STAT GR5205 – Section 005 HW 2

Bo Rong br2498

Oct. 2nd, 2016

```r
#1.(a) I don't agree. Because there maybe exist negative linear association between X and Y.
```

```r
#(b)
#No,I don't agree. The standard error for this case is necessarily larger
#as prediction is subject to two sources of uncertainty: (1) the true mean
#E(Y |X = xh) = beta0 + beta1xh is unknown, so we use instead b0 + b1xh
#which is subject to estimation error; and (2) even if the values of beta0
#and beta1 were known exactly, the mean of m observations will not equal its true mean.
#The estimation problem is subject only to the first source.
```

```r
#(c)
#No,I don't agree.95% prediction interval must account for the variance as estimated by
#the MSE which is Yhath+-t(.975;n-2) {MSE[1/n+(xh-xbar)^2/(sum(xi-xbar)^2) +MSE}^1/2 .
```

```r
#2.(a)
filename <- "~/Downloads/copier_maintenance.txt"
copier_maintenance<- read.table(file=filename, header=T)
#H0 :beta1 =0, H1:beta1!=0.
x <- copier_maintenance$copiers
y <- copier_maintenance$minutes
xbar <- mean(x); ybar <- mean(y)
b1 <- sum( (x - xbar)*(y - ybar) ) / sum( (x - xbar)^2 )
b1
```

```
## [1] 15.03525
```

```r
b0 <- ybar - b1*xbar
b0
```

```
## [1] -0.5801567
```

```r
yhat <- b0 + b1*x
e <- y - yhat
n <- length(y)
MSE <- sum(e^2)/(n-2)
se.b1 <- sqrt( MSE / sum( (x - xbar)^2 ) )
se.b1
```

```
## [1] 0.4830872
```

```r
t.star <- b1 / se.b1
t.star
```

```
## [1] 31.12326
```

```r
2 * (1 - pt(t.star, df=n-2))  # p-value
```

```
## [1] 0
```

```r
#Since p value is 0,there exists a linear association between X and Y.
```

```r
#(b)
fit <- lm(minutes ~ copiers, data=copier_maintenance)
confint(fit, "copiers", level=.95)
```

```
##            2.5 %   97.5 %
## copiers 14.06101 16.00949
```

```r
#95% CI is between 14.06101 and 16.00949 minutes.
```

```r
#(c)
#H0 :beta1 <=14, Ha :beta1 >14
t.star <- (b1 - 14) / se.b1
t.star
```

```
## [1] 2.142984
```

```r
1 - pt(t.star, df=n-2)  # p-value
```

```
## [1] 0.01890766
```

```r
#The P-value is about .01890766, so the manager's claim
#is wrong,
#we would reject H0 at alpha = .05 but not alpha = .01).
```

```r
#(d)
summary(fit)
```

```
##
## Call:
## lm(formula = minutes ~ copiers, data = copier_maintenance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.7723  -3.7371   0.3334   6.3334  15.4039
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.5802     2.8039  -0.207    0.837
## copiers      15.0352     0.4831  31.123   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.914 on 43 degrees of freedom
## Multiple R-squared:  0.9575, Adjusted R-squared:  0.9565
## F-statistic: 968.7 on 1 and 43 DF,  p-value: < 2.2e-16
```

```r
#The test ofH0 :beta0 =0, Ha beta0!=0 gives a P -value of .837. So the data are consistent with
#the time required for start-up work is zero.
```

```r
#3.(a)
predict(fit, data.frame(copiers=6), interval="confidence")
```

```
##        fit     lwr      upr
## 1 89.63133 86.8152 92.44746
```

```r
#The 95% CI is between 86.82 and 92.45 minutes.
```

```r
#(b)
predict(fit, data.frame(copiers=6), interval="prediction")
```

```
##        fit      lwr      upr
## 1 89.63133 71.43628 107.8264
```

```r
#The 95% prediction interval for the next service call, in which six copiers are serviced
#will be between 71.43 and 107.83 minutes.
```

```r
#(c)
c(86.82/6,92.45/6)
```

```
## [1] 14.47000 15.40833
```

```r
#The 95% confidence interval of the expected service time per copier on calls
#in which six copiers are to be serviced is between 14.47 and 15.41 minutes.
```

```r
#4.(a)
SST <- sum( (y - ybar)^2 )
SST #sum of square of total
```

```
## [1] 80376.8
```

```r
SSE <- sum( (y - yhat)^2 )
SSE #sum of square of error
```

```
## [1] 3416.377
```

```r
SSR <- sum( (yhat - ybar)^2 )
SSR #sum of square of Residuals
```

```
## [1] 76960.42
```

```r
MSR <- SSR/1
MSR #mean of square of Residuals with degree of freedom 1
```

```
## [1] 76960.42
```

```r
MSE <- SSE/(n-2)
MSE #mean of square of error with degree of freedom n-2
```

```
## [1] 79.45063
```

```r
#(b)
#H0 :beta1 =0, Ha:beta1!=0
F.star <- MSR / MSE
F.star
```

```
## [1] 968.6572
```

```r
1 - pf(F.star, df1=1, df2=n-2)
```

```
## [1] 0
```

```r
#The P-value is zero, there were no linear association between the variables.
