SAS – Assignment 3.

牛圣杰 1191090l

Problem 2 (Writting Section)

(b). We are testing 48 null hypotheses $H_{1,0}, \cdots, H_{48,0}$ simultaneously, and the corresponding P-values are $P_1, \cdots, P_{48}$.

We can compute the probability of making at least one Type I error (FWER): $FWER = 1 - (1 - 0.05)^{48} \approx 0.9147$.

$\Rightarrow$ We still have $1 - 0.9147 = 0.0853$'s chance of making no Type I error even without adjustment for multiple comparsion.

- Interpretation: the probability of incorrectly rejecting at least one null hypothesis from 48 total simultaneous without any adjustment for multiple comparisons is 91.47%

Problem 3

Note that $\underset{\sim}{y} = \underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$. $Var(\underset{\sim}{y}) = \underset{\sim}{W}^{-1} = diag\{b_i^2\}$.

Let $\underset{\sim}{H} = (\underset{\sim}{X}^T \underset{\sim}{W} \underset{\sim}{X})^{-1} \underset{\sim}{X}^T \underset{\sim}{W}$.

Suppose $\underset{\sim}{\lambda}' \underset{\sim}{y}$ is a linear function of observations and an estimator of $\underset{\sim}{\beta}$. if $\underset{\sim}{\lambda}\underset{\sim}{y}$

is b.l.u.e. of $\underset{\sim}{\beta}$, we have

1°. $E(\underset{\sim}{A}\underset{\sim}{y}) = \underset{\sim}{\beta} \Rightarrow \underset{\sim}{A}\underset{\sim}{X}\underset{\sim}{\beta} = \underset{\sim}{\beta} \Rightarrow \underset{\sim}{A}\underset{\sim}{X} = \underset{\sim}{I}$

2°. We need to satisfy that $\underset{\sim}{A}\underset{\sim}{y}$ has the minimum variance

$\Rightarrow \mathrm{Var}(\underset{\sim}{A}\underset{\sim}{y}) = \underset{\sim}{A}\,\mathrm{Var}(\underset{\sim}{y})\underset{\sim}{A} = \underset{\sim}{A}\underset{\sim}{W}^{-1}\underset{\sim}{A} = (\underset{\sim}{A}-\underset{\sim}{H}+\underset{\sim}{H})\underset{\sim}{W}^{-1}(\underset{\sim}{A}-\underset{\sim}{H}+\underset{\sim}{H})'$

$= (\underset{\sim}{A}-\underset{\sim}{H})\underset{\sim}{W}^{-1}(\underset{\sim}{A}-\underset{\sim}{H}) + \underset{\sim}{H}'\underset{\sim}{W}^{-1}\underset{\sim}{H} + (\underset{\sim}{A}-\underset{\sim}{H})\underset{\sim}{W}^{-1}\underset{\sim}{H}' + \underset{\sim}{H}\underset{\sim}{W}^{-1}(\underset{\sim}{A}-\underset{\sim}{H})$

Note that $(\underset{\sim}{A}-\underset{\sim}{H})\underset{\sim}{W}^{-1}\underset{\sim}{H}' = \underset{\sim}{A}\underset{\sim}{W}^{-1}\underset{\sim}{H}' - \underset{\sim}{H}\underset{\sim}{W}^{-1}\underset{\sim}{H}' = \underset{\sim}{A}\underset{\sim}{W}^{-1}\underset{\sim}{W}'\underset{\sim}{X}(\underset{\sim}{X}'\underset{\sim}{W}'\underset{\sim}{X})^{-1}$

$-(\underset{\sim}{X}'\underset{\sim}{W}\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{W}\underset{\sim}{W}^{-1}\underset{\sim}{W}'\underset{\sim}{X}(\underset{\sim}{X}'\underset{\sim}{W}'\underset{\sim}{X})^{-1} = \underset{\sim}{A}\underset{\sim}{X}(\underset{\sim}{X}'\underset{\sim}{W}'\underset{\sim}{X})^{-1} - (\underset{\sim}{X}'\underset{\sim}{W}'\underset{\sim}{X})^{-1} = \underset{\sim}{0}$ (Since $\underset{\sim}{A}\underset{\sim}{X} = \underset{\sim}{I}$)

Similarly, we have $\underset{\sim}{H}\underset{\sim}{W}^{-1}(\underset{\sim}{A}-\underset{\sim}{H}) = \underset{\sim}{0}$

Thus, $\mathrm{Cov}(\underset{\sim}{A}\underset{\sim}{y}) = (\underset{\sim}{A}-\underset{\sim}{H})\underset{\sim}{W}^{-1}(\underset{\sim}{A}-\underset{\sim}{H}) + \underset{\sim}{H}'\underset{\sim}{W}^{-1}\underset{\sim}{H} \geqslant \underset{\sim}{H}'\underset{\sim}{W}^{-1}\underset{\sim}{H} = \mathrm{Cov}(\underset{\sim}{H}'\underset{\sim}{y})$

Note that $(\underset{\sim}{A}-\underset{\sim}{H})\underset{\sim}{W}^{-1}(\underset{\sim}{A}-\underset{\sim}{H})$ is positive semi-definite $\Rightarrow$ this equality holds only when $\underset{\sim}{A} = \underset{\sim}{H}$.

