Department of Statistics • The Chinese University of Hong Kong
STAT 5102 Regression in Practice (Term 1, 2018-19)

Lab Demonstration #3 (26th November 2018)

[Example 1] Creatinine clearance (Y) is an important measure of kidney function, but is difficult to obtain in a clinical office setting because it requires 24-hour urine collection. To determine whether this measure can be predicted from some data that are easily available, a kidney specialist obtained the data that follow for 33 male subjects. The predictor variables are serum creatinine concentration (X1), age (X2), and weight (X3).

Theoretical arguments suggest use of the following regression function:

$$E\{\ln Y\} = \beta_0 + \beta_1 \ln X_1 + \beta_2 \ln(140 - X_2) + \beta_3 \ln X_3$$

a) Fit the regression function based on theoretical considerations.

Edit > Protect Data (to Update mode)
(Right Click) > Insert Column > Name: logY, Expression: LOG(Y)
(Right Click) > Insert Column > Name: logX1, Expression: LOG(X1)
(Right Click) > Insert Column > Name: logX3, Expression: LOG(X3)
(Right Click) > Insert Column > Name: logX2, Expression: LOG(140-X2)

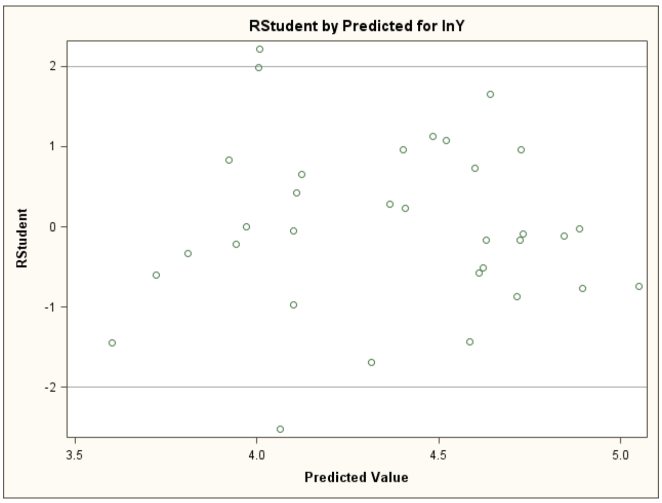
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-2.04269	1.01919	-2.00	0.0545
lnX1	1	-0.71195	0.09203	-7.74	<.0001
lnX3	1	0.75745	0.15923	4.76	<.0001
ln(140-X2)	1	0.74736	0.15696	4.76	<.0001

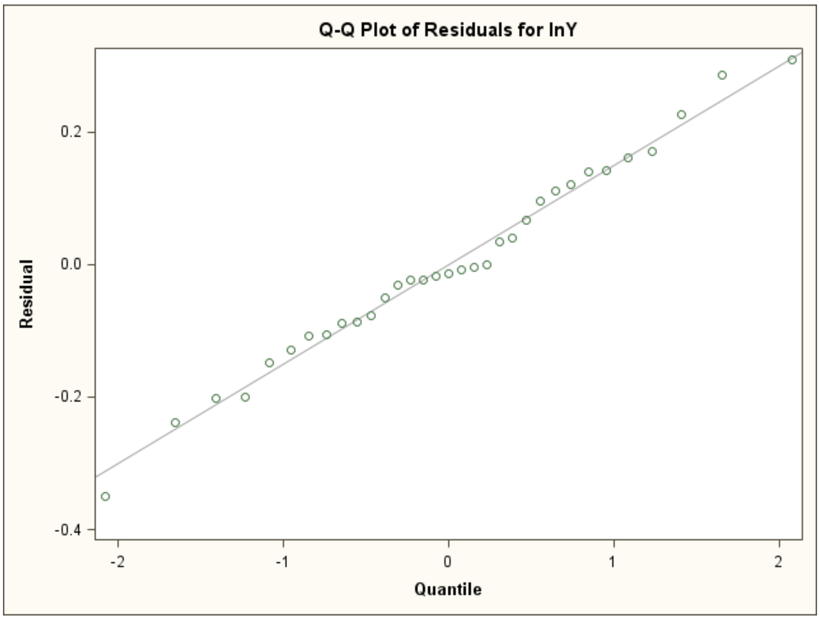
Refer to the above SAS output, the regression equation is

$$\hat{Y} = -2.0427 - 0.712 \log(X1) + 0.7474 \log(140 - X2) + 0.7575 \log(X3)$$

Residual plots (Studentised (deleted) residuals): Outliers?



