STAT4520 HW1

Anton Yang

2024-09-02

Problem 1 Linear Model Review

Generate data set with x1 be a binary categorical variable with 0 and 1, and x2 be normally randomly generated with mean of 0 and standard deviation of 1. y is the response variable of x1+x2+rnorm(200).

```
set.seed(123)
library(MASS)
library(olsrr)
```

```
##
## Attaching package: 'olsrr'

## The following object is masked from 'package:MASS':
##
##     cement

## The following object is masked from 'package:datasets':
##
##     rivers
```

```
x1<-sample(c(0,1), 200, replace = TRUE)
x2<-rnorm(200)
y<-3+5*x1+7*x2+rnorm(200)
model1<-lm(y~x1+x2)
summary(model1)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.79640 -0.60044  0.03345  0.70939  2.49242
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.13646    0.10008   31.34   <2e-16 ***
## x1           4.79028    0.14391   33.29   <2e-16 ***
## x2           7.01105    0.07488   93.62   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.014 on 197 degrees of freedom
## Multiple R-squared:  0.9795, Adjusted R-squared:  0.9793
## F-statistic:  4716 on 2 and 197 DF,  p-value: < 2.2e-16
```

Now we'll add more predictor with x3 be a exponential randomly generated number with a rate of 1/20. We'll also include x4 be a binary categorical variable of 0 and 1 but with uneven probability. Lastly, let x5 be the sum of x1 and x3 with rnorm(200). Therefore, our true model now will be $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \epsilon$ with $\beta_3 = \beta_4 = \beta_5 = 0$

```r
x3<-rexp(200,1/20)
x4<-sample(c(0,1), prob = c(0.6,0.4), replace = T, size = 200)
x5<-x1+x3+rnorm(200)
model2<-lm(y~x1+x2+x3+x4+x5)
summary(model2)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4 + x5)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.83635 -0.61282  0.06305  0.64914  2.55474
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.13090    0.13519  23.160   <2e-16 ***
## x1           4.89570    0.15677  31.228   <2e-16 ***
## x2           7.00006    0.07366  95.030   <2e-16 ***
## x3           0.08546    0.06967   1.227   0.2215
## x4           0.33389    0.14453   2.310   0.0219 *
## x5          -0.09140    0.06978  -1.310   0.1918
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9948 on 194 degrees of freedom
## Multiple R-squared:  0.9806, Adjusted R-squared:  0.9801
## F-statistic:  1961 on 5 and 194 DF,  p-value: < 2.2e-16
```

```r
anova(model1, model2)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x1 + x2
## Model 2: y ~ x1 + x2 + x3 + x4 + x5
##   Res.Df    RSS Df Sum of Sq     F  Pr(>F)
## 1    197 202.48
## 2    194 191.97  3    10.512 3.541 0.01568 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To check the effectiveness of the model, we'll use the F-test to check on the importance of the parameters x3, x4, and x5. For the Null Hypothesis, H_0: the simpler model with x1 and x2 is correct, and the alternative hypothesis, H_1: our larger model with additional parameters x3, x4, x5 is correct. According to the ANOVA Table, we can see that the p-value is significantly lower than 0.05, so model 2 is a better model. Therefore, we reject the Null Hypothesis.

```
initial_model<-lm(y~1)
forward_model<-stepAIC(initial_model, direction = "forward", scope = formula(model2), trace = FALSE)

summary(forward_model)
```

```
##
## Call:
## lm(formula = y ~ x2 + x1 + x4 + x5)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.04320 -0.66329  0.09895  0.67484  2.55673
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.127186   0.135328  23.108   <2e-16 ***
## x2           7.004392   0.073672  95.075   <2e-16 ***
## x1           4.813138   0.141772  33.950   <2e-16 ***
## x4           0.324186   0.144495   2.244    0.026 *
## x5          -0.005893   0.003246  -1.816    0.071 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.996 on 195 degrees of freedom
## Multiple R-squared:  0.9805, Adjusted R-squared:  0.9801
## F-statistic:  2445 on 4 and 195 DF,  p-value: < 2.2e-16
```

```
backward_model<-stepAIC(model2, direction = "backward", trace = FALSE)

summary(backward_model)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x4 + x5)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.04320 -0.66329  0.09895  0.67484  2.55673
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.127186   0.135328  23.108   <2e-16 ***
## x1           4.813138   0.141772  33.950   <2e-16 ***
## x2           7.004392   0.073672  95.075   <2e-16 ***
## x4           0.324186   0.144495   2.244    0.026 *
## x5          -0.005893   0.003246  -1.816    0.071 .
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.996 on 195 degrees of freedom
## Multiple R-squared:  0.9805, Adjusted R-squared:  0.9801
## F-statistic:  2445 on 4 and 195 DF,  p-value: < 2.2e-16
```

```
full_model_summary<-summary(model2)
reduced_model_summary<-summary(model1)
forward_model_summary<-summary(forward_model)
backward_model_summary<-summary(backward_model)


table<-data.frame(
  Model = c("Full Model", "Reduced Model", "Forward Model", "Backward Model"),
  `Adj R^2`= c(full_model_summary$adj.r.squared, reduced_model_summary$adj.r.squared, forward_model_summ
  )
)

print(table)
```

```
##            Model   Adj.R.2
## 1     Full Model 0.9801019
## 2  Reduced Model 0.9793319
## 3  Forward Model 0.9800504
## 4 Backward Model 0.9800504
```

Based on the result, we can see that the both forward and backward selection chose the best model as $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 + \beta_5 x_5 + \epsilon$. After comparing all the model with adjusted R^2, we can see that the best model is the full model.

