Linear Model Project

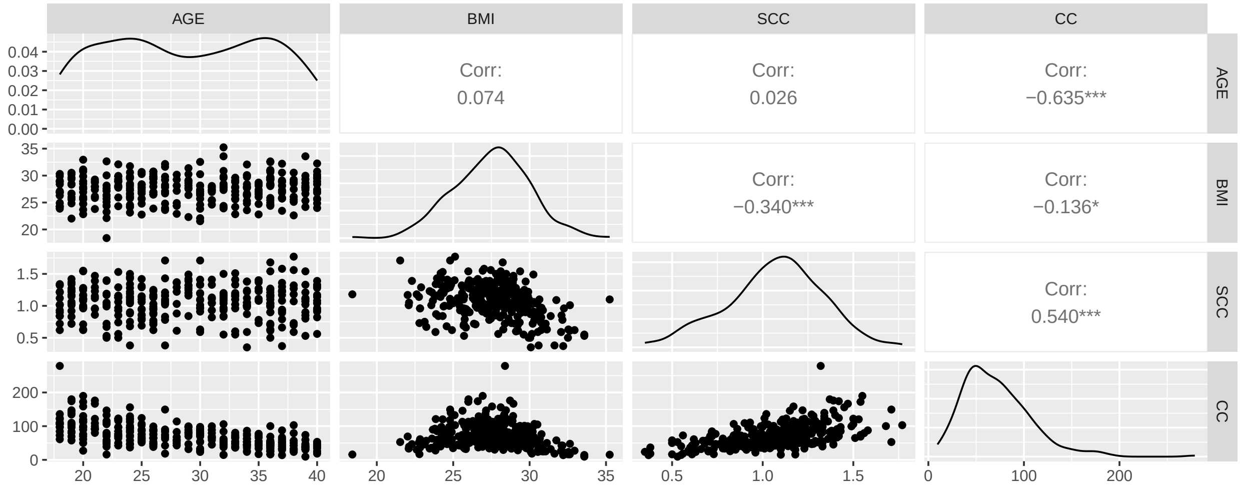
Chris Corona

Introduction

Creatinine clearance (CC) is a measure of kidney function, but is clinically difficult to obtain. A researcher wishes to study whether CC can be predicted from other easily available characteristics from patients. The researcher randomly selected 300 patients between 18 and 40 years old and recorded data. There are 300 observations of nine variables (1) `ID`: identity number for each patient (2) `CC`: creatinine clearance measurement, this is the variable of interest, (3) AGE: the age of patients when taking the measurements, (4) `BMI`: body mass index, (5) `SCC`: serum creatinine concentration, (6) `SEX`: 0 female, 1 male, (7) `BLOOD`: 1 the patient has high blood pressure, 0 otherwise, (9) `DISEASE`: 1 patient develops kidney disease (including death), 0 otherwise. The goal of this analysis is to find the best model that can predict creatinine clearance using all the other variables (except disease). In the last section, we include disease and try to predict whether a new patient will develop some kidney disease (including death) within the next year.

Exploratory Data Analysis

We perform some exploratory data analysis, but first we remove the disease column as we will return to it later. Next, we want to understand whether any of the covariates have a relationship with the response variable, creatinine clearance (CC). After some exploration we notice that BMI appears to have a quadratic relationship with CC, so we transform the response, CC, by squaring it. Figure 1 shows this relationship with BMI before and after transforming the response. After transforming there still appears to be a slightly non-linear relationship, but it is better. In Figure 2 we plot a correlation matrix of all the numeric variables with CC.SQ to see the other relationships. It looks like age has a slightly negative relationship with CC.SQ. It also appears that SCC has a slightly positive relationship with CC.SQ, but with potential non-constant variance - we will explore this further during model selection.