```

```r
#(c)
R_square<-SSR/SST
R_square  #relative reduction
```

```
## [1] 0.9574955
```

```r
#Since the result is 0.9574955 ,it's a large reduction.
```

```r
#5.(a)
filename <- "~/Downloads/crime_rates.txt"
crime_rates <- read.table(file=filename, header=T)
attach(crime_rates)
```

```
## The following objects are masked _by_ .GlobalEnv:
##
##     x, y
```

```r
#H0 :beta1 =0 , Ha :beta1!=0
fit <- lm(y ~ x, data=crime_rates)
summary(fit)
```

```
##
## Call:
## lm(formula = y ~ x, data = crime_rates)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5278.3 -1757.5  -210.5  1575.3  6803.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 20517.60    3277.64   6.260 1.67e-08 ***
## x            -170.58      41.57  -4.103 9.57e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2356 on 82 degrees of freedom
## Multiple R-squared:  0.1703, Adjusted R-squared:  0.1602
## F-statistic: 16.83 on 1 and 82 DF,  p-value: 9.571e-05
```

```r
#t_star is -4.103.
n <- nrow(crime_rates)
2 * pt(-4.103, df=n-2)
```

```
## [1] 9.567866e-05
```

```r
#The p value is 0.00009567866,there were no linear association between crime rate and percentage of hig
```

```r
#(b)
confint(fit, "x", level=.99)
```

```
##       0.5 %    99.5 %
## x -280.2118 -60.93856
```

```r
#The 99% CI is (-280.2118, -60.93856),which means for every additional percentage point
#of residents with HS diploma,expected crime rate for a county decreases by
#between 61 and 280 crimes per 100,000 residents.
```

```r
#(c)
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df    Sum Sq  Mean Sq F value    Pr(>F)
## x          1  93462942 93462942  16.834 9.571e-05 ***
## Residuals 82 455273165  5552112
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#(d)
#The F value in the ANOVA table is 16.83,and P = .000086. The P -values are the same. And F value = squ
```

```r
#(e)
SST=93462942+455273165
SSE=93462942
R_square<-SSE/SST
R_square  #The proportion
```

```
## [1] 0.170324
```

```r
#R_square=0.170324.It's a relatively small reducton.
```

```r
#6.
filename <- "~/Downloads/SENIC.txt"
SENIC <- read.table(file=filename, header=T)
fit.Risk <- lm(Stay ~ Risk, data=SENIC)
summary(fit.Risk)
```

```
##
## Call:
## lm(formula = Stay ~ Risk, data = SENIC)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0587 -0.7776 -0.1487  0.7159  8.2805
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.3368     0.5213  12.156  < 2e-16 ***
## Risk          0.7604     0.1144   6.645 1.18e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.624 on 111 degrees of freedom
## Multiple R-squared:  0.2846, Adjusted R-squared:  0.2781
## F-statistic: 44.15 on 1 and 111 DF,  p-value: 1.177e-09
```

```r
fit.AFS <- lm(Stay ~ AFS, data=SENIC)
summary(fit.AFS)
```

```
##
## Call:
## lm(formula = Stay ~ AFS, data = SENIC)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2712 -1.0716 -0.2816  0.7584  9.5433
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.71877    0.51020  15.129  < 2e-16 ***
## AFS          0.04471    0.01116   4.008 0.000111 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.795 on 111 degrees of freedom
## Multiple R-squared:  0.1264, Adjusted R-squared:  0.1185
## F-statistic: 16.06 on 1 and 111 DF,  p-value: 0.0001113
```

```r
fit.Xray <- lm(Stay ~ Xray, data=SENIC)
summary(fit.Xray)
```

```
##
## Call:
## lm(formula = Stay ~ Xray, data = SENIC)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9226 -1.0810 -0.2708  0.8200  8.7008
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.566373   0.726094   9.043 5.67e-15 ***
## Xray        0.037756   0.008657   4.361 2.91e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.774 on 111 degrees of freedom
## Multiple R-squared:  0.1463, Adjusted R-squared:  0.1386
## F-statistic: 19.02 on 1 and 111 DF,  p-value: 2.906e-05
```

```
#The multiple R-squared are 0.2846,0.1264 and 0.1463 for the regression of Stay on
#Risk, AFS, and Xray respectively. The multiple R-squared is the R^2 value.
#Hence Infection risk accounts for the largest reduction to variability in average length of stay.
#Compare to the conclusion reached based on MSE,they are same. Because of the equation:
#R^2=1-(n-2)/[sum(y-ybar)^2]*MSE
```