STA 101 HW 3 Solutions

Dr. Erin K. Melcon

"By Hand" Problems

1. (a) $H_0 : \beta_2 = \beta_3 = 0$ vs. H_A : At least one β_2 or β_3 does not equal zero.

   (b) $F_S = \frac{\frac{SSE_R - SSE_F}{d.f.(SSE_R) - d.f.(SSE_F)}}{\frac{SSE_F}{d.f.(SSE_F)}} = \frac{\frac{12744.8829 - 10819.3756}{(1436-2)-(1436-4)}}{\frac{10819.3756}{1436-4}} = 127.4254$
   By R, the p-value is ≈ 0.
   By the table, p-value < 0.0001 at $d.f.\{num\} = 2$, $d.f.\{denom\} \approx 140$

   (c) We reject the null and conclude that the model that includes information about the type of fuel is significantly better than one that does not (assuming information on total distance is included). In other words, we should not remove the variable with type of fuel from the model.

   (d) $H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 \neq 0$

   (e) $F_S = \frac{\frac{SSE_R - SSE_F}{d.f.(SSE_R) - d.f.(SSE_F)}}{\frac{SSE_F}{d.f.(SSE_F)}} = \frac{\frac{18795.4714 - 10819.3756}{(1436-3)-(1436-4)}}{\frac{10819.3756}{1436-4}} = 1055.6773$
   By R, the p-value is ≈ 0.
   By the table, p-value < 0.0001 at $d.f.\{num\} = 1$, $d.f.\{denom\} \approx 140$

   (f) We reject the null and conclude that the model that includes information total distanceis significantly better than one that does not (assuming information on about the type of fuel is included). In other words, we should not remove the variable with total distance from the model.

---

2. (a) This is $R^2\{X_1|\cdot\} = \frac{SSE(\cdot) - SSE(X_1)}{SSE(\cdot)} = \frac{18877.2415 - 12744.8829}{18877.2415} = 0.3249$

   (b) This is $R^2\{X_2|\cdot\} = \frac{SSE(\cdot) - SSE(X_2)}{SSE(\cdot)} = \frac{18877.2415 - 18795.4714}{18877.2415} = 0.0043$

   (c) $R^2\{X_1|X_2\} = \frac{SSE(X_2) - SSE(X_1, X_2)}{SSE(X_2)} = \frac{18795.4714 - 10819.3756}{18795.4714} = 0.4244$
   When we add information on distance traveled to a linear model that has information on fuel type in it, we reduce the error by 42.44%.

   (d) $R^2\{X_2|X_1\} = \frac{SSE(X_1) - SSE(X_1, X_2)}{SSE(X_1)} = \frac{12744.8829 - 10819.3756}{12744.8829} = 0.1511$
   When we add information on fuel type to a linear model that has information on distance traveled in it,, we reduce the error by 15.11%.

   (e) I would pick the model with both X_1 and X_2, since they both significantly reduce the error in the model.

---

3. (a) $H_0 : \beta_2 = \beta_3 = 0$ vs. H_A : At least one β_2 or β_3 does not equal zero.

   (b) $F_S = \frac{\frac{SSE_R - SSE_F}{d.f.(SSE_R) - d.f.(SSE_F)}}{\frac{SSE_F}{d.f.(SSE_F)}} = \frac{\frac{5093.92 - 4248.84}{(46-2)-(46-4)}}{\frac{4248.84}{46-4}} = 4.1768$
   By R, the p-value is ≈ 0.0221612
   By the table, $0.02 <$ p-value < 0.05 at $d.f.\{num\} = 2$, $d.f.\{denom\} \approx 40$

   (c) If in reality a model without information on severity of illness and anxiety level fits better (or there was no significant linear relationship between patient satisfaction and severity of illness and anxiety level score), we would observe our data or more extreme with probability between 0.02 and 0.05.

   (d) We fail to reject the null, and conclude that we can drop both variables from the model. In other words, a model without information on severity of illness and anxiety level is statistically no worse than a model with those variables (including information on age).

   (e) $R^2\{X_2, X_3|X_1\} = \frac{SSE(X_1) - SSE(X_1, X_2, X_3)}{SSE(X_1)} = \frac{5093.92 - 4248.84}{5093.92} = 0.1659$
   When we add information on severity of illness and anxiety level to a model with information on age in it, we reduce our overall error by 16.59%.