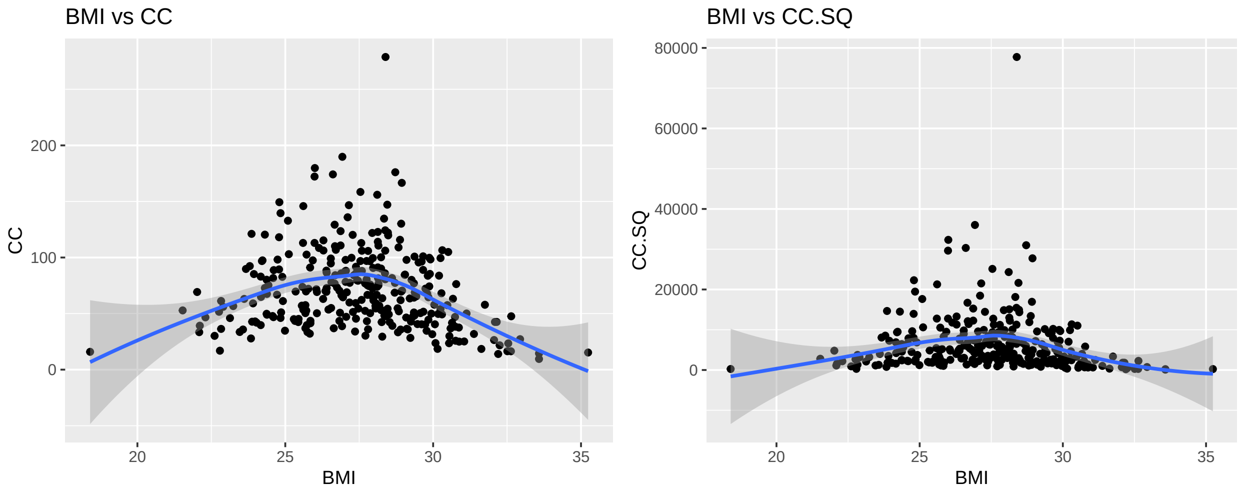


Figure 1: Comparison of plots before and after transforming the response, BMI vs CC and CC.SQ

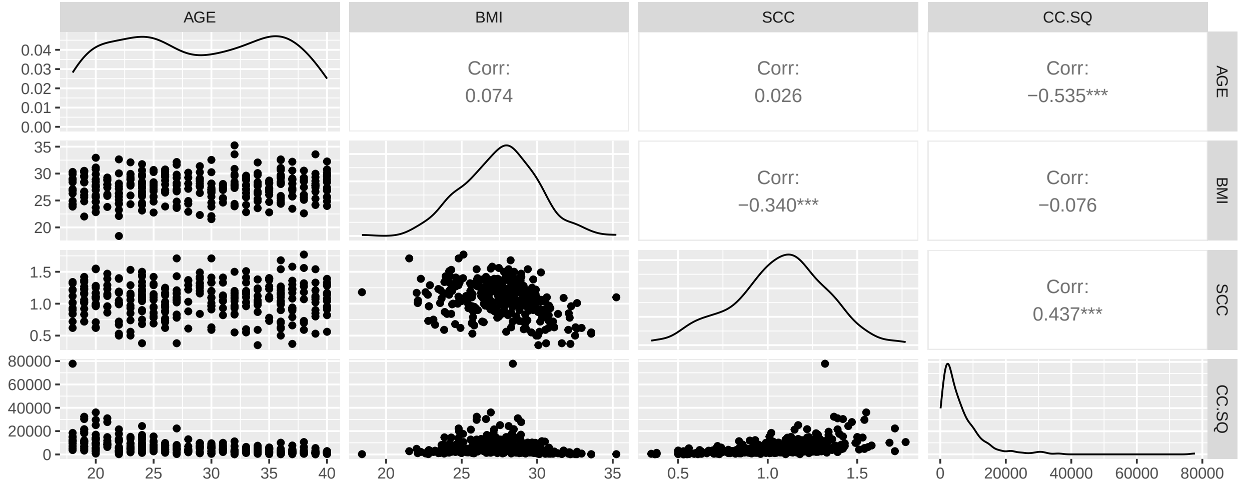


Figure 2: Correlation matrix of all numeric variables withh CC.SQ

Methods

In this section we perform a forward selection procedure to fit the best model for predicting CC from age, BMI, SSC, sex, and blood. As mentioned in the EDA section, we have squared CC to have a more linear relationship with the covariates. We start the forward selection process by including only one covariate and choosing the best model by comparing BIC scores. At the next step, we keep the covariate from the best single-variable model and add covariates one at a time to find the best two-, three-, four-, and five-variable models. We make sure to check for interactions at each step, again choosing the model with the lowest BIC score. Once all models are fit, we choose the best overall model according to its BIC score. This best overall model has four variables: the main effects from BMI, blood, age, and SCC, and also the interaction between age and SCC. To verify this interaction, Figure 3 shows the effects plot for this model. The diagnostics for this best overall model are shown in Figure 4. There is still a slight amount of non-linearity in the residuals vs fitted plot even after transforming the response, but it is not a glaring violation. The residuals vs fitted plot also indicates a small amount of non-constant variance. To ameliorate this, we tried fitting several non-constant variance models including: varIdent on the categorical variables, and then varFixed, varPower, varExp, and varConstPower on the numerical variables. We even tried interactions with different categorical variables in the variance structure. The best non-constant variance model, with the lowest BIC score, is varConstPower with form=~SCC|SEX. Figure 5 shows the comparison of the normalized residuals vs fitted for the best overall model before and after fitting the best non-constant variance structure. The fanning in this figure still has a similar shape as before, but now the scale is significantly reduced. The last step is to center the data to allow for more sensible interpretations. After the forward selection process and the non-constant variance structure has been fitted, the final best model is

