}_e^2 = \hat{\mathbf{e}}'\hat{\mathbf{e}}\left[\sum_{i=1}^{T}\left(n_i - 1\right) - 2\right]^{-1}$$

where $\hat{\mathbf{e}}'\hat{\mathbf{e}}$ is the residual sum of squares for the regression of the y deviations, $y_{ij} - \overline{y}_{i.}$, on the x deviations, $\mathbf{x}_{ij} - \overline{\mathbf{x}}_i$ for the counties with more than one samples. Then

$$d_e\frac{\hat{\sigma}_e^2}{\sigma_e^2} \sim \chi^2(d_e)$$

where $d_e \equiv \sum_{i=1}^{T}\left(n_i - 1\right) - 2$. In order to get an estimator of σ_v, consider the average of the ordinary least-squares residuals for county i:

$$\overset{\Delta}{u}_{i.} = \overline{y}_{i.} - \overline{\mathbf{x}}_i\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}$$

and the weighted sum of squares of the average residuals for the counties

$$\hat{m}_{..} \equiv \left(\sum_{i=1}^{T} n_i b_i\right)^{-1}\left(\sum_{i=1}^{T} n_i \overset{\Delta}{u}_{i.}^2\right)$$

then the estimator of σ_v^2 is

$$\hat{\sigma}_v^2 = \max\left\{\hat{m}. - c\hat{\sigma}_e^2, 0\right\}$$

Thus, an accessable predictor for the mean crop area in county i is

$$\hat{y}_i \equiv \overline{\mathbf{x}}_{i(p)}\hat{\boldsymbol{\beta}} + \hat{u}_i\hat{g}_i$$

where

$$\hat{g}_i = \left(\hat{\sigma}_v^2 + n_i^{-1}\hat{\sigma}_e^2\right)^{-1}\hat{\sigma}_v^2$$

The authors also tried to modify the nested-error model of SUPER CARP, by which the variance components are first estimated and generalized least-squares estimators for the β parameters are obtained. The results showed that all coefficients for corn model is significant but only the coefficient of soybeans pixels is significantly different from 0. Moreover, the among-county variance is more significant for soybeans than for corn. The mean estimator retains desirable properties for non-normal assumption but the variance estimator can be seriously biased without normal assumption. The hypothesis testing of normality based on this data showed no reason to reject the normality assumption.

Denote the multiple regression estimators and the generalized least-squares estimators for β_1 and β_2 as $\hat{\boldsymbol{\beta}}_{\boldsymbol{W}}$ and $\hat{\boldsymbol{\beta}}_{\boldsymbol{G}}$, respectively. Correspondingly, the estimated covariance matrix are $\hat{\sum}_{\boldsymbol{W}}$ and $\hat{\sum}_{\boldsymbol{G}}$, respectively. Then the approximate distribution of the statistic

$$F = 2^{-1}\left(\hat{\boldsymbol{\beta}}_W - \hat{\boldsymbol{\beta}}_G\right)' \left(\hat{\boldsymbol{\Sigma}}_W - \hat{\boldsymbol{\Sigma}}_G\right)^{-1} \left(\hat{\boldsymbol{\beta}}_W - \hat{\boldsymbol{\beta}}_G\right)$$

Under the null hypothesis that the slope parameters are the same within and among counties, $F \sim F(2, 22)$ and it can not be rejected based on this dataset.

Based on the predictor $\overline{y}_i$, the predictions for the mean crop hectares per segment along with the estimated standard errors can be computed, as well as the standard errors for the survey regression predictor and the sample mean of the survey data. From the result, we can see that as the number of sample segments increases, the differences between the predicted hectares of two crops and the corresponding sample means decrease. The standard errors of the sample mean are greater than those of the survey regression predictor, while the ratio of the standard error of the best predictor to that of the survey regression predictor increases as the sample segments increases. The improvement of the precision is modest while the sample segments increases from 3 to 5.

The survey regression predictor is unbiased and has relatively small variance. Based on this, the renewable predictor is defined by

$$\overset{*}{y}_i = \hat{y}_i + a_i \left[\sum_{j=1}^{T} W_j \left(\overline{y}_i - \overline{\mathbf{x}}_{i\cdot}\hat{\boldsymbol{\beta}}\right)(1 - \hat{g}_i)\right]$$

where $a_i = \left[\sum_{j=1}^{T} W_j^2 \hat{V}\left(\hat{y}_j\right)\right]^{-1} W_i^2 \hat{V}\left(\hat{y}_i\right)$. This adjustment produces a very small increase in the variance.

In all, the nested-error regression model in the paper provides a promising approach to predicting crop areas in small domains, and the USDA allows people to use supplementary information including estimates of variances form other areas and other years.

1.2 Derivation of formula (3.1) and (3.2)

Proof. Based on the assumptions above, we have

$$\begin{pmatrix} v_i \\ \overline{u}_{i\cdot} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_v^2 & \rho\sigma_v^2 \\ \rho\sigma_v^2 & \sigma_v^2 + n_i^{-1}\sigma_e^2 \end{pmatrix}\right)$$

Then the expectation of v_i given $\overline{u}_{i\cdot}$ is

$$E\left(v_i|\overline{u}_i\right) = 0 + \sigma_v^2 * (\sigma_v^2 + n_i^{-1}\sigma_e^2)^{-1}(\overline{u}_{i\cdot}) = \overline{u}_i g_i$$

where $g_i = m_i^{-1}\sigma_v^2$ and $m_i = \left(\sigma_v^2 + n_i^{-1}\sigma_e^2\right)$. Consider matrix $A = (1, -g_i)$, then $A(v_i, \overline{u}_{i\cdot}) \sim N(0, A\begin{pmatrix} \sigma_v^2 & \rho\sigma_v^2 \\ \rho\sigma_v^2 & \sigma_v^2 + n_i^{-1}\sigma_e^2 \end{pmatrix} A^T) = N(0, \sigma_v^2(1-g_i) - g_i(\sigma_v^2 - g_i m_i)) = N(0, \sigma_v^2(1-g_i))$. So the error variance in this best predictor is

$$\mathrm{E}\left\{\left(v_i - \overline{u}_i g_i\right)^2\right\} = \sigma_v^2\left(1 - g_i\right) = n_i^{-1}\sigma_e^2 - n_i^{-2}\sigma_e^2 m_i^{-1}\sigma_e^2$$

□

1.3 Derivation of formula (3.6)

Proof. The predictor $\tilde{y}_i$ is

$$\tilde{y}_i = \overline{\mathbf{x}}_{i(p)}\tilde{\boldsymbol{\beta}} + \tilde{v}_i$$

which is the best linear unbiased predictor (BLUP) of y_i and have the variance of the error as

$$\mathrm{E}\left\{\left(\tilde{y}_i - y_i\right)^2\right\}$$

Consider the model by matrix notation as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{v} + \boldsymbol{\epsilon} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$$

where $\mathbf{V} = \text{diag}(\mathbf{V_1}, \mathbf{V_2}, \cdots, \mathbf{V_{12}}), \mathbf{V_i} = \sigma_v^2 \mathbf{1_{n_i}} \mathbf{1_{n_i}^T} + \sigma_e^2 \mathbf{1_{n_i}}$

Since the generalized least-squares estimator of β is $\tilde{\boldsymbol{\beta}} = \left(\mathbf{X'V^{-1}X}\right)^{-1} \mathbf{X'V^{-1}Y}$, we have

$$\tilde{\boldsymbol{u}} = \mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}$$

and

$$\tilde{v_i} = g_i(\overline{y}_{i\cdot} - \overline{x}_{i\cdot}\tilde{\boldsymbol{\beta}})$$

Then we have

$$\begin{aligned}
\text{E}\left\{(\tilde{y}_i - y_i)^2\right\} &= \text{Var}\left\{(\tilde{y}_i - y_i)^2\right\} \\
&= \text{Var}(\overline{\mathbf{x}}_{i(p)}\tilde{\boldsymbol{\beta}} + \tilde{v}_i - \overline{\mathbf{x}}_{1(p)}\boldsymbol{\beta} - v_i) \\
&= \text{Var}[\overline{\mathbf{x}}_{i(p)}\tilde{\boldsymbol{\beta}} + g_i(\overline{y}_{i\cdot} - \overline{\mathbf{x}}_i) - v_i] \\
&= \text{Var}[(\overline{\mathbf{x}}_{i(p)} - g_i\overline{x}_i)\tilde{\boldsymbol{\beta}} + g_i\overline{y}_{i\cdot} - v_i]
\end{aligned}$$

Now let

$$Z_i = (0, \cdots, 0, \frac{1}{n_i}, \cdots, \frac{1}{n_i}, 0, \cdots, 0)^T$$

where only the elements for the ith county is $\frac{1}{n_i}$ and others are zero. Also let $\mathbf{c}_i = \overline{\mathbf{x}}_{i(p)} - g_i\overline{\mathbf{x}}_i$ and also notice that $\mathbf{V}(\tilde{\boldsymbol{\beta}}) = (X^TV^{-1}X)^{-1}$, so

$$\begin{aligned}
\text{E}\left\{(\tilde{y}_i - y_i)^2\right\} &= \text{Var}(\mathbf{c}_i\mathbf{V}(\tilde{\boldsymbol{\beta}})X^TV^{-1}y + g_iZ_i^Ty - Z_i^Tv) \\
&= [\mathbf{c}_i\mathbf{V}(\tilde{\boldsymbol{\beta}})X^TV^{-1} + g_iZ_i^T]V[V^{-1}XV(V\tilde{\boldsymbol{\beta}})\mathbf{c}_i^T + g_iZ_i] + \sigma_v^2 - 2\text{Cov}((\mathbf{c}_i\mathbf{V}(\tilde{\boldsymbol{\beta}})X^TV^{-1} + g_iZ_i^T)y, Z_i^Tv) \\
&= \mathbf{c}_i\mathbf{V}(\tilde{\boldsymbol{\beta}})\mathbf{c}_i^T + 2g_iZ_i^TX\mathbf{V}(\tilde{\boldsymbol{\beta}})\mathbf{c}_i^T + g_i^2Z_i^TVZ_i + \sigma_v^2 - 2\sigma_v^2(\mathbf{c}_i\mathbf{V}(\tilde{\boldsymbol{\beta}})X^TV^{-1} + g_iZ_i^T)n_iZ_i \\
&= \mathbf{c}_i\mathbf{V}(\tilde{\boldsymbol{\beta}})\mathbf{c}_i^T + 2g_i\overline{x}_{i\cdot}V(\tilde{\boldsymbol{\beta}})\mathbf{c}_i^T + g_i^2\frac{\sigma_v^2}{g_i} + \sigma_v^2 - 2\sigma_v^2(\mathbf{c}_i\mathbf{V}(\tilde{\boldsymbol{\beta}})\frac{g_i}{\sigma_v^2}\overline{x}_{i\cdot}^T + g_i) \\
&= \mathbf{c}_i\mathbf{V}(\tilde{\boldsymbol{\beta}})c_i^T + \sigma_v^2 - \sigma_v^2g_i \\
&= \sigma_v^2(1 - g_i) + \mathbf{c}_i\mathbf{V}(\tilde{\boldsymbol{\beta}})\mathbf{c}_i'
\end{aligned}$$

where $\mathbf{V}(\tilde{\boldsymbol{\beta}}) = \left(\mathbf{X'V^{-1}X}\right)^{-1}$, $\mathbf{c}_i = \overline{\mathbf{x}}_{i(p)} - g_i\overline{\mathbf{x}}_i$ and $\overline{\mathbf{x}}_\mathbf{i} = n_i^{-1}\sum_{j=1}^{n_i}\mathbf{x}_{ij}$ □

1.4 Derivation of formula (3.10)

Proof. Considering the ordinary least-squares for county i, the average of the residuals is

$$\overset{\triangle}{u}_{i\cdot} = \overline{y}_{i\cdot} - \overline{x}_{i\cdot}(\mathbf{X'X})^{-1}\mathbf{X'Y} = \frac{1}{n_i}\mathbf{1}'_{n_i}\mathbf{Y}_i - \overline{\mathbf{x}}_{i\cdot}(\mathbf{X'X})^{-1}\mathbf{X'Y}$$

Since

$$\text{E}(\mathbf{Y}_i) = \mathbf{0}, \text{E}(\mathbf{Y}) = \mathbf{0}, Var(\mathbf{Y}) = \text{diag}(\text{Var}(\mathbf{Y}_i)) = \text{diag}(\mathbf{J}_i\sigma_v^2 + \mathbf{I}_i\sigma_e^2)_{i=1,\cdots,T},$$

we have

$$\begin{aligned}
\mathrm{E}(\overset{\triangle}{u}_{i\cdot}^{2}) &= \mathrm{Var}(\overset{\triangle}{u}_{i\cdot}) + \{\mathrm{E}(\overset{\triangle}{u}_{i\cdot})\}^2 \\
&= \frac{1}{n_i^2}\mathbf{1}'_{n_i}\mathrm{Var}(\mathbf{Y}_i)\mathbf{1}_{n_i} + \overline{\mathbf{x}}_{i\cdot}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{Var}(\mathbf{Y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\overline{\mathbf{x}}'_{i\cdot} \\
&\quad - \frac{2}{n_i}\mathbf{1}'_{n_i}\mathrm{Cov}(\mathbf{Y}_i, \mathbf{Y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\overline{\mathbf{x}}'_{i\cdot} \\
&= \frac{1}{n_i^2}\mathbf{1}'_{n_i}(\mathbf{J}_i\sigma_v^2 + \mathbf{I}_i\sigma_e^2)\mathbf{1}_{n_i} + \overline{\mathbf{x}}_{i\cdot}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{diag}(\mathbf{J}_i\sigma_v^2 + \mathbf{I}_i\sigma_e^2)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\overline{\mathbf{x}}'_{i\cdot} \\
&\quad - \frac{2}{n_i}\mathbf{1}'_{n_i}(\mathbf{J}_i\sigma_v^2 + \mathbf{I}_i\sigma_e^2)\mathbf{x}_i(\mathbf{X}'\mathbf{X})^{-1}\overline{\mathbf{x}}'_{i\cdot} \\
&= b_i\sigma_v^2 + d_i\sigma_e^2,
\end{aligned}$$

where

$$b_i = 1 - 2n_i\overline{\mathbf{x}}_{i\cdot}(\mathbf{X}'\mathbf{X})^{-1}\overline{\mathbf{x}}_{i\cdot} + \overline{\mathbf{x}}_{i\cdot}(\mathbf{X}'\mathbf{X})^{-1}\{\sum_{j=1}^{T} n_j^2\overline{\mathbf{x}}'_{j\cdot}\overline{\mathbf{x}}_{j\cdot}\}(\mathbf{X}'\mathbf{X})^{-1}\overline{\mathbf{x}}'_{i\cdot},$$

and $d_i = n_i^{-1}\{1 - n_i\overline{\mathbf{x}}_{i\cdot}(\mathbf{X}'\mathbf{X})^{-1}\overline{\mathbf{x}}'_{i\cdot}\}$ □

1.5 Proof of the statement after (3.12)

Under the assumptions of the model (2.1)-(2.2), the weighted sum of squares $\hat{m}_{\cdot\cdot}$ is independent of $\hat{\sigma}_e^2$.

Proof.

$$\hat{m}_{\cdot\cdot} \equiv \left(\sum_{i=1}^{T} n_i b_i\right)^{-1}\left(\sum_{i=1}^{T} n_i \overset{\Delta}{u}_i^{2}\right)$$

where

$$\overset{\Delta}{u}_{i\cdot} = \overline{y}_{i\cdot} - \overline{\mathbf{x}}_{i\cdot}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}$$

So

$$\begin{aligned}
\hat{m}_{\cdot\cdot} &= \text{constant} * \left(\sum_{i=1}^{T} n_i \overset{\Delta}{u}_i^{2}\right) \\
&= \text{constant} * \left(\sum_{i=1}^{T} n_i(\overline{y}_{i\cdot} - \overline{\mathbf{x}}_{i\cdot}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y})^2\right) \\
&= \text{constant} * (\bar{\mathbf{Y}} - \hat{\mathbf{Y}})^T(\bar{\mathbf{Y}} - \hat{\mathbf{Y}}) \\
&= \text{constant} * \mathbf{Y}^T(\mathbf{H} - \frac{1}{n}J_n)\mathbf{Y}
\end{aligned}$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$.

$$\hat{\sigma}_e^2 = \hat{\mathbf{e}}^T\hat{\mathbf{e}}[\sum_{i=1}^{T}(n_i - 1) - 2]^{-1}$$

where $\hat{\mathbf{e}}^T\hat{\mathbf{e}} = (\hat{\mathbf{Y}}^* - \mathbf{Y}^*)^T(\hat{\mathbf{Y}}^* - \mathbf{Y}^*) = \mathbf{Y}^{*T}(\mathbf{H}^* - I_n)\mathbf{Y}^*$ with

$$\mathbf{Y}^* = \mathbf{Y} - \mathrm{diag}(\frac{1}{n_1}J_{n_1}, \frac{1}{n_2}J_{n_2}, \cdots, \frac{1}{n_{12}}J_{n_{12}})\mathbf{Y}$$

$$\mathbf{X}^* = \mathbf{X} - \mathrm{diag}(\frac{1}{n_1}J_{n_1}, \frac{1}{n_2}J_{n_2}, \cdots, \frac{1}{n_{12}}J_{n_{12}})\mathbf{X}$$

$$\mathbf{H}^* = \mathbf{X}^*(\mathbf{X}^{*T}\mathbf{X}^*)^{-1}\mathbf{X}^{*T}$$

Denote $I_n - \text{diag}(\frac{1}{n_1}J_{n_1}, \frac{1}{n_2}J_{n_2}, \cdots, \frac{1}{n_{12}}J_{n_{12}}) = K_n$, then we have

$$\hat{\sigma}_e^2 = \text{constant} * \mathbf{Y}^{*T}(\mathbf{H}^* - I_n)\mathbf{Y}^* = \text{constant} * \mathbf{Y}^T K_n^T(\mathbf{H} - I_n)K_n\mathbf{Y}$$

Since $(\mathbf{H} - \frac{1}{n}J_n)K_n^T(\mathbf{H} - I_n)K_n = 0$, by the Cochran's theorem, we have that $\mathbf{Y}^T(\mathbf{H} - \frac{1}{n}J_n)\mathbf{Y}$ and $\mathbf{Y}^T K_n^T(\mathbf{H} - I_n)K_n\mathbf{Y}$ are independent, i.e. the weighted sum of squares $\hat{m}_{..}$ is independent of $\hat{\sigma}_e^2$. □

1.6 Derivation of (A.1) (A.2)

Proof. The predictor $\hat{y}_i$ is

$$\hat{y}_i = \overline{\mathbf{x}}_{(p)}\hat{\boldsymbol{\beta}} + \left(\overline{y}_{i.} - \overline{\mathbf{x}}_i\hat{\boldsymbol{\beta}}\right)\hat{g}_i$$

where

$$\hat{g}_i = 1 - \hat{h}_i$$

$$\hat{h}_i = \left[\hat{m}_i + \hat{k}_i + \left(n_i^{-1} - c\right)^2 \hat{w}_i\right]^{-1}\left[n_i^{-1}\hat{\sigma}_e^2 + \left(n_i^{-1} - c\right)n_i^{-1}\hat{w}_i\right]$$

$$\hat{m}_i = \hat{m}_{..} + \left(n_i^{-1} - c\right)\hat{\sigma}_e^2$$

$$\hat{w}_i = 2d_e^{-1}\hat{m}_i^{-1}\hat{\sigma}_e^4$$

$$\hat{k}_i = 2\hat{\sigma}_e^2\left(\ddot{\sigma}_{ff} + n_i^{-1}\right)^{-1}\left[\sum_{j=1}^{T} n_j b_j\right]^{-2}\left[\sum_{j=1}^{T} n_j^2 b_j\left(\ddot{\sigma}_{ff} + n_j^{-1}\right)^2\right]$$

$$\ddot{\sigma}_{ff} = \max\left[0, (T-5)^{-1}(T-3)\hat{\sigma}_e^{-2}\hat{m}_{.} - c\right]$$

Then we can compute the variance of the error of the predictor above as

$$\begin{aligned}
V(\hat{y}_i - y_i) &= \text{Var}(\overline{\mathbf{x}}_{(p)}\hat{\boldsymbol{\beta}} + \left(\overline{y}_{i.} - \overline{\mathbf{x}}_i\hat{\boldsymbol{\beta}}\right)\hat{g}_i - \overline{x}_i\beta - v_i) \\
&= \text{Var}[(\overline{\mathbf{x}}_{i(p)} - \hat{g}_i\overline{x}_{i.})\hat{\boldsymbol{\beta}} + \hat{g}_i\overline{y}_{i.} - v_i] \\
&= \text{Var}(\hat{c}_i\hat{\beta} + \hat{g}_i\overline{y}_{i.} - v_i) \\
&= \text{Var}([\hat{c}_i(X^T\hat{V}^{-1}X)^{-1}X^T\hat{V}^{-1} + \hat{g}_i Z_i^T]Y - Z_i^T v) \\
&= [\hat{c}_i(X^T\hat{V}^{-1}X)^{-1}X^T\hat{V}^{-1} + \hat{g}_i Z_i^T]V[\hat{V}^{-1}X(X^T\hat{V}^{-1}X)^{-1}\hat{c}_i^T + \hat{g}_i Z_i] + \sigma_v^2 \\
&- \text{Cov}(\hat{c}_i(X^T\hat{V}^{-1}X)^{-1}X^T\hat{V}^{-1} + \hat{g}_i Z_i^T, Z_i^T V) \\
&= \hat{c}_i(X^T\hat{V}^{-1}X)^{-1}X^T\hat{V}^{-1}V\hat{V}^{-1}X(X^T\hat{V}^{-1}X)^{-1}\hat{c}_i^T + 2\hat{g}_i Z_i^T V\hat{V}^{-1}X(X^T\hat{V}^{-1}X)^{-1}c_i^T + {\hat{g}_i}^2 Z_i^T V Z_I + \sigma_v^2 \\
&- 2\text{Cov}(\hat{c}_i(X^T\hat{V}^{-1}X)^{-1}\frac{1}{n_i\hat{\sigma}_v^2 + \hat{\sigma_e^2}}\overline{x}_{i.}^T + \hat{g}_i Z_i^T)n_{Zi} \\
&= \hat{c}_i(X^T\hat{V}^{-1}X)^{-1}X^T\hat{V}^{-1}V\hat{V}^{-1}X(X^T\hat{V}^{-1}X)^{-1}\hat{c}_i^T + 2\hat{g}_i Z_i^T V\hat{V}^{-1}X(X^T\hat{V}^{-1}X)^{-1}c_i^T + {\hat{g}_i}^2\frac{\sigma_v^2}{g_i} + \sigma_v^2 \\
&- 2\sigma_v^2(\hat{c}_i(X^T\hat{V}^{-1}X)^{-1}\frac{\overline{x}_{i.}^T}{n_i\hat{\sigma}_v^2 + \hat{\sigma}_e^2} + \hat{g}_i) \\
&= \hat{c}_i(X^T\hat{V}^{-1}X)^{-1}X^T\hat{V}^{-1}V\hat{V}^{-1}X(X^T\hat{V}^{-1}X)^{-1}\hat{c}_i^T + 2(1 - [\hat{m}_i \\
&+ \hat{k}_i + \left(n_i^{-1} - c\right)^2\hat{w}_i^{-1}\left[n_i^{-1}\hat{\sigma}_e^2 + \left(n_i^{-1} - c\right)n_i^{-1}\hat{w}_i\right])Z_i^T V\hat{V}^{-1}X(X^T\hat{V}^{-1}X)^{-1}c_i^T
\end{aligned}$$

2 Part I: Model Selection

2.1 AIC and BIC Criteria

The nested regression model proposed here is

$$Y_{ij} = x_{ij}^{'}\beta + v_i + e_{ij}, i = 1, 2, .., 12, j = 1, 2, .., n_i$$

where y_{ij} is the reported hectares of corn(soybean) and the fixed effect is given by

$$x_{ij}^{'}\beta = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij}$$

The county-specific random effect is $v_i \sim N(0, \sigma_v^2)$ and the sample-specific random error is $e_{ij} \sim N(0, \sigma_e^2)$ and they are independent with both variance unknown.

In the original paper, the author discussed about the possibility to incorporate quadratic and interaction effects. So for the fixed effects we have the following candidate variable set in which we select the optimal model based on AIC/BIC:

$$\Theta = \{x_{1ij}, x_{2ij}, x_{1ij}^2, x_{2ij}^2, x_{1ij}x_{2ij}\}$$

As for preprocessing the data, as the original paper argues that the 33rd observation may be problematic, we delete this sample and fit our model on the rest of the data.

For implementation, we traverse all the possible cases including the all possible subsets in Θ for corn hectares and soybean hectares separately. We apply **AIC** and **BIC** criteria to perform the model comparison and select the model with the smallest AIC/BIC.

$$\text{AIC} = -2\log\text{Lik} + 2k \tag{1}$$

$$\text{BIC} = -2\log\text{Lik} + k\log(n) \tag{2}$$

where k is the number of parameters in the model and n is the number of samples.

Here we use the n_{total} as the "effective" sample size. However, it may not be the optimal choice for this data as we did not consider the effect size innate to the random effects.

In 2014, Maud Delattre proposed an alternative BIC for mixed effect model [1] with a more comprehensive consideration of the effective sample size for both random effects and fixed effects.

$$\text{BIC}_h = -2\log\text{Lik} + \dim(\theta_R)\log(N) + \dim(\theta_F)\log(n_{total}) \tag{3}$$

where $\dim(\theta_R)$ and $\dim(\theta_F)$ denotes the number of parameters of random effects and fixed effects and N denotes the number of subjects (corresponding to the number of county in this application).

In the following sections, we will use (1)(3) to calculate AIC and BIC and select our corresponding models.

2.2 Results

The results below show that for corn hectares both AIC and BIC select the same model with CornPix and CornPix:SoyBean as fixed effect, and for soybeans hectares both AIC and BIC select the same model with only SoyBeans as fixed effect.

Criterion	Selected Model
AIC	CornHec $\sim$ CornPix + CornPix:SoyBeanPix + (1|County)
BIC	CornHec $\sim$ CornPix + CornPix:SoyBeanPix + (1|County)
AIC	SoyBeanHec $\sim$ SoyBeansPix + (1|County)
BIC	SoyBeanHec $\sim$ SoyBeansPix + (1|County)

Table 1: Model Selection

2.3 Comments

- In real application BIC generally tends to select a smaller model (when the sample size n is relatively large) as it impose more penalty on the model complexity than AIC does. **However** in this data application we can see that they select the same model for both response variables.

Note that in fitting the linear mixed model with higher order terms especially with quadratic terms, some of the predictor variables are on very different scales which may introduce some kinds of numerical instabilities.

3 Part II: Application of Sumca

3.1 Sumca Method

Recently, a new method was proposed for estimating the MSPE of a complex predictor, known as **Sumca**: a Simple, Unified, Monte-Carlo Assisted Approach to Second-order Unbiased MSPE Estimation (Jiang, Torabi, 2018). [2]

Here, we apply Sumca method for the problem under two different situations. The discrepancy between these situations depends on our unknown parameters ψ.

In the first situation, ψ only includes fixed effects and variance components i.e.

$$\psi = (\beta, \sigma_v^2, \sigma_e^2) \tag{4}$$

We refer to this method as "Plain" Sumca.

In the second situation, ψ includes model (M), fixed effects and variance components i.e.

$$\psi = (M, \beta, \sigma_v^2, \sigma_e^2) \tag{5}$$

We refer to this method as "M-parameterized" Sumca.

3.1.1 Plain Sumca

The post model selection predictor $\hat{\theta}_i$ be a predictor of θ_i, where $\theta_i = \beta_0 + \beta_1 \bar{X}_{1i(p)} + \beta_2 \bar{X}_{2i(p)} + v_i$ and $\hat{\theta}_i$ is a function of Y_i. Then, the MSPE of the post model selection prediction is

$$\text{MSPE}(\hat{\theta}_i) = E(\hat{\theta}_i - \theta)^2 = E[E(\hat{\theta}_i - \theta_i)^2 | Y_i] \tag{6}$$

Let

$$\begin{aligned} a(Y_i, \psi) = \text{E}[(\hat{\theta}_i - \theta_i)^2 | Y_i] &= (\hat{\theta}_i - \text{E}(\theta_i | Y_i))^2 + Var(\theta_i | Y_i) \\ &= (\hat{\theta}_i - a_1(Y_i, \psi))^2 + a_2(Y_i, \psi) \end{aligned} \tag{7}$$

where $a_1(Y_i, \phi) = \text{E}(\theta_i | Y_i)$ and $a_2(Y_i, \phi) = Var(\theta_i | Y_i)$

Under the first situation where $\psi = (\beta, \sigma_e^2, \sigma_v^2)$ and $a_1(Y_i, \hat{\psi}) \neq 0$, thus we need to derive $E(\theta_i | Y_i)$ and $\hat{\theta}_i$, and $E(\theta_i | Y_i)$ should be derived under the full model and $\hat{\theta}_i$ should be derived under the selected model.

$$\begin{aligned} a_1(Y_i, \psi) &= \text{E}(\theta_i | Y_i) \\ &= \beta_0 + \beta_1 \bar{X}_{1i} + \beta_2 \bar{X}_{2i} + \sigma_v^2 1_{n_i}^T (\sigma_e^2 I + \sigma_v^2 1_{n_i} 1_{n_i}^T)^{-1} (Y_i - x_i \beta) \\ &= \beta_0 + \beta_1 \bar{X}_{1i} + \beta_2 \bar{X}_{2i} + \frac{\sigma_v^2 n_i (\bar{Y}_i - \bar{x}_i \beta)}{\sigma_v^2 n_i + \sigma_e^2} \end{aligned} \tag{8}$$

$$a_2(Y_i, \psi) = \text{Var}(\theta_i | Y_i) = \frac{\sigma_v^2 \sigma_e^2}{\sigma_v^2 n_i + \sigma_e^2} \tag{9}$$

$$\hat{\theta}_i = [\beta_0 + \beta_1 \bar{X}_{1i} + \beta_2 \bar{X}_{2i} + \frac{\sigma_v^2 n_i (\bar{Y}_i - \bar{x}_i^* \beta)}{\sigma_v^2 n_i + \sigma_e^2}]_{\psi = \hat{\psi}^S} \tag{10}$$

where $x_i \beta = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{1ij}^2 + \beta_4 x_{2ij}^2 + \beta_5 x_{1ij} x_{2ij}$, $\quad x_i^* \beta$ is the linear predictor of the selected model and $\hat{\psi}^S$ is obtained under the selected model.

Let $a(Y_i, \hat{\psi}^F)$ be an estimate of $a(Y_i, \psi)$ where $\hat{\psi}^F$ is estimated parameters under the full model. Thus, the **leading term** $a(Y_i, \hat{\psi}^F)$ is

$$a(Y_i, \hat{\psi}^F) = (\hat{\theta}_i - a_1(Y_i, \hat{\psi}^F))^2 + a_2(Y_i, \hat{\psi}^F) \tag{11}$$

where $\hat{\theta}_i$ in (10) is the predictor under the selected model.

$$a_1(Y_i, \hat{\psi}^F) = [\beta_0 + \beta_1 \bar{X}_{1i} + \beta_2 \bar{X}_{2i} + \frac{\sigma_v^2 n_i (\bar{Y}_i - \bar{x}_i \beta)}{\sigma_v^2 n_i + \sigma_e^2}]_{\psi=\hat{\psi}^F} \tag{12}$$

$$a_2(Y_i, \hat{\psi}^F) = [\frac{\sigma_v^2 \sigma_e^2}{\sigma_v^2 n_i + \sigma_e^2}]_{\psi=\hat{\psi}^F} \tag{13}$$

According to Sumca, therefore, a second-order unbiased estimator of MSPE$(\hat{\theta}_i)$ is

$$\hat{\text{MSPE}}(\hat{\theta}_i) = a(Y_i, \hat{\psi}^F) + d(\hat{\psi}^F) \tag{14}$$

where $d(\hat{\psi}^F) = \text{E}[a(Y_i, \psi) - a(Y_i, \hat{\psi})]_{\psi=\hat{\psi}^F}$ can be estimated by Monte-Carlo approach.

$$d(\hat{\psi}^F) = \frac{1}{K} \sum_{k=1}^{K} (a(Y_{i[k]}, \hat{\psi}^F) - a(Y_{i[k]}, \hat{\psi}^F_{[k]})) \tag{15}$$

where

$$a(Y_{i[k]}, \hat{\psi}^F) = (\hat{\theta}_i^{[k]} - a_1(Y_{i[k]}, \hat{\psi}^F))^2 + a_2(Y_{i[k]}, \hat{\psi}^F) \tag{16}$$

$$a(Y_{i[k]}, \hat{\psi}^F_{[k]}) = (\hat{\theta}_i^{[k]} - a_1(Y_{i[k]}, \hat{\psi}^F_{[k]}))^2 + a_2(Y_{i[k]}, \hat{\psi}^F_{[k]}) \tag{17}$$

where $Y_{[k]}$ is simulated under the full model as $\psi = \hat{\psi}^F$, $Y_{i[k]}$ is the corresponding response of the i-th county, and $\hat{\psi}^F_{[k]}$ is estimated based on $Y_{[k]}$ under the full model.

Algorithm 1: Plain Sumca

Input: Response Y, Predictor X, Population mean $\overline{X}_{(p)}$, Index i;
Select the optimal submodel according to AIC or BIC and obtain $\hat{\theta}_i$ by (10);
Fit the full model to obtain $\hat{\psi}^F$ and $a_1(Y_i, \hat{\psi}^F), a_2(Y_i, \hat{\psi}^F)$ by (12)(13);
Obtain the leading term $a(Y_i, \hat{\psi}^F)$ by (11);
for *k=1:K* **do**
  Simulate $Y_{[k]}$ based on the full model;
  Fit a new full model with response $Y_{[k]}$ and obtain the corresponding $\hat{\psi}^F_{[k]}$;
  Obtain $d_{[k]} = a(Y_{i[k]}, \hat{\psi}^F) - a(Y_{i[k]}, \hat{\psi}^F_{[k]})$
end
Set $\hat{\text{MSPE}} = a(Y_i, \hat{\psi}^F) + \sum_{k=1}^{K} d_{[k]}/K$;
if $\hat{\text{MSPE}} < 0$ **then**
  return $a(Y_i, \hat{\psi}^F)$
else
  return $\hat{\text{MSPE}}$
end

3.1.2 M-parameterized Sumca

Under the second situation in which we view the model as a parameter in ψ, we have the poste model selection predictor $\hat{\theta}_i = a_1(Y_i, \hat{\psi}^S)$. Thus, the **leading term** $a(Y_i, \hat{\psi}^S)$ is

$$\begin{aligned} a(Y_i, \hat{\psi}^S) = a_2(Y_i, \hat{\psi}^S) &= \text{Var}(\theta_i | Y_i)_{\psi=\hat{\psi}^S} \\ &= [\frac{\sigma_v^2 \sigma_e^2}{\sigma_v^2 n_i + \sigma_e^2}]_{\psi=\hat{\psi}^S} \end{aligned} \tag{18}$$

where $\text{Var}(\theta_i|Y_i)_{\psi=\hat{\psi}^S}$ is the conditional variance under the selected model.

Based on the M-parameterized Sumca, a second-order unbiased estimator of $\text{MSPE}(\hat{\theta}_i)$ is

$$\hat{\text{MSPE}}(\hat{\theta}_i) = a(Y_i, \hat{\psi}^S) + d(\hat{\psi}^S) = a_2(Y_i, \hat{\psi}^S) + d(\hat{\psi}^S) \tag{19}$$

where $d(\hat{\psi}^S) = E[a(Y_i, \psi) - a(Y_i, \hat{\psi})]_{\psi=\hat{\psi}^S}$ and it can be estimated by Monte-Carlo approach.

$$d(\hat{\psi}^S) = \frac{1}{K}\sum_{k=1}^{K}[a(Y_{i[k]}, \hat{\psi}^S) - a(Y_{i[k]}, \hat{\psi}^S_{[k]})] \tag{20}$$

where

$$a(Y_{i[k]}, \hat{\psi}^S) = (\hat{\theta}_i^{[k]} - a_1(Y_{i[k]}, \hat{\psi}^S))^2 + a_2(Y_{i[k]}, \hat{\psi}^S) \tag{21}$$

$$a(Y_{i[k]}, \hat{\psi}^S_{[k]}) = a_2(Y_{i[k]}, \hat{\psi}^S_{[k]}) = \text{Var}(\theta_i|Y_{i[k]})_{\psi=\hat{\psi}^S_{[k]}} \tag{22}$$

Here $Y_{[k]}$ is simulated under the selected model as $\psi = \hat{\psi}^S$, $Y_{i[k]}$ is the corresponding response of the i-th county, and $\hat{\psi}^S_{[k]}$ is estimated based on $Y_{[k]}$ after model selection.

Algorithm 2: M-parameterized Sumca

Input: Response Y, Predictor X, Population mean $\overline{X}_{(p)}$, Index i;
Select the optimal submodel according to AIC or BIC and obtain $\hat{\psi}^S$;
Obtain the leading term $a(Y_i, \hat{\psi}^S)$ by (18);
for *k=1:K* **do**
  Simulate $Y_{[k]}$ based on the $\hat{\psi}^S$;
  Select the optimal submodel with response $Y_{[k]}$ and obtain the corresponding $\hat{\psi}^S_{[k]}$;
  Obtain $d_{[k]} = a(Y_{i[k]}, \hat{\psi}^S) - a(Y_{i[k]}, \hat{\psi}^S_{[k]})$
end
Set $\hat{\text{MSPE}} = a(Y_i, \hat{\psi}^S) + \sum_{k=1}^{K} d_{[k]}/K$;
if $\hat{\text{MSPE}} < 0$ **then**
  return $a(Y_i, \hat{\psi}^S)$
else
  return $\hat{\text{MSPE}}$
end

3.1.3 The Leading Term

The leading term for Sumca method above is guaranteed positive, since it is the summation of the squared conditional bias and the conditional variance. This is a desirable property for an MSPE estimator. If there any negative estimates of the MSPEs by Sumca, an alternative estimator is the leading term of the Sumca estimator.

- **Plain Sumca**: The leading term is

$$a(Y_i, \hat{\psi}^F) = (\hat{\theta}_i - a_1(Y_i, \hat{\psi}^F))^2 + a_2(Y_i, \hat{\psi}^F)$$

  where $\hat{\theta}_i$ is the predictor under the selected model.

$$\hat{\theta}_i = [\beta_0 + \beta_1\bar{X}_{1i} + \beta_2\bar{X}_{2i} + \frac{\sigma_v^2 n_i(\bar{Y}_i - \bar{x}_i^*\beta)}{\sigma_v^2 n_i + \sigma_e^2}]_{\psi=\hat{\psi}^S}$$

$$a_1(Y_i, \hat{\psi}^F) = [\beta_0 + \beta_1\bar{X}_{1i} + \beta_2\bar{X}_{2i} + \frac{\sigma_v^2 n_i(\bar{Y}_i - \bar{x}_i\beta)}{\sigma_v^2 n_i + \sigma_e^2}]_{\psi=\hat{\psi}^F}$$

$$a_2(Y_i, \hat{\psi}^F) = [\frac{\sigma_v^2\sigma_e^2}{\sigma_v^2 n_i + \sigma_e^2}]_{\psi=\hat{\psi}^F}$$

where $x_i\beta = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{1ij}^2 + \beta_4 x_{2ij}^2 + \beta_5 x_{1ij} x_{2ij}$, $x_i^*\beta$ is the linear predictor of the selected model and $\hat{\psi}^S$ is obtained under the selected model.

- M**-parameterized Sumca**: The leading term is

$$a(Y_i, \hat{\psi}^S) = a_2(Y_i, \hat{\psi}^S) = \text{Var}(\theta_i|Y_i)_{\psi=\hat{\psi}^S} = [\frac{\sigma_v^2\sigma_e^2}{\sigma_v^2 n_i + \sigma_e^2}]_{\psi=\hat{\psi}^S}$$

where $\text{Var}(\theta_i|Y_i)_{\psi=\hat{\psi}^S}$ is the conditional variance under the selected model.

3.1.4 The choice of K

not need to large

- computational cost
-

3.2 Results

- Graph
- Explaination: methods, legends
- Appendix

3.3 Comments

- Inconsistency in θ: Jiming vs. Non-Jiming: discussions: 1. colinearity; 2. Population quadratic interaction
- Stablization: K size does not make difference to stablization. discussions: 1. data size-¿ simulation not stable
- AIC better than BIC.

References

[1] Maud Delattre, Marc Lavielle, Marie-Anne Poursat, et al. A note on bic in mixed-effects models. *Electronic journal of statistics*, 8(1):456–475, 2014.

[2] Jiming Jiang. *Linear and generalized linear mixed models and their applications*. Springer Science & Business Media, 2007.

