S&DS 563 / F&ES 758b - Multivariate Statistics Homework #5 MANOVA and Multivariate GLM

Lanxin Jiang (lj345), Grace Sun (ys544), Chenglin Lu (cl939)

2018-04-07

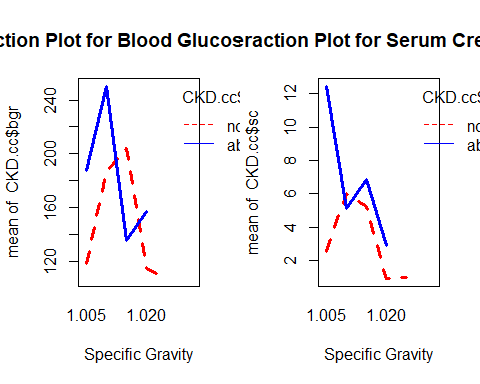
```  
## Loading required package: rpanel  
```  
  
```  
## Loading required package: tcltk  
```  
  
```  
## Package `rpanel', version 1.1-3: type help(rpanel) for summary information  
```  
  
```  
## Loading required package: tkrplot  
```  
  
```  
## Loading required package: lattice  
```  
  
```  
## Loading required package: SpatialEpi  
```  
  
```  
## Loading required package: sp  
```  
  
```  
## ---  
## biotools version 3.1  
```  
  
```  
##   
```

library(ggplot2)  
multiplot <- function(..., plotlist=NULL, file, cols=1, layout=NULL) {  
 library(grid)  
  
 # Make a list from the ... arguments and plotlist  
 plots <- c(list(...), plotlist)  
  
 numPlots = length(plots)  
  
 # If layout is NULL, then use 'cols' to determine layout  
 if (is.null(layout)) {  
 # Make the panel  
 # ncol: Number of columns of plots  
 # nrow: Number of rows needed, calculated from # of cols  
 layout <- matrix(seq(1, cols \* ceiling(numPlots/cols)),  
 ncol = cols, nrow = ceiling(numPlots/cols))  
 }  
   
 if (numPlots==1) {  
 print(plots[[1]])  
  
 } else {  
 # Set up the page  
 grid.newpage()  
 pushViewport(viewport(layout = grid.layout(nrow(layout), ncol(layout))))  
  
 # Make each plot, in the correct location  
 for (i in 1:numPlots) {  
 # Get the i,j matrix positions of the regions that contain this subplot  
 matchidx <- as.data.frame(which(layout == i, arr.ind = TRUE))  
  
 print(plots[[i]], vp = viewport(layout.pos.row = matchidx$row,  
 layout.pos.col = matchidx$col))  
 }  
 }  
}

```r  
Sys.setenv(JAVA\_HOME="C:\\Program Files\\Java\\jre-9.0.4\\")  
library(rJava)  
library(RWeka)  
CKD <- read.arff("C:/Users/lanxin/Documents/GitHub/Chronic\_Kidney\_Disease/Chronic\_Kidney\_Disease/chronic\_kidney\_disease\_full.arff")  
```  
  
1. For two categorical factors, make interaction plots for each of your response variables. Discuss what you see. If you only have one categorical factor, then provide some plot/discussion of which means are different for which response variables  
Two categorical factors: Specific Gravity(sg) of urine,Red Blood Cells(rbc)  
Two continous:Blood Glucose Random(bgr), Serum Creatinine(sc)

CKD\_hw5<-CKD[,c(3,6,10,12,15)]  
 CKD.cc<-CKD\_hw5[complete.cases(CKD\_hw5),]  
 CKD.cc$rbc<-as.factor(CKD.cc$rbc)  
 write.csv(CKD.cc,file="C:/Users/lanxin/Documents/GitHub/Chronic\_Kidney\_Disease/HW5/CKDCC.csv")

par(mfrow=c(1,2))  
#this makes the plots  
interaction.plot(CKD.cc$sg,CKD.cc$rbc,CKD.cc$bgr, lwd=3,col=c("red","blue"),xlab="Specific Gravity",main="Interaction Plot for Blood Glucose Random")  
interaction.plot(CKD.cc$sg,CKD.cc$rbc,CKD.cc$sc, lwd=3,col=c("red","blue"),xlab="Specific Gravity",main="Interaction Plot for Serum Creatinine")

 These plots suggest that there may be an interaction between specific graviity and red blood cells on serum creatinine and blood glucose. Also suggests that there may not be much difference of RBC on serum creatinine and glucose value. Overall, it seems that normal red blood cells have higher serum creatinine concentration and specific gravity of urine. Two plots also indicate that When SG=1.025, there is no observation where red blood cell is abnormal.

