STAT 602 Final Project
Analysis Based on Medical Clinic Data

WENJIN LI, SHURONG GU, WEN HUANG
May 12, 2017

1 Introduction

The paper is based on data sets **e** and **o**. There are 568 rows and 201 columns in each dataset, which implies 568 patients are recorded and tracked after they came to hospital for the first diagnose. Concentrations of 200 kinds of genes are chosen as prediction variables. The two response variables are *daystolastfollowup*, which is time of duration between the date of their first visit date and the date they die and *TP53*, which implies concentration of TP53. In our datasets TP53 is not included in the predictors. Since there are too many predictors in raw data sets, the given R code is computed to select variables which has relatively larger GMC with respect to each response. When the response variable is *daystolastfollowup*, 31 and 30 estimators are selected from data set **e** and **o**, while 29 and 30 variables are considered as significant if *TP53* is the response.

The rest of the paper is organized as follows. Section 2 describes the linear relationship between gene concentration and responses. Section 3 provides a simple framework to understand the effect of penalty on different functions, and a relatively best fitted model. Section 4 describes the data as binomial distributed, and applies logistic regression on data. Section 5 concludes the paper.

2 Linear Regression

2.1 Days to Last Follow Up as Response

2.1.1 Based on Dataset e

Using the 31 screened variables, we performed an ordinary linear regression on the response variable *daystolastfollowup*. As the residuals have heavy tails, we did a log transformation on the response. However, in case of zero, the transformation function adds 1 to all the response values first and take the logarithms. The 496th observation is considered as a high influential and high leverage point, and the 467th observation is a high residual point based on model diagnosis. Notice the 467th observation is also a special case with $daystolastfollowup = 0$, which can be considered as an abnormal record.

After discarding two outliers, for the remaining 31 variables, we performed Mallow's Cp and stepwise selection method, using both AIC and BIC as criteria to realize the further variable

selection and reduce the number of variables to 6, 11, 3, respectively. Notice that the model using AIC as criteria has the largest number of variables and it contains all the variables in other models. In order to fit a best model, we saw that the differences between three models are significant, which means all the 11 variables are important to improve our model. Therefore, we decided to select 11 genes (TMEFF1, TRIM36, FLG, SPINK1, CDKN2AIP, MYLPF, PYY, HPCA, DUSP8, FNDC4 and PPAP2B) as final predictors. Model detail can be found in Table 1 and Figure 1.

2.1.2 Based on Dataset o

The behavior of response in set o does not have a heavy tail trend, so we simply used the original data to fit a linear regression model. The diagnosis plots illustrate that the 496th observation is a high influential and high leverage point, while the 310th observation is a high residual point. Both of the outliers are removed in the following analysis.

Use the same method in the last section, that is, stepwise selection with AIC, BIC criteria and Mallow's Cp, to do further variable selection. Since the P-value is pretty small, the result of analysis of variance gives out enough evidence to accept the more complicate model rather than the simpler one. Thus, the AIC model with 10 prediction variables (UGDH, SLC25A23, ITIH1, TRIM36, PYY, GJA8, FLG, FNDC4, GGT1 and CAMK2G) is chosen to be our final model. Model detail can be found in Table 1 and Figure 1.

2.2 TP53 as Response

2.2.1 Based on Dataset e

First we obtained a full model based on 29 selected estimators to predict TP53 and the QQ plot of residuals performs well. However, the 392th and the 574th points have high residual, which should be discarded from the model. We then use stepwise selection with AIC, BIC criteria and Mallow's Cp to do further selection. After comparison of variance, a model with 10 prediction variables (HIST1H4I, TMEM87A, RBM4, KIAA0556, CTCF, IFT122, AHI1, CWF19L1, FGFR1OP and STAT5B) is chosen as our final model. Model detail can be found in Table 2 and Figure 2.

2.2.2 Based on Dataset o

The precedure to infer a reasonable model is similar to the last section. First notice that there is no significant outliers in full model based on dataset o. After comparing the three models derived from stepwise selection with AIC, BIC criteria and Mallow's Cp, ten variables are considered as significant estimators. Model detail can be found in Table 2 and Figure 2.

3 GMC Variable Selections

According to investigate the correlation between response and predictors, models are chosen from different criterions with diverse penalty terms against $GMC(Y|\hat{Y})$. We are going to fit the

following five function and find reasonable values of penalties. Denote $X = \beta_0 + \sum_{i=1}^{p} \beta_{Xi}$ where X_i is the scaled estimator, and

- $g(X) = X$
- $g(X) = X^2$
- $g(X) = X^3$
- $g(X) = e^X$
- $g(X) = 1/X$

First note that both multivariate and univariate function should be taken into consideration. That is, for example, for $g(X) = X^2$, the corresponding multivariate function is

$$Y = \beta_0 + \beta_1 X_1^2 + \cdots + \beta_p X_p^2 + \epsilon$$

while the univariate function is

$$Y = X^2 = (\beta_0 + \sum_{i=1}^{p} \beta_{Xi})^2.$$

A fact is, from our R output, the optimization of $GMC(Y|\hat{Y})$ is always achieved when we use the multivariate function.

Since all of the coefficients derived from the previous analysis are moderately small, in this case, penalty terms can be chosen from a pretty narrow range and have small values. To maximize

$$\frac{Var(g(X))}{Var(g(X)) + Var(\epsilon)} - \lambda_1 |Cov(g(X), \epsilon)| - \lambda_2 (Lasso) \tag{1}$$

and

$$GMC(Y|G(X)) - \lambda (Lasso) \tag{2}$$

seperately, the algorithm is following:

Step 1 Fit a full model $Y = \beta_0 + \beta_1 g(X_1) + \cdots + \beta_p g(X_p)$ or $Y = g(\beta_0 + \sum_{i=1}^{p} \beta_i X_i)$ by least square error method to find $\beta_{i.LSE}$.

