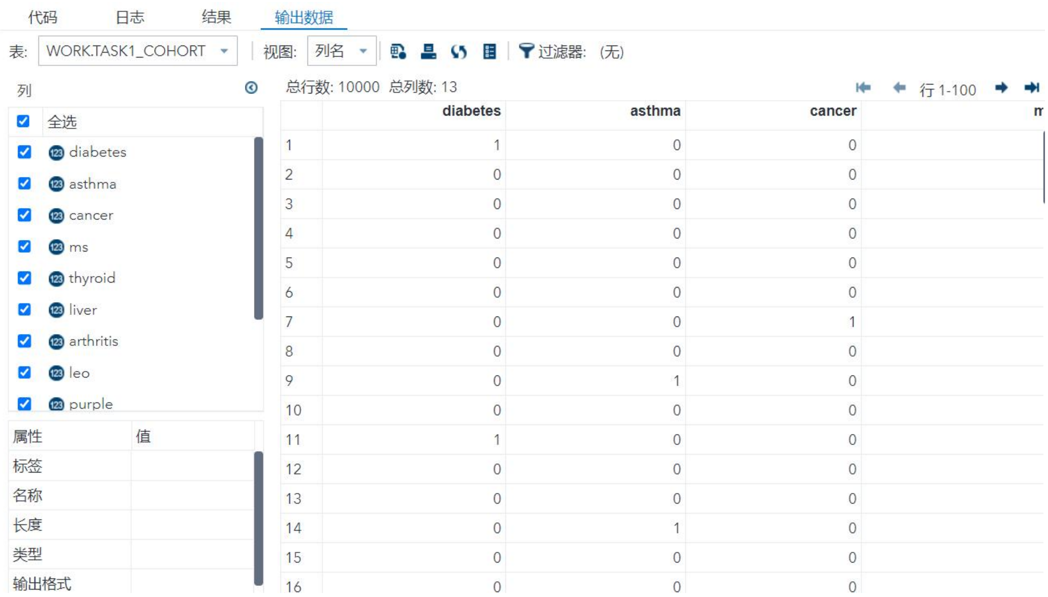
STA409

Answer to Assignment 3

12112627
李乐平

1. Solution.

(1).



	diabetes	asthma	cancer
1	1	0	0
2	0	0	0
3	0	0	0
4	0	0	0
5	0	0	0
6	0	0	0
7	0	0	1
8	0	0	0
9	0	1	0
10	0	0	0
11	1	0	0
12	0	0	0
13	0	0	0
14	0	1	0
15	0	0	0
16	0	0	0

(2). The family-wise error rate (FWER) is computed by

$$\text{FWER} = 1-(1-0.05)^{42} = 0.8840$$

That means we only have a chance of 11.6% that we do not make any Type I Error without any adjustment.

(3).

	Diabetes	Asthma	Cancer	Multiple Sclerosis	Thyroid Disease	Liver Disease	Arthritis
Leo	0.7953	0.8187	0.9528	0.1092	0.4827	0.1701	0.9060
Purple	0.6143	0.2577	0.4849	0.2487	0.8314	0.7887	0.1878
0:00-1:00	0.0628	0.4728	0.1874	0.0088	0.8947	0.0103	0.5419
Red Hair	0.2353	0.5705	0.7993	0.2556	0.3873	0.3861	0.7048
First Name C	0.4391	0.3479	0.3406	0.7265	0.2134	0.6235	0.8320
Summer	0.6329	0.3941	0.8162	0.3259	0.9229	0.3629	0.3189

In the table above, 2 combinations of disease and risk factor are determined to be dependent, which are actually Type I Errors. That implies it makes sense to do adjustment to reduce the chance of Type I Error happens under the increment of number of experiments.

2. Solution.

(1). The ANOVA table and coefficients table are given by

ANOVA Table					
Source	DF	Sum of Squares	Mean Square	F-value	p-value
Model	4	23665352	5916338	22.9782	5.0715×10^{-13}

Error	88	22657938	257476.5682		

where MS = sum_of_squares / DF; F-value = MS_{Model} / MS_{Error}; p-value = Pr(F(DF_{Model}, DF_{Error})>F).