(f) When you add X's to a model, in reality they all share a small correlation with Y, so will help lower the error by some amount. In other words, the model with the most X values always has the lowest SSE.

---

4. (a) The best model by AIC is the model that includes X_1, X_3 (the lowest AIC).
   (b) The best model by BIC is the model that includes X_1, X_3 (the lowest BIC).
   (c) The best model by PRESS is the model with the lowest PRESS, which is the model that includes X_1, X_3.
   (d) I would pick the model with X_1, X_3, since most of the criteria agree this is the "best" model.

---

5. (a) FALSE. SSE relies on the units of Y, and if we transform Y, the SSE's will have different unit and not be comparable.
   (b) FALSE. An underfit model is too small, so exluded important X variables.
   (c) TRUE. Prediction models tend to be a bit larger than "correct" models, for example.
   (d) TRUE. SSE will always decrease, so R^2 will always increase.

R problems

I. (a) $H_0 : \beta_4 = \beta_5 = 0$ vs. H_A : at least one β_i is non zero, $i = 4, 5$.
   The test-statistic is 19.3946, with corresponding p-value 0.0000008

   (b) We reject the null, and conclude we cannot drop X_4 (which contains information about the rank of the subject) from the model.

   (c) $H_0 : \beta_2 = \beta_3 = 0$ vs. H_A : at least one β_i is non zero, $i = 2, 3$.
   The test-statistic is 0.6869, with corresponding p-value 0.5081955

   (d) We fail to reject the null, and conclude we can drop both X_2 (information on the highest degree earned) and X_3 (gender) from the model.

   (e) The best linear model is one with X_1 and X_4, which has model fit :
   $\hat{Y} = 17166.46 + 95.08X_1 + 4209.65X_{4,associate} + 10310.30X_{4,full}$

II. (a) This is found to be 57.24%.

   (b) $R^2\{X_1|X_4\} = 0.0528$. When we add information about the years of experience to a model that contains information about the rank of the subject, we reduce our overall error by 5.28%.

   (c) This is found to be 2.9%.

   (d) $R^2\{X_1, X_4|X_2, X_3\} = 0.7578$. When we add information about the years of experience and the rank of the subject to a model that contains information about the gender of the subject and their highest degree earned, we reduce our overall error by 75.78%.

   (e) It does, since X_1, X_2 appear to have little effect on reducing error, but X_1, X_4 have a large effect.

III. (a) The values are:

   (b) The best model according to BIC is: Y X4

   (c) The best model according to AIC is: Y X1 + X4

   (d) The best model according to PRESS is: Y X1 + X4

   (e) The best model according to R^2_{adj} is: Y X1 + X3 + X4

   (f) I would use the model with X_1, X_3, and X_4, since it is the largest of the "best" models selected. Either model according to PRESS or R^2_{adj} would have been accepted.

	LL	p	n	AIC	BIC	PRESS	R2adj
Y ˜ X1	-509.131	2.000	52.000	1022.263	1026.165	1060038083.758	0.445
Y ˜ X2	-524.806	2.000	52.000	1053.612	1057.515	1906063823.539	-0.015
Y ˜ X3	-523.216	2.000	52.000	1050.432	1054.335	1814851461.505	0.045
Y ˜ X4	-488.450	3.000	52.000	982.899	988.753	489746475.084	0.744
Y ˜ X1 + X3	-507.270	3.000	52.000	1020.540	1026.394	1036192780.605	0.472
Y ˜ X1 + X2	-504.688	3.000	52.000	1015.377	1021.231	942056188.036	0.522
Y ˜ X1 + X4	-487.040	4.000	52.000	982.080	989.885	486639795.342	0.753
Y ˜ X1 + X2 + X4	-487.031	5.000	52.000	984.062	993.818	524240790.833	0.747
Y ˜ X1 + X3 + X4	-486.276	5.000	52.000	982.551	992.307	501206381.483	0.755
Y ˜ X1 + X2 + X3 + X4	-486.275	6.000	52.000	984.550	996.257	539051398.515	0.749

(g) I would use the model with X_1, X_4, since it is the smallest of the “best” models and has the most agreement. You also could have used the model suggested by BIC.