Q-Q plot: Normality



b) *Obtain the variance inflation factors. Are there indications that serious multicollinearity problems exist here? Explain.*

Refer to the following SAS output, the VIFs do not indicate much problem in terms of multicollinearity (all VIFs are well below 5).

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	**1**	-2.04269	1.01919	-2.00	0.0545	0
lnX1	**1**	-0.71195	0.09203	-7.74	<.0001	***1.33932***
lnX3	**1**	0.75745	0.15923	4.76	<.0001	***1.01603***
ln(140-X2)	**1**	0.74736	0.15696	4.76	<.0001	***1.33011***

[Example 2] (Logistic Regression) Health study to investigate an epidemic outbreak of a disease

In a health study to investigate an epidemic outbreak of a disease that is spread by mosquitoes, individuals were randomly sampled within two sectors in a city to determine if the person had recently contracted the disease under study. This was ascertained by the interviewer, who asked pertinent questions to assess whether certain specific symptoms associated with the disease were present during the specified period. The response variable Y was coded 1 if this disease was determined to have been present, and 0 if not.

Three predictor variables

X1: Age
X2, X3: Indicator variables for socioeconomic status
Upper: X2 = 0, X3 = 0
Middle: X2 = 1, X3 = 0
Lower: X2 = 0, X3 = 1
X4: Indicator variable for city sector (0 for sector 1, 1 for sector 2)

Regression of Y on X1, X2, X3 and X4.

Analyze > Regression > Logistic >
Dependent variable: Y
Quantitative variables: X1, X2, X3, X4 >
Model > Response > Fit Model to level > 1
Model > Effects > Effects: Main X1, X2, X3, X4 > Run

SAS results
(A)

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

(B)

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	21.2635	4	0.0003
Score	20.4067	4	0.0004
Wald	16.6437	4	0.0023

(C)

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.3127	0.6426	12.9545	0.0003
X1	1	0.0297	0.0135	4.8535	0.0276
X2	1	0.4088	0.5990	0.4657	0.4950
X3	1	-0.3051	0.6041	0.2551	0.6135

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
X4	1	1.5746	0.5016	9.8543	0.0017

What is the estimated logistic response function?

$$\hat{\pi} = [1 + \exp(2.3129 - 0.0297X1 - .4088X2 + .3051X3 - 1.5746X4)]^{-1}$$

(D) Odds ratio estimates

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
X1	1.030	1.003	1.058
X2	1.505	0.465	4.868
X3	0.737	0.226	2.408
X4	4.829	1.807	12.907

Explanation:

- The odds of a person having contracted the disease increase by about 3 percent with each additional year of age, for given socioeconomic status and city sector location.
- The odds of a person in sector 2 (X_4) having contracted the disease are almost
- five times as great for a person in sector 1, for given socioeconomic status.

[Example 3] (Multicollinearity)

An assistant in the district sales office of a national cosmetics firm obtained data on advertising expenditures and sales last year in the district's 44 territories. X1 denotes expenditures for point-of-sales displays in beauty salons and department stores (in thousand dollars), and X2 and X3 represent the corresponding expenditures for local media advertising and prorated share of national media advertising, respectively. Y denotes sales (in thousand cases). The assistant was instructed to estimate the increase in expected sales when X1 is increased by 1 thousand dollars and X2 and X3 are held constant, and was told to use an ordinary multiple regression model with linear terms for the predictor variables and with independent normal error terms.

a) *State the regression model to be employed and fit it to the data.*

Analyze > Regression > Linear > … > Statistics > Variance inflation values > Run

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation

Parameter Estimates							
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Intercept	Intercept	1	1.02325	1.20287	0.85	0.4000	0
X1	point-of-sales-exp	1	0.96569	0.70922	1.36	0.1809	***20.07203***
X2	local-exp	1	0.62916	0.77830	0.81	0.4237	***20.71610***
X3	national-exp	1	0.67602	0.35574	1.90	0.0646	1.21797

