Homework 8

Yunting Chiu

2021-04-16

1. (**9.1**) A speaker stated: "In well-designed experiments involving quantitative explanatory variables, a procedure for reducing the number of explanatory variables after the data are obtained is not necessary." Do you agree? Discuss.

- An explanatory variable is a type of independent variable.
- Assume that these variables have met the assumptions and that there is no collinearity in the well-designed experiments. The better approach is to carry out procedures to reduce the number of explanatory variables that are not significant, because these variables cannot explain the model and may reduce the precision of the outcome.

2. (**9.5**) In forward stepwise regression, what advantage is there in using a relatively small α to-enter value for adding variables? What advantage is there in using a larger α-to-enter value?

- The main point of stepwise regression method is to obtain the best relationship between the independent variables and the dependent variable.
- Advantage of small α to-enter value:

4. (**Continuing 6.27 from an earlier homework**) In a small-scale regression study, the following data were obtained,

Y	X1	X2
42.0	7.0	33.0
33.0	4.0	41.0
75.0	16.0	7.0
28.0	3.0	49.0
91.0	21.0	5.0
55.0	8.0	31.0

Make a data frame

```
Y <- c(42, 33, 75, 28, 91, 55)
X1 <- c(7, 4, 16, 3, 21, 8)
X2 <- c(33, 41, 7, 49, 5, 31)

df <- data.frame(Y, X1, X2)
df
```

```
##    Y X1 X2
## 1 42  7 33
## 2 33  4 41
## 3 75 16  7
## 4 28  3 49
## 5 91 21  5
```

```
## 6 55  8 31
```

Model selection

1. Exhaustive Search

```
library(leaps)
df.fit <-  regsubsets(Y ~ X1 + X2, data = df)
summary(df.fit)
```

```
## Subset selection object
## Call: regsubsets.formula(Y ~ X1 + X2, data = df)
## 2 Variables  (and intercept)
##    Forced in Forced out
## X1     FALSE      FALSE
## X2     FALSE      FALSE
## 1 subsets of each size up to 2
## Selection Algorithm: exhaustive
##          X1  X2
## 1  ( 1 ) "*" " "
## 2  ( 1 ) "*" "*"
```

1.1 Find out the largest adjusted R squares

- Because R^2 is not a fair measurement. As the number of parameters increases, so does the $R2$.

```
summary(df.fit)$adjr2
```

```
## [1] 0.9724995 0.9663230
```

```
which.max(summary(df.fit)$adjr2)
```

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

1.2 Find out the smallest Mallows Cp

```
summary(df.fit)$cp
```

```
## [1] 1.266385 3.000000
```

```
which.min(summary(df.fit)$cp)
```

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

1.3 Find out the smallest BIC (penalized-likelihood criteria)

```
summary(df.fit)$bic
```

```
## [1] -19.31664 -18.03531
```

```
which.min(summary(df.fit)$bic)
```

