343 HW 2

Madhuri Raman

10/12/2021

1.

```
library(MASS)
library(faraway)
set.seed(819)

N <- 1000

rho <- 0.95

mu1 <- 0; s1 <- 1
mu2 <- 0; s2 <- 1
mu <- c(mu1, mu2)
sigmaMat <- matrix(c(s1^2, s1*s2*rho, s1*s2*rho, s2^2), 2) # cov matrix

bvn1 <- mvrnorm(N, mu = mu, Sigma = sigmaMat)
colnames(bvn1) <- c("X1", "X2")

eps <- rnorm(N, 0, 1)

beta0 <- 0
beta1 <- 1
beta2 <- -1000

Y_m1 <- beta0 + beta1*bvn1[,"X1"] + eps
Y_m2 <- beta0 + beta1*bvn1[,"X1"] + beta2*bvn1[,"X2"] + eps

m1 <- lm(Y_m1 ~ bvn1[,"X1"])
m2 <- lm(Y_m2 ~ bvn1[,"X1"] + bvn1[,"X2"])
```

a)

We see that the coefficient on X1 is negative when modeled vs Y and then negative when X2 is included in the model. For example, this phenomenon occurs with parameters beta0 = 0, beta1 = 1, a highly negative beta2 = -1000, sigma squared of 1, and a high correlation rho between X and Y of 0.95.

b)

Weight vs Cardio + Calorie_Intake

There is a negative relationship between one's body weight (Y) and the amount of cardio exercise they do per week (X1). Doing more exercise typically results in lower body weight. However, when we also take into account calorie intake (X2), that is highly correlated with amount of cardio per week because as one does more and more cardio, they will have higher calorie intake to compensate. Because of this, there can now be a positive relationship between cardio and weight.

2.

```
df_prostate <- data.frame(prostate)
lm.out.1 <- lm(lpsa ~ ., data = df_prostate)
```

a)

```
# 90% CI for age parameter
confint(lm.out.1, "age", level = 0.90)
```

```
##             5 %         95 %
## age -0.0382102 -0.001064151
```

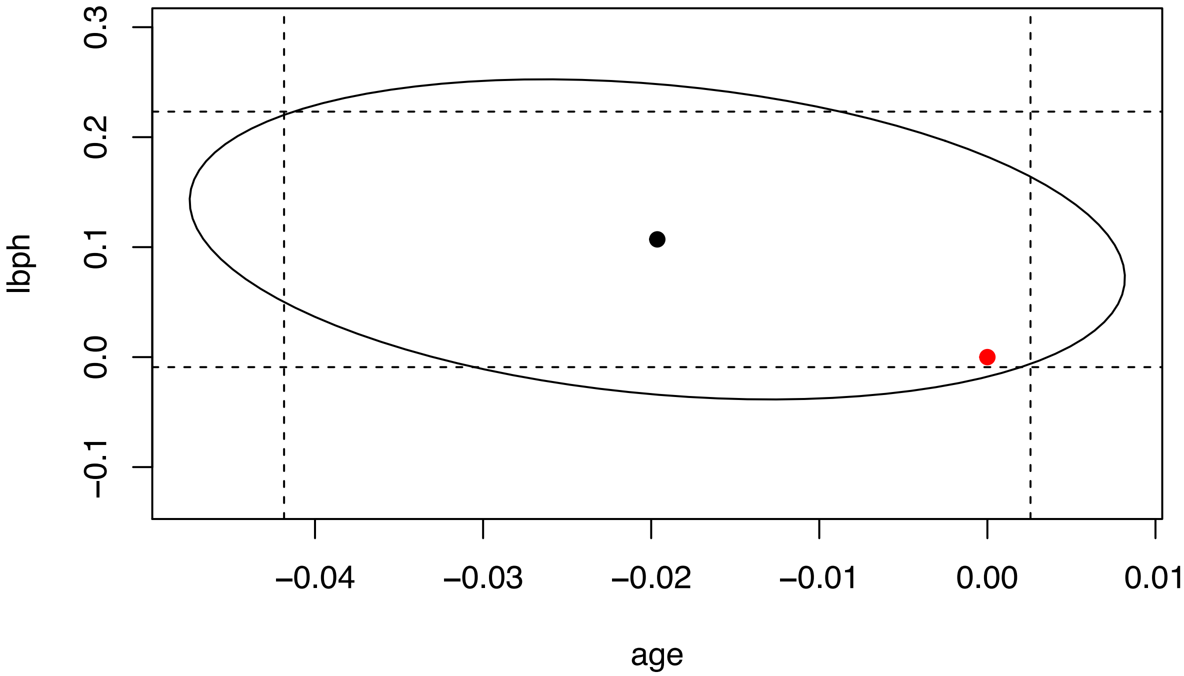
```
# 95% CI for age parameter
confint(lm.out.1, "age", level = 0.95)
```

```
##            2.5 %      97.5 %
## age -0.04184062 0.002566267
```