A Appendix

A.1 Model Selection

CPix	SPix	$CPix^2$	$SPix^2$	CPix:SPix	CHec (AIC/BIC)	SHec (AIC/BIC)
0	0	0	0	0	357.90 / 360.46	365.04 / 367.60
1	0	0	0	0	307.53 / 311.66	349.43 / 353.57
0	1	0	0	0	330.74 / 334.88	**314.23** / **318.37**
1	1	0	0	0	304.03 / 309.75	316.01 / 321.73
0	0	1	0	0	305.72 / 309.86	348.39 / 352.53
1	0	1	0	0	307.58 / 313.30	350.28 / 356.00
0	1	1	0	0	302.83 / 308.55	316.11 / 321.83
1	1	1	0	0	304.66 / 311.96	317.78 / 325.08
0	0	0	1	0	336.58 / 340.72	323.06 / 327.20
1	0	0	1	0	306.68 / 312.40	324.87 / 330.59
0	1	0	1	0	331.04 / 336.76	315.76 / 321.48
1	1	0	1	0	303.41 / 310.71	317.48 / 324.78
0	0	1	1	0	303.97 / 309.69	324.31 / 330.03
1	0	1	1	0	305.97 / 313.27	322.23 / 329.54
0	1	1	1	0	304.60 / 311.90	317.47 / 324.78
1	1	1	1	0	305.41 / 314.30	319.47 / 328.36
0	0	0	0	1	358.24 / 362.37	352.67 / 356.80
1	0	0	0	1	**301.93** / **307.65**	323.44 / 329.17
0	1	0	0	1	320.86 / 326.58	315.85 / 321.57
1	1	0	0	1	303.88 / 311.18	317.84 / 325.14
0	0	1	0	1	304.63 / 310.35	327.72 / 333.44
1	0	1	0	1	303.93 / 311.23	320.49 / 327.79
0	1	1	0	1	304.83 / 312.13	317.83 / 325.14
1	1	1	0	1	305.76 / 314.64	319.77 / 328.66
0	0	0	1	1	335.67 / 341.39	321.57 / 327.29
1	0	0	1	1	303.81 / 311.12	319.14 / 326.44
0	1	0	1	1	306.09 / 313.40	317.63 / 324.93
1	1	0	1	1	305.41 / 314.30	319.10 / 327.98
0	0	1	1	1	305.45 / 312.75	319.44 / 326.74
1	0	1	1	1	305.62 / 314.51	320.92 / 329.81
0	1	1	1	1	305.52 / 314.41	319.36 / 328.25
1	1	1	1	1	307.41 / 317.88	320.23 / 330.70

Table 2: Caption

A.2 MSPE

A.3 R Code

STA 224 Final Project:
Beta-Carotene and Skin Cancer Prevention
-Application of Longitudinal Analysis

Heqiao Ruan
email:hruan@ucdavis.edu
Instructor: Prof Xiaodong Li

June 3, 2018

1 Abstract

In this project, we apply various techniques in longitudinal data analysis:Generalized Linear Mixed model(GLMM),Generalized Estimating Equation(GEE) to explore the effect of beta carotene on preventing the appearance of the skin cancer as well as corresponding diagnostics to identify the effect of beta-carotene on skin cancer restricted in the first center in the whole data. For these model fitting, we also perform the corresponding model diagnostics to identify the outliers(241th object) which is validated by fitting trajectory visualization. Then we compare the model performance after removing the outliers. Finally we conclude that in this case, Generalized Linear Mixed Model(GLMM) seems to be better than the Generalized Estimating Equation.

2 Introduction

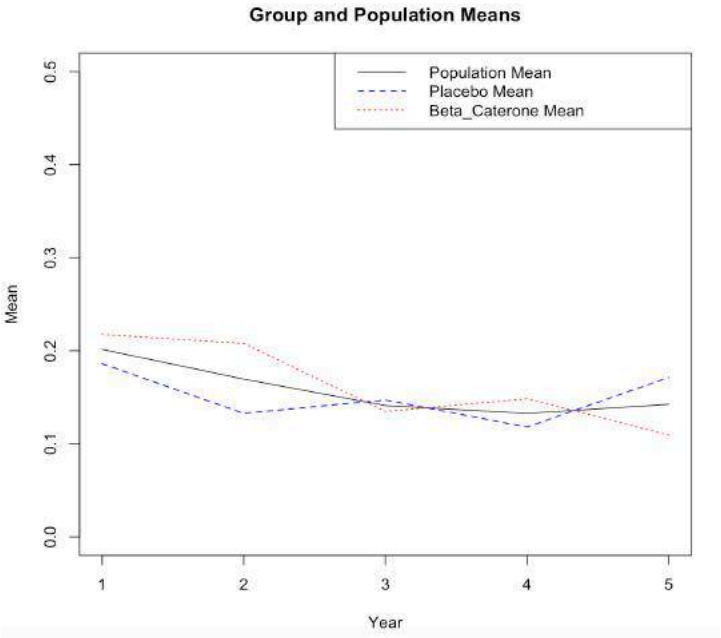
2.1 Background

Skin cancer is among one of the most dangerous cancer in the modern society and many research has focused on identifying some specific chemical materials to help prevent the appearance of skin cancer. In 1990, Greenberg([1]) et.al conducted a clinical trial on randomly assigned 1805 people with previous history of relative cancer(denote as high risk subjects) to placebo and beta-catotene group and observe the number of patients' new skin cancer since the previous observation each year in a duration of five years. The ultimate goal of this clinical trial can be interpreted as that beta-carotene has somewhat degree of positive effect to prevent non-melanoma skin cancer in high risk subjects. The beta-carotene is an important

kind of nutrients and some previous research shows that it can help cure some kinds of cancer or even prevent the cancer. So here we conduct various methods in longitudinal data analysis to explore the effect of beta-carotene on the skin cancer prevention.

2.2 Dataset and Exploratory Analysis

Here the original dataset we get from the Internet consists of 7081 measurements,1683 subjects with 6 covariates:**Center,Age,Skin,Gender,Exposure,Treatment** where the response variable is in the count form which denotes the number of new skin cancer occurs since the last observation.
One thing we should point out is that it seems impossible for us to accurately identify the difference of effect among different centers which means the covariate **Center** should be excluded from our downstream analysis.So we may need to remove some of the samples from the whole dataset with 4 centers. We solve this issue by restricting our analysis into center 1 which has 422 objects and 1883 measurements in total after pre-processing.
Then we draw the plot of the mean of response among different treatments(group means for placebo group and beta-caterone group) and the whole population mean in one plot after all of the missing data points:



By only observing the plot, we can't tell whether the beta-carotene treatment has significant effect on prevention of skin cancer so we need some quantitative tools to fit the data and perform corresponding diagnostics.

3 GLMM

Here the design is random which means individuals are assigned randomly to two treatment groups and the response variable Y_{ij} here denote the number of new skin cancers for j th year's observation of the ith subject.

3.1 Model selection

The response variable denotes the number of the skin cancers observed from the previous exam and it can be treated as the count variable so the reasonable link function here is chosen as log link(Poisson). Here in fitting the generalized linear mixed model, we may need to start from the full model and prune it by the local wald test(or alternatively, likelihood ratio test). Then we first fit the full model with random slope and random intercept (including most possible covariates),here we use the random intercept and slope b_i which follows a bivariate normal distribution(covariance matrix 2 by 2)

$$log(E[Y_{ij}|b_i]) = \beta_0+\beta_1 trt_i+\beta_2 year_{ij}+\beta_3 age_i+\beta_4 skin_i+\beta_5 gender_i+ \beta_6 exposure_i + \beta_7 trt_i * year_{ij} + b_{0i} + b_{1i} * year_{ij} \ \{1\}$$

Then from the coefficient table we can see that effect of $trt_i, year, skin, trt_i * year_{ij}$ are not significant(z test p value much larger than 0.05). What's more,we also conduct a likelihood ratio test which null hypothesis is $H_0 : \beta_1 = \beta_2 = \beta_4 = \beta_7 = 0$ and then the p value for this test is 0.956 which means we can't reject H_0 so then we get the second model:

$$log(E[Y_{ij}|b_i]) = \beta_0 + \beta_1 age_i + \beta_2 gender_i + \beta_3 exposure_i + b_{0i} + b_{1i} * year_{ij} \ \{2\}$$

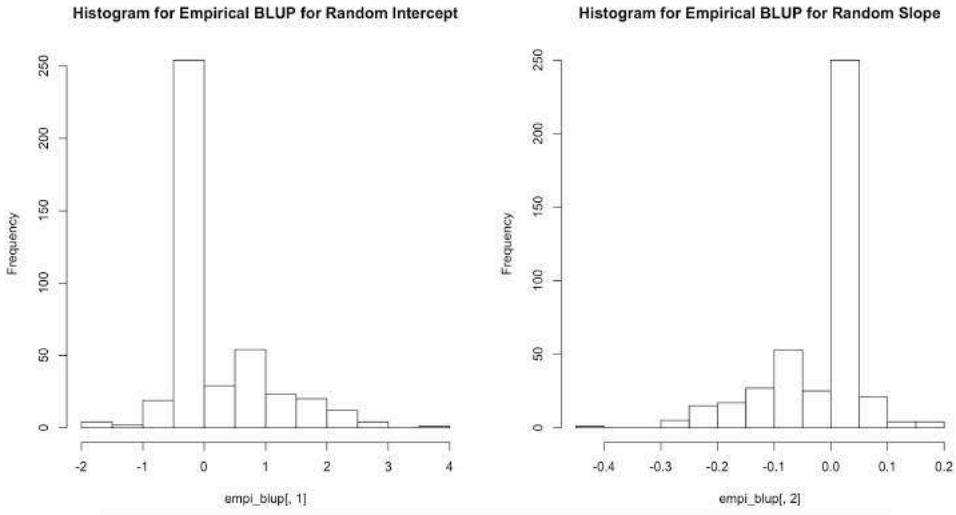
After fitting this model,we observe that all coefficient is significant. Then we need to test the significance of random slope, so we conduct another likelihood ratio test where the reduced model in which only have random intercept but no random slope so here the null hypothesis is $H_0 : b_{1i} = 0$. Then the p value for this likelihood ratio test is 0.000552 which means we reject H_0. So we use the model {2} as our final model for GLMM. Then the model estimation including both the fixed effect and random effect for the generalized linear mixed model is shown as below:

	Estimate	Standard.error	p value
Intercept	-4.758	0.655	3.66e-13
age	0.02126	0.00969	0.028
gender	0.6533	0.1882	0.000518
exposure	0.1714	0.0213	7.46e-16
$var(b_{0i})$	1.9657	_	_
$var(b_{1i})$	0.0225	_	_
$cov(b_{0i}, b_{1i})$	-0.210397	_	_

In terms of model interpretation, we can see that for GLMM, older patients are more likely to develop the cancer no matter what kind of treatment they receive. Similarly, we can see that the log of the expected number of new skin cancers for subjects with more previous skin cancers are larger. What's more, we observe that the male subject has a significant larger log of the expected number of skin cancer than the women subject by the degree of 0.6533. For the random effect, approximately 95% of subjects have changes in expected number of new skin cancers varies from [-0.294,0.294] and b_{0i} here indicates that the subject-wise variability is significant for their log expected number of new skin cancer.

3.2 Diagnostics for GLMM

For diagnostic part of GLMM, we first use the χ^2 test by the sum of square of pearson residual to check the model adequancy(i.e. whether overdispersion exist).Then the test statistic here is $\sum_{i,j} r^2_{P,ij} = \sum_{i,j} \frac{(Y_{ij}-\hat{\mu_{ij}})^2}{\hat{\mu_{ij}}}$. Here we know that under the null hypothesis(model is adequate), it follows χ^2_{n-p}. For the degree of freedom, it is the total number of parameters in the model which include both the random effect and fixed effect. Then the p value here is 1 which means we don't have enough evidence to reject the null hypothesis. So we conclude that the model is adequate.
What's more, we plot the empirical BLUP to see whether there are outliers:



We can see that there may be an outlier and after investigation, it's the 241th object. The following analysis in GEE validate this observation.

4 GEE

For fitting the generalized estimating equation(GEE), we also use the poisson family(count response).Similarly we will start from a full model which include all possible covariates and their interactions and we assume the cubic time trend.What's more, WLOG, we assume our

correlation pattern as unstructured. So the first model here is

$$log(E[Y_{ij}]) = \beta_0+\beta_1 year_{ij}+\beta_2 age_i+\beta_3 skin_i+\beta_4 gender_i+\beta_5 exposure_i + \beta_6 trt_i * year_{ij} + \beta_7 trt_i * age_i + \beta_8 trt_i * skin_i + \beta_9 trt_i * gender_i + \beta_{10} year_{ij}^2 + \beta_{11} year_{ij}^3 + \beta_{12} trt_i + \beta_{13} trt_i * exposure_i \ \{3\}$$

Then from the coefficient table we can see that only $trt, exposure, I(trt * age), I(trt * gender)$ are significant. So here we conduct the likelihood ratio test for the null hypothesis $H_0 : \beta_{rest} = 0$ where β_{rest} denote all of the rest variables apart from the four variables. So here comes our second model:

$$log(E[Y_{ij}]) = \beta_0 + \beta_1 trt_i + \beta_2 exposure_i + \beta_3 trt_i * age_i + \beta_4 trt_i * gender_i \ \{4\}$$

Then the p value for the LR test is 0.34 which favors the reduced model.
After reaching a reasonable model, we would like to try different correlation patterns: AR(1), exchangeable, unstructured and independent.From [3](Lab Note of UNC ECOL562) we know that, to select the best pattern, we can just choose the one with the smallest deviation between the naive correlation estimation and the robust correlation estimation.(Alternative methods includes QIC or empirically observe the data or domain knowledge) To evaluate the difference, we may use the l_1 norm of the matrix which sums the absolute value of all elements in the working correlation matrix. Here a better fit means that our model-based estimator is close to our sandwich estimator achieved by the GEE procedure:difference of Robust(sandwich) s.e and naive s.e(Sandwich estimator most approximate the NAIVE estimator). Then the difference of the two estimators for different correlation structures are shown as below.

Unstructured	AR(1)	Exchangeable	Independent
0.242	0.662	0.426	0.838

So here we may choose the unstructured covariance structure.Then model summary is shown as below:

Variable	Estimate	Standard.error	p value
Intercept	-2.4838	0.1324	$< 2e-16$
trt	-3.9569	1.0547	0.000176
exposure	0.13668	0.01197	$< 2e-16$
I(trt*age)	0.0521	0.0152	0.000589
I(trt*gender)	1.0419	0.3044	0.000620
ϕ	1.27	0.278	_

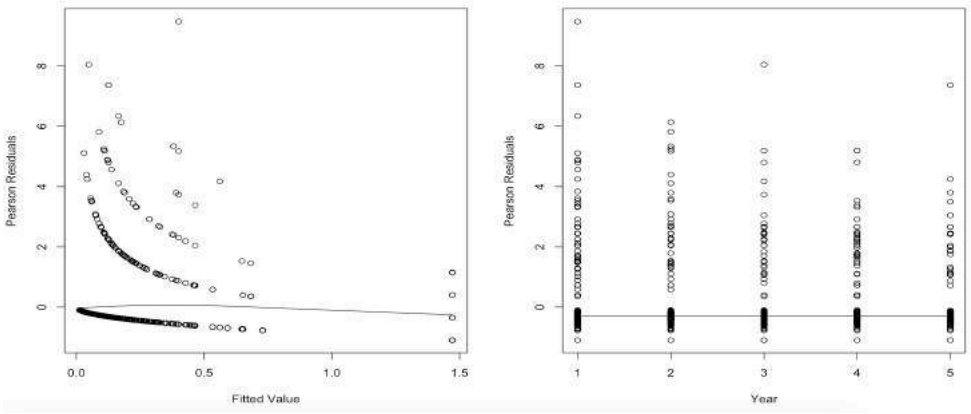
The correlation structure here $Corr(Y_{ij}, Y_{ij})$ is estimated as a unstructured matrix and variance of sample is defined as $Y_{ij} = \phi E[Y_{ij}]$.Then the estimated correlation structure is shown as below:

$$\begin{bmatrix} 1.00000 & 0.32462 & 0.15279 & 0.29707 & 0.18767 \\ 0.32462 & 1.00000 & 0.09347 & 0.13870 & -0.00246 \\ 0.15279 & 0.09347 & 1.00000 & -0.00953 & 0.13381 \\ 0.29707 & 0.13870 & -0.00953 & 1.00000 & -0.04172 \\ 0.18767 & -0.00246 & 0.13381 & -0.04172 & 1.00000 \end{bmatrix}$$

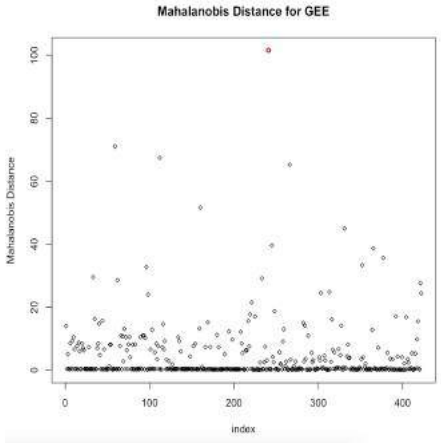
So here we can see that the beta-carotene treatment has significant effect on reducing the expected number of new skin cancers detected while the high exposure will contribute to the relapse of skin cancer. What's more, we can see that beta-carotene works better to reduce the risk of developing new cancers in female and younger patients. Here the $\hat{\phi}$ is only slightly larger than 1 means that there are no significant degree of overdispersion.

4.1 Diagnostics for GEE

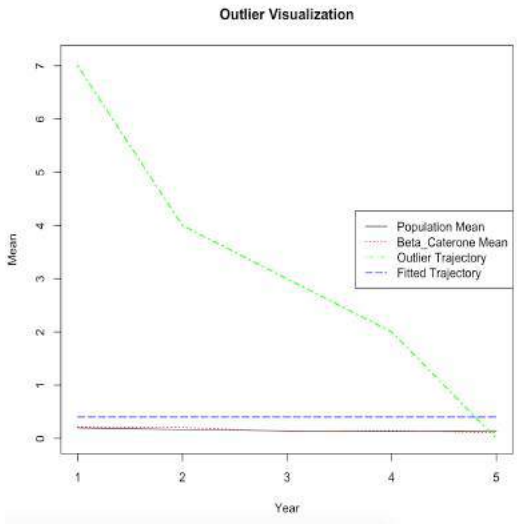
Here from CH13,[2], we use studentized pearson residuals $e_{ij} = \frac{Y_{ij} - g^{-1}(X_{ij}^T \hat{\beta})}{\sqrt{\phi v(\hat{\mu_{ij}})}}$ and the mahalanobis distance between the observed samples and fited values $d_i = r_i^T \hat{V^{-1}} r_i$ (where $\hat{V^{-1}}$ is the estimated covariance structure) to perform model diagnostics. Then we draw the residual versus fitted value and residual versus year plot. From the plots below we can see that there's no systematic trend for the fitted curves so it indicates that there's no lack of fit.



Then we plot the mahalanobis distance(it just remove the heterogeneity) to do the outlier detection. We treat the observation with the largest mahalanobis distance as red:



We observe that the 241th object has the largest mahalanobis distance so it may probably be the outlier case. Then we plot the trajectory for this object comparing with the global and its treatment groups' mean and the plot is shown as below:



We can see that obviously, the 241th observation is an outlier. Then we remove this object and fit the GEE again. Then the final model we choose is similar to that before removing outliers, mainly differ in the covariance structure, here we use the exchangeable covariance structure by applying the similar procedure of structure choosing as above from [3].

Variable	Estimate	Standard.error	p value
Intercept	-2.3893	0.1315	$< 2e-16$
trt	-3.1604	0.9479	0.00086
exposure	0.1289	0.0117	$< 2e-16$
I(trt*gender)	0.9087	0.2879	0.00160
I(trt*age)	0.0402	0.0136	0.00317
ϕ	1.13	_	_
ρ	0.104	_	_

The model fitted after removing outliers are similar to that before removing because the sign of variables are all the same and the magnitude of coefficients and the estimated scale parameter $\hat{\phi}$ are also very similar.

5 Model Comparison

Here for both the GLMM and GEE models, we do some visualizations which can also be denoted as trajectory display, we randomly select 16 objects from the dataset and draw the fitted lines by both the GLMM and GEE model and compare them(See Appendix), we can see that the model by GLMM is slightly more efficient than the GEE because the fitted trajectory is more like the observation points(for example the 3th, the 7th and the 11th random selected observation. Indeed it is reasonable because GLMM also include the random effect for each subject which is not explored by GEE. **What's more, from CH13[2] we know that if the missing pattern is MAR(missing at random) not MCAR(missing completely at random), we can see that in this case the GEE model may not be as efficient as in usual. But in this case, GLMM in R indeed use the maximum likelihood method which is not influenced by the missing pattern.**

So we conclude here that Generalized Linear Mixed Model fits the dataset better than the Generalized Estimating Equation.

6 Discussion

In this report for analyzing the skin cancer prevention study data, we compare the two methods very popular in the longitudinal data analysis: GLMM and GEE. We identify that the 241 th observation as the outlier case and validate it by visualizing the trajectory. Then not surprisingly, after conducting these two methods, we get pretty similar results. However, from the trajectory visualization, we can see that GLMM also include the subject-wise effect which borrows more information from the data points than the GEE which can be interpreted as marginal model. For the effect of beta-carotene, we can see that it can significantly reduce the risk of developing new skin cancer for patients. What's more, we can see that patients with more previous cancers are significantly more likely to develop new cancer.

References

[1] Greenberg,E.R.,Baron,J.A.,Stukel,T.A.,Stevens,M.M.,Mandel,J.S.,Spencer,S.K.,Elias,P.M.,Lowe,N.,Nierenberg,D.W.,Bayrd,G.,Vance,J.C.,Freeman,D.H.,Clendenning,W.E.,Kwan,T. and the Skin Cancer Group(1990) *A clinical trial of beta carotene to prevent basal-cell and squamous-cell cancers of the skin.* New England Journal of Medicine,323,789-795.

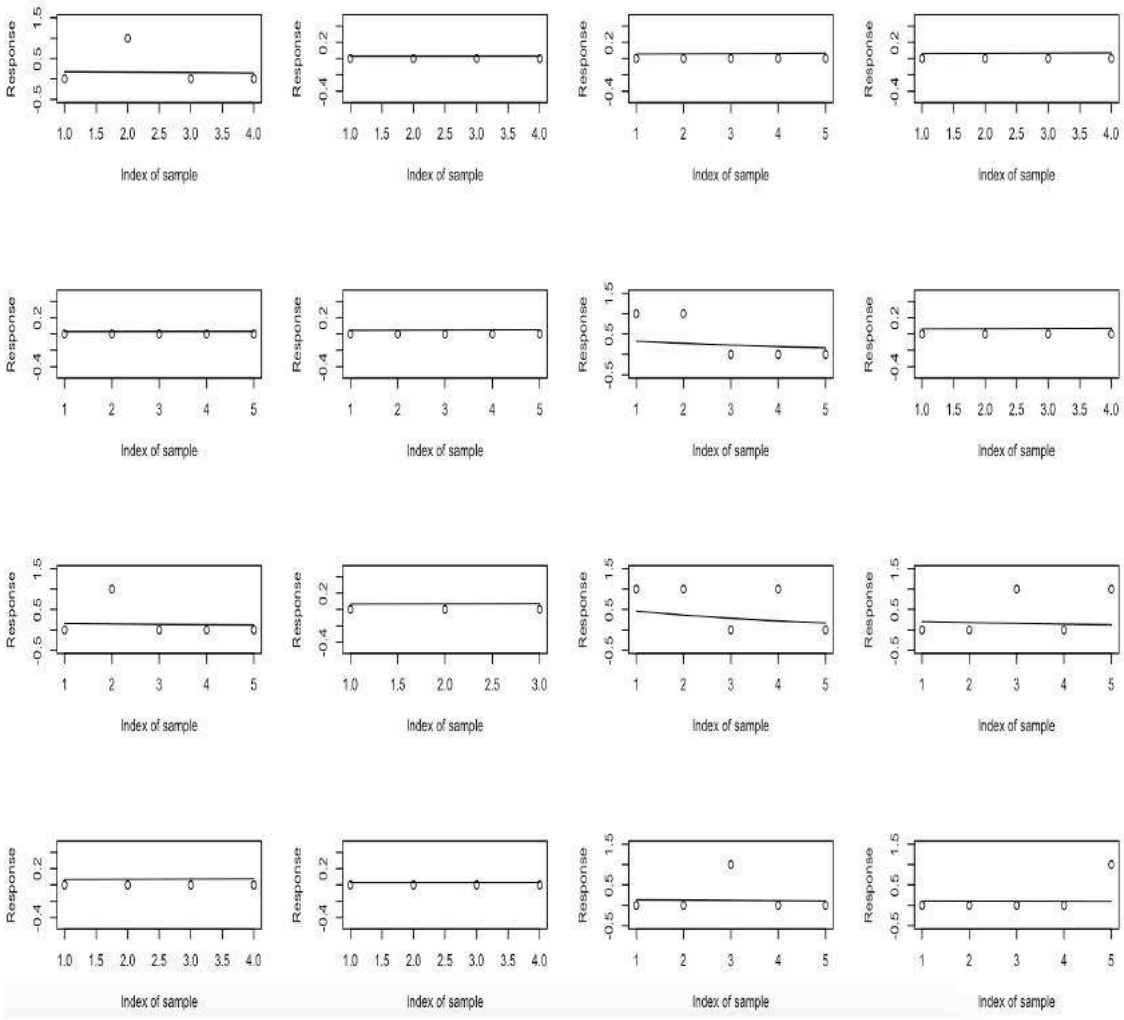
[2] Garrett M.Fitzmaurice,Nan M.Laird and James H.Ware. 2011. *Applied Longitudinal Analysis* John Wiley,Sons,Inc.,Hoboken,New Jersey.

[3] UNC ECOL 562 *Lecture 13 https* : *//www.unc.edu/courses/*2010*spring/ecol/*562/001 */docs/lectures/lecture*13*.htm#choosing*

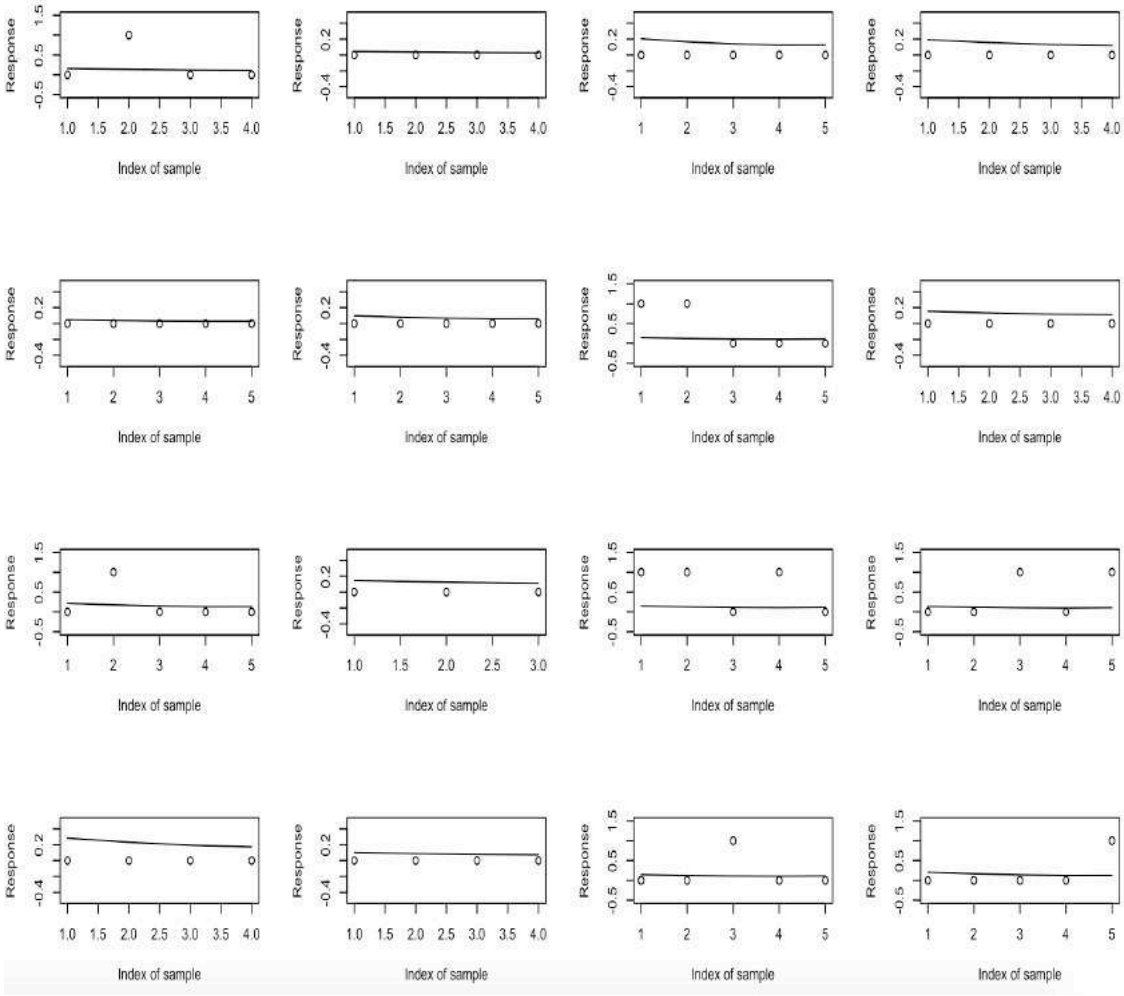
7 Appendix

7.1 Plots

Trajectory Visualization for GLMM models:



Trajectory Visualization for GEE models:



7.2 R code

```
#####################################################################
#
#Project for STA 224: Here we explore the and do corresponding diagnosis
#as well as comparing different possible models.
################
#load data and required packages:
library(foreign)
library(geepack)
library(lme4)
library(MASS)
library(lattice)
library(VIM)
library(mi)
skin_cancer<-read.dta("skin.dta")

#Here we restrict the analysis in the first group:
skin<-skin_cancer[skin_cancer$center==1,]
#Number of patients observed in center 1
dim(skin)
length(unique(skin[,1]))

##Exploratory analysis:
```

```
#Group mean and population(global) mean:
#Transform the data into wide format:
skin_wide<-reshape(skin,v.names="y",idvar="id",timevar="year",direction="wide")
response_placebo<-skin_wide[skin_wide[,7]==0,8:12]
response_beta_caterone<-skin_wide[skin_wide[,7]==1,8:12]
#calculate the group mean after removing the outliers:
population_mean<-apply(skin_wide[,8:12],2,mean,na.rm=TRUE)
placebo_mean<-apply(response_placebo,2,mean,na.rm=TRUE)
beta_caterone_mean<-apply(response_beta_caterone,2,mean,na.rm=TRUE)
#plot the mean:
plot(population_mean,type="l",ylim=c(0,0.5),lty=1, lwd=0.89,xlab="Year",ylab="Mean",main="Group and Population Means",)
lines(placebo_mean,col="blue",lty=2,lwd=1.20)
lines(beta_caterone_mean,col="red",lty=3,lwd=1.10)
legend("topright",legend=c("Population Mean","Placebo Mean",
"Beta_Caterone Mean"),col=c("black","blue","red"),lty=c(1,2,3))

##################
#GEE:
model_gee1<-geeglm(y~year+trt+I(trt*year)+age+skin+gender+exposure+ I(trt*age)+I(trt*skin)+I(trt*exposure)+I(trt*gender)
        +I(year^2)+I(year^3),id=id,data=skin,family= poisson("log"),corstr="unstructured")
model_gee2<-geeglm(y~trt+exposure+I(trt*age)+I(trt*gender),id=id,data=skin,family=poisson("log"),corstr="unstructured")
model_gee3<-geeglm(y~trt+exposure,id=id,data=skin,family= poisson("log"),corstr="unstructured")
model_gee4<-geeglm(y~trt+exposure+I(trt*age),id=id,data=skin,family= poisson("log"),corstr="unstructured")
anova(model_gee1,model_gee2)
anova(model_gee2,model_gee4)
anova(model_gee4,model_gee3)
model1_gee<-geeglm(y~trt+exposure+I(trt*age)+I(trt*gender),id=id,data=skin,family=poisson("log") ,corstr="unstructured")
model2_gee<-geeglm(y~trt+exposure+I(trt*age)+I(trt*gender),id=id,data=skin,family=poisson("log") ,corstr="ar1")
model3_gee<-geeglm(y~trt+exposure+I(trt*age)+I(trt*gender),id=id,data=skin,family=poisson("log") ,corstr="exchangeable")
model4_gee<-geeglm(y~trt+exposure+I(trt*age)+I(trt*gender),id=id,data=skin,family=poisson("log") ,corstr="independence")
#select the best model by the difference between the naive  Correlation
estimation and the robust correlation estimation:
dif1<-sum(abs(model1_gee["geese"]]$vbeta-model1_gee["geese"]]$vbeta.naiv))
dif2<-sum(abs(model2_gee["geese"]]$vbeta-model2_gee["geese"]]$vbeta.naiv))
dif3<-sum(abs(model3_gee["geese"]]$vbeta-model3_gee["geese"]]$vbeta.naiv))
dif4<-sum(abs(model4_gee["geese"]]$vbeta-model4_gee["geese"]]$vbeta.naiv))
#So here we use the unstructured covariance structure:



#Diagnostic w.r.t GEE and identification of outliers:
#Calculate the pearson residuals:
fitted_value<-fitted(model1_gee)
#Pearson residuals:
res_0<-skin$y-fitted_value
resid<-(skin$y-fitted_value)/sqrt(1.21*fitted_value)
#Perform chisquare test:
sum(resid^2)
#Residual plot:
par(mfrow=c(1,2))
plot(fitted_value,resid,xlab="Fitted Value",ylab="Pearson Residuals")
lines(smooth.spline(fitted_value,resid,spar=1.3))
scatter.smooth(skin$year,resid,xlab="Year",ylab="Pearson Residuals")

#Calculate and plot the mahalanobis distance for each one:
#transform it to wide:
skin_wide<-reshape(skin,v.names="y",idvar="id",timevar="year",direction="wide")
respons<-skin_wide[,8:12]
nn<-apply((1-is.na(respons)),1,sum)
m<-length(nn)
pval<-rep(0,m)
d<-rep(0,m)
cormat<-matrix(c(1,0.32462,0.15279,0.29707,0.18767,
                        0.32462,1,0.09347,0.13870,-0.00246,
                        0.15279,0.09347,1,-0.00953,0.13381,
                        0.29707,0.13870,-0.00953,1,-0.04172
                        ,0.18767,-0.00246,0.13381,-0.04172,1),5,5)
for(i in 1:m){
        end<-sum(nn[1:i])
        start<-end-nn[i]+1
        #C is the estimated correlation structure
        na_index<-which(is.na(skin_wide[i,8:12]))
        C<-cormat[is.na(pmatch(c(1:5),na_index)),is.na(pmatch(c(1:5),na_index))]
        if(start!=end){
                A<-diag(sqrt(1.27*fitted_value[start:end]))
        }
        else{
                A<-sqrt(1.27*fitted_value[start:end])
        }
        V=A%*%C%*%A
        d[i]<-t(as.matrix(res_0[start:end]))%*%solve(V)%*%as.matrix(res_0[start:end])
        pval[i]<-pchisq(d[i],nn[i],lower.tail=FALSE)
}

plot(d,cex=0.6,main="Mahalanobis Distance for GEE",xlab="index",ylab="Mahalanobis Distance",pch=5)
points(order(pval)[1],max(d),col="red")
#Identify the largest mahalanobis distance:
```

```
index_outlier<-which.max(d) #241
fitted_gee<-fitted(model1_gee)
end_point<-sum(nn[1:index_outlier])
start_point<-end_point-nn[index_outlier]+1
#Plot the trajectory of this object with the global mean and beta-carotene group mean:
plot(population_mean,type="l",ylim=c(0,7.1),lty=1,lwd=0.89,xlab="Year", ylab="Mean",main="Outlier Visualization")
lines(beta_caterone_mean,col="red",lty=3,lwd=1.10)
lines(c(1:5),skin_wide[index_outlier,8:12],col="green",lty=4,lwd=1.85)
lines(fitted_gee[start_point:end_point],col="blue",lty=5,lwd=1.41)
legend("right",legend=c("Population Mean","Beta_Caterone Mean","Outlier Trajectory",
"Fitted Trajectory"),col=c("black","red","green","blue"),lty=c(1,3,4,5))

#Then we remove the outlier object and fit the GEE again:
skin_wide<-skin_wide[-241,]
skin<-skin[-(start_point:end_point),]
model5_gee<-geeglm(y~year+trt+I(trt*year)+age+skin+gender+exposure+I(trt*age) +I(trt*skin)+I(trt*exposure)+I(trt*gender)
        +I(year^2)+I(year^3),id=id,data=skin,family=poisson("log"),corstr="unstructured")
summary(model_gee5)
model6_gee<-geeglm(y~trt+exposure+I(trt*gender)+I(trt*age),id=id,data=skin,family=poisson("log"),corstr="unstructured")
anova(model_gee6,model_gee5)

model7_gee<-geeglm(y~trt+exposure+I(trt*gender)+I(trt*age),id=id,data=skin,family=poisson("log"),corstr="ar1")
model8_gee<-geeglm(y~trt+exposure+I(trt*gender)+I(trt*age),id=id,data=skin,family=poisson("log"),corstr="exchangeable")
model9_gee<-geeglm(y~trt+exposure+I(trt*gender)+I(trt*age),id=id,data=skin,family=poisson("log"),corstr="independent")
dif1<-sum(abs(model6_gee[["geese"]]$vbeta-model1_gee[["geese"]]$vbeta.naiv))
dif2<-sum(abs(model7_gee[["geese"]]$vbeta-model2_gee[["geese"]]$vbeta.naiv))
dif3<-sum(abs(model8_gee[["geese"]]$vbeta-model3_gee[["geese"]]$vbeta.naiv))
dif4<-sum(abs(model9_gee[["geese"]]$vbeta-model4_gee[["geese"]]$vbeta.naiv))
#Choose exchangeable covariance structure:
summary(model8_gee)

##################
#GLMM:
model_glmm1<-glmer(y~year+trt+age+skin+gender+exposure+I(trt*year)+(year|id),data=skin,family=poisson("log"))
model_glmm2<-glmer(y~age+gender+exposure+(year|id),data=skin,family=poisson("log"))
model_glmm3<-glmer(y~year+age+gender+exposure+(1|id),data=skin,family=poisson("log"))
#apply the likelihood ratio test to compare different models:
anova(model_glmm1,model_glmm2,test="Chi")
anova(model_glmm2,model_glmm3,test="Chi")
#So model_glmm2 is our final model for GLMM.
#Diagnostics:
empi_blup<-ranef(model_glmm2)$id
par(mfrow=c(1,2))
hist(empi_blup[,1],main="Histogram for Empirical BLUP for Random Intercept")
hist(empi_blup[,2],main="Histogram for Empirical BLUP for Random Slope")


#First calculate the degree of freedom:
m_random<-VarCorr(model_glmm2)
m_fixed<-fixef(model_glmm2)
df<-3+length(m_fixed)
test_stat<-sum(residuals(model_glmm2,type="pearson")^2)
pvalue_glmm<-pchisq(test_stat,nrow(skin)-df,lower.tail=FALSE)
#Here we have identified that the 52th object is the outlier by the histogram of EBLUP:
outlier<-which.max(empi_blup[,1])
plot(1:5,skin_wide[52,8:12],type="l",xlab="year",ylab="Response(# of new skin cancer)",col="red")
lines(apply(skin_wide[skin_wide$trt==0,8:12],2,mean,na.rm=TRUE))

######################
#Trajectory Visualization and comparison:
#Here we randomly choose 16 objects randomly and display their trajectory and
#the fitted curve by GLMM and GEE.
samp<-sample(1:m,16)
fitted_gee<-fitted(model_gee1)
fitted_glmm<-fitted(model_glmm2)
yy<-skin_wide[,8:12]
par(mfrow=c(4,4))
for(i in 1:16){
        end<-sum(nn[1:samp[i]])
        start<-end-nn[samp[i]]+1
        ylow<-min(yy[samp[i],(1:nn[samp[i]])])-0.5
        yhigh<-max(yy[samp[i],(1:nn[samp[i]])])+0.5
        plot(fitted_gee[start:end],ylim=c(ylow,yhigh), type="l",xlab="Index of sample",ylab="Response")
        points(1:nn[samp[i]],yy[samp[i],(1:nn[samp[i]])],type="p")
}
par(mfrow=c(4,4))
for(i in 1:16){
        end<-sum(nn[1:samp[i]])
        start<-end-nn[samp[i]]+1
        ylow<-min(yy[samp[i],(1:nn[samp[i]])])-0.5
        yhigh<-max(yy[samp[i],(1:nn[samp[i]])])+0.5
 plot(fitted_glmm[start:end],ylim=c(ylow,yhigh),type="l",xlab="Index of sample",ylab="Response")
        points(1:nn[samp[i]],yy[samp[i],(1:nn[samp[i]])],type="p")
}

######################
```

Improving High School Math Education in Portugal in GLM Framework

Heqiao Ruan Instructor: Hans-Georg Mueller

March 14, 2018

1 Abstract

Portugal is a country located in southwest Europe and its educational level has improved over the last decades. However the statistics keep the Portugal at the Europe's tail end in education due to its high student failure rate and dropping out rate in fundamental subjects such as Math and Portuguese in secondary school. So the official has realized this serious problem and it becomes pretty inspirational to identify what improve the students' math grade so that we can give corresponding advice to improve the math education in Portugal. Here we use the student performance dataset in UC Irvine machine learning dataset repository to conduct our research. Current research typically applied some machine learning algorithms such as ANN and SVM to perform binary/multi-label classification. We plan to extend their research by conducting techniques in generalized linear model and transformed model. Finally we will try to provide some insightful suggestions to improve middle school math education in Portugal standing in school's position.
For details of variables in the dataset,refer to Appendix A.

2 Background and Introduction

The Generalized Linear Model which is also named GLM is to examine the non-linear relationship between the response variables and predictors. The GLM has form $g(E[Y|X = x]) = \beta_0+\beta_1X_1+..+\beta_pX_p = X\beta = \eta$ where g is the link function and $\eta = X\beta$ is the linear predictor. What's more, $\mu = E[Y]$ is another component. So it concludes the three components.
The dataset we use depicts the Portugal secondary school students' performance in math. It consists of 395 observations and 33 features. In the 33 features, the first 30 describes the students' status during the semester and the other three are the students' math grades for stage I, stage II and the final stage which is denoted as G1,G2 and G3 respectively. Here we apply various models in the framework of GLM and identify what influences the performance of the student in math exam. We point out that among the 30 predictors, most of them are categorical and for these predictors we use indicator variables to fit.