Refer to the above SAS output, the regression equation is

$\hat{Y} = 1.02325 + .96569X_1 + .62916X_2 + .67603X_3$

b) *Test whether there is a regression relation between sales and the three predictor variables (use α = 0.05). State the alternatives, decision rule, and conclusion.*

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	3	382.65880	127.55293	38.28	<.0001
Error	40	133.28632	3.33216		
Corrected Total	43	515.94512			

Define $E(y\,|\,\boldsymbol{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

H_0: $\beta_1 = \beta_2 = \beta_3 = 0$
H_a: not all $\beta_k = 0$ (k = 1, 2, 3).
Test Statistics: MSR = 127.553 MSE = 3.33216 $F = 127.553/3.33216 = 38.28$

Critical Value
$F_{(.05;\,3,40)} = 2.84$. If $F \leq 2.84$ conclude H_0, otherwise H_a. Therefore, conclude H_a.
[Can use the p-value and arrive at the same conclusion]

c) *Test for each of the regression coefficients (equal to zero?) individually (use α = 0.05 each time). Do the conclusions of these tests correspond to that obtained in part (b)?*

Refer to the SAS output of part (a), all regression coefficients are not significant at the given alpha level. Therefore, these tests do not yield the same conclusion as in part (b). This is a consequence of the multi-collinearity problem.

d) *Obtain the correlation matrix of the X variables and comment on the suitability of the data for the research objective.*

From the correlation matrix below, we observe that the independent variables (X1 and X2) are highly correlated and the regression model is therefore not quite appropriate.

Analyze > Multivariate > Correlations > Analysis variables > Y, X1, X2, X3 > Run

Pearson Correlation Coefficients, N = 44 Prob > |r| under H0: Rho=0				
	Y	**X1**	**X2**	**X3**
Y	1.00000	0.84173	0.84248	0.47406
Sales		<.0001	<.0001	0.0012
X1	0.84173	1.00000	0.97443	0.37595
point-of-sales-exp	<.0001		<.0001	0.0119
X2	0.84248	0.97443	1.00000	0.40992
local-exp	<.0001	<.0001		0.0057
X3	0.47406	0.37595	0.40992	1.00000
national-exp	0.0012	0.0119	0.0057	

e) *Obtain the three variance inflation factors. What do these suggest about the effects of multicollinearity here?*

From the SAS output, we have

$(VIF)_1 = 20.072$

$(VIF)_2 = 20.716$

$(VIF)_3 = 1.218$

The problem is quite serious since two of the VIF are large than 5.

f) *The assistant eventually decided to drop variables X2 and X3 from the model to clear up the picture. Fit the assistant's revised model. Is the assistant now in a better position to achieve the research objective?*

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	3.16277	0.67118	4.71	<.0001
X1	Point-of-sales-exp	1	1.65806	0.16410	10.10	<.0001

Refer to the above SAS output, the regression equation is

$\hat{Y} = 3.16277 + 1.65806X_1$

Using only X1 is not an appropriate measure since X3 is not very highly correlated with other variables. Therefore, we should at least try to perform a regression analysis with X1 and X3.