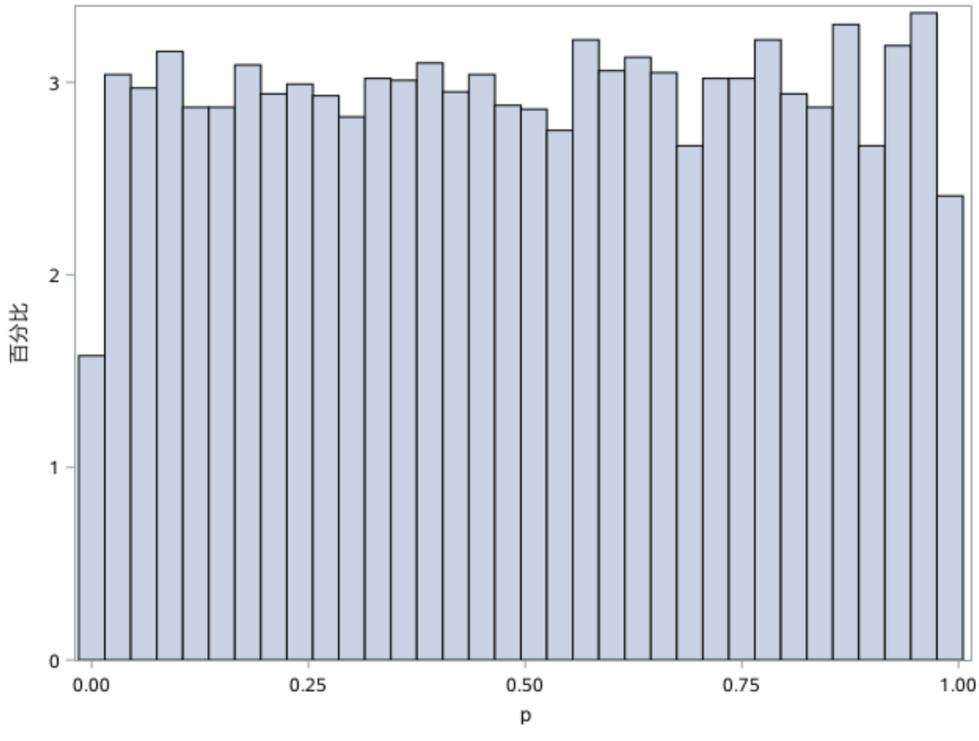
$\Rightarrow$ thus we have $\hat{\underset{\sim}{\beta}}^{WLS} = \underset{\sim}{A}\underset{\sim}{y} = \underset{\sim}{H}\underset{\sim}{y} = (\underset{\sim}{X}^T\underset{\sim}{W}\underset{\sim}{X})^{-1}\underset{\sim}{X}^T\underset{\sim}{W}\underset{\sim}{y}$ is b.l.u.e. of $\underset{\sim}{y}$

Problem 1

Code of SAS

```
%LET SampleSize = 10000;
DATA Data1;
  CALL STREAMINIT(12321);
  DO i = 1 TO &SampleSize;
    ARRAY x(*) x1-x20;
    DO j = 1 TO 20;
      x(j) = RAND("NORMAL");
    END;
    * Compute t statistic value;
    Mean = MEAN(OF x(*));Std= std(Of x(*));
    t=SQRT(20)*Mean/Std;p = 2 - 2 * CDF("T",abs(t),19);
    OUTPUT;
  END;
  DROP i j;
RUN;
* Draw the histogram;
PROC SGPLOT data=Data1;
  histogram p;
run;
```

The histogram of the 10000 p-values are given below:



From this figure, we can observe that P appears to obey $U(0,1)$, and this can also be proven in theory.

Proof

Note that the t-statistic $T = \frac{\sqrt{n}(\overline{X}-\mu_0)}{S}$ follows the t-distribution with degrees of freedom $n-1=19$ under H_0.

Thus for an obs t under H_0, we have $p = Pr(|T| \geq |t|) = 1 - |F(t) - F(-t)|$ where F is the cdf function of t-distribution with with degrees of freedom $n-1=19$.

Then consider the random variable under H_0, we have $P = 1 - |F(T) - F(-T)| = 1 - |2F(T) - 1|$.

We have $F(T) \sim U(0,1)$, and $2F(T) - 1 \sim U(-1,1)$, then $|2F(T) - 1| \sim U(0,1)$

Thus we have $P = 1 - |2F(T) - 1| \sim U(0,1)$.