3 Methods

3.1 Logistic Regression

Previous research claims that the serious problem in Portugal education partly comes from the students' high failure rate in a couple of key subjects. So we fit the logistic regression model to identify what determines the students' failure rate in math. As we all know, logistic regression can be treated as a binary classification we also perform ten fold cross validation to examine the model's prediction ability as well checking the fit.

3.2 Multinomial Regression

Multinomial Regression stands for the GLM dealing with multiple response.As previous research states, students' grade are divided into five levels:A,B,C,D,F**(table[1])**. First we try to build two kinds of multinomial models. However, the prediction performance error of proportional odds model and baseline odds model is pretty high**(table[3] and table [4])**, we merge the BCD as medium and A as high and F as low to explore the relationship**(See table [5])**. Then we fit both baseline odds model and proportional odds models and compare their performance and decide a final model after performing model selection identifying the important elements decide students' performance.

3.3 Transformed Model

Based on the nature of the data we can see that the grade is divided into three categories: G1,G2 and G3.From previous research we know that it will be more efficient to predict G3 with G1 and G2. So we would like to extract information from the three components and we perform principal component analysis here and extract the first principal component which explains most of the variability in the data. As the first component represents more than 90% of the total variance(table[10]), this technique is well-rounded in this case.
We scale the combined grade(weighted average of G1,G2 and G3) into percentile which range from ϵ to $1-\epsilon$(continuous). Then we apply logit transformation on the combined grade which is denoted as Y for the distribution of $logit(Y) = log(\frac{Y}{1-Y})$ is more like normal distribution(**graph[2]**)**(It indeed help alleviate the heavy tail)**. Then we apply the general linear regression model(special case of GLM with identity link) which has the form $logit(Y) = \beta X$. Note that this is a transformed model, not a generalized linear model.

4 Main Results

First from the **Logistic Regression** model, we get the classification error of 24.4%**(Table 16)** which is a pretty decent classification.(after cross validation, still decent,**Graph 3(II)**) Then for the model fitting, we observe that **Age,Failures, Schoolsupport ,FamilySupport,Goout** are significant. We can see that students' passing chance are negatively related

to age,goout, familysupport,failures. However, as we all know, schoolsupport will highly different whether the students' fail before. So we add the interaction terms **Schoolsup:Failures** and observe that it is significant. Then from **Table 2** we can see that students' probability of fail in math exam is positively related to their **past failures, time of going out with friends** and negatively related to the students' age. Then from **Table 2** we also find the interaction effect between school support and failure is significant. What's more, outlier analysis including leverage value and cook distance can be seen in **(Table 15 and Graph 8)** . **Then we fit the multinomial regression model**, if we fit the 5 category proportional odds model and baseline odds model respectively, the prediction error are 56.2% and 53.9% respectively.(**See Table 3 and Table 4**). So we merge the categories as three,(**See Table 5**), low, medium and high. So after fitting the proportional odds model and baseline odds model, the prediction errors are 36.2%,31.9%(see **Table6,Table7**) respectively. From **Graph 5** we can see that there are no obvious evidence of lack of fit in the proportional odds model while from **Graph 6** we can see that there is somewhat lack of fit in the baseline odds model(Pearson residuals are messy). What's more, various metrics measuring the multi-label classification performance is shown in **Table 1 (II)** which all indicates that the prediction fitting is fairly decent. Then we perform model selection by AIC criterion and the final model is shown in **Table2(I)**.

Then to compare the proportional odds model and the baseline odds model, we can see that in the baseline odds model none of the predictors in $Medium|Low$ is significant level 0.05, and the number of parameters in baseline odds model are pretty high,**(table 9)** what's more the baseline odds model shows obvious evidence of lack of fit**(Graph 6)**. **So we tend to prefer proportional odds type model to baseline odds model.(From Table 1(III),Graph 5,proportional odds type model fits decently)** Then we perform 10-fold cross validation to predict, the mean CV error is 41.6% which is normal for multi-label classification.**(Graph 3)**

So from the proportional odds type model we can see that **Age, Fjob, Failures, goouts, school support** are significant in determining the probability of getting a better math grade(Especially A). So we can see that the probability of students getting a better math grade is positively related to Fjobteacher and are negatively related to Age, Failures, Goouts. **For school support, its sign is not consistent to our common sense, so we add interaction term** $Failures : Schoosupport$ **and it's significant.** Then we can see that What's more we can also draw various conclusions. First, boys tends to perform better than girls and surprisingly we can see that students live in rural areas performs no worse than students living in urban areas, secondly, the students studying longer tend to get a better grade. What's more, students those mother has a higher education with internet access at home are more probable to get a higher math grade.

Then we want to include the influence of grade in the first two stages then we perform the logit transformed model. Then after fitting the model, we use AIC criterion to do model selection and from **Graph 7** we can see that no obvious pattern in residuals and the qq plot shows roughly normal which means a somewhat decent fit.Then we can see that **sexM, studytime, failures, schoolsupyes, famsupyes, goout,schoolsupyes:failures** are significant and the

interaction effect between school support:failure is significant as well. We can see that the result of transformed model is similar to the two previous models which means it is pretty meaningful to track students along the whole process.

5 Conclusion

The **significant predictors and the sign of coefficient** is shown as below in the three model(logistic regression, proportional odds model,logit transformed model) are shown as below.**(For detail, please refer to Appendix B)**

Logistic Model	Coef	ProportionOdd Model	Coef	Transform Model	Coef
sexM	0.569	sexM(BOY)	0.563	SexM(BOY)	0.006
failures	-1.233	Failures	-1.278	failures	-0.418
schoolsupyes	-1.334	schoolsupyes	–1.404	schoolsupyes	-0.465
goout	-0.346	goout	-0.346	goout	-0.1195
famsupyes	-0.615	age	-0.196	famsupyes	-0.2123
age	-0.217	health	0.161	studytime	0.1466
scsup:fail	1.429	scsup:fail	1.478	scsup:fail	0.403

We can see that the significant factors have the same sign and similar scale of coefficients which means the models are consistent in terms of significant predictors. It indeed validate the conclusion in [1] that predicting G3 will be more efficient with the information of G1 and G2.We can see that young men performs significantly better than young girls and young students tend to perform better in math **(It is indeed explanable that 15-16 is a adequate age for high school so older students may have difficulties in study even before high school)**. What's more, the students who fail more times before, go out to party too frequently tend to get a worse grade which is pretty straight forward to understand. Then for the influence of schoolsupport, we can see that for students with no failure history, students' grade are negatively related with school support(**Trivial by the definition of Indicator variable**) while for students with serious failure history, school support will have a better effect to students. Then some of the models indicate that healthier students are more probable to get a better grade which fits our common sense.
What's more, there are some other predictors we may particularly interest in which can help us to adjust the policy shown in the below table(Although may not that significant):

Logistic Model	Coef	ProportionOdd Model	Coef	TransformModel	Coef
higheryes	0.750	higheryes	0.874	higheryes	0.367
famsup	-0.4626	famsup	-0.439	famsup	-0.221
health	-0.148	health	-0.161	health	-0.062

From the above table we can see that students in a better health state are more likely to get a good grade. What's more, we can see that students willing to pursue a higher degree are more probable to get a good grade. It is pretty insightful that **math is an important**

prerequisite subject for most area in science and technology so the students willing to pursue a higher degree are more motivated so they will not only spend more time in studying(see **Table 12**) but also getting a higher grade**(see Table 13)** we can see that all students get A are willing to pursue a higher degree. Then for the influence of family support, we can see that the students' probability getting a better grade is negatively related to the family support and this is somewhat explanable as most parents are not experts in math education.**(See Table 14 which fit famsup individually)**.
So here we can see that the significant(important) predictors in all of the three models are similar.

6 Discussion and Corresponding Advice

In this investigation about **1/3 of all students fails in math which is one of the most important subject in high school** so we want to find what is significant in resulting to a higher grade and then propose corresponding advice to improve education quality in Portugal. From the point view of school, first we should guide students to arrange their time independently if they are competent **(school sup coef neg for no fail)** and provide help only for those who have failed before as well as prevent students from distracting from study during the process**(goout coef neg)**.More importantly, to decrease the failure rate of students and ensure everyone keep up with the course, we should particularly focus on those students who have failed before.Obviously,if you fail to follow math course this quarter, you will never understand it in the next quarter which will potentially lead to even higher failure rate in the future.For parents, we advice them not intervene on children's study as most of them they aren't expert in this area**(famsup coef neg,Table14)**. Another issue we have to point out is that in a well-developed country, university is one indispensable part of education so students should not only have motivation to pursue higher degree than secondary school but also the qualification to keep up with college level study. Then to arouse students' motivation to study math**(higheryes positive)** is pretty important too. What's more, as young men tend to perform better than young women, we should pay more attention to the girls' study especially for those have some difficulties in studying.**(SexM coef Positive)**. Finally we have to point out that the data we use to fit the model is somewhat limited for it ignores the differential influence of time. So further study may require the longitudinal type of data to repeatedly measure the students' performance and the predictors which may varies by time along the whole process of their study.

References

[1] Paulo Cortez and Alice Silva. *Using Data Mining to Predict Secondary School Student Performance*. University of Minho, Guimaraes, Portugal

[2] Hans-Georg Mueller. *Generalized Linear Models Lecture Notes* UC Davis Winter 2018

7 Appendix

7.1 Appendix A:Description of Datasets

Predictor Variables:
1.School- Student's School (Binary:"GP"-Gabriel Pereira or "MS"-Mousinho da Silveira)
2.Sex- Student's sex (Binary:"F"-female or "M"-male)
3 age - student's age (numeric: from 15 to 22)
4 address - student's home address type (binary: "U" - urban or "R" - rural)
5 famsize - family size (binary: "LE3" - less or equal to 3 or "GT3" - greater than 3)
6 Pstatus - parent's cohabitation status (binary: "T" - living together or "A" - apart)
7 Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
8 Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
9 Mjob - mother's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "*athome*" or "other")
10 Fjob - father's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "*athome*" or "other")
11 reason - reason to choose this school (nominal: close to "home", school "reputation", "course" preference or "other")
12 guardian - student's guardian (nominal: "mother", "father" or "other")
13 traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
14 studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15 failures - number of past class failures (numeric: n if $1 <= n < 3$, else 4)
16 schoolsup - extra educational support (binary: yes or no)
17 famsup - family educational support (binary: yes or no)
18 paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19 activities - extra-curricular activities (binary: yes or no)
20 nursery - attended nursery school (binary: yes or no)
21 higher - wants to take higher education (binary: yes or no)
22 internet - Internet access at home (binary: yes or no)
23 romantic - with a romantic relationship (binary: yes or no)
24 famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25 freetime - free time after school (numeric: from 1 - very low to 5 - very high)
26 goout - going out with friends (numeric: from 1 - very low to 5 - very high)
27 Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28 Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29 health - current health status (numeric: from 1 - very good to 5 - very bad)
30 absences - number of school absences (numeric: from 0 to 93)

Response Variables:
G1:Math Grade for first stage
G2:Math Grade for second stage
G3:Math Grade for final stage

7.2 Appendix B:Chosen Model for three types

1.Logistic regression Model:
$Logit(E[Y|X]) \sim \beta_0+\beta_1 I(sexM)+\beta_2 age+\beta_3 I(Mjobhealth)+\beta_4 I(Mjobother)+\beta_5 I(Mjobservice)+\beta_5 I(Mjobteacher)+\beta_6 failures+\beta_7 I(schoolsupYes)+\beta_8 I(famsupYes)+\beta_9 I(higherYes)+\beta_{10} goout+\beta_{11} health+\beta_{12} failures*I(schoolsupYes)$
2.Proportional Odds Type Model:
$logit(P(Y \leq k))=\beta_{0,k}+\beta_1 I(sexM)+\beta_2 age+\beta_3 I(PstatusT)+\beta_4 I(Mjobhealth)+\beta_5 I(Mjobother)+\beta_6 I(Mjobservice)+\beta_7 I(Mjobteacher)+\beta_8 studytime+\beta_9 failures+\beta_{10} I(schoolsupYes)+\beta_{11} I(famsupYes)+\beta_{12} I(higheryes)+\beta_{13} freetime+\beta_{14} goout+\beta_{15} health+\beta_{16} failures*I(schoolsupYes)$
3.Logit Transformed Model:
$E[Y|X]=E[ScaledScore|X]=\beta_0+\beta_1 sexM+\beta_1 I(Mjobhealth)+\beta_2 I(Mjobother)+\beta_3 I(Mjobservice)+\beta_4 I(Mjobteacher)+\beta_5 studytime+\beta_6 failures+\beta_7 I(schoolsupYes)+\beta_8 I(famsupYes)+\beta_9 I(higherYes)+\beta_{10} goout+\beta_{11} failures*schoolsupyes+\beta_{12} freetime+\beta_{13} health$
For detailed coefficients, please refer to Appendix E.

7.3 Appendix C:Significant Predictors

1.Logistic Regression Model:

Predictors	Coefficient	Standard Error	z value	P Value
age	-0.217	0.108	-2.002	0.045
sexM	0.569	0.268	2.126	0.033
failures	-1.233	0.226	-5.460	4.76e-8
schoolsupyes	-1.334	0.385	-3.462	5.36e-4
goout	-0.346	0.114	-3.039	2.37e-3
higheryes	0.965	0.588	1.641	0.100
failures:schoolsupyes	1.412	0.475	2.982	2.87e-3

2.Proportional Odds Type Model after model selection

Predictors	Coefficient	Standard Error	z value	PVALUE
sexM	0.562	0.239	2.345	0.0195
age	-0.196	0.095	-2.065	0.396
failures	-1.278	0.221	-5.77	1.62e-8
schoolsupyes	-1.404	0.361	-3.888	1.19e-4
goout	-0.346	0.107	-2.047	0.041
higheryes	0.904	0.572	1.58	0.057
failures:schoolsupyes	1.478	0.447	3.309	1.027e-3

2(I):Intercepts

Prediction	Value	Std.Error	t value	pvalue
1|2	-4.706	1.863	-2.526	0.012
2|3	-1.163	1.842	-0.633	0.5270

3.Logit Transformed Model after model selection

Predictors	Coefficient	Std.Error	t value	pvalue
sexM	0.251	0.092	2.735	6.53e-3
studytime	0.147	0.053	2.784	5.64e-3
failures	-0.418	0.0643	-6.506	2.5e-10
schoolsupyes	-0.465	0.137	-3.401	7.45e-4
famsupyes	-0.212	0.0872	-2.433	0.0154
romanticyes	-0.218	0.0890	-2.446	0.015
goout	-0.119	0.039	-3.098	0.002
health	-0.063	0.0299	-2.087	0.038
higher	0.750	0.632	1.186	0.236
failures:schoolsupyes	0.403	0.167	2.408	0.017

7.4 Appendix D:Tables

Table 1: Distribution of levels

A(G3≥16)	$B(13 < G3 < 16)$	C$(11 < G3 < 14)$	D$(9 < G3 < 12)$	F$(G3 < 10)$
40	60	62	103	130

Table1(II):Measure of prediction performance of proportional odds type model with respect to merged categories:

Accuracy	Precision	Recall	F-Score	F-Score$(\beta = 0.5)$
0.638	0.649	0.703	0.676	0.683

Table1(III):Runs test for proportional odds type model of merged categories:

```
> runs.test(rdi_high)

        Runs Test - Two sided

data:  rdi_high
Standardized Runs Statistic = -1.1586, p-value = 0.2466

> runs.test(rpi_high)

        Runs Test - Two sided

data:  rpi_high
Standardized Runs Statistic = -1.2594, p-value = 0.2079

> runs.test(rdi_medium)

        Runs Test - Two sided

data:  rdi_medium
Standardized Runs Statistic = 0.45355, p-value = 0.6502

> runs.test(rpi_medium)

        Runs Test - Two sided

data:  rpi_medium
Standardized Runs Statistic = -0.15101, p-value = 0.88
```

Table 2(I): Full coefficient table of Logistic Regression Model after model selection:

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	5.55492	2.07353	2.679	0.007385	**
sexM	0.56958	0.26795	2.126	0.033528	*
age	-0.21678	0.10828	-2.002	0.045281	*
Mjobhealth	0.77209	0.57963	1.332	0.182844	
Mjobother	-0.18179	0.36658	-0.496	0.619949	
Mjobservices	0.72741	0.41422	1.756	0.079075	.
Mjobteacher	-0.47103	0.44765	-1.052	0.292702	
failures	-1.23304	0.22583	-5.460	4.76e-08	***
schoolsupyes	-1.33381	0.38524	-3.462	0.000536	***
famsupyes	-0.46258	0.26588	-1.740	0.081891	.
higheryes	0.96506	0.58799	1.641	0.100740	
goout	-0.34590	0.11381	-3.039	0.002372	**
health	-0.14809	0.09191	-1.611	0.107115	
failures:schoolsupyes	1.41762	0.47541	2.982	0.002865	**

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

Table 3:Proportional Odds Model(5 Category) Prediction Result

Prediction/Level	A	B	C	D	F
A	7	5	2	3	1
B	16	24	10	10	4
C	0	0	0	0	0
D	16	27	27	56	39
F	1	4	23	34	86

Table 4:Baseline Odds Model(5 Category) Prediction Result

Prediction/Level	A	B	C	D	F
A	21	7	2	4	1
B	7	28	9	12	9
C	1	6	14	7	8
D	7	9	17	48	25
F	4	10	20	32	87

Table 5: Merged Categories

Low(G3<10)	Medium($10 \leq G3 < 16$)	High($G3 \geq 16$)
130	225	40

Table 6:Merged Proportional odds Model Prediction:

Predic/Level	Low	Medium	High
Low	57	32	0
Medium	73	193	38
High	0	0	2

Table 7:Merged Baseline Odds Model Prediction:

Predic/Level	low	Medium	High
Low	63	27	1
Medium	67	193	26
High	0	5	13

Table 8:Merged Proportional Odds Model after model selection:

	Value	Std. Error	t value	PVALUE
sexM	0.5623557	0.23980239	2.3450795	1.953820e-02
age	-0.1960426	0.09493438	-2.0650328	3.959949e-02
PstatusT	-0.5747717	0.35072678	-1.6388019	1.020843e-01
Mjobhealth	0.8086347	0.47814796	1.6911810	9.162431e-02
Mjobother	-0.1971477	0.33410197	-0.5900824	5.554870e-01
Mjobservices	0.9190661	0.36646843	2.5078999	1.256228e-02
Mjobteacher	-0.1408073	0.40782700	-0.3452624	7.300887e-01
studytime	0.2214303	0.13714921	1.6145209	1.072467e-01
failures	-1.2778355	0.22131517	-5.7738271	1.615042e-08
schoolsupyes	-1.4039719	0.36112172	-3.8878081	1.194119e-04
famsupyes	-0.4383335	0.23090447	-1.8983325	5.841172e-02
higheryes	0.9040662	0.57168980	1.5813930	1.146224e-01
freetime	0.1858861	0.11710992	1.5872786	1.132835e-01
goout	-0.3463535	0.10666424	-3.2471379	1.269591e-03
health	-0.1608674	0.07860043	-2.0466475	4.138206e-02
failures:schoolsupyes	1.4775380	0.44654813	3.3087990	1.026689e-03
1|2	-4.7060509	1.86272531	-2.5264331	1.192908e-02
2|3	-1.1662618	1.84196615	-0.6331613	5.270102e-01

Table 9:Baseline Odds Model(We ignore this model due to severe lack of fit):

(Intercept)	4.58362118	6.269436e-06	-10.94312390	0.0000000000
schoolMS	0.24737426	8.047565e-01	0.07045897	0.9438666257
sexM	0.46728876	6.405699e-01	0.77622095	0.4381171248
age	-0.24232833	8.086608e-01	-0.22082475	0.8253513563
addressU	0.31105183	7.559373e-01	-0.86530717	0.3874339513
famsizeLE3	0.10179867	9.189719e-01	0.43674153	0.6625549525
PstatusT	-0.53891908	5.902686e-01	-0.67917484	0.4974540256
Medu	0.02700642	9.784693e-01	0.74426068	0.4571937551
Fedu	0.15287676	8.785793e-01	-0.41261896	0.6801260881
Mjobhealth	0.46518106	6.420769e-01	0.79692363	0.4260092315
Mjobother	-0.31661063	7.517186e-01	-0.90162694	0.3678447884
Mjobservices	0.38457061	7.007777e-01	1.55723716	0.1202737404
Mjobteacher	-0.87951071	3.796984e-01	-0.90831459	0.3643064810
Fjobhealth	-0.18182527	8.558200e-01	0.93059982	0.3526706731
Fjobother	0.13071636	8.960711e-01	-0.06167560	0.9508546571
Fjobservices	-0.05621325	9.552024e-01	-0.25308321	0.8003452330
Fjobteacher	0.36884464	7.124556e-01	2.53042415	0.0118086660
reasonhome	0.28085369	7.789805e-01	0.58058047	0.5618785883
reasonother	0.33304118	7.392929e-01	0.42719634	0.6694861686
reasonreputation	0.50900033	6.110570e-01	0.26454778	0.7915059582
guardianmother	-0.08830207	9.296846e-01	-0.72307686	0.4700919795
guardianother	0.11987435	9.046481e-01	1.03981192	0.2991099729
traveltime	0.11225697	9.106809e-01	-0.34150253	0.7329202177
studytime	0.19401019	8.462749e-01	0.35721050	0.7211389868
failures	-1.09482063	2.743113e-01	-3.10179403	0.0020719934
schoolsupyes	-1.17825789	2.394551e-01	-3.43449575	0.0006613172
famsupyes	-0.63168210	5.279868e-01	-0.61313037	0.5401688827
paidyes	0.39371663	6.940183e-01	-1.14543582	0.2527730193
activitiesyes	-0.15250935	8.788688e-01	-0.60960252	0.5425013335
nurseryyes	-0.45366562	6.503369e-01	0.56047009	0.5754998811
higheryes	0.62012420	5.355598e-01	14.03167383	0.0000000000
internetyes	0.13872557	8.897429e-01	1.49103779	0.1368082297
romanticyes	-0.23149785	8.170567e-01	-0.67681433	0.4989489631
famrel	0.15941320	8.734308e-01	0.32870540	0.7425653398
freetime	0.12155331	9.033191e-01	0.13800656	0.8903107160
goout	-0.50608123	6.131027e-01	-0.39721859	0.6914365538
Dalc	-0.04955655	9.605027e-01	-0.20889932	0.8346424198
Walc	0.22800481	8.197691e-01	0.19654899	0.8442889900
health	-0.10216668	9.186800e-01	-0.38282725	0.7020688773
absences	-0.01130261	9.909881e-01	-0.05874724	0.9531853203
failures:schoolsupyes	1.33983821	1.811247e-01	-17.63425728	0.0000000000

Table 10:PCA of G1,G2,G3:

Factor	PC1	PC2	PC3
Proportion of Variance	0.9095	0.06162	0.02892
Cumulative Proportion	0.9095	0.9711	1
G1	0.4629	0.8024	-0.3764
G2	0.5614	0.0632	0.8251
G3	0.6859	-0.5933	-0.4212

Table 11:Logit Transformed model after AIC model selection:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-0.149374	0.360263	-0.415	0.678655	
sexM	0.251252	0.091857	2.735	0.006531	**
famsizeLE3	0.141826	0.091605	1.548	0.122419	
Medu	0.083697	0.052076	1.607	0.108859	
Mjobhealth	0.369433	0.204967	1.802	0.072291	.
Mjobother	-0.061151	0.131829	-0.464	0.643015	
Mjobservices	0.232519	0.148542	1.565	0.118353	
Mjobteacher	-0.201697	0.192928	-1.045	0.296494	
Fjobhealth	0.075793	0.269443	0.281	0.778642	
Fjobother	-0.150223	0.191309	-0.785	0.432816	
Fjobservices	-0.049219	0.198940	-0.247	0.804731	
Fjobteacher	0.328818	0.242108	1.358	0.175240	
studytime	0.146624	0.052661	2.784	0.005638	**
failures	-0.418060	0.064262	-6.506	2.5e-10	***
schoolsupyes	-0.464579	0.136612	-3.401	0.000745	***
famsupyes	-0.212256	0.087231	-2.433	0.015433	*
higheryes	0.367196	0.199698	1.839	0.066748	.
romanticyes	-0.217933	0.089099	-2.446	0.014909	*
freetime	0.064686	0.044093	1.467	0.143208	
goout	-0.119513	0.038582	-3.098	0.002099	**
health	-0.062546	0.029975	-2.087	0.037602	*
absences	0.009724	0.005226	1.861	0.063566	.
failures:schoolsupyes	0.403040	0.167363	2.408	0.016518	*

Table 12: Students' willing to pursue higher degree vs studytime:

Highyes/studytime	1	2	3	4
No	12	8	0	0
Yes	93	190	65	27

Table 13: Students graded A vs Willing to pursue higher degree:

Highyes/studytime	1	2	3	4
No	0	0	0	0
Yes	11	15	9	5

Table 14: Family Support coefficient with only predictor:

```
Call:
polr(formula = factor(level2) ~ famsup, data = stude_three)

Coefficients:
            Value Std. Error t value
famsupyes -0.2117     0.2032  -1.042

Intercepts:
    Value   Std. Error t value
1|2 -0.8430  0.1659    -5.0802
2|3  2.0586  0.2040    10.0920
```

Table 15: Outliers identified by leverage and cook's distance in logistic model:

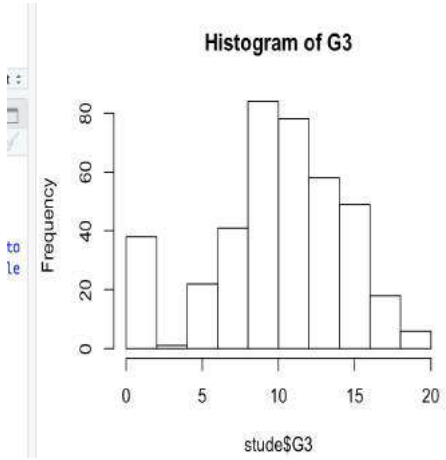
```
> intersect(index,index2)
[1]   3  62 184 277
```

Table 16:Logistic Regression Classification confusion matrix:

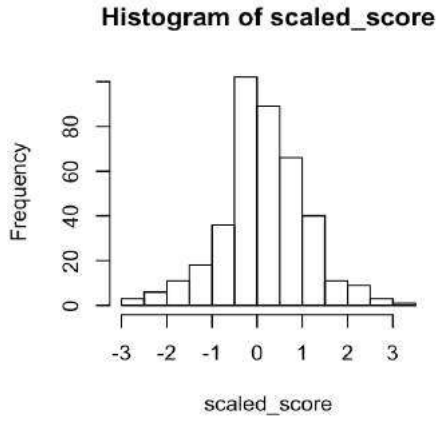
Predic/Level	Fail	Pass
Fail	60	25
Pass	70	240

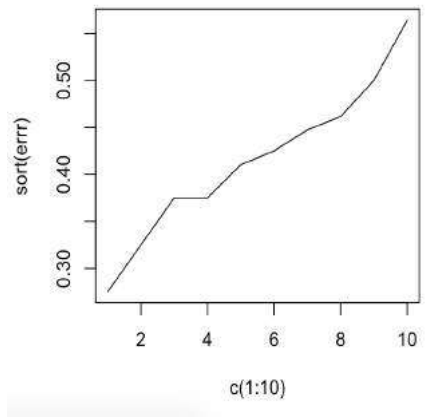
7.5 Appendix E:Graphs

Graph 1:Histogram of G3

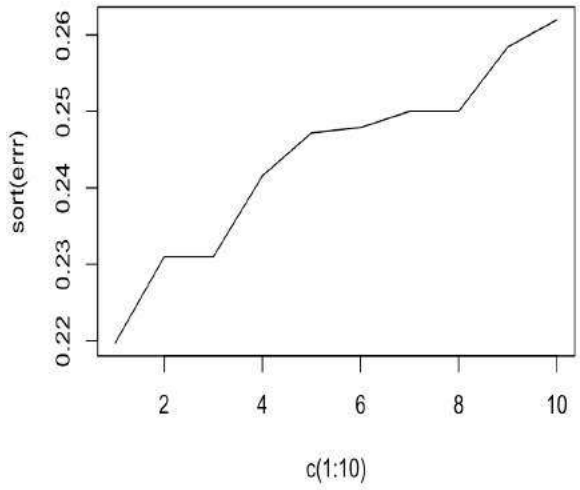


Graph 2:Histogram of Transformed First Principal Component



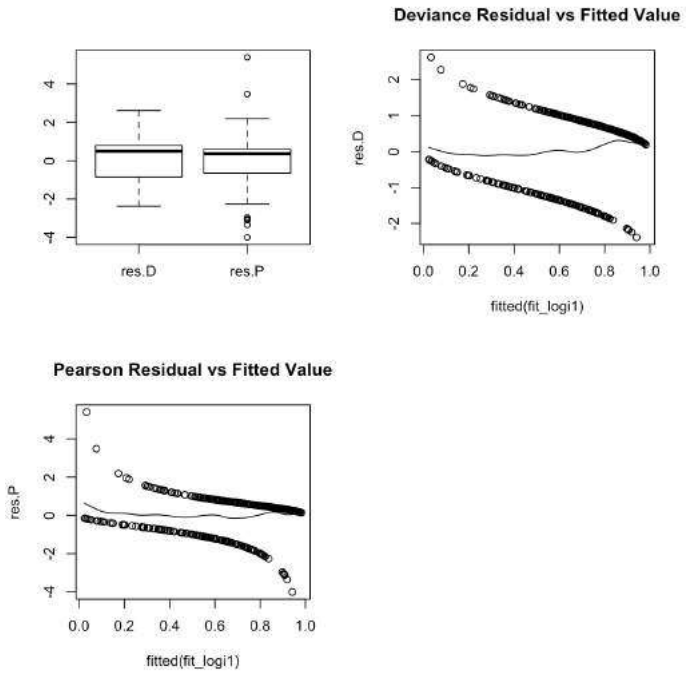


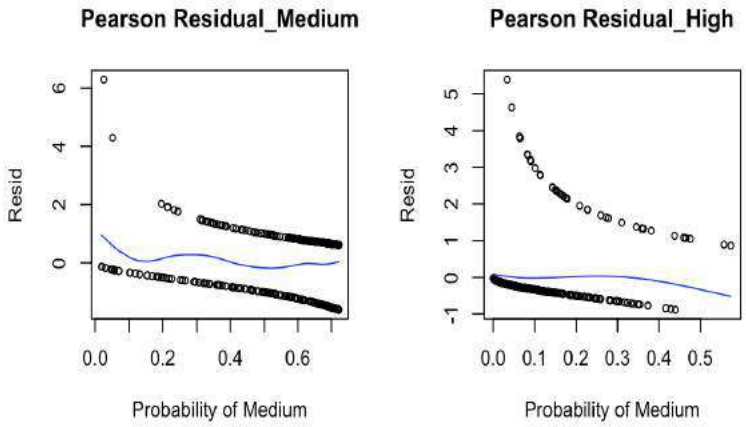
Graph 3:10-fold Cross validation error of proportional odds type model



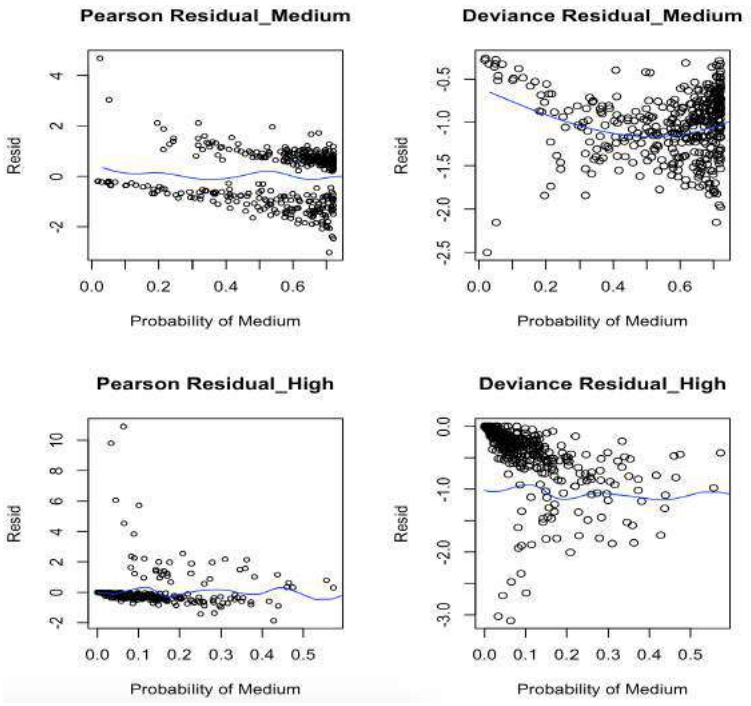
Graph 3(II):10-fold Cross validation error of logistic regression model

Graph 4:Diagnostic Plot for Logistic Regression

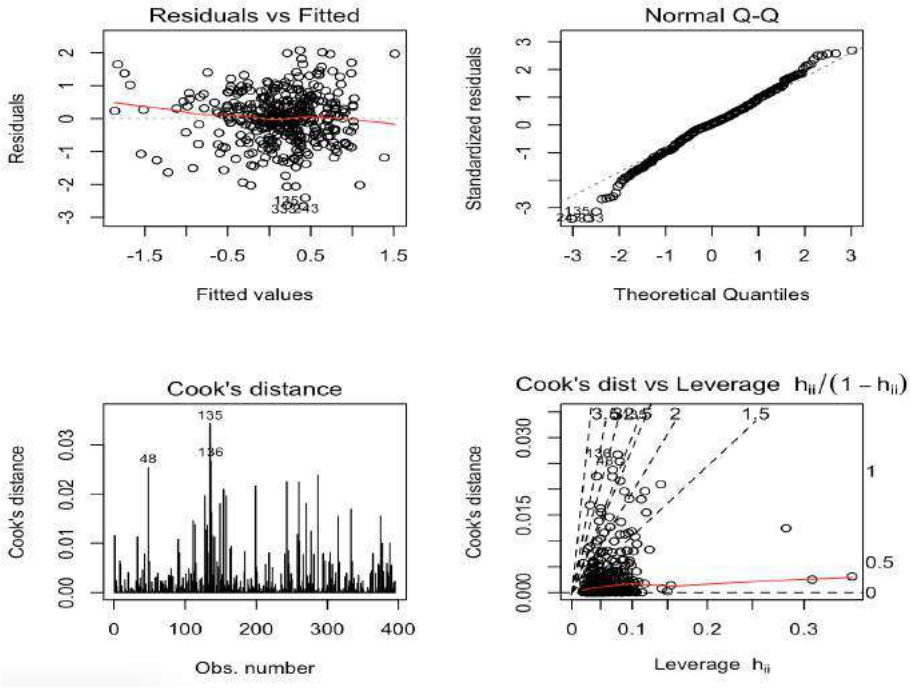




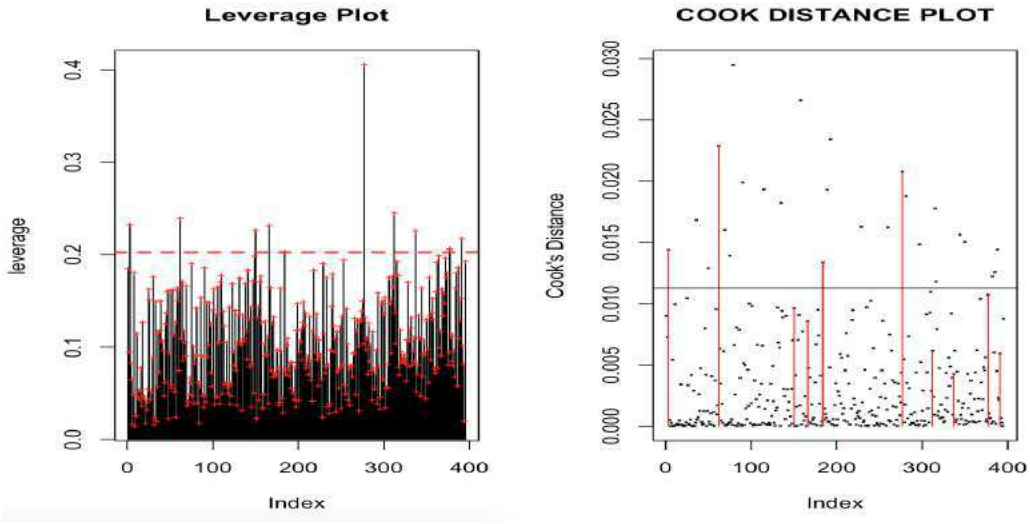
Graph 5:Diagnostic Plot for Merged Proportional Odds Model



Graph 6:Diagnostic Plot for Merged Baseline Odds Model



Graph 7:Diagnostic Plot for transformed linear model



Graph 8:Leverage and Cook Distance Plot in logistic model.

7.6 Appendix D:Acknowledgement

First I want to express my gratefulness to the instructor Prof Hans Georg Mueller who has taught me such a lot during this quarter and his passion in statistics constantly inspiring me. Secondly I would like to express my thankfulness to the TA Yaqing Chen. She has helped me a lot in both the assignments and the project.

7.7 Appendix F:R code

```
library(MASS)
library(mgcv)
library(caret)
library(lawstat)
stu<-read.table("student-mat.csv",sep=";",header=TRUE)
head(stu)
stude<-stu[,-c(31,32)]
head(stude)
n<-nrow(stu)
#G3 is the final grade. Here we aim to explore the functional relationship between the fi
n<-nrow(stu)
#histogram of the students' grade:
hist(stude$G3,main="Histogram of G3")
table(stude$G3)
#Here we may choose various parametrization of the grades:
#(1): We choose the A,B,C,D,F parametrization: 16-20 means(A), 14-15 means good(B), 12-13
#(2):We split the grade into pass(over 10) and fail(below 10)
```

ECS 289N Project Report: Pseudo time reconstruction and evaluation in single cell RNA-seq analysis: Application to NKT cell Dataset

Heqiao Ruan SID:915490857
email:hruan@ucdavis.edu

March 18, 2018

1 Abstract

The recent technological improvement allows researchers to measure transcriptome in level of individual cells. One efficient way to gain biological insights is to quantitatively order the cells according to the transition status and relative expression along the whole process. In recent 5 years more than 10 methods are proposed to construct the trajectory while most of them are only perfectly applicable in limited cases. So new principal technique is yet to develop.

In this project we apply three popular methods in trajectory inference with some modifications in some key steps:TSCAN, Mpath and Monocle to generate the landscape of the gene expression level in the biological process and compare their performance and compare them. What's more, we will try to extract some biological insight and the key regulator controlling NKT cell differentiation in mouse.

2 Introduction

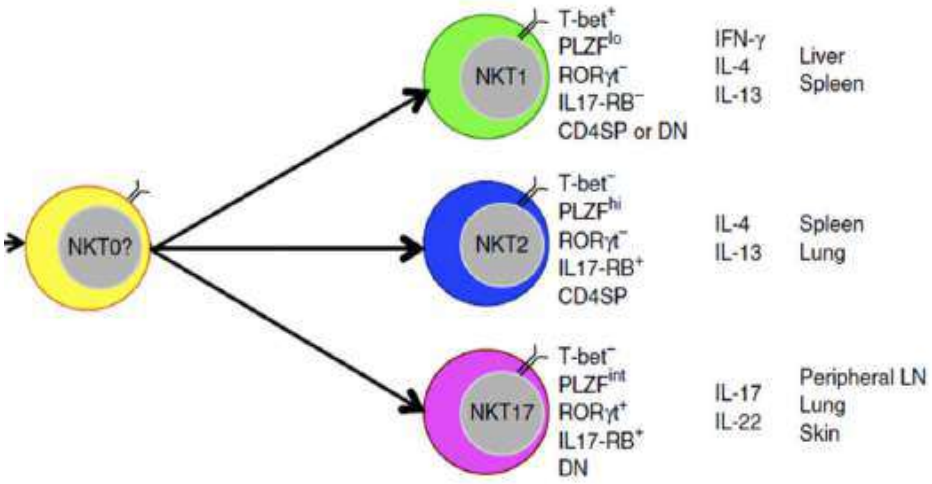
Traditionally in biological experiment we tend to measure on a bulk of. Single cell RNA-seq is a relatively new technology that allows researcher to measure the expression based on every individual cell. It has a couples of advantages compared with the traditional RNA seq techniques which are based on the average of gene expression level. First it can construct a more resolute picture to capture signals conveyed by some single cells which can be easily ignored in bulk RNA-sequence. Secondly, the single-cell RNA-seq is capable of generating a well-rounded picture of the whole gene expression landscape in a highly heterogeneous cell population.
Here we apply various trajectory inference methods including Monocle,TSCAN and Mpath

to re-order the cells according to their relative expression status and try to reconstruct the trajectory reflecting the cell differentiation process.

3 Dataset and Preprocessing

3.1 Dataset

Here we use the NKT cell dataset which was first used in [2]. The dataset contains **22694 rows and 203 columns** where each row stands for a gene while each column represent a single cell. The cells are sequences ranging from 445 to 647 and we just make the assumption that the cells with smaller number are collected prior to the cells with larger number. For example, the cell coded as 550 is collected before the cell coded as 551. We have to point out this is not a very rigorous assumption but in general the cells follow this sequence. **The dataset consists of four types of cells:NKT0,NKT1,NKT2,NKT17. We assume that the four types of cells are collected in time sequence which means NKT0 is the earliest and NKT2 is the latest collected and this is the most fundamental assumption of our downstream analysis**.
The biological path of differentiation of Natural Killer T cell in mouse is shown as below:



We can see that the common sense is NKT0 usually differentiate to NKT1, NKT2, NKT17 cells and here in this dataset the cell collection time is **NKT0− >NKT1− >NKT17− >NKT2**. However, the NKT1,NKT17 and NKT2 may have some transition relation which is yet to be identified.