Step 2.a Choose λ_1 and λ_2.

Step 2.b Choose λ.

Step 3.a Start from $\beta_{i.LSE}$, optimize Equation (1) and obtain the new parameters $\hat{\beta}_i$.

Step 3.b Start from $\beta_{i.LSE}$, optimize Equation (2) and obtain the new parameters $\hat{\beta}_i$.

Step 4 For each function g, find the model which gives out the best $GMC(Y|\hat{Y})$ and its corresponding λ.

3.1 Days to Last Follow Up as Response

When the response is *Daystolastfollowup*, we choose λ_1 change between $\{0.001, 0.002, \cdots, 0.009\}$ and $\lambda_2 \in \{0.01, 0.02, \cdots, 0.09\}$. For each function $g(X)$, for each combination of λ_1 and λ_2, a model

maximizes (1) is obtained. We derive 405 models from the algorithm in total. Comparing these models by $GMC(Y|\hat{Y})$, a most reasonable one is selected for each g, which is provided by Table 3 - 6. The largest value of GMC is reached by $g(X) = e^X$ for both dataset and the corresponding models are

$$\begin{aligned}\textbf{Set e}: Days &= 919.03 + 25.49\exp(TMEFF1) + 3.14\exp(CEBPA) + 1.78\exp(CADM3) + 14.11\exp(UBXD8) \\ &\quad - 13.52\exp(AHI1) - 0.52\exp(TRIM36) + 0.24\exp(HPCA) + 0.94\exp(CEL) + 0.04\exp(SPINK1) + \epsilon\end{aligned}$$

and

$$\begin{aligned}\textbf{Set o}: Days &= 803.52 + 32.60\exp(ITIH1) + 49.21\exp(UGDH) + 2.34\exp(GGT1) + 9.27\exp(SDK2) \\ &\quad + 5.36\exp(SLC25A23) - 8.34\exp(RAD1) + 7.54\exp(GSN) + 1.57\exp(CADM3) + 0.98\exp(TGDS) + \epsilon\end{aligned}$$

Similarly, to maximize (2), the best tion is $g(X) = e^X$ for set e and X^3 for set o, and models are

$$\begin{aligned}\textbf{Set e}: Days &= 1048.87 + 0.05\exp(SPINK1) + 1.42\exp(CEL) - 0.26\exp(SOX2) + 0.12\exp(HPCA) \\ &\quad - 30.61\exp(BTG1) + 1.79\exp(CADM3) - 8.85\exp(4.62) - 21.24\exp(PSMA1) + \epsilon \\ \textbf{Set o}: Days &= 990.41 + 85.59SPINK1^3 + 56.78CEL^3 + 33.11UGDH^3 + 25.08SLC25A23^3 \\ &\quad - 13.71RAD1^3 + 1.31CHORDC1^3 + 90.67SDK2^3 - 17.92NUPL2^3 + 28.73GSN^3 + \epsilon\end{aligned}$$

3.2 TP53 as Response

As in the previous section, from the result in Table 7 - 10, to maximize (1) we have $g(X) = X$ and models are

$$\begin{aligned}\textbf{Set e}: TP53 &= 5.22 + 0.16TMEM87A + 0.15HIST1H4I + 0.08IFT122 + 0.11CWF19L1 \\ &\quad + 0.10FGFR1OP + 0.11KIAA0556 + 0.06CTCF + 0.06NAG + 0.06STAT5B + 0.06BTG1 + \epsilon \\ \textbf{Set o}: TP53 &= 5.22 + JMJD3 + 0.14MYH9 + 0.13NUP88 + 0.07PDK2 + 0.07TMEM5 \\ &\quad + 0.05KCTD5 + 0.06SNRPC + 0.05RYK + 0.05CASP2 + 0.04UNC119B + \epsilon\end{aligned}$$

and to maximize (2), $g(X) = X$, with best models

$$\begin{aligned}\textbf{Set e}: TP53 &= 5.22 + 0.05NFYC + 0.05EXOSC5 + 0.09CEP135 + 0.10BTG1 \\ &\quad + 0.16HIST1H4T - 0.04FOLR3 + 0.11CWF19L1 + 0.14NAG - 0.03AHI1 + \epsilon \\ \textbf{Set o}: TP53 &= 5.22 + 0.05NFYC + 0.05EXOSC5 + 0.09CEP135 + 0.10BTG1 \\ &\quad + 0.16HIST1H4T - 0.04FOLR3 + 0.11CWF19L1 + 0.14NAG - 0.03AHI1 + \epsilon\end{aligned}$$

4 Logistic Regression Based on Days

For days to last follow up, each observations can be divided into one of the two groups: stay alive for more than n days and less than n days. A question is, is there a reasonable critical point n such that genes of individuals live more than n days have significant different concentrations, comparing to genes of the other group? Logistic regression is applied here to solve it.