Problem 2

1(1)

Code of SAS

```
%LET SampleSize = 10000;
DATA Sim1;
  CALL STREAMINIT(12321);
  ARRAY Disease(*) ar as CHD db liv lun;
  ARRAY Factor(*) leo tm Kb pur hair fn ID sum;
  DO i=1 TO &SampleSize;
    DO j= 1 TO 6;
      Disease(j) = RAND("Bernoulli", 0.1);
    END;
    DO j= 1 TO 8;
      Factor(j) = RAND("Bernoulli", 0.05);
    END;
    OUTPUT;
  END;
  DROP i j;
RUN;
```

2(2)

See the writing section.

2(3)

The p-values of the chi-square tests are given below:

	Arthritis	Asthma	CHD	Diabetes	Liver cirrhosis	Lung cancer
Leo	0.2547	0.6160	0.5447	0.3114	0.9013	0.3564
6:00-8:00am	0.3682	0.8012	0.4060	0.6837	0.6051	0.2045
Kobe fan	0.5229	0.2197	0.2677	0.2085	0.2055	0.9008
love purple	0.6200	0.5861	0.9840	0.7711	0.8054	0.9091
red hair	0.3622	0.3747	0.7012	0.0325	0.7950	0.0256
first name C	0.4643	0.6074	0.1696	0.1388	0.0368	0.7499
ID number 1	0.2281	0.6768	0.4149	0.5230	0.4221	0.2577
summer	0.5666	0.8767	0.1546	0.6984	0.5463	0.5115

From these results, without adjusting for multiple comparisons, we found three significant associations: red hair & Diabetes, first name C & Liver cirrhosis and red hair & Lung cancer.

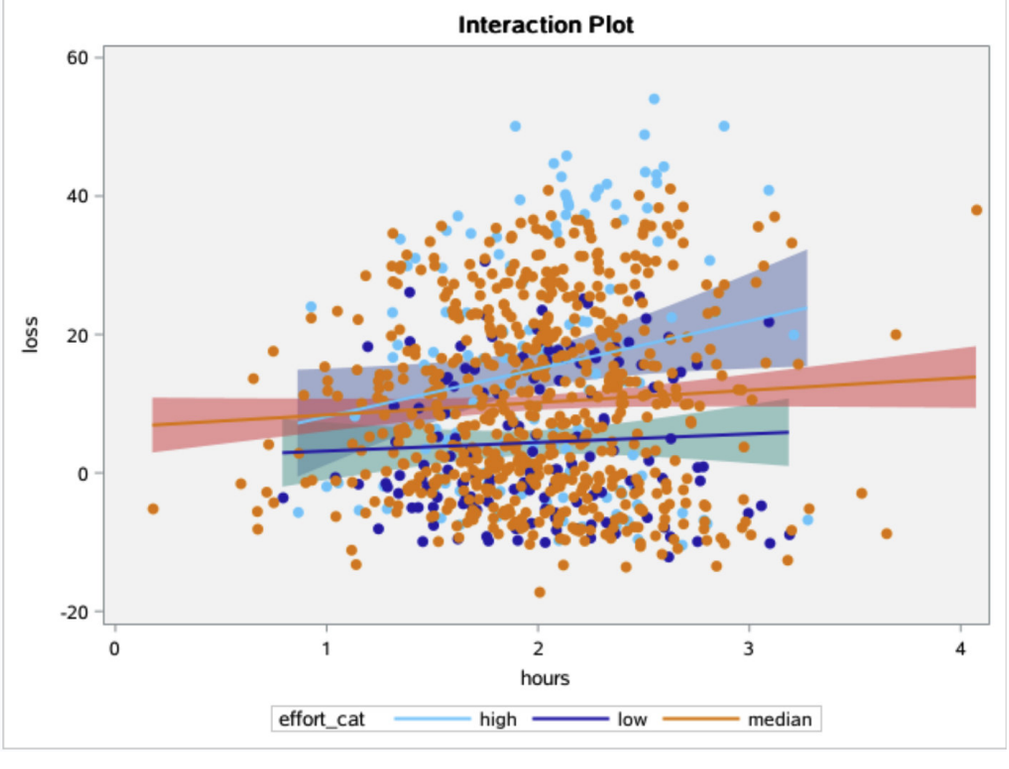
These findings are actually Type I errors.

Problem 4

4(1)

The mean of *effort* is 29.6592189, and the standard deviation (sd) of *effort* is 5.1427635.

The scatterplot of *loss* vs. *hours* with a regression line by *effort_cat* (i.e., the interaction plot) are given below



And from the plot, we can find:

- The relationship between *loss* and *hours* does not change direction based on *effort_cat*.
- The slopes of several straight lines are not seriously different, which indicates that their interaction

effect between *loss* and *hours* on *effort_cat* is not strong.

From above, we think the interaction term between *loss* and *hours* should not be added to our linear regression model.