3.2 Preprocessing

In the real life setting, dataset are not always in a good shape so it requires preprocessing and quality control. Especially in gene inference framework, the dataset is sparse and in general, 70% of all observations are 0. So as our ultimate goal is to explore the trend of gene expression and cell fate process based on them, it becomes particular important for us to filter out some

low or constant expressed genes as well as deleting cells that doesn't express any gene.
First, we perform log transformation on the dataset to deal with extreme value and in practice, we would like to add a pseudo smoother 1 on the original value to deal with 0 observation so that makes it still 0 after transformation. In math formulation, it's: $EXPRESSION = log_2(expressionlevel + 1)$.
After performing log transformation, we are going to perform quality control on cells and genes. Unfortunately, as we know, there's no universal method for us up to now to set the preprocessing threshold corresponding to a specific dataset. Here we just artificially specify the threshold. First we filter out the genes that are expressed in less than 5% of all cells. Secondly we filter out the genes that the coefficient of variation across all cells less than 0.5. The coefficient of variation is $v = \frac{var(Y)}{E[Y]^2}$ which represents the variance divided by the square of mean.
After performing preprocessing, the dataset are of **5138 rows(genes) and 197 columns(cells)**.
Here we have to point out that we just set one threshold for all of the methods we use but we do acknowledge there are some limitations on this setting for in practice, the threshold may makes a big difference.

3.3 Preliminary Analysis

First as we know that the gene naming system is pretty confusing and we don't have access to the specific symbol name of each gene, we have to artificially choose some "gold standard" genes which can also be named as so-called "marker genes"(it maybe a somewhat clumsy name but it does has this kind of property). Here the gold standard genes are identified as the most differential expressed genes across the time.
As we know that the most of the observation is 0, the relationship between the expression level of a single gene and the time is obviously nonlinear we can only explore the functional trend of expression across time. So here we use generalized additive model(GAM) to explore this relationship(Maybe using trend here is more appropriate).
Assuming Y as the expression level of a single gene, x is the collection order. So the generalized additive model is $g(E[Y|X = x]) = s(x, k)$, the s denotes the smoothing function with degree k. Note that normally k is no more than 3 to avoid the curse of dimensionality. Here link function g is either identity(normal additive model) or log link(Poisson additive model). After fitting, we can identify the trend of response variable. We have to point out that it can only generally check the functional relationship instead of fitting the specific regression coefficients for the observation is pretty sparse.
Then we try to identify the degree of differential expression by applying likelihood ration test.Here we use the asymptotic property: $Deviance(Nullmodel) - Deviance(fullmodel) -> \chi_1^2$ where the Nullmodel is fitted as we treat the gene expression level along time as constant. Then we can easily see that the smaller the p value, the more significant of this differential expression effect is. Then we use **holm-Bonferroni** procedure to control the familywise error rate to get a more powerful p value denoting as q value. Then we sort the genes by q value by increasing order to help us find the gold standard genes.
We artificially specify the gold standard genes as the top 100 differential genes

which has the smallest q value. These genes are particular important in our downstream analysis. What's more, the gene id of the first 20 differential expressed genes are shown in Appendix B.

4 Main Methods

4.1 TSCAN and Modifications

TSCAN(also named as Tool for Single Cell Analysis) is originally brought out by Zhicheng and Hongkai in JHU Department of Biostat in 2016 which is based on connecting the clusters after performing dimension reduction. It is a unsupervised learning technique.
This methods can be summarized into three steps. Firstly, it clusters the cells with similarly expression profiles. Then the minimal spanning tree is constructed to connect all the cluster centers as well as specifying the order of the cluster centers. Finally cells are projected to the backbone of the tree so that we can determine their pseudotime and order.
One important feature of **TSCAN** is that it cluster the similar expressed genes together to alleviate the drop out event by using the average expression profile of the clustered genes. Some research have demonstrated that clustering genes before conducting the main algorithm can improve the performance and we will validate this in the downstream analysis(potentially alleviate the dropout effect).
In the original paper, Zhicheng and Hongkai applied **PCA** in dimension reduction step and used mclust based on mixed gaussian assumptions in clustering step. PCA is among the most popular dimension reduction techniques to extract a few number of features with most of the variability of the data and in the original paper they use the LS technique to find the optimal number of principal components extracted. Mclust is a model based clustering algorithm which is optimized by EM algorithm by assuming that very cell follows a multivariate normal distribution.
Then after clustering the cells we are going to construct the trajectory and project each cell onto the trajectory. We start by constructing the minimal spanning tree which has the smallest sum of length of the edges connecting each vertex. Then while the trajectory may often be branching we find the longest path of the minimal spanning tree which has the largest numbers of clusters with the largest total numbers of cells and then in terms of choosing a origin, **we use the gene expression profile of the second marker gene(ENSMUSG00000001025.8) as it is minimal expressed at NKT0 cells compared to other three types of cells**. Then we first exhaust the main path and then add the branches onto the tree iteratively. We want to point out that usually the numbers of clusters won't excess 6 or 7 which greatly reduces the variability and complexity of the tree space comparing with Monocle and alleviate the risk of being contaminated by the various source of noise arisen from the previous analysis and even in the biological experiment. Then after ordering the clusters we project the cells onto the edges of the tree and for cell A in C_m we project it to $C_m - C_{m+1}$ is $d(A, center(C_{m+1})) < d(A, center(C_{m-1}))$ and project it to C_{m-1} vice versa. Then the cell

ordering are determined following these steps, first for cells in the same cluster projecting onto the same edge, their order is determined by the projected values on the edge(for cells in C_m,$C_{m-1} - C_m$ is negative while $C_m - C_{m+1}$ is positive). Then the order of cells in each cluster is determined by the order of edges. Finally we use the order of clusters to order them together. We can see that TSCAN greatly reduces the complexity and variability of the minimal spanning tree. **However, the good performance of TSCAN is pretty highly depend on the appropriateness of MoG clustering optimized by the EM algorithm** and it has huge potential in terms of developing even more delicate clustering techniques.
Here in actual implementation, we make several modifications and compare the new algorithm with the original one. First as some research indicate that the dimension reduction technique diffusion map which based on markovian transition matrix may performs perfect in some cases. Secondly, besides mclust, kmeans has been demonstrated a particularly efficient algorithm to combine nearby data points together(which means similarly expressed cells in gene inference our framework). What's more, we artificially set 4 clusters in mclust to perform the original version of algorithm. So we try another three techniques **Nonclu-TSCAN**,**Kmeans-TSCAN**,**Diffusionmap-TSCAN**.
The **Kmeans-TSCAN** starts from clustering similarly expressed genes following by PCA to reduce the dimension. Then we perform kmeans clustering and set the clusters as 4 after that we use TSCAN to project the cells onto the path.
The **Diffusionmap-TSCAN** starts from clustering similarly expressed genes following by dimension reduction technique Diffusion Map. Other steps are the same as the original TSCAN technique. The matrix W is given by $w_{ij} = \frac{exp(-\frac{||X_i - X_j||^2}{2*\sigma^2})}{\sum_j exp(-\frac{||X_i - X_j||^2}{2*\sigma^2})}$ and the t th step diffusion distance is given by $D_{ij}^t = \sum_k \lambda_k^{2t}(\phi^k(X_i) - \phi^k(X_j))^2) = ||\Phi(X_i) - \Phi(X_j)||^2$ where λ_k denotes the k th eigenvalue and $\Phi_k(X)$ is the k th eigenvector of the markov transition matrix. Then we can extract the first k diffusion coordinates $\P = (\lambda_1^t\phi_1(x), ..., \lambda_d^t\phi_d(x))^T$ which contain most of the information in the data. There are two main issue in the implementation,**the first is to find the tuning variance σ^2 while the second is to determine the number of diffusion components we extract**.
For the first problem, indeed there will be a range of parameter σ which the markovian transition matrix defines an ergodic diffusion process on the data as a connected graph and still the diffusion distances between the cells are informative. Here we use the **median of distance to the k th nearest neighbor of every cell** which helps us try best to maintain the local property. It indeed reaches a cost-accuracy tradeoff for the more neighbors we select, the more noise it will contain. Empirically we set **k as 10% of the total number of cells**. For the second problem, we use the rank-based criteria proposed in [12] by John A.Lee which measure the degree of consistency local property between the original data and embedding space. They use the **quality measure** $Q_{NX}(k) = \frac{1}{NK}\sum_{i=1}^{N}|v_i^k \cap n_i^k|$ **and the adjusted quality measure** $R_{NX}(k) = \frac{(N-1)Q_{NX}(k)-k}{N-1-k}$**.** We try the plot the $R_{NX}(k), Q_{NX}(k)$ versus k(search from 2 to 30 empirically) and observe them to get the best k based on the cost-accuracy tradeoff.**We implement Diffusion Map algorithm by following the above description.**
Another alternative technique here to do the dimension reduction is the **t-SNE TSCAN**.

t-SNE(also named as t-Stochastic Neighbor Embedding) is a popular dimension reduction method which has been demonstrated particular useful in visualization of high dimensional dataset. Here we aim to map the d dimension data $(X_1, ..., X_n)$ into 2 dimension in which maintain its local similarity. The similarity matrix is constructed by $p_{ij} = \frac{p_{j|i}+p_{i|j}}{2N}$ and $p_{j|i}$ is calculated by $p_{j|i} = \frac{exp(-||X_i-X_j||^2/2\sigma_i^2)}{\sum_{k\neq i} exp(-||X_i-X_j||^2/2\sigma_i^2)}$. It depicts the similarity between the two points in high dimension. We assume the transformed data points are $Y_1, ..., Y_n$ and $q_{ij} = \frac{\frac{1}{1+||Y_i-Y_j||^2}}{\sum_{k\neq i}\frac{1}{1+||Y_i-Y_k||^2}}$. Here the loss function is given by Kullback-Leibler Divergence $DL(P||Q) = \sum_{i\neq j} p_{ij} log(\frac{p_{ij}}{q_{ij}})$ which depicts the dispersion of one distribution from the other. Then we use the MoG(Mixture of Gaussian) clustering methods to cluster the similar expressed cells and use the same technique as described in TSCAN to order the cells. Then we can also apply various performance measures to evaluate the efficiency of ordering and trajectory.
Here we have to point out that actually the perplexity makes a big difference in the performance. The number of perplexity means the information they gain from the nearest neighbors so too small perplexity will bring out severe information loss while too large perplexity will contain too much noise that can't be identified by us. Typically in implementation, we use 30 and here we try three different perplexity and compare their performance.
TSCAN has its own advantage in greatly reducing the number of vertices in the minimal spanning tree so that it will be more robust than Monocle which will be demonstrated later. What's more, we use some prior information of marker gene expression and it's not completely unsupervised learning technique.

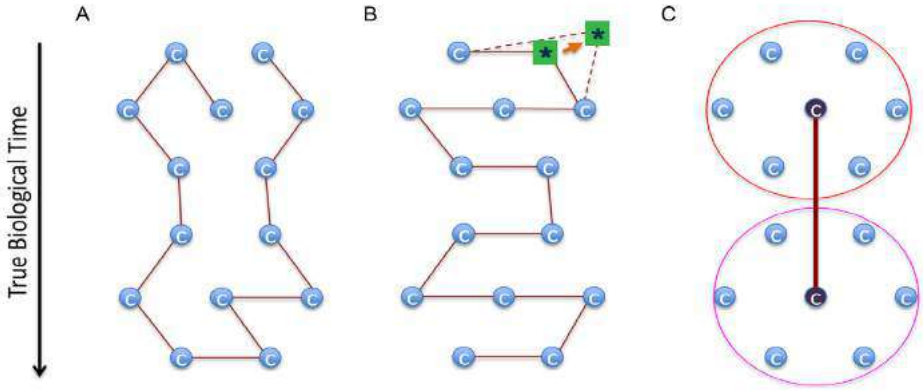
4.2 Mpath

Mpath is a relative recent algorithm that can help us construct a branching trajectory by choosing the so-called landmark cluster centers and then find the minimal spanning tree connecting them. Its idea is from the empirical assumption that the likelihood of two clusters of cells are higher if the number "between" this two clusters are large. So we construct the transition network based on this assumption.
Mpath starts by hierarchically clustering the cells after preprocessing in **ward** distance which is given by $\Delta(A, B) = \frac{n_A n_B}{n_A+n_B}||m_A - m_B||^2$ where m_A and m_B are cluster center of A,B respectively. Then the problem become how to choose the number of cut on the dendrogram which is equivalent to choose the number of clusters. Here we use two measure to perform the quality control and we denote the clusters passing this two as landmark clusters. First we use the size of 5% of total number of sizes as the threshold and the cluster sizes larger than this passing the quality control. Secondly about the purity measure, we use the shannon diversity $H = -\sum_{i=1}^{n} p_i ln(p_i)$ where p_i is the proportion of category i in the whole sample and the threshold is 0.6. **We have to point out that smaller diversity means the cluster is 'pure' which means most of cells in this cluster are from the same type which is pretty important to construct the cell fate branching network.** Then we search from 4 to 20 and plot the number of clusters versus the total number of clusters to determine the optimal number of clusters.

After that we construct the graph with directed weights on them where each vertex is a cluster center. Assume two vertices A,B The weights of $A->B$ represents the number of cells which its nearest neighbor is B and the second nearest neighbor is A while the weights of $B->A$ is calculated by the number of cells which its nearest cluster center is A and the second nearest cluster center is B. Then we use the weight to estimate the likelihood of cells transitioning between stage A and stage B.As a result, we build a complete graph(edges with weight 0 automatically) where. Then we trim the network so that the remaining tree(no circle) has the highest sum of weights(highest likelihood of state transitioning).
Finally we project every cell onto the paths, we assign cells to the cluster which center is nearest to them and project them to the edge between the nearest neighbor and second nearest neighbor. **Note that we requires some prior information about the cell collection time. What's more, we have to point out that in Mpath algorithm, we can only order the cells along several branches and observe the potential biological process along these path so it will be pretty difficult to measure the POS score and the Robustness. We measure the performance of Mpath by counting the gold standard genes among the top differential expressed genes.**
Here we implement the algorithm Mpath by ourselves and the R code for the whole algorithm and the downstream simulation can be seen in the supplementary materials.

4.3 Monocle

Monocle is a insightful method proposed by Trapnell lab which help us identify the branches based on constructing the minimal spanning tree based on all cells.
It can be easily demonstrated that Monocle is not as robust as TSCAN which is based on constructing minimal spanning tree between several clusters of cells.(Pretty sensitive to random noise which will be validated in the downstream analysis)



From the above plot we can see that both A and B are possible solution of minimal spanning tree however obviously solution B is more reasonable for it is more consistent to the true biological time.What's more, the cell ordering constructed based on every cell level is not as robust as that constructed by TSCAN given a small pertubation shown as the green node in solution B.
The Monocle algorithm first performs **ICA**(Independent Component Analysis) which aims to

extract the independent signals from the data to reduce the dimension to 2 or 3. However, ICA may be **pretty slow** in real implementation and a R package called **fastICA** demonstrate this. Then we construct the minimal spanning tree based on the reduced data points in two or three dimensions. **Although there is an alternative to reduce the time cost by reducing the number of genes using a differential expression test but it will introduce significant bias in terms of heterogeneities within subpopulations and deviation from the existing population groupings. So we won't use this trick.**
Here we replace ICA by tSNE to perform the dimension reduction the reason can be found in section(5.2.2) and compare it with that conducted by ICA.
Then we find the longest connected path in the tree after which we assign every cell to its nearest neighbor. We implement it by constructing the PQ tree and search the tree recursively along the main path.
Here we use the existing R package monocle(also needs some slight modifications in real implementation) to do this analysis.

4.4 Performance Measurement

Here we apply various techniques in measuring the efficiency of ordering and trajectory.
The first and foremost evaluation approach is to measure the to detect the differential expressed genes across the cell ordering constructed by pseudo-time. Then we can either **measure the numbers of gold standard genes detected before** in the top differential expressed genes along the pseudo-time axis(degree of deviation) or calculate **the mean rank of the gold standard genes** in the top differential expressed genes along the pseudo-time axis.
Alternatively we use the **POS score** to the efficiency of ordering. Assume the cells are collected in v time points $T_1, ..., T_v$. The POS score is given by $POS_\pi = \sum_i \sum_j g(\pi, i, j)$. If two cells are originally collected at the same time,$g(\pi, i, j) = 0$, otherwise, if the ith cell is collected from time point T_u and the jth cell is collected from time point T_v,then $g(\pi, i, j) = \frac{u-v}{D_\pi}$. The scaling factor D_π is chosen to restrict the POS score between -1 and 1. So $POS_\pi = 1$ means the order of cells produced by pseudo-time reconstruction perfectly matches the order of the cell collection time and $POS_\pi = -1$ means the cell ordering by the pseudo-time reconstruction is exactly the reverse order comparing to the original cell collection time. So in this way does POS score measure the efficiency of cell ordering.**We have to point out here that using POS score to evaluate of pseudo-time reconstruction is based on a fairly strong assumption that the cell collection time indeed reflect the true biological process.For example, we assume that the cell collected later is in a later stage of the true biological process(cell differentiation or cell apoptosis).** . We will test this on all methods.
Another insightful evaluation is the robustness of the cell ordering. We artificially add some perturbation onto the single cell RNA-seq dataset. We have two methods to perturb the dataset which denoted as cell-level and gene expression level. **For cell-level perturbation, we subsample 75%,90% and 95% of the whole sample of cells while for gene expression level perturbation, we retain all of the cells but now we add random simulated**

noise onto the original expression level of every cell. In [1],for each gene,they add the random noise as $(Y - E[Y]) * \zeta$ where ζ is chosen as 5%,10% and 25%. **However,here we use another perturbation method. We add normal distributed random noise $\epsilon_i \sim N(0, k\sigma_i^2)$ where σ_i^2 is the variance of the expression level and k can either be 5%,10% and 25%. To deal with the negative observation, we truncate it by 0. We use this because we use mclust to cluster the cells which is based on optimize of the mixture gaussian distribution so we add the normal noise.** We have to point out that in [1] they repeat each procedure 100 times but we only implement 15 times for each procedure considering the time limitation.
We have to point out that POS score and Robustness themselves are not sufficient to claim a good trajectory reconstruction, it should have some biological meaning.

5 Results

5.1 Performance Measure

Here totally we use 8 methods to do the trajectory inference and the measurements are shown in the below table.

Methods/Measure	markergeneexp	POS	Robustness	meanrank	numofmarkergene
t-SNE TSCAN	Y	Y	Y	Y	Y
Nonclu TSCAN	Y	Y	Y	Y	Y
Kmeans TSCAN	Y	Y	Y	Y	Y
Diff-map TSCAN	Y	Y	Y	Y	Y
ordinary TSCAN	Y	Y	Y	Y	Y
Monocle	Y	Y	N	Y	Y
tSNE-Monocle	Y	Y	N	Y	Y
Mpath	P	N	N	P	P

Y denotes yes and P denotes partly while N denotes not.
Here for t-SNE TSCAN we try three perplexities(30,40,50 and 30 is the default parameter in the R implementation) and choose the one with the highest POS score and the smallest mean rank of gold standard genes in the top differential expressed genes.
The choice of parameter in t-SNE is shown as below:

Method/Measure	POS score	meanrank of gold standard genes
t-SNE 30	0.5013164	532.55
t-SNE 40	0.546984	352.2
t-SNE 50	0.4664693	430.06

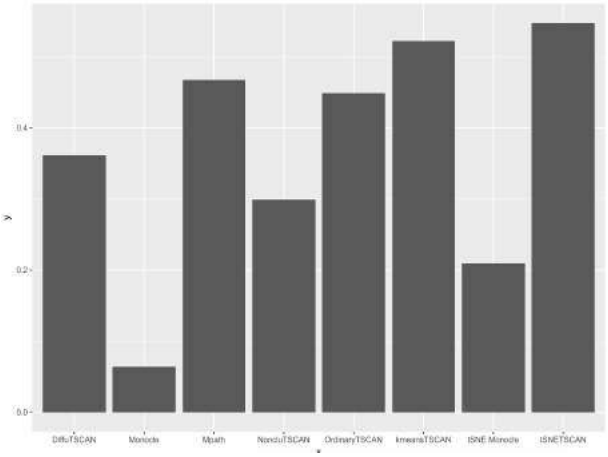
So we choose 40 as the perplexity parameter in t-SNE which indeed reaches a noise-information trade-off.

5.1.1 Marker Gene Expression

First we would like to check whether the expression trend of our artificially specified marker genes is consistent before and after cell-ordering and we only check the first 8 most differentially expressed genes. The graphs are shown as **Graph1-Graph7 in Appendix A**. We can see that in terms of consistency in trend with the original order, noncluTSCAN, kmeans TSCAN and Monocle obviously doesn't do well(the trend of the fitted values is greatly dispersed from the original one) and tSNE Monocle performs much better than the ordinary Monocle algorithm while still not that satisfactory. Diffusion map TSCAN performs decently in the first 2 marker gens while fail to maintain the trend in marker gene 3,5,6,7. Then tSNE TSCAN performs as well as the ordinary TSCAN which may not be treated as a large surprise.

5.1.2 POS Score

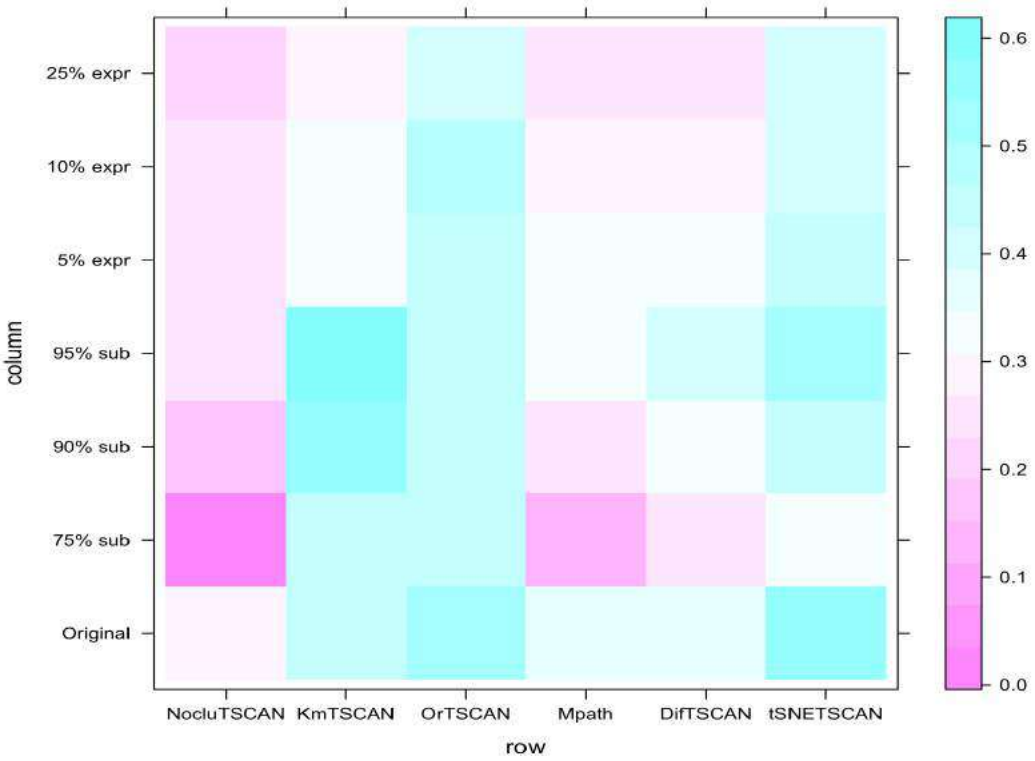
Here we compare the efficiency of ordering of the 6 trajectory inference techniques and the result is shown as below:



We can see that among the various modifications of TSCAN methods, the ordinary TSCAN(PCA+ clustering+mclust) performs the best and the diffusionmap and t-SNE also perform decently. However, Monocle is again. What's more, the ordinary Monocle doesn't perform well and even we replace ICA by tSNE in Monocle, although some improvement observed,the performance is still unsatisfactory.

5.1.3 Robustness

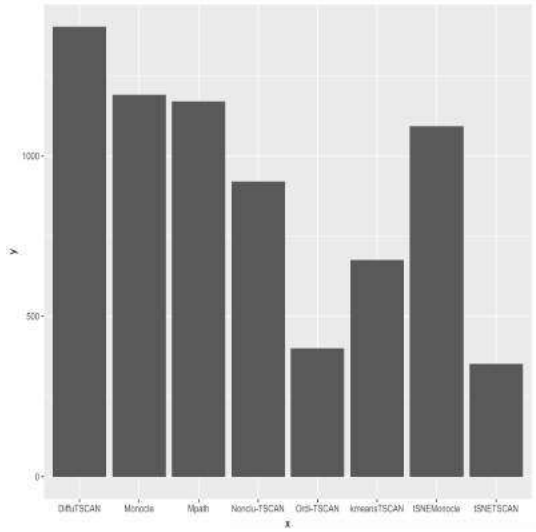
Here we compare **Ordinary TSCAN, Nonclustering TSCAN, Diffusionmap TSCAN, Kmeans TSCAN**. As we can see from the previous part, **Monocle and tSNE-Monocle** doesn't perform well in terms of POS score so we won't test its robustness **(Indeed it cost too much time to perform fastICA function in R for hundreds of times..)**. The result shown in levelplot is shown as below:



Here we can see that in terms of robustness, among the several modifications of TSCAN algorithm, **OrdinaryTSCAN and t-SNE TSCAN** does better and **kmeans-TSCAN** performs slightly worse than the previous two methods. Then we observe **Non clustering TSCAN** doesn't do well in this case. It's not surprising that non-clustering TSCAN not perform well due to the prevalent dropout effect in all kinds of biological experiments. For diffusion map TSCAN, it obviously down-performs compared with t-SNE. We guess it may arise from the choice of the ϵ and the number of diffusion components extracted and here we only use the threshold first developed in another field. Then to compare t-SNE TSCAN and the ordinary TSCAN, we conclude that the tSNE TSCAN is not that robustness comparing with the ordinary TSCAN maybe it's because we use the same perplexity for subsampling and random-noise perturbation and for each case, a specific optimum perplexity requires to be found artificially.
We have to point out a limitation because we only repeat 8-15 times(for t-SNE we only do 6 times each, for other methods we repeat for 15 times and get their average value).

5.1.4 Mean Rank of Gold Standard Genes

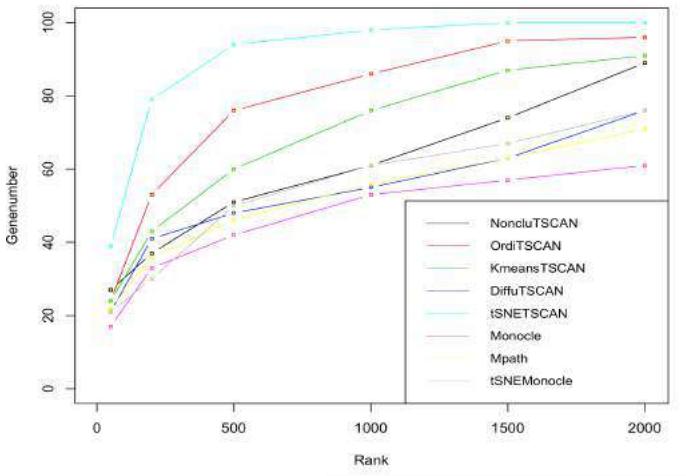
Here we compare the mean rank of the 100 marker genes across the seven trajectory inference techniques to compare their innate quality. The result is shown as below:



We can see that t-SNE are performs even a little bit better than ordinary TSCAN, however, the diffusionmap TSCAN doesn't perform well in terms of innate consistency of the cell ordering. We can also see that clustering similar expressed genes help improve the performance (ordinary TSCAN does better than non-clustering TSCAN) What's more,Monocle maintain too much noise on the individual cell level and fail to get a innate consistent ordering and for tSNE-Monocle, we still fail to observe a satisfactory performance. For Mpath, it may not fair to claim it doesn't perform well just based on one type of ordering for the branching nature of its reconstructed trajectory. We will clarify the result of Mpath and evaluate its performance in the trajectory visualization part.

5.1.5 Number of Gold Standard Genes in the top Differential Expressed Genes

Here we compare the number of gold standard genes(our pre-specified top 100 differential expressed genes as marker genes) in the top 50,100,500,1000,1500,2000 differential expressed genes to our reconstructed ordering to check the innate quality of the ordering.The result is shown as below:



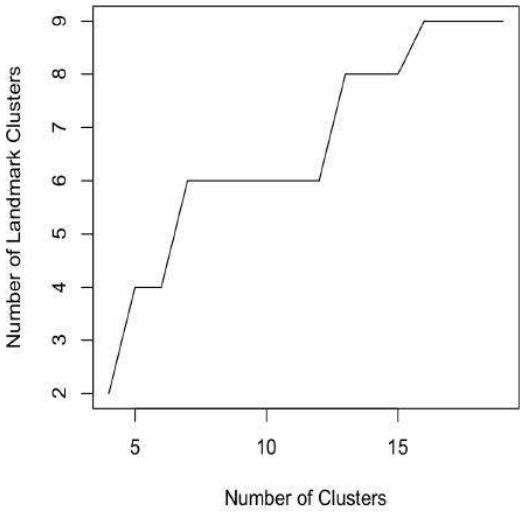
We can see that monocle doesn't perform well here because of the unidentifiable noise in constructing the trajectory based on each individual cell level and even after replacing ICA by tSNE, the performance still unsatisfactory. What's more we can see that nonclustering TSCAN and kmeans TSCAN performs slightly worse than ordinary TSCAN which is not surprising. Then what is amazing is that **tSNE-TSCAN** performs better than the ordinary TSCAN method with respect to this measure. What's more, Monocel doesn't do well and t-SNE Monocle does much better than the ordinary one although still not satisfactory. Diffusion Map TSCAN also doesn't do pretty well.

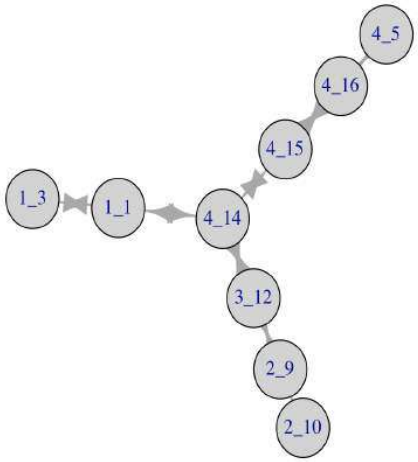
5.2 Visualization

Here we visualize the result of methods with good performance**(tSNE-TSCAN, Ordinary-TSCAN)** as well as Mpath.

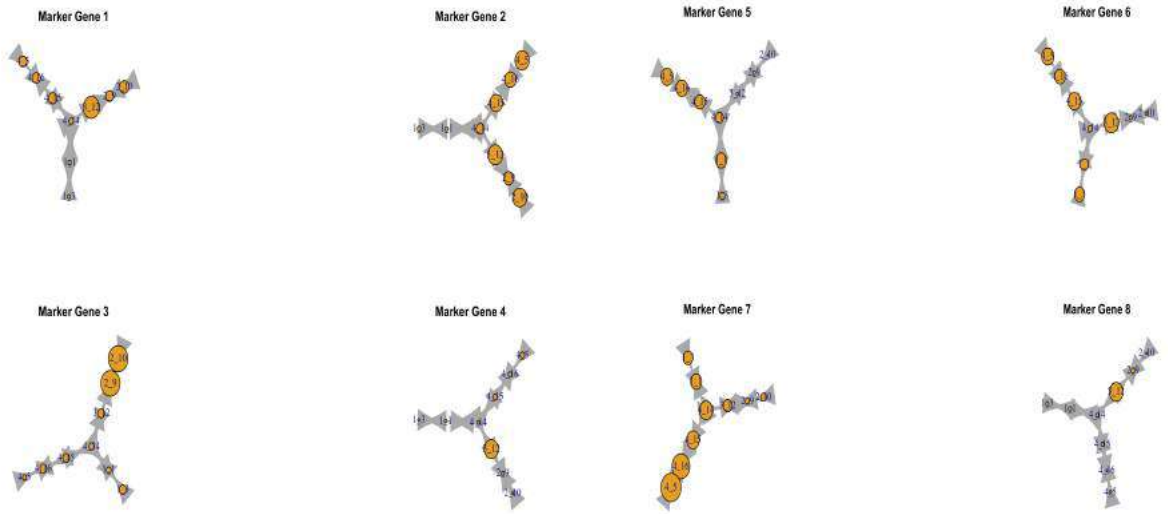
5.2.1 Mpath

Here in Mpath algorithm, most of the previous performance measure may not that adequate because in real implementation, we may need to determine the direct by ourselves because the nature of the reconstructed trajectory. Here after implementation we construct the branched path and the plot of number of landmark clusters versus the total number of clusters is shown as below:





Here the optimal number of clusters cut from the hierarchical clustering result is 16 and the branched trajectory indicate that a subpopulation of **NKT2** cell is important in controlling the further differentiation and we explore the trend of some of our pre-specified marker genes. Like **section 5.1.1** we visualize with respect to the first 8 differential expressed genes.



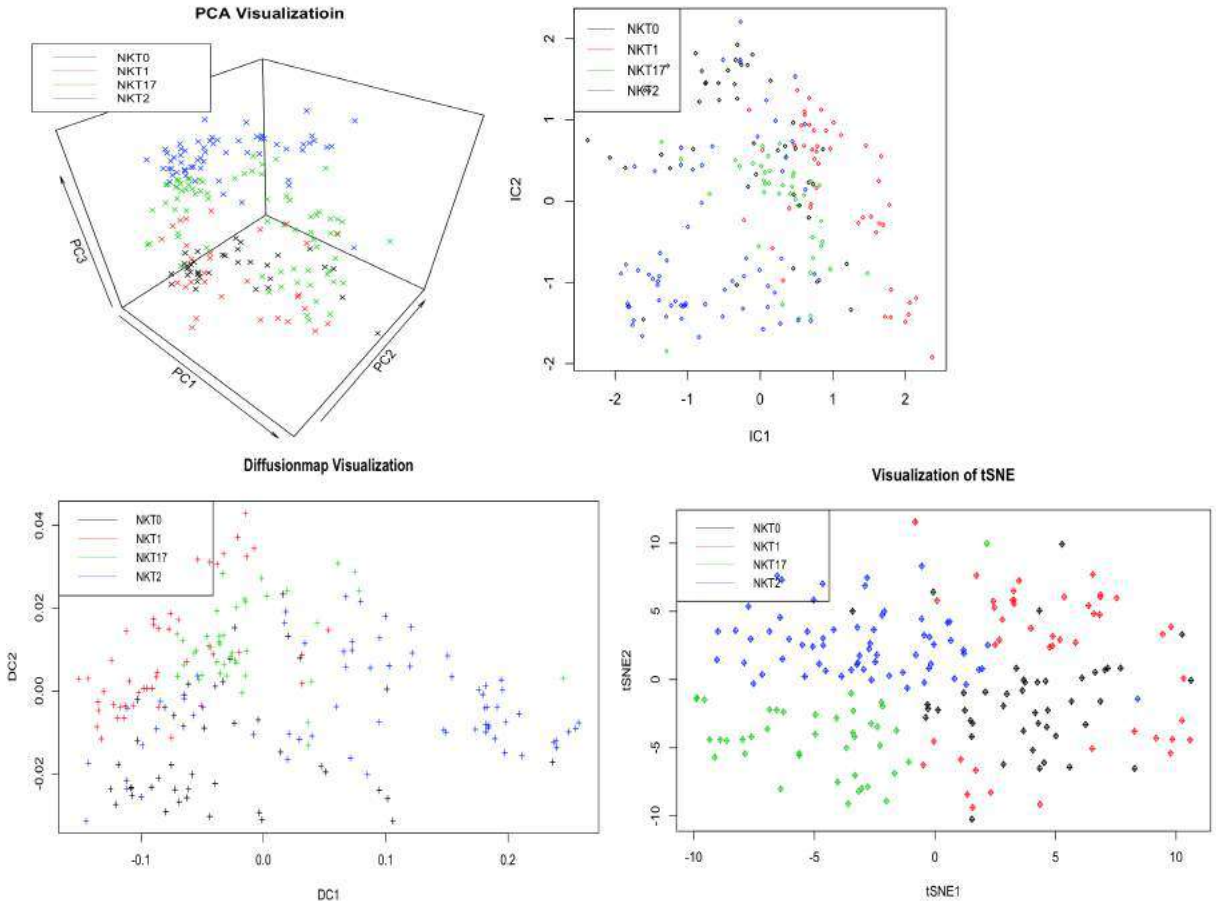
Here for marker gene 1**(ENSMUSG00000001020.8)**, the NKT 17 cell has a much higher expression profile than other types of cells; For marker gene 2**(ENSMUSG00000001025.8)**, it is higher expressed in NKT 1,2,17 cells than NKT0 cells; For marker gene 3**(ENSMUSG00000004612.9)**, it is up-expressed in NKT1 cell while for marker gene 4,8**(ENSMUSG00000014453.3, ENSMUSG00000023367.14)**, it is up-expressed in NKT 17 cell. Then for marker gene 5 and 7**(ENSMUSG00000 015314.10 and ENSMUSG00000023004.8)**, they are upregulated in NKT2 cells. For marker gene 6**(ENSMUSG00000021728.7)**, we can see that it is up-expressed in NKT 17 and NKT 2 cells.

For one branch 3-12,2-9,2-10, we can see that NKT 1 and NKT 17 cells have some transition relation. What's more, we are particularly interested to the landmark cluster **4-14**, a subpopulation of NKT 2 cells, which is the beginning of two branches and we may want to do more biological experiments to see it. NKT1 and NKT 17 may have some transitional relations which requires validation in the future. One more thing we want to point out is that in trimming the network, we ignore some of the minor interactions between different states which can also biological meaningful.

An limitation of Mpath is that when we see **Graph 8,Appendix A**,in trimming the constructed network based on the landmark clusters,we ignore the edge between cluster 3-12 and 4-15 which has 16 cells between them which may have a biological meaning. So we may loss information about the biological process we reconstruct.

5.2.2 Dimension Reduction

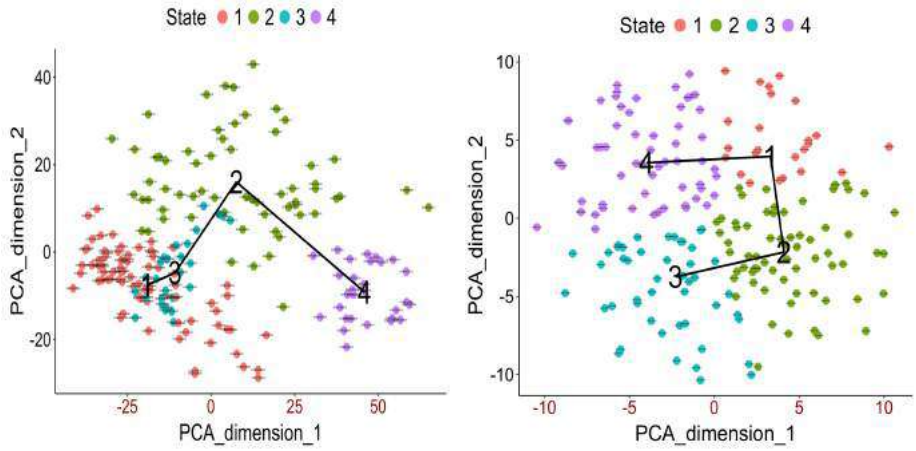
First we would like to see the result of various dimension reduction techniques(ICA,PCA,t-SNE,Diffusion Map) and they are shown as below. The right of the first row is ICA.



We can see that **PCA** and **t-SNE** performs much better than **Diffusionmap** and **ICA** here in separating different types of cells. We find that NKT 2 cell and NKT17 cell is identifiable from the plot. What's more, we find that t-SNE performs even slightly better than the ordinary PCA which is consistent to the previous result in various measurement of performance. To get a more satisfactory result of Diffusion, we may need to make some modifications to the algorithms like in [7]. Here ICA also doesn't perform well and what's more, this algorithm is too **SLOW**. **So we consider replacing ICA by t-SNE in monocle.** So here we can see that t-SNE and PCA may performs better than ICA in single cell sequencing dataset.
From the visualization we can see that the tSNE can separate NKT17,NKT2 and NKT0 while it may still fail to identify the cell population NKT 1(NKT 1 cell population pattern seems to be vague which we may be particularly interested in in the future research).

5.2.3 TSCAN

Here as demonstrated from **section 5.1**, ordinary TSCAN and t-SNE TSCAN performs among the best in various methods. So we will only visualize these two methods while the left is ordinary TSCAN and the right is tSNE-TSCAN:



Actually there are slight difference between these two trajectories and we find that indeed, the trajectory constructed by the tSNE-TSCAN is consistent to the biological nature that **the differentiation should start from NKT0 cell and may probably be a branching path** compared with that constructed via ordinary TSCAN. So here we can see that tSNE TSCAN pretty well to identify the biological path. For the ordinary TSCAN, we can identify the differentiation path $1- > 3- > 2$, however the path $2- > 4$ doesn't seem to have biological meaning(their may be interaction between the process from NKT0 to NKT1 and the process from NKT1 to NKT2). So here we can see that robustness itself doesn't ensure a good performance if from the minimal spanning tree we can't get a reasonable biological explanation we still can't claim a good result.

6 Conclusion and Discussion

We compare 8 methods of trajectory inference in our analysis and compare their performances in terms of various measurements.

First we can see that the fact that kmeans-TSCAN performs slightly worse than the ordinary TSCAN is not surprising because the kmeans can be seen as a restriction of the MoG clustering method. The euclidean distance used in calculating distance in kmeans is indeed the same as that in the log likelihood of the MoG optimized function.

