Data Analysis on Video Streaming QoE over Mobile Networks

Ticao Zhang

***Abstract*—Video streaming is becoming one of the most popular services over mobile networks. The growing demands on video streaming over mobile networks brings challenges to the network optimization in order to improve the user perceptual experience. Many previous studies mainly focus on improving quality of service (QoS) of video steaming over mobile networks. Typical QoS measures includes throughput, bandwidth, outage and delay. However, most of the QoS fails to characterize the user perceptual experience, which is called quality of experience (QoE). In real systems, people tend to use QoE to assess the video quality. However, QoE is a subjective metric and is difficult to obtain compared with QoS, in this paper, we will investigate the relationship between QoE and QoS parameters with multiple linear regression models. We perform model adequacy checking, multicollinearity diagnostics and variable selections. We find that the video initial buffering latency and the video playing time has the largest impact on the QoE while the initial video maximum downloading rate doesn't impact the QoE. The developed reduced model can more accurately reflect the impact of QoS parameter on the value of QoE. These results will assist the optimization in network streaming design.**

***Index Terms*—QoE, Data analysis, Linear regression, Video streaming.**

List of Figures

List of Tables

I. Introduction

According to the Cisco's report, the traffic caused by video streaming will account for more than 77% percent by 2021, among which the mobile video traffic will be more than 55% [2]. How to optimize the network resources to improve the *users' perceptual* experience becomes a big challenge.

A lot of existing works mainly focuses on how to improve the quality of service (QoS) of video streaming networks. The measuring metrics include throughput, bandwidth, delay, video compression rate and outage probability. However, the existing objective metrics for assessing the video quality can not fully characterize the users's perceptual experience, which is quite subjective. Besides, the increase in the QoS doesn't necessarily improve users' perceptual experience.

A more convincing criterion to assess the video quality would be the quality of experience (QoE) [3]. However, QoE-based video quality assessment is difficult because the subjective user experience is hard to quantify and measure. As a comparison, QoS-based video quality assessment is quite easy since the metrics such as bandwidth, delay and throughput can be obtained quantitively. Generally, there are three kinds of QoE assessment methods [3]. They are subjective tests, objective assessments and data-driven approaches. Of the three methods, the data-driven approach is gaining an increased attention due to the fact that this method is capable of fully utilizing the information of the massive datasets and doesn't suffer from the influence of the human visual system knowledge. It is also cheaper and more convenient.

T. Zhang is with the Department of Electrical and Computer Engineering, Auburn University, Auburn, AL 36849-5201, USA. Email: tzz0031@tigermail.auburn.edu

In this paper, we adopt a data-driven approach to measure the video QoE. We use the U-vMOS (User/Unified/Ubiquitous video Mean opinion Score) which is proposed by Huawei in 2016 [1] assess the QoE with discrete grades from 1 to 5. We call these grades the Mean Opinion Score (MOS) which is standardized by the International Telecommunication Union (ITU). Grades 1 to 5 represent *bad, poor, fair, good and excellent*, respectively. In [1], it is shown that the vMOS is affected by a series of QoS factors such as *video quality, initial buffering delay* and so on. However, the accurate impact that the QoS factors have on the QoE score has not been fully revealed.

Recently, [4] proposed a data-based method to investigate the relationship between QoE and other QoS parameters based on the Huawei's vMOS assessment model. In particular, the author uses a realistic dataset on video QoE based on SpeedVideo global Operating Platform (SVGOP) established by Huawei. The proposed analysis uses the k-means clustering to divide QoE score into two categories: good and bad. To deal with the binary classification problem, a logistic multiple linear regression approach is used to investigate the relationship between the binary QoE and other QoS parameters. However, the developed model has a bias of regarding "good" QoE values as "bad" and the relaxation to a binary QoE value will inevitably lose a lot of information.

In contrast to the existing studies, in this paper, we will try to fit the largest model possible to the data and try to retain as much information as possible. We will provide a rigorous thorough regression analysis on the QoE value and the QoS parameters. The analysis will include the model adequacy checking, diagnostics of the leverage and influence points, variable selection and multicollinearity analysis.

The remainder of this paper is organized as follows. Section II describes the data used in this paper and identifies the challenges. The multiple linear regression model and a detailed data analysis is presented in section III. Section IV gives a further discussion and finally section V concludes this paper.

II. Dataset description

A. Mobile U-vMOS overview

U-vMOS is a standard to describe the users' video experience, which covers a wide array of video services on mobile terminals, PCs, TVs and video calls. According to the Huawei's survey, there are three factors that may affect video experience: video quality, interactive experience and the viewing experience. As defined in Mobile U-vMOS, the interactive experience at video playback startup is determined by the initial buffering delay (sLoading) and the interactive experience during video playback is determined by the video freeze duration (sStalling).

B. Dataset description

We obtain the datasets[1] from the SpeedVideo Global Operating Platform (SVGOP) established by Huawei. The dataset contains 2001 samples. The scoring factor vMOS is the dependent parameter to measure the QoE while the remaining 11 QoS factors are the regression parameters. A detailed description of the dataset is given in table I.

TABLE I
Description of the dataset

Parameter	Features	variables
QoS	Average rate of playing phase (kbps)	x_1
	Video total download (DL) rate (kbps)	x_2
	Video bitrate (kbps)	x_3
	Initial max DL rate (kbps)	x_4
	End-to-End (E2E) round-trip time (RTT) (ms)	x_5
	Initial buffering latency (ms)	x_6
	Video Initial buffer downloaded (byte)	x_7
	Playing time (ms)	x_8
	Playing total duration	x_9
	Stalling times	x_{10}
	Stalling ratio	x_{11}
QoE	vMOS	y

Typical U-vMOS Value	Initial Buffering Latency (s)	Video Source		Requirements for E2E Network KPIs			
		Resolution	Typical Bit Rate (Mbps)	Playback Rate (Mbps)	Peak Rate for Initial Buffering	RTT (ms)	PLR Threshold
3	2	480P	0.7	0.9	4.6	80	1.0E-03
3.3	2	720P	1.5	2.0	8.8	70	3.6E-04
3.5	1.5	720P	1.5	2.0	18.9	60	1.1E-04
3.8	1.5	1080P	3	3.9	19.7	45	1.7E-04
4	1	1080P	3	3.9	29.7	30	1.7E-04
4.2	1	2K	6	7.8	65.3	30	3.6E-05
4.5	1	4K	13.5	17.6	96.7	20	3.6E-05

Figure. 1. Typical E2E KPI requirements corresponding to varying Mobile U-vMOS scores (The table is taken from [1])

Of the 11 QoS parameters, x_1, x_2 and x_3 are video quality related parameters. They describe the video bitrate, download rate and playing rate. x_4, x_5, x_6 and x_7 are video initial loading parameters. They describe the status of the videos. $x8, x_9, x10$ and x_{11} are video stalling parameters. They describe the video freeze duration time and playing time.