Parameter Estimates

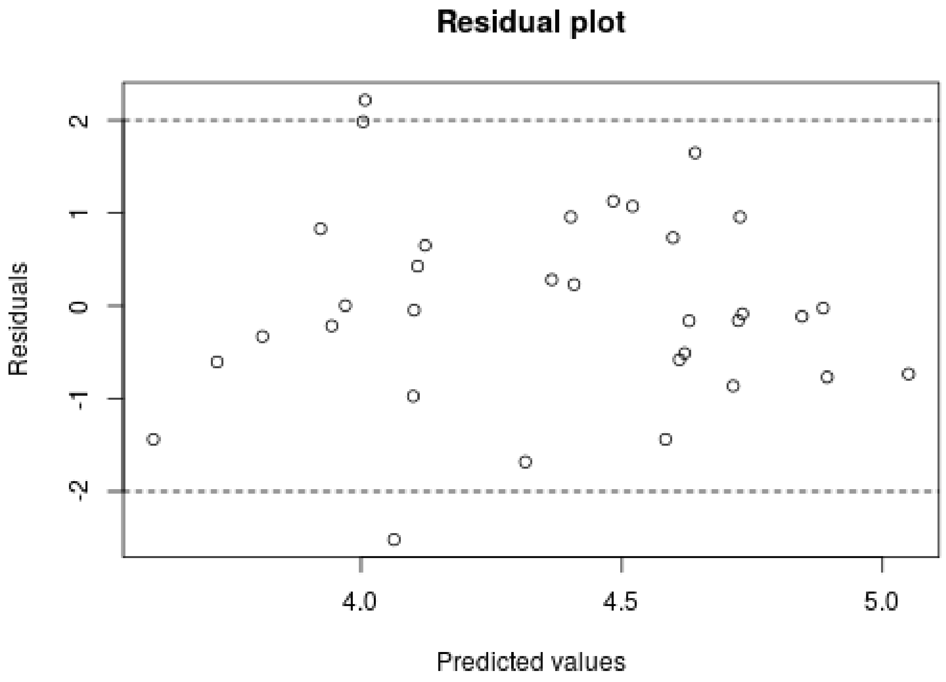
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	1.01729	1.19776	0.85	0.4006	0

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
X1	Point-of-sales-exp	1	1.52213	0.17011	8.95	<.0001	*1.16460*
X3	National-exp	1	0.73622	0.34639	2.13	0.0396	*1.16460*

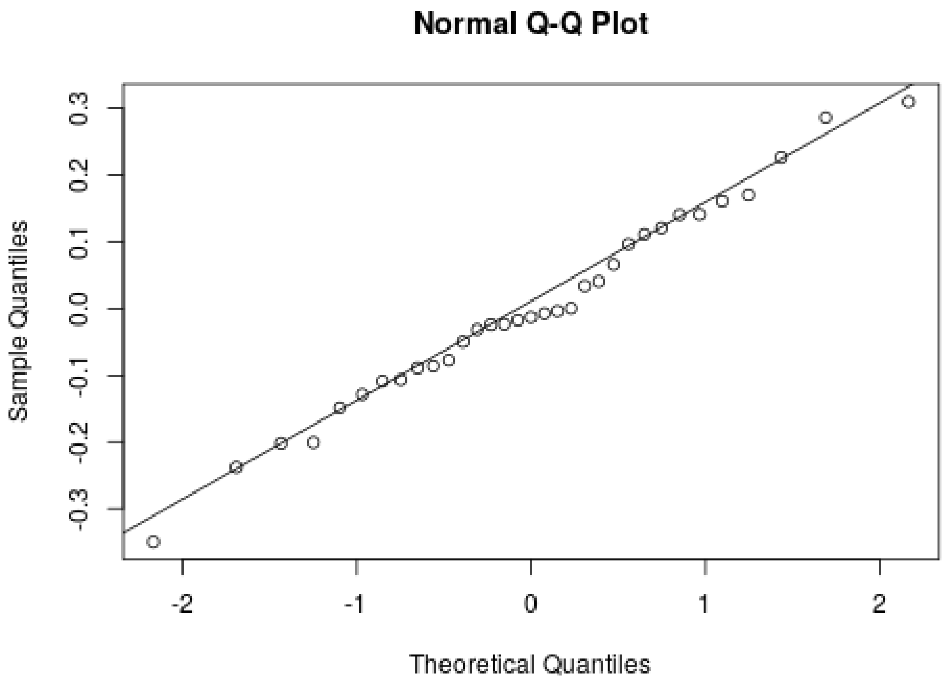
With reference to the SAS output, the VIF indicates that the problem of multicollinearity disappears (less than 5).