Based on these confidence intervals for the coefficient of `age`, we could deduce that its p-value in the regression summary will be greater than 0.05 but less than 0.10. This is because the 95% confidence interval for this coefficient includes 0 so the test is not significant at the 0.05 alpha level, but the 90% confidence interval does include 0 so the test would be significant at the 0.10 alpha level.

b)

```
library(ellipse)
plot(ellipse(lm.out.1, c(4,5)), type = "l", ylim = c(-0.13,0.3))
points(coef(lm.out.1)[4], coef(lm.out.1)[5], pch = 19)
points(0, 0, pch = 19, col = "red")
abline(v = confint(lm.out.1)[4,], lty = 2)
abline(h = confint(lm.out.1)[5,], lty = 2)
```



The location of the origin point (red) in the 95% confidence region plot above tells us the outcome of the following hypothesis test:

$H_0 : \beta_{age} = \beta_{lbph} = 0$

H_A : either of β_{age} or $\beta_{lbph} \neq 0$

Since the origin point lies inside the 95% confidence ellipse, we fail to reject the null hypothesis. There is not sufficient evidence to conclude that either of β_{age} or $\beta_{lbph} \neq 0$.

c)

```
tt <- summary(lm.out.1)$coef[4,3] # t-statistic for age in original model
nreps <- 4000
set.seed(819)
tstats <- numeric(nreps)
for (i in 1:nreps){
  lmods <- lm(lpsa ~ sample(age) + ., df_prostate)
  tstats[i] <- summary(lmods)$coef[2,3]
}
mean(abs(tstats) > abs(tt))
```

```
## [1] 0.08075
```

```
summary(lm.out.1)$coef[4,4]
```

```
## [1] 0.08229321
```

Note that the outcome of the permutation test corresponding to the t-test for age in this model (0.08075) is very similar to the observed normal-based p-value of 0.0823, so both methods agree.

d)

```r
summary(lm.out.1)$coef[,4]
```

```
##  (Intercept)       lcavol      lweight          age         lbph          svi 
## 6.069335e-01 2.110698e-09 8.955363e-03 8.229321e-02 7.039846e-02 2.328749e-03 
##          lcp      gleason        pgg45 
## 2.496377e-01 7.750328e-01 3.088604e-01
```

Note that `age`, `lbph`, `lcp`, `gleason`, `pgg45` are the predictors that are not significant at the 5% level, so we will now test a simpler model with these predictors removed against our original full model.

```r
lm.out.2 <- lm(lpsa ~ lcavol + lweight + svi, data = df_prostate)
anova(lm.out.2, lm.out.1)
```

```
## Analysis of Variance Table
## 
## Model 1: lpsa ~ lcavol + lweight + svi
## Model 2: lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason + 
##     pgg45
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     93 47.785                           
## 2     88 44.163  5    3.6218 1.4434 0.2167
```

Since the p-value of the F test is much larger than $\alpha = 0.05$, we fail to reject the null hypothesis at the 5% level. There is not enough evidence to conclude that the model including `age`, `lbph`, `lcp`, `gleason`, and `pgg45` as predictors is significantly better or different than the model without these predictors.