4(2)

From problem 4(1), we do not consider the interaction effect of *hours* and *effort* and the estimated regression coefficients are given below

参数	估计	标准误差	t 值	Pr > |t|
截距	-15.60248713	3.20106924	-4.87	<.0001
effort	0.70525760	0.08812083	8.00	<.0001
hours	2.35014374	0.91638026	2.56	0.0105

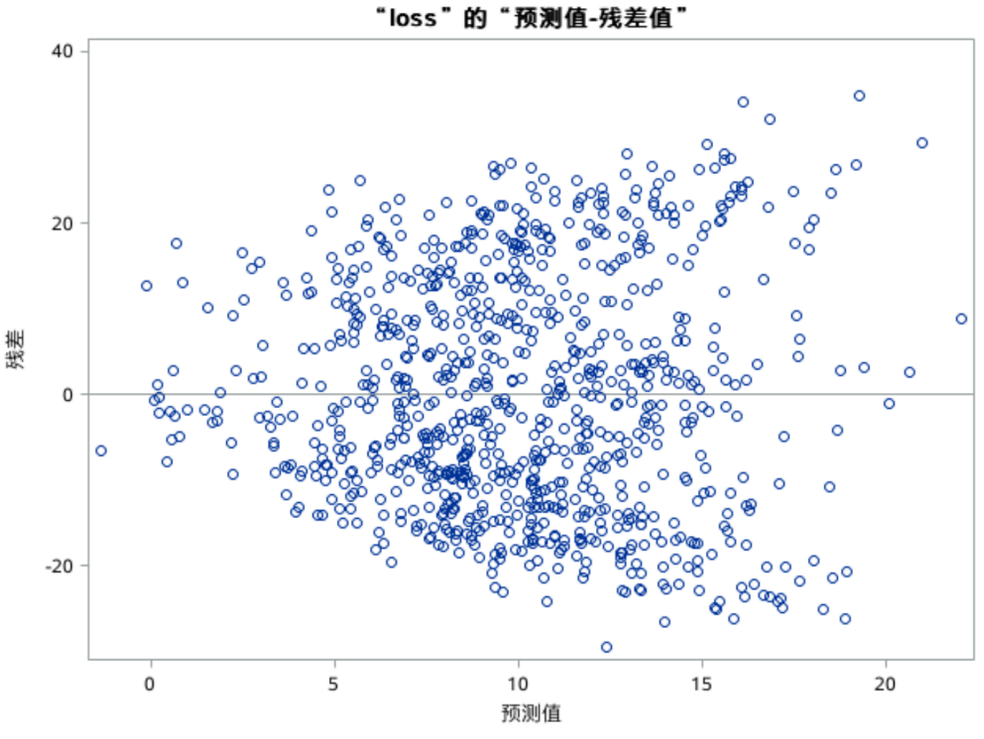
Conclusion:

- If we take the significance level $\alpha = 0.05$, then all the intercept, coefficients of effort and hours are significant.
- The linear regression model is
  $loss = 0.70525760 * effort + 2.35014374 * hours - 15.60248713 + \epsilon$

4(3)

Homoscedasticity assumption

Fitted vs. residual plot



White test

异方差性检验					
方程	检验	统计量	自由度	Pr > 卡方	变量
loss	White 检验	137.8	5	<.0001	所有变量的叉积
	Breusch-Pagan	129.7	2	<.0001	1, effort, hours