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

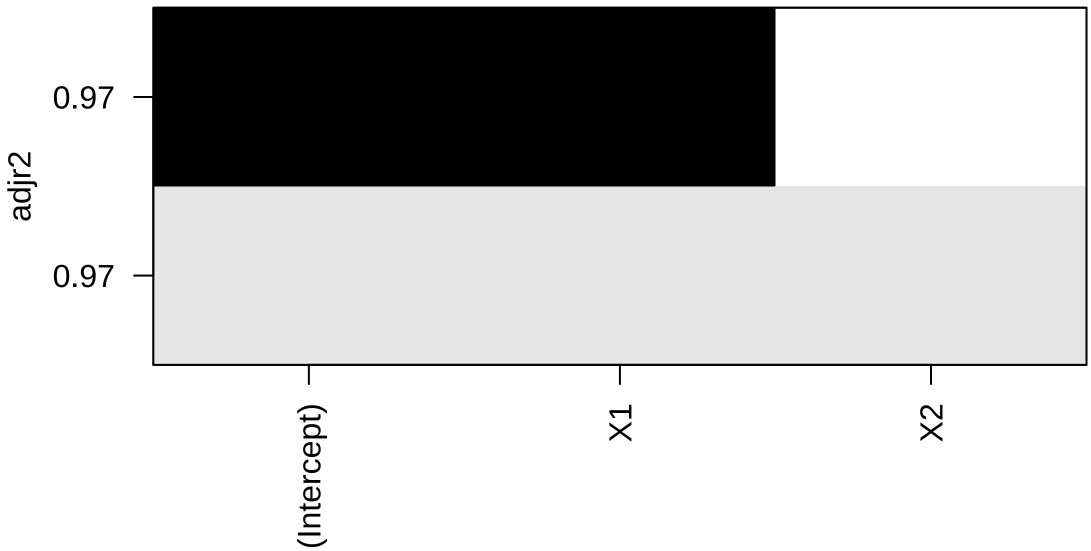
Conclusion

$$\hat{Y} = \beta_0 + \beta_1 X_1 + e$$

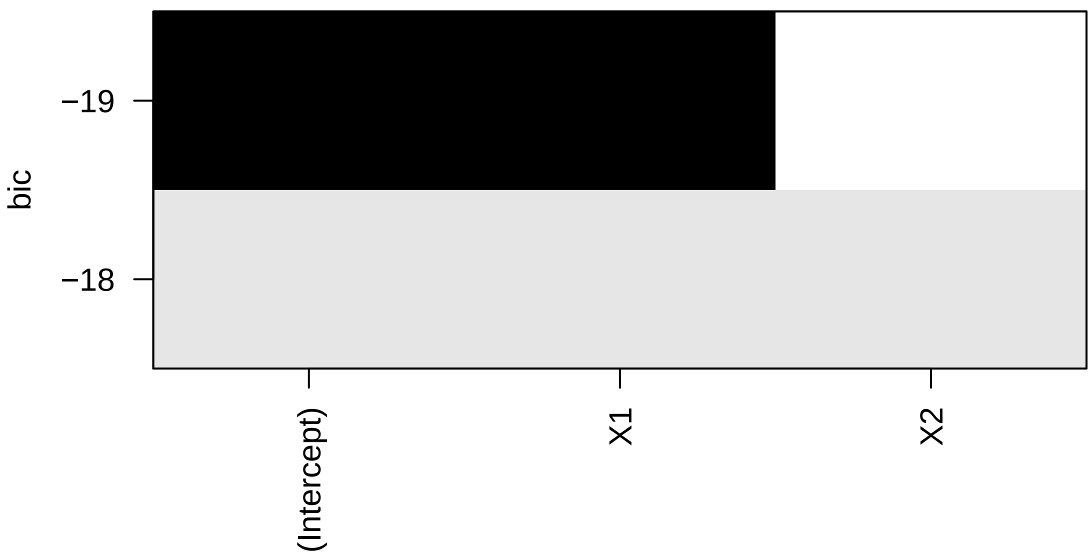
2. Sequential Search

2.1 Find out the proper adjusted R and BIC using plot

```
reg.backward <-  regsubsets( Y ~ ., data = df, method = "backward" )
plot(reg.backward, scale = "adjr2")
```




```
plot(reg.backward, sclae = "bic")
```



Conclusion

According to the result above, the best model is

$$\hat{Y} = \beta_0 + \beta_1 X_1 + e$$

3. Choosing the best model by means of a stepwise procedure

Forward selection

```r
null <-  lm( Y ~ 1, data = df )
full <-  lm( Y ~ ., data = df )

step(null, scope = list(lower = null, upper = full), direction = "forward" )
```

```
## Start:  AIC=39.43
## Y ~ 1
##
##        Df Sum of Sq     RSS    AIC
## + X1    1    3004.4   67.59 18.530
## + X2    1    2913.4  158.64 23.649
## <none>              3072.00 39.430
##
## Step:  AIC=18.53
## Y ~ X1
##
##        Df Sum of Sq    RSS    AIC
## <none>              67.585 18.530
## + X2    1    5.5118 62.074 20.019

##
## Call:
## lm(formula = Y ~ X1, data = df)
##
## Coefficients:
## (Intercept)           X1
##      20.236        3.434
```

Backward elimination

```r
step(null, scope = list(lower = null, upper = full), direction = "backward" )
```

```
## Start:  AIC=39.43
## Y ~ 1

##
## Call:
## lm(formula = Y ~ 1, data = df)
##
## Coefficients:
## (Intercept)
##          54
```

Or using algorithm

```r
step(null, scope = list(lower = null, upper = full), direction = "both" )
```

```
## Start:  AIC=39.43
## Y ~ 1
##
##        Df Sum of Sq     RSS    AIC
## + X1    1    3004.4   67.59 18.530
## + X2    1    2913.4  158.64 23.649
## <none>              3072.00 39.430
```

```
##
## Step:  AIC=18.53
## Y ~ X1
##
##        Df Sum of Sq     RSS    AIC
## <none>                 67.59 18.530
## + X2    1      5.51   62.07 20.019
## - X1    1   3004.41 3072.00 39.430
```

```
##
## Call:
## lm(formula = Y ~ X1, data = df)
##
## Coefficients:
## (Intercept)           X1
##      20.236        3.434
```