To help the readers have a basic knowledge of what the data looks like, we give the typical vMOS values in Figure 1 [1]. It can be seen that when the RTT time is short, the Bit rate is high and the initial buffering latency is small, the users will have a good QoE. The corresponding vMOS value will be high.

C. A scatter plot

We present a basic scatter plot of y and x_i (For the space limitation, we only list several scatter plot in Figure 2-Figure 4). It can be seen that there seems to be a positive relationship between y and x_8 and a negative relationship between y and x_{11}. Also, we give a matrix of scatter plot in Figure 26 (see appendix), it can be seen that x_1 and x_2 are strongly correlated. x_8 and x_{11} are strongly linear correlated. This shows that if we do multiple linear regression with all the regressors, there will be multicollinearity problems.

[1]The dataset can be downloaded from http://speedvideo.huawei.com/mobilevmos/

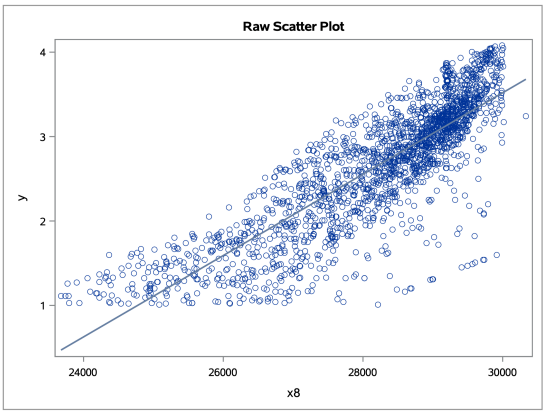

Figure. 2. A raw scatter plot between y and x_8

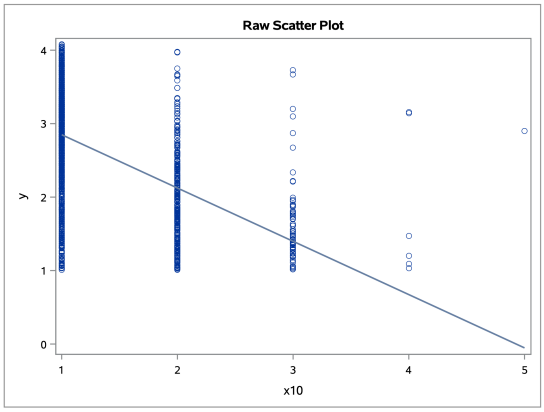

Figure. 3. A raw scatter plot between y and x_{10}

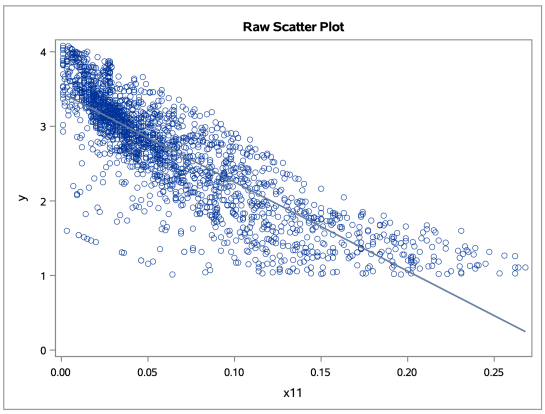

Figure. 4. A raw scatter plot between y and x_{11}

Parameter Estimates								
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	Intercept	1	1.38248	3.28486	0.42	0.6739	.	0
x1	x1	1	-0.00025778	0.00002917	-8.84	<.0001	0.00294	340.11847
x2	x2	1	0.00026343	0.00003001	8.78	<.0001	0.00290	345.14601
x3	x3	1	0.00015006	0.00009806	1.53	0.1261	0.36915	2.70895
x4	x4	1	0.00000284	3.285074E-7	8.66	<.0001	0.48830	2.04793
x5	x5	1	-0.00061668	0.00013008	-4.74	<.0001	0.91088	1.09783
x6	x6	1	-0.00020780	0.00000267	-77.72	<.0001	0.47970	2.08462
x7	x7	1	-2.77877E-7	1.475106E-7	-1.88	0.0597	0.30403	3.28912
x8	x8	1	0.00018259	0.00004215	4.33	<.0001	0.00297	336.87520
x9	x9	1	-0.00009109	0.00011158	-0.82	0.4144	0.92942	1.07594
x10	x10	1	-0.02684	0.00748	-3.59	0.0003	0.54780	1.82549
x11	x11	1	-7.12040	1.05567	-6.74	<.0001	0.00297	337.16945

Figure. 5. Test for significance of regression

III. Data Analysis

We give a general expression of the multiple regression equation as follows

$$y = \beta_0 + \sum_{i=1}^{11} \beta_i x_i + \epsilon, \tag{1}$$

where β_i is the regression coefficient and ϵ is an error term with zero mean and variance σ^2. For different observations, we assume that the errors are uncorrelated and normally distributed. We call this assumption the *normality assumption*.

