STAT 3690 Lecture 19

zhiyanggeezhou.github.io

Zhiyang Zhou (zhiyang.zhou@umanitoba.ca)

Mar 14, 2022

Multivariate influence measures

- Hat/projection matrix $\mathbf{H} = [h_{ij}]_{n\times n} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$
  - $|h_{ij}| \le 1$
- $\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$
  - the ith row of $\hat{\mathbf{Y}}$: $\hat{\mathbf{Y}}_{i\cdot} = \sum_{j=1} h_{ij}\mathbf{Y}_{j\cdot} = h_{ii}\mathbf{Y}_i + \sum_{j\neq i} h_{ij}\mathbf{Y}_{j\cdot}$
- Leverage: the influence of observation $\mathbf{Y}_{i\cdot}$ on $\hat{\mathbf{Y}}_{i\cdot}$
  - Observation $\mathbf{Y}_{i\cdot}$ is said to have a high leverage if h_{ii} is large compared to the other elements on the diagonal of $\mathbf{H}$.
- (Externally) Studentized residuals

$$T_i^2 = \frac{\hat{\mathbf{E}}_{i\cdot}^\top \mathbf{\Sigma}_{\mathrm{LS},(i)}^{-1}\hat{\mathbf{E}}_{i\cdot}}{1-h_{ii}}$$

  - $\hat{\mathbf{E}}_{i\cdot}^\top$: the ith row of $\hat{\mathbf{E}} = (\mathbf{I}-\mathbf{H})\mathbf{Y}$
  - $\hat{\mathbf{E}}_{(i)\cdot}^\top$: remaining part of $\hat{\mathbf{E}}$ with row i removed
  - $\mathbf{\Sigma}_{\mathrm{LS},(i)} = (n-q-2)^{-1}\hat{\mathbf{E}}_{(i)\cdot}^\top\hat{\mathbf{E}}_{(i)\cdot}$: LS estimator of $\mathbf{\Sigma}$ where we have removed row i from the residual matrix
  - Observation $\mathbf{Y}_i$ may be considered as a potential outlier if

$$T_i^2 > \frac{p(n-q-2)}{n-p-q-1}F_{1-\alpha,p,n-q-2}$$

    - $F_{1-\alpha,p,n-q-2}$: the $1-\alpha$ quantile of $F(p, n-q-2)$
- (Multivariate) Cook's distance

$$D_i = \frac{h_{ii}}{(1-h_{ii})^2(q+1)}\hat{\mathbf{E}}_{i\cdot}^\top\mathbf{\Sigma}_{\mathrm{LS}}^{-1}\hat{\mathbf{E}}_{i\cdot}$$

  - The Cut-off is far from unique even for univariate linear regression ($p = 1$)
  - Pay attention to a small set of observations that has substantially higher values than the remaining observations

---

```
install.packages(c("car","EnvStats"))
options(digits = 4)
tear <- c(
  6.5, 6.2, 5.8, 6.5, 6.5, 6.9, 7.2, 6.9, 6.1, 6.3,
  6.7, 6.6, 7.2, 7.1, 6.8, 7.1, 7.0, 7.2, 7.5, 7.6
)
gloss <- c(
  9.5, 9.9, 9.6, 9.6, 9.2, 9.1, 10.0, 9.9, 9.5, 9.4,
  9.1, 9.3, 8.3, 8.4, 8.5, 9.2, 8.8, 9.7, 10.1, 9.2
```

```
)
opacity <- c(
  4.4, 6.4, 3.0, 4.1, 0.8, 5.7, 2.0, 3.9, 1.9, 5.7,
  2.8, 4.1, 3.8, 1.6, 3.4, 8.4, 5.2, 6.9, 2.7, 1.9
)
n = length(opacity)
rate <- factor(gl(2,10,length=n), labels=c("Low", "High"))
additive <- factor(gl(2,5,length=n), labels=c("Low", "High"))
(plastic = data.frame(tear=tear, gloss=gloss, opacity=opacity,
                      rate=rate, additive=additive))

fit0 <- lm(cbind(tear, gloss, opacity) ~ rate*additive, data = plastic)
resids <- residuals(fit0)

# Leverage
X <- model.matrix(fit0)
H <- X %*% solve(crossprod(X)) %*% t(X)
Hii = diag(H)
hist(Hii, 50)

# Externally Studentized residuals
n <- nrow(X)
p = ncol(resids)
T_square = numeric(n)
for (i in 1:n){
  SigmaHatLS_i <- crossprod(resids[-i,])/(n-1-ncol(X))
  T_square[i] = t(resids[i,]) %*% solve(SigmaHatLS_i) %*% resids[i,]
}
hist(T_square, 50)
which(T_square > p*(n-1-ncol(X))/(n-p-ncol(X))*qchisq(.95, p, n-1-ncol(X)))

# Cook distance
SigmaHatLS <- crossprod(resids)/(n - ncol(X))
cook_values <- Hii/((1 - Hii)^2*ncol(X)) * diag(resids %*% solve(SigmaHatLS) %*% t(resids))
hist(cook_values, 50)
which(cook_values>0.4)
```