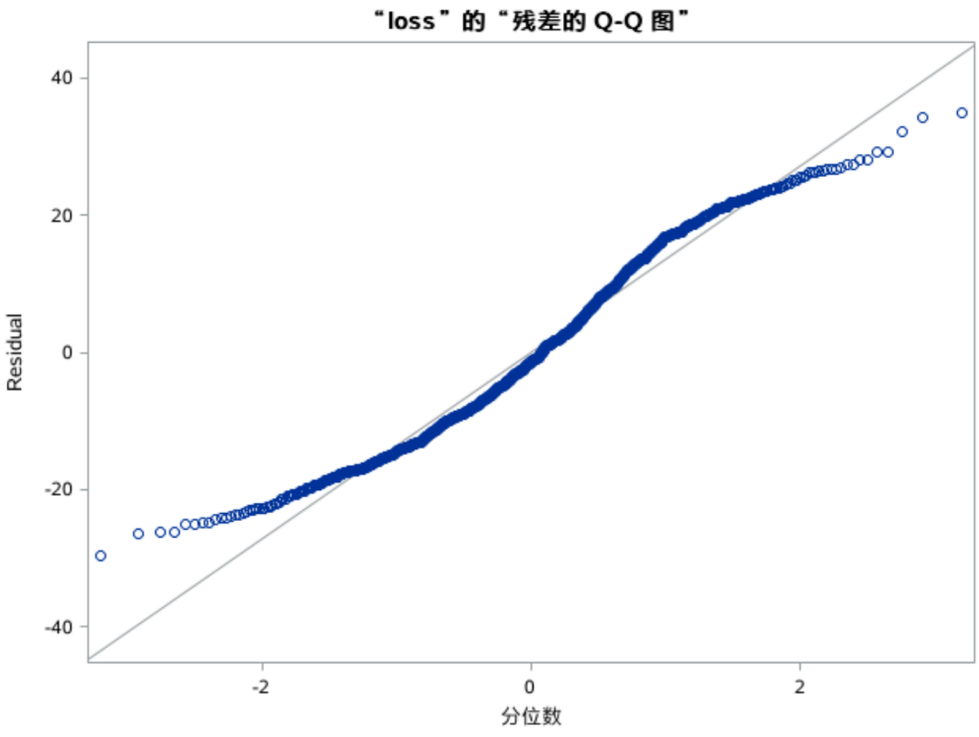
Conclusion

This model can not satisfy the homoscedasticity assumption, since

- The absolute value of residuals shows an increasing trend with the increase of the predicted value.
- The p-value obtained from White's test is significant, indicating that we reject the assumption of homoscedasticity.

Normality assumption

Q-Q plot



Shapiro-Wilk test

正态性检验			
方程	检验统计量	值	概率
resid	Shapiro-Wilk W	0.97	<.0001
系统	Mardia 偏度	8.40	0.0037
	Mardia 峰度	-5.61	<.0001
	Henze-Zirkler T	11.84	<.0001

Conclusion

This model can not satisfy the normality assumption, since

- The upper and lower quantiles of QQplot deviate significantly from the line.
- The p-value obtained from Shapiro-Wilk test is significant, indicating that we reject the assumption of normality.

But we can not judge Whether the cause of the error of non-normal residuals is caused by heteroskedasticity or the data itself.

Problem 5

5(1)

Obtain the pairwise Pearson correlation coefficients and rank them in descending order by the absolute value.

And we can have that X4&X5, X4&X6 and X5&X6 are highly linearly correlated (The corresponding correlation coefficients are 0.98456 and 0.97708 respectively).

Pearson 相关系数, N = 3045															
X1	X1 1.00000	deathRate 0.46272	X12 0.11403	X7 -0.11021	X8 0.10227	X9 0.10042	X10 0.05420	X14 -0.04731	X13 -0.01460	X11 -0.01252	X3 0.00834	X5 -0.00572	X4 0.00469	X6 0.00322	X2 -0.00223
X2	X2 1.00000	X3 -0.78895	X10 -0.75519	X9 0.72451	X14 0.70605	X8 -0.45295	deathRate -0.42832	X13 0.42612	X7 0.35573	X12 -0.27037	X11 0.16740	X6 -0.15354	X4 -0.11812	X5 -0.09180	X1 -0.00223
X3	X3 1.00000	X9 -0.82326	X2 -0.78895	X8 0.65508	X10 0.65178	X7 -0.64365	X14 -0.53283	X12 0.51161	X11 -0.50955	deathRate 0.42935	X5 -0.21385	X4 -0.19264	X13 -0.15765	X6 -0.14798	X1 0.00834
X4	X4 1.00000	X5 0.98456	X6 0.97708	X7 0.42702	X10 0.42663	X11 0.37690	X13 -0.24822	X12 -0.20560	X3 -0.19264	X14 -0.14858	X8 -0.12912	X2 -0.11812	X9 0.07175	deathRate -0.00577	X1 0.00469
X5	X5 1.00000	X4 0.98456	X6 0.93361	X7 0.44832	X10 0.39808	X11 0.39807	X12 -0.24267	X13 -0.23716	X3 -0.21385	X8 -0.14276	X14 -0.12965	X2 -0.09180	X9 0.08395	deathRate -0.02251	X1 -0.00572
X6	X6 1.00000	X4 0.97708	X5 0.93361	X10 0.45454	X7 0.37266	X11 0.33987	X13 -0.25735	X14 -0.17851	X12 -0.15654	X2 -0.15354	X3 -0.14798	X8 -0.11126	X9 0.04915	deathRate 0.01122	X1 0.00322
X7	X7 1.00000	X11 0.67853	X3 -0.64365	X12 -0.62340	X8 -0.55241	X9 0.45278	X5 0.44832	X4 0.42702	X6 0.37266	X2 0.35573	deathRate -0.26787	X10 -0.24974	X13 -0.14699	X1 -0.11021	X14 0.10722
X8	X8 1.00000	X3 0.65508	X9 -0.63470	X7 -0.55241	X10 0.53013	X11 -0.50192	X12 0.46938	X2 -0.45295	deathRate 0.37823	X14 -0.37374	X5 -0.14276	X4 -0.12912	X6 -0.11126	X1 0.10227	X13 -0.02211
X9	X9 1.00000	X3 -0.82326	X2 0.72451	X10 -0.71967	X8 -0.63470	X14 0.60290	X7 0.45278	X11 0.42971	deathRate -0.38578	X12 -0.34569	X13 0.18848	X1 0.10042	X5 0.08395	X4 0.07175	X6 0.04915
X10	X10 1.00000	X2 -0.75519	X9 -0.71967	X3 0.65178	X14 -0.63581	X8 0.53013	X6 0.45454	X4 0.42663	deathRate 0.40428	X5 0.39808	X13 -0.30489	X7 -0.24974	X12 0.19604	X11 -0.13421	X1 0.05420
X11	X11 1.00000	X12 -0.82845	X7 0.67853	X3 -0.50955	X8 -0.50192	X9 0.42971	X5 0.39807	X4 0.37690	X6 0.33987	X13 -0.26551	deathRate -0.17783	X2 0.16740	X10 -0.13421	X14 0.04909	X1 -0.01252
X12	X12 1.00000	X11 -0.82845	X7 -0.62340	X3 0.51161	X8 0.46938	X9 -0.34569	X2 -0.27037	deathRate 0.25739	X5 -0.24267	X4 -0.20560	X10 0.19604	X6 -0.15654	X14 -0.14690	X1 0.11403	X13 0.01634
X13	X13 1.00000	X14 0.43721	X2 0.42612	X10 -0.30489	X11 -0.26551	X6 -0.25735	X4 -0.24822	X5 -0.23716	X9 0.18848	deathRate -0.18611	X3 -0.15765	X7 -0.14699	X8 -0.02211	X12 0.01634	X1 -0.01460
X14	X14 1.00000	X2 0.70605	X10 -0.63581	X9 0.60290	X3 -0.53283	deathRate -0.48608	X13 0.43721	X8 -0.37374	X6 -0.17851	X4 -0.14858	X12 -0.14690	X5 -0.12965	X7 0.10722	X11 0.04909	X1 -0.04731
deathRate	deathRate 1.00000	X14 -0.48608	X1 0.46272	X3 0.42935	X2 -0.42832	X10 0.40428	X9 -0.38578	X8 0.37823	X7 -0.26787	X12 0.25739	X13 -0.18611	X11 -0.17783	X5 -0.02251	X6 0.01122	X4 -0.00577