A. Multiple linear regression

1) Parameter estimation: With the least-square (LS) estimation, we get the fitted multiple linear regression full model as

$$\begin{aligned} y = {} & 1.38248 - 0.00025778x_1 + 0.00026343x_2 + 0.00015006x_3 \\ & + 0.00000284x_4 - 0.00061668x_5 - 0.00020780x_6 \\ & - 2.77877 \times 10^{-7}x_7 + 0.00018259x_8 \\ & - 0.00009109x_9 - 0.02684x_{10} - 7.12040x_{11} \end{aligned} \tag{2}$$

From Figure 5, we get the parameter estimation. With the standard error $\mathrm{se}(\hat{\beta}_j)$, we can also get the confidence interval for each parameters estimation.

$$[\hat{\beta}_j - t_{\alpha/2,p,n-p}\mathrm{se}(\hat{\beta}_j), \hat{\beta}_j + t_{\alpha/2,p,n-p}\mathrm{se}(\hat{\beta}_j)] \tag{3}$$

where n is the number of samples, $p = k + 1$ and k is the number of regressors and α is the confidence level.

2) Test for significance of regression: We test if there is a linear relationship between the response variable y and any of the regressor variables $x_i (i = 1, 2, ..., 11)$. The analysis of variance table is shown in Figure 6. Since F value is quite large (5440.44), we reject the null hypothesis and conclude that at least one of the regressors contributes significantly to the model. In Figure 7, the R value and the adjusted R value is 0.9678 and 0.9677, respectively. Hence, about 96% variability in y can be explained by this linear regression model. However,

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	1095.27703	99.57064	5440.55	<.0001
Error	1989	36.40185	0.01830		
Corrected Total	2000	1131.67887			

Figure. 6. Analysis of variance table for the full model

Root MSE	0.13528	R-Square	0.9678
Dependent Mean	2.65396	Adj R-Sq	0.9677
Coeff Var	5.09741		

Figure. 7. Coefficient of determination for the full model

this does not necessarily imply that the relationship found is an appropriate one for predicting y as a function of $x_i(i = 1, 2, ..., 11)$. Further tests of model adequacy are required.

3) Test on individual regression coefficients and subsets of coefficients: Now we perform the partial or marginal t-test for each coefficient $\beta_j(j = 1, 2, ..., 11)$. This is a test of the contribution of x_j given the other regressors in the model. From Figure 5, we find that the regressor x_3, x_7, x_9 may do not contribute significantly to the model given other parameters in the model.

However, this is only the initial results. Since there may be severe multicollinearity problems, the parameter estimation and its corresponding confidence intervals here may be not accurate. Further diagnostics need to be performed.

B. Model adequacy checking

1) normal probability plot of residuals: In our multiple linear regression model, we made the assumption that the errors are uncorrelated and normally distributed. To check the normality assumption, a normal probability plot of the residuals is given in Figure 8.

From Figure 8, we see that most of the points lie approximately on a straight line except several points. These points may be outliers. Compared with the size of the total sample points, the number of departure points is small. We therefore conclude that the normality assumption is reasonable here.

2) plot of the residuals versus the predicted response: We give the plot of the residuals versus the predicted response in Figure 9. It can be seen that there are several points which have very large residuals. They may be the outliers. For the rest points, there is a light pattern, which shows that here may be a nonlinear relationship here. It is very likely that the model is not adequate.

3) Detection of outliers: An outlier is an extreme observation; one that is considerably different from the majority of the data. Residuals that are considerably larger in absolute value than the others, say three or four standard deviations

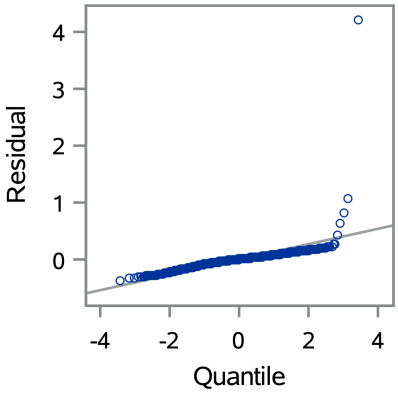

Figure. 8. A normal probability plot of the residuals for the full model

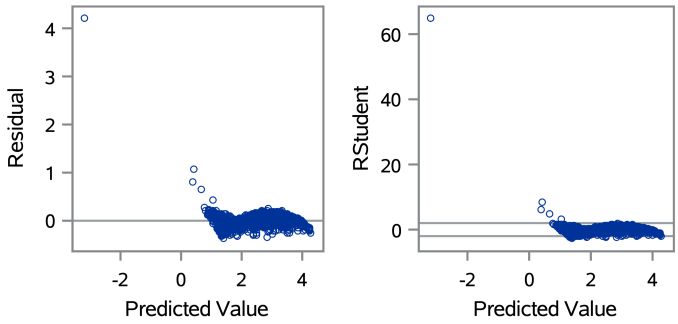

Figure. 9. A plot of the residuals versus the predicted response for the full model

from the mean, indicate potential space outliers. We sort the residual values of different observations in a descending order in Figure 10.

It can be seen that the first 5 values have a residual larger than 3. Thus we call them outliers. Now we remove the 5 outliers and refit the regression model. The adjusted R^2 is 0.9937 and the F value is 28551.7. This shows that the regression relationship is significant. The residual and normality plot are presented in Figure 11 and 12, respectively. It seems that there is a nonlinear pattern and the normality is violated

Unless specified, the following analysis are all based on the dataset where the 5 outliers are removed.

C. Diagnostics for leverage and influence points

1) Leverage points: A leverage point is a point that is remote in x space from the rest of the sample, but it lies almost on the regression line passing through the rest of the sample points. To diagnose the leverage points, we focus on the diagonal element h_{ii} of the hat matrix

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}, \tag{4}$$

. If the hat diagonal exceeds twice the average $2p/n$, then the point can be considered as leverage points. Here, $p = k+1 = 12$ and $n = 2000 - 5 = 1995$. Thus $2p/n \approx 0.012$. We sort the datas' h value in Figure 13. We find that the first 6 points are potential leverage points.

