Statistical Modeling 1
Problem set 2 2024

1. To all karate enthusiasts, it would be nice to find such a punching board (makiwara board) that will withstand the blows but which would not be so rigid or hard that training would then harm hands. The makiwara board can be made in different kinds of wood. In study, it was examined how much a makiwara board bends (in millimeters) of the force of the strike in different tree species. The makiwara boards used in study were made in two different ways. Dataset is given in file makiwaraboard.txt.

```
    WoodType BoardType Deflection
1     Cherry   Stacked      144.3
2     Cherry   Stacked      125.9
3     Cherry   Stacked      263.2
4     Cherry   Stacked      114.6
.
.
335      Oak   Tapered       73.3
336      Oak   Tapered       44.9

Description: Results of experiments measuring deflection (mm) of
makiwara boards of two types (stacked and tapered) and of four
wood types (Cherry, Ash, Fir, and Oak).

Source: P.K. Smith, T. Niiler, and P.W. McCullough (2010). "Evaluating
Makiwara Punching Board Performance," Journal of Asian Martial Arts,
Vol 19, #2, pp. 34-45.
```

Denote explanatory variables as X_1=WoodType and X_2=BoardType. Consider modeling the response variable Y=Deflection by the model

$$\begin{aligned} \mathcal{M}_{1|2}: \quad Y_i &\sim N(\mu_{jh}, \sigma^2), \\ \mu_{jh} &= \beta_0 + \beta_j + \alpha_h, \end{aligned}$$

where index j is related to the categories of the variable X_1=WoodType and index h is related to the categories of the variable X_2=BoardType.

(a) Test the hypotheses

$$\begin{aligned} H_0 &: \beta_2 = 0 \text{ and } \beta_4 = 0, \\ H_1 &: \beta_2 \neq 0 \text{ or } \beta_4 \neq 0. \end{aligned}$$

Use the Wald statistic

$$\begin{aligned} W &= \frac{(\mathbf{K}'\hat{\boldsymbol{\beta}})'(\tilde{\sigma}^2\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K})^{-1}\mathbf{K}'\hat{\boldsymbol{\beta}}}{q} \\ &= \frac{(\mathbf{K}'\hat{\boldsymbol{\beta}})'\left(\widetilde{\mathrm{Cov}}(\mathbf{K}'\hat{\boldsymbol{\beta}})\right)^{-1}\mathbf{K}'\hat{\boldsymbol{\beta}}}{q} \sim F_{q,n-(p+1)}, \end{aligned}$$

to test the above hypotheses. Construct appropriate matrix $\mathbf{K}$ and then calculate the value of the test statistic. Return the value of the Wald statistic as your response.

(2 points)

(b) Under the model $\mathcal{M}_{1|2}$, consider the predictive effect size $Y_{2f} - Y_{1f}$ in situation where the explanatory variables are changed from the values

$$X_1 = \mathsf{Cherry},$$
$$X_2 = \mathsf{Stacked}.$$

to the values

$$X_1 = \mathsf{Oak},$$
$$X_2 = \mathsf{Tapered}.$$

Test the hypotheses

$$H_0 : y_{1f} = y_{2f},$$
$$H_1 : y_{1f} \neq y_{2f}.$$

Report the so-called d-value as your result.

(2 points)

(c) Make yourself familiar with the R code of the Assignment 2.4. Then test the pairwise average difference

$$H_0 : \left(\frac{\mu_{11} + \mu_{21} + \mu_{31} + \mu_{41}}{4}\right) - \left(\frac{\mu_{12} + \mu_{22} + \mu_{32} + \mu_{42}}{4}\right) = 0,$$

i.e., test whether there is average difference on means in the levels of $X_2 = \mathsf{BoardType}$ variable. Particularly, report the Wald test statistic value obtained from the comparison of Stacked and Tapered treatments.

(2 points)

2. In biodiesel study, methyl ester was produced from waste canola oil. In experiments, it was measured what kind of effect the factors $X_1 =$ Time (15,30,45min), $X_2 =$ Temperature (240,255,270C), and level of Methanol/Oil weight ratio (1,1.5,2), $X_3 =$ Methanol, have on yield of methyl ester, $Y =$ Yield. Data obtained from experiments is available in a file canoladiesel.txt.

	Time	Temp	Methanol	Yield
1	15	240	1.0	1.5
2	15	240	1.5	3.2
3	15	240	2.0	3.8
4	15	255	1.0	2.2
5	15	270	1.0	8.9
6	15	270	2.0	13.6
7	30	240	1.0	1.5
8	30	240	2.0	12.3
9	30	255	1.5	11.4
10	30	255	1.5	13.6
11	30	255	1.5	12.7
12	30	255	2.0	18.5
13	30	270	1.5	60.9
14	45	240	1.0	4.4
15	45	240	1.5	16.5
16	45	240	2.0	24.5
17	45	255	1.5	62.8
18	45	270	1.0	96.4
19	45	270	2.0	102.0

Source: S. Lee, D. Posarac, N. Ellis (2012). "An Experimental Investigation of Biodiesel Synthesis from Waste Canola Oil Using Supercritical Methanol," Fuel, Vol. 91, pp. 229-237.