4.1 Based on Dataset e

After picking out the 31 estimators given in previous section, we first traverse all potential values of n and fit the full model. The best critical points of link function logit, probit and cloglog are 3400, 3200 and 3400 based on AIC, respectively. In order to reduce the number of variables, we used stepwise method and obtained three final generalized linear models, which are shown in Table 11.

$$\begin{aligned}
Logit(I(DaysToLastFollowUp > 3400)) &= -64.18 + 3.09ACCN1 + 1.50HABP2 + 1.51BTG1 \\
&\quad + 0.83PYY + 1.69RB1CC1 + 3.73HPCA \\
Probit(I(DaysToLastFollowUp > 3200)) &= -14.69 + 0.90ACCN1 - 0.59UBXD8 + 0.67HABP2 \\
&\quad + 0.31PYY + 0.69RB1CC1 + 0.22HOXC10 + 3.73HPCA \\
Loglog(I(DaysToLastFollowUp > 3400)) &= -64.47 + 3.67ACCN1 + 0.73TMEFF1 + 1.45HABP2 \\
&\quad + 1.54BTG1 + 0.94PYY + 1.52RB1CC1 - 1.36AHI1 + 4.08HPCA
\end{aligned}$$

4.2 Based on Dataset o

Similar to section 4.1, When the critical value equals to 3000, AIC is minimized for all of the link function. The result is shown in Table 12.

$$\begin{aligned}
Logit(I(DaysToLastFollowUp > 3000)) &= -20.25 + 0.45HOXC10 + 0.77UGDH - 1.63SLC25A23 \\
&\quad + 7.09ITIH1 - 6.04PROP1 - 1.49NSD1 + 0.35CEL + 0.75PYY + 2.36GGT1 + 1.48GAMK2G \\
Probit(I(DaysToLastFollowUp > 3000)) &= -15.48 + 0.20HOXC10 + 0.56UGDH + 3.61ITIH1 \\
&\quad - 3.05PROP1 - 0.75NSD1 + 0.37PYY + 0.45GJA8 + 1.16GGT1 + 0.65GAMK2G \\
Loglog(I(DaysToLastFollowUp > 3000)) &= -25.43 + 0.44HOXC10 + 0.87UGDH + 2.96SDK2 - 1.46SLC25A23 \\
&\quad + 5.68ITIH1 - 5.49PROP1 - 1.55NSD1 + 0.36CEL + 0.69PYY + 2.00GGT1 + 1.10GAMK2G
\end{aligned}$$

5 Conclusion

For ordinary linear regression model in Section 2, significant estimators of *Daystolastfollowup* and *TP53* are quite different, which may illustrate the two response variables are not relative. Here is an interesting finding that when the response is *Daystolastfollowup*, the concentrations of genes PYY and FNDC4 are significant for both datasets, and their symbols are consistent, + and -, respectively. Hence, it can be considered that gene PYY has a positive effect on the survival time, while for FNDC4, the conclusion is opposite. Output of logistic regression confirms that a higher concentration of gene PYY increases the probability that survival time is longer than 3000 days.

The results in Section 3 are clear and encouraging. We see that all coefficients in the models show significant correlation of the concentration of gene CEL and CADM3 with the survival times. Let us emphasize that in pancreatic tumoral cells, encoded protein on gene CEL is thought to be probably not secreted, which is an evidence consistent with the finding.

There are several limitations in our algorithms. Since *optim* function we used in Section 3 can only find the best local value based on the initial one, we can't obtain a best value on the whole real number field. Also, we can improve the result by applying stepwise or other variable selection method into this part.

6 Figures

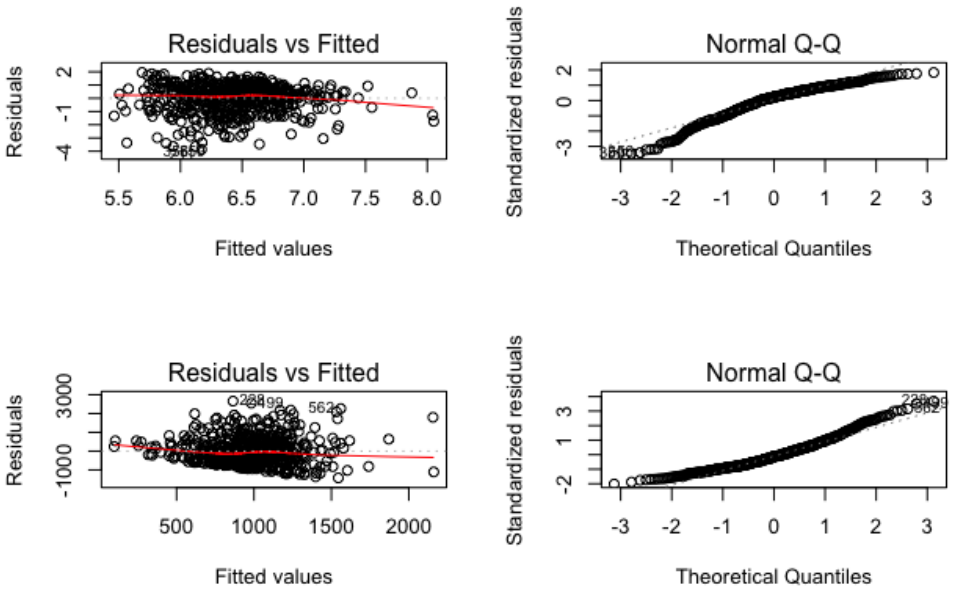

Figure 1: Lm(Dayslastfollowup ∼ estimators) Based on Data eo.