What's more, we find that nonclustering TSCAN doesn't perform as well as the other methods. So clustering similarly expressed genes before analysis is really helpful to reduce the dropout effect.

Then we can see that diffusion map doesn't perform as well as the ordinary TSCAN. However, some research such as [7] has demonstrated that diffusion map may be a good choice in single cell sequencing data analysis because its ability in handling the density heterogeneity and decent robustness to noise. As [7] mentioned, they made various modifications on the original algorithm such as density normalization and consideration for missing value and uncertain observations. So we only use a developed threshold before to test and it's unfair to claim that diffusion map TSCAN is not good.

For Monocle, we can claim that it may not be a good choice compared with TSCAN because

it construct the MST based on every individual cell which may potentially introduce a lot of noise that can't be identified by researchers. So the ordering may be easily contaminated and become pretty unstable and vague. What's more, ICA is not a optimal choice of dimension reduction compared with the other dimension reduction techniques in the analysis because it does not scale well with an increasing number of genes[5]. When we replace ICA by t-SNE, Monocle's performance improve slightly while still unsatisfactory.
What's more, for t-SNE TSCAN and ordinary TSCAN, we can see that t-SNE TSCAN can perform slightly better than the ordinary TSCAN by a careful choice of the tuning parameter **perplexity** but it has lots of limitations in real application. The key is that there's no optimal technique to decide a perplexity(can be interpreted as the number of nearest neighbors).In real application, we may still prefer the ordinary TSCAN technique as we doesn't need to adjust the parameter artificially for each case, each loop. However, some research([8]) has demonstrated that t-SNE will be a good technique in single cell sequencing data analysis and it can also delicately handle density heterogeneity and random noise.
Another important issue we have to point out is that the mclust doesn't perform well here in the ordinary TSCAN and t-SNE TSCAN performs much better after mclust. We can see that the TSCAN has its own limitation in terms of difficultie in visualization.
What's more, Mpath is a really insightful method which is based on the idea that the likelihood of state is proportional to the number of cells between the two landmark clusters and actually we get some useful biological insight. What's more, the use of ward distance which depicts the cost of merging clusters in hierarchical clustering is also a highlight. For future extensions, we can extend the algorithm to identify the multi start developmental processes for more heterogeneous single-cell sequencing dataset.
We have found that there are probability that a subpopulation of NKT2 cells can potentially differentiate to NKT 17 cells and it may also control whether to differentiate(the beginning point of two branches).
We have to point out that our analysis actually have some limitations. First we artificially ignore the variability in isoform which can potentially makes somewhat a difference but in our dataset, each gene only contain one isoform.
The second limitation of our analysis lies in the simulation process, we only repeat 5 or 10 times in various measurements of robustness due to the time limitation so it may not be that accurate compared with that in [1].
The third limitation is that we have made a strong assumption prior to analysis that the cells are collected along the time sequence and the cells collected at the same time are only one type. For this dataset, however, it may not always the case because the differentiation mechanics of natural killer cells in mouse are yet to clarified.
Another limitation arises from the awful gene naming system which prevent us from getting the gene id from the gene symbol. Then we can't use any information from [2] which explored the various property of this dataset.So in our analysis we have to personally specify the marker genes by using the generalized additive model which is indeed pretty limited.

References

[1] Zhicheng Ji and Hongkai Ji. *TSCAN:Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis*. Nucleic Acids Research, May, 2016

[2] Isaac Engel, Gregory Seumois et.al *Innate-like functions of natural killer T cells subsets result from highly divergent gene programs.* Nat Immunol, June 2016.

[3] Cole Trapnell, Davide Cacchiarelli et.al *The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells* Nature Biotechnology, 2014

[4] Laurens van der Maaten and Eric Postma. *Dimensionality Reduction:A Comparative Review* TiCC,Tilburg University,Oct 2009.

[5] Robrecht Cannoodt, Wouter Saelens and Yvan Saeys. *Computational methods for trajectory inference from single-cell transcriptomics* European Journal of Immunology, 2016.46:2496-2506

[6] Jinmiao Chen, Andreas Schlitzer, Svetoslav Chakarov, Florent Ginhoux and Micheal Poidinger. *Mpath maps multi-branching single-cell trajectories revealing progenitor cell progression* Nature Communications, Jun 2016.

[7] Laleh Haghverdi, Florian Buettner and Fabian J.Theis. *Diffusion maps for high-dimensional single-cell analysis of differentiation data Bioinformatics* 31(18),2015

[8] Laurens van der Maaten, Geoffrey Hinton. *Visualizing Data using t-SNE* Journal of Machine Learning Research 9(2008) 2579-2605

[9] Sam T.Roweis and Lawrence K.Saul. *Nonlinear Dimensionality Reduction by Locally Linear Embedding* SCIENCE,Dec 2000

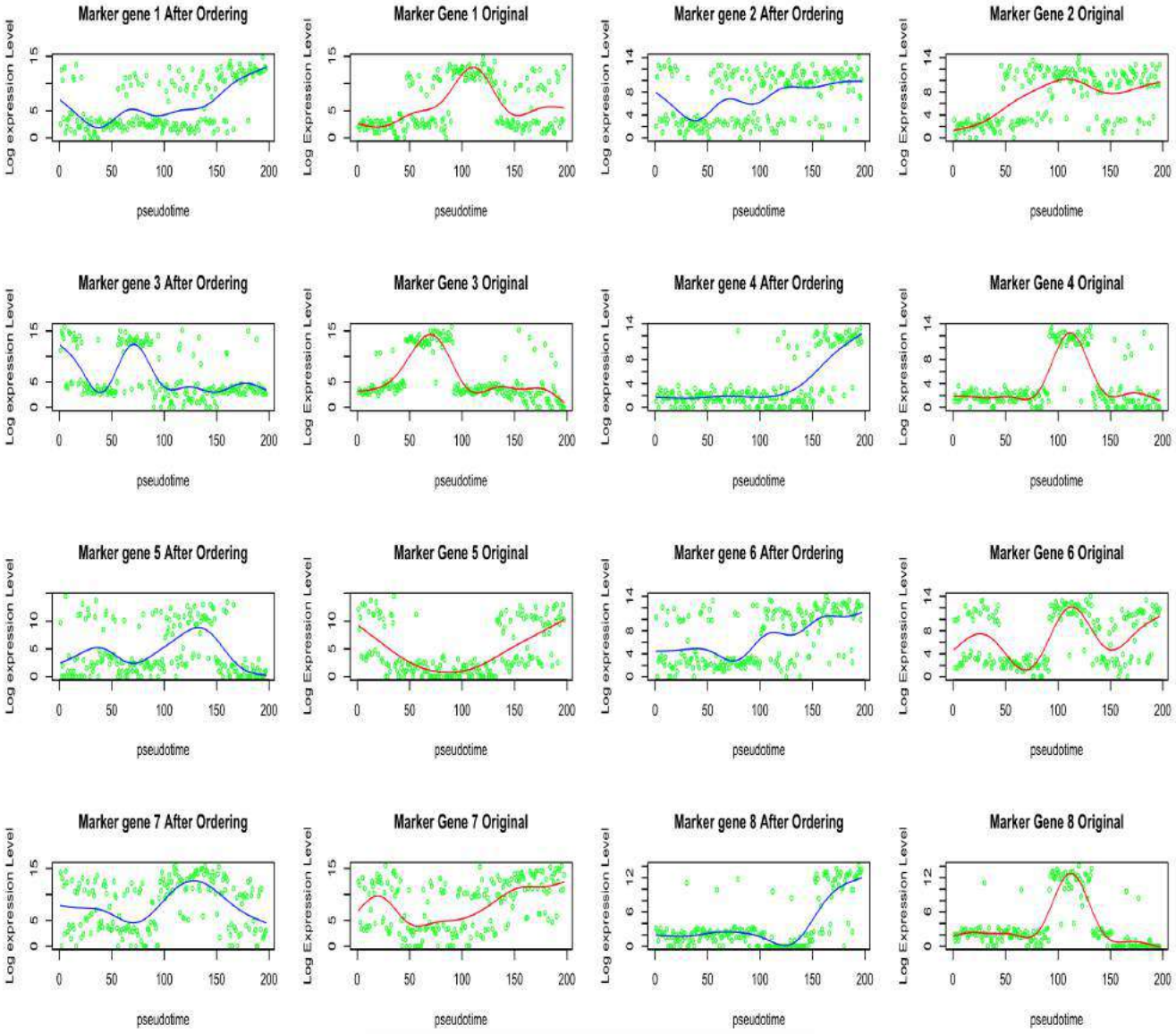
[10] Joshua B.Tenenbaum, Vin de Silva, John C.Langford *A Global Geometric Framework for Nonlinear Dimensionality Reduction* SCIENCE,Dec 2000

[11] John A.Lee,Michel Verleysen *Quality assessment of dimensionality reduction:Rank-based criteria*

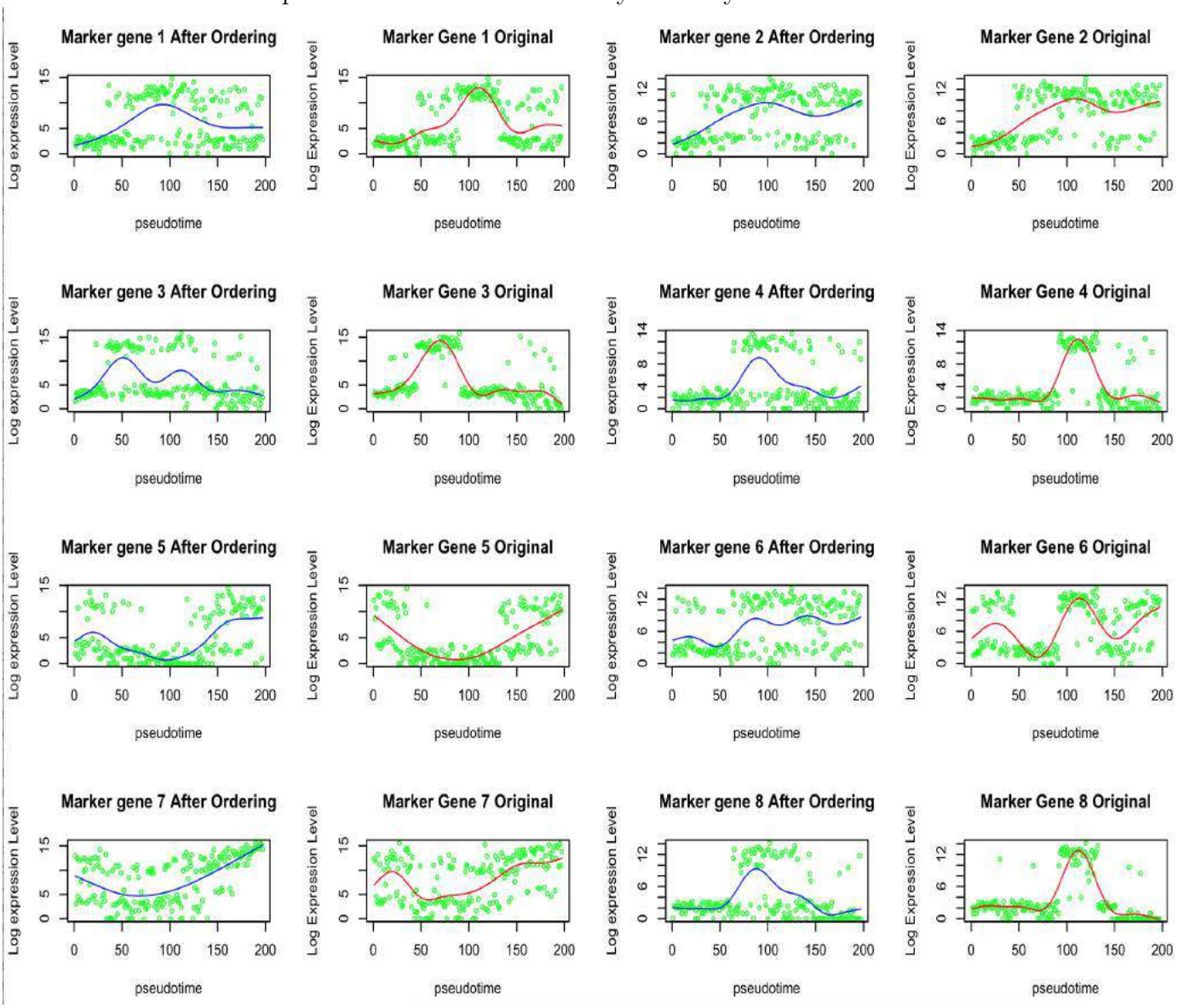
7 Appendix

7.1 Appendix A:Graphs

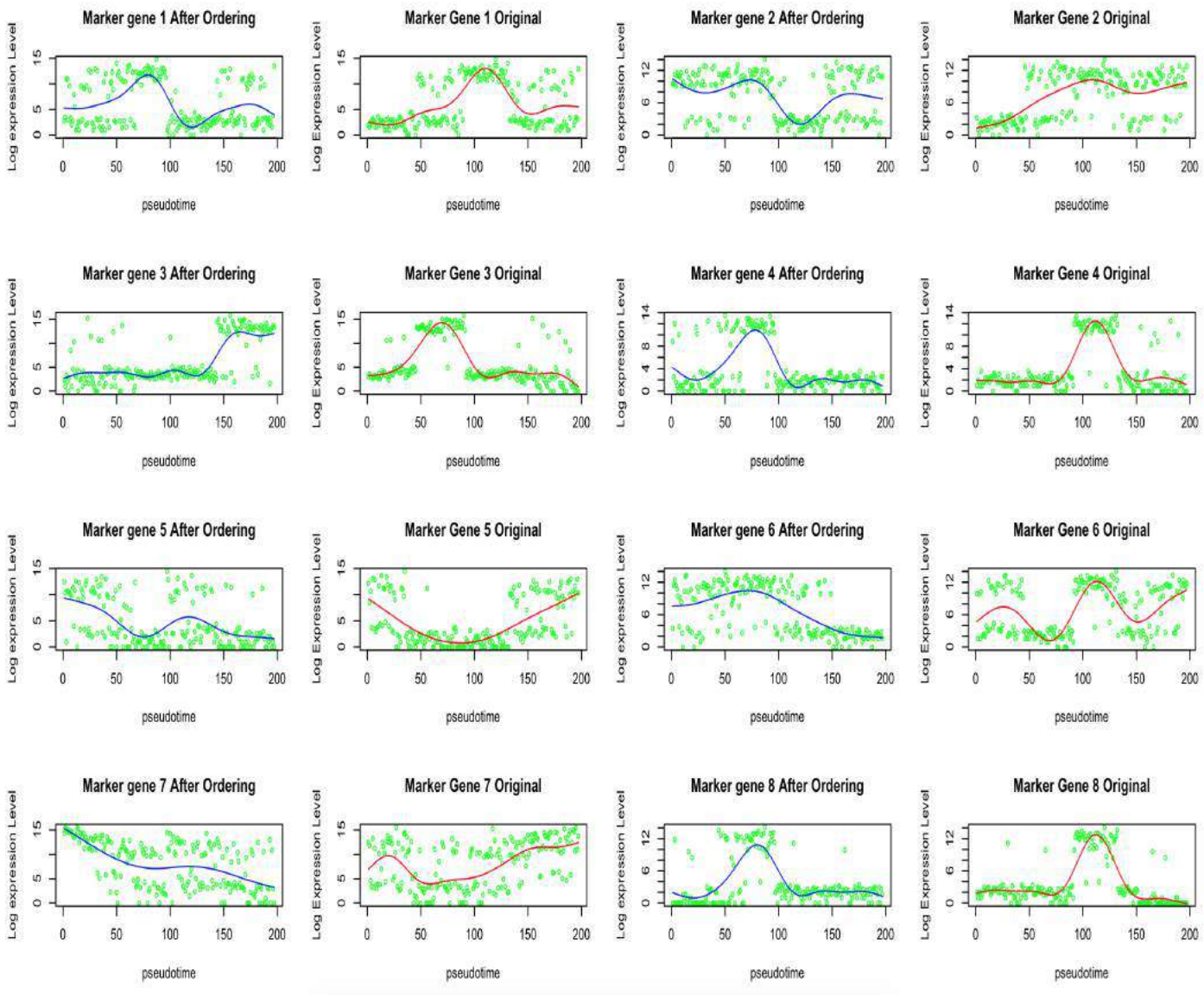
Graph1:Marker-Gene Consistency Nonclustering TSCAN:



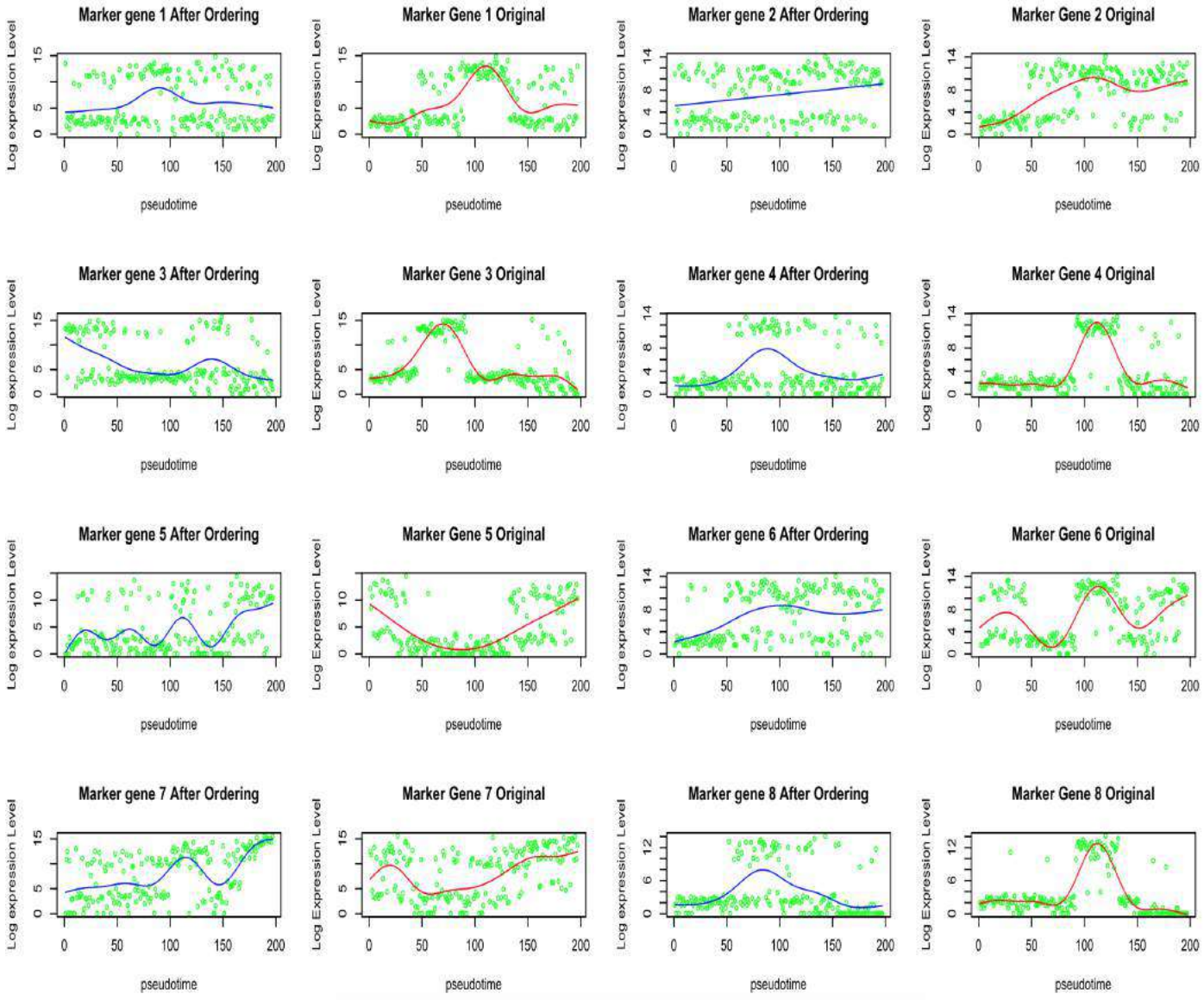
Graph2:Marker-Gene Consistency ordinary TSCAN:



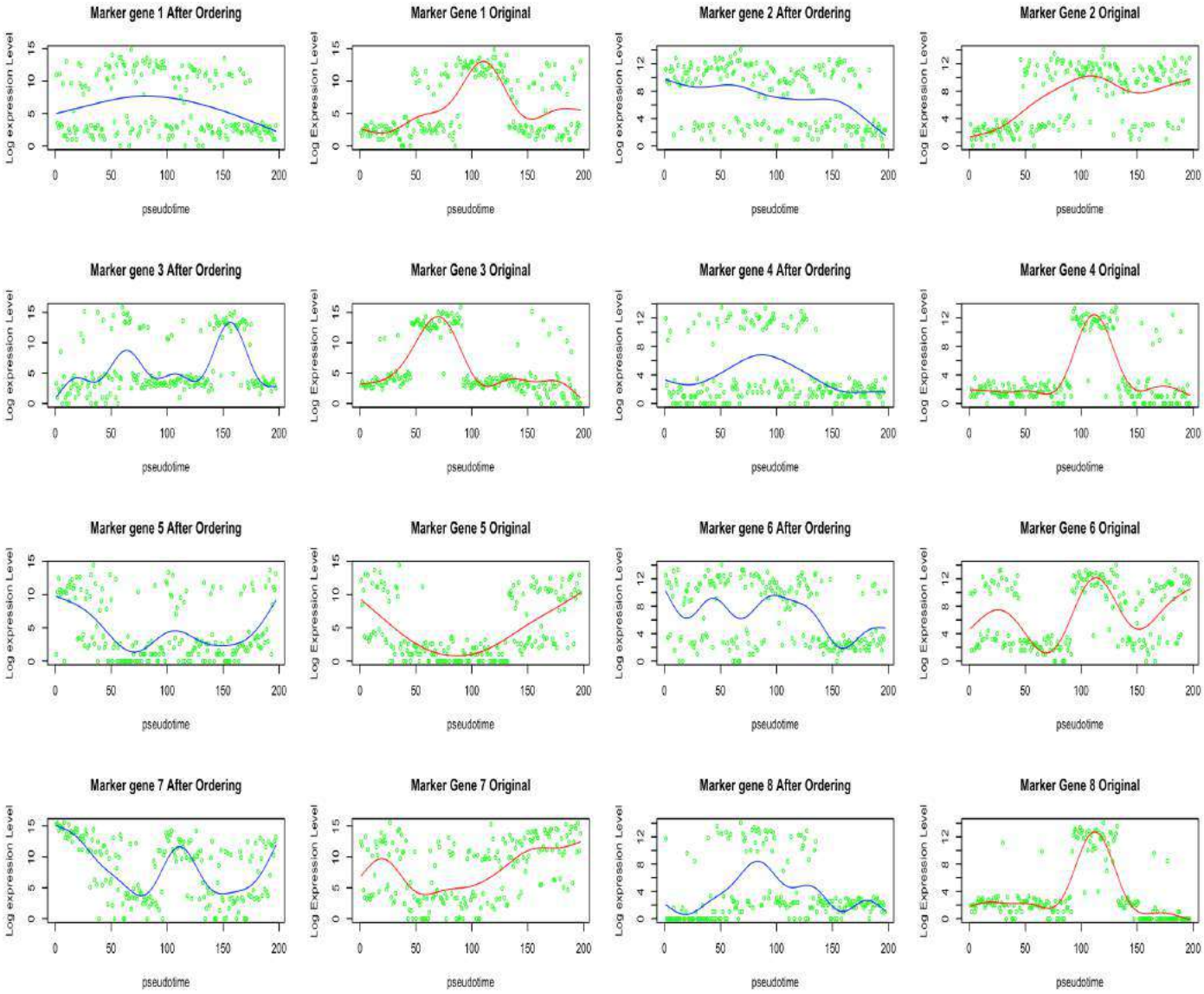
Graph3:Marker-Gene Consistency Diffusionmap TSCAN:



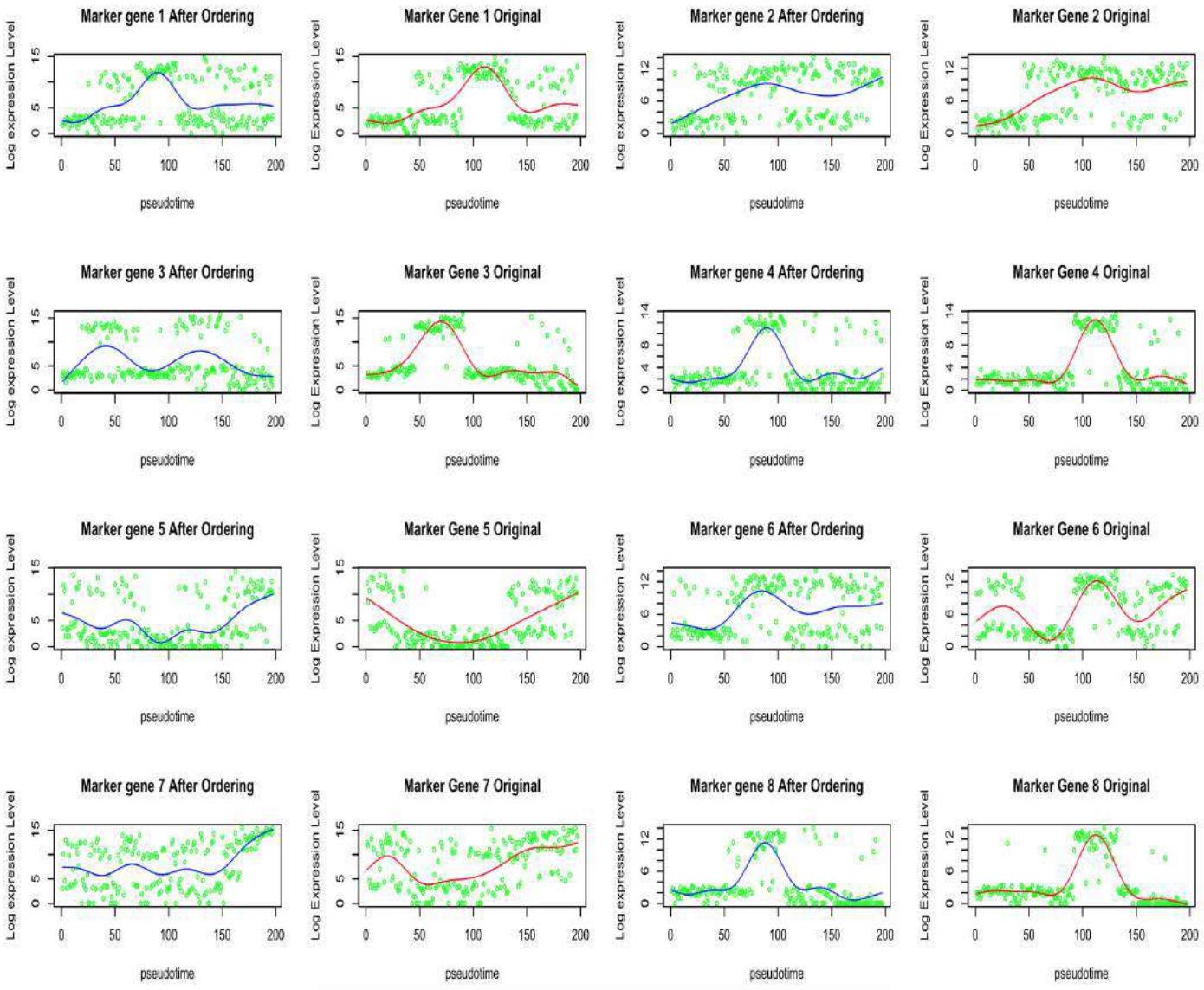
Graph4:Marker-Gene Consistency Kmeans TSCAN:



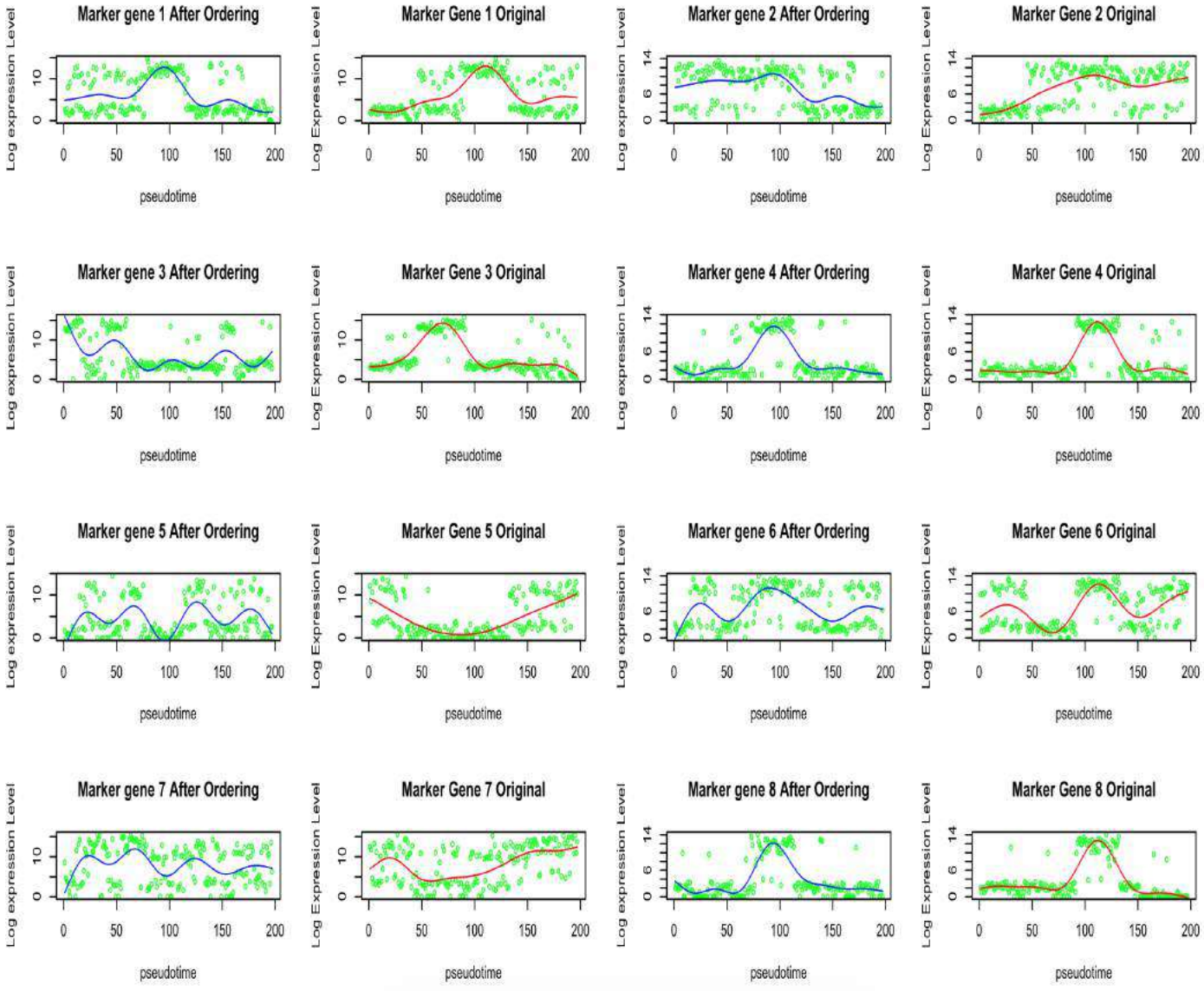
Graph5:Marker-Gene Consistency Monocle:



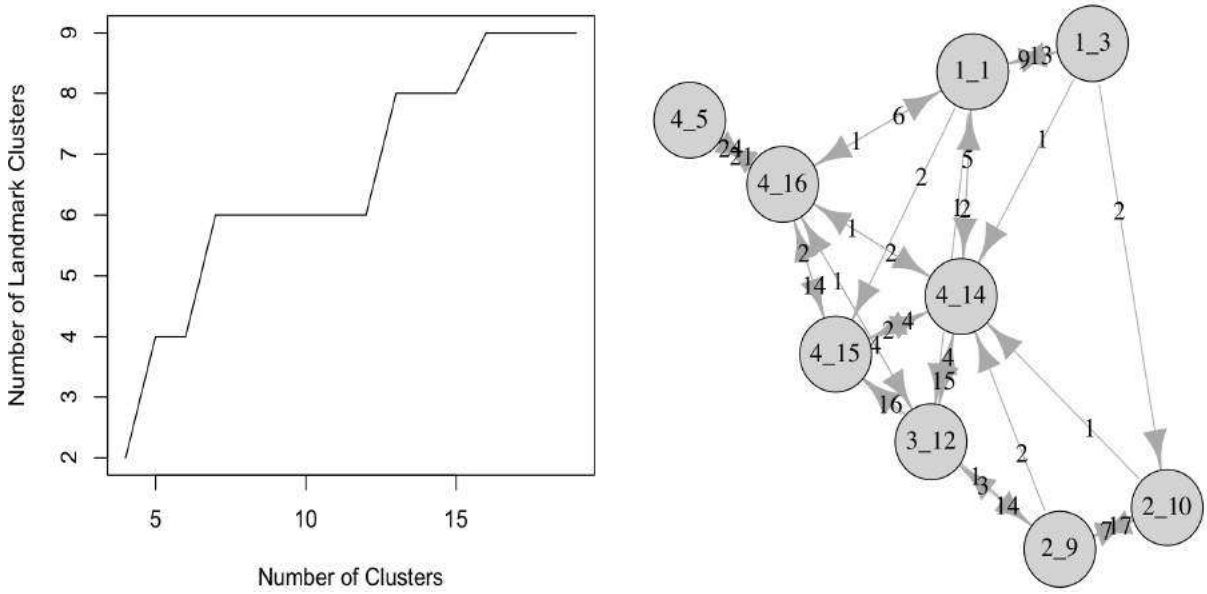
Graph6:Marker-Gene Consistency t-SNE TSCAN:



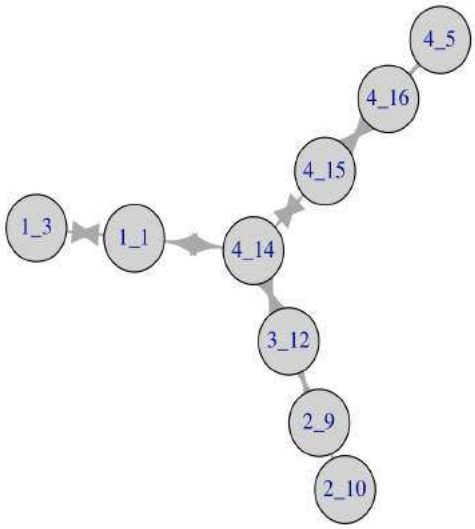
Graph7:Marker Gene Consistency t-SNE Monocle:



Graph8:Mpath clusters vs landmarker-clusters and original network construction



Graph9: Mpath Branching trajectory .



7.2 Appendix B:Artificially Specified Marker Genes

The first 20 differential expressed genes detected by us along the cell collection time are shown as below.

```
 [1] "ENSMUSG00000001020.8"  "ENSMUSG00000001025.8"  "ENSMUSG00000004612.9"
 [4] "ENSMUSG00000014453.3"  "ENSMUSG00000015314.10" "ENSMUSG00000021728.7"
 [7] "ENSMUSG00000023004.8"  "ENSMUSG00000023367.14" "ENSMUSG00000025163.6"
[10] "ENSMUSG00000026009.14" "ENSMUSG00000027863.8"  "ENSMUSG00000027985.14"
[13] "ENSMUSG00000028832.11" "ENSMUSG00000029810.15" "ENSMUSG00000030149.15"
[16] "ENSMUSG00000030165.16" "ENSMUSG00000030325.16" "ENSMUSG00000031239.5"
[19] "ENSMUSG00000031933.17" "ENSMUSG00000032026.6"
```

Modeling survival data: Application to Larynx dataset

Heqiao Ruan SID:915490857
email:hruan@ucdavis.edu

December 9, 2017

Contents

1 Abstract

Survival Analysis is a branch of statistics for analyzing the expected duration of time until one or more events happen. The core difference of this and traditional statistic is that the data is incomplete due to the limitation of observational studies. Our project primarily focus on applying as many as techniques and advanced methods in survival analysis in analyzing the dataset *larynx* by using R. For the first part we will apply various exploratory analysis on this and extract some certain features so that we can validate them in further modeling. Then we construct the estimated survival function through *Kaplan-Meier* estimator and cumulative hazard *Nelson-Aalen* through estimator.We guess the survival function's distribution and validate them via goodness of fit test. Furthermore,we will apply various non-parametric methods to explore the difference of survival time among different treatments and other covariates. For the second part, we will model this survival data by cox proportional hazard model and do model diagnostics via various techniques. Then we fit semi-parametric accelerated failure time model to examine the covariate effects on event times in censored data regression. as well as some parametric regression models and compare them and do model diagnostics. Finally we will compare these methods and try to identify the factors to influence the survival time and discuss some potential advanced topics such as additive hazard regression models.