1. Run Two-Way MANOVA for these two categorical factors. Discuss your results, both univariate and multivariate. If only one categorical predictor, do one-way MANOVA.

#fit linear model  
mod1=manova(as.matrix(CKD.cc[,3:4])~CKD.cc$sg + CKD.cc$rbc +CKD.cc$sg\*CKD.cc$rbc)  
#get univariate results   
summary.aov(mod1)

## Response bgr :  
## Df Sum Sq Mean Sq F value Pr(>F)   
## CKD.cc$sg 4 301474 75368 21.6350 8.944e-15 \*\*\*  
## CKD.cc$rbc 1 16413 16413 4.7116 0.031174 \*   
## CKD.cc$sg:CKD.cc$rbc 3 58801 19600 5.6264 0.001019 \*\*   
## Residuals 194 675824 3484   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Response sc :  
## Df Sum Sq Mean Sq F value Pr(>F)   
## CKD.cc$sg 4 946.77 236.692 39.1767 < 2.2e-16 \*\*\*  
## CKD.cc$rbc 1 35.49 35.493 5.8748 0.0162770 \*   
## CKD.cc$sg:CKD.cc$rbc 3 131.88 43.960 7.2762 0.0001193 \*\*\*  
## Residuals 194 1172.08 6.042   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#get multivariate results summary.manova(mod1)   
summary.manova(mod1,test="Wilks")

## Df Wilks approx F num Df den Df Pr(>F)   
## CKD.cc$sg 4 0.40169 27.8795 8 386 < 2.2e-16 \*\*\*  
## CKD.cc$rbc 1 0.94029 6.1279 2 193 0.002629 \*\*   
## CKD.cc$sg:CKD.cc$rbc 3 0.82604 6.4506 6 386 1.704e-06 \*\*\*  
## Residuals 194   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

mod1$coefficients

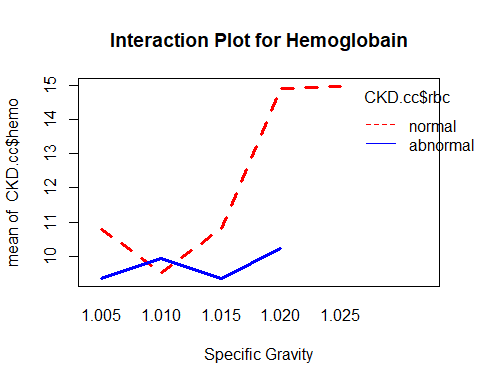
## bgr sc  
## (Intercept) 118.666667 2.533333  
## CKD.cc$sg1.010 68.196970 3.471212  
## CKD.cc$sg1.015 85.500000 2.700000  
## CKD.cc$sg1.020 -3.857843 -1.602451  
## CKD.cc$sg1.025 -11.393939 -1.524242  
## CKD.cc$rbcabnormal 68.833333 9.916667  
## CKD.cc$sg1.010:CKD.cc$rbcabnormal -5.543124 -10.798135  
## CKD.cc$sg1.015:CKD.cc$rbcabnormal -137.750000 -8.300000  
## CKD.cc$sg1.020:CKD.cc$rbcabnormal -26.864379 -7.958660  
## CKD.cc$sg1.025:CKD.cc$rbcabnormal NA NA

Univariate Results : For blood glucose, there are significant differences between either sg or rbc. Also, there, there is evidence on an interaction effect. But for serum creatinine, rbc is not a significant factor. Multivariate Results : Both covariate and the interaction term is significant. ….

The coefficient for sg=1.025\*rbc=abnormal is NA, which corresponds to that When SG=1.025, there is no observation where red blood cell is abnormal.

1. Perform (multivariate) contrasts to compare levels of a particular factor or combinations of factors. Discuss your results. If you’re using SPSS/R, you won’t be able to do the multivariate contrasts. See SAS code and rtf output for details
2. If applicable, add a continuous variable to your model and fit as a multiple-response generalized linear model. Before you fit the model, make some plots to see if there are linear relationships between your covariates and your responses.