(a) Let us assume $Y_i \sim N(\mu_i, \sigma^2)$. Model the expected value μ_i of the response variable $Y =$ Yield by the following model

$$\mathcal{M}_{1|2|3}: \quad Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i.$$

Consider the predictive effect size $Y_{2f} - Y_{1f}$ in situation where the explanatory variables are changed from the values

$$\begin{aligned} x_{1f} &= 15, \\ x_{2f} &= 240, \\ x_{3f} &= 1, \end{aligned}$$

to the values

$$\begin{aligned} x_{1f} &= 45, \\ x_{2f} &= 270, \\ x_{3f} &= 2, \end{aligned}$$

Test the hypotheses

$$\begin{aligned} H_0 &: y_{1f} = y_{2f}, \\ H_1 &: y_{1f} \neq y_{2f}. \end{aligned}$$

Select the appropriate test statistic to test the above hypotheses. Calculate the value of the test statistic.

(2 points)

(b) Next, model the expected value μ_i of the response variable $Y = \mathsf{Yield}$ by the following model

$$\mathcal{M}_{123}: \quad Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i1} x_{i2} + \beta_5 x_{i1} x_{i3} + \beta_6 x_{i2} x_{i3} + \beta_7 x_{i1} x_{i2} x_{i3} + \varepsilon_i.$$

Calculate the maximum likelihood estimate for the parameter β_7.

(1 point)

(c) Continue to model the expected value μ_i of the response variable $Y = \mathsf{Yield}$ by the following model

$$\mathcal{M}_{123}: \quad Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i1} x_{i2} + \beta_5 x_{i1} x_{i3} + \beta_6 x_{i2} x_{i3} + \beta_7 x_{i1} x_{i2} x_{i3} + \varepsilon_i.$$

Test the hypotheses

$$\begin{aligned} H_0 &: \beta_4 = 0 \text{ and } \beta_5 = 0 \text{ and } \beta_6 = 0 \text{ and } \beta_7 = 0, \\ H_1 &: \beta_4 \neq 0 \text{ or } \beta_5 \neq 0 \text{ or } \beta_6 \neq 0 \text{ or } \beta_7 \neq 0. \end{aligned}$$

Use the Wald statistic

$$\begin{aligned} W &= \frac{(\mathbf{K}'\hat{\boldsymbol{\beta}})'(\tilde{\sigma}^2 \mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K})^{-1}\mathbf{K}'\hat{\boldsymbol{\beta}}}{q} \\ &= \frac{(\mathbf{K}'\hat{\boldsymbol{\beta}})' \left(\widetilde{\mathrm{Cov}}(\mathbf{K}'\hat{\boldsymbol{\beta}})\right)^{-1} \mathbf{K}'\hat{\boldsymbol{\beta}}}{q} \sim F_{q,n-(p+1)}, \end{aligned}$$

to test the above hypotheses. Construct appropriate matrix $\mathbf{K}$ and then calculate the value of the test statistic. Return the value of the Wald statistic as your response.

(2 points)

(d) Create scatterplots between fitted values $\hat{\mu}_i$ and residuals e_i in case of models $\mathcal{M}_{1|2|3}$ and $\mathcal{M}_{123}$. Based on your plots, which one you feel is estimating the expected values with less systematic bias on fit?

i. The model $\mathcal{M}_{1|2|3}$.
ii. The model $\mathcal{M}_{123}$.

(1 point)

3. (a) Let us assume $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$. It is shonw that the maximum likelihood estimator of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. Consider then the estimator $\mathbf{X}\hat{\boldsymbol{\beta}}$. Calculate the expected value $\mathrm{E}(\mathbf{X}\hat{\boldsymbol{\beta}})$ and the covariance matrix $\mathrm{Cov}(\mathbf{X}\hat{\boldsymbol{\beta}})$. What distribution the fitted values $\mathbf{X}\hat{\boldsymbol{\beta}}$ are following?

(2 points)

(b) Consider the linear model

$$\begin{aligned}\mathbf{y} &\sim N(\boldsymbol{\mu}, \sigma^2\mathbf{I}),\\ \boldsymbol{\mu} &= \mathbf{1}\beta_0,\end{aligned}$$

where $\mathbf{1}$ is a vector of ones $\mathbf{1} = (1, 1, \ldots, 1)'$. Use the fundamental equation of the BLUE to show that the sample mean

$$\bar{y} = \frac{y_1 + y_2 + \cdots + y_n}{n} = \frac{1}{n}\mathbf{1}'\mathbf{y}$$

is the best linear unbiased estimator for the parameter β_0, i.e., $\hat{\beta}_0 = \bar{y}$.