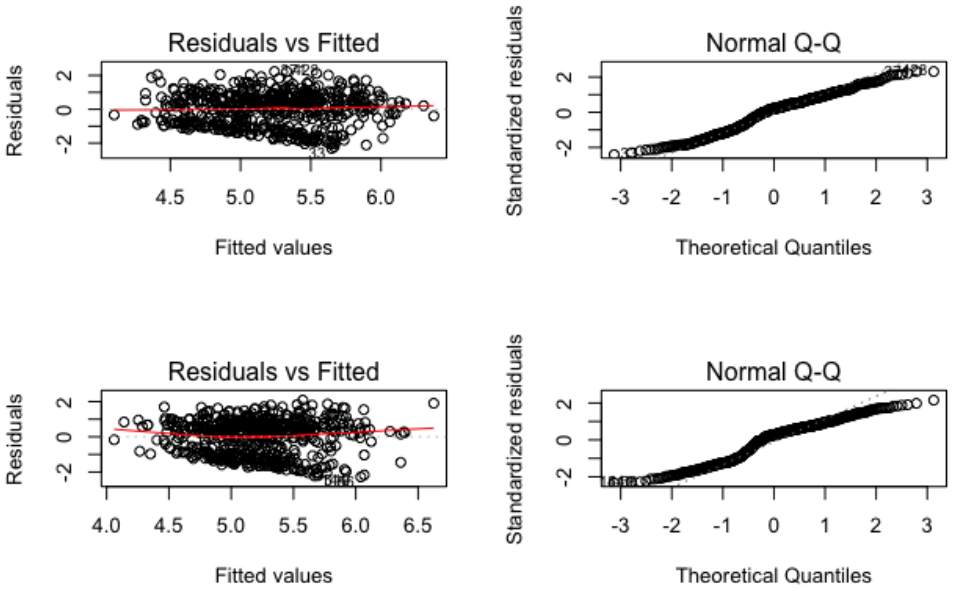

Figure 2: Lm(TP53 ∼ estimators) Based on Data eo.