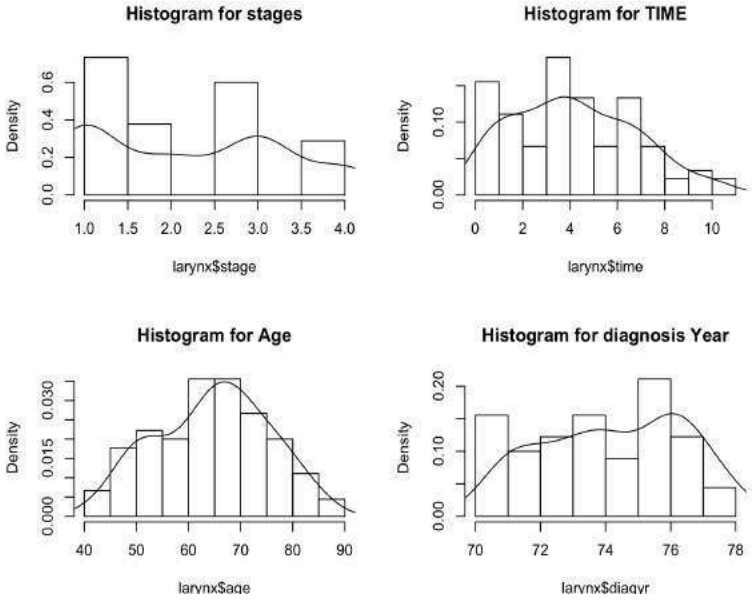
2 Preliminary Analysis

2.1 Introduction and Preprocessing

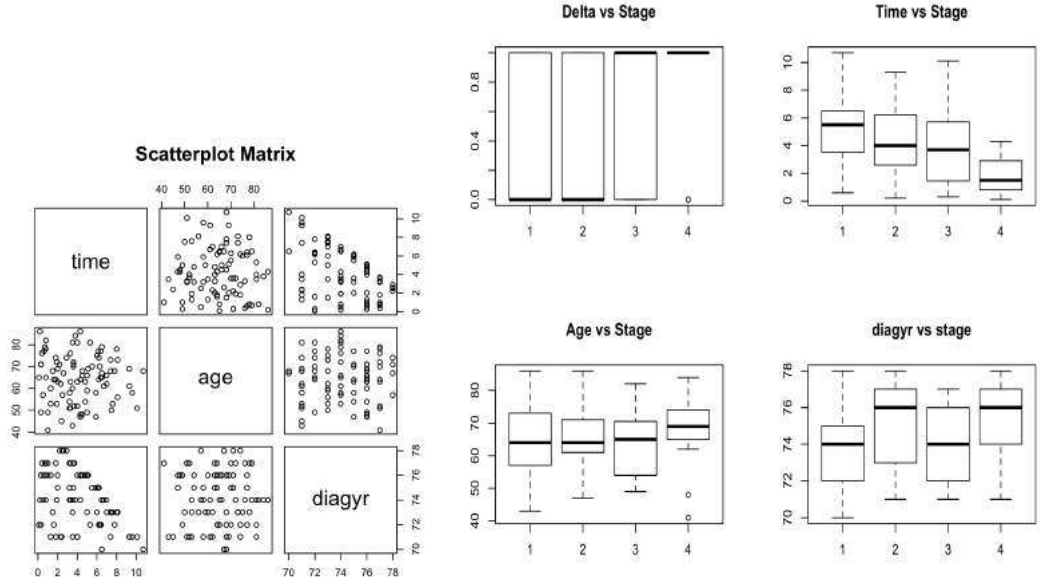
We choose the larynx dataset to conduct our analysis. This is a well-known dataset and it contains 90 rows(observations) and 5 columns(variables).This dataset originally appear in *Survival Analysis Techniques for Censored and truncated data* by Klein and Moeschberger(1997). The attributes are shown in the below table:

Stage	Stage of diseases(1,2,3,4)
Time	Time to death or on-study time,months
Age	Age at diagnosis of larynx cancer
diagyr	Year of diagnosis of larynx cancer
delta	Death indicator(0=alive,1=dead)

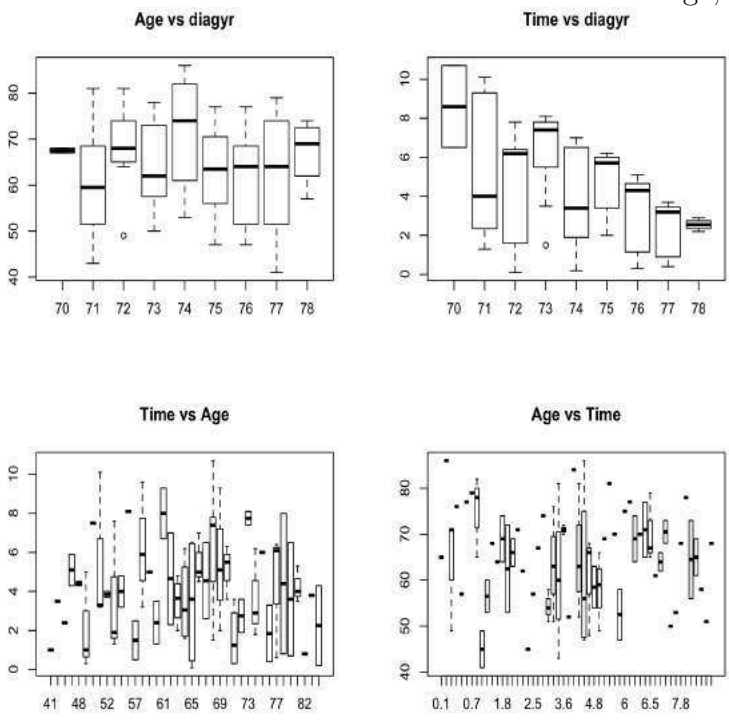
Here we model the survival data as $(\delta_i, X_i)_{i=1}^n$ and denote the delta variable as δ_i and denote the Time variable as X_i. For delta=0 means alive which means this case is censored during the observational study.
Here we can denote the variable *Age*, *Diagyr*, *Stage* as covariates. We can see the age and diagyr as qualitative variables for we can divide them into several groups so that we can examine the influence of them on survival time by using various modeling techniques.
First we are going to do some preliminary data analysis to examine the association between variables intuitively. We draw the histogram of the three covariates and time with its estimated density line:



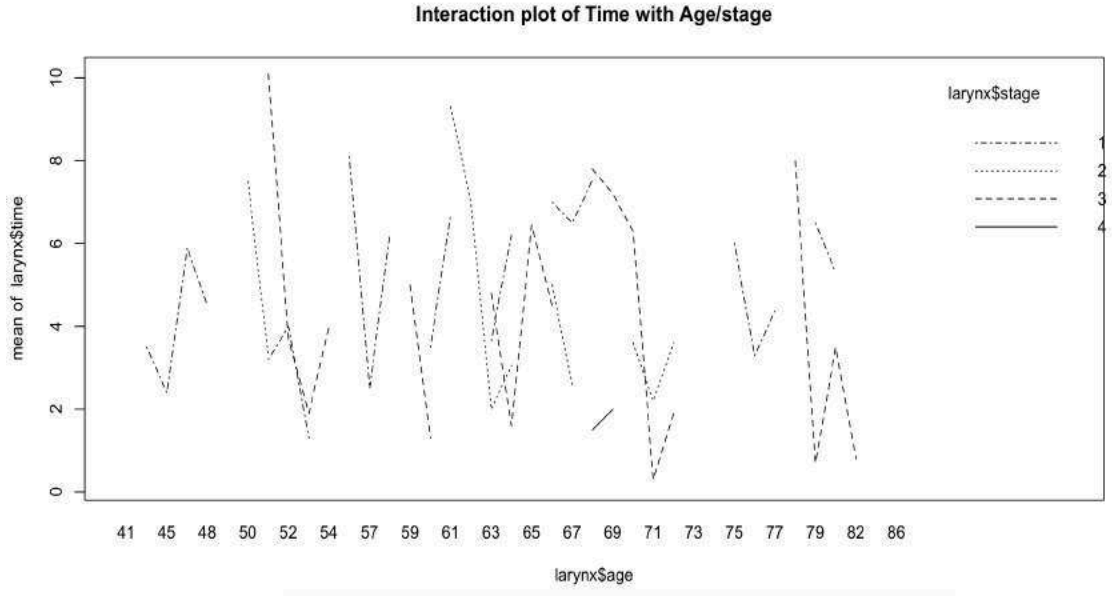
We can see that the distribution of age seems likes normal. There seems no obvious trend of the distribution of diagyr and stage.From the distribution of time,we will guess it follows some distribution later.
Then we draw the scatterplot matrix of the variables age,diagyr and censoring time:



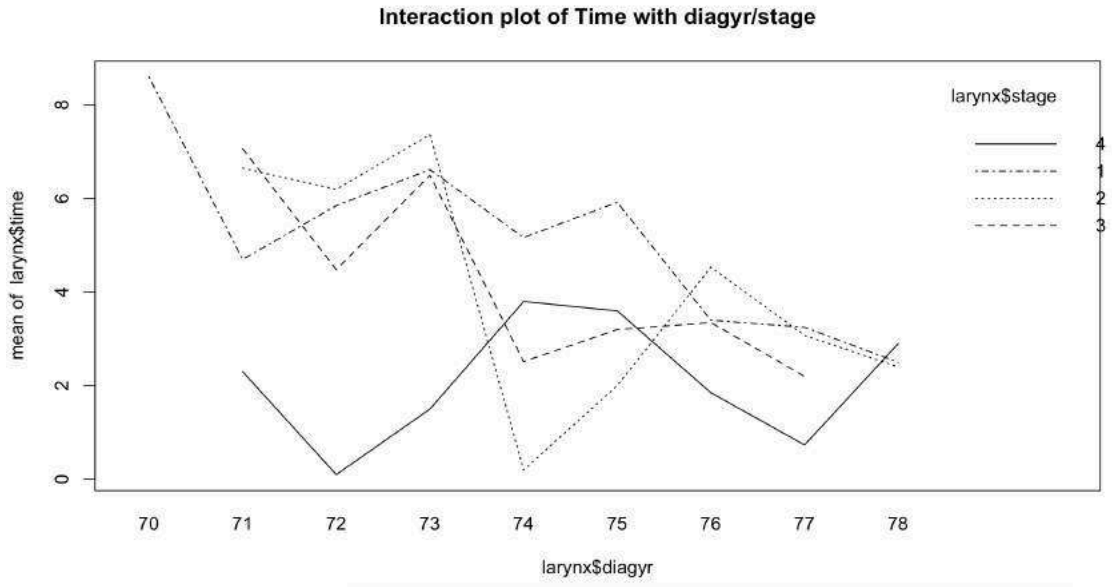
We can see that there's no obvious direct linear relation between these variables.
Then we plot the boxplot between the other variables and stages.We can see that survival time,diagyr and age are different among different stages. Observations with stage 4 have the smallest censoring time and latter the stage the shorter the censoring time. It matches our intuition and we will validate later. What's more, observations with stage 3 and stage 4 have more death cases than censoring cases which means they tend to die earlier. It also obeys our intuition.
Then we examine the relation between covariates age,diagyr and time:



From the plot we can see that, generally, later the cancer is diagnosed, the shorter the censoring/survival time. However,the older age when diagnosed does not necessarily mean this observed people will have shorter censoring time. It also depend on which stage the cancer is when they are diagnosed.So here we have to use the interaction plot to examine the relation between time and age diagnostics depending on different stages of cancer:



Here we still can't see obvious influence of stage1,2,3 on the censoring time of observations with different stages. However,observations diagnosed as stage 4, it tends to have shorter censoring time. What's more, the four lines(may not even continuous) are not parallel so we can conclude that there are relation between age and time depend on different diagnostic stages. Similarly we can draw the interaction plot of relation between time and diagyr's dependence on different stages.



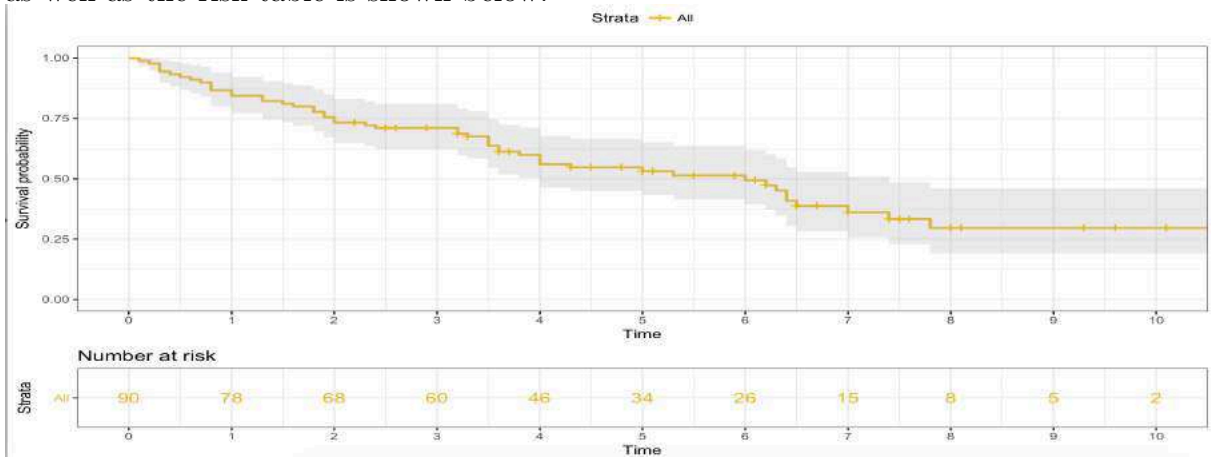
We can see that the lines are not parallel which means the relation between time and diagyr is depend on different stages. It also validate there are interaction effects among them. So there are interaction effects between these covariates.

For further model fitting and comparing survival curves between different groups,we should transform the continuous covariate to categorical variable which means we should divide them into several groups and denote them as category 1-4 and change their class as factor in r.We group them both by 4 quantiles, for variable age: if $Age \leq 55$ then denote as category 1,if $55 < Age \leq 65$,then denote as category 2,if $65 < Age \leq 75$,then denote as category 3,else denote as category 4. For variable diagyr: if $diagyr \leq 72$ then denote as category 1,if $72 < diagyr \leq 74$ then denote as category 2,if $74 < diagyr \leq 76$ then denote as category 3,else denote as category 4. By letting them as class factor,we transform them into categorical variables and add at the right side of our dataset for further survival model fitting(we maintain the original variables for we can fit a model with continuous variables).
Then we see the basic summary statistics for every group, first we see that the mean survival time for every group are different:5.25 for stage I,4.38 for stage II,3.93 for stage III and 1.83 for stage IV.

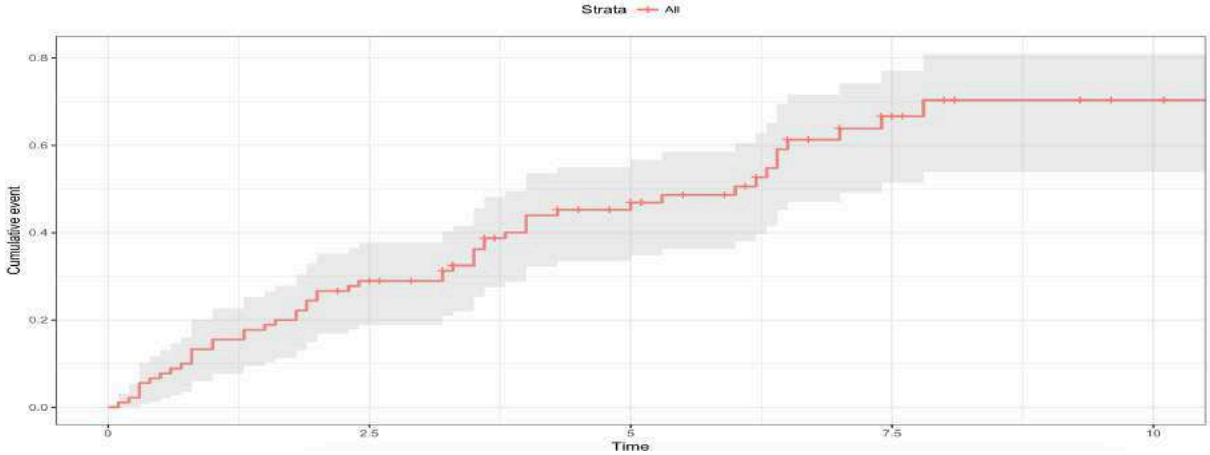
2.2 Estimating survival function and relevant exploration

First we get the risk set and the event set for every observation and then we can derive kaplan-meier estimator and Nelson-Aalen estimator based on the information in the dataset. Then we estimate the mean and median of the survival function.
As we all know, the kaplan-meier estimator is given by $\coprod_{x_i \leq t}(1-\frac{1}{n-i+1})^{\delta_i}$ and in this problem we have some ties which means the death case and censoring case for a specific observed time is not unique(tied).So the formula become: $\hat{S}_i = \coprod_{j \leq i}(1-\frac{d_j}{r_j})^{\delta_j}$ Here r_j is the risk set at time j and d_j is the number of death observations at time j.
For Nelson-Aalen estimator, the formula is $\Lambda_i = \sum_{i \leq j} \frac{d_j}{r_j}$. Then we would like to derive the interval estimation of Kaplan-Meier estimator. Here if we only use the traditional technique which indicates that $var(\hat{S}_t) = \int_0^\infty \frac{dF_u}{(1-H(S))^2}(\int_t^\infty S(s)ds)^2) = (\hat{S}_t)^2 \sum_{j:\tau_j \leq t} \frac{d_j}{(r_j-d_j)r_j}$(Greenwood formula) we will find that some of the estimators is more than 1,it's not accurate.Note that this formula is indeed derived from a method we applied many times in empirical process including matrix transformation.
So here we apply log-log technique to modify this.Here we define $L(t) = log(-log(S(t)))$,then $\hat{L(t)} = log(-log(\hat{S}(t)))$. The confidence interval for L(t) is $[\hat{L}(t)-H, \hat{L}(t)+H]$,here H is determined by $\phi(0.95)sd(\hat{L}(t))$ So by delta method, the variance become $\frac{1}{[log\hat{S}(t)]^2}\sum_{j:\tau_j \leq t}\frac{d_j}{(r_j-d_j)r_j}$.
Then the confidence interval for $\hat{S}(t)$ become:$([\hat{S}(t)]^{e^H}, [\hat{S}(t)]^{e^{-H}})$. Then we can see that none of the interval estimator will contain 1. Similary,for Nelson Aalen estimator the interval estimator is becoming: $[\hat{A}(t)e^{\phi(0.95)\frac{\hat{\sigma}(t)}{\hat{A}(t)}}, \hat{A}(t)e^{-\phi(0.95)\frac{\hat{\sigma}(t)}{\hat{A}(t)}}]$ where $\hat{\sigma}(t) = \sum_{t_j \leq t}\frac{(r_j-d_j)d_j}{(r_j-1)r_j^2}$.
Then comes one of our core part,we fit the survival function by the *survfit* function in R

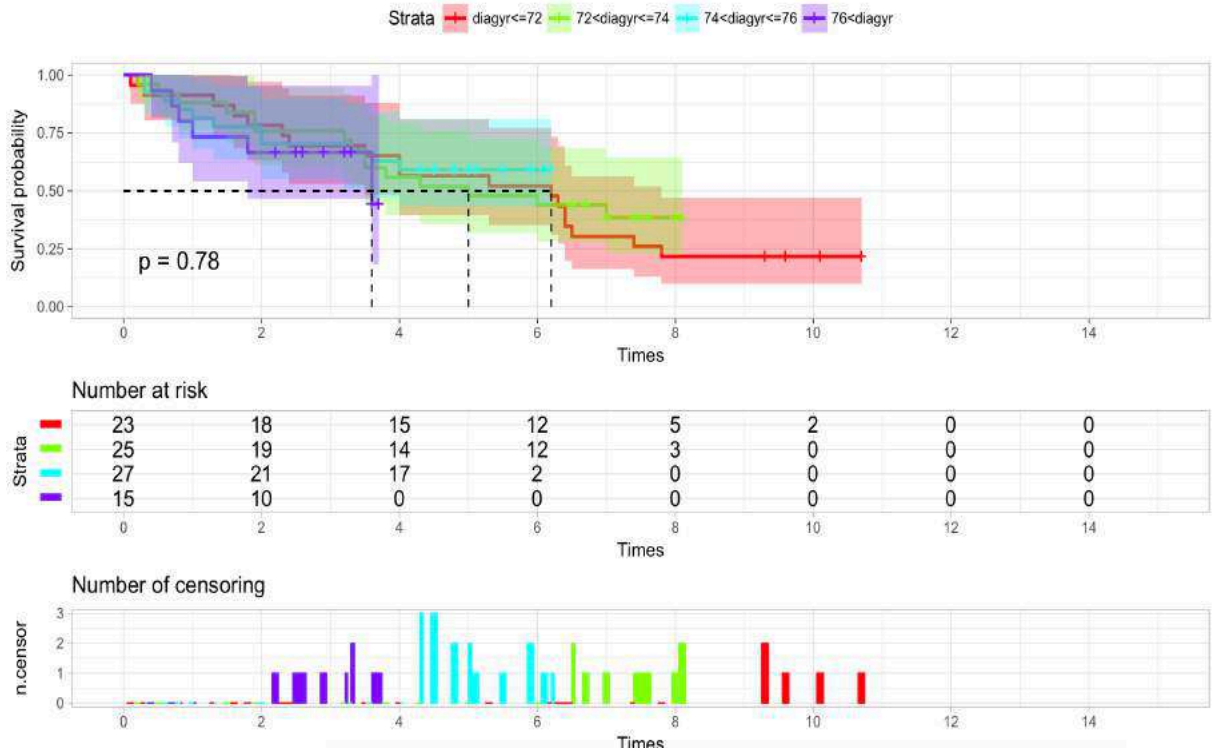
package *survival*. Here we would like to use the function *ggsurv* and *ggsurvplot* in R package *survminer*. For we want to compare the survival time and survival function among different categories of categorical variables, we fit them with respect to *factor(stage)*,*factor(age) factor(diagyr)* as well as the full model. Firstly we fit the full model and the survival curve as well as the risk table is shown below:



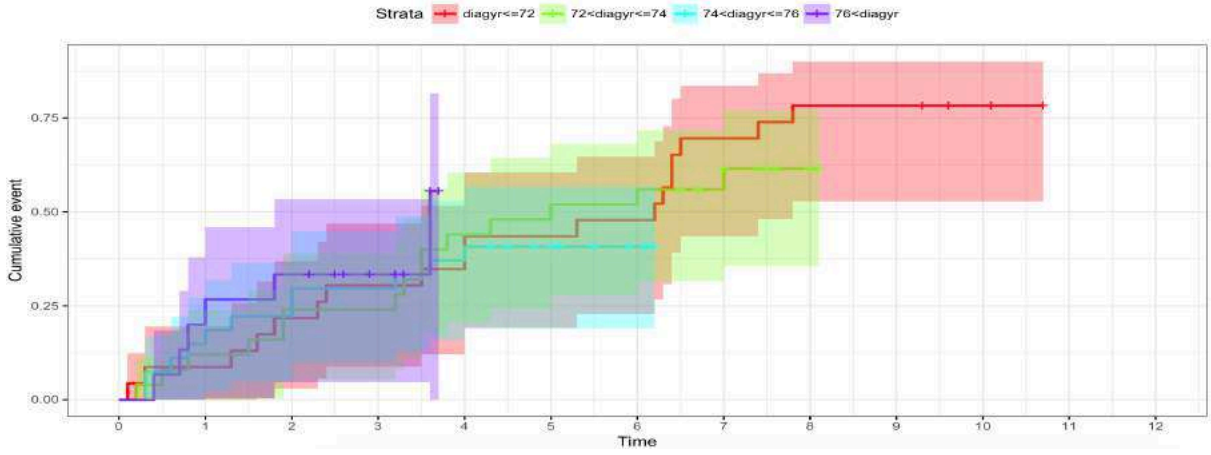
Here we guess the distribution of the hazard rate as exponential,weibull and distribution .We will do goodness of fit test to see which guess will have more probability to be right in the next part. Then we plot the cumulative hazard of the full model:



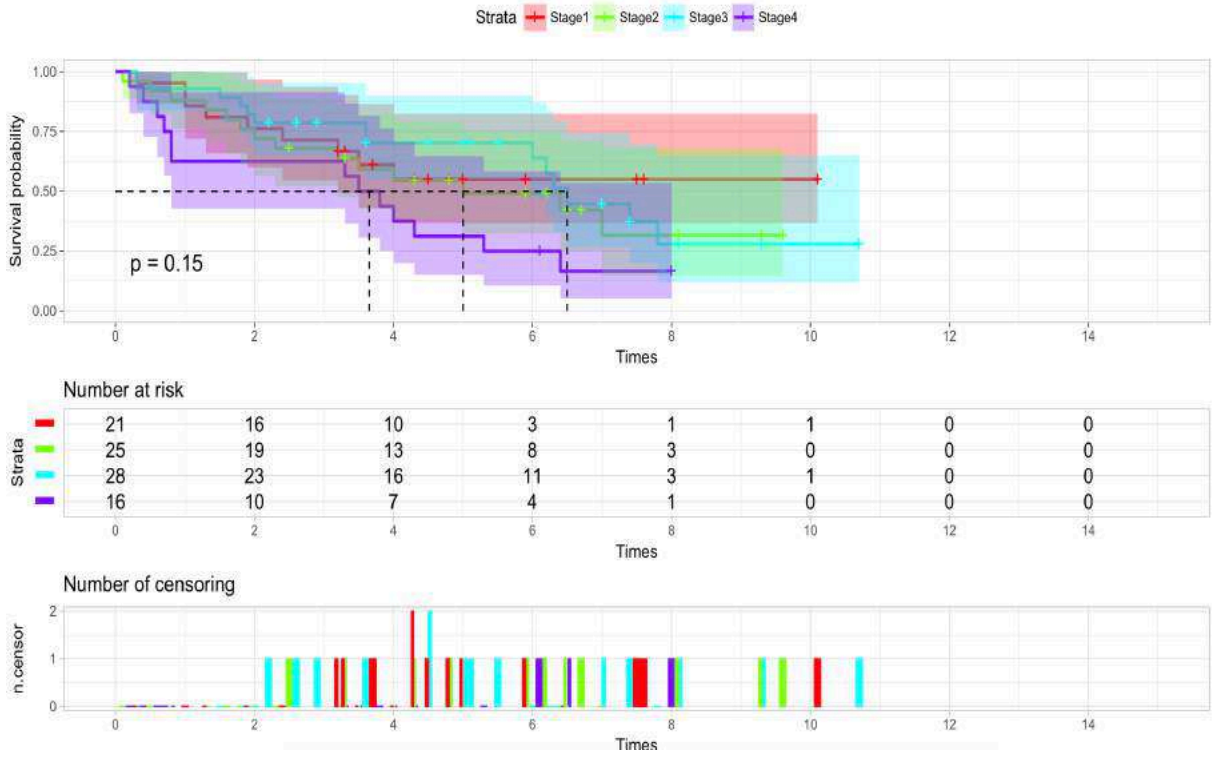
From the plot we can see that the Then we fit the model with respect to the four diagyr group and the survival curve as well as the risk table and table of number of censoring is shown below:



Here we can see that the group with the youngest diagnostic year of cancer has the longest survival time and the longest median survival time and the censoring status is perfectly separable by the four groups of diagnostic year.We see that there are certain kinds of difference among these survival curves but the difference may not be very significant.
Then we plot the cumulative hazard curve for diagyr groups:

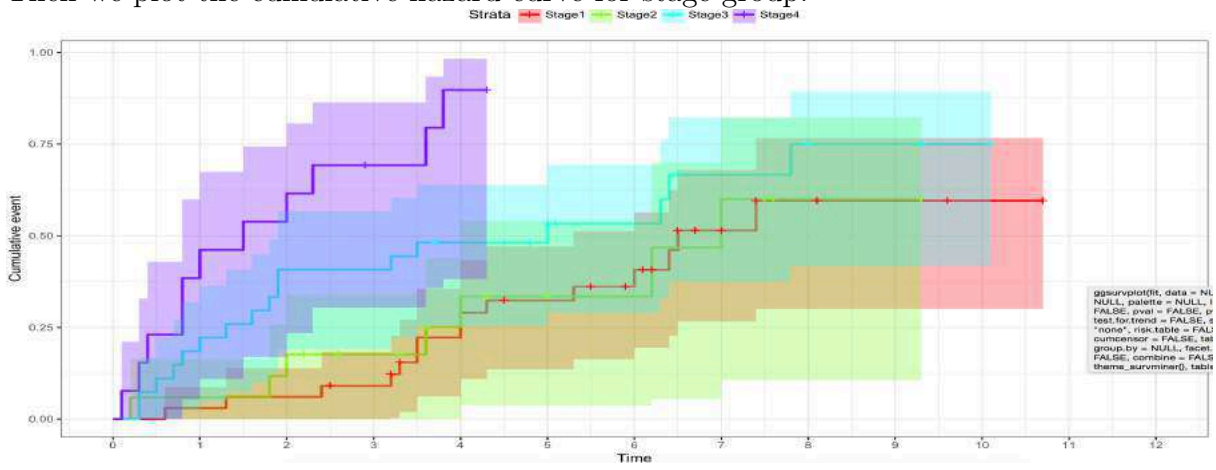


From here we can see that the four curves are fairly similar in their scope of existence and further test later may validate this. Then we fit the model with respect to the four stages of cancer and the survival curve as well as the risk table and the table of number of censoring is shown below:



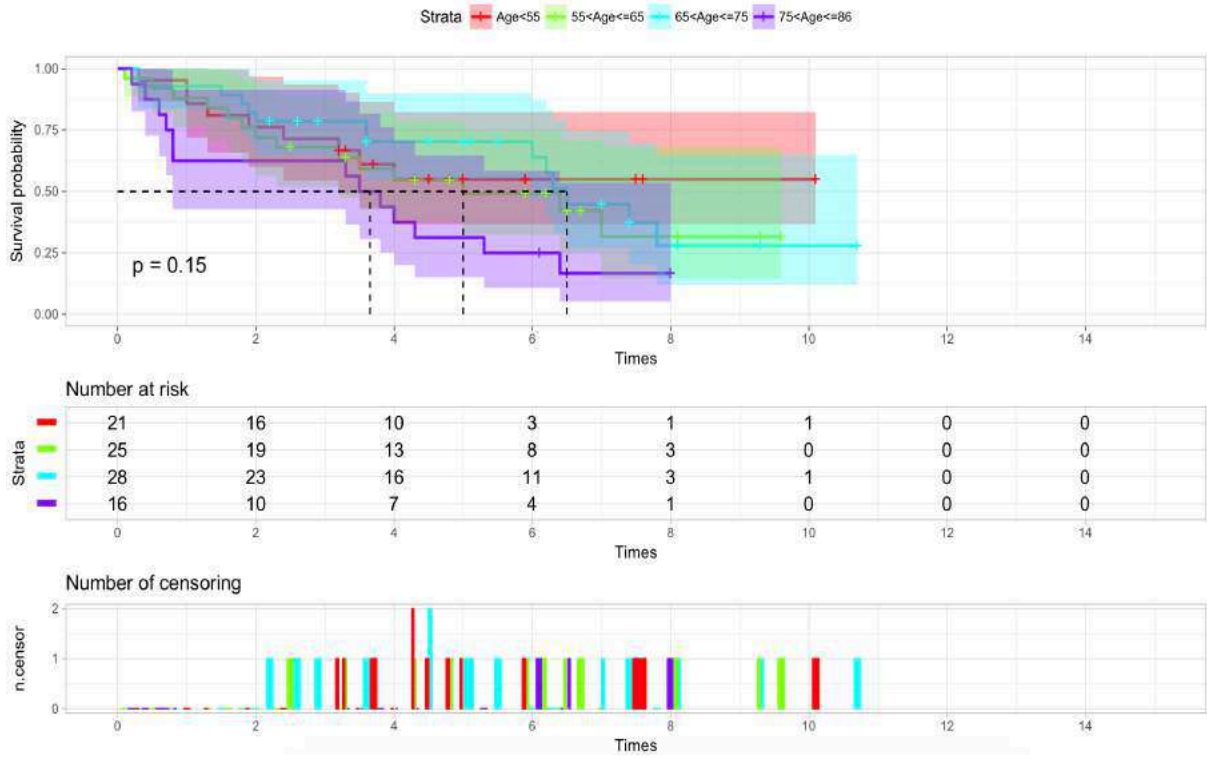
Here we can see that the difference of survival curves of stage1,stage2 and stage 3 is not as obvious as the difference between survival curves of stage4 cancer and them. What's more, the censoring observations are not so perfectly separable among different stages as among different diagnosed years. The median survival time is becoming smaller as the stage become later(stage 4 is a later stage 1 of the cancer i.e).It matches the common sense.

Then we plot the cumulative hazard curve for stage group:

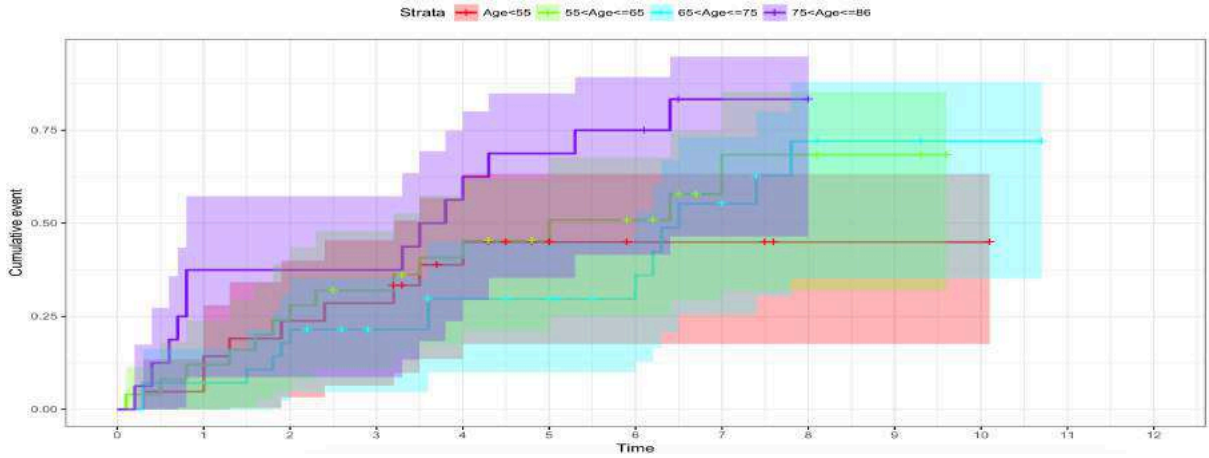


We can conclude that the four cumulative hazard are really different which means the four survival curves are also significantly different in their scope of existence and further test will validate this.

Then we fit the model with respect to the four groups of age and the survival curve as well as the risk table and the table of number of censoring is shown as below:



In terms of survival functions in different age groups, we can see that older people tend to have lower median survival time and we can see that as the time flies the estimator interval become larger and larger(this is trivial for the formula of the variance).
Then we plot the cumulative hazard curve for age group:

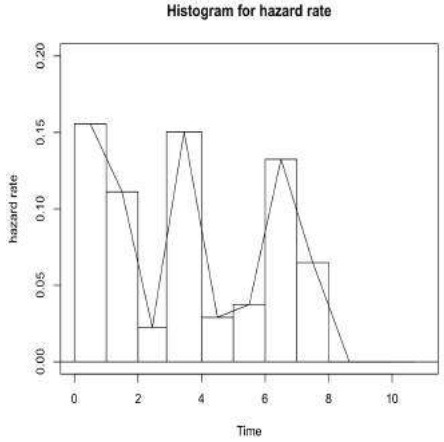


Then we estimate the mean and median of the survival time. According to the lecture,the mean survival time point estimator is $\hat{\mu} = E[\hat{T}] = \int_0^\infty S(\hat{S})ds = \sum_{i=0}^{n-1}(X_{i+1} - X_i)S(\hat{S})$ and its variance is $\sum_{i=1}^{n}[\int_{t_i}^{\tau} \hat{S}(t)dt]^2 \frac{d_i}{Y_i(Y_i - d_i)}$.This is derived by the idea of empirical process. Then we know that in asymptotic point of view it follows normal distribution. Then in the problem,we construct 95% confidence interval,its result is 5.689754 for the estimated mean survival time and [4.827,6.552] for the interval estimate for the mean survival time. We implement this method by R without calling the survfit function in the survival package and we can check it

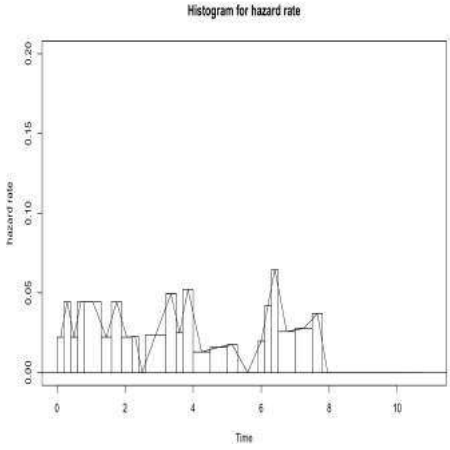
in appendix.
Then for the median survival time: $median = \hat{S}^{-1}(\frac{1}{2})$ We see that $\hat{S}(k) > \frac{1}{2}, \hat{S}(k+1) < \frac{1}{2}$ then(for more accurate),the point estimator is $\frac{\hat{S}(k+1)+\hat{S}(k)}{2}$. What's more,we know that its variance is $\sum_{i=1}^{n} \frac{d_i}{Y_i(Y_i-d_i)}$. Similarly,we have asymptotic normal so we can get the 95% interval estimation. Its result is 5.95 for the estimated median survival time and [5.439,6.464] for the interval estimation. Notice that it's much larger than the mean survival time. This is usual for the estimator gives increasing jump sizes with increasing t and due to censored observations dropping out,the gaps between uncensored observations tend to increase with t. So $\hat{\theta}$ tends to be larger.
Then we estimate the mean survival time and median survival time of each stage:
For the different diagnosed years the median survival time of four categories are 6.2 for category 1,5.0 for category 2,3.6 for category 4 and for category 3,the survival function will never reach 0.5 so no median. We can see that the earlier the cancer diagnosed,the longer is the survival time.
For the different age of diagnosis, the median survival time of four categories are 5 for category 2,6.5 for category 3,3.65 for category 4. For category 1,the survival function never reaches 0.5,no median.
For the different stages of cancer,the median survival time of four stages are 6.5 for stage1,7.0 for stage 2,5.0 for stage 3,1.5 for stage 4. Here we can see that the median estimator of stage 2 is not that accurate,it's much larger than the sample mean.So later the stage when diagnosed,the shorter survival time.
So we can see that after comparing the median estimators of every groups,it is fairly consistent with what we conclude in preliminary analysis(boxplot). However the difference of median and mean survival time does not necessarily indicate the significant difference between different curves. This may result from the difference of their domain.

2.3 Estimating Survival Function and Goodness of Fit test

A very important topic is to what distribution the survival function follows and if we can show that the survival function nearly follows certain distribution,we can do parametric regression models on the dataset and it will greatly help us.In doing this the core part is to do the goodness of fit test,we can first draw the histogram for the survival function and hazard rate and guess what the distribution may look like. We plot the histogram for the hazard rate and is shown as below(the breaks are almost same time interval).
However,in real life, the dataset often not do well in helping us identifying what the distribution of survival function is. Then if the survival function shows no obvious pattern(no distribution will make sense here), the minimum chi-square technique will not work. That's what I have to point out.
First after drawing the histogram to see what kind of distribution it may follows.However,it's not easy for us to identify what kinds of distribution it may follows purely from the histogram.The histogram shows no obvious pattern of the distribution of the survival function.



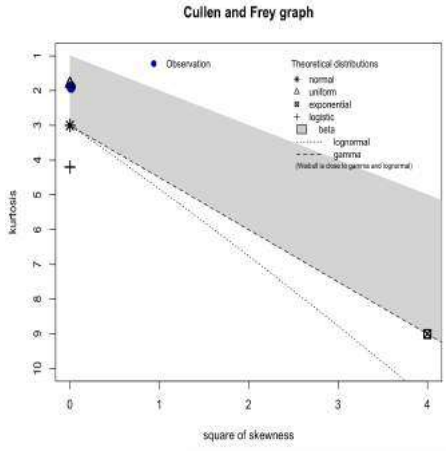
Unfortunately, we can guess any specific kind of distribution just based on this histogram, if we narrower our breaks to every two points, the histogram is shown as below:



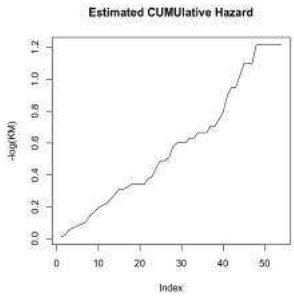
I still can't guess which specific distribution it may follows. So we do can try some parametric distributions such as exponential,weibull and even gampertz but these are all return to very negative results.

However, we should somewhat apply the methods mentioned in this course to do the goodness of fit test by using minimum chi square technique. Here the minimum chi square technique construct a gaussian process which $Z(y_1), ..., Z(y_k)$ which $Z(y_1)$ is independent to $Z(y_i) - Z(y_{i-1})$,i=2,3,...,k. Here $Z(y_i) = \hat{\Lambda}_n(y_i)$ is the cumulative hazard. Then the test statistic we construct is $\sum_{k=1}^{K} \frac{(\Delta Z_k)^2}{var(\Delta_k Z_n(y_k))}$. It is a function of the parameter of our assumed distribution. If the assumed distribution is exponential(α), $\Lambda(t) = \alpha t$.Then we do find the parameter that minimizes this statistic and the distribution of the minimized chi-square statistic is χ^2_{K-1}. Here we may need to do simulations to find the optimized parameter α. Note that this procedure is better than the ordinary chi square test $A = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i^2}$ for here we use the independent increment property of the gaussian process and thus wipe out the covariance approximately among different terms on statistic A. Here we may assume the distribution of the survival function is exponential and do this procedure. We find that the optimized α is 0.004 and at this time the minimum chi square statistic is 133. It's obviously an awful fit.

Then we use the *descdist* function in the fitdistrplus package in r to try to see which distribution fits the survival function $\hat{S}_t$) best based on skewness and kurtosis.



From the plot we can see that the uniform distribution fits the survival function(our KM estimator) best. It totally makes non-sense.
Then we try to see whether the cumulative hazard follows certain pattern:we draw the graph of $\hat{\Lambda}_t$ and it is shown as below:



The only thing we can see from this very graph is that the cumulative hazard looks like a straight line so may be we can guess is that the survival function follows exponential distribution which has been proven awful fitting by the previous minimum chi-square technique. So unfortunately, from both the minimum chi-square technique and even some more advanced methods, we can't identify any kinds of parametric models based on this dataset. So maybe all of the parametric techniques may not works very well in this case.

2.4 Non-Parametric Techniques

Here we apply Mantel-Haenszel test and wilcoxon rank test to this multi-sample problems. Here we have four groups which denotes the four stages.
First we apply the Cochran-Mantel-Haenszel test. We know that in two sample problem we have a series of contingency table. In the multi-sample problem we can just extend it naturally and our test statistic still follows chi-square distribution.
For convenience we obtain the risk-event table first reflecting the event set and risk set for each group in every observation.
Then if we want to compare the survival function of each stage group with the whole dataset.Now

the null-hypothesis become $H_0 : F_1 = F_2 = .. = F_p$ which means there's no difference of survival function between each group. The test statistic here become: where r_j is the number of elements in the jth risk set of full dataset.$r_{i,j}$ is the number of elements in the jth risk set for stage i. Here for the multi-sample problem, let $O_j = (d_{1,j}, ..., d_{p,j})$ bet the vector of observed number of failures in group $1- > p$,$E_j = (\frac{d_j r_{i,j}}{r_j}, ..., \frac{d_j r_{p,j}}{r_j})^T$ denotes the mean of O_j. This is in fact similar to one kind of goodness of fit test $\sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$Then they have a covariance matrix: $V_j = \begin{bmatrix} v_{11,j} & v_{12,j} & ... & v_{1p,j} \\ ... & v_{22,j} & ... & v_{2p,j} \\ ... & ... & ... & ... \\ ... & ... & ... & v_{pp,j} \end{bmatrix}$ where diagonal element is $v_{kk,j} = \frac{r_{k,j}(r_j - r_{k,j})d_j(r_j - d_j)}{r_j^2(r_j - 1)}$ and non-diagonal element is $v_{km,j} = \frac{r_{k,j} r_{m,j} d_j(r_j - d_j)}{r_j^2(r_j - 1)}$. For all observations we sum all the v matrix and all the vector $O_j - Ej$ up. According to asymptotic theory we know that $(O - E)^T V^{-1}(O - E)$ follows $\chi^2(p - 1)$ distribution. Then after calculation the test statistic become:22.8 and the p-value is $4.53e - 05$. So we reject the null hypothesis which means we have enough evidence to state that the survival functions are different among the 4 stages.
Similarly when we conduct this test in age group and diagyr group,it doesn't works well for the p-value of the CMH test are 0.145,0.783 respectively which means we should state that the survival functions are the same among these two kinds of grouping in significance level 0.1. There are also some non-parametric tests such as Gehan-test,tarone-wane test. However, they are indeed very similar to the CMH test but only differences in the weights in each observation(each contigency table) and CMH test is the most commonly used test here.We have to point out that this kind of test is limited to location shift and in scale shift,it won't work well.

3 Main modeling and exploration

3.1 Cox Proportional Hazard Model

Cox model is one of the most fundamental part of survival analysis. The idea of proportional hazard model comes from comparing the two survival curves and getting the relative risk. $S_1(t) = (S_0(t))^\theta$. So in terms of hazard rate we have equation: $h(t|Z) = h_0(t)exp(\sum_{k=1}^{p} \beta_k Z_k)$,in another form we have:$\frac{h(t|Z)}{h_0(t)} = exp(\sum_{k=1}^{p} \beta_k Z_k)$ where $\theta = exp(\sum_{k=1}^{p} \beta_k Z_k)$
Here the cumulative hazard are also proportional:
Here,if we want to fit the cox model with the covariate stage(4 categories) and ages. Then the model become: $\lambda_l(t) = e^{\beta H_l}\lambda_0(t)$ where $\lambda_0(t)$ is the baseline hazard. The exponential term is $exp(\beta H_l) = exp(\beta_1 Z_1 + \beta_2 Z_2 + ... + \beta_{p-1} Z_{p-1})$. In the modeling of age and stages,the covariates Z become:
Z_1:1 If stage II of cancer,0 otherwise;$H_l = (1, 0, 0, Z_4)$: If stage II of cancer
Z_2:2 If stage III of cancer,0 otherwise;$H_l = (0, 1, 0, Z_4)$: If stage II of cancer.
Z_3:3 If stage Iv of cancer,0 other wise;$H_l = (0, 0, 1, Z_4)$:If stage III of cancer.
Z_4:Age variabels.$H_l = (0, 0, 0, Z_4)$: If Stage I of cancer.