Obs	ID	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	y	studresids	cook	leverage	studdelresid
1	2800909	1839	1844	2966	5522	124	31182	1565776	28214	30016	1	0.064	1.01	36.7623	44.4434	0.28296	64.9169
2	3720944	6281	4402	3005	14328	36	14717	1577901	29565	30008	1	0.015	1.49	8.3604	0.7694	0.11668	8.5092
3	3092358	4801	3645	2903	19264	118	13619	1624327	28681	30012	1	0.046	1.2	6.1494	0.1730	0.05204	6.2071
4	3089819	6035	5247	2966	6569	78	13503	1566408	28997	30005	1	0.035	1.31	4.8080	0.0560	0.02824	4.8350
5	3839056	3236	2531	2975	24609	50	13152	1564845	29499	30005	1	0.017	1.47	3.2572	0.0357	0.03883	3.2651
6	4003666	3248	3245	2975	13731	39	1581	1568072	23797	30025	3	0.262	1.03	2.0151	0.0098	0.02811	2.0166
7	2770465	13890	13788	3065	24940	45	1262	1644684	27790	30006	4	0.08	3.14	1.9407	0.0108	0.03317	1.9420
8	4002305	1845	1849	3088	18200	62	1865	1640892	24181	30003	3	0.241	1.14	1.6935	0.0046	0.01878	1.6943
9	3469618	2217	2232	3092	38593	69	1322	1628992	23787	30009	2	0.262	1.11	1.6272	0.0071	0.03105	1.6279
10	3057363	3011	3018	2903	24828	83	1864	1568582	23939	30004	2	0.253	1.03	1.6020	0.0055	0.02489	1.6026
11	2765847	13541	13141	3065	30803	75	2213	1621336	27833	30009	3	0.078	2.87	1.5934	0.0053	0.02434	1.5940

Figure. 10. Detection of outliers

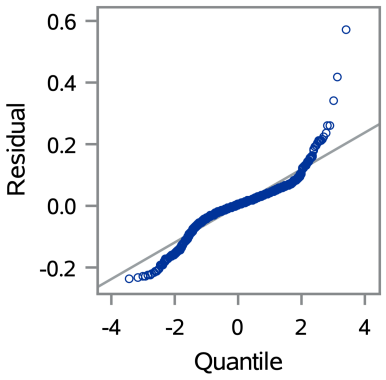

Figure. 11. A normal probability plot of the residuals for the full model(after removing 5 outliers)

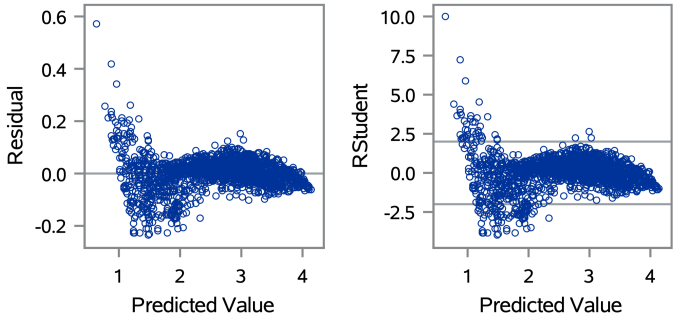

Figure. 12. A plot of the residuals versus the predicted response for the full model (after removing 5 outliers)

2) Influence points: An influence point is a point that has a moderately unusual x coordinate, and the y value is unusual as well. We use cook's D value to detect influence points. As shown in Figure 14, there seems to be no influential points for their cook's D value are all less than 1.

D. Multicollinearity

1) Diagnose with Multicollinearity: From Figure 26, we know that there seems to be a linear relationship between variable x_1 and x_2. In this section, we present methods to detect multicollinearity and techniques to deal with the problem.

Recall that the diagonal elements of the matrix $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}$ are useful in detecting multicollinearity. Also,

$$C_{jj} = \frac{1}{1 - R_j^2}, \tag{5}$$

where R_j^2 is the coefficient of determination obtained when x_j is regressed on the remaining k regressors. If x_j is nearly orthogonal to the remaining regressors, R_j^2 is small and C_{jj} is close to unity. On the contrary, if R_j^2 is near unity, C_{jj} will be large. We use the variance inflation factor (VIF) to determine if there is a multicollinearity.

$$\text{VIF}_j = C_{jj}. \tag{6}$$

From Figure 15, we know that the VIF value of x_1, x_2, x_8 and x_{11} are quite large. It is an indication that the associated regression coefficients are poorly estimated because of multicollinearity.

2) Dealing with Multicollinearity: There are several ways to deal with multicollinearity. One is model re-specification. Since x_1 and x_2 are highly correlated, we can either redefine the regressors such as $x_a = (x_1+x_2)/2$ and $x_b = (x_8+x_{11})/2$ or we can simply eliminate one regressor (say x_2 and x_{11}). The other methods include ridge regression, principal component regression and so on. Since we know from Figure 26 that x_1 and x_2 are strongly related, we can simply eliminate the regressor x_2 and x_{11} and we get the resulting reduced model.

$$\begin{aligned} y = {} & \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 \\ & + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \beta_{10} x_{10} \end{aligned} \tag{7}$$

E. Variable selection and model building

Note that in this model, we still have a large set of possible candidate regressors. However, only a few are likely to be important. Hence we need to do the variable selection to determine an appropriate subset of important regressors.

Various methods have been developed to determine the regression models by either adding or deleting regressors one at a time. These methods are generally referred to as stepwise-type procedures.They include the forward selection, backward selection and stepwise regression.