3.

```r
df_teengamb <- data.frame(teengamb)
lm.out.3 <- lm(gamble ~ ., data = df_teengamb)
sumary(lm.out.3)
```

```
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  22.555651  17.196803  1.3116   0.19677
## sex         -22.118330   8.211115 -2.6937   0.01011
## status        0.052234   0.281112  0.1858   0.85349
## income        4.961979   1.025392  4.8391 1.792e-05
## verbal       -2.959493   2.172150 -1.3625   0.18031
## 
## n = 47, p = 5, Residual SE = 22.69034, R-Squared = 0.53
```

a)

We see that of the four predictors, `sex` and `income` are the only ones that are significant at the 5% level.

b)

The coefficient of `sex` is about -22.12. Note that in this data set, `sex = 0` indicates males and `sex = 1` indicates females. So, this coefficient means that with all other variables in the model constant, gambling expenditure for females is about 22.12 pounds per year less than gambling expenditure for males.

c)

```r
lm.out.4 <- lm(gamble ~ income, data = df_teengamb)
anova(lm.out.4, lm.out.3)
```

```
## Analysis of Variance Table
##
## Model 1: gamble ~ income
## Model 2: gamble ~ sex + status + income + verbal
##   Res.Df   RSS Df Sum of Sq      F  Pr(>F)
## 1     45 28009
## 2     42 21624  3    6384.8 4.1338 0.01177 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p-value of this F test is less than $\alpha = 0.05$, we can reject the null hypothesis at the 5% level. There is sufficient evidence to conclude that the inclusion of `sex`, `status`, and `verbal` as additional predictors significantly improves the model fit compared to the model with only `income` as a predictor.

4.

$$\underset{n\times 1}{Y_i} = \underset{n\times p}{X_i^T}\underset{p\times 1}{\beta} + \underset{n\times 1}{\epsilon_i}$$

$$= \beta_1 X_{i1} + \beta_2 + X_{i2} + \dots \beta_p X_{ip} + \epsilon_i \quad \text{w/ } \epsilon_i \sim N(0,\sigma^2)$$

just assume no intercept term for now (usually β_0)

$\hat{\beta}, \hat{\sigma}^2$ usual least sq. estimates of β, σ^2

$$\left.\begin{aligned} &y^{(1)},\ \vec{X}^{(1)} = X_1^{(1)}, X_2^{(1)}, \dots X_p^{(1)} \\ &y^{(2)},\ \vec{X}^{(2)} = X_1^{(2)}, X_2^{(2)}, \dots X_p^{(2)} \\ &y^{(3)},\ \vec{X}^{(3)} = X_1^{(3)}, X_2^{(3)} \dots X_p^{(3)} \end{aligned}\right\} \text{3 new data points}$$

a) estimate sample mean of these 3 y's

estimate $\frac{1}{3}(y^{(1)} + y^{(2)} + y^{(3)})$:

$$\underset{1\times 1}{\hat{y}^{(1)}} = \underset{1\times p}{\vec{X}^{(1)T}} \cdot \underset{p\times 1}{\hat{\beta}}$$

$$\hat{y}^{(1)} = \hat{\beta}_1 \cdot X_1^{(1)} + \hat{\beta}_2 \cdot X_2^{(1)} + \dots \hat{\beta}_p \cdot X_p^{(1)} = \vec{X}^{(1)T} \cdot \hat{\beta}$$

Similarly,

$$\hat{y}^{(2)} = \hat{\beta}_1 \cdot X_1^{(2)} + \hat{\beta}_2 \cdot X_2^{(2)} + \dots \hat{\beta}_p \cdot X_p^{(2)} = \vec{X}^{(2)T} \cdot \hat{\beta}$$

$$\hat{y}^{(3)} = \hat{\beta}_1 \cdot X_1^{(3)} + \hat{\beta}_2 \cdot X_2^{(3)} + \dots \hat{\beta}_p \cdot X_p^{(3)} = \vec{X}^{(3)T} \cdot \hat{\beta}$$