$$y_i = \beta_0 + \beta_1 x_{AGE} + \beta_2 x_{SCC} + \beta_3 I(x_{BLOOD}) + \beta_4 x_{BMI} + \beta_5 x_{AGE} \times x_{AGE} + \epsilon_i$$
$$\epsilon_i \sim N\Big(0, \sigma^2(\delta_1 + |x_{SCC} \times I(x_{BLOOD})|^{\delta_2})\Big)$$

y_i is the squared creatinine clearance measurement of the patient,

x_{AGE} is the age of the patient

x_{SCC} is serum creatinine concentration

x_{BLOOD} is 1 if the patient has high blood pressure, 0 otherwise

x_{BMI} is the body mass index of the patient

β_0 is the expected squared creatinine clearance for the average patient (age=29, SCC=1.1, BMI=28, no high blood pressure)

β_1 is the expected change in squared creatinine clearance for a 1 unit increase in age, holding all other variables constant

β_2 is the expected change in squared creatinine clearance for a 1 unit increase in SCC, holding all other variables constant

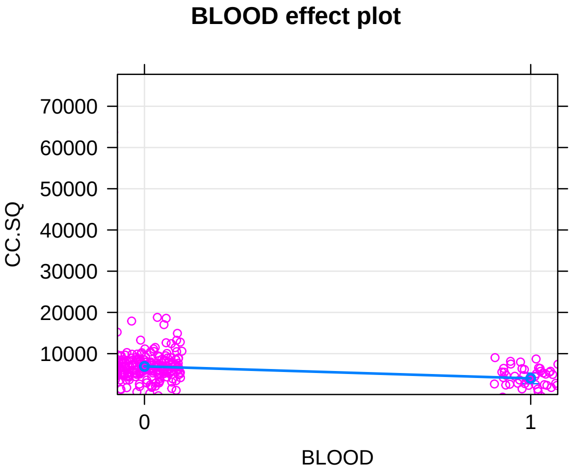
β_3 is the expected change in squared creatinine clearance for a patient with high blood pressure, holding all other variables constant

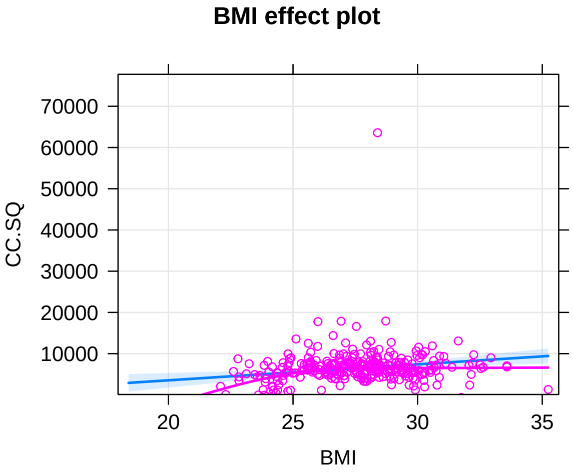
β_4 is the expected change in squared creatinine clearance for a 1 unit increase in BMI, holding all other variables constant

β_5 is the expected change in squared creatinine clearance for a 1 unit increase in the interaction between age and SCC, holding all other variables constant

ϵ_i is the error term

δ_1 and δ_2 are two additional estimated parameters for the non-constant variance structure