Obs	ID	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	y	studresids	cook	leverage	studdelresid	var	outlier	studresids2	cook2	leverage2	studdelresid2
1	3884567	4476	4443	2975	11265	82	3359	1577892	29433	30626	2	0.019	3.13	0.37702	0.004064	0.25544	0.37694	.	0	-0.56261	0.00906	0.25568	-0.56251
2	2810588	2013	2016	2966	8555	85	3276	1567181	30332	30591	1	0.009	3.24	-0.97713	0.023301	0.22652	-0.97712	.	0	-0.82430	0.01660	0.22674	-0.82423
3	3092061	1385	1647	3065	10125	68	3225	1615904	30013	30590	1	0.019	3.17	-1.24641	0.037581	0.22498	-1.24658	.	0	-0.50794	0.00627	0.22575	-0.50785
4	3672918	6581	6518	2975	17196	79	2084	1595192	28944	30519	2	0.037	3.34	1.46051	0.039122	0.18039	1.46092	.	0	0.61254	0.00692	0.18122	0.61245
5	2784571	30619	30351	3065	34935	48	1136	1658174	27249	30009	5	0.101	2.9	1.55687	0.036541	0.15319	1.55743	.	0	2.23364	0.07608	0.15469	2.23589
6	3884231	7823	7904	3088	29756	62	1573	1951692	29624	30003	1	0.013	3.72	0.58646	0.003631	0.11243	0.58636	.	0	-0.11666	0.00015	0.11359	-0.11663
7	2784344	25360	25341	3065	42286	54	1012	1623792	27686	30013	4	0.084	3.16	1.07514	0.010563	0.09882	1.07518	.	0	2.23321	0.04645	0.10053	2.23545
8	2879396	3747	3692	2903	35576	86	4535	1824482	29988	30006	1	0.001	2.93	-1.56518	0.020414	0.09091	-1.56575	.	0	-1.57422	0.02094	0.09205	-1.57481
9	3657818	9381	9464	2912	96853	41	718	1818692	29205	30004	1	0.027	3.84	-0.23253	0.000340	0.07010	-0.23247	.	0	0.13741	0.00012	0.07070	0.13738
10	2903994	1078	1194	3065	11846	47	2026	1635508	28696	30329	1	0.057	3.16	0.82349	0.004209	0.06931	0.82342	.	0	1.01524	0.00641	0.06941	1.01525
11	3357099	7031	6915	3065	37546	60	2467	1863664	29595	30005	1	0.014	3.44	0.03869	0.000008	0.05841	0.03868	.	0	-0.57317	0.00171	0.05891	-0.57307

Figure. 13. Detection of the leverage points

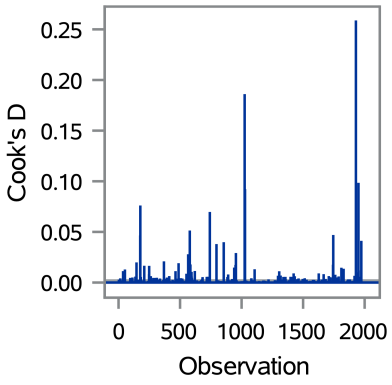

Figure. 14. Diagnostics for influence points

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	4.29625	1.44731	2.97	0.0030	0
x1	x1	1	0.00000215	0.00001429	0.15	0.8806	419.53910
x2	x2	1	-7.22297E-7	0.00001469	-0.05	0.9608	425.23627
x3	x3	1	0.00010883	0.00004336	2.51	0.0122	2.71851
x4	x4	1	4.534711E-7	1.488974E-7	3.05	0.0024	2.16671
x5	x5	1	-0.00021068	0.00005776	-3.65	0.0003	1.10883
x6	x6	1	-0.00027749	0.00000142	-195.89	<.0001	2.32287
x7	x7	1	-6.15985E-8	6.52654E-8	-0.94	0.3454	3.31311
x8	x8	1	-0.00005574	0.00001876	-2.97	0.0030	343.66064
x9	x9	1	0.00004960	0.00004918	1.01	0.3133	1.07721
x10	x10	1	-0.02937	0.00329	-8.92	<.0001	1.82484
x11	x11	1	-12.94283	0.46970	-27.56	<.0001	343.60778

Figure. 15. Multicollinearity diagnostics

1) Forward selection: We apply the forward selection procedure to the reduced model (7). We specify the cutoff value for entering variables by choosing the alpha-to-enter value as $\alpha = 0.10$. From Figure 16, we see that the regressor x_8 enters the model first, then $x_2, x_{10}, x_9, x_5, x_3$ enters the model. Then no other variable met the $\alpha = 0.1$ significance level for entry into the model.

2) Backward selection: Backward elimination attempts to find a good model by working in the opposite direction. By

Summary of Forward Selection							
Step	Variable Entered	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	x8	x8	1	0.7239	0.7239	60850.5	5227.13	<.0001
2	x6	x6	2	0.2667	0.9905	162.219	56200.5	<.0001
3	x10	x10	3	0.0005	0.9910	55.8448	105.63	<.0001
4	x9	x9	4	0.0001	0.9911	31.2824	26.22	<.0001
5	x5	x5	5	0.0001	0.9912	14.1110	19.09	<.0001
6	x3	x3	6	0.0000	0.9913	9.0315	7.07	0.0079

Figure. 16. Forward selection results

Summary of Backward Elimination								
Step	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	x7	x7	8	0.0000	0.9913	8.9571	0.96	0.3280
2	x1	x1	7	0.0000	0.9913	8.9660	2.01	0.1565
3	x4	x4	6	0.0000	0.9913	9.0315	2.06	0.1509

Figure. 17. Backward selection results

choosing the significance level as 0.1, x_7 is removed from the model first and then x_1 and x_4 are removed as shown in Figure 17.

3) Stepwise regression: Stepwise regression is a modification of forward selection in which at each step all regressors entered into the model previously are reassessed via their partial F (or t) statistics. A regressor added at an earlier step may now be redundant because of the relationships between it and regressors now in the equation. As shown in Figure 18, x_8 enters the model first. Finally, all variables left in the model are significant at the 0.1 level. No other variable met the 0.05 significance level for entry into the model.

As can be seen, all of the three methods produce the same result. We retain $x_3, x_5, x_6, x_8, x_9, x_{10}$ in our model.