5(2)

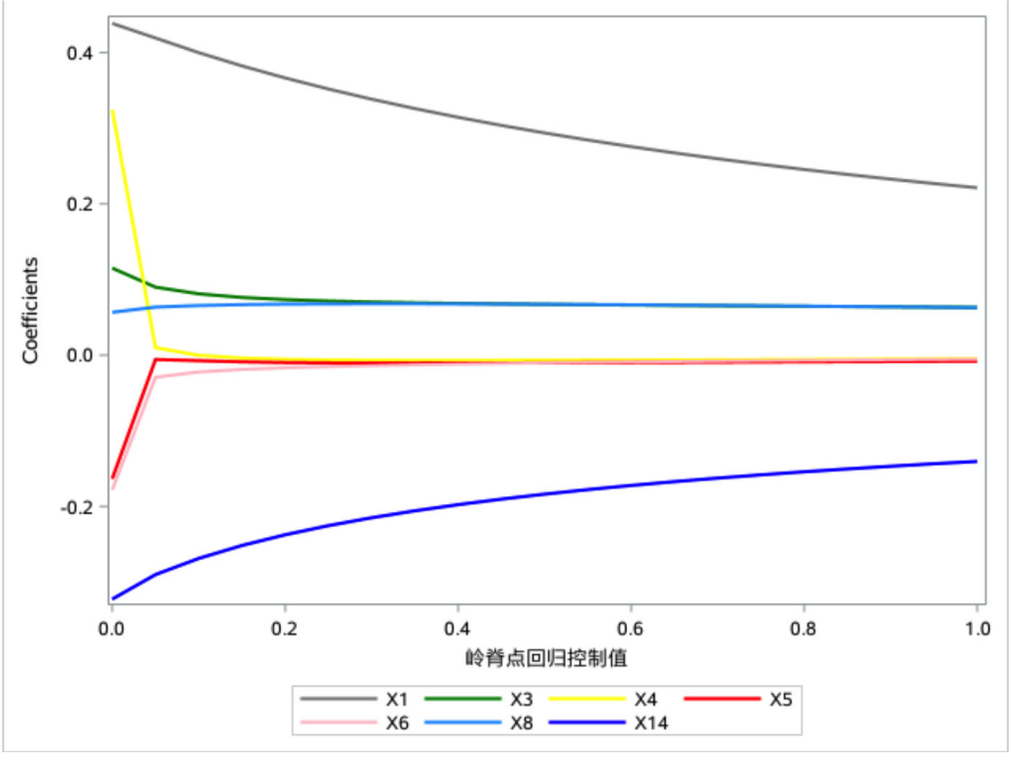
We can check the tolerance or the variance inflation factor (tolerance < 0.1 or VIF > 10) to detect the multicollinearity.

And we can see that multicollinearity exists for X4, X5 and X6.