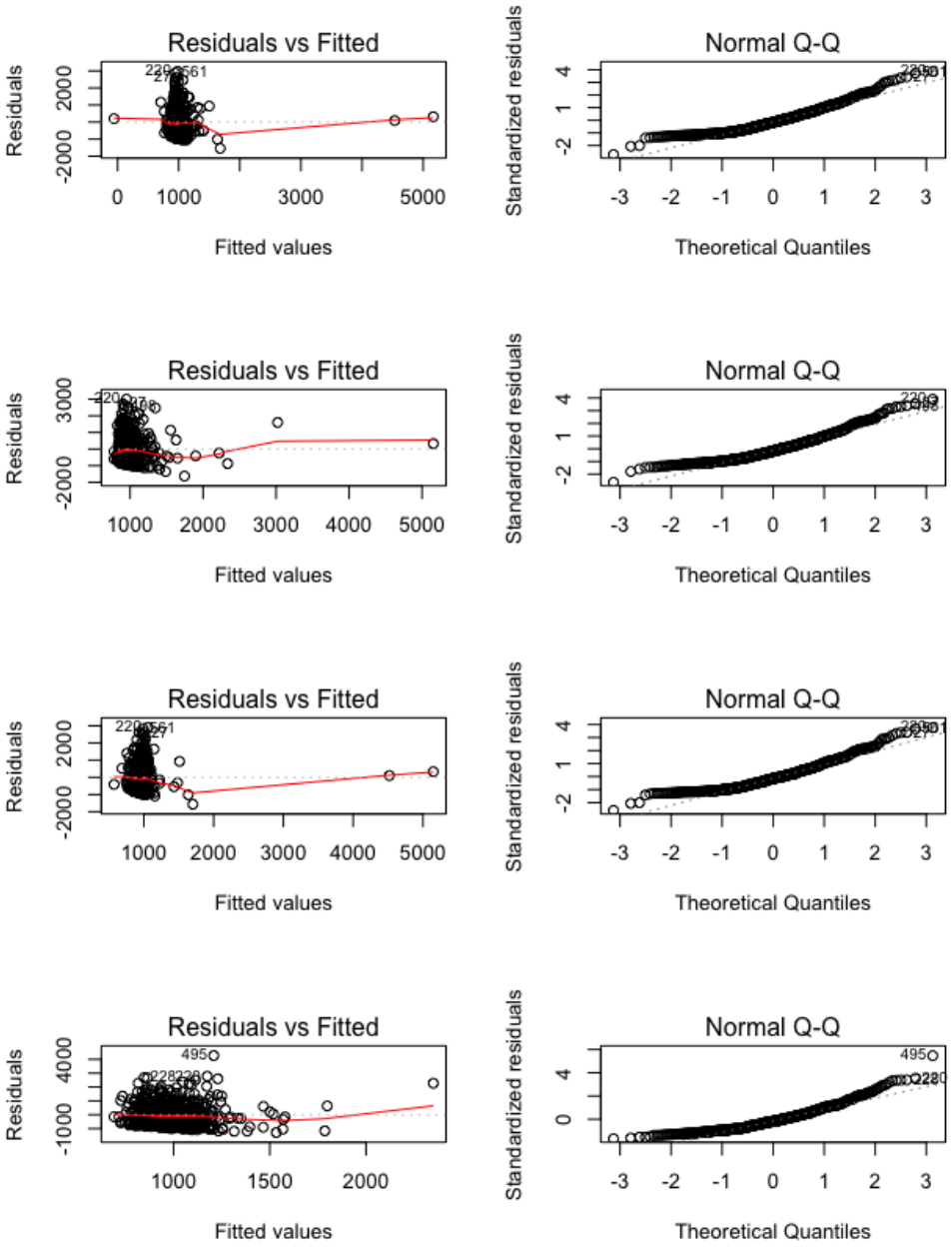

Figure 3: Models maximize (1), (2) and GMC, response = Daystolastfollowup

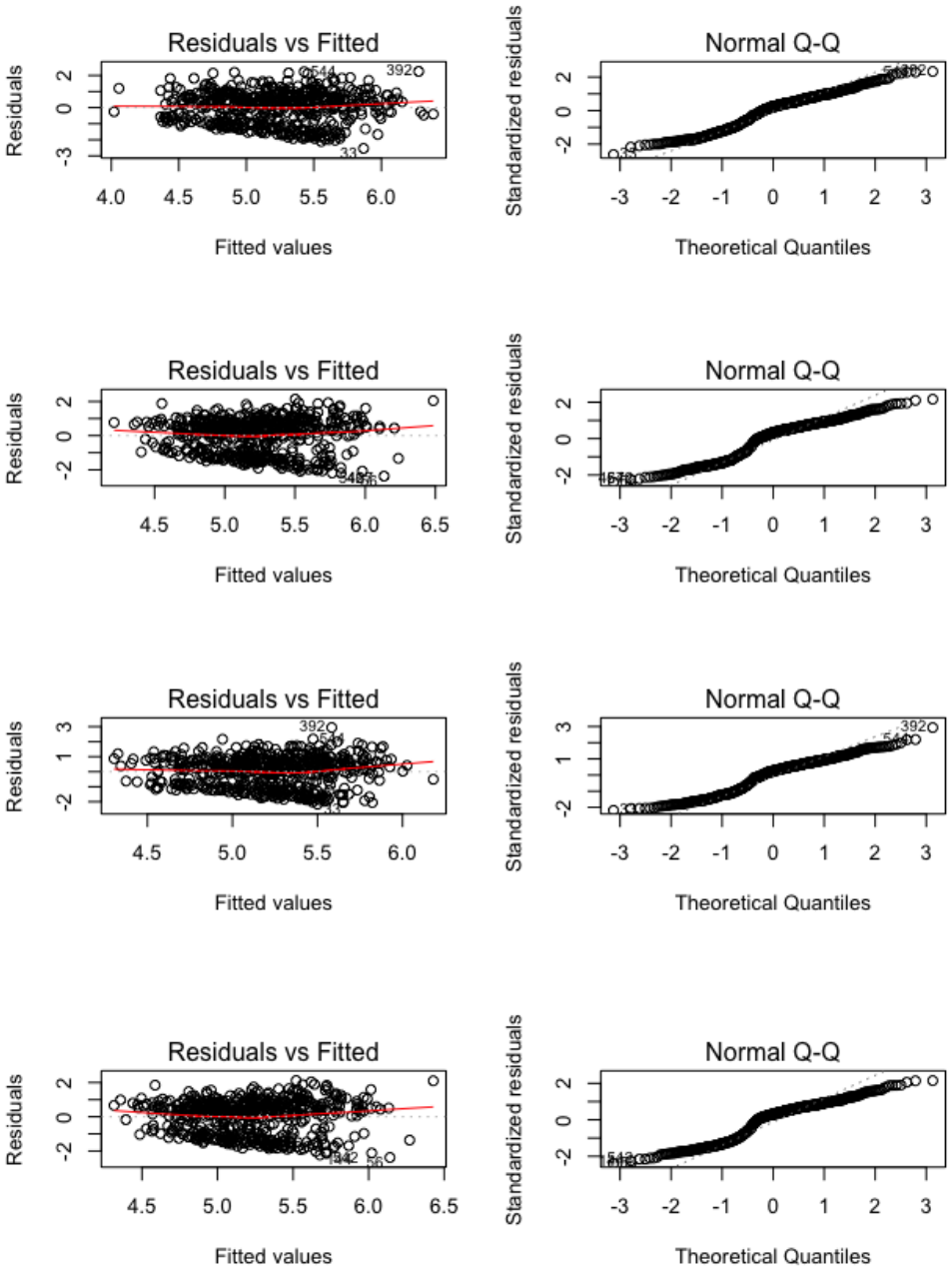

Figure 4: Models maximize (1), (2) and GMC, response = TP53

7 Tables

Table 1: Ordinary Linear Regression on Daystolastfollowup

lm(formula = response ~ TMEFF1 + TRIM36 + FLG + SPINK1 + CDKN2AIP + MYLPF + PYY + HPCA + DUSP8 + FNDC4 + PPAP2B, data = **Data.E**)

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	4.52210	1.37604	3.286	0.001079	**
TMEFF1	0.08052	0.04539	1.774	0.076600	.
TRIM36	-0.11720	0.07013	-1.671	0.095246	.
FLG	-0.08542	0.04210	-2.029	0.042935	*
SPINK1	0.10107	0.04871	2.075	0.038442	*
CDKN2AIP	0.14131	0.08125	1.739	0.082571	.
MYLPF	0.30447	0.11458	2.657	0.008102	**
PYY	0.12990	0.06990	1.858	0.063650	.
HPCA	0.82078	0.21863	3.754	0.000192	***
DUSP8	-0.15881	0.11170	-1.422	0.155669	
FNDC4	-0.51969	0.16356	-3.177	0.001569	**
PPAP2B	-0.10818	0.04214	-2.567	0.010512	*

Residual standard error: 1.075 on 554 degrees of freedom
Multiple R-squared: 0.113, Adjusted R-squared: 0.09542
F-statistic: 6.418 on 11 and 554 DF, p-value: 4.556e-10
AIC: 1702.088, BIC:1758.489

lm(formula = response ~ UGDH + SLC25A23 + ITIH1 + TRIM36 + PYY + GJA8 + FLG + FNDC4 + GGT1 + CAMK2G, data = **Data.O**)

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2217.92	1203.08	-1.844	0.065783	.
UGDH	122.39	48.07	2.546	0.011163	*
SLC25A23	84.82	59.92	1.415	0.157485	
ITIH1	763.83	230.68	3.311	0.000989	***
TRIM36	-77.79	46.95	-1.657	0.098087	.
PYY	148.69	47.95	3.101	0.002028	**
GJA8	151.88	77.64	1.956	0.050942	.
FLG	-50.65	28.56	-1.774	0.076640	.
FNDC4	-452.73	114.13	-3.967	8.23e-05	***
GGT1	378.38	146.90	2.576	0.010258	*
CAMK2G	-100.37	69.27	-1.449	0.147893	

Residual standard error: 728.9 on 555 degrees of freedom
Multiple R-squared: 0.1136, Adjusted R-squared: 0.09763
F-statistic: 7.113 on 10 and 555 DF, p-value: 1.405e-10
AIC: 9080.727, BIC:9132.79