BLOOD effect plot


BMI effect plot

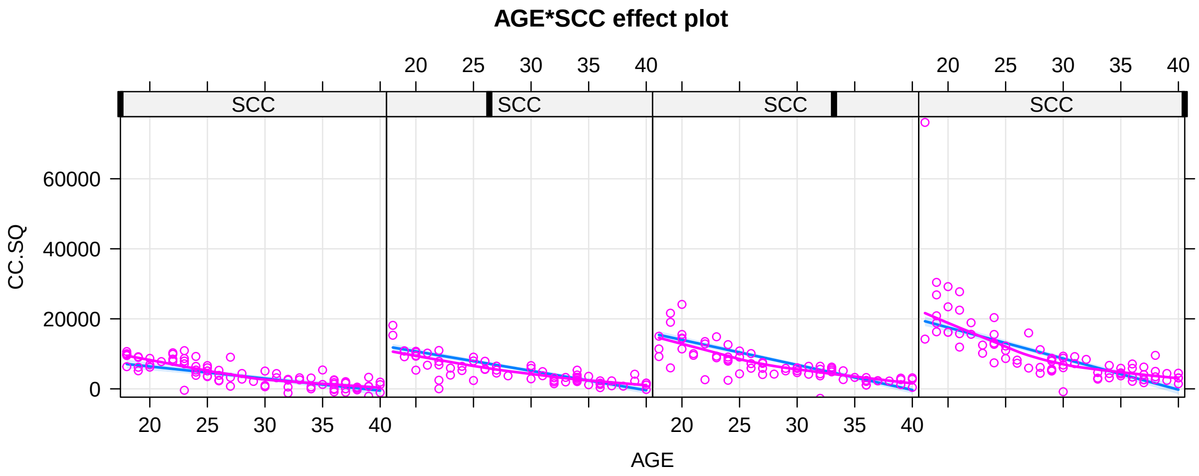


Figure 3: Effects plot for the best overall model

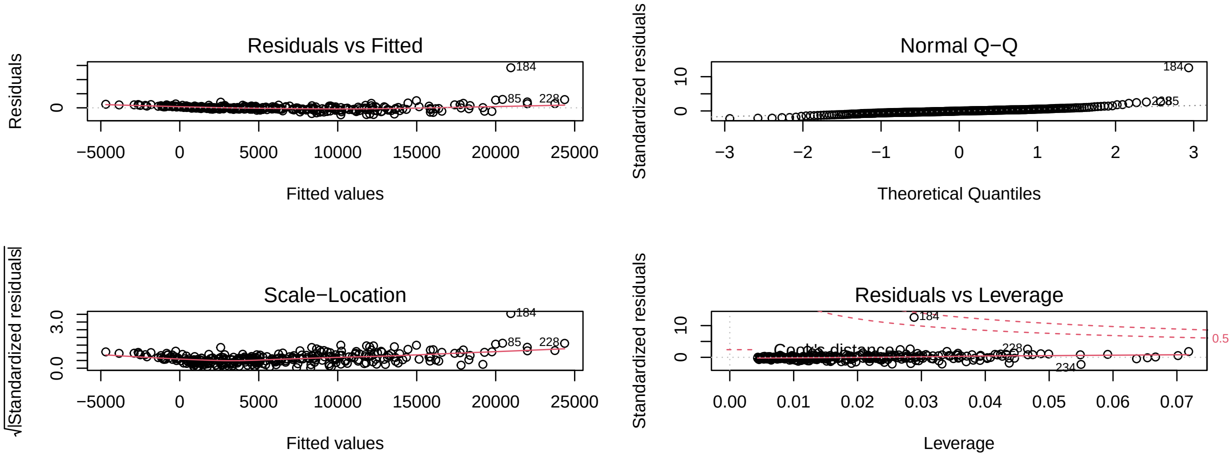


Figure 4: Diagnostics plots for the overall best model

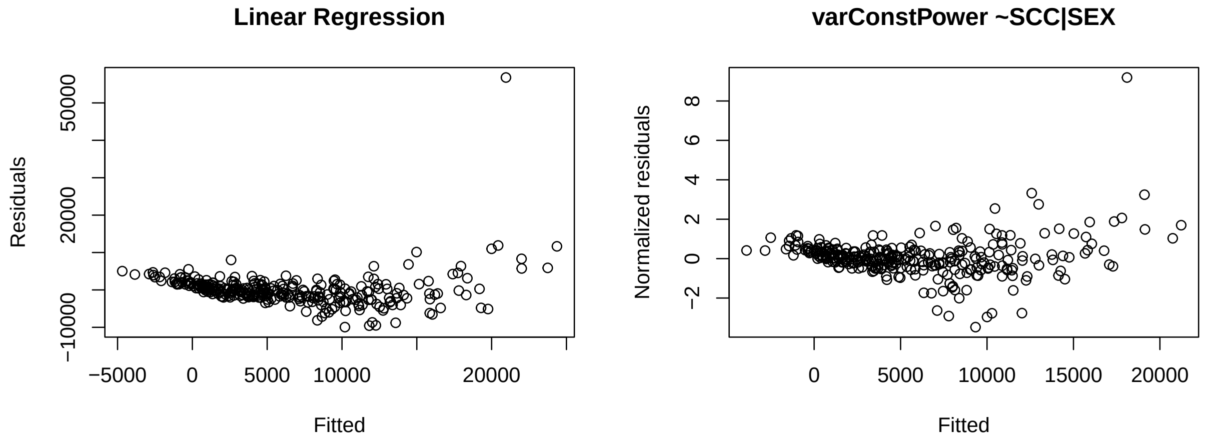


Figure 5: Comparison of the normalized residuals vs fitted for the best overall model before and after fitting the best non-constant variance structure