Conclusion

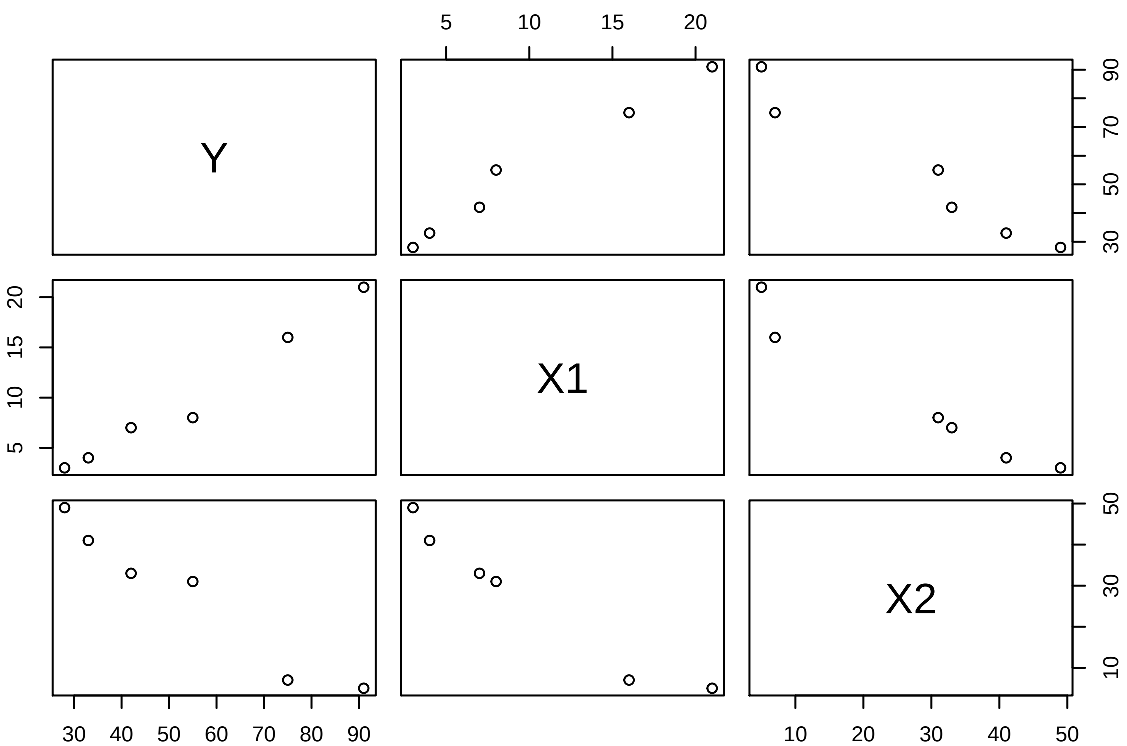
In summary, the smallest RSS is `Y~X1`, so the best performance of this model is:

$$\hat{Y} = \beta_0 + \beta_1 X_1 + e$$

4 Visualization – scatterplot matrix

- We can also see if there is a linear relationship between the independent and dependent variables in scatterplot. According to the plot, `Y` and `X1` does have a linear relationship, but `X2` and `Y` does not. Also, the model may have a multicollinearity problem as `X1` and `X2` appear to have a linear relationship so consider keeping `lm(Y ~ X1)` to run a linear model.

```
plot(df)
```



Conclusion

$$\hat{Y} = \beta_0 + \beta_1 X_1 + e$$

5. (**9.10–9.11, 9.18, 9.21–9.22**) A personnel officer in a governmental agency administered four newly developed aptitude tests to each of 25 applicants for entry-level clerical positions in the agency. For purpose of the study, all 25 applicants were accepted for positions irrespective of their test scores. After a probationary period, each applicant was rated for proficiency on the job, and scores of the four tests (X1, X2, X3, X4) and the job proficiency score (Y) were recorded.

```r
A1 <- read_table("./data/CH09PR10.txt", col_names = FALSE)
```

```
## Parsed with column specification:
## cols(
##   X1 = col_double(),
##   X2 = col_double(),
##   X3 = col_double(),
##   X4 = col_double(),
##   X5 = col_double()
## )
```

```r
A2 <- read_table("./data/CH09PR22.txt", col_names = FALSE)
```

```
## Parsed with column specification:
## cols(
##   X1 = col_double(),
##   X2 = col_double(),
##   X3 = col_double(),
##   X4 = col_double(),
##   X5 = col_double()
## )
```