Table 2: Ordinary Linear Regression on Daystolastfollowup

lm(formula = response ∼ HIST1H4I + TMEM87A + RBM4 + KIAA0556 + CTCF + IFT122 + AHI1 + CWF19L1 + FGFR1OP + STAT5B, data = **Data.E**)

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-5.60109	1.37777	-4.065	5.49e-05	***
HIST1H4I	0.74443	0.20207	3.684	0.000252	***
TMEM87A	0.26335	0.06803	3.871	0.000121	***
RBM4	-0.25829	0.11637	-2.220	0.026850	*
KIAA0556	0.26477	0.09409	2.814	0.005066	**
CTCF	0.25014	0.09733	2.570	0.010431	*
IFT122	0.13975	0.05774	2.420	0.015830	*
AHI1	-0.18245	0.09625	-1.896	0.058535	.
CWF19L1	0.31395	0.10441	3.007	0.002758	**
FGFR1OP	0.33037	0.10893	3.033	0.002535	**
STAT5B	0.28552	0.15103	1.891	0.059210	.

Residual standard error: 0.9692 on 555 degrees of freedom
Multiple R-squared: 0.1578, Adjusted R-squared: 0.1426
F-statistic: 10.4 on 10 and 555 DF, p-value: 3.48e-16
AIC: 1583.77, BIC:1635.833

lm(formula = response ∼ MYH9 + PDK2 + JMJD3 + RPP38 + NUP88 + SNRPC + UNC119B + RAD1 + CASP2 + MRS2L,data = **Data.O**)

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.87019	1.01153	-2.837	0.004713	**
MYH9	0.18683	0.06547	2.854	0.004481	**
PDK2	0.17264	0.09539	1.810	0.070851	.
JMJD3	0.36965	0.11668	3.168	0.001618	**
RPP38	0.13025	0.08211	1.586	0.113251	
NUP88	0.16130	0.06993	2.307	0.021438	*
SNRPC	0.37077	0.10084	3.677	0.000259	***
UNC119B	0.09642	0.06592	1.463	0.144117	
RAD1	-0.23406	0.08121	-2.882	0.004101	**
CASP2	0.36197	0.13672	2.647	0.008339	**
MRS2L	-0.25614	0.08207	-3.121	0.001897	**

Residual standard error: 0.9853 on 557 degrees of freedom
Multiple R-squared: 0.1432, Adjusted R-squared: 0.1279
F-statistic: 9.312 on 10 and 557 DF, p-value: 2.393e-14
AIC: 1607.96, BIC:1660.046

Table 3: **GMC Optimization:**Lm(Daystolastfollowup $\sim$.), Data = Data.E

g(x)	λ_1	λ_2	GMC(Y | g(X))	Variable Selection
x	0.001	0.01	0.322	HPCA, MYLPF, PYY, SOX2, SPINK1 TMEM30A, CEL, HABP2, TMEFF1
x^2	0.001	0.08	0.53480575	SPINK1, DUSP8, SOX2, TRIM36, CEL HABP2, CADM3, MYLPF, HOXC10
x^3	0.001	0.07	0.6707302	TMEFF1, CEBPA, CADM3, UBXD8, AHI1 TRIM36, CEL, HPCA, SPINK1
e^x	0.001	0.03	0.8565	TMEM30A, TMEFF1, UBXD8, CDKN2AIP, ACCN1 RB1CC1, BTG1, SOX2, CADM3
$\frac{1}{x}$	0.002	0.04	0.34163	TMEM30A, TMED1, HABP2, CEBPA, UBXD8 RB1CC1, CADM3, SOX2, PYY

Table 4: **GMC Optimization:**Lm(Daystolastfollowup $\sim$.), Data = Data.O

g(x)	λ_1	λ_2	GMC(Y | g(X))	Variable Selection
x	0.001	0.01	0.28579	ITIH1, UGDH, PYY, GJA8, GGT1 SDK2, SPINK1, CADM3, SLC25A23
x^2	0.001	0.06	0.58386921	SPINK1, ITIH1, KIAA1045, NR0B2, NSD1 CEL, CAMK2G, NUPL2, CHORDC1
x^3	0.001	0.09	0.5645333	UGDH, PROP1, GSN, ITIH1, FNDC4 CHORDC1, SPINK1, C16orf45, SLC25A23
e^x	0.001	0.06	0.8804978	ITIH1, UGDH, GGT1, SDK2, SLC25A23 RAD1, GSN, CADM3, TGDS
$\frac{1}{x}$	0.004	0.09	0.34163	ITIH1, PRRX2, NSD1, CADM3, FLG PROP1, KIAA1045, SCGB1A1, HABP2

Table 5: **GMC Optimization:**Lm(TP53 $\sim$.), Data = Data.E

g(x)	λ_1	λ_2	GMC(Y | g(X))	Variable Selection
x	0.008	0.01	0.2126656	TMEM87A, HIST1H4I, IFT122, CWF19L1, BTG1 SDK2, SPINK1, CADM3, SLC25A23
x^2	0.009	0.01	0.08718180	FGFR1OP, SLC7A4, STAT5B, RGL2, RBBP5 IFT122, KIAA0556, NFYC, NAG, GPRASP1
x^3	0.009	0.01	0.1609107	FGFR1OP, KIAA0556, TMEM87A, RBBP5, HIST1H4I C16orf58, EXOSC5, CTCF, CWF19L1, IFT122
e^x	0.004	0.01	0.1777479	TMEM87A, FGFR1OP, ABCF1, KIAA0556, CTCF EXOSC5, HIST1H4I, RBBP5, NAG, IFT122
$\frac{1}{x}$	0.009	0.09	0.03477509	ABCF1, AHI1, RBM4, TMEM30A, NAG KIAA0556, ZBTB5, C16orf58, TMEM87A