Results

From the estimated model, all the estimated parameters are significant with p-values < 0.05. All confidence intervals in this report are 95% confidence intervals and are calculated as $CI = \hat{y_i} \pm 1.96 \times \hat{SE_i}$. Note: we reverse-transformed (square-root) the parameter values back into the original scale of the response. For the average patient: 29 years old, with SCC 1.1, no high blood pressure, and a BMI of 28, the estimated baseline CC is 78.5 with a lower bound of 75.8 and an upper bound of 81.0. For an increase of 1 year in age, holding all other variables constant, we estimate the CC to change by -22.5 with a lower bound of -21.3 and an upper bound of -23.7. That is, as the patient ages their estimated CC decreases. For a 0.1 unit increase in SCC, holding all other variables constant, we estimate the CC to change by 11.4 with a lower bound of 10.6 and an upper bound of 12.2. That is, as the patient increases SCC their estimated CC also increases. But there is also an interaction between age and SCC that estimates a net negative effect as these two variables increase. For a patient with high blood pressure, holding all other variables constant, we estimate CC to change by -44.5 with a lower bound of -33.1 and an upper bound of 53.4. That is, having high blood pressure decreases the estimated CC. And for a 1 unit increase in BMI, holding all other variables constant, we estimate the CC to change by 16.9 with a lower bound of 11.6 and an upper bound of 21.0. That is, a higher BMI also increases the estimated CC. See Table 1 for all the parameter estimates.

parameter	value	se
Intercept	6155.4199	209.039
AGE	-508.1274	27.181
SCC	12976.9485	936.667
BLOOD	-1975.8560	449.385
BMI	286.5215	77.904
AGE:SCC	-1035.4328	135.482

Table 1: Summary of the estimated parameters from the best model. Note: the parameters are on the transformed scale of CC.SQ and for the centered data

Now we use the estimated model to make predictions. For a new male patient, 32 years old, has BMI 33, SCC measurement 1.2, and has high blood pressure, we estimate his CC to be 64.8 with a lower bound of 61.2 and an upper bound of 68.3.

We also examine whether or not there is an important relationship with the gender: female and male patients. That is, whether the relevant variables have roughly the same effects on female and male patients. To determine this, we include the sex variable in the model (not previously included in the best model). The estimated parameter for sex does not have a significant effect in the best model, with a high p-value of 0.85. Therefore we can conclude that there is not strong evidence against the null model that there is no difference between sex for this model.

Hypothetically, if the researchers actually collected multiple measurements for each of the 300 patients at different time points, then the assumption of independence is clearly violated. To still create a valid model, we could include the ID as a covariate in the model. It would make the most sense to include it as a random effect. With 300 patients, a random effect would only estimate one extra parameter for the model. In addition, we could use this mixed effect model to make predictions for patients outside of the original study.

In a follow up study, the researcher recorded whether or not the patient has developed some kidney disease (including death) within the next year. The theoretical model for predicting DISEASE becomes a logistic regression model:

$$
\begin{aligned}
y_i &\sim Bernoulli(p_i) \\
\eta_i &= \overline{\beta}\,\overline{\mathbf{X}} \\
logit(p_i) &= \eta_i
\end{aligned}
$$

Using the new data DISEASE and the same significant variables in the model we obtained previously, we can answer how likely the new patient from above will develop some kidney disease (including death) within the next year. To calculate this, we need to calculate the inverse logit of the prediction. $\hat{p}_i = logit^{-1}(\hat{y}_i) = \frac{\exp(\hat{y}_i)}{1+\exp(\hat{y}_i)}$. The resulting probability of the patient developing some kidney disease (including death) is 32.1% with a lower bound of 11.3% and an upper bound of 63.6%. If the same patient has a BMI=30 while keeping other variables unchanged, then the probability becomes 41.7% with a lower bound of 20.5% and an upper bound of 66.5%.