#relationship  
interaction.plot(CKD.cc$sg,CKD.cc$rbc,CKD.cc$hemo, lwd=3,col=c("red","blue"),xlab="Specific Gravity",main="Interaction Plot for Hemoglobain")



#fit linear model  
mod1=lm(as.matrix(CKD.cc[,3:5])~CKD.cc$sg + CKD.cc$rbc +CKD.cc$sg\*CKD.cc$rbc)  
#get univariate results   
mod1

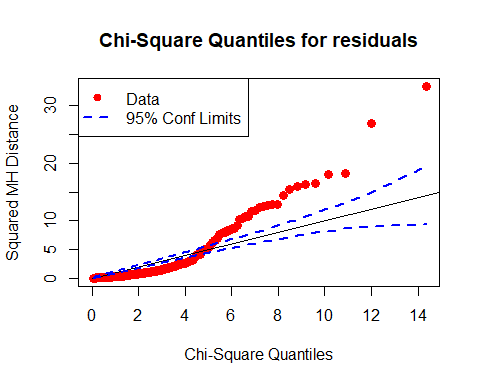
##   
## Call:  
## lm(formula = as.matrix(CKD.cc[, 3:5]) ~ CKD.cc$sg + CKD.cc$rbc +   
## CKD.cc$sg \* CKD.cc$rbc)  
##   
## Coefficients:  
## bgr sc hemo   
## (Intercept) 118.66667 2.53333 10.76667  
## CKD.cc$sg1.010 68.19697 3.47121 -1.24394  
## CKD.cc$sg1.015 85.50000 2.70000 0.04167  
## CKD.cc$sg1.020 -3.85784 -1.60245 4.14510  
## CKD.cc$sg1.025 -11.39394 -1.52424 4.20152  
## CKD.cc$rbcabnormal 68.83333 9.91667 -1.41667  
## CKD.cc$sg1.010:CKD.cc$rbcabnormal -5.54312 -10.79814 1.84009  
## CKD.cc$sg1.015:CKD.cc$rbcabnormal -137.75000 -8.30000 -0.04167  
## CKD.cc$sg1.020:CKD.cc$rbcabnormal -26.86438 -7.95866 -3.27288  
## CKD.cc$sg1.025:CKD.cc$rbcabnormal NA NA NA

summary(mod1)