Table 6: **GMC Optimization:**Lm(TP53 $\sim$.), Data = Data.O

g(x)	λ_1	λ_2	GMC(Y | g(X))	Variable Selection
x	0.007	0.01	0.2783180	JMJD3, MYH9, NUP88, PDK2, TMEM5 KCTD5, SNRPC, RYK, CASP2, UNC119B
x^2	0.007	0.01	0.06865822	PDK2, KCTD5, MYH9, LYRM2, MRPS18A JMJD3, RPP38, NUP88, CASP2, SLC2A1
x^3	0.002	0.01	0.1930473	SNRPC, PDK2, NUP88, RPP38, MYH9 UNC119B, KCTD5, CACNB3, CASP2, TM9SF1
e^x	0.001	0.06	0.2738609	MJD3, MRPS18A, C17orf71, NUP88, BTBD3 PIK3R4, SNRPC, TMEM5, CACNB3,RYK
$\frac{1}{x}$	0.007	0.09	0.03631872	UGDH, UNC119B, RYK, SLC2A1, NUP88 BTBD3, RAD1, PRKRA, PDK2, TMEM5

Table 7: **GMC Optimization:**Lm(Daystolastfollowup $\sim$.), Data = Data.E

g(x)	λ	GMC(Y | g(X))	Variable Selection
x	0.01	0.3285067	SPINK1, CEL, SOX2, PYY, HABP2 BTG1, RB1CC1, CDC37L1, UBXD8
x^2	0.01	0.2508411	PYY, MYLPF, HABP2, HPCA, TMEM30A MPHOSPH8, CDKN2AIP, TRIM36, PSMA1
x^3	10	0.2872726	HPCA, MYLPF, PYY, SOX2, SPINK1 TMEM30A, CEL, HABP2, TMEFF1
e^x	0.1	0.690312	SPINK1, CEL, SOX2, HPCA, BTG1 CADM3, CDC37L1, RB1CC1, PSMA1
$\frac{1}{x}$	0.001	0.2463075	PYY, HABP2, TRIM36, CDKN2AIP, CADM3 CDC37L1, TMEFF1, TMEM30A, MYLPF

Table 8: **GMC Optimization:**Lm(Daystolastfollowup $\sim$.), Data = Data.O

g(x)	λ	GMC(Y | g(X))	Variable Selection
x	0.01	0.337402	CEL, SPINK1, PYY, HABP2, SCGB1A1 UGDH, SLC25A23, CHORDC1, SDK2
x^2	0.1	0.5911118	SPINK1, CEL, UGDH, RAD1, SLC25A23 CHORDC1, TGDS, NUPL2, SDK2
x^3	1	0.6603686	SPINK1, CEL, UGDH, SLC25A23, RAD1 CHORDC1, SDK2, NUPL2, GSN
e^x	10	0.1245866	TIH1, UGDH, PYY, GJA8, GGT1 SDK2, SPINK1, CADM3, SLC25A23
$\frac{1}{x}$	0.01	0.3500089	SDK2, HABP2, CADM3, CHORDC1, SCGB1A1 GSN, UGDH, ITIH1, RAD1

Table 9: **GMC Optimization:**Lm(TP53 $\sim$.), Data = Data.E

g(x)	λ	GMC(Y | g(X))	Variable Selection
x	0.001	0.2365294	NFYC, EXOSC5, CEP135, BTG1, HIST1H4I FOLR3, CWF19L1, NAG, AHI1
x^2	0.1	0.02756474	CTCF, TMEM87A, NAG, ABCF1, FGFR1OP HIST1H4I, EXOSC5, BTG1, CEP76
x^3	10	0.1342547	TMEM87A, FGFR1OP, HIST1H4I, KIAA0556 CWF19L1, IFT122, CTCF, STAT5B, CEP135
e^x	10	0.1390977	TMEM87A, FGFR1OP, HIST1H4I, KIAA0556, CWF19L1 IFT122, CTCF, STAT5B, CEP135
$\frac{1}{x}$	0.001	0.02295907	TMEM87A, ABCF1, FGFR1OP, RBBP5, IFT122 STAT5B, BTG1, BAT1, EXOSC5

Table 10: **GMC Optimization:**Lm(TP53 $\sim$.), Data = Data.O

g(x)	λ	GMC(Y | g(X))	Variable Selection
x	10	0.3078768	SNRPC, NUP88, JMJD3, MYH9, PDK2 RYK, CASP2, RPP38, KCTD5
x^2	10	0.07807312	SNRPC, NUP88, JMJD3, MYH9, PDK2 RYK, CASP2, RPP38, KCTD5
x^3	10	0.16072633	SNRPC, NUP88, JMJD3, MYH9, PDK2 RYK, CASP2, RPP38, KCTD5
e^x	10	0.13297034	SNRPC, NUP88, JMJD3, MYH9, PDK2 RYK, CASP2, RPP38, KCTD5
$\frac{1}{x}$	0.001	0.0164724	NUP88, LYRM2, RAD1, JMJD3, RPP38 VPS37C, UNC119B, KCTD5, MRPL46, NGDN

Table 11: GLM Output Based on Dataset e

glm(formula = I(y>3400) ∼ ACCN1 + HABP2 + BTG1 + PYY + RB1CC1 + HPCA, family = binomial(link = "logit"), data = **Data.E**)

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-64.1811	15.1189	-4.245	2.18e-05	***
ACCN1	3.0898	1.1460	2.696	0.00701	**
HABP2	1.4954	0.6135	2.437	0.01479	*
BTG1	1.5116	0.7374	2.050	0.04037	*
PYY	0.8332	0.2106	3.956	7.63e-05	***
RB1CC1	1.6908	0.5996	2.820	0.00481	**
HPCA	3.7252	1.4468	2.575	0.01003	*