Coefficients Table				
Parameter	Estimate	Standard Error	t-value	p-value
Intercept	3526.4	327.7	10.7610	0
Gender	722.5	117.8	6.1333	2.3917×10^{-8}
Education	90.02	24.69	3.6460	4.5033×10^{-4}
Experience	1.2690	0.5877	2.1593	3.3547×10^{-2}
Months	23.406	5.201	4.5003	2.0765×10^{-5}

where t-value = Estimate / SE; p-value = Pr(|t(DF_{Error})| > |t|)

In this question, the model for test is given by

$$\text{Salary} = \beta_0 + \beta_1\text{Gender} + \beta_2\text{Education} + \beta_3\text{Experience} + \beta_4\text{Months} + \varepsilon$$

The overall F-test is rejected, indicating at least one of βs are significantly different from 0. And t-tests show that every predictor variable is significant.

(2). The R^2 and adjusted R^2 are computed by

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} = 1 - \frac{23665652}{46323290} = 0.5109$$

$$\text{Adjusted} - R^2 = 1 - \frac{\text{SSE}/(n-p-1)}{\text{SST}/(n-1)} = 1 - \frac{23665652/88}{46323290/92} = 0.4886$$

(3). A partial F-test is utilized to test which of the full model and the reduced model is better.

$$F = \frac{(SST_R - SST_F)/(p-k)}{SST_F/(n-p-1)} = \frac{(38460756 - 22657938)/3}{22657938/88} = 20.4586$$

Because $\Pr(F(3, 88) > F) = 3.8052\times10^{-10}$, hence the reduced model is rejected.

3. Solution.

Now we want to show that $\hat{\boldsymbol{\beta}}^{\text{WLS}} = (X^TWX)^{-1}X^TW\boldsymbol{y}$ is the best linear unbiased estimate of $\boldsymbol{\beta}$, where $W = [\text{Var}(\boldsymbol{y})]^{-1}$.

The variance of $\hat{\boldsymbol{\beta}}^{\text{WLS}}$ is

$$\begin{aligned}\text{Var}(\hat{\boldsymbol{\beta}}^{\text{WLS}}) &= ((X^TWX)^{-1}X^TW)Var(\boldsymbol{y})((X^TWX)^{-1}X^TW)^T \\ &= (X^TWX)^{-1}\end{aligned}$$

Assume there is another linear estimation of $\boldsymbol{\beta}$: $\tilde{\boldsymbol{\beta}} = A\boldsymbol{y}, A = (X^TWX)^{-1}X^TW + B$, then it must satisfies

$$\text{E}(\tilde{\boldsymbol{\beta}}) = \text{E}[((X^TWX)^{-1}X^TW + B)(X\boldsymbol{\beta} + \boldsymbol{\varepsilon})] = (I + BX)\boldsymbol{\beta} = \boldsymbol{\beta},$$

i.e. $BX = 0$, or it would not be unbiased estimator of $\boldsymbol{\beta}$.

Yet

$$\begin{aligned}\text{Var}(\tilde{\boldsymbol{\beta}}) &= ((X^TWX)^{-1}X^TW + B)W^{-1}((X^TWX)^{-1}X^TW + B)^T \\ &= (X^TWX)^{-1} + BW^{-1}B^T \\ &= \text{Var}(\hat{\boldsymbol{\beta}}^{\text{WLS}}) + BW^{-1}B^T \\ &\ge \text{Var}(\hat{\boldsymbol{\beta}}^{\text{WLS}})\end{aligned}$$

then it is clear that $\hat{\boldsymbol{\beta}}^{\text{WLS}}$ is the best linear unbiased estimate of $\boldsymbol{\beta}$.

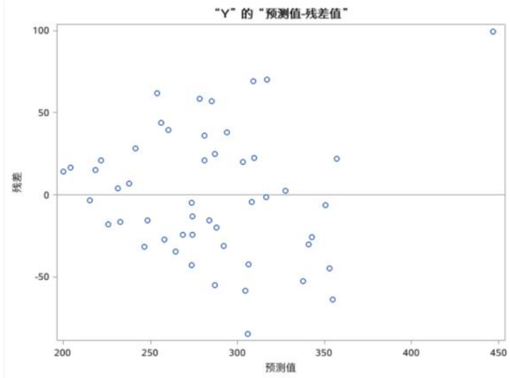
4. Solution.