Appendix

```r
knitr::opts_chunk$set(echo = T, warning=F, message=F, fig.height=4, fig.width=10)
options(show.signif.stars = FALSE)
library(tidyverse)
library(lme4)

library(GGally)
library(ggplot2)
kidney <- read.table("kidney.dat", header=T)
kidney$SEX <- factor(kidney$SEX)
kidney$BLOOD <- factor(kidney$BLOOD)
kidney$DISEASE <- factor(kidney$DISEASE)

#ggpairs(kidney[c(3:5,2)])

library(gridExtra)
kidney$CC.SQ <- kidney$CC^2

p1 <- ggplot(data=kidney, aes(x=BMI, y=CC)) +
  geom_point() +
  geom_smooth(method="loess") +
  ggtitle("BMI vs CC")

p2 <- ggplot(data=kidney, aes(x=BMI, y=CC.SQ)) +
  geom_point() +
  geom_smooth(method="loess") +
  ggtitle("BMI vs CC.SQ")

grid.arrange(p1, p2, nrow=1)

ggpairs(kidney[c(3:5,9)])

# forward selection CC
# 1-variable models
m1a <- lm(CC ~ AGE, data=kidney)
m1b <- lm(CC ~ BMI, data=kidney)
m1c <- lm(CC ~ SCC, data=kidney)
m1d <- lm(CC ~ SEX, data=kidney)
m1e <- lm(CC ~ BLOOD, data=kidney)
BIC(m1a,m1b,m1c,m1d,m1e)
# best 1-variable model
m1 <- m1a
# 2-variable models
```

```
m2a <- lm(CC ~ AGE + BMI, data=kidney)
m2b <- lm(CC ~ AGE + SCC, data=kidney)
m2c <- lm(CC ~ AGE + SEX, data=kidney)
m2d <- lm(CC ~ AGE + BLOOD, data=kidney)
# 2-variable models (interactions)
m2e <- lm(CC ~ AGE * BMI, data=kidney)
m2f <- lm(CC ~ AGE * SCC, data=kidney)
m2g <- lm(CC ~ AGE * SEX, data=kidney)
m2h <- lm(CC ~ AGE * BLOOD, data=kidney)
BIC(m2a,m2b,m2c,m2d,m2e,m2f,m2g,m2h)
# best 2-variable model
m2 <- m2f
# 3-variable models
m3a <- lm(CC ~ AGE * SCC + BMI, data=kidney)
m3b <- lm(CC ~ AGE * SCC + SEX, data=kidney)
m3c <- lm(CC ~ AGE * SCC + BLOOD, data=kidney)
m3d <- lm(CC ~ AGE * SCC * BMI, data=kidney)
m3e <- lm(CC ~ AGE * SCC * SEX, data=kidney)
m3f <- lm(CC ~ AGE * SCC * BLOOD, data=kidney)
BIC(m3a,m3b,m3c,m3d,m3e,m3f)
# best 3-variable model
m3 <- m3c
# 4-variable models
m4a<- lm(CC ~ AGE * SCC + BLOOD + BMI, data=kidney)
m4b<- lm(CC ~ AGE * SCC + BLOOD + SEX, data=kidney)
m4c<- lm(CC ~ AGE * SCC + BLOOD * BMI, data=kidney)
m4d<- lm(CC ~ AGE * SCC + BLOOD * SEX, data=kidney)
BIC(m4a, m4b,m4c, m4d)
# best 4-variable model
m4 <- m4a
# 5-variable model
m5a <- lm(CC ~ AGE * SCC + BLOOD + BMI + SEX, data=kidney)
m5b <- lm(CC ~ AGE * SCC + BLOOD + BMI * SEX, data=kidney)
BIC(m5a, m5b)
# best 5-variable model
m5 <- m5a
# compare all best models
BIC(m1,m2,m3,m4,m5)
# overall best model from forward selection
mbest <- m4

# forward selection CC.SQ
# 1-variable models
m1a <- lm(CC.SQ ~ AGE, data=kidney)
m1b <- lm(CC.SQ ~ BMI, data=kidney)
m1c <- lm(CC.SQ ~ SCC, data=kidney)
m1d <- lm(CC.SQ ~ SEX, data=kidney)
m1e <- lm(CC.SQ ~ BLOOD, data=kidney)
#BIC(m1a,m1b,m1c,m1d,m1e)
# best 1-variable model
m1 <- m1a
# 2-variable models
m2a <- lm(CC.SQ ~ AGE + BMI, data=kidney)
```

```r
m2b <- lm(CC.SQ ~ AGE + SCC, data=kidney)
m2c <- lm(CC.SQ ~ AGE + SEX, data=kidney)
m2d <- lm(CC.SQ ~ AGE + BLOOD, data=kidney)
# 2-variable models (interactions)
m2e <- lm(CC.SQ ~ AGE * BMI, data=kidney)
m2f <- lm(CC.SQ ~ AGE * SCC, data=kidney)
m2g <- lm(CC.SQ ~ AGE * SEX, data=kidney)
m2h <- lm(CC.SQ ~ AGE * BLOOD, data=kidney)
#BIC(m2a,m2b,m2c,m2d,m2e,m2f,m2g,m2h)
# best 2-variable model
m2 <- m2f
# 3-variable models
m3a <- lm(CC.SQ ~ AGE * SCC + BMI, data=kidney)
m3b <- lm(CC.SQ ~ AGE * SCC + SEX, data=kidney)
m3c <- lm(CC.SQ ~ AGE * SCC + BLOOD, data=kidney)
m3d <- lm(CC.SQ ~ AGE * SCC * BMI, data=kidney)