(2 points)

(c) Let us consider the large sample situation where we have simulated $n_j = 10000$ observations from the normal distributions $Y_{iA} \sim N(\mu, \sigma^2)$ and $Y_{iB} \sim N(\mu + \delta, \sigma^2)$. The aim is investigate in which values $\delta > 0$ the predictive effect size between Y_{iA} and Y_{iB} is large enough that you would be ready to declare that there is a real effect size between the random variables Y_{iA} and Y_{iB}. Copy the R-code given below. Start changing the value of δ in code and based on the histograms and estimated density curves, decide yourself which values of δ are such that the values of the random variables Y_{iA} and Y_{iB} are "mostly" different. What are your corresponding p-value and d-value when you are feeling that there is a effect size difference?

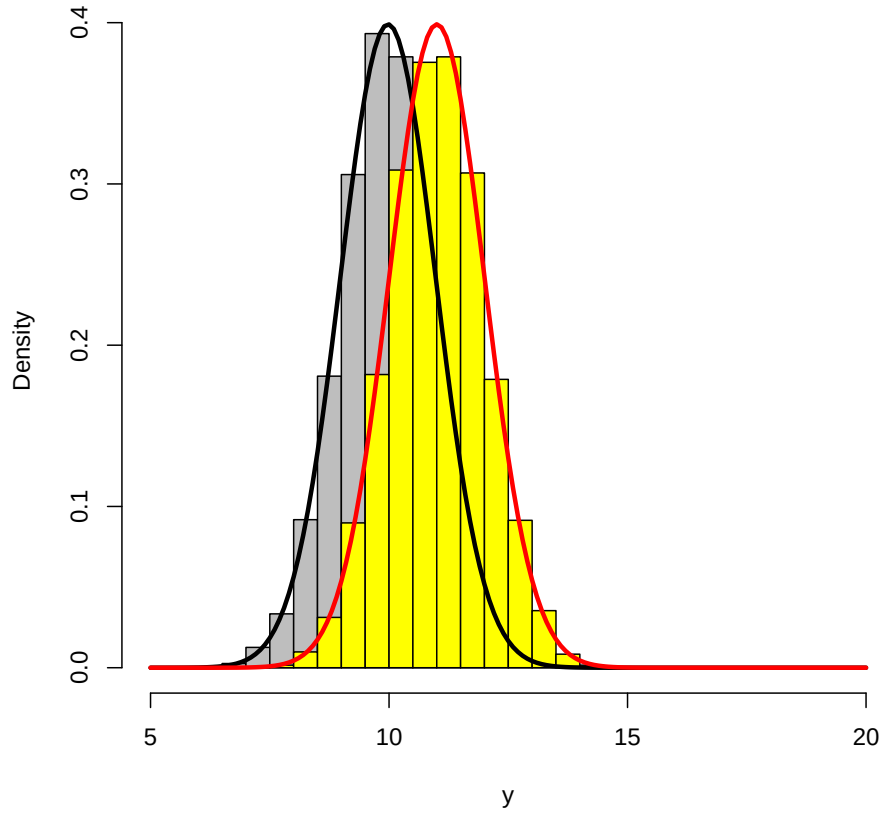
```
muA<-10
delta<-1         # You should change this value
muB<-muA+delta
x<-rep(c("A","B"),each=10000)
yA<-rnorm(10000, mean=muA, sd=1)
yB<-rnorm(10000, mean=muB, sd=1)
y<-c(yA,yB)
model<-lm(y~factor(x))
betahat<-coef(model)

k1<-c(0,1)
K<-cbind(k1)
q<-1
Wald<-(t(t(K)%*%betahat)%*%solve(t(K)%*%vcov(model)%*%K)%*%t(K)%*%betahat)/q
Wald
p.value<-pf(Wald, 1, 19998, lower.tail = FALSE)
p.value

x1<-cbind(c(1,0))
x2<-cbind(c(1,1))
pred<-(t(x2)-t(x1))%*%betahat
sigma2<-sigma(model)^2
X<-model.matrix(model)
T<-pred/sqrt(sigma2*(2+(t(x2)-t(x1))%*%solve(t(X)%*%X)%*%(x2-x1)))
```

```
d.value<-2*pt(abs(T),df=19998, lower.tail = FALSE)
d.value

hist(yA, xlim=c(5,20), col="grey", main="", freq=FALSE, xlab="y")
hist(yB, xlim=c(5,20),add=TRUE, col="yellow", freq=FALSE)
lines(seq(5,20,0.1),dnorm(seq(5,20,0.1), mean=betahat[1],sd=sigma(model)),
col="black",lwd=3)
lines(seq(5,20,0.1),dnorm(seq(5,20,0.1), mean=betahat[1]+betahat[2],sd=sigma(model)),
col="red",lwd=3)
```



(2 points)