STA 108 Discussion 3: Prediction

Yuanyuan Li

Reference: Textbook Chapter 2.4-2.6.

Background (Problem 1.21 from textbook): A substance used in biological and medical research is shipped by airfreight to users in cartons of 1,000 ampules. The data below, involving 10 shipments, were collected on the number of times the carton was transferred from one aircraft to another over the shipment route (X) and the number of ampules found to be broken upon arrival (Y).

```
#Read data of Problem 1.21 in textbook
setwd("~/books/108s21/UCDSTA108-master/datasets")#set working directory to "datasets" folder
data1 = read.table("airfreight+breakage.txt")
Y = data1[,1]
X = data1[,2]
n = length(X)
```

1. Confidence interval for $E(Y_h)$ given x_h

The $100(1-\alpha)\%$ c.i. for $E(Y_h)$ is

$$\hat{Y}_h \pm t_{n-2;1-\alpha/2} \text{ s.e.}(\hat{Y}_h),$$

where

$$\hat{Y}_h = \hat{\beta}_0 + \hat{\beta}_1 X_h,$$

$$s.e.(\hat{Y}_h) = \sqrt{MSE\left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)}.$$

Let us see how to obtain the confidence interval in R. Suppose we are interested in the mean breakage when the carton was transferred 3 times ($X_h = 3$).

```
#Get least square estimates:
fit=lm(Y~X)
b0hat = fit$coefficients[[1]]
b1hat = fit$coefficients[[2]]
```

Compute MSE, the point estimate $\hat{Y}_h$ and standard error $s.e.(\hat{Y}_h)$.

```
#MSE
mse= summary(fit)$sigma^2
#estimated mean response given xh=3:
xh = 3
yhat = b0hat + b1hat*xh#xh is a new observation
yhat
```

```
## [1] 22.2
```

```r
se.yhat = sqrt(mse*(1/n+ (xh - mean(X))^2/sum((X - mean(X))^2)))
se.yhat
```

```
## [1] 1.048809
```

Calculate c.i given $\alpha = 0.05$,

```r
alpha = 0.05
p = 1-alpha/2
critical.value = qt(p, n-2)
yhat.lb = yhat - critical.value*se.yhat
yhat.ub = yhat + critical.value*se.yhat
c(yhat.lb, yhat.ub)
```

```
## [1] 19.78144 24.61856
```

A more convenient way is to use `predict` function,

```r
predict(fit, newdata=data.frame(X=3), interval="confidence",level = 0.95)
```

```
##    fit      lwr      upr
## 1 22.2 19.78144 24.61856
```

Interpretation: We are 95% confident that *the mean breakage of shipments* transferred 3 times is between 19.78 and 24.62.

2. Prediction interval for Y_h given X_h

The $100(1-\alpha)\%$ prediction interval for Y_h is

$$\hat{Y}_h \pm t_{n-2;1-\alpha/2} \text{ p.s.e.}(\hat{Y}_h)$$

where $\hat{Y}_h$ is the same with the above point estimator, p.s.e.$(\hat{Y}_h)$ is the prediction standard error,

$$\text{p.s.e}(\hat{Y}_h) = \sqrt{MSE\left(1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)}$$

Let us predict the breakage of a shipment transferred 3 times ($X_h = 3$).

```r
pse.yhat = sqrt(mse*(1+1/n+ (xh - mean(X))^2/sum((X - mean(X))^2)))
pse.yhat
```

```
## [1] 1.81659
```

```r
#Construct prediction interval using pse.yhat
ypred.lb = yhat - critical.value*pse.yhat
ypred.ub = yhat + critical.value*pse.yhat
c(ypred.lb, ypred.ub)
```

```
## [1] 18.01094 26.38906
```

We can also use the `predict` function.

```r
predict(fit, newdata=data.frame(X=3), interval="prediction",level = 0.95)
```

```
##    fit      lwr      upr
## 1 22.2 18.01094 26.38906
```

Interpretation: We are 95% confident that *the breakage of a shipment* transferred 3 times is between 18.01 and 26.39.

3. Prediction interval for the mean of m new observations for given X_h

The $100(1-\alpha)\%$ prediction interval for the mean of m new observations is

$$\hat{Y}_h \pm t_{n-2;1-\alpha/2}\ s(\text{predmean}),$$

where

$$s(\text{predmean}) = \sqrt{MSE\left(\frac{1}{m}+\frac{1}{n}+\frac{(X_h-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\right)}$$

Suppose we have $m = 4$ new shipments transferred 3 times each$(X_h = 3)$, and we want to predict the mean breakage of these 4 shipments.

```r
m=4
s.predmean = sqrt(mse*(1/m + 1/n + (xh - mean(X))^2/sum((X - mean(X))^2)))
s.predmean
```

```
## [1] 1.284523
```

```r
#lower bound for average of predicted y
mean.pred.lb = yhat - critical.value*s.predmean
#upper bound for average of predicted y
mean.pred.ub = yhat + critical.value*s.predmean
#prediction confidence interval for the mean
c(mean.pred.lb, mean.pred.ub)
```

```
## [1] 19.23788 25.16212
```

```r
#prediction confidence interval for total
c(m*mean.pred.lb, m*mean.pred.ub)
```

```
## [1]  76.95154 100.64846
```

Interpretation: We are 95% confident that *the mean breakage of 4 new shipments* transferred 3 times each is between 19.24 and 25.16. Hence, the total breakage of the 4 new shipments transferred 3 times is between 76.95 and 100.65 with 95% confidence.

4. Prediction plots for a gid of new X values

Given a gid of new X values on the original X range $X_h = (0, 0.05, \cdots, 3)$, let draw the point estimators $\hat{Y}_h$ with the c.i. for $E(Y_h)$ and p.i for Y_h.

```
#Fitted line
plot(X,Y,ylim=c(5,30),xlab = "Number of transfer",ylab = "Number of broken ampules")
abline(fit, col="red")

newx <- seq(0,3, by=0.05)#create a gid of new x values
#Confidence interval for E(Yh)
conf_interval <- predict(fit, newdata=data.frame(X=newx), interval="confidence", level = 0.95)
lines(newx, conf_interval[,2], col="orange", lty=2)
lines(newx, conf_interval[,3], col="orange", lty=2)

#Prediction interval for Yh
pred_interval <- predict(fit, newdata=data.frame(X=newx), interval="prediction", level = 0.95)
lines(newx, pred_interval[,2], col="blue", lty=4)
lines(newx, pred_interval[,3], col="blue", lty=4)
legend("topleft", legend=c("fitted values", "c.i. for E(Yh)", "p.i. for Yh"),
       col=c("red", "orange", "blue"), lty=c(1,2,4),cex=0.8)
points(mean(X),mean(Y), col="purple",pch=5)
```