(1). The fitted linear model is stated by

$$Y = -556.5680 + 0.0724X_1 + 1.5521X_2 - 0.0043X_3 + \varepsilon$$

And the assumptions are checked as follows.

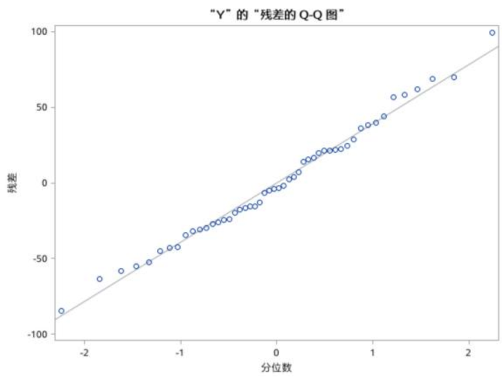
Linearity assumption: The predicted value - residual plot shows that the mean of the residuals are roughly 0 at any place, which implies the validity of linearity assumption.

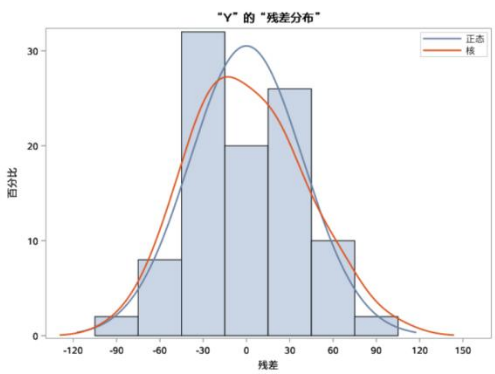


Homoscedasticity assumption: The predicted value - residual plot does not show homoscedaticity since it is narrow at left side and wide at right side. Also, the White test and Breusch-Pagan test reject the homoscedasticity assumption with p-values less than 0.01.

异方差性检验					
方程	检验	统计量	自由度	Pr > 卡方	变量
Y	White 检验	22.68	9	0.0070	所有变量的叉积
	Breusch-Pagan	15.59	3	0.0014	1, X1, X2, X3

Normality: The Q-Q plot as well as distribution of residuals does not show violation of the normality assumption. Also, the Shapiro-Wilk test does not reject the normality assumption.



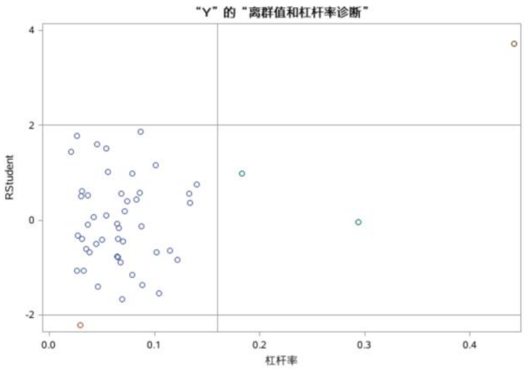


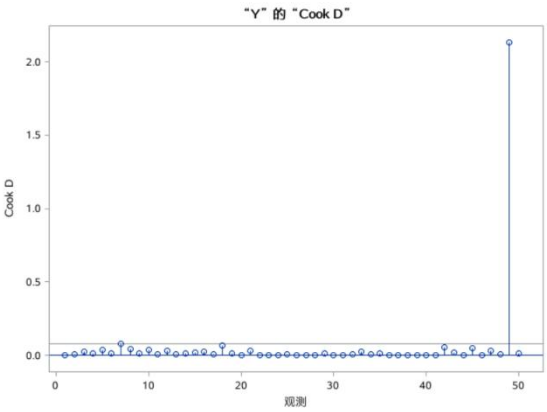
正态性检验				
检验	统计量		p 值	
Shapiro-Wilk	W	0.962619	Pr < W	0.1145
Kolmogorov-Smirnov	D	0.077587	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.0368	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.295345	Pr > A-Sq	>0.2500

Independence: The data are collected from 50 states independently, hence we can assume the independence of the data.

(2). Notice that the state Alaska has both high leverage, Cook's distance and studentized residual, which shows Alaska is an outlier and influential observation. Also shown in the plot.