$$\Rightarrow \frac{\underset{1\times p}{\vec{X}^{(1)T}}\underset{p\times 1}{\hat{\beta}} + \underset{1\times p}{\vec{X}^{(2)T}}\underset{p\times 1}{\hat{\beta}} + \underset{1\times p}{\vec{X}^{(3)T}}\underset{p\times 1}{\hat{\beta}}}{3}$$

$$= \frac{\left(\underset{1\times p}{\vec{X}^{(1)T}} + \underset{1\times p}{\vec{X}^{(2)T}} + \underset{1\times p}{\vec{X}^{(3)T}}\right)\underset{p\times 1}{\hat{\beta}}}{3}$$

$$= \frac{\left(\overset{p\times 1}{\vec{X}^{(1)}} + \overset{p\times 1}{\vec{X}^{(2)}} + \overset{p\times 1}{\vec{X}^{(3)}}\right)^T \hat{\beta}}{3} = \boxed{\frac{1}{3}\left(X^{(1)} + X^{(2)} + X^{(3)}\right)^T \cdot \hat{\beta}}$$

Prediction interval for $\dfrac{y^{(1)}+y^{(2)}+y^{(3)}}{3}$

b) with coverage level $1-\alpha$:

"y" $= \dfrac{y_1+y_2+y_3}{3}$

$\hat{y} = \dfrac{(\vec{x}_1+\vec{x}_2+\vec{x}_3)^T}{3} \cdot \hat{\beta}$

each $\vec{x} = 1 \times p$

$Var(y-\hat{y})$

$= Var(y) + var(\hat{y})$

$= Var\left[\dfrac{y_1+y_2+y_3}{3}\right] + Var\left[\dfrac{(\vec{x}_1+\vec{x}_2+\vec{x}_3)^T}{3} \cdot \hat{\beta}\right]$

$= \dfrac{1}{9} var\left[\sum_{i=1}^{3} y_i\right] + \sigma^2 \cdot \dfrac{(\vec{x}_1+\vec{x}_2+\vec{x}_3)^T}{3} \cdot (X^TX)^{-1} \cdot \dfrac{(\vec{x}_1+\vec{x}_2+\vec{x}_3)}{3}$

Let $\dfrac{(\vec{x}_1+\vec{x}_2+\vec{x}_3)}{3} = \bar{x}$

$\Rightarrow \dfrac{1}{9} \cdot 3 \cdot \sigma^2 + \sigma^2 \cdot \bar{x}^T \cdot (X^TX)^{-1} \cdot \bar{x}$

$\Rightarrow \dfrac{\sigma^2}{3} + \sigma^2 \cdot \bar{x}^T (X^TX)^{-1} \bar{x}$

$\Rightarrow \sigma^2 \left[\dfrac{1}{3} + \bar{x}^T (X^TX)^{-1} \bar{x}\right]$

so $(1-\alpha)\%$ Prediction Interval:

$$\boxed{\bar{x}\hat{\beta} \pm t^*_{1-\alpha/2,\, n-p} \cdot \sigma \sqrt{\dfrac{1}{3} + \bar{x}^T (X^TX)^{-1} \bar{x}}}$$

5. $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

Assume X_{i1} and X_{i2} are highly correlated.
Assume they are basically = for every i.
New data pt $i = n+1$, with $X_{i1} = 0$ but X_{i2} large.

a) This point is potentially more useful for analyzing the data compared to a more typical point with highly correlated X_{i1} & X_{i2} because it does not follow the general trend of the rest of the data. Thus, as a leverage point, it can potentially have large influence on the slope of the regression line if we do include the point in our model.

b) If this new unusual data point is actually an error in data entry, then that will likely be very problematic. As previously mentioned, this data point is a high leverage point, so if its value in the hat matrix of the regression model (H_{ii}) is large, that means our regression line is overfitting to this point, which we really don't want since the point is an error. A more typical point like the original data with highly correlated X_{i1}, X_{i2} would be very close to the general trend and regression line anyways, so its addition or removal would not have as large an effect on the slope/coefficients as an unusual data point.