IV. Some discussions

A. A discussion on the reduced model

In the previous section, we first fit a full model and perform a thorough analysis of this model, including a full residual analysis. We find that at least one of the regressors contributes significantly to the model. In the model adequacy test, we find

Summary of Stepwise Selection									
Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	x8		x8	1	0.7239	0.7239	60850.5	5227.13	<.0001
2	x6		x6	2	0.2667	0.9905	162.219	56200.5	<.0001
3	x10		x10	3	0.0005	0.9910	55.8448	105.63	<.0001
4	x9		x9	4	0.0001	0.9911	31.2824	26.22	<.0001
5	x5		x5	5	0.0001	0.9912	14.1110	19.09	<.0001
6	x3		x3	6	0.0000	0.9913	9.0315	7.07	0.0079

Figure. 18. Stepwise regression

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-1.29696	1.68384	0.00293	0.59	0.4412
x3	0.00008313	0.00003126	0.03491	7.07	0.0079
x5	-0.00028796	0.00006757	0.08965	18.16	<.0001
x6	-0.00026954	0.00000115	272.48186	55198.6	<.0001
x8	0.00046006	0.00000158	418.79791	84838.9	<.0001
x9	-0.00027795	0.00005604	0.12144	24.60	<.0001
x10	-0.03868	0.00379	0.51459	104.24	<.0001

Figure. 19. Parameter estimation

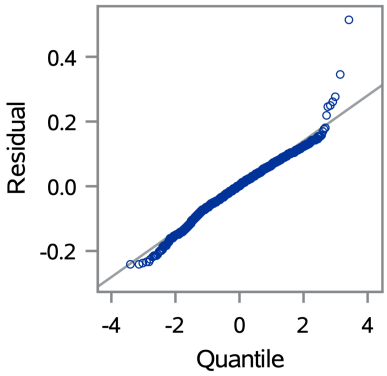

Figure. 20. A normal probability plot of the residuals for the reduced model

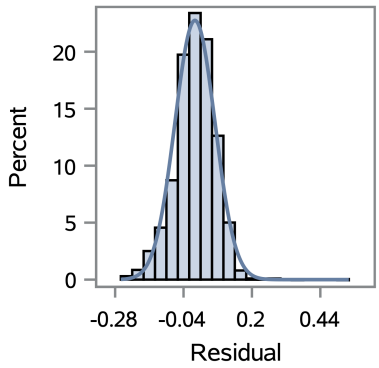

Figure. 21. The distribution of the residuals for the reduced model

that the normality assumption looks good but there seems to be a pattern in the residual plot. Then we diagnose the possible outliers, leverage points and influence points. We find that there are 5 possible outliers. To better analyze the behavior of the data, we remove the possible outliers. In the remaining data samples, there are 6 potential leverage points and there seems to be no influential points. Based on the remaining data samples, we further diagnose the multicollinearity problem, we find that x_1, x_2 are highly correlated and x_8, x_{11} are highly correlated. We simply delete the regressor x_2 and x_{11} and we get the reduced model (7).

Note that not all of the regressors contribute significantly to the model, hereafter we perform a variable selection. The three methods all show that $x_2, x_{10}, x_9, x_5, x_3$ and x_8 should be retained in the model. We thus get the reduced model based on the parameter estimation in Figure 19,

$$y = -1.2967 + 0.00008313x_3 - 0.00028796x_5 - 0.00026954x_6 + 0.00046006x_8 - 0.00027795x_9 - 0.03868x_{10}. \quad (8)$$

1) Model adequacy checking: Now we perform the model adequacy checking to see if there is any problem with the normality assumption in (8). We present the normal probability plot in Figure 20 and 21, we find that there is no problem with the normality assumption since all points almost lie in a straight line.

We then give the plot of the residuals versus the predicted response in Figure 22, we find that the residuals of almost all points lie between the -2 and 2. There seems to be a slight pattern which indicates there may be a nonlinear relationship in the fitted model. However, the problem is not that serious. Despite that, proper transformation may still be needed. We leave this for our future work.

2) Diagnostics for leverage, outlier and influence: Finally, Figure 23 suggests that no observations shows as influential. Also, we have already removed the possible outliers from the dataset, there is no outliers now. We can do similar procedure as previous to determine if there is any leverage points.

3) Multicollinearity checking: From Figure 24, we see that the VIF values are quite small, which indicates that there is no

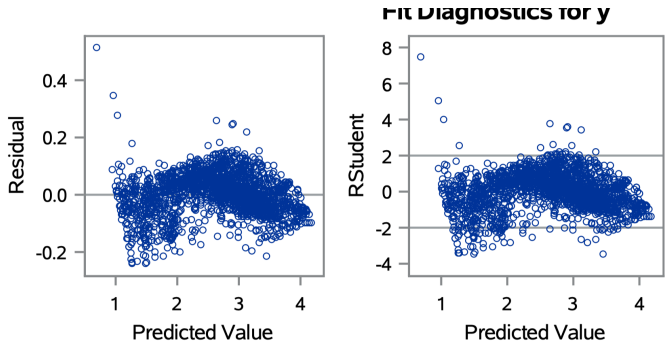

Figure. 22. A plot of the residuals versus the predicted response for the reduced model

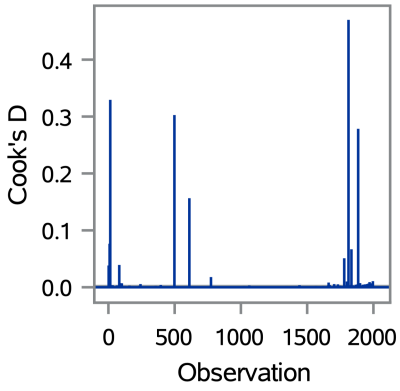

Figure. 23. diagnostics of influential points

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	-1.29696	1.68384	-0.77	0.4412	0
x3	x3	1	0.00008313	0.00003126	2.66	0.0079	1.01636
x5	x5	1	-0.00028796	0.00006757	-4.26	<.0001	1.09148
x6	x6	1	-0.00026954	0.00000115	-234.94	<.0001	1.09603
x8	x8	1	0.00046006	0.00000158	291.27	<.0001	1.75171
x9	x9	1	-0.00027795	0.00005604	-4.96	<.0001	1.00606
x10	x10	1	-0.03868	0.00379	-10.21	<.0001	1.73689

Figure. 24. A plot of the residuals versus the predicted response for the reduced model

multicollinearity problem in the reduced model. Moreover, the t statistics for x_6 and x_8 are quite large, hence the contribution of x_6 and x_8 may be significant.