	State	leverage	cookD	rstudent
1	AK	0.4419099163	2.1327949114	3.7099223829
2	UT	0.2941338295	0.0002247544	-0.04594205
3	NM	0.1829094976	0.054304552	0.9847410811
4	VT	0.1400044576	0.0235191903	0.7566427892
5	WV	0.1334181929	0.0053186159	0.3682198446
6	FL	0.1329380512	0.0121187746	0.5580649664
7	MS	0.121715957	0.024280392	-0.834377609
8	LA	0.1150154266	0.0134134888	-0.638374691
9	ND	0.1046879429	0.0682663128	-1.551362037
10	CT	0.101921737	0.01292351	-0.670860476
11	RI	0.1011891856	0.0370314221	1.1510950224
12	NJ	0.088828501	0.0452504612	-1.375748431
13	TX	0.0872943074	0.0003731912	-0.123585963
14	NY	0.0864580927	0.077778339	1.8610093979
15	CA	0.0861878909	0.0078594908	0.57311144





Considering Alaska is away from the mainland of U.S., it may be dropped from the analysis.

(3). If we only refit the model using weighted GLM without dropping outlier to solve the heteroscedasticity, the refitted model is stated as

$$Y = -423.0778 + 0.0649X_1 + 1.1954X_2 + 0.0223X_3 + \varepsilon$$

参数	估计	标准误差	t 值	Pr > |t|
截距	-423.0777720	116.1533088	-3.64	0.0007
X1	0.0648833	0.0114096	5.69	<.0001
X2	1.1953871	0.2978225	4.01	0.0002
X3	0.0222879	0.0468463	0.48	0.6365

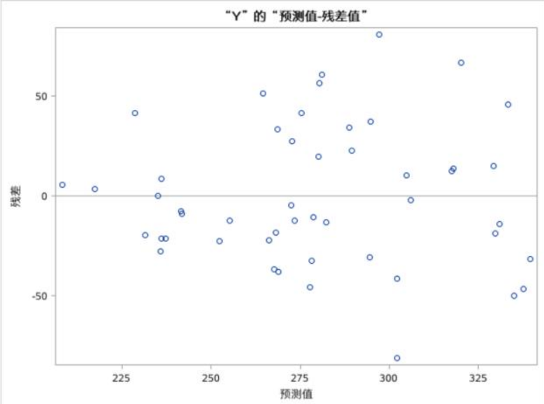
However, after dropping the data of Alaska, the fitted linear model is stated as

$$Y = -277.5773 + 0.0483X_1 + 0.8869X_2 + 0.0668X_3 + \varepsilon$$

This time the homoscedasticity assumption is accepted. And the other 3 assumptions are also valid (omitted).

参数估计

变量	标签	自由度	参数估计	标准误差	t 值	Pr > |t|	方差膨胀
Intercept	Intercept	1	-277.57731	132.42286	-2.10	0.0417	0
X1	X1	1	0.04829	0.01215	3.98	0.0003	2.23892
X2	X2	1	0.88693	0.33114	2.68	0.0103	1.26399
X3	X3	1	0.06679	0.04934	1.35	0.1826	1.89792



异方差性检验

方程	检验	统计量	自由度	Pr > 卡方	变量
Y	White 检验	8.69	9	0.4665	所有变量的叉积
	Breusch-Pagan	5.03	3	0.1698	1, X1, X2, X3

Compared with the former fitted model,

$$Y = -556.5680 + 0.0724X_1 + 1.5521X_2 - 0.0043X_3 + \varepsilon$$

all the assumptions are accepted, and the fitted coefficients differs significantly, which indicates the Alaska row is indeed an outlier.

5. Solution.

(1). By the ranked table of pairwise Pearson correlations, we can observe that X_{12} and X_{13} are highly correlated, whose correlation coefficient is 0.98384.

