Solution: Homework 6

STA 207
Winter Quarter, 2016

1. (a) The correlation matrix is

	Y	X_1	X_2	X_3	X_4	X_5
Y	1					
X_1	−0.1145	1				
X_2	0.9235	−0.0142	1			
X_3	0.7414	−0.1886	0.8000	1		
X_4	0.2250	−0.3627	0.2241	0.1661	1	
X_5	0.9681	0.0270	0.8779	0.6727	0.0893	1

All the variables have been standardized and the results given in this part and below are for these standardized variables.

In the graph given below, note that
$V1 = Y$, $V2 = X_1$, $V3 = X_2, V4 = X_3, V5 = X_4$ and $V6 = X_5$.

It seems that the distribution of all the variables are right skewed and hence appropriate transformations would be helpful (though not done here). The response is well correlated with variables X_2, X_3 and X_5, and these three independent variables are well correlated among themselves.

(b) Eigenvalues of X^TX are $63.2195, 31.9245, 15.7690, 7.0980, 1.9890$. Since trace$(X^TX) = 120$, the first two eigenvalues explain about 79% of the total variability of the independent variables whereas the first three explain about 92 %. This points out there is multicollinearity.

(c) Estimated regression is

$$\begin{aligned} \hat{Y} &= -0.1046X_1 + 0.2466X_2 + 0.0185X_3 + 0.0629X_4 + 0.7364X_5, \\ s(b_1) &= 0.0365, s(b_2) = 0.0886, s(b_3) = 0.0569, s(b_4) = 0.0367, s(b_5) = 0.0692. \end{aligned}$$

ANOVA table

Source	df	SS	MS	F
Regression	5	23.5325	4.7065	191.30
Error	19	0.4675	0.0246	
Total	24	24		

(d) The variance inflation factors are $\{D_{jj}A_{jj}\}$, where $A = D^{-1}$and $D = X^TX$, are given in the table below

j	1	2	3	4	5
VIF_j	1.2987	7.6579	3.1576	1.3166	4.6702

VIF for X_2 is high. Even though there is some multicollinearity, but it is not severe if we use the standard rule of thumb (none of variance inflation factor is above 10).

(e) Optimal value of the penalty is $\hat{k}^* = 0.31$ (the GCV has been calculated on $[0, 1]$ with a grid 0.01).

Estimate of σ^2 is $\hat{\sigma}^2 = 0.023587$.

Estimated beta parameters and their standard errors are

$\hat{\beta}_j$	-0.1020	0.2634	0.0242	0.0609	0.7088
$s(\hat{\beta}_j)$	0.0348	0.0753	0.0517	0.0348	0.0607

(f) The plots of Y against $\hat{Y}$ shows a good fit as the scatterplot is well concentrated around the 45^o line. The plot of residuals against the fitted do not indicate any serious departure from the model assumptions such as linearity and equal variance. The normal probability plot does not show any serious departure from normality. However, skewness of the distribution of Y are apparent here and this has been alluded to in part (a). Variable transformations would have improved the quality of analysis here.
(g) The variance inflations factors are $\{D_{jj}A_{jj}\}$ where $A_{jj} = (X^TX+\hat{k}^*I)^{-1}X^TX(X^TX+\hat{k}^*I)$. Note that there is a mistake in part (g) since it is wrongly stated that the variance inflation factors are $\{A_{jj}\}$. The calculated variance inflation factors are given in the table below

j	1	2	3	4	5
VIF_j	1.2300	5.7736	2.7246	1.2353	3.7495

The variance inflation factors are lower than those in part (d) and this is due to the ridge regression approach.

2 and 3. Detailed R notes are given since these questions are quite computer intensive.

4.
(a) We have

$$\begin{aligned} E(Y_{ij}) &= \mu + \beta_1 x_i + \gamma_1 t_j, \\ Var(Y_{ij}) &= Var(\rho_i) + Var(\varepsilon_{ij}) = \sigma_\rho^2 + \sigma^2, \\ Cov(Y_{ij}, Y_{ij'}) &= \sigma_\rho^2, j \neq j', \\ Corr(Y_{ij}, Y_{ij'}) &= \frac{\sigma_\rho^2}{\sigma_\rho^2 + \sigma^2}, j \neq j'. \end{aligned}$$