参数估计							
变量	自由度	参数估计	标准误差	t 值	Pr > |t|	容差	方差膨胀
Intercept	1	105.06360	10.64377	9.87	<.0001	.	0
X1	1	0.22746	0.00734	31.00	<.0001	0.85533	1.16913
X2	1	-0.02214	0.07085	-0.31	0.7547	0.18114	5.52062
X3	1	0.49805	0.15528	3.21	0.0014	0.13307	7.51509
X4	1	1.73487	1.07601	1.61	0.1070	0.00423	236.36600
X5	1	-0.86791	0.62807	-1.38	0.1671	0.01226	81.57354
X6	1	-0.93525	0.52369	-1.79	0.0742	0.01722	58.07256
X7	1	-0.11051	0.09337	-1.18	0.2367	0.31928	3.13209
X8	1	0.45534	0.16062	2.83	0.0046	0.42862	2.33305
X9	1	-0.22009	0.07959	-2.77	0.0057	0.18370	5.44356
X10	1	-0.01005	0.11719	-0.09	0.9317	0.15622	6.40119
X11	1	0.05969	0.05389	1.11	0.2681	0.16912	5.91292
X12	1	0.10414	0.05434	1.92	0.0554	0.21124	4.73405
X13	1	0.00290	0.18131	0.02	0.9872	0.58885	1.69823
X14	1	-1.66058	0.10730	-15.48	<.0001	0.39410	2.53744

5(3)

The plot of lines showing the parameter estimates of X1, X3, X4, X5, X6, X8, X14 against λ are given below:



From the plot we can find:

- When λ increases by a small amount, the coefficients of X4 (Median age of residents), X5(Median age of male residents) and X6 (Median age of female residents) decrease rapidly in absolute value.

  This is not unexpected because the sample correlation coefficients of X4&X5, X4&X6 and X5&X6 are 0.98456, 0.97708 and 0.93361respectively (from problem5(1)). The coefficients of X4, X5 and X6 are quickly driven towards zero and are almost mirror images of each other about the zero line.

- The effects of X1 (per capita(100,000) cancer diagnosis), X3 (Percent of populace in poverty) and X14 (Percent of resident (age 25 and over) with bachelor's degree) also appear to be overestimated in absolute value. The coefficients decrease in absolute value as λ increases, and level off at non-zero values.

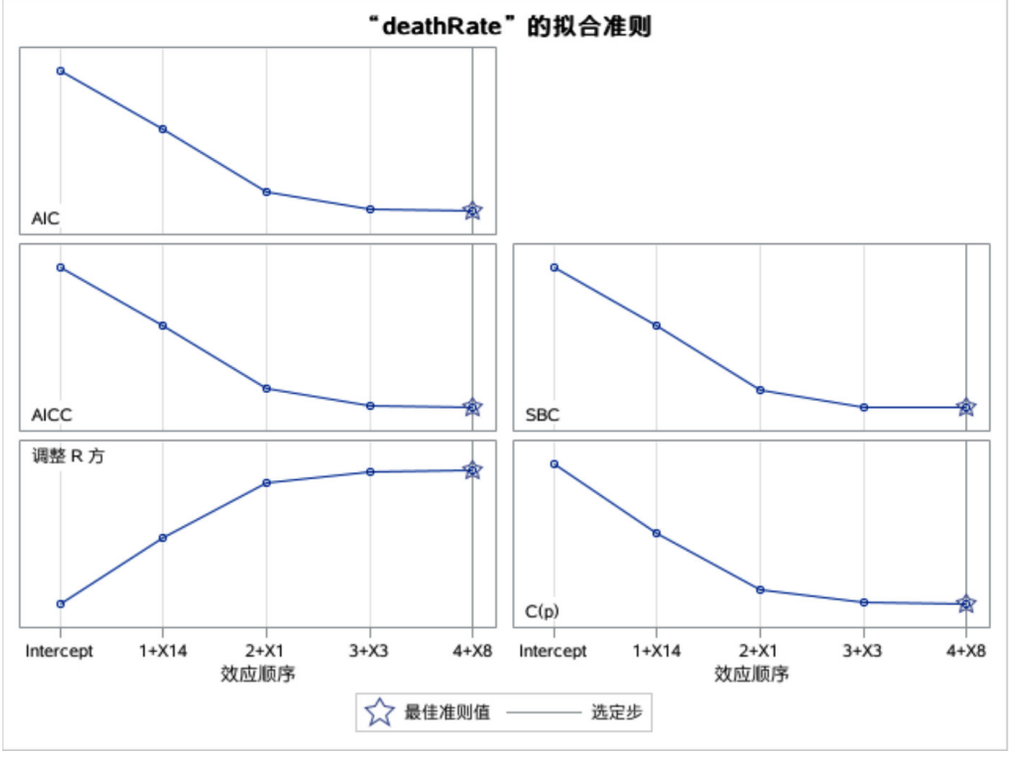
- The effect of X8 (Percent of resident (age 16 and over) unemployed), is likely to be originally underestimated. The coefficient increases slightly as λ increases.
- The coefficients appear to stabilize in the neighborhood of $\lambda = 0.4$. So we expect coefficients estimated at $\lambda = 0.4$ to be more suitable for estimation of the effects of the explanatory variables.