Normality of residuals

- Apply techniques in Lecture 7 to checking the normality of residuals
- Apply Box-Cox transformation to colums of $\mathbf{Y}$

```
install.packages(c("car","EnvStats"))
options(digits = 4)
tear <- c(
  6.5, 6.2, 5.8, 6.5, 6.5, 6.9, 7.2, 6.9, 6.1, 6.3,
  6.7, 6.6, 7.2, 7.1, 6.8, 7.1, 7.0, 7.2, 7.5, 7.6
)
gloss <- c(
  9.5, 9.9, 9.6, 9.6, 9.2, 9.1, 10.0, 9.9, 9.5, 9.4,
  9.1, 9.3, 8.3, 8.4, 8.5, 9.2, 8.8, 9.7, 10.1, 9.2
)
```

```r
opacity <- c(
  4.4, 6.4, 3.0, 4.1, 0.8, 5.7, 2.0, 3.9, 1.9, 5.7,
  2.8, 4.1, 3.8, 1.6, 3.4, 8.4, 5.2, 6.9, 2.7, 1.9
)
n = length(opacity)
rate <- factor(gl(2,10,length=n), labels=c("Low", "High"))
additive <- factor(gl(2,5,length=n), labels=c("Low", "High"))
(plastic = data.frame(tear=tear, gloss=gloss, opacity=opacity,
                     rate=rate, additive=additive))

fit0 <- lm(cbind(tear, gloss, opacity) ~ rate*additive, data = plastic)

# Normal Q-Q plots of residuals
res = residuals(fit0)
name = colnames(res)
op <- par(mfrow = c(2,2),
          oma = c(5,4,0,0),
          mar = c(1,1,2,2))
for (i in 1:ncol(res)){
  car::qqPlot(res[,i], main = name[i], id = F)
}
title(xlab = "Normal quantiles",
      ylab = "Sample quantiles",
      outer = TRUE, line = 3)
par(op)

# Box-Cox transformation
fit1 = lm(tear ~ rate*additive, data = plastic)
fit2 = lm(gloss ~ rate*additive, data = plastic)
fit3 = lm(opacity ~ rate*additive, data = plastic)

(lambda1 = EnvStats::boxcox(fit1 , optimize=T, lambda=c(-10,10))$lambda)
plastic$tear.new = (plastic$tear^lambda1-1)/lambda1
(lambda2 = EnvStats::boxcox(fit2 , optimize=T, lambda=c(-10,10))$lambda)
plastic$gloss.new = (plastic$gloss^lambda2-1)/lambda2
(lambda3 = EnvStats::boxcox(fit3 , optimize=T, lambda=c(-10,10))$lambda)
plastic$opacity.new = (plastic$opacity^lambda3-1)/lambda3

fit0.new <- lm(cbind(tear.new, gloss.new, opacity.new) ~ rate*additive, data = plastic)
```