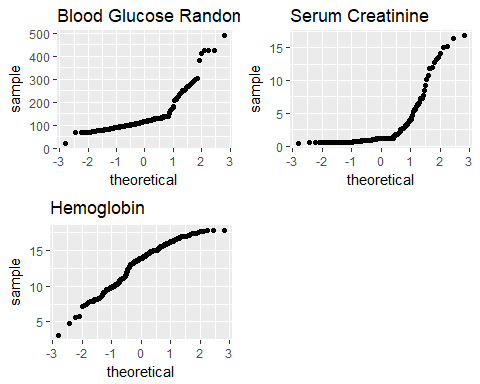
## Response bgr :  
##   
## Call:  
## lm(formula = bgr ~ CKD.cc$sg + CKD.cc$rbc + CKD.cc$sg \* CKD.cc$rbc)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -171.154 -23.309 -5.809 17.459 309.191   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 118.667 34.077 3.482 0.000614 \*\*\*  
## CKD.cc$sg1.010 68.197 36.326 1.877 0.061967 .   
## CKD.cc$sg1.015 85.500 38.099 2.244 0.025951 \*   
## CKD.cc$sg1.020 -3.858 34.820 -0.111 0.911895   
## CKD.cc$sg1.025 -11.394 34.842 -0.327 0.744010   
## CKD.cc$rbcabnormal 68.833 53.880 1.278 0.202939   
## CKD.cc$sg1.010:CKD.cc$rbcabnormal -5.543 57.700 -0.096 0.923566   
## CKD.cc$sg1.015:CKD.cc$rbcabnormal -137.750 60.239 -2.287 0.023292 \*   
## CKD.cc$sg1.020:CKD.cc$rbcabnormal -26.864 57.804 -0.465 0.642633   
## CKD.cc$sg1.025:CKD.cc$rbcabnormal NA NA NA NA   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 59.02 on 194 degrees of freedom  
## Multiple R-squared: 0.3579, Adjusted R-squared: 0.3314   
## F-statistic: 13.52 on 8 and 194 DF, p-value: 1.705e-15  
##   
##   
## Response sc :  
##   
## Call:  
## lm(formula = sc ~ CKD.cc$sg + CKD.cc$rbc + CKD.cc$sg \* CKD.cc$rbc)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.1500 -0.4309 -0.1309 0.1909 10.8955   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.533 1.419 1.785 0.07580 .   
## CKD.cc$sg1.010 3.471 1.513 2.295 0.02283 \*   
## CKD.cc$sg1.015 2.700 1.587 1.702 0.09041 .   
## CKD.cc$sg1.020 -1.602 1.450 -1.105 0.27049   
## CKD.cc$sg1.025 -1.524 1.451 -1.050 0.29481   
## CKD.cc$rbcabnormal 9.917 2.244 4.420 1.64e-05 \*\*\*  
## CKD.cc$sg1.010:CKD.cc$rbcabnormal -10.798 2.403 -4.494 1.20e-05 \*\*\*  
## CKD.cc$sg1.015:CKD.cc$rbcabnormal -8.300 2.509 -3.309 0.00112 \*\*   
## CKD.cc$sg1.020:CKD.cc$rbcabnormal -7.959 2.407 -3.306 0.00113 \*\*   
## CKD.cc$sg1.025:CKD.cc$rbcabnormal NA NA NA NA   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.458 on 194 degrees of freedom  
## Multiple R-squared: 0.4873, Adjusted R-squared: 0.4662   
## F-statistic: 23.05 on 8 and 194 DF, p-value: < 2.2e-16  
##   
##   
## Response hemo :  
##   
## Call:  
## lm(formula = hemo ~ CKD.cc$sg + CKD.cc$rbc + CKD.cc$sg \* CKD.cc$rbc)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.4227 -1.0118 0.0318 1.1428 6.7500   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 10.76667 1.05275 10.227 < 2e-16 \*\*\*  
## CKD.cc$sg1.010 -1.24394 1.12224 -1.108 0.269041   
## CKD.cc$sg1.015 0.04167 1.17701 0.035 0.971797   
## CKD.cc$sg1.020 4.14510 1.07572 3.853 0.000158 \*\*\*  
## CKD.cc$sg1.025 4.20152 1.07641 3.903 0.000131 \*\*\*  
## CKD.cc$rbcabnormal -1.41667 1.66454 -0.851 0.395772   
## CKD.cc$sg1.010:CKD.cc$rbcabnormal 1.84009 1.78258 1.032 0.303234   
## CKD.cc$sg1.015:CKD.cc$rbcabnormal -0.04167 1.86102 -0.022 0.982161   
## CKD.cc$sg1.020:CKD.cc$rbcabnormal -3.27288 1.78579 -1.833 0.068375 .   
## CKD.cc$sg1.025:CKD.cc$rbcabnormal NA NA NA NA   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.823 on 194 degrees of freedom  
## Multiple R-squared: 0.6416, Adjusted R-squared: 0.6268   
## F-statistic: 43.42 on 8 and 194 DF, p-value: < 2.2e-16

1. Check model assumptions by making a chi-square quantile plot of the residuals. Modify your model as appropriate based on your findings.

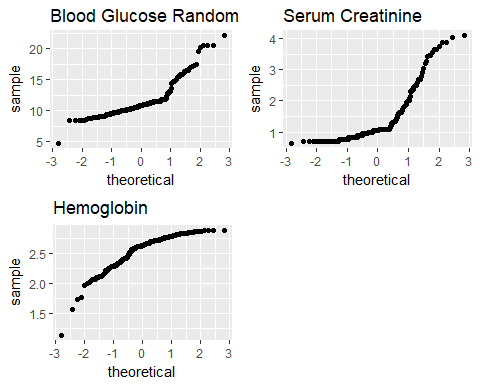
source("http://www.reuningscherer.net/STAT660/R/CSQPlot.r.txt")  
#run the function  
CSQPlot(mod1$residuals,label="residuals")