Code Appendix

```
library(MPV)
library(MASS)
library(rcompanion)
#Problem I
library(car)
alcohol <- read.csv("C:/Github/Teaching-Materials/STA-108-2017-Fall/Datasets/HW02/alcohol.csv")
names(alcohol) = c("X2","X1","Y")
the.model = lm(Y ~ X1, data = alcohol)

tukeyY = transformTukey(alcohol$Y,plotit = FALSE)
tukeyX = transformTukey(alcohol$X1, plotit = FALSE)
lambdaY = 0.3; lambdaX = 0.175
tukey.data = data.frame(Y = tukeyY, X1 = tukeyX)
tukey.model  = lm(Y ~ X1, data = tukey.data)
TSW.pval = shapiro.test(tukey.model$residuals)$p.val

#(b)
BC = boxcox(the.model,lambda = seq(-6,6,0.1),plotit = FALSE)
lambda = BC$x[which.max(BC$y)]
BC.Y = (alcohol$Y^lambda - 1)/lambda
BC.data = data.frame(Y = BC.Y, X1 = alcohol$X1)
BC.model  = lm(Y ~ X1, data = BC.data)
BCSW.pval = shapiro.test(BC.model$residuals)$p.val

BC.data$ei = BC.model$residuals
Group = rep("Lower",nrow(BC.data)) #Creates a vector that repeats "Lower" n times
Group[BC.data$Y > median(BC.data$Y)] = "Upper" #Changing the appropriate values to "Upper"
Group = as.factor(Group) #Changes it to a factor, which R recognizes as a grouping variable.
BC.data$Group = Group
the.FKtest= fligner.test(BC.data$ei, BC.data$Group)

tukey.data$Group = Group
tukey.data$ei = tukey.model$residuals
the.BFtest = leveneTest(ei~Group, data=tukey.data, center=median)
p.val = the.BFtest[[3]][1]
#par(mfrow = c(1,2))
#plot(BC.data,pch = 19,cex = 2, main = "Box-Cox I(a)",font = 2,font.lab = 2)
#plot(tukey.data,pch = 19,cex = 2, main = "Tukey I(c)",font = 2,font.lab = 2)

#Problem II
the.data = read.csv("c:/Github/Teaching-Materials/STA-108-2017-Fall/Datasets/HW05/salary2.csv")[,c(6,5,4,1
names(the.data) = c("Y","X1","X2","X3","X4")
full.model = lm(Y ~ X1 + X2 + X3 + X4, data = the.data )
R.model1  = lm(Y ~ X1 + X2 + X3 , data = the.data )
R.model2  = lm(Y ~ X1 + X4, data = the.data )
Test1 = anova(R.model1,full.model)
Test2 = anova(R.model2,full.model)

#Problem III
Partial.R2 = function(small.model,big.model){
  SSE1 = sum(small.model$residuals^2)
  SSE2 = sum(big.model$residuals^2)
  PR2 = (SSE1 - SSE2)/SSE1
  return(PR2)
}
X1X4 = lm(Y ~ X1 + X4,data = the.data)
X1 = lm(Y ~ X1,data = the.data)
X4 = lm(Y ~ X4, data = the.data)
```

```r
X2X3 = lm(Y ~ X2 + X3, data = the.data)

part.a = round(Partial.R2(X1, X1X4),4)
part.b = round(Partial.R2(X4, X1X4),4)
part.c = round(Partial.R2(X1X4, full.model),4)
part.d = round(Partial.R2(X2X3, full.model),4)

#Problem IV
library(MPV)
all.models = c("Y ~ X1", "Y ~ X2", "Y ~ X3", "Y ~ X4",
                "Y ~ X1 + X3", "Y ~ X1 + X2", "Y ~ X1 + X4",
                "Y ~ X1 + X2 + X4", "Y ~ X1 + X3 + X4",
                "Y ~ X1 + X2 + X3 + X4")
All.Criteria = function(the.model){
  p = length(the.model$coefficients)
  n = length(the.model$residuals)
  the.LL = logLik(the.model)
  the.BIC =  -2*the.LL + log(n)*p
  the.AIC =  -2*the.LL + 2*p
  the.PRESS = PRESS(the.model)
  the.R2adj = summary(the.model)$adj.r.squared
  the.results = c(the.LL,p,n,the.AIC,the.BIC,the.PRESS,the.R2adj)
  names(the.results) = c("LL","p","n","AIC","BIC","PRESS","R2adj")
  return(the.results)
}

all.model.crit = t(sapply(all.models,function(M){
  current.model = lm(M,data = the.data)
  All.Criteria(current.model)
}))
RC = round(all.model.crit,4)
```