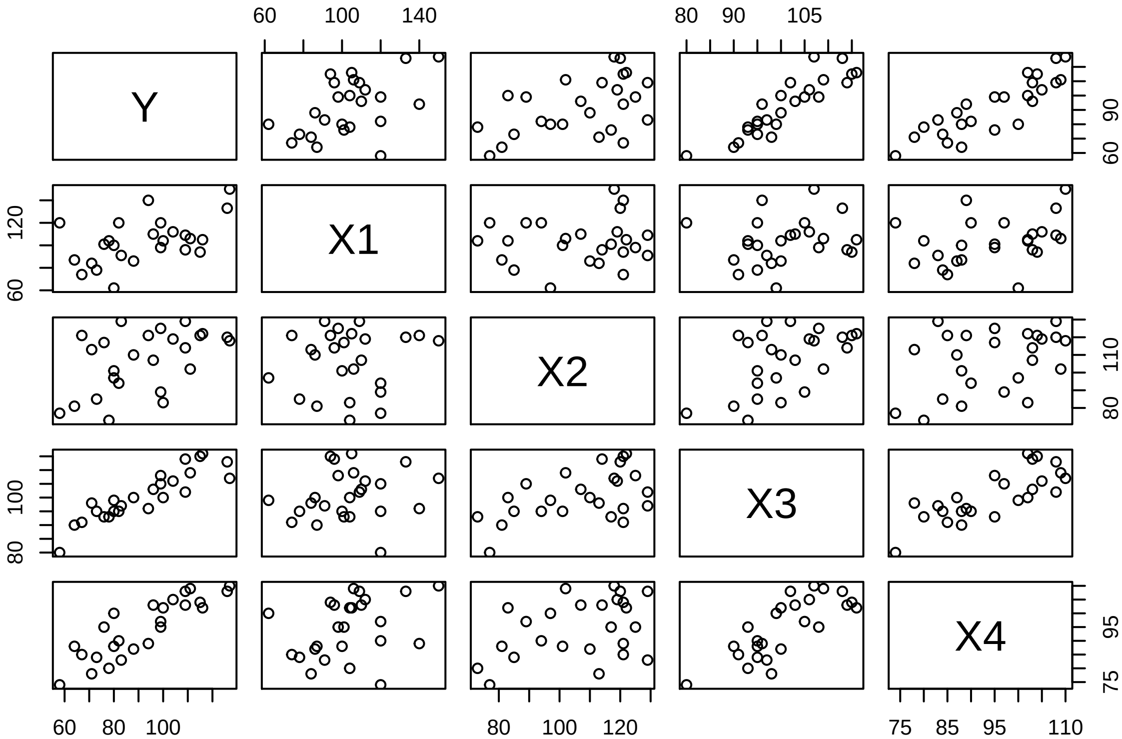
```r
A1 %>%
  rename(Y = X1, X1 = X2, X2 = X3, X3 = X4, X4 = X5) -> A1 # Original Data
A2 %>%
  rename(Y = X1, X1 = X2, X2 = X3, X3 = X4, X4 = X5) -> A2 # Additional Data
```

The resulting **Job Proficiency** data set is available on our Blackboard in "Data sets" and on the next page of this homework assignment.

(a) Obtain the scatter plot matrix of these data. What do the scatter plots suggest about the nature of the functional relationship between the response variable and each of the predictor variables? Do you notice any serious multicollinearity problems?

- `X1` and `X2` do not appear to have a linear relationship with the response variable `Y`. In contrast, `X3` and `X4` appear to have a linear relationship with the response variable `Y`.
- The model may have a multicollinearity problem as `X3` and `X4` appear to have a linear relationship.

```r
plot(A1)
```



(b) Fit the multiple regression function containing all four predictor variables as first-order (linear) terms. Does it appear that all predictor variables should be retained?

- According to the table, the `X2` variable is not in the significant level. In other words, we fail to reject the null: $\beta 2 = 0$ so we consider removing the `X2` variable.

```
# Full model for (b)
mul.reg <- lm( Y ~ ., data = A1)
summary(mul.reg)
```

```
##
## Call:
## lm(formula = Y ~ ., data = A1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9779 -3.4506  0.0941  2.4749  5.9959
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -124.38182    9.94106 -12.512 6.48e-11 ***
## X1             0.29573    0.04397   6.725 1.52e-06 ***
## X2             0.04829    0.05662   0.853  0.40383
## X3             1.30601    0.16409   7.959 1.26e-07 ***
## X4             0.51982    0.13194   3.940  0.00081 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.099 on 20 degrees of freedom
## Multiple R-squared:  0.9629, Adjusted R-squared:  0.9555
## F-statistic: 129.7 on 4 and 20 DF,  p-value: 5.262e-14
```

(c) Using only first-order terms for the predictor variables in the pool of potential X variables, find the best regression models according to different criteria - adjusted R^2, Cp, and BIC.