Null deviance: 84.090 on 567 degrees of freedom
Residual deviance: 53.586 on 561 degrees of freedom
AIC: 67.586

glm(formula = I(y>3200) ∼ ACCN1 + UBXD8 + HABP2 + PYY + RB1CC1 + HOXC10 + HPCA, family = binomial(link = "probit"), data = **Data.E**)

	Estimate	Std.	Error	z value	Pr(>|z|)
(Intercept)	-14.6873	4.3290	-3.393	0.000692	***
ACCN1	0.8978	0.4937	1.818	0.069024	.
UBXD8	-0.5857	0.2571	-2.278	0.022721	*
HABP2	0.6745	0.2795	2.413	0.015811	*
PYY	0.3135	0.1030	3.045	0.002330	**
RB1CC1	0.6917	0.2464	2.808	0.004993	**
HOXC10	0.2236	0.1288	1.736	0.082587	.
HPCA	0.9761	0.5211	1.873	0.061071	.

Null deviance: 108.559 on 567 degrees of freedom
Residual deviance: 76.324 on 560 degrees of freedom
AIC: 92.324

glm(formula = I(y>3400) ∼ ACCN1 + TMEFF1 + HABP2 + BTG1 + PYY + RB1CC1 + AHI1 + HPCA, family = binomial(link = "cloglog"), data = **Data.E**)

	Estimate	Std.	Error	z value	Pr(>|z|)
(Intercept)	-64.4694	13.8278	-4.662	3.13e-06	***
ACCN1	3.6661	1.0996	3.334	0.000856	***
TMEFF1	0.7251	0.4407	1.645	0.099883	.
HABP2	1.4537	0.5639	2.578	0.009943	**
BTG1	1.5434	0.5843	2.641	0.008254	**
PYY	0.9390	0.2186	4.296	1.74e-05	***
RB1CC1	1.5195	0.5222	2.910	0.003617	**
AHI1	-1.3568	0.9342	-1.452	0.146409	
HPCA	4.0771	1.4043	2.903	0.003693	**

Null deviance: 84.090 on 567 degrees of freedom
Residual deviance: 49.956 on 559 degrees of freedom
AIC: 67.956

Table 12: GLM Output Based on Dataset o

glm(formula = I(y>3000) ∼ HOXC10 + UGDH + SLC25A23 + ITIH1 + PROP1 + NSD1 + CEL + PYY + GGT1 + CAMK2G, family = binomial(link = "logit"), **Data.O**)

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-20.2498	12.7130	-1.593	0.111194	
HOXC10	0.4464	0.2585	1.727	0.084114	.
UGDH	0.7712	0.4787	1.611	0.107187	
SLC25A23	-1.6296	0.9637	-1.691	0.090828	.
ITIH1	7.0867	2.6174	2.708	0.006779	**
PROP1	-6.0449	2.8490	-2.122	0.033857	*
NSD1	-1.4850	0.8040	-1.847	0.064747	.
CEL	0.3558	0.2083	3.638	0.000275	***
GGT1	2.3640	1.3023	1.815	0.069491	.
CAMK2G	1.4806	0.5876	2.520	0.011748	*

Null deviance: 116.318 on 567 degrees of freedom
Residual deviance: 74.373 on 557 degrees of freedom
AIC: 96.373

glm(formula = I(y>3000) ∼ HOXC10 + UGDH + ITIH1 + PROP1 + NSD1 + PYY + GJA8 + GGT1 + CAMK2G, family = binomial(link = "probit"), data = **Data.O**)

	Estimate	Std.	Error	z value	Pr(>|z|)
(Intercept)	-15.4790	6.4007	-2.418	0.01559	*
HOXC10	0.1991	0.1374	1.449	0.14740	
UGDH	0.5603	0.2430	2.306	0.02113	*
ITIH1	3.6089	1.1854	3.044	0.00233	**
PROP1	-3.0470	1.3707	-2.223	0.02622	*
NSD1	-0.7452	0.3823	-1.949	0.05125	.
PYY	0.3729	0.1143	3.264	0.00110	**
GJA8	0.4515	0.2532	1.783	0.07457	.
GGT1	1.1566	0.6639	1.742	0.08146	.
CAMK2G	0.6547	0.2905	2.254	0.02422	*

Null deviance: 116.318 on 567 degrees of freedom
Residual deviance: 75.794 on 558 degrees of freedom
AIC: 95.794

glm(formula = I(y>3000) ∼ HOXC10 + UGDH + SDK2 + SLC25A23 + ITIH1 + PROP1 + NSD1 + CEL + PYY + GGT1 + CAMK2G, family = binomial(link = "cloglog"), data = **Data.O**)

	Estimate	Std.	Error	z value	Pr(>|z|)
(Intercept)	-25.4251	11.8572	-2.144	0.032011	*
HOXC10	0.4399	0.2163	2.034	0.041958	*
UGDH	0.8685	0.4400	1.974	0.048403	*
SDK2	2.9578	1.7892	1.653	0.098306	.
SLC25A23	-1.4579	0.9230	-1.579	0.114233	
ITIH1	5.6788	2.5395	2.236	0.025337	*
PROP1	-5.4897	2.5953	-2.115	0.034411	*
NSD1	-1.5520	0.7484	-2.074	0.038096	*
CEL	0.3563	0.1755	2.030	0.042371	*
PYY	0.6893	0.1832	3.763	0.000168	***
GGT1	2.0045	1.1839	1.693	0.090411	.
CAMK2G	1.1000	0.5099	2.157	0.030988	*

Null deviance: 116.318 on 567 degrees of freedom
Residual deviance: 70.926 on 556 degrees of freedom
AIC: 94.926