m3e <- lm(CC.SQ ~ AGE * SCC * SEX, data=kidney)
m3f <- lm(CC.SQ ~ AGE * SCC * BLOOD, data=kidney)
#BIC(m3a,m3b,m3c,m3d,m3e,m3f)
# best 3-variable model
m3 <- m3c
# 4-variable models
m4a<- lm(CC.SQ ~ AGE * SCC + BLOOD + BMI, data=kidney)
m4b<- lm(CC.SQ ~ AGE * SCC + BLOOD + SEX, data=kidney)
m4c<- lm(CC.SQ ~ AGE * SCC + BLOOD * BMI, data=kidney)
m4d<- lm(CC.SQ ~ AGE * SCC + BLOOD * SEX, data=kidney)
#BIC(m4a, m4b,m4c, m4d)
# best 4-variable model
m4 <- m4a
# 5-variable model
m5a <- lm(CC.SQ ~ AGE * SCC + BLOOD + BMI + SEX, data=kidney)
m5b <- lm(CC.SQ ~ AGE * SCC + BLOOD + BMI * SEX, data=kidney)
#BIC(m5a, m5b)
# best 5-variable model
m5 <- m5a
# compare all best models
#BIC(m1,m2,m3,m4,m5)
# overall best model from forward selection
mbest <- m4

# plot interaction effects
library(effects)
plot(allEffects(mbest, residuals=T)[1:2],
     residuals.pch=1,
     residuals.cex=0.75,
     lwd=2, grid=T, rug=T)

plot(allEffects(mbest, residuals=T)[3],
     residuals.pch=1,
     residuals.cex=0.75,
     lwd=2, grid=T, rug=T)

# diagnostic plots for the best model
```

```
#summary(mbest)
par(mfrow=c(2,2))
plot(mbest)

# non-constant variance models
par(mfrow=c(1,2))
library(nlme)
formula = CC.SQ ~ AGE * SCC + BLOOD + BMI
# varFixed
mbestVFa <- gls(formula, data=kidney, weights=varFixed(~SCC))
mbestVFb <- gls(formula, data=kidney, weights=varFixed(~BMI))
#plot(resid(mbestVFa, type = "normalized") ~ fitted(mbestVFa))
#plot(resid(mbestVFb, type = "normalized") ~ fitted(mbestVFb))

# varIdent
#boxplot(resid(mbest) ~ kidney$SEX, varwidth = T)
#boxplot(resid(mbest) ~ kidney$BLOOD, varwidth = T)
mbestVIa<-gls(formula, data=kidney, weights=varIdent(form = ~1|SEX))
mbestVIb<-gls(formula, data=kidney, weights=varIdent(form = ~1|BLOOD))
#plot(resid(mbestVIa, type = "normalized") ~ fitted(mbestVIa))
#plot(resid(mbestVIb, type = "normalized") ~ fitted(mbestVIb))

# varPower
mbestVPa <- gls(formula, data=kidney, weights=varPower(form=~SCC))
mbestVPb <- gls(formula, data=kidney, weights=varPower(form=~BMI))
#plot(resid(mbestVPa, type = "normalized") ~ fitted(mbestVPa))
#plot(resid(mbestVPb, type = "normalized") ~ fitted(mbestVPb))
# varPower interaction
mbestVP2a <- gls(formula, data=kidney, weights=varPower(form=~SCC|SEX))
mbestVP2b <- gls(formula, data=kidney, weights=varPower(form=~SCC|BLOOD))
mbestVP2c <- gls(formula, data=kidney, weights=varPower(form=~BMI|SEX))
mbestVP2d <- gls(formula, data=kidney, weights=varPower(form=~BMI|BLOOD))
#plot(resid(mbestVP2a, type = "normalized") ~ fitted(mbestVP2a))
#plot(resid(mbestVP2b, type = "normalized") ~ fitted(mbestVP2b))
#plot(resid(mbestVP2c, type = "normalized") ~ fitted(mbestVP2c))
#plot(resid(mbestVP2d, type = "normalized") ~ fitted(mbestVP2d))

# varExp
mbestVEa <- gls(formula, data=kidney, weights=varExp(form=~SCC))
mbestVEb <- gls(formula, data=kidney, weights=varExp(form=~BMI))
#plot(resid(mbestVEa, type = "normalized") ~ fitted(mbestVEa))
#plot(resid(mbestVEb, type = "normalized") ~ fitted(mbestVEb))
# varExp interaction
mbestVE2a <- gls(formula, data=kidney, weights=varExp(form=~SCC|SEX))
mbestVE2b <- gls(formula, data=kidney, weights=varExp(form=~SCC|BLOOD))
mbestVE2c <- gls(formula, data=kidney, weights=varExp(form=~BMI|SEX))
mbestVE2d <- gls(formula, data=kidney, weights=varExp(form=~BMI|BLOOD))
#plot(resid(mbestVE2a, type = "normalized") ~ fitted(mbestVE2a))