R Demonstration

```
# L3E1
# install.packages("car")
library(MASS)
library(car)
lab3_d1 = read.sas7bdat("ex1_data.sas7bdat")
# part a
lm_lab3_ex1 = lm(logY ~ logX1 + logX3 + logX2, data=lab3_d1)
coef(summary(lm_lab3_ex1))
#                Estimate Std. Error   t value     Pr(>|t|)
# (Intercept) -2.0426862 1.01919186 -2.004221 5.446473e-02
# logX1       -0.7119509 0.09202975 -7.736095 1.569202e-08
# logX3        0.7574464 0.15923410  4.756810 4.985002e-05
# logX2        0.7473627 0.15696014  4.761481 4.920793e-05
plot(lm_lab3_ex1[['fitted.values']], studres(lm_lab3_ex1), xlab="Predicted values", ylab="Residuals", main="Residual plot")
abline(h=c(-2,2), lty=2)
```




```
qqnorm(resid(lm_lab3_ex1))
```




```
qqline(resid(lm_lab3_ex1))

# part b
vif(lm_lab3_ex1)
#    logX1    logX3    logX2
# 1.339318 1.016032 1.330109
```

```
# L3E2
lab3_d2 = read.sas7bdat("ex2.sas7bdat")
# part a
lm_lab3_ex2_1 = glm(Y ~ Age + X2 + X3 + X4, data=lab3_d2, family="binomial")
coef(summary(lm_lab3_ex2_1))
#                Estimate Std. Error    z value     Pr(>|z|)
# (Intercept) -2.31293482 0.64258788 -3.5994062 0.0003189446
# Age          0.02975009 0.01350281  2.2032516 0.0275770178
# X2           0.40879024 0.59900377  0.6824502 0.4949543235
# X3          -0.30525456 0.60412836 -0.5052810 0.6133615163
# X4           1.57474923 0.50162060  3.1393233 0.0016933852
anova(lm_lab3_ex2_1, test='LRT')
# Analysis of Deviance Table
#
# Model: binomial, link: logit
#
# Response: Y
#
# Terms added sequentially (first to last)
#
#
#      Df Deviance Resid. Df Resid. Dev Pr(>Chi)
# NULL                    97     122.32
# Age   1   7.4050        96     114.91 0.006504 **
# X2    1   1.8040        95     113.11 0.179230
# X3    1   1.6064        94     111.50 0.205003
# X4    1  10.4481        93     101.05 0.001228 **
# ---
# Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
# L3E3
lab3_d3 = read.sas7bdat("cosmetics.sas7bdat")
# library(car)
# part a
lm_lab3_ex3_1 = lm(lab3_d3)
coef(summary(lm_lab3_ex3_1))
#              Estimate Std. Error   t value   Pr(>|t|)
# (Intercept) 1.0232513  1.2028750 0.8506714 0.40001560
# X1          0.9656902  0.7092217 1.3616197 0.18093842
# X2          0.6291644  0.7783009 0.8083820 0.42365267
# X3          0.6760246  0.3557408 1.9003291 0.06461444
vif(lm_lab3_ex3_1)
#        X1        X2        X3
# 20.072031 20.716101  1.217973

# part b
# summary(lm_lab3_ex3_1)
# p-value at the bottom in summary(lm_lab3_ex3_1)

# part c
# p-values from each coefficients in summary(lm_lab3_ex3_1)

# part d
cor(lab3_d3)
#            Y        X1        X2        X3
# Y  1.0000000 0.8417342 0.8424849 0.4740581
# X1 0.8417342 1.0000000 0.9744313 0.3759509
# X2 0.8424849 0.9744313 1.0000000 0.4099208
# X3 0.4740581 0.3759509 0.4099208 1.0000000

# part e
# refer to the VIFs derived in part a

# part f
lm_lab3_ex3_2 = lm(Y ~ X1, data=lab3_d3)
summary(lm_lab3_ex3_1, correlation=T)$correlation
#              (Intercept)          X1           X2         X3
# (Intercept)  1.000000000 -0.03838193  0.006127475 -0.8253631
# X1          -0.038381931  1.00000000 -0.970555905  0.1146127
# X2           0.006127475 -0.97055590  1.000000000 -0.2093274
# X3          -0.825363075  0.11461266 -0.209327402  1.0000000
lm_lab3_ex3_3 = lm(Y ~ X1 + X3, data=lab3_d3)
vif(lm_lab3_ex3_3)
#       X1       X3
# 1.164604 1.164604
```