Another model we consider(called model3) is only consider the stage variable and then the model become: $\lambda_l(t) = e^{\beta H_l}\lambda_0(t)$ where $\lambda_0(t)$ is the baseline hazard. The exponential term is $exp(\beta H_l) = exp(\beta_1 Z_1 + \beta_2 Z_2 + ... + \beta_{p-1} Z_{p-1})$. In the modeling of age and stages,the covariates Z become:
Z_1:1 If stage II of cancer,0 otherwise;$H_l = (1, 0, 0)$: If stage II of cancer
Z_2:2 If stage III of cancer,0 otherwise;$H_l = (0, 1, 0)$: If stage II of cancer.
Z_3:3 If stage Iv of cancer,0 other wise;$H_l = (0, 0, 1)$:If stage III of cancer.
If Stage I of cancer,$H_l = (0, 0, 0)$.
Our dataset has various ties so we should use another approach to compute the partial likelihoods. The most common one is conducted by Breslow(1974), $L(\beta) = \coprod_{i=1}^{n} \frac{exp^{(\beta^T \sum_{j \in D_i} z_j)}}{[\sum_{j \in R_i} exp^{\beta^T z_j}]^{d_i}}$,here d_i denotes the number of deaths at t_i and D_i be the set of all individuals who die at time t_i.Here we use this method to handle the ties. This likelihood considers each of the d_i events at a given time as distinct and construct their contribution to the likelihood function and obtains the contribution to the likelihood by multiplying over all events at time t_i. The estimator $\hat{\beta} = argmax_\beta L(\beta)$ maximize the partial likelihood. Here the log-likelihood function is $l(\beta) = \sum_{i=1}^{n} \delta_i(\beta^T z_i) - \sum_{i=1}^{n} \delta_i ln(\sum_{j \in R(t_i)} exp(\beta^T z_j))$ .So score function is $U_\beta = \frac{\partial l}{\partial \beta} = \delta^T Z - \sum_{i=1}^{n} \delta_i \frac{\sum_{j \in R(t_i)} exp(\beta^T z_j) z_j^T}{\sum_{j \in R(t_i)} exp(\beta^T z_j)}$.In real framework,we can solve it numerically.
Here,various tests are applied in evaluating different models. First: wald test is to test a hypothesis about a subset of the parameter β,the null hypothesis $H_0 : \beta_1 = \beta_{10}$ where $\beta = (\beta_1^T, \beta_2^T)^T$ and here β_1 is $q \times 1$ vector and β_2 is $(p - q) \times 1$ vector.For we know that $(\hat{\beta} - \beta)^T I^{-1}(\hat{\beta})(\hat{\beta} - \beta)$ follows $\chi^2(p)$ distribution, then the test statistic now for the subset become: $\chi^2_{subset} = (\hat{\beta}_1 - \beta_{10})^T [I^{11}(\hat{\beta})]^{-1}(\hat{\beta}_1 - \beta_{10})$,it follows χ^2_q distribution and we reject the null hypothesis when $\chi^2_{subset} > \chi_q(\alpha)$ in significance level α.Usually we often denote β_{10} as 0 vector and in real problems we often test the global assumption: $\beta_1 = \beta_2 = .. = \beta_p$. Other tests include likelihood ratio test and score test. The likelihood ratio test here is also applied in evaluating difference among models. The null hypothesis is $H_0 : \beta_k = 0$,the alternative hypothesis is $H_a : \beta_k \neq 0$(the variable added in the full model from the reduced model) The test statistic is $L = -2 \times log-likelihood(reduced\ model) - (-2 \times log-likelihood(full\ model)$, which asymptotically follows χ^2_{k-1} distribution. Then if $L > \chi^2_{k-1}(1 - \alpha)$ we reject the null hypothesis which means the full model may be correct.If we want to test the whole model's validity,we can also apply the same test. For the score test,it is also called the logrank test and its test statistic is $T_{SC} = U_1(\hat{\beta}_0)^T I^{11}(\hat{\beta}_0) U_1(\hat{\beta}_0)$ here $U_1(\beta)$ is the subvector of first q elements of the score function $U(\beta)$ and $\hat{\beta}_0 = (0^T, \hat{\beta}_2^{\ T})$ is the mle of β under null hypothesis,and in null hypothesis,the test statistic follows χ^2_q distribution.
Here we use the *coxph* function in package *survival* to fit the model and get the coefficients and relevant testing statistics.
Now let's fit start to fit the model: first we have the null model which only include the intercept: β_0 in $e^{\beta H_l}$ and a full model which include all of the linear terms and the interaction terms among these covariates(age,stage,diagyr) which equivalent to the full model. Then we apply *step* function in r to choose the model,this function indeed choose AIC criterion which denoted as: $-2 \times log - likelihood + 2 \times n_{par}$ which combines the model complexity

and goodness of fit on the data. We find that the model we find via *step* function including *stage,age,diagyr and age:diagyr*.We denote this as *model*1 1. Naturally we consider another model with no interaction terms which also denoted as first-order model. We denote this as *model*2.
Then we fit the full model with 2 order(including interaction terms) and then we apply model selection technique AIC(Akaike Information Criterion)(we choose not to use BIC for BIC often returns to a smaller model and may lose some importance information) to select the best submodel to fit the cox regression model.We denote our this chosen model as model1.
Then if we are interested in examine the effects of diagyr on survival time, we fit the first-order model here with age,stage(II,III,IV) and diagyr as predictor variables.However,we can see that the pvalue for the diagyr variable is 0.80,which is pretty large. So this variable is not significant and this is pretty consistent with our common sense which may because there won't be any significant change in the environment in this period of time which will impose such a great influence in the survival time of patients with all ages with various stages of the cancer diagnosed.The coefficients are shown as below:

Variables	Coefficients	P value
Age	0.01869	0.1922
factor(stage)2	0.15164	0.7442
factor(stage)3	0.64473	0.0703
factor(stage)4	1.73211	7.09e-5
diagyr	-0.01819	0.8120

This indeed coincides with our previous tests which indicates that there's no difference of survival time among diagyr(CMH test).So we just ignore this model.
Then after deleting the diagyr variable from the model, we naturally fit the model with age and stage which appear at the top of this subsection($Z_1 - Z_4$). We denote this as model3 this is a model we are really interested for we know that the most important variables influencing the survival time are age and different stages. Then another model we may interest in is only to include the stages covariate in the model which helps us explore how the stage of the cancer influence the survival time exactly. We denote this as model4 and its coefficients is shown as below:

Variables	Coefficient	Standard Error	p value
Z1:Stage II	0.06481	0.45843	0.8876
Z2:Stage III	0.6218	0.35519	0.0835
Z3:Stage IV	1.7349	0.4194	3.52e-5

From this model we can see that the relative risk of stage II over stage I is$exp(\beta_1) = e^{0.06481} = 1.06696$ while the relative risk of Stage III of the cancer over stage I is 1.8493 and the relative risk of Stage IV of the cancer over stage I is 5.688,we can see that in stage Iv of the cancer,the risk of death improves drastically than stage I while stage II 's relative risk over stage I is almost one which means there may not be large difference between this two stages.However,it

depends on another covariate age according to later model interpretation. Note that we can compare this with that in AFT model.
According to the previous exploration, there are certainly interaction effects between stage and age, so we should fit the model with stage(i-iv) and age and there interaction effect terms to examine the effect of different stages on age,we denote this as *model3.int*.Then the model has 7 parameters(age,stage2,3,4,three interaction terms). After examing the summary we can see that the interaction terms:*age:factor(stage)3* and *age:factor(stage)4* are not significant(pvalue are more than 0.7). So we should drop these two interaction terms. Then the model become *model3.updates*.Here we show the coefficient of the model for sake of interpretation:

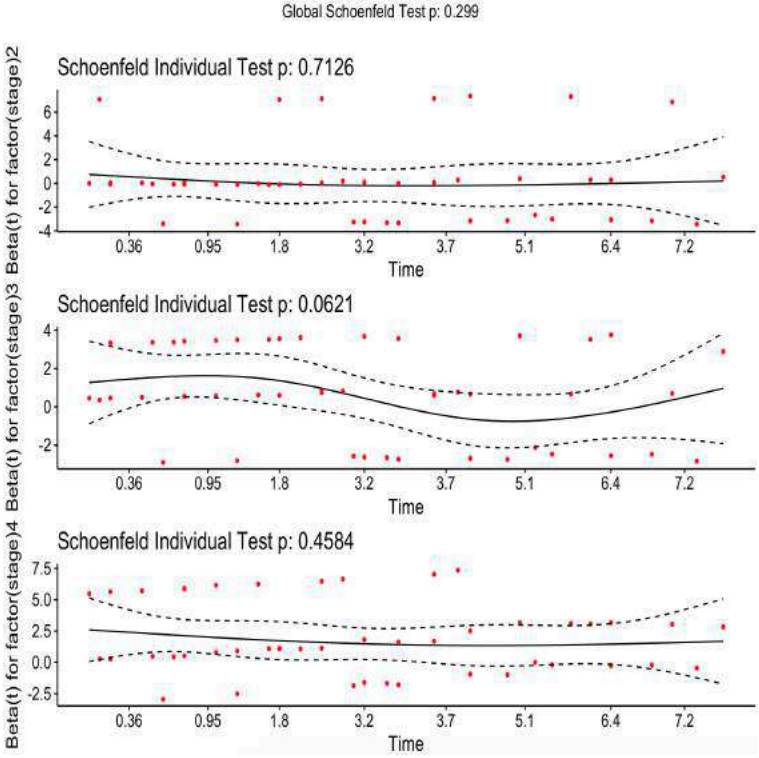
Variables	Coefficient	Standard Error	p value
Z1:Stage II	-7.3820	3.4027	0.03
Z2:Stage III	0.6218	0.3558	0.08
Z3:Stage IV	1.7534	0.4240	¡0.0001
Z4:Age	0.0060	0.0149	0.69
Z5:Z1× Z4(Interaction term)	0.1117	0.0477	0.02

Here we can see that that there is a significant interaction between age and stage II disease. Then the relative risk of dying for a stage patient of age Z_4 as compared to a stage I patient of the same age depend on the age. The relative risk is $e^{\beta_1+\beta_5 Z_4} = e^{-7.382+0.1117\times Age}$. For a 80 years old patient, the relative risk is 4.730 while for a 65 years old patient,the relative risk is 0.886 and for a 50 years old patient, the relative risk is only 0.17. Then with this model we can also test whether the risk of dying is the same for different ages. The null hypothesis here is $H_0 : \beta_1 + \beta_5 \times Age = 0$.To test this we use a technique very similar to the local wald test: the test statistic is $\chi^2_{RISK} = \frac{(\beta_1+\beta_5\times Age)^2}{V(b_1)+age^2\times V(b_5+2\times cov(b_1,b_5)}$ Under null hypothesis it follows χ^2_1 distribution. Then the test statistic for a 60 years old patient is 0.99 with a p-value 0.32 while for a 75 years old patient the test statistic is 4.09 with pvalue 0.04. This test do suggest that for young age there's little difference in survival between stage I and II patients while for old patients the stage II greatly effects the lifetime of the patients.
According to the previous part,the survival curves between different diagyr are not significantly different due to the CMH test(which is also validated with score test and local wald test,we fit the model only consisting of age and factor(stage). Finally, we choose model3 and model4 which reflects the survival time among different stages(model4) and different ages with stages(model3).The several tests we mentioned before's statistic are shown below in the table:

Models	Wald Test	Likelihood Ratio Test	Score test
Model 3	21.15	18.31	24.78
P value	0.0002958	0.001072	5.57e-5
Model 4	19.24	16.49	22.88
P value	0.0002433	0.0009016	4.28e-5

We can see that all of the tests indicates that the model is significant.Then we do the model diagnostics for the two models.

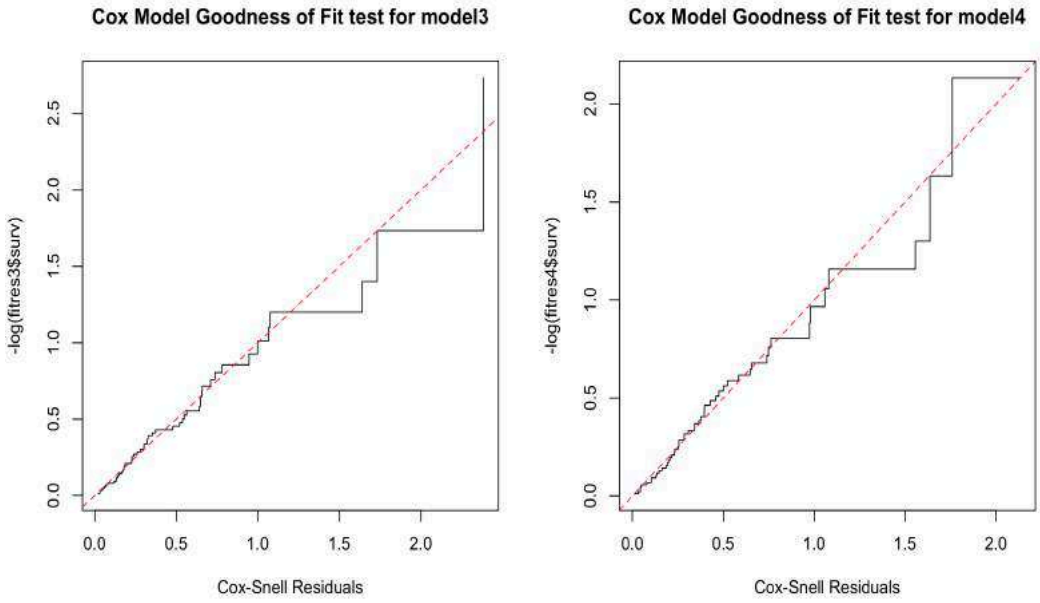
To diagnose the cox model we fit,we should first test whether it follows the assumption of cox-model.One assumption is the proportional hazards assumption which means the survival function of different groups are nearly proportional which means $\Lambda_l(t) = \Lambda_0(t)\theta$,it doesn't vary over time. Here we use the *ggcoxzph* function in *survminer* package to draw the residual vs time plot for the model3(with only stage).Here we use a graphical diagnostics based on scaled schoenfeld residuals. If the residual doesn't have obvious pattern over time, it may be consistent to the proportional assumption.The plot is shown as below:



We can see that there's no obvious pattern of these residuals over time. So the model is fairly consistent with the proportion hazard assumption.

Another diagnostic method is to do the goodness of fit test: according to the lecture, $r_j = \hat{\Lambda_0}(x_i)exp(\beta z_i)$ follows exponential distribution with parameter 1.This is because $P(\Lambda_0(x) > t) = P(x > \Lambda_0^{-1}(e^{\beta z_i}t)) = exp(-\Lambda_0(\Lambda_0^{-1}(e^{\beta z_i}t))) = e^{-t}$.Here $\hat{\Lambda_0}(x_i)$ is the baseline hazard rate. Here we use cox-snell residual to examine the fit of the model. Cox and Snell(1968) To check whether the r_j behave as a sample from a unit exponential,we can compute the Nelson-Aalen estimator of the cumulative hazard rate of $r_j's$(We use (r_j, δ_j) to fit the model and find the cumulative hazard function). If it fits well, this estimator should be approximately equal to the cumulative hazard rate of the unit exponential $\lambda_t = t$,which means $\hat{H}_r(r_j)$ versus r_j plot should looks like a straight line through the origin with slope 1.So this is indeed equivalent to the method in the lecture and it will be more convenient to test intuitively by the graph.

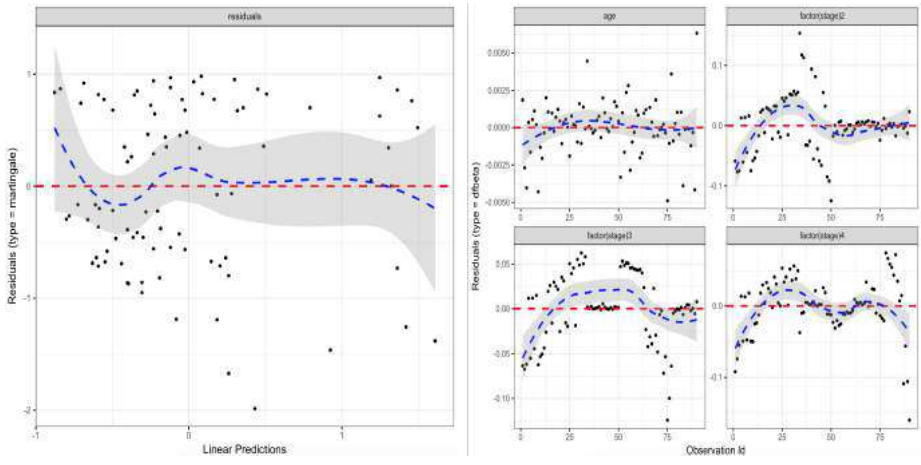
We examine this for our chosen model model3(w.r.t age and stage) and model4(w.r.t stage) and the plots are shown as below:



From the plot we can see that it's indeed a decent fit but there are some outliers(doesn't so badly) and model4 fits slightly better than model3.

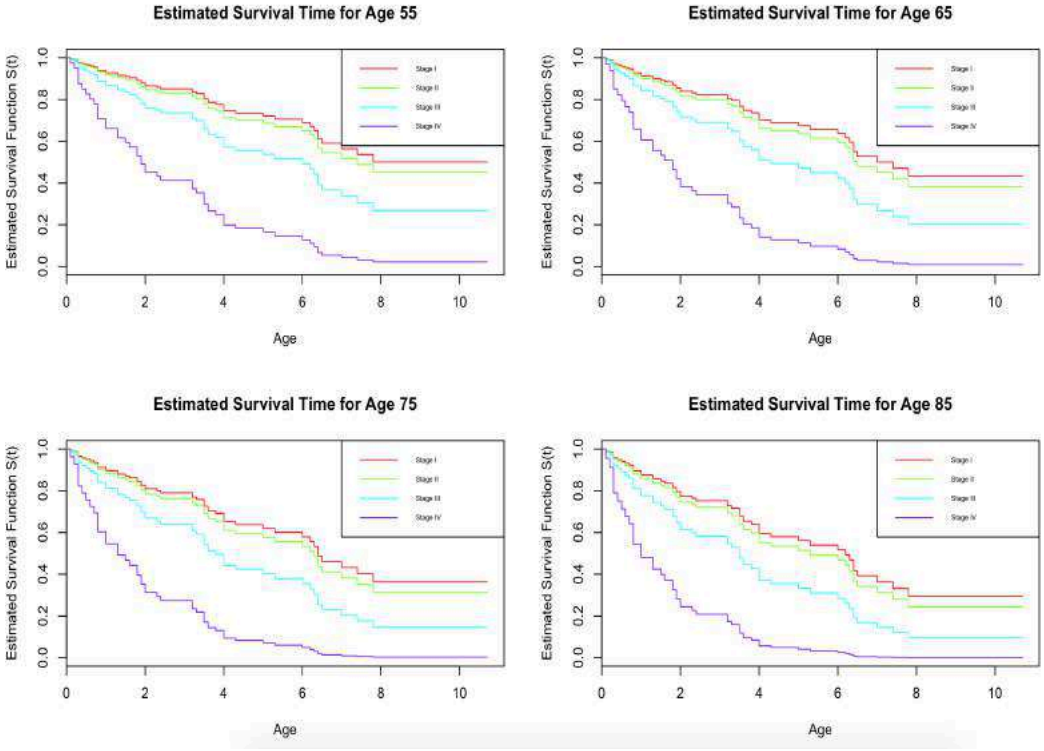
However,there are still various different methods to do the model diagnostics via residuals diagnostics: deviance residuals, martingale residuals. The deviance residuals can be used to test influential observations and we can use *ggcoxdiagnostics* function in *survminer* package to do the exploration.We do this for model3(with stage and age covariates).What's more,the martingale residual can be used to test the non-linearity and we won't implement it here.What's more,we can also visualize the dfbeta values which can plot the estimated changes in the regression coefficients upon deleting each observation in turn;Following are the two graphs of the diagnostics:

From the first plot we can see that the pattern looks not really symmetric around 0 and from



the second plot we can see that the some of the observations are influential but the dispersion are not extreme which means there are some outliers(especially for stage covariate) but these

won't make our modeling nonsense. Here the negative one means that the patient may live too long compared to their expected survival time and the positive residuals correspond to individuals that died too soon compared to expected survival time. So from this plot we can see that there are some patients lives much longer than they expected which indicates that there are certain amount of outliers(can't be perfectly predicted by the model).I have to say it's very usual in the practical case for the real data often doesn't looks really like what we derive in theory.That's one of the most important that I have learned from this project.
After modeling, we can do prediction on the survival time of a individual of a certain age: here we use the model3 to do it,we predict the survival time for individual aged 55,65,75,85 and the plot is shown as below:



From the model3 we can see that the estimated survival time for stage I cancer patients is $S_0(t)^{exp(0.019*age)}$, for stage II patient the survival time is $S_0(t)^{exp(0.019*age+0.140)}$,for stage III patient the survival time is $S_0(t)^{exp(0.019*age+0.6423)}$ while for stage IV the survival time become $S_0(t)^{exp(0.019*age+1.7059)}$. From the plot we can see that for every age, the survival time tends to decrease as the latter stage of the cancer diagnosed.What's more,the survival time for patients stage IV seems not very different among different ages,they are all much more shorter than the earlier stages. It's consistent with our guess and observation in preliminary analysis.

3.2 AFT model

Accelerated failure time model are alternative to relative risk models which are used extensively to examine the covariate effects on event times in censored data regression. The

model is: $T_i = e^{\beta z_i} U_i$ after log-transformation, the model become $logT_i = z_i^T \beta + U_i$,by re-parametrization,the equation become $Y_i = z_i^T \beta + e_i(\beta)$ Here β is an unknown $p \times 1$ vector of regression parameters and U_i are completely unknown and z_i is the covariate. So the cumulative hazard of T_i: $\Lambda_{T_i}(t) = \Lambda_U(e^{\beta z_i} t)$.In the presence of right censoring,the observed data are independent copies of (),where $Y_i = min(T_i, C_i), \delta_i = I(T_i < C_i)$ and $I(.)$ is the indicator function.

For the error terms' distribution is completely unknown, we have two methods to fit this model, the first is to assume that the error term follows some distribution and do the parametric modeling. The second is to fit the semi-parametric model based on the observed data and censoring status. Now first we assume that the error term follows certain distributions. From the above part we can see that the *diagyr* variable is not significant in regression. So we try to fit the model:$Y_i = log(X_i) = \mu + \sum_{i=1}^{4} \beta_k Z_k + \epsilon_i$ Here $Z_1, .., Z_4$ are defined as:

Z_1:1 if stage II cancer,0 otherwise;

Z_2:2 if stage III cancer,0 otherwise;

Z_3:3 if stage IV cancer,0 otherwise;

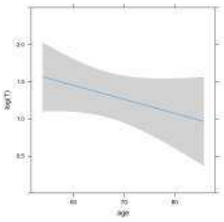
Z_4:Ages variable(denote the patient's age at diagnosis).

Empirically, the common used distributions include weibull,log-logistic,log-normal,rayleigh and exponential distribution. Then we use the *survreg* function in package *survival* to fit the parametric regression model with respect to each distribution of error terms. After fitting these models we get the log-likelihood of these models,for AIC=-2*(log-likelihood)+2*(number of predictor variables+number of parameters in assumed distribution). If we want to test the whole model's validity,we apply

Criterion,Distribution	Exponential	Weibull	log-logistic	log-normal	Rayleigh
log-likelihood	-141.9	-141.4	-141.6	-141.4	-155.3
AIC	293.8	294.8	295.2	294.8	322.6

Obviously,model with exponential distribution has the smallest AIC. So now our parametric regression model become: $Y = log(X) = 3.755 - 0.1456 Z_1 - 0.6483 Z_2 - 1.6350 Z_3 - 0.0197 Z_4 + \epsilon$. The chi-square test statistic is 18.44(indeed the wald test mentioned in part 3.1) and its p-value is 0.001 which means the coefficients(regression relation) are significant.In terms of interpreting this model: we can see that using the accelerated failure-time model for the exponential model, we see that the acceleration factor for stage II is $exp^{(-(-0.1456))} = 1.1567$,the acceleration factor for stage III of cancer is $exp^{(-(-0.6483))} = 1.9122$,the acceleration factor for stage IV of cancer is $exp^{(-(-1.6350))} = 5.1294$.It's a little bit similar to the relative risk in the cox model(model 4) This may suggests that the survival time for stage I patients is about 5.13 times than stage IV patients,1.91 times that of Stage III patients,1.12 times that of Stage II patients.So individuals with stage II,III,IV cancer tend to have shorter life times than individuals with stage I cancer. This is consistent with our exploration in preliminary analysis. However,there may be some limitations among this method:the exponential distribution doesn't fit our model very well, we can see this from the following plot(using *fitdist* function in r to fit the best possible distribution).We can see that the distribution fitted doesn't make any sense. So here the parametric regression model may not fit very well for this dataset although we may get similar conclusion like in the previous part of cox model fitting and preliminary

exploration.
Secondly we apply the semi-parametric technique based on the observed data when we have totally no information about the residual term.We use the R package *aftgee* proposed by Sy Han Chiou and lss function in R package *lss* to help our modeling. The *aftgee* package implements the rank-based procedures and least-square estimators based on the semi-parametric model which greatly alleviate the limitations of usage of Semi-parametric AFT model.(Based on the GEE non-parametric technique).Here the function's implementation also account for multivariate dependence through working correlation structures to improve efficiency.A well known method of estimating T_i is the Buckley James estimator(1979). $\hat{Y_i}(\beta) = \delta_i Y_i^{'} + (1-\delta_i)[\frac{\int_{e_i(\beta)}^{\infty} u d\hat{F}_\beta(u)}{1-\hat{F}_\beta(e_i\beta)} + Z_i^T\beta]$. We do the buckley james estimator for the survival time vs age, diagyr and stage and graph of log(T) and age is shown as below:



The coefficients are shown as below:

Model	Coefficients	P value(wald test)
Age only	-0.0167	0.1942
Stage only	-0.5832	¡0.0001
diagyr only	-0.0621	0.3781

Here to examine the significance,we use wald test(similar to that in cox model).The estimator of age are only slightly different from that in the cox model and the ordinary AFT model. However,the r function bj is very limited here because we can't even plot the log(T) vs diagyr and stage. So this is what greatly hinge the use of this technique greatly.
From the plot we can see that the survival time tend to decrease as the age increases(it is consistent to our preliminary analysis). Here we plan to fit two accelerated failure time models: with stage and age covariates(like model3 in subsection 3.2),only with stage covariate(like model4 in subsection 3.2).
We apply the lss function in R package *lss* to fit the AFT model and compare this to that in *aftgee* function,here we indeed use the least square method.The model we want to fit become: $log(X_i) = \sum_{i=1}^{k} \beta_i Z_i + \epsilon_i = \sum_{i=1}^{k} \beta_i Z_i + log(U_i)$. After fitting this model we should check the model's adequacy. To check the model's adequacy,we use goodness of fit test here.We know that if the model fits well the residual term U_i follows unit exponential distribution(with parameter 1).To test whether the term U_i follows the unit exponential distribution, we can first use the formula $A = \frac{\sum_{i=1}^{k}(O_i - E_i)^2}{E_i^2}$,k is the number of breaks we choose for convenience, we know that if the null hypothesis is true, the statistic A follows χ^2_{k-1} distribution. For the model1.1,we can see that the test statistic is 1.4 and we absolutely accept the null hypothesis so we claim it fits the unit exponential distribution well.So model1.1 is adequate. Then

for model1.2,for we can see that there are several severe outliers then the test statistic is pretty large and we reject the null hypothesis thus we claim that model1.2 may not a good model.This may because some regression coefficients of model1.2 are not significant(the p value for stage 2 is 0.804) and the algorithm may not accurate enough in terms of this specific structure.However,we will still see the model's parameter and try to extract some information from them.
When we fit the model with stage only(we denote here as model 1.1),the parameter is shown as below:

Least Square estimator for model1.1	Estimate	Std.Error
Stage II	-0.247	0.4334
Stage III	-0.9466	0.369
Stage Iv	-1.89	0.43
Aftgee estimator for model1.1	Estimate	Std.Error
Intercept	2.086	0.227
Stage II	-0.247	0.441
Stage III	-0.946	0.385
Stage IV	-1.895	0.465

We can see that the estimator for both techniques are roughly the same.So the algorithm is pretty consistent in this model.(This two methods works equally well). In terms of interpretation of this model we can see that there appears to be accelerated failure(death) time on later stage(II,III,IV) compared with stage I. The acceleration factor for stage II is $e^{-(-0.247)} = 1.280$,for stage III is $e^{-(-0.946)} = 2.576$ while for stage IV is $e^{-(-1.895)} = 6.651$. It's different from that in the parametric regression model and this accelerated factor should be more accurate than the previous one. When we fit the model with stage and age(we denote here as model 1.2),the parameter is shown as below:

LSS Square estimator for model1.2	Estimate	Std.Error
Age	-0.0393	0.0384
Stage II	-0.461	1.217
Stage III	-2.571	1.091
Stage Iv	-4.138	0.907
Aftgee estimator for model1.2	Estimate	Std.Error
Age	-0.030	0.023
Stage II	-0.133	0.536
Stage III	-1.054	0.456
Stage IV	-2.070	0.523

There are some different in parameter estimation between this two methods but the difference may not extreme. We can see that the coefficient of age is larger than that in the cox proportional hazard model. What's more, we can see that the different ages will surely influence the survival time of the patients in different stages. So it naturally indicate us to add some cross-product terms to represent this kind of interaction terms.

Then we try to fit the AFT model with interaction terms between age and stage. However, both methods in r not work well. The model fitted by *aftsrr* function are not reasonable for most of the regression coefficients are not significant while the model fitted by *lss* doesn't converge in 1000 iterations. So it requires further research and investigation in actually proposing a more robust algorithm to implement this model fitting with various interaction terms or other more complex structures.

3.3 Additive Hazard model

Here we point out an alternative method to compare the four stages of laryngeal cancer, the additive hazard model. The additive hazard model is an alternative to the semi-parametric multiplicative hazard model.We want to express the hazard rate and the cumulative hazard as a(t)+X(B(t)) and a and X are both time-dependent. We have an event time X whose distribution depends on a vector of covariates which may be time-dependent and we denote this as $Z(t) = [Z_1(t), ..., Z_p(t)]$. We assume that the hazard rate at time t for an individual with covariate vector $Z(t)$ is a linear combination of the $Z_k(t)$'s then the model become: $h[t|Z(t)] = \beta_0(t) + \sum_{k=1}^{p} \beta_k(t) Z_k(t) + \epsilon$.We can see that this is indeed a regression model too and the error term's distribution is totally unknown which indicates that it's a semi-parametric model.

Here we consider the model with covariates same as the model3 in subsection3.1:note here we center the age covariate at its mean.

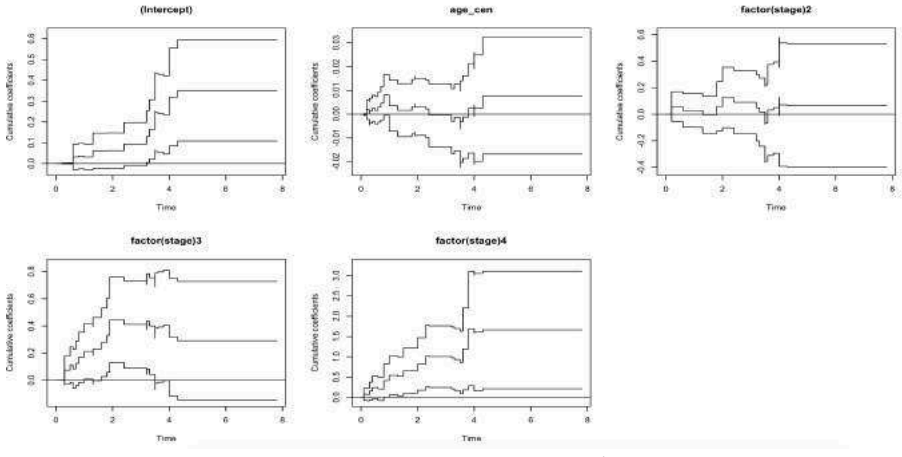
Z_1:1 If stage II of the disease,0 otherwise.

Z_2:2 If stage III of cancer,0 otherwise

Z_3:3 If stage Iv of cancer,0 other wise.

Z_4:Age at diagnosis

Here we can get the excess risk of stage II,III,IV cancer as compared to stage I and we can also see a 95% pointwise confidence interval for the patients.The plot is shown as below:

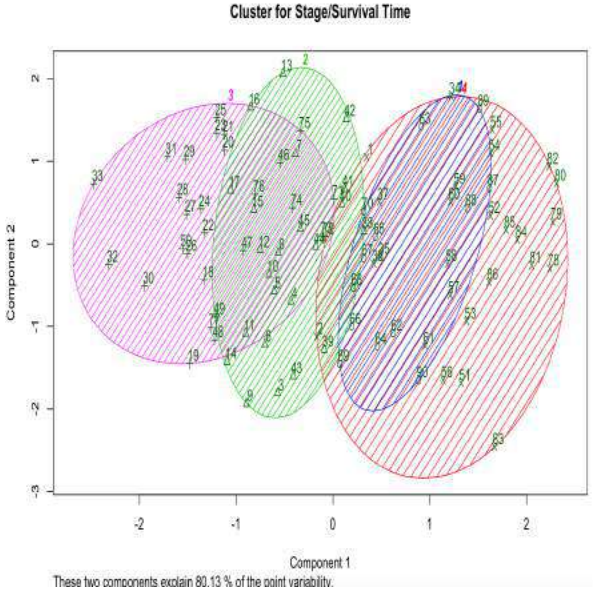


This figure show the estimate of $B_k(t) = \int_0^t \beta_k(u)du, k = 0, 1, 2, ..., p$,the first for intercept is indeed the baseline hazard. Here the baseline hazard is an estimate of the cumulative hazard rate of stage I patient in the mean age(64.1 years old). We can see from the plot that there

appears little excess risk of stage II patients while for stage III and stage IV patients, they tend to have an much more risk in the first two years of diagnosis and the excess risk tends to be invariant after 2 years of diagnosis of the cancer.What's more,the age plot shows that the excess risk has relation with age but the relation seems not so significant.What's more we can also test whether there are no difference in survival time between four stages of larynx cancer adjusting for age.(Similar to wald test and we won't implement this) instead of observing the plot intuitively.
What's more,there are so many related topics we can do with this model. For example we can test $H_0 : \beta_k(t) = 0$ for all $t \leq \tau$ which means the covariate is zero before a certain time. Various research have been conducted among these. What's more,there is another kind of additive hazards model proposed by Lin and Ying which replace the time-varying regression coefficients in the Aalen model by constants.The further derivation of the parameter estimation and model diagnostics are somewhat based on non-parametric techniques.

3.4 Advanced Topics

Now apart from the two most important models mentioned in this course,we plan to do some advanced analysis on this dataset.
I:Clustering Analysis: we want to see whether there are significant difference in pattern among the four stages of the cancer.We would like to validate whether the patients of four stages of the cancer are distinguishable.(With respect to the Here we use the hierarchical clustering based on euclidean distance matrix and kmeans technique. The number of clusters are 4. Then we draw the clusplot via *fvizcluster* in *factoextra* package to show whether the patients of four stages of the disease are distinguishable and the cluster plot is shown as below the points are labelled as their



From the plot we can see that the patients are not perfectly distinguishable which means

there's no obvious pattern between survival time and the stage however the survival time for patients in different stages are different.It indeed tells us that all of our modeling with making the stage as a categorical variable seems reasonable.
II:

4 Conclusion and Discussion

In this project we use the larynx project to do survival modeling.Firstly,we observe that there are obvious difference of survival time among different stages of cancer and we conclude that later the stage of the cancer,the shorter will the patients survival.We conclude that there are interaction effects between age and stage in explaining survival time.
In terms of limitations, here the distribution of survival function is fairly weird and the parametric techniques doesn't work well. Maybe it's fairly normal in real life dataset and that's we are passionate on the non-parametric techniques with foundation of empirical process.
In cox model fitting we do the model fitting,model selection and model diagnostics and we have found that the the difference of survival time between stages are affected by ages(interaction) which indicates that for younger age there's little difference in survival between stage I and stage II but for old people, the difference will be more significant while for stageIII vs stageI and stageIV vs stageI the effect of age is not that significant. It is indeed inspiring for the scientists to do more research to extend the survival time of older patients with later stage of cancer.What's more,in the part of prediction, we see that the survival time for stage IV(last stage)
In the AFT model fitting we can see that the accelerated factor estimator will be more accurate in using non-parametric techniques for none of the parametric techniques work well based on our exploration of the dataset.
From the clustering part we conclude that the patients of four stages can't be perfectly distinguishable and I believe that further research will also help us identify the most influential effect on survival time for the different patients.
In terms of limitations, here the distribution of survival function and the minimum chi-square technique not makes sense too.Alternatively,we use *fitdist* function in r to try to approximate a reasonable distribution but it still doesn't work very well which means the parametric techniques won't work well. Maybe it's very normal in the real-life data analyzing.The parametric model is limited to a very small part of cases and for most of the time,applying non-parametric techniques will help us more.What's more, the current implementation of semi-parametric AFT model and the buckley-james estimating function *bj* in r is not really robust to all kinds of linear models(the algorithm even not converge ever when we add a interaction term of stage and age) so it greatly limit our analysis and interpretation in AFT modeling.
Survival analysis and modeling survival data is such a broad topic and there are so many kinds of models and lots of research has been conducted in the past 40 to 50 years and although we can't exhaust all methods in one project because of the restriction of the data, I still find there are a lot to explore further in this subject.

5 Acknowledgements

First I would like to express my gratitude to Prof Fushing. In this quarter,he not only tells us how to model the survival data but also tells us how to explore a certain topic this quarter.His lectures always make the somewhat very tricky problems much easier to deal with.His passion in the related research also greatly moves me and I can see that Prof Fushing hope everyone can get a good understanding of the course.
Then I would like to express my thankfulness to Mrs Roy Tania. She is really patient and always willing to help us and her discussion session every week tells us about how to exactly implement modeling in r and it helps me a lot.
What's more, I would like to express my sincere thankfulness to everyone writing R packages. For example, the *survminer*, *ggplot2* packages enables me to draw such an elegant and clear plot via functional plotting.I won't have been here without them!What's more,the *lss* package allows us to implement the so complex AFT model in real life and makes life much more easier.
Finally, I have to express my gratitude to D.R.Cox, in my opinion he is a great statistician and his intuition in constructing proportional hazard model inspire me in how to conduct research.

6 Bibliography

Cox,D.R.(1972).Regression models and life tables(with discussion).Journal of Royal Statistical Society Serie B,34,187-220
Anderson,P.K.,Gill,R.D.(1982).Cox's regression model for counting process:A large sample study. Annals of Statistics,10,1100-1120.
Aalen,O.O.,Gjessing,H.K.(2001).Understanding the shape of the hazard rate:A process point of view.Statistical Science,16,1-22
L.J.Wei,D.Y.Lin,L.Weissfeld(1989). Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions. Journal of the American Statistical Association,Vol.84, No.408.(Dec,1989),pp.1065-1073.
David M Diez(2013).Survival Analysis in R.OpenIntro.org
Sy.Han Chiou,Sangwook Kang,Jun Yan(2014).Fitting Accelerated Failure Time Models in Routine Survival Analysis with R Package aftgee. Journal of Statistical Software,Vol 61,Issue 11,Nov,2014.
John Fox,Sanford Weisberg(2013). Cox Proportional-Hazards Regression for Survival Data in R. An Appendix to *An R Companion to Applied Regression, Second Edition*
Terry Therneau, Cynthia Crowson, Elizabeth Atkinson. Using Time Dependent Covariates and Time Dependent Coefficients in the Cox Model. 2017
D.R.Cox,E.J.Snell.A General Definition of Residuals(1968). Journal of the Royal Statistical Society.Series B. Vol.30,No.2(1968),pp.248-275
Zhezhen Jin,D.Y.Lin,Zhiliang Ying.On least-squares regression with censored data. Biometrika(2006),93, 161

STA 207 Project:Determinant of Crime rate in America

Heqiao Ruan
hruan@ucdavis.edu
Rafee Musabbir
rafeemusabbir@ucdavis.edu

March 21, 2018

1 Introduction

In this project we plan to do extensive data analysis by stepwise regression, ridge regression, lasso and glmnet and cross validation and PLS to perform the analysis on the crime dataset which depicts the crime-related statistics for 47 US states in 1960 and the goal is to relate crime rate to the other socio-economic variables. Here we use various approach to conduct the analysis.

2 Introduction to dataset

The dataset is crime-related statistics for 47 US states in 1960 are given and there are 47 observations and 13 features.
Column 1:Crime rate, number of offenses known to the police per 1000000 population.
Column 2:Age, number of males aged 24-24 per 100 of total population.
Column 3:Ed, mean number of years of schooling times 10 of the population, 25 years or older.
Column 4: PE, police expenditures- per capita expenditure on police protection by state and local governments in 1960.
Column 5: PE-1, police expenditures- per capita expenditure on police protection by state and local governments in 1959.
Column 6: LF, labor force participation rate per 1000 civilian urban males in the age group 14-24.
Column 7:M, number of males per 1000 females.
Column 8: Pop, state population size in 100000.
Column 9: NW, number of nonwhites per 1000.
Column 10: UE1, unemployment rate of urban males per 1000 in the age group 14-24.
Column 11: UE2, unemployment rate of urban males per 1000 in the age group 35-49.
Column 12: Wealth, median value of transferable goods and assets or family income(units 10 dollars).
Column 13: IncIneq, income inequality- number of families per 1000 earning below one-half of the median income.