(b) Assume that $\gamma_{i1} \sim N(0, \sigma_\gamma^2)$. Then

$$\begin{aligned} E(Y_{ij}) &= \mu + \beta_1 x_i + \gamma_1 t_j, \\ Var(Y_{ij}) &= Var(\rho_i) + Var(\gamma_{i1} t_j + Var(\varepsilon_{ij}) \\ &= \sigma_\rho^2 + \sigma_\gamma^2 t_j^2 + \sigma^2, \\ Cov(Y_{ij}, Y_{ij'}) &= \sigma_\rho^2 + \sigma_\gamma^2 t_j t_{j'}, j \neq j', \\ Corr(Y_{ij}, Y_{ij'}) &= \frac{\sigma_\rho^2 + \sigma_\gamma^2 t_j t_{j'}}{\sigma_\rho^2 + \sigma_\gamma^2 t_j^2 + \sigma^2}, j \neq j' \end{aligned}$$

Figure 1: Problem 1a HW6

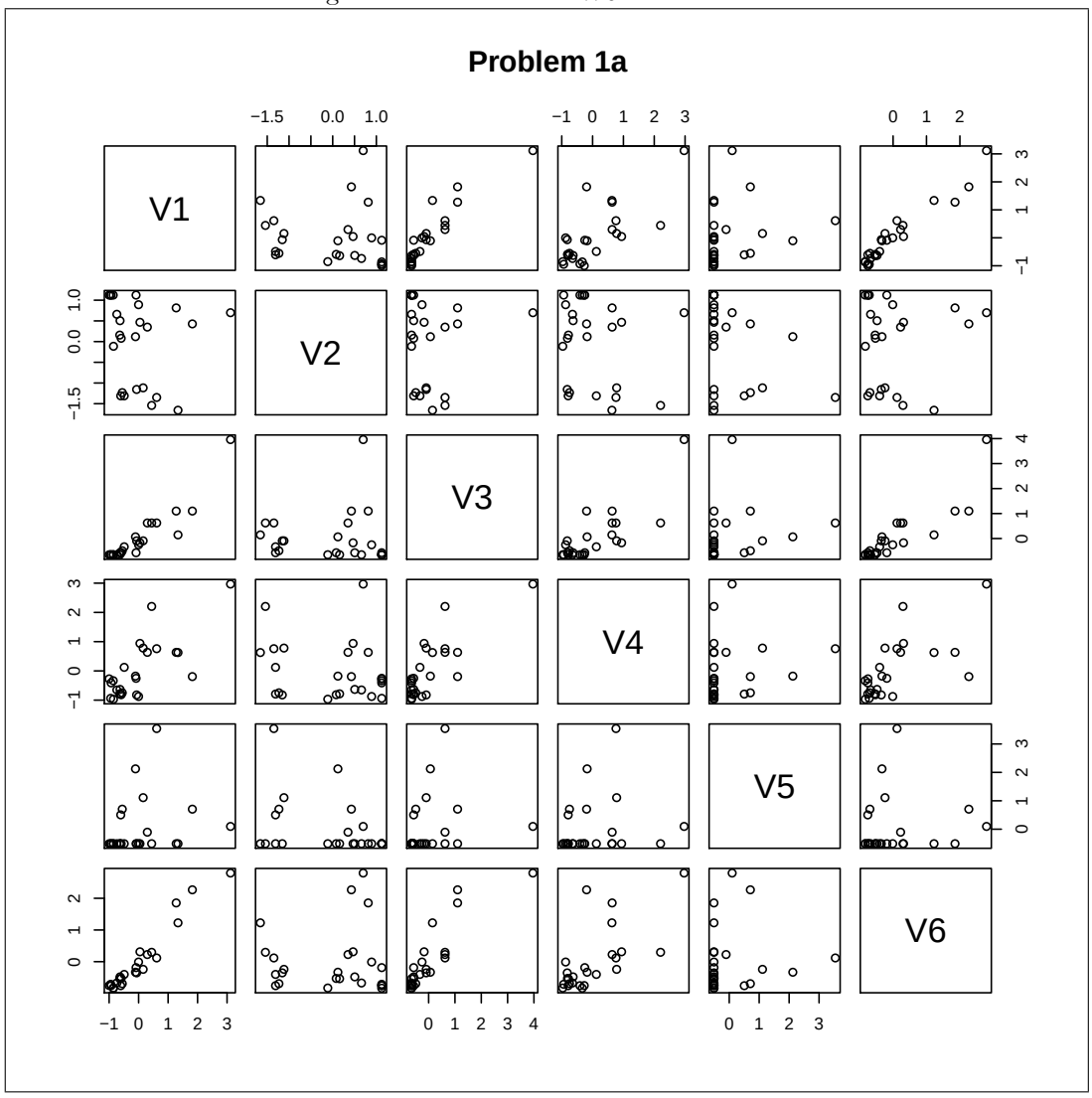

Figure 2: Problem 1f HW6