Pearson 相关系数, N = 60
Prob > |r|, H0: Rho=0

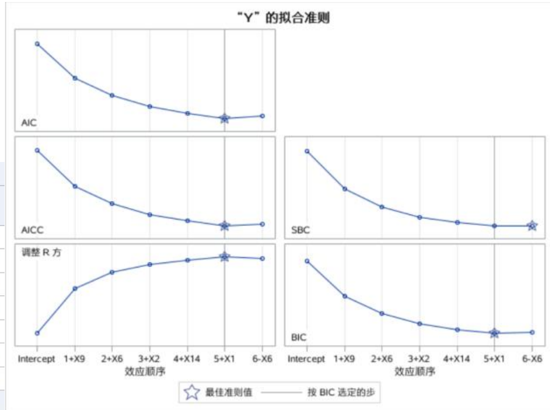
Y **Y**	Y 1.00000	X9 0.64374 <.0001	X6 -0.51099 <.0001	X1 0.50950 <.0001	X7 -0.42682 0.0007	X14 0.42590 0.0007	X11 0.41049 0.0011	X5 0.35731 0.0051	X10 -0.28480 0.0274	X3 0.27702 0.0321	X8 0.26550 0.0403	X12 -0.17724 0.1755	X4 -0.17459 0.1821	X15 -0.08850 0.5013	X13 -0.07738 0.5568	X2 -0.03002 0.8199
X1 **X1**	X1 1.00000	X12 -0.53176 <.0001	Y 0.50950 <.0001	X11 0.50659 <.0001	X3 0.50327 <.0001	X7 -0.49076 <.0001	X6 -0.49043 <.0001	X13 -0.48732 <.0001	X9 0.41320 0.0010	X10 -0.29729 0.0211	X5 0.26344 0.0420	X14 -0.10692 0.4161	X4 0.10111 0.4421	X2 0.09221 0.4835	X15 -0.07734 0.5570	X8 -0.00352 0.9787
X2 **X2**	X2 1.00000	X11 0.56531 <.0001	X9 0.45377 0.0003	X4 -0.39810 0.0016	X12 0.35081 0.0060	X3 0.34628 0.0067	X13 0.32101 0.0124	X10 0.23799 0.0671	X5 -0.20921 0.1087	X6 0.11628 0.3763	X14 -0.10781 0.4123	X8 -0.10005 0.4469	X1 0.09221 0.4835	X15 0.06787 0.6064	Y -0.03002 0.8199	X7 0.01485 0.9103
X3 **X3**	X3 1.00000	X11 0.61931 <.0001	X9 0.57531 <.0001	X1 0.50327 <.0001	X15 -0.45281 0.0003	X4 -0.43404 0.0005	X7 -0.41503 0.0010	X12 -0.35649 0.0052	X2 0.34628 0.0067	X13 -0.33767 0.0083	Y 0.27702 0.0321	X5 0.26228 0.0429	X6 -0.23854 0.0664	X14 -0.09935 0.4501	X8 -0.06099 0.6434	X10 -0.02141 0.8710
X4 **X4**	X4 1.00000	X9 -0.63782 <.0001	X5 -0.50909 <.0001	X3 -0.43404 0.0005	X2 -0.39810 0.0016	X11 -0.30977 0.0160	Y -0.17459 0.1821	X8 0.16199 0.2162	X6 -0.13886 0.2900	X10 -0.11771 0.3704	X15 0.11243 0.3924	X1 0.10111 0.4421	X7 0.06501 0.6217	X12 -0.02049 0.8765	X14 0.01725 0.8959	X13 -0.00208 0.9874
X5 **X5**	X5 1.00000	X4 -0.50909 <.0001	X10 -0.42572 0.0007	X9 0.41941 0.0009	X7 -0.41059 0.0011	X6 -0.39507 0.0018	X12 -0.38821 0.0022	X13 -0.35843 0.0049	Y 0.35731 0.0051	X1 0.26344 0.0420	X3 0.26228 0.0429	X11 0.25990 0.0449	X2 -0.20921 0.1087	X8 -0.18433 0.1586	X15 -0.13574 0.3011	X14 -0.00408 0.9753
X6 **X6**	X6 1.00000	X10 0.70320 <.0001	X7 0.55224 <.0001	Y -0.51099 <.0001	X1 -0.49043 <.0001	X11 -0.40334 0.0014	X5 -0.39507 0.0018	X12 0.28683 0.0263	X8 -0.24388 0.0604	X3 -0.23854 0.0664	X14 -0.23435 0.0715	X13 0.22440 0.0848	X9 -0.20877 0.1094	X15 0.17649 0.1774	X4 -0.13886 0.2900	X2 0.11628 0.3763
X7 **X7**	X7 1.00000	X11 -0.68068 <.0001	X6 0.55224 <.0001	X1 -0.49076 <.0001	Y -0.42682 0.0007	X3 -0.41503 0.0010	X5 -0.41059 0.0011	X9 -0.41033 0.0011	X12 0.38677 0.0023	X13 0.34825 0.0064	X10 0.33875 0.0081	X8 0.18188 0.1643	X15 0.12190 0.3535	X14 0.11795 0.3694	X4 0.06501 0.6217	X2 0.01485 0.9103