Recall that x_6 is the initial buffering latency and x_8 is the total video playing time, we therefore conclude that the the initial buffering latency and the total video playing time affects the users's experience most. Other factors such as total video download time (x_2), RTT (x_5), Playing total duration (x_9) and stalling times x_{10} also has an affect on the QoE. Since x_1, x_2 and x_8, x_{11} are highly correlated, we can also say that The average rate of playing phase (x_1) and the video stalling ratio (x_{11}) affect the QoE. The initial maximum download rate (x_4) and the video stalling ratio (x_7) may not have much affect on QoE.

4) Residual plot for the regressors: We plot the residuals versus the regressors in Figure 25. We observe that there is a significant pattern in the plot for x_8. It appears that adding the regressor x_8 to the model help to explain some of the variability in the response. On the other hand, if the plot shows no pattern or trend, it is suggested that there is no relationship between the residuals and regressor.

B. Strategy summary

We employed the classical approach to regression model selection. Our basic strategy is as follows

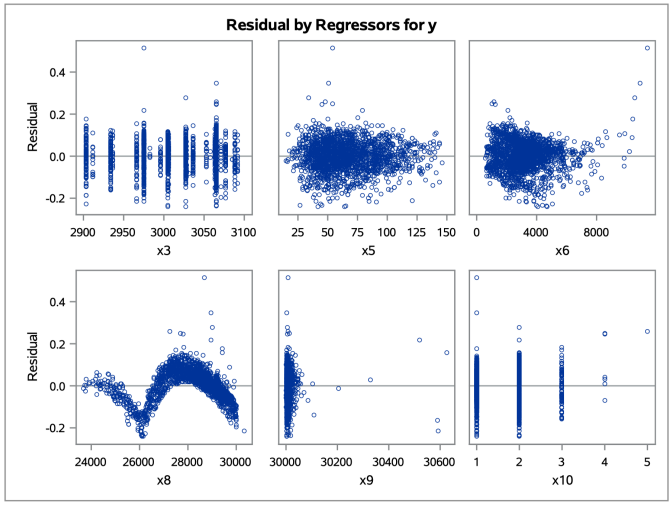

Figure. 25. A plot of the residuals versus the regressors for the reduced model

- Fit the full model
- Perform a thorough analysis of the full model, including residual analysis, influence points analysis and multi-collinearity diagnostics.
- Use t test on the individual regressors to edit the model and perform variable selection to find an appropriate subset of regressors.
- Check the reduced model again.

By performing the above strategies iteratively, we find that the residual plot, normality test and significance of regression test look good (although not perfect). Hence we feel comfortable in recommending our reduced linear regression model (7).

C. Utility of results

- There is a linear relationship between QoE and QoS parameters.

$$y = -1.2967 + 0.00008313x_3 - 0.00028796x_5 - 0.00026954x_6 + 0.00046006x_8 - 0.00027795x_9 - 0.03868x_{10}. \quad (9)$$

- The initial buffering latency and the total video playing time affects the users's experience most.
- The initial maximum download rate (x_4) and the video stalling ratio (x_7) may not have much affect on QoE.
- The total video playing time and the video bitrate have positive impact on QoE while other QoS parameters have negative impact on QoE value.
- The developed model can be used to predict the QoE value based on the measured QoS metric. These results can be used in the network optimization so that the resources can be utilized more effectively to help the users enjoy a better perceptual experience.

D. Subjects for further study

- We still need to consider if possible transformation is needed to mitigate the slight nonlinear pattern in the residual plot.

- We still need to investigate the C_p value, PRESS, DFFITS information and so on to get a better regression model.
- We may use a training set to train our model and use a testing set to check the accuracy of the developed model. We want to ensure that the results also hold for large dataset.

V. Conclusion

In this paper we conduct a data-driven analysis on video Mean Opinion Score (vMOS), which is an important measure of user quality of experience (QoE) of video streaming. In particular, our study is based on a realistic dataset consisting of 2001 samples and 11 features. This dataset has a dependency feature and consists of various data types, which brings us many challenges. To handle these difficulties, we employ the classical approach to effectively analyze the relationship between QoE and QoS parameters. We find that the QoE is affected by many QoS factors together and the initial buffering latency and the total video playing time affects the users's experience most. Our results paved a way for the improvement of video streaming. Several optimizations can be performed to help mitigate the QoS bottlenecks so as to improve the video streaming QoE more effectively.

VI. Appendix

References

[1] huawei, "Requirements of mobile bearer network for mobile video service," white paper 1-8, 2016.

[2] Cisco, "Cisco visual networking index: Global mobile data traffic forecast update, 2016–2021 white paper." [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html

[3] Y. Chen, K. Wu, and Q. Zhang, "From qos to qoe: A tutorial on video quality assessment," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 2, pp. 1126–1165, 2015.

[4] Q. Wang, H.-N. Dai, D. Wu, and H. Xiao, "Data analysis on video streaming qoe over mobile networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2018, no. 1, p. 173, 2018.