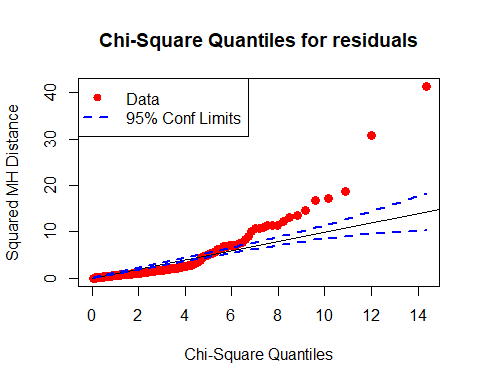
#check normality for each variable  
  
#check QQ plot for some numeric variables  
gg.qqplot <- function (variable, title, data) {  
ggplot(data, aes\_string(sample=variable)) + ggtitle(title) + stat\_qq()  
}  
p2 <- gg.qqplot("bgr", "Blood Glucose Random",CKD.cc)  
p3 <- gg.qqplot("hemo", "Hemoglobin",CKD.cc)  
p4 <- gg.qqplot("sc", "Serum Creatinine",CKD.cc)  
multiplot( p2, p3, p4, cols=2)



CKD.cc2=CKD.cc  
CKD.cc2$bgr <-sqrt(CKD.cc$bgr)  
CKD.cc2$sc <-sqrt(CKD.cc$sc)  
CKD.cc2$hemo <-log(CKD.cc$hemo)  
p5 <- gg.qqplot("bgr", "Blood Glucose Random",CKD.cc2)  
p6 <- gg.qqplot("hemo", "Hemoglobin",CKD.cc2)  
p7 <- gg.qqplot("sc", "Serum Creatinine",CKD.cc2)  
multiplot( p5, p6, p7, cols=2)



CKD.cc[,6:8]=log(CKD.cc[,3:5])  
mod2=lm(as.matrix(CKD.cc[,6:8])~CKD.cc$sg + CKD.cc$rbc +CKD.cc$sg\*CKD.cc$rbc)  
CSQPlot(mod2$residuals,label="residuals")



Boxcox transformation

library(car)  
trans <- powerTransform(as.matrix(CKD.cc[,3:5])~CKD.cc$sg + CKD.cc$rbc +CKD.cc$sg\*CKD.cc$rbc, family="bcPower")  
summary(trans)

## bcPower Transformations to Multinormality   
## Est Power Rounded Pwr Wald Lwr bnd Wald Upr Bnd  
## bgr -0.1808 -0.33 -0.3587 -0.0028  
## sc -0.3608 -0.33 -0.4848 -0.2368  
## hemo 1.5556 1.56 1.1639 1.9473  
##   
## Likelihood ratio tests about transformation parameters  
## LRT df pval  
## LR test, lambda = (0 0 0) 124.7737 3 0  
## LR test, lambda = (1 1 1) 630.0641 3 0

testTransform(trans, c(0, 0,-1))

## LRT df pval  
## LR test, lambda = (0 0 -1) 328.4855 3 0

testTransform(trans, c(-1, 0,0))

## LRT df pval  
## LR test, lambda = (-1 0 0) 214.3969 3 0

testTransform(trans, c(0, -1,0))

## LRT df pval  
## LR test, lambda = (0 -1 0) 173.1624 3 0

transformedY <- bcPower(with(CKD.cc, cbind(sc,hemo,bgr)),coef(trans, round=TRUE))  
 colnames(transformedY)<-c("sc","hemo","bgr")  
 mod3=lm(transformedY~CKD.cc$sg + CKD.cc$rbc +CKD.cc$sg\*CKD.cc$rbc)  
 summary(mod3)