X8 **X8**	X8 1.00000	X14 0.43209 0.0006	Y 0.26550 0.0403	X6 -0.24388 0.0604	X5 -0.18433 0.1586	X7 0.18188 0.1643	X13 0.16531 0.2069	X11 -0.16295 0.2135	X4 0.16199 0.2162	X15 -0.12498 0.3414	X12 0.12028 0.3600	X2 -0.10005 0.4469	X3 -0.06099 0.6434	X10 -0.03177 0.8096	X9 -0.00568 0.9657	X1 -0.00352 0.9787
X9 **X9**	X9 1.00000	X11 0.70492 <.0001	Y 0.64374 <.0001	X4 -0.63782 <.0001	X3 0.57531 <.0001	X2 0.45377 0.0003	X5 0.41941 0.0009	X1 0.41320 0.0010	X7 -0.41033 0.0011	X6 -0.20877 0.1094	X14 0.15929 0.2241	X15 -0.11796 0.3694	X12 -0.02586 0.8445	X13 0.01839 0.8891	X8 -0.00568 0.9657	X10 -0.00439 0.9735
X10 **X10**	X10 1.00000	X6 0.70320 <.0001	X5 -0.42572 0.0007	X7 0.33875 0.0081	X1 -0.29729 0.0211	Y -0.28480 0.0274	X2 0.23799 0.0671	X12 0.20367 0.1186	X11 -0.18516 0.1567	X13 0.16003 0.2219	X4 -0.11771 0.3704	X14 -0.06846 0.6032	X15 0.06071 0.6449	X8 -0.03177 0.8096	X3 -0.02141 0.8710	X9 -0.00439 0.9735
X11 **X11**	X11 1.00000	X9 0.70492 <.0001	X7 -0.68068 <.0001	X3 0.61931 <.0001	X2 0.56531 <.0001	X1 0.50659 <.0001	Y 0.41049 0.0011	X6 -0.40334 0.0014	X4 -0.30977 0.0160	X5 0.25990 0.0449	X10 -0.18516 0.1567	X8 -0.16295 0.2135	X15 -0.15222 0.2456	X12 -0.12978 0.3230	X13 -0.10254 0.4356	X14 -0.09648 0.4633
X12 **X12**	X12 1.00000	X13 0.98384 <.0001	X1 -0.53176 <.0001	X5 -0.38821 0.0022	X7 0.38677 0.0023	X3 -0.35649 0.0052	X2 0.35081 0.0060	X6 0.28683 0.0263	X14 0.28230 0.0289	X10 0.20367 0.1186	Y -0.17724 0.1755	X11 -0.12978 0.3230	X8 0.12028 0.3600	X9 -0.02586 0.8445	X4 -0.02049 0.8765	X15 -0.02018 0.8784
X13 **X13**	X13 1.00000	X12 0.98384 <.0001	X1 -0.48732 <.0001	X14 0.40939 0.0012	X5 -0.35843 0.0049	X7 0.34825 0.0064	X3 -0.33767 0.0083	X2 0.32101 0.0124	X6 0.22440 0.0848	X8 0.16531 0.2069	X10 0.16003 0.2219	X11 -0.10254 0.4356	Y -0.07738 0.5568	X15 -0.04591 0.7276	X9 0.01839 0.8891	X4 -0.00208 0.9874
X14 **X14**	X14 1.00000	X8 0.43209 0.0006	Y 0.42590 0.0007	X13 0.40939 0.0012	X12 0.28230 0.0289	X6 -0.23435 0.0715	X9 0.15929 0.2241	X7 0.11795 0.3694	X2 -0.10781 0.4123	X1 -0.10692 0.4161	X15 -0.10255 0.4356	X3 -0.09935 0.4501	X11 -0.09648 0.4633	X10 -0.06846 0.6032	X4 0.01725 0.8959	X5 -0.00408 0.9753
X15 **X15**	X15 1.00000	X3 -0.45281 0.0003	X6 0.17649 0.1774	X11 -0.15222 0.2456	X5 -0.13574 0.3011	X8 -0.12498 0.3414	X7 0.12190 0.3535	X9 -0.11796 0.3694	X4 0.11243 0.3924	X14 -0.10255 0.4356	Y -0.08850 0.5013	X1 -0.07734 0.5570	X2 0.06787 0.6064	X10 0.06071 0.6449	X13 -0.04591 0.7276	X12 -0.02018 0.8784