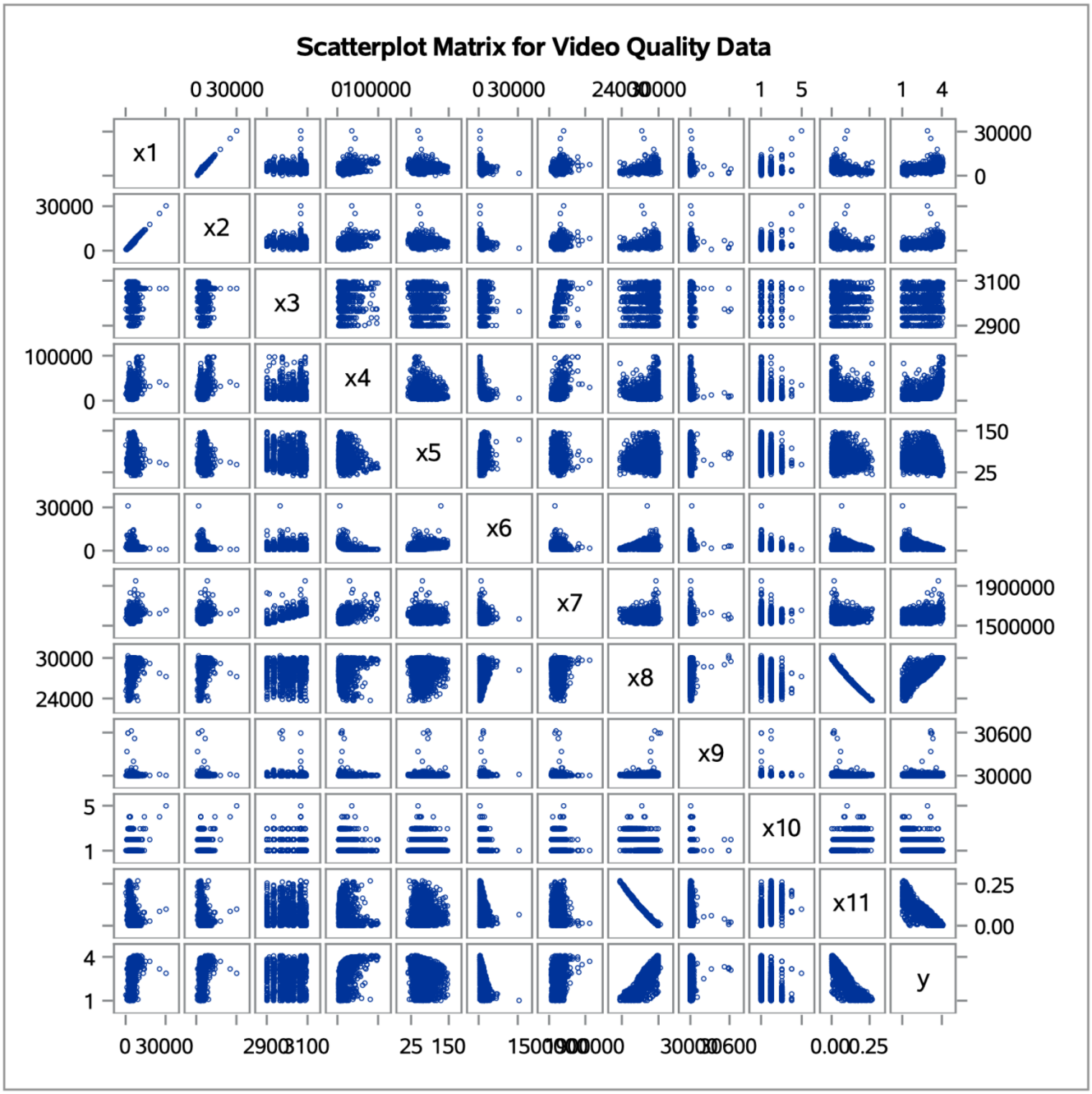

Figure. 26. a matrix of scatter plot

Here is the SAS code for the project

```
/* Gas Mileage Data (B.3)
y : vMOS
x1 : Average rate of playing phase (kbps)
x2 : Video total download (DL) rate (kbps)
x3 : Video bitrate (kbps)
x4 : Initial max DL rate (kbps )
x5 : E2E RTT(ms)
x6 : Initial buffering latency (ms)
x7 : Video Initial buffer download (byte)
x8 : Playing time(ms)
x9 : Playing total duration(ms)
x10: Stalling times
x11 : Stalling ratio
*/

* Read an Excel spreadsheet using PROC IMPORT;
PROC IMPORT DATAFILE = '/home/tzz00310/code/HW/Project/CaseData.xls'
DBMS=XLS OUT = VideoQuality;
RUN;

/* 1. scatter plot */
proc sgscatter data=VideoQuality;
  title "Raw Scatter Plot ";
  plot y*x1/ reg;
run;
proc sgscatter data=VideoQuality;
  title "Raw Scatter Plot ";
  plot y*x2/ reg;
run;


/* 2. scatter plot matrix */
proc sgscatter data=VideoQuality;
  title "Scatterplot Matrix for Video Quality Data";
  matrix x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 y;
run;


/* 3. regression analysis */
Proc reg DATA=VideoQuality;
TITLE 'Results of Regression Analysis';
model y = x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11/p clm cli tol vif
collin partial influence corrb;
plot rstudent.*(predicted. x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11  obs.);
```

```
plot npp.*rstudent.;
output out=demo1 student=studresids rstudent=studdelresid
cookd=cook h=leverage;
run;


/* 4. outlier points; */
proc sort data=demo1; by DESCENDING studresids;
proc print data=demo1;
run;
data demo1;
set demo1;
if -3<= studresids <=3 then outlier=0;
else outlier=1;    /*label the outlier*/
run;
data RMOutlier; /*here is the data that removes the outliers*/
set demo1;
where outlier=0;
run;

/* 5. regression analysis */
Proc reg DATA=RMOutlier;
TITLE 'Results of Regression Analysis';
model y = x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11/p clm cli tol vif
collin partial influence corrb;
plot rstudent.*(predicted. x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11  obs.);
plot npp.*rstudent.;
output out=demo2 student=studresids rstudent=studdelresid
cookd=cook h=leverage;
run;


/* 6. influence points */
proc sort data=demo2; by DESCENDING cook;
proc print data=demo2;
run;

/* 7. leverage points  */
proc sort data=demo2; by DESCENDING leverage;
proc print data=demo2;
run;


/* 8. multicollinearity analysis */
Proc reg DATA=RMOutlier;
```

```
TITLE 'Results of Regression Analysis';
model y = x1  x3 x4 x5 x6 x7 x8 x9 x10 / corrb vif collin;
run;


/* 9. variable selection */
PROC REG data=RMOutlier;
  MODEL y = x1 x3 x4 x5 x6 x7 x8 x9 x10/ selection=forward
   sle=0.10;  *default sle=0.5;
RUN;

PROC REG data=RMOutlier;
  MODEL y = x1 x3 x4 x5 x6 x7 x8 x9 x10/ selection=backward
  sls=0.10;  *default sle=0.5;
RUN;

PROC REG data=RMOutlier;
  MODEL y = x1 x3 x4 x5 x6 x7 x8 x9 x10/ selection=stepwise
  sle=0.05 sls=0.10;  *default sle=0.5;
RUN;
```