## Response sc :  
##   
## Call:  
## lm(formula = sc ~ CKD.cc$sg + CKD.cc$rbc + CKD.cc$sg \* CKD.cc$rbc)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.97649 -0.32389 0.04737 0.33153 1.63983   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.6234 0.2553 2.442 0.01551 \*   
## CKD.cc$sg1.010 0.4998 0.2722 1.836 0.06786 .   
## CKD.cc$sg1.015 0.3951 0.2855 1.384 0.16790   
## CKD.cc$sg1.020 -0.8084 0.2609 -3.098 0.00223 \*\*  
## CKD.cc$sg1.025 -0.7780 0.2611 -2.980 0.00325 \*\*  
## CKD.cc$rbcabnormal 1.0572 0.4037 2.619 0.00952 \*\*  
## CKD.cc$sg1.010:CKD.cc$rbcabnormal -1.2039 0.4323 -2.785 0.00589 \*\*  
## CKD.cc$sg1.015:CKD.cc$rbcabnormal -0.7664 0.4514 -1.698 0.09110 .   
## CKD.cc$sg1.020:CKD.cc$rbcabnormal -0.1700 0.4331 -0.392 0.69519   
## CKD.cc$sg1.025:CKD.cc$rbcabnormal NA NA NA NA   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4422 on 194 degrees of freedom  
## Multiple R-squared: 0.6475, Adjusted R-squared: 0.633   
## F-statistic: 44.55 on 8 and 194 DF, p-value: < 2.2e-16  
##   
##   
## Response hemo :  
##   
## Call:  
## lm(formula = hemo ~ CKD.cc$sg + CKD.cc$rbc + CKD.cc$sg \* CKD.cc$rbc)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.61918 -0.02703 0.00592 0.04029 0.27639   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 1.646261 0.048654 33.836 < 2e-16  
## CKD.cc$sg1.010 -0.082886 0.051865 -1.598 0.11165  
## CKD.cc$sg1.015 -0.005325 0.054397 -0.098 0.92213  
## CKD.cc$sg1.020 0.138253 0.049715 2.781 0.00596  
## CKD.cc$sg1.025 0.139641 0.049747 2.807 0.00551  
## CKD.cc$rbcabnormal -0.068429 0.076928 -0.890 0.37483  
## CKD.cc$sg1.010:CKD.cc$rbcabnormal 0.109832 0.082384 1.333 0.18404  
## CKD.cc$sg1.015:CKD.cc$rbcabnormal -0.029846 0.086009 -0.347 0.72896  
## CKD.cc$sg1.020:CKD.cc$rbcabnormal -0.105898 0.082532 -1.283 0.20098  
## CKD.cc$sg1.025:CKD.cc$rbcabnormal NA NA NA NA  
##   
## (Intercept) \*\*\*  
## CKD.cc$sg1.010   
## CKD.cc$sg1.015   
## CKD.cc$sg1.020 \*\*   
## CKD.cc$sg1.025 \*\*   
## CKD.cc$rbcabnormal   
## CKD.cc$sg1.010:CKD.cc$rbcabnormal   
## CKD.cc$sg1.015:CKD.cc$rbcabnormal   
## CKD.cc$sg1.020:CKD.cc$rbcabnormal   
## CKD.cc$sg1.025:CKD.cc$rbcabnormal   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.08427 on 194 degrees of freedom  
## Multiple R-squared: 0.5628, Adjusted R-squared: 0.5448   
## F-statistic: 31.22 on 8 and 194 DF, p-value: < 2.2e-16  
##   
##   
## Response bgr :  
##   
## Call:  
## lm(formula = bgr ~ CKD.cc$sg + CKD.cc$rbc + CKD.cc$sg \* CKD.cc$rbc)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3254.1 -368.2 -122.2 207.4 6771.7   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1137.92 693.06 1.642 0.1022   
## CKD.cc$sg1.010 1230.37 738.80 1.665 0.0975 .  
## CKD.cc$sg1.015 1593.37 774.86 2.056 0.0411 \*  
## CKD.cc$sg1.020 -52.91 708.18 -0.075 0.9405   
## CKD.cc$sg1.025 -198.74 708.63 -0.280 0.7794   
## CKD.cc$rbcabnormal 1150.54 1095.82 1.050 0.2951   
## CKD.cc$sg1.010:CKD.cc$rbcabnormal 310.21 1173.52 0.264 0.7918   
## CKD.cc$sg1.015:CKD.cc$rbcabnormal -2460.62 1225.16 -2.008 0.0460 \*  
## CKD.cc$sg1.020:CKD.cc$rbcabnormal -447.76 1175.63 -0.381 0.7037   
## CKD.cc$sg1.025:CKD.cc$rbcabnormal NA NA NA NA   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1200 on 194 degrees of freedom  
## Multiple R-squared: 0.334, Adjusted R-squared: 0.3065   
## F-statistic: 12.16 on 8 and 194 DF, p-value: 4.834e-14

no other way to do it better