(2). By checking the tolerances and VIFs of X_{12} and X_{13}, with the criterion Tolerance < 0.1 and VIF > 10, multicollinarity does exist between X_{12} and X_{13}.

参数估计								
变量	标签	自由度	参数估计	标准误差	t 值	Pr > |t|	容差	方差膨胀
Intercept	Intercept	1	1763.99793	437.33031	4.03	0.0002	.	0
X1	X1	1	1.90536	0.92374	2.06	0.0451	0.24308	4.11389
X2	X2	1	-1.93762	1.10839	-1.75	0.0874	0.16277	6.14355
X3	X3	1	-3.10040	1.90167	-1.63	0.1102	0.25203	3.96777
X4	X4	1	-9.06517	8.48622	-1.07	0.2912	0.13387	7.47004
X5	X5	1	-106.83103	69.78007	-1.53	0.1329	0.23215	4.30762
X6	X6	1	-17.15689	11.86012	-1.45	0.1551	0.20574	4.86054
X7	X7	1	-0.65111	1.76777	-0.37	0.7144	0.25033	3.99478
X8	X8	1	0.00360	0.00403	0.89	0.3761	0.60303	1.65828
X9	X9	1	4.45958	1.32721	3.36	0.0016	0.14750	6.77960
X10	X10	1	-0.18715	1.66169	-0.11	0.9108	0.35192	2.84158
X11	X11	1	-0.16741	3.22730	-0.05	0.9589	0.11472	8.71707
X12	X12	1	-0.67216	0.49102	-1.37	0.1780	0.01014	98.63993
X13	X13	1	1.34010	1.00559	1.33	0.1895	0.00953	104.98240
X14	X14	1	0.08626	0.14752	0.58	0.5617	0.23647	4.22893
X15	X15	1	0.10674	1.16943	0.09	0.9277	0.52436	1.90709

(3). The process of model selection is shown as follows. X_9, X_6, X_2, X_{14} and X_1 are added to model sequentially, then X_6 is removed from 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	BIC	CP	SBC
0	Intercept		1	0.0000	558.6471	496.3652	129.1367	498.7414
1	X9		2	0.4043	528.5396	466.3202	53.5866	470.7283
2	X6		3	0.5473	513.0205	451.2771	27.8371	457.3036
3	X2		4	0.6198	503.4928	442.5757	15.5320	449.8702
4	X14		5	0.6605	497.6139	437.7450	9.2216	446.0856
5	X1		6	0.6907*	492.9298*	434.4851*	4.9784*	443.4958
6		X6	5	0.6751	494.9860	435.5642	6.6837	443.4577*
* 准则的最佳值								



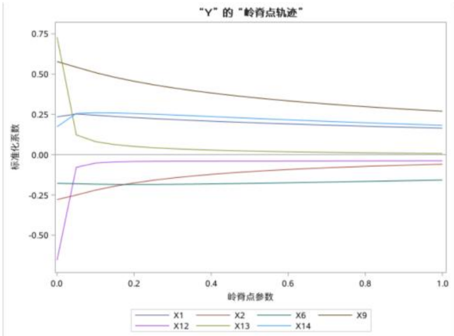
The final model is stated as

$$Y = 857.43 + 2.06\ X_1 - 1.77\ X_2 + 4.08\ X_9 + 0.33\ X_{14} + \varepsilon$$

参数估计				
参数	自由度	估计	标准误差	t 值
Intercept	1	857.431124	26.234966	32.68
X1	1	2.059246	0.524393	3.93
X2	1	-1.771640	0.527540	-3.36
X9	1	4.078669	0.671468	6.07
X14	1	0.330551	0.077299	4.28