5(4)

The model selection results according to the BIC criterion are presented below, X14, X1, X3, X8 are added sequentially to the model.

逐步选择汇总							
步	进入的效应	删除的效应	引入效应数	调整 R 方	AIC	CP	SBC
0	Intercept		1	0.0000	23285.4967	2795.6228	20244.5180
1	X14		2	0.2360	22466.7135	1418.0861	19431.7561
2	X1		3	0.4297	21577.4697	288.6053	18548.5335
3	X3		4	0.4737	21333.9233	32.8040	18311.0084
4	X8		5	0.4773*	21314.2926*	13.0743*	18297.3989*
* 准则的最佳值							

Generate the plots of these criteria by step.



参数估计				
参数	自由度	估计	标准误差	t 值
Intercept	1	80.166233	3.685285	21.75
X1	1	0.226680	0.006860	33.05
X3	1	0.849711	0.082680	10.28
X8	1	0.654495	0.140590	4.66
X14	1	-1.700383	0.079852	-21.29

From above, we have the final model after model selection is

$$Y = 80.166233 + 0.226680X_1 + 0.849711X_3 + 0.654495X_8 - 1.700383X_{14} + \epsilon$$

- On average, after accounting for the effect of the other three variables, the deathRate(per capita cancer mortalities) with 1 more per capita cancer diagnosis is 0.2267 higher.
- On average, after accounting for the effect of the other three variables, the deathRate(per capita cancer mortalities) with 1% more population in poverty is 0.8497 higher.
- On average, after accounting for the effect of the other three variables, the deathRate(per capita cancer mortalities) with 1% more residents (age 16 and over) unemployed is 0.6545 higher.
- On average, after accounting for the effect of the other three variables, the deathRate(per capita cancer mortalities) with 1% more residents (age 25 and over) with bachelor's degree is 1.7004 lower.

5(5)

When we filter out the regions with $|r_{stu}| > 3$ as outliers, the filtered regions sorted by X1 in ascending order are given below

观测	Region	deathRate	X1	X3	X8	X14	r
1	Aleutians West Census Area, Alask	203.3	201.3	9.9	2.1	10.3	4.27942
2	Presidio County, Texas	66.3	211.1	21.8	9.6	14	-3.14242
3	Zapata County, Texas	112.5	340.9	32.6	12.2	5.3	-3.58358
4	Atchison County, Missouri	222.4	372.9	12.2	4.3	15.8	3.56789
5	Wibaux County, Montana	214.4	373.3	11.7	3	16.4	3.27764
6	Randolph County, Georgia	127	374.8	36.3	10.4	7.1	-3.18806
7	Crowley County, Colorado	110.7	389.1	40.5	27	7.8	-4.85664
8	Crittenden County, Arkansas	242.9	389.9	27.3	9.9	11.7	3.22666
9	Anderson County, Texas	245.2	394.3	20.4	4.3	7.5	3.41220
10	Dade County, Georgia	238.2	423.9	16.4	6.1	10.2	3.06410
11	Esmeralda County, Nevada	262.1	453.5494221	14	11.9	12.6	4.04464
12	Lander County, Nevada	123.8	453.5494221	11	12	4.7	-3.42247
13	Pershing County, Nevada	121.8	453.5494221	18.5	9.5	7.5	-3.51546
14	Rush County, Kansas	114.3	453.5494221	12.1	5.6	12.8	-3.03888
15	Woodson County, Kansas	293.9	453.5494221	18.4	4.6	12.7	5.70904
16	Cimarron County, Oklahoma	258.7	456.9	18	2.2	15.3	4.21594
17	Coahoma County, Mississippi	266.7	460.1	35	20.3	12.6	3.03551
18	Madison County, Mississippi	292.5	460.5	13.1	5.4	26.7	6.97678
19	Mississippi County, Missouri	270.5	476.9	32.2	12.2	8	3.02300
20	Franklin Parish, Louisiana	140.5	479.6	29	11.5	6.4	-3.48220
21	Baker County, Georgia	127.6	488.4	27.3	7.1	4.3	-4.19506
22	Greene County, Alabama	154.5	495.8	33.2	20.4	7.8	-3.32126
23	Nome Census Area, Alaska	277.6	499.3	26.1	16.5	10.3	3.43978
24	North Slope Borough, Alaska	256.9	501.1	11.5	9.3	10.4	3.24662
25	Robertson County, Kentucky	274	530.3	23	6.7	9.8	3.31704
26	Echols County, Georgia	126.4	538.1	27.1	10.1	6.5	-4.72418

The 2712th, 2724th and 2725th observation is the datapoint for Alaska, which is a state not connected to the mainland of the US. So it may be considered an outlier and removed from the analysis.