3 Preliminary Analysis

First the response is crime rate and for detailed description of the variables in the dataset, see **Appendix A**. First we want to see the summary statistic of the dataset including and shown in **Table 1**. What's more, we plot the histogram of every variable and try various transformation techniques:**square, square root, log transformation** and choose the one that most like normal distributed, what's more log transformation is preferred in the response variable **Crime**(Graph 2.V)
Then from the plot we can see that log transformation is preferred for **UE.2,PE.1 and NW** while squareroot transformation is preferred for **POP** and square transformation is preferred

for **UE.1**. What's more we perform standardization to each variables so that everyone of them is in the same scale.
Then we explore the inter-correlation among the variables**(Graph 3)**. We can observe correlation present in this dataset, **wealth and PE,PE.1,ED** are highly correlated,**PE and PE.1** are highly correlated,**UE1 and UE2** are highly correlated. What's more we can see that the variance inflation factor of **(Table 4(I))** PE and PE.1 and Wealth are larger than 10 which means somewhat serious multicollinearity.
So as we can see that the correlation between PE and PE.1 are almost 1, we would like to choose only one of them in our downstream analysis and use PE.1 in our analysis then after dropping this we can see that no severe multicollinearity appear.**(Table 4(II))**.
Then we draw the pairwise scatterplot**(Graph 16)** and perform the lm model on the original data and the diagnostic plot is shown in **(Graph 17)**.

4 Model Building

4.1 Stepwise regression

Here we start from the full model and use **StepAIC** function to perform the model selection and see whether the model perform well.
Then from **Graph 5(I)** we can see that there's no obvious nonlinear pattern in the residual plot and the q-q plot shows no obvious dispersion from normal distribution and the summary of the model selected by stepwise regression is shown as **Graph 6**.
The model we select here is $log(crime)^* = -1.81 * 10e-15 + 0.367 * Age^* + 0.496 * Ed^* + 0.866 * log(PE.1)^* + 0.204 * log(UE2)^* + 0.403 * Wealth^* + 0.947 * IncInequ^*$, here $*$ means standardized variables.

4.2 Ridge regression

Here we use a shrinkage method using L2 penalty to alleviate the multicollinearity. After using the GCV criterion and cross validation, we get the optimal parameter k=1.2548**(Graph 7)**.
Then here from **(Graph 8)** we can see that there's no obvious nonlinear pattern in residual plot and only slight light tail appear in the qqplot which means the model assumption is decent.
Then the coefficient and estimation table is shown in **(Graph 9)** and the variance inflation factor is shown in **(Graph 10,Graph 5(II))** we can see that it is smaller than that in the stepwise regression and we can see that the ridge regression somewhat alleviate the multicollinearity.
Our final model get by ridge regression is $log(Crime)^* = 0.327 * age^* + 0.441 * Ed^* + 0.690 * log(PE.1)^* + 0.879 * LF^* - 0.0157 * M - 0.0569 * sqrt(Pop)^* + 0.187 * log(NW)^* - 0.131 * UE1^{2,*} + 0.324 * log(UE2)^* + 0.258 * Wealth^* + 0.589 * IncInequ^*$

4.3 Lasso

Here we use another regularization technique called lasso to do the variable selection and shrinkage. Then we use 10 fold cross validation and the MSE versus log-lambda curve is shown in **Graph 11**.Then our fitted model here become: $log(Crime)^* = 0.244 * age^* + 0.33 * Ed^* + 0.769 * log(PE.1)^* + 0.06 * LF^* + 0.195 * log(NW)^* - 0.013 * UE1^{2,*} + 0.166 * log(UE2)^* + 0.363 * IncInequ^*$ see also **Graph 12**.
Then we perform model diagnostic on lasso and we have seen that there are somewhat nonlinear pattern in the residual plot and the qq plot shows a little bit dispersion from the normal distribution. So here lasso may not be the optimal choice.

4.4 Partial Least Square

Here at first we apply PLS(partial least square) to build the model with 8 components and check the F statistic. Then we calculate the adjusted coefficient of determination and coefficient of determination and CV and the F statistic. The table is shown as below:

k	1	2	3	4	5	6	7	8	9	10
SSE	25.22	16.79	14.17	11.80	10.99	10.13	9.76	9.64	9.62	9.61
R-square	0.452	0.635	0.692	0.743	0.761	0.780	0.788	0.790	0.7908	0.7909
F(k)	37.08	22.07	7.97	8.42	3.02	3.40	1.46	0.490	0.073	0.0147
qF	4.057	4.062	4.067	4.073	4.079	4.085	4.091	4.098	4.105	4.113
CV	0.8062	0.7261	0.6958	0.6773	0.6758	0.6548	0.6457	0.6502	0.6512	0.6
Varpro	0.290	0.499	0.658	0.733	0.862	0.919	0.937	0.955	0.976	0.987

Then we can see that based on the CV criterion, we should choose 7 components which can explain 93.7% of the total variance and the loadings of the first 7 components is shown as **(Graph 14)**. Then our model here become

4.5 Model Selection

Here we use the 10-fold cross validation method(7 of the 10 samples are 5 while the other 3 samples are 4) and compare ridge, lasso, partial least square, stepwise regression. Here the CV error is given by $CV^{(-1)} = \frac{1}{n}\sum_{j=1}^{n}\sum_{i\in I_j}(Y_i - x_i^T\hat{\beta^{(j)}})^2$ and we want to choose the model with the smallest one which means the best fitting and predictive ability for the downstream analysis.
Here the CV error are shown as below:

Type	Stepwise Regression	Ridge	Lasso	PLS
MMSE	0.6289	0.3287	0.4210	0.6038

So here we find that the ridge regression is the best choice in terms of CV criterion.
Then our final model is the same as what we choose in ridge regression: $log(Crime)^* = 0.327 * age^* + 0.441 * Ed^* + 0.690 * log(PE.1)^* + 0.879 * LF^* - 0.0157 * M - 0.0569 * sqrt(Pop)^* + 0.187 * log(NW)^* - 0.131 * UE1^{2,*} + 0.324 * log(UE2)^* + 0.258 * Wealth^* + 0.589 * IncInequ^*$

5 Conclusion and Discussion

Here we have concluded that the ridge regression is the optimal one among the four types of fitted models with the best predictive ability. Here after comparing the sign of the coefficients in several models, we can see that they are consistent among these models. Then we can see that the crime rate is positively related to age, Ed, PE.1, LF, NW, UE2, Wealth, IncInequ and is negatively related to Pop, M and UE1.
Here we interpret this relationship. First the crime rate is positively related to the ratio of young men in the whole population and it matches our common sense. Then we can see that the crime rate is positively related to the police expenditure on police protection but it may not be a causal however, as we all know, higher crime rate will lead to higher police expenditure for police protection. Then we can see that the relation between crime rate and M,LF,Pop are not that significant(same scale, compare the coefficient), however we can see that the crime rate is negatively related to the population size and the number of men in the whole population and are positively related to the labor force participation rate per 1000 civilian urban males in the age group 14-24. Then we can see that crime rate is positively related to the number of nonwhites in the population which may somewhat match our common sense. What's more, the crime rate is positively related to the unemployment rate of urban males older than 35 which matches our common sense because men at that age are supposed to work. However, the crime rate is negatively related to the unemployment rate of urban males aged between 14 and 24 which is also explainable because young men at that age are supposed to be in the school and if they work when 15 or 16 years old, probably they have not received a decent education which potentially lead a higher crime rate.
What's more, the crime rate is positively related to the median value of transferable goods and assets or family income which is pretty interesting because we have found that even wealthy men tend to crime and we can see that the desire of some people can never be fulfilled.
Then we can see that the crime rate is positively related to the degree of income inequality which is pretty insightful because people are more worried about others being wealthier than themselves than worrying about their own poverty. So as government, we should provide help to those who fail to get a decent education.
So in conclusion, we can see that the crime rate are mainly related to the men's education level and the degree of inequality of economical status in this city as well as the unemployment rate of the men who are supposed to work to support their family in that age.
For the limitation we can see that even after transformation, there are still very slight dispersion from the normal assumption(either light or heavier tail). Another issue is that Lasso doesn't seem to be a particular optimal model here for the somewhat severe dispersion from the fundamental assumption. What's more, the CV criterion may be also limited in terms of comparing these kinds of models and other more delicate techniques are yet to be developed. The most important is that the number of observations in the dataset is not enough for us to conduct a very very compelling result and I think repeated measurements in the 47 states in US along the last a couple of years are required. For future work, maybe more delicate model selection techniques and longitudinal data analysis are required.

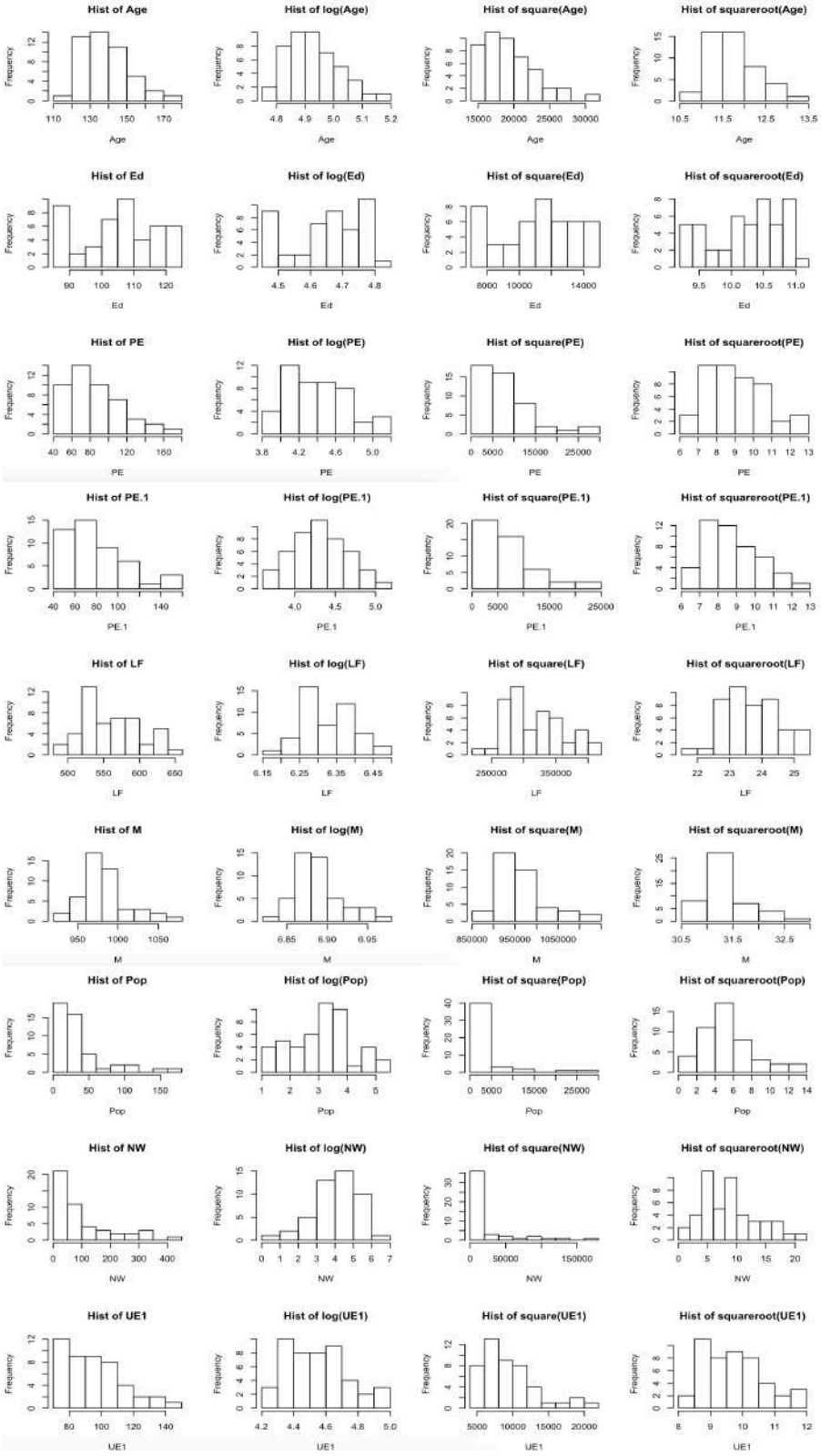
6 Appendix A:Tables and Graphs

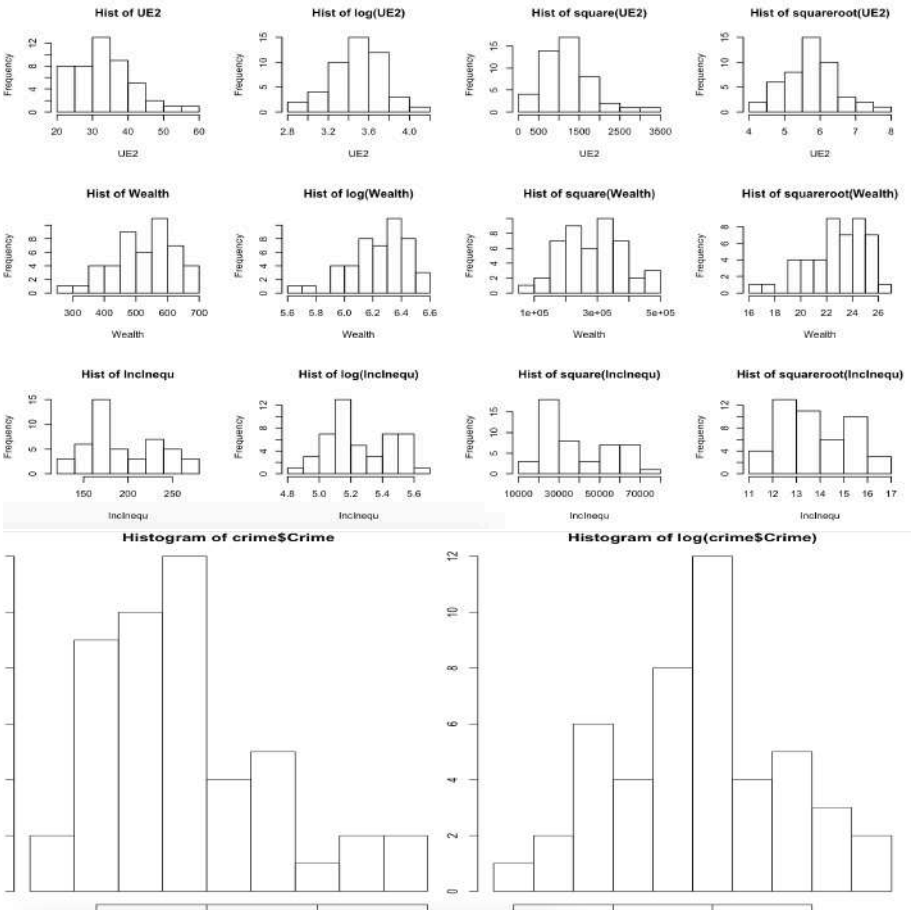
Table 1: Summary statistics:

	Crime	Age	Ed	PE	PE.1	LF	M	Pop
Min.	34.20000	119.0000	87.0000	45.0	41.00000	480.0000	934.0000	3.00000
1st Qu.	65.85000	130.0000	97.5000	62.5	58.50000	530.5000	964.5000	10.00000
Median	83.10000	136.0000	108.0000	78.0	73.00000	560.0000	977.0000	25.00000
Mean	90.50851	138.5745	105.6383	85.0	80.23404	561.1915	983.0213	36.61702
3rd Qu.	105.75000	146.0000	114.5000	104.5	97.00000	593.0000	992.0000	41.50000
Max.	199.30000	177.0000	122.0000	166.0	157.00000	641.0000	1071.0000	168.00000

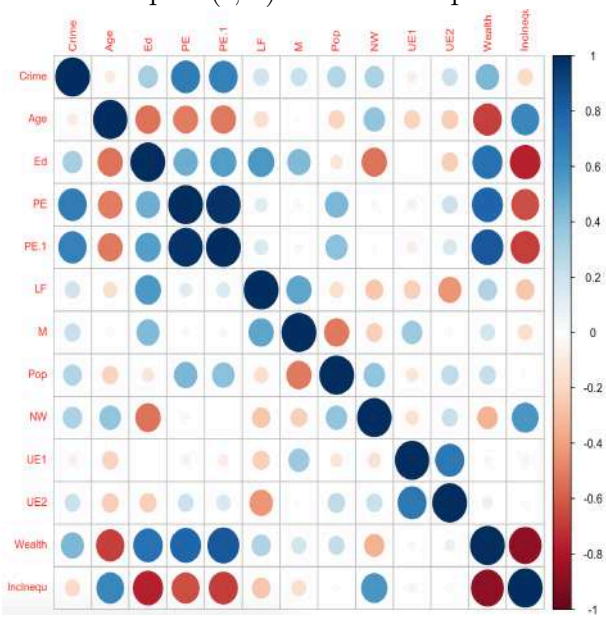
	NW	UE1	UE2	Wealth	IncInequ
Min.	2.0000	70.00000	20.00000	288.000	126.0
1st Qu.	24.0000	80.50000	27.50000	459.500	165.5
Median	76.0000	92.00000	34.00000	537.000	176.0
Mean	101.1277	95.46809	33.97872	525.383	194.0
3rd Qu.	132.5000	104.00000	38.50000	591.500	227.5
Max.	423.0000	142.00000	58.00000	689.000	276.0

Graph 2(I,II,III,IV,V):Histograms of variables:





Graph 3(I,II):Correlation plot:



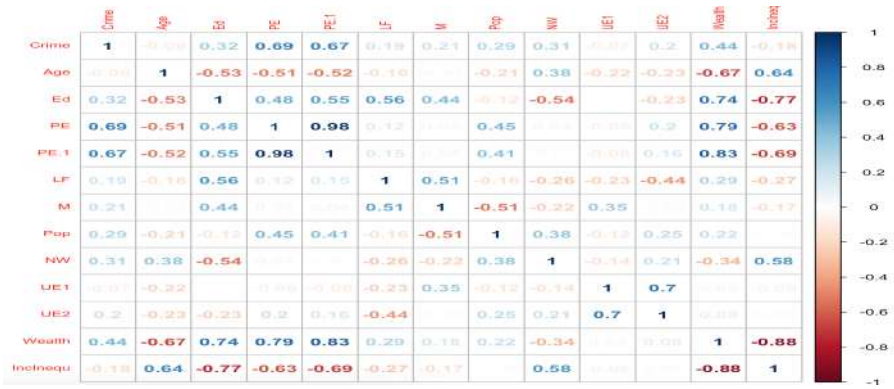
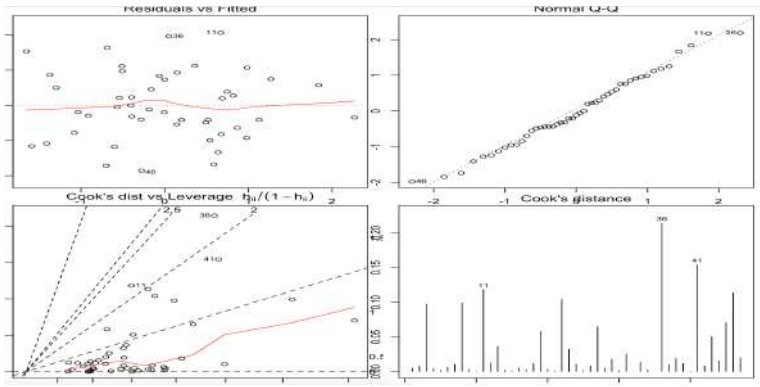


Table 4(I,II): Eigen values of the inverse of the design matrix:

	Variables	Tolerance	VIF
	<chr>	<dbl>	<dbl>
1	Age	0.410	2.44
2	Ed	0.203	4.93
3	PE	0.0405	24.7
4	PE.1	0.0330	30.3
5	LF	0.384	2.60
6	M	0.259	3.87
7	Pop	0.359	2.79
8	NW	0.322	3.10
9	UE1	0.236	4.24
10	UE2	0.244	4.10
11	Wealth	0.0968	10.3
12	IncInequ	0.106	9.40

	Variables	Tolerance	VIF
	<chr>	<dbl>	<dbl>
1	Age	0.415	2.41
2	Ed	0.206	4.85
3	PE.1	0.173	5.77
4	LF	0.385	2.60
5	M	0.272	3.68
6	Pop	0.380	2.63
7	NW	0.323	3.10
8	UE1	0.238	4.20
9	UE2	0.246	4.06
10	Wealth	0.0977	10.2
11	IncInequ	0.106	9.40

Graph 5(I): Diagnostics for Stepwise Regression:



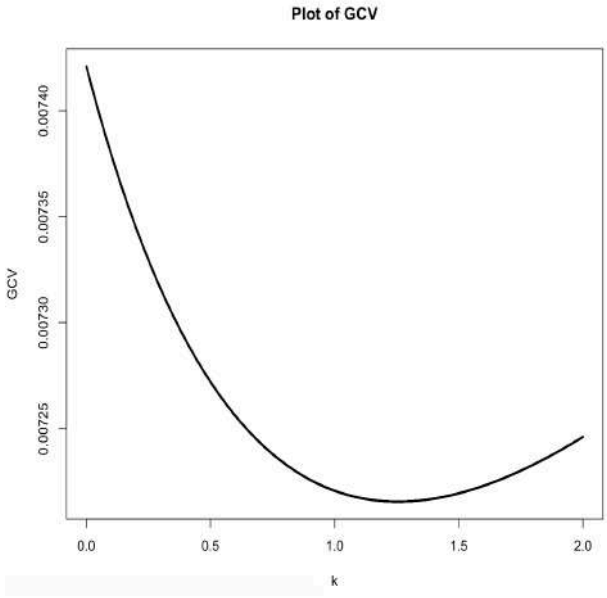
Graph 5(II): Variance Inflation Factor for stepwise regression.

	Variables	Tolerance	VIF
	<chr>	<dbl>	<dbl>
1	Age	0.481	2.08
2	Ed	0.326	3.07
3	PE.1	0.279	3.59
4	UE2	0.713	1.40
5	Wealth	0.107	9.32
6	IncInequ	0.182	5.51

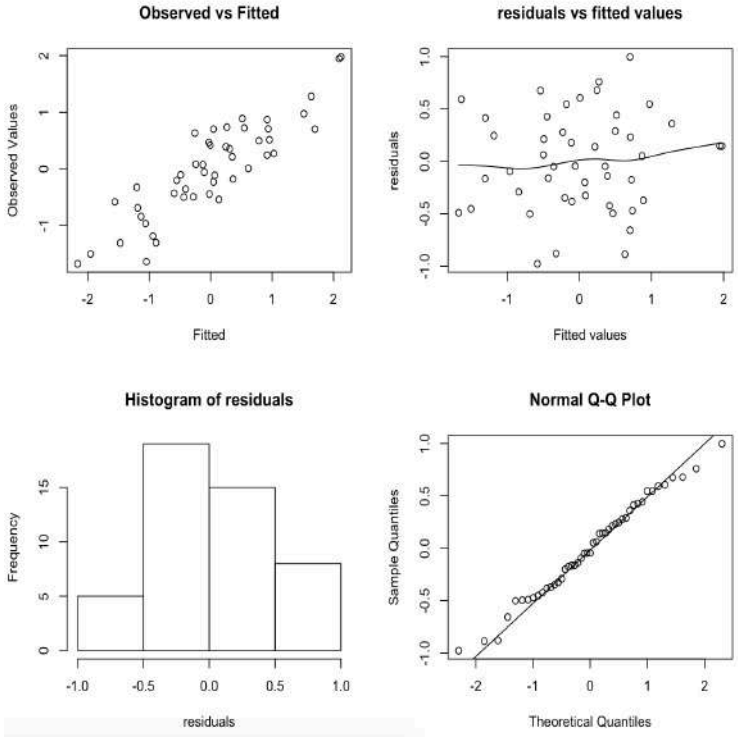
Graph 6: Summary of the selected model of stepwise regression:

```
Call:
lm(formula = Crime ~ Age + Ed + PE.1 + UE2 + Wealth + IncInequ,
    data = crime_std)

Residuals:
     Min       1Q   Median       3Q      Max
-0.92650 -0.29577 -0.05818  0.36837  1.02899

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.810e-15  7.496e-02   0.000 1.000000
Age          3.671e-01  1.093e-01   3.359 0.001729 **
Ed           4.963e-01  1.327e-01   3.741 0.000574 ***
PE.1         8.663e-01  1.435e-01   6.035 4.21e-07 ***
UE2          2.043e-01  8.972e-02   2.277 0.028184 *
Wealth       4.030e-01  2.313e-01   1.742 0.089175 .
IncInequ     9.471e-01  1.778e-01   5.326 4.17e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5139 on 40 degrees of freedom
Multiple R-squared:  0.7704,    Adjusted R-squared:  0.7359
F-statistic: 22.37 on 6 and 40 DF,  p-value: 2.344e-11
```

Graph 7: Ridge coefficient versus GCV:



Graph 8: Ridge regression diagnostics:



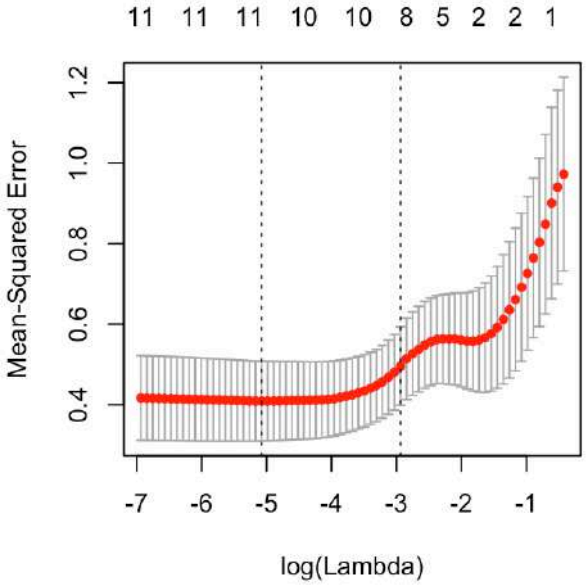
Graph 9: Ridge regression coefficient table and estimation:

	estimation	std_error
	-1.535339e-15	0.07548353
Age	3.270995e-01	0.11009174
Ed	4.412504e-01	0.14712647
PE.1	6.900085e-01	0.15057723
LF	8.788386e-02	0.11231313
M	-1.579924e-02	0.12989121
Pop	-5.696460e-02	0.11567178
NW	1.870250e-01	0.10813399
UE1	-1.311716e-01	0.11944912
UE2	3.236700e-01	0.12123126
Wealth	2.580289e-01	0.18064839
IncInequ	5.898477e-01	0.17399318

Graph 10: Variance inflation factor for ridge regression:

Age	1.9750571
Ed	3.5273766
PE.1	3.6947818
LF	2.0555652
M	2.7493482
Pop	2.1803440
NW	1.9054371
UE1	2.3250704
UE2	2.3949665
Wealth	5.3178780
IncInequ	4.9332675

Graph 11: LASSO MSE vs log(lambda):



Graph 12:Coefficient of Lasso:

```
12 x 1 sparse Matrix of class "dgCMatrix"
                      1
(Intercept)  .
Age          0.24445613
Ed           0.33056884
PE.1         0.76866809
LF           0.06352784
M            .
Pop          .
NW           0.19525084
UE1         -0.01349814
UE2          0.16612782
Wealth       .
IncInequ     0.36262584
```

Graph 13: Model diagnostic for Lasso:

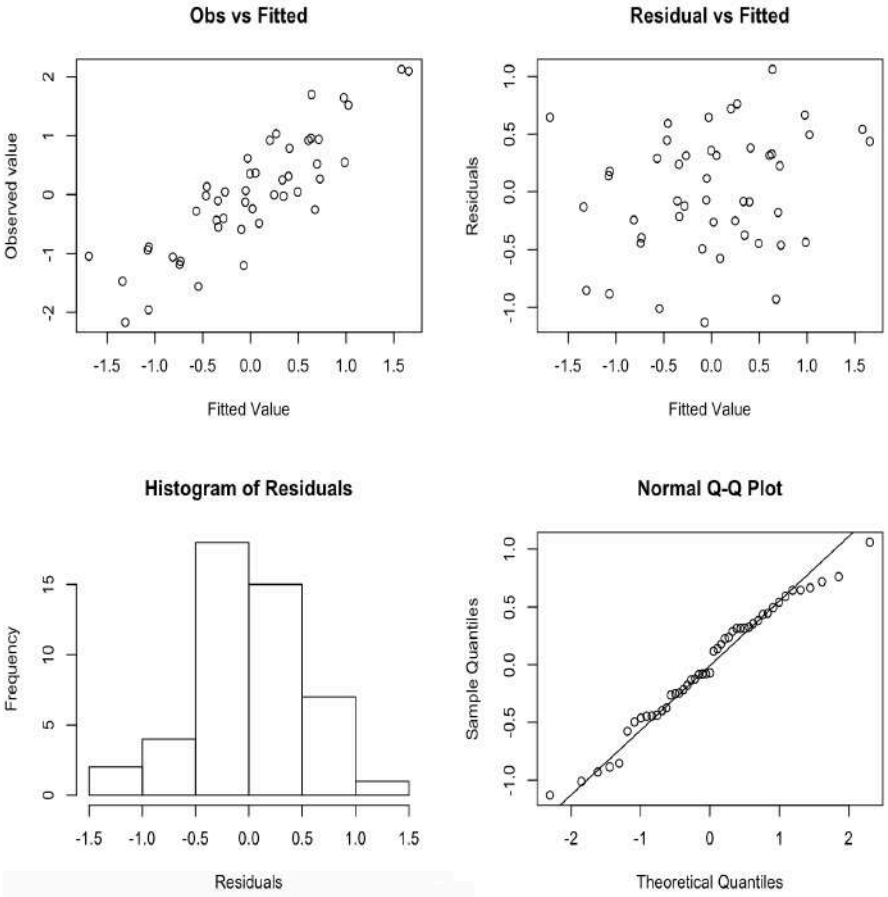
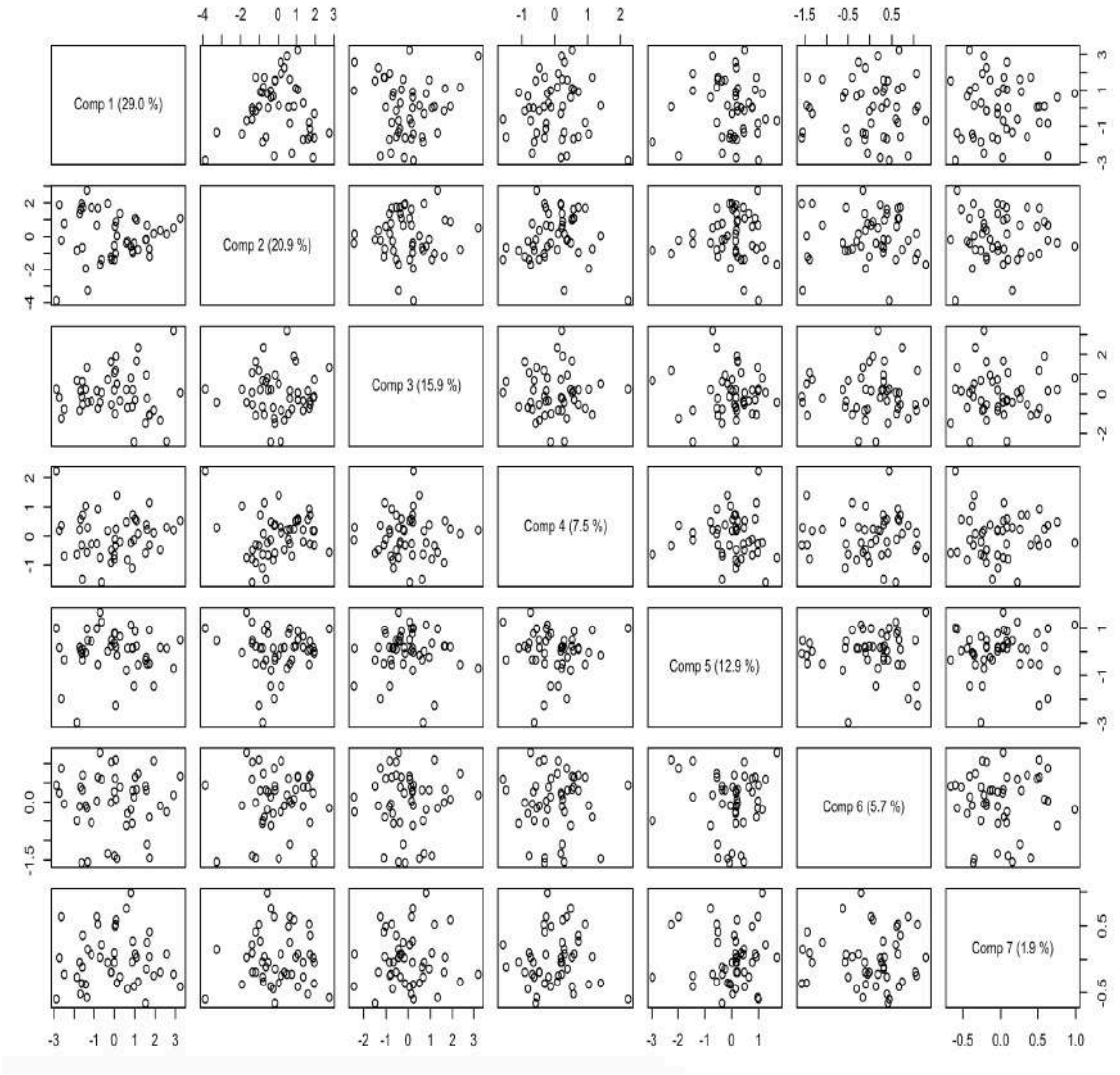


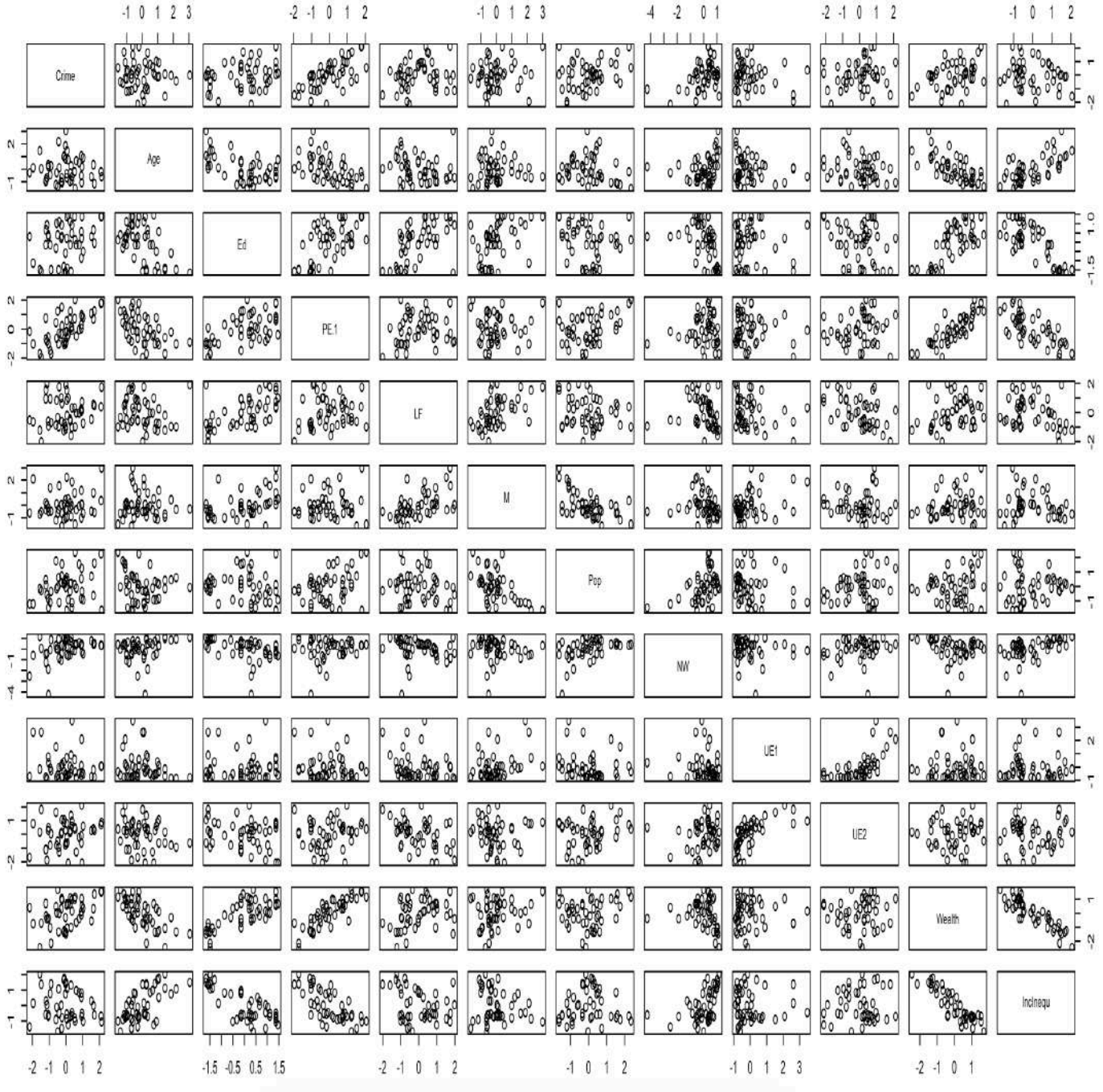
Table 14: Loadings of PLS under chosen number of components:

	Comp 1	Comp 2	Comp 3	Comp 4	Comp 5	Comp 6	Comp 7
Age	-0.36879709	0.40520860	0.305120816	0.15524608	0.19317997	-0.381217943	-0.2936111
Ed	0.40614824	-0.40555865	0.337740944	-0.06521700	0.20545524	0.139018412	0.2166889
PE.1	0.60293707	-0.02370157	-0.019382995	0.31487510	0.04621064	-0.141454678	0.1632475
LF	0.21163552	-0.15077678	0.430395779	-0.82163779	0.35585916	0.284600344	-0.5211128
M	0.09595830	-0.09887740	0.738143952	-0.33608706	-0.39566775	0.189155020	-0.1981444
Pop	0.25729486	0.29269976	-0.630283648	-0.06977315	0.15984282	0.148903228	-0.5236562
NW	0.11657059	0.64891826	-0.106815597	-0.31951963	-0.16719050	-0.001293357	0.4647805
UE1	-0.05583545	-0.12504848	0.036509759	0.17978720	-0.90466834	0.558780534	0.1432412
UE2	0.09297922	0.14162046	-0.191728532	0.66820480	-0.68402654	0.567010317	-0.5754360
Wealth	0.56362584	-0.28385225	0.008879955	0.08577938	-0.08192454	-0.164686517	0.1188680
IncInequ	-0.43704932	0.47097883	-0.033415129	-0.12536577	0.09102583	0.359246702	0.2067510

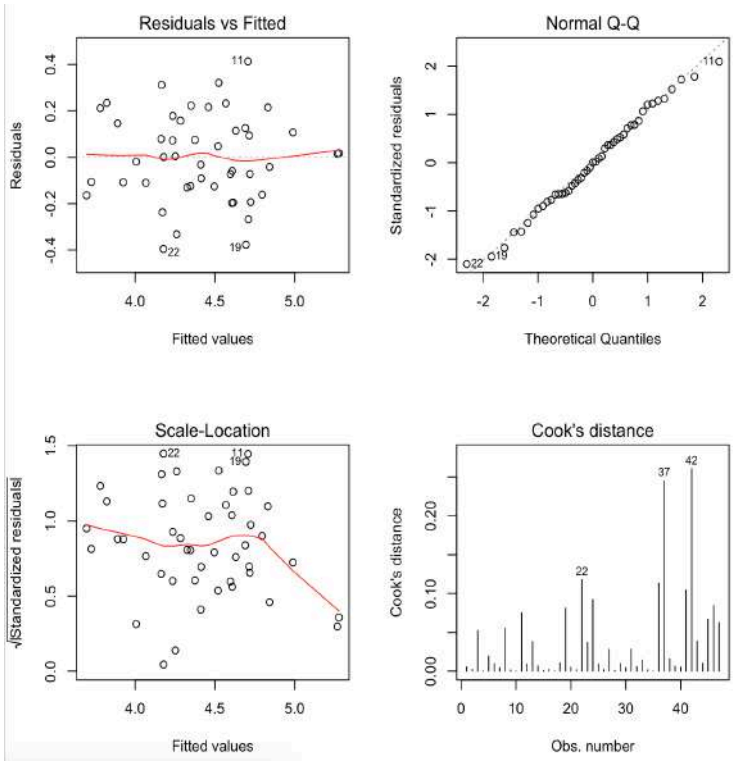
Graph 15: Scores plot for PLS:



Graph 16: Pairwise scatterplot:



Graph 17:Initial lm model diagnostics:



6.1 Appendix B: R code

```
library(MASS)
library(glmnet)
library(pls)
library(DAAG)
library(e1071)
crime<-read.csv("crime.csv",header=TRUE)
n<-nrow(crime)
#summary statistics:
apply(crime,2,summary)
#Preprocess:
fit0<-lm(Crime~.,data=crime)
par(mfrow=c(2,2))
plot(fit0,which=1)
plot(fit0,which=2)
plot(fit0,which=3)
plot(fit0,which=4)
#Pairwise scatterplot:
plot(crime)
#Histogram and possible transformation:
par(mfrow=c(3,4))
```