Exhaustive Search

```
best <- regsubsets(Y ~ ., data = A1)
summary(best)
```

```
## Subset selection object
## Call: regsubsets.formula(Y ~ ., data = A1)
## 4 Variables  (and intercept)
##    Forced in Forced out
## X1     FALSE      FALSE
## X2     FALSE      FALSE
## X3     FALSE      FALSE
## X4     FALSE      FALSE
## 1 subsets of each size up to 4
## Selection Algorithm: exhaustive
##          X1  X2  X3  X4
## 1  ( 1 ) " " " " "*" " "
## 2  ( 1 ) "*" " " "*" " "
## 3  ( 1 ) "*" " " "*" "*"
## 4  ( 1 ) "*" "*" "*" "*"
```

find out the largest adjusted R squares

- If the model includes `X1`, `X3`, and `X4`, the adjusted R squares will be the highest: 0.9560482.

```
summary(best)$adjr2
```

```
## [1] 0.7962344 0.9269043 0.9560482 0.9554702
```

```
which.max(summary(best)$adjr2)
```

```
## [1] 3
```

find out the smallest Mallows Cp

- If the model includes `X1`, `X3`, and `X4`, the CP will be the smallest: 3.727399.

```
summary(best)$cp
```

```
## [1] 84.246496 17.112978  3.727399  5.000000
```

```
which.min(summary(best)$cp)
```

```
## [1] 3
```

find out the smallest Bayesian Information Criterion

- If the model includes X1, X3, and X4, the BIC will be the smallest: -68.57933.

```
summary(best)$bic
```

```
## [1] -34.39587 -57.91831 -68.57933 -66.25356
```

```r
which.min(summary(best)$bic)
```

```
## [1] 3
```

(d) Using **forward** stepwise selection, find the best subset of predictor variables to predict job proficiency. Use the α-to-enter limit of 0.05.

- Forward and backward selection algorithms with partial F-tests at each step.

```r
library(SignifReg) # significance testing in regression model building
null <- lm(Y~1, data = A1)
# summary(null)
full <- lm(Y~., data = A1)

SignifReg(null, alpha = 0.05, direction = "forward")
```

```
##
## Call:
## lm(formula = Y ~ X3 + X1 + X4, data = A1)
##
## Coefficients:
## (Intercept)           X3           X1           X4
##   -124.2000       1.3570       0.2963       0.5174
```

```r
# step(null, scope = list(lower = null, upper = full), direction = "forward", alpha = 0.05)
```

(e) Repeat the previous question using the backward elimination method and the α-to remove limit of 0.10.

```r
SignifReg(full, alpha = 0.1, direction = "backward")
```

```
##
## Call:
## lm(formula = Y ~ X1 + X3 + X4, data = A1)
##
## Coefficients:
## (Intercept)           X1           X3           X4
##   -124.2000       0.2963       1.3570       0.5174
```

Compared to `forward` and `backward` two methods, the slope and the value of b_1, b_3, and b_4 are the same. Also, they both also give up the `X2` variable.

(f) To assess and compare internally the predictive ability of our models, split the data into training and testing subsets and estimate the mean squared prediction error `MSPE` for all regression models identified in (b–e).

- `A1` is a training data, and `A2` is a testing data
- After some methods of model selection, we will use `lm(formula = Y ~ X1 + X3 + X4, data = A1)` to run a regression model.

```r
pd1 <- lm(formula = Y ~ X1 + X3 + X4, data = A1)  # training
```

```r
library(cvTools)
```

```
## Loading required package: lattice
```

```
## Loading required package: robustbase
```

```r
Yhat_pd1 <- predict(pd1, A2) # testing
MSPE_pd1 <- mspe(A2$Y, Yhat_pd1, includeSE = FALSE)
MSPE_pd1
```

```
## [1] 15.70972
```

(g) To assess and compare externally the validity of our models, 25 additional applicants for entry level clerical positions were similarly tested and hired. Their data are below, in the table on the right. Use these data as the testing set and estimate MSPE for all regression models identified in (b–e).

- Adding `X2` in the previous model as full model.

```r
pd2 <- lm(formula = Y ~ ., data = A1) # training
Yhat_pd2 <- predict(pd2, A2) # testing
MSPE_pd2 <- mspe(A2$Y, Yhat_pd2, includeSE = FALSE)
MSPE_pd2
```

```
## [1] 13.95808
```

- Reference: http://finzi.psych.upenn.edu/library/cvTools/html/cost.html