This linear model describes the relationship between the total age-adjusted mortality rate (per 100,000) and five explanatory variables. The intercept term of 857.43 indicates that when all explanatory variables are zero, the expected total age-adjusted mortality rate is 857.43. For the coefficients of each explanatory variable, the interpretations are as follows: The coefficient of X_1 is 2.06, meaning that when the mean annual precipitation increases by one unit (in inches), the total age-adjusted mortality rate is expected to increase by 2.06 units. The coefficient of X_2 is -1.77, indicating that when the mean January temperature increases by one unit (in degrees Fahrenheit), the total age-adjusted mortality rate is expected to decrease by 1.77 units. The coefficient of X_9 is 4.08, suggesting that when the percent of nonwhite population increases by one unit, the total age-adjusted mortality rate is expected to increase by 4.08 units. Lastly, the coefficient of X_{14} is 0.33, signifying that for each unit increase in the relative pollution potential of sulfur dioxide, the total age-adjusted mortality rate is expected to increase by 0.33 units. ε represents the error term, which accounts for the unexplained portion of the model.

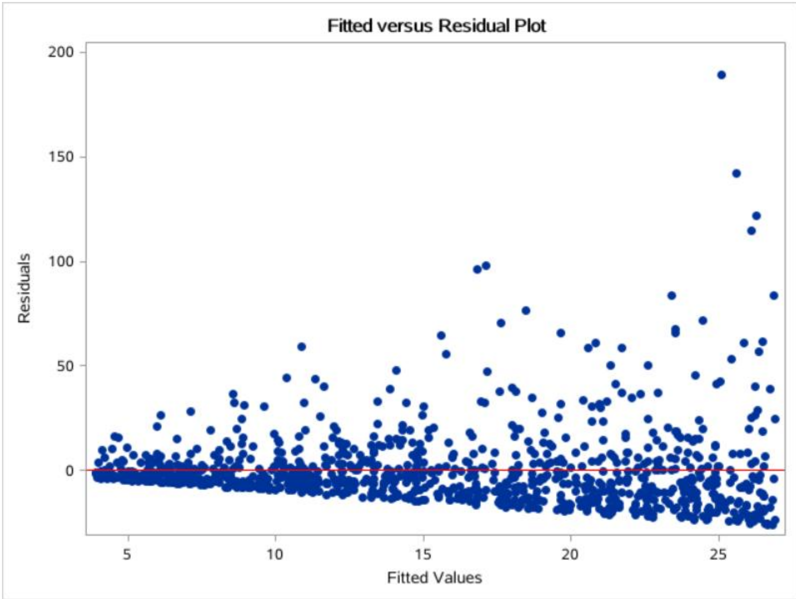
(4). The lines showing the parameter estimates of X_1, X_2, X_6, X_9, X_{12}, X_{13} and X_{14} against λ is plotted as follows.



From the plot, it's evident that as λ increases slightly, the coefficients of X_{12} and X_{13} experience rapid absolute value decreases, which aligns with their high sample correlation coefficient of 0.98 as noted in question 5.(1). These coefficients quickly converge towards zero, exhibiting a nearly symmetrical pattern about the zero line. Additionally, the effects of X_1 , X_2 , X_6 , and X_9 seem to be initially overestimated, with their absolute values decreasing as λ increases and stabilizing at non-zero values. Conversely, the effect of X_{14} appears to be initially underestimated, with its coefficient slightly increasing as λ increases. Notably, coefficients tend to stabilize around λ = 0.2, suggesting that estimates at this level of regularization are more suitable for assessing the explanatory variables' effects.

6. Solution.

(1). The Fitted versus Residual plot is shown below. Obviously it does not follows the homoscedasticity assumption.



(2). After the transformation of $Y_i' = \log Y_i$, the transformed plot is as follows. And now it seems more like a distribution of homoscedasticity.