#plot(resid(mbestVE2b, type = "normalized") ~ fitted(mbestVE2b))
#plot(resid(mbestVE2c, type = "normalized") ~ fitted(mbestVE2c))
#plot(resid(mbestVE2d, type = "normalized") ~ fitted(mbestVE2d))

# varConstPower
```

```
mbestVCPa <- gls(formula, data=kidney, weights=varConstPower(form=~SCC))
mbestVCPb <- gls(formula, data=kidney, weights=varConstPower(form=~BMI))
#plot(resid(mbestVCPa, type = "normalized") ~ fitted(mbestVCPa))
#plot(resid(mbestVCPb, type = "normalized") ~ fitted(mbestVCPb))
# varConstPower interaction
mbestVCP2a <- gls(formula, data=kidney, weights=varConstPower(form=~SCC|SEX))
mbestVCP2b <- gls(formula, data=kidney, weights=varConstPower(form=~SCC|BLOOD))
mbestVCP2c <- gls(formula, data=kidney, weights=varConstPower(form=~BMI|SEX))
mbestVCP2d <- gls(formula, data=kidney, weights=varConstPower(form=~BMI|BLOOD))
#plot(resid(mbestVCP2a, type = "normalized") ~ fitted(mbestVCP2a))
#plot(resid(mbestVCP2b, type = "normalized") ~ fitted(mbestVCP2b))
#plot(resid(mbestVCP2c, type = "normalized") ~ fitted(mbestVCP2c))
#plot(resid(mbestVCP2d, type = "normalized") ~ fitted(mbestVCP2d))

# compare non-constant variance models
#BIC(mbestVFa, mbestVFb, mbestVIa, mbestVIb, mbestVPa, mbestVPb, mbestVP2a, mbestVP2b, mbestVP2c, mbest

# the best best model
mbestbest <- mbestVCP2a
# mean-center the numeric columns for easier interpretation
kidney$AGE.SCALE <- kidney$AGE - mean(kidney$AGE)
kidney$BMI.SCALE <- kidney$BMI - mean(kidney$BMI)
kidney$SCC.SCALE <- kidney$SCC - mean(kidney$SCC)
mbestbestscale <- gls(CC.SQ ~ AGE.SCALE * SCC.SCALE + BLOOD + BMI.SCALE, data=kidney,
                      weights=varConstPower(form=~SCC.SCALE|SEX))

par(mfrow=c(1,2))
plot(resid(mbest) ~ fitted(mbest),
     main="Linear Regression", ylab="Residuals", xlab="Fitted")
plot(resid(mbestbestscale, type = "normalized") ~ fitted(mbestbestscale),
     main="varConstPower ~SCC|SEX", ylab="Normalized residuals", xlab="Fitted")

library(knitr)
mbestbestsummary <- summary(mbestbestscale)
df <- data.frame(parameter=c("Intercept","AGE","SCC","BLOOD","BMI","AGE:SCC"),
                 value=mbestbestsummary$coefficients,
                 se=c(209.039,27.181,936.667,449.385,77.904,135.482))
rownames(df)=NULL

kable(df)

library(AICcmodavg)
new.patient <-  data.frame(SEX=1, AGE=32, BMI=33, SCC=1.2, BLOOD=factor(1))
pred <- predictSE.gls(mbestbest, newdata=new.patient, se.fit=T)
lwr <- sqrt(pred$fit-1.96*pred$se.fit)
upr <- sqrt(pred$fit+1.96*pred$se.fit)

mbest.test.sex <- gls(CC.SQ ~ AGE * SCC + BLOOD + BMI + SEX, data=kidney, weights=varConstPower(form=~SC
#summary(mbest.test.sex)

mlogit <- glm(DISEASE ~ AGE * SCC + BLOOD + BMI, family="binomial", data=kidney)

predlogit <- predict(mlogit, newdata=new.patient, se.fit=T)
```

```
lwrlogit <- predlogit$fit-1.96*predlogit$se.fit
uprlogit <- predlogit$fit+1.96*predlogit$se.fit
p <- exp(predlogit$fit)/(1+exp(predlogit$fit))
lwrp <- exp(lwrlogit)/(1+exp(lwrlogit))
uprp <- exp(uprlogit)/(1+exp(uprlogit))

new.patient2 <- data.frame(SEX=1, AGE=32, BMI=30, SCC=1.2, BLOOD=factor(1))
predlogit2 <- predict(mlogit, newdata=new.patient2, se.fit=T)
lwrlogit2 <- predlogit2$fit-1.96*predlogit2$se.fit
uprlogit2 <- predlogit2$fit+1.96*predlogit2$se.fit
p2 <- exp(predlogit2$fit)/(1+exp(predlogit2$fit))
lwrp2 <- exp(lwrlogit2)/(1+exp(lwrlogit2))
uprp2 <- exp(uprlogit2)/(1+exp(uprlogit2))
```