```
final_model<-lm(y~x1+x2+x4+x5)
confint(final_model)
```

```
##                   2.5 %       97.5 %
## (Intercept)  2.86029242 3.3940803328
## x1           4.53353457 5.0927421211
## x2           6.85909490 7.1496888056
## x4           0.03921217 0.6091598073
## x5          -0.01229458 0.0005079766
```

Therefore, at the end, we arrived at the little different model without x3, and it was shown by both hypothesis testing and best model selected by adjusted R^2. Therefore, we did not retrieve the true model.

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 + \beta_5 x_5 + \epsilon$

Problem 2 More Observations

Now we'll create the models with 2000 observations. We'll create the same variables as question 1.

```r
x1<-sample(c(0,1), 2000, replace = TRUE)
x2<-rnorm(2000)
y<-3+5*x1+7*x2+rnorm(2000)
model1<-lm(y~x1+x2)
summary(model1)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3264 -0.6847  0.0127  0.6716  3.3762
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.93734    0.03150   93.25   <2e-16 ***
## x1           5.12382    0.04461  114.86   <2e-16 ***
## x2           7.01215    0.02278  307.86   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.997 on 1997 degrees of freedom
## Multiple R-squared:  0.9822, Adjusted R-squared:  0.9822
## F-statistic: 5.518e+04 on 2 and 1997 DF,  p-value: < 2.2e-16
```

```r
x3<-rexp(2000,1/20)
x4<-sample(c(0,1), prob = c(0.6,0.4), replace = T, size = 2000)
x5<-x1+x3+rnorm(2000)
model2<-lm(y~x1+x2+x3+x4+x5)
summary(model2)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4 + x5)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3161 -0.6843  0.0135  0.6737  3.3536
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.92979    0.04250  68.930   <2e-16 ***
## x1           5.14089    0.05002 102.776   <2e-16 ***
## x2           7.01171    0.02281 307.412   <2e-16 ***
## x3           0.01765    0.02275   0.776    0.438
## x4           0.01112    0.04521   0.246    0.806
## x5          -0.01753    0.02274  -0.771    0.441
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9976 on 1994 degrees of freedom
```

```
## Multiple R-squared:  0.9822, Adjusted R-squared:  0.9822
## F-statistic: 2.205e+04 on 5 and 1994 DF,  p-value: < 2.2e-16
```

```r
anova(model1, model2)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x1 + x2
## Model 2: y ~ x1 + x2 + x3 + x4 + x5
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   1997 1985.0
## 2   1994 1984.3  3   0.66443 0.2226 0.8808
```

We have the same hypothesis as the question 1. We can see that with the increasing observations, we have p-value as higher than 0.05 significantly. Therefore, we will fail to reject the null hypothesis.

```r
initial_model<-lm(y~1)
forward_model<-stepAIC(initial_model, direction = "forward", scope = formula(model2), trace = FALSE)

summary(forward_model)
```

```
##
## Call:
## lm(formula = y ~ x2 + x1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3264 -0.6847  0.0127  0.6716  3.3762
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.93734    0.03150   93.25   <2e-16 ***
## x2           7.01215    0.02278  307.86   <2e-16 ***
## x1           5.12382    0.04461  114.86   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.997 on 1997 degrees of freedom
## Multiple R-squared:  0.9822, Adjusted R-squared:  0.9822
## F-statistic: 5.518e+04 on 2 and 1997 DF,  p-value: < 2.2e-16
```

```r
backward_model<-stepAIC(model2, direction = "backward", trace = FALSE)

summary(backward_model)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3264 -0.6847  0.0127  0.6716  3.3762
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.93734    0.03150   93.25   <2e-16 ***
## x1           5.12382    0.04461  114.86   <2e-16 ***
## x2           7.01215    0.02278  307.86   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.997 on 1997 degrees of freedom
## Multiple R-squared:  0.9822, Adjusted R-squared:  0.9822
## F-statistic: 5.518e+04 on 2 and 1997 DF,  p-value: < 2.2e-16
```

```
full_model_summary<-summary(model2)
reduced_model_summary<-summary(model1)
forward_model_summary<-summary(forward_model)
backward_model_summary<-summary(backward_model)


table<-data.frame(
  Model = c("Full Model", "Reduced Model", "Forward Model", "Backward Model"),
  `Adj R^2`= c(full_model_summary$adj.r.squared, reduced_model_summary$adj.r.squared, forward_model_summ
  )
)

print(table)
```

```
##            Model   Adj.R.2
## 1     Full Model 0.9821881
## 2  Reduced Model 0.9822089
## 3  Forward Model 0.9822089
## 4 Backward Model 0.9822089
```

This time we see that both forward and backward model has the same model with x1 and x2. We can see that the best model is the reduced and backward model with $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$

```
final_model<-lm(y~x1+x2)
confint(final_model)
```

```
##                2.5 %   97.5 %
## (Intercept) 2.875564 2.999109
## x1          5.036336 5.211311
## x2          6.967485 7.056824
```

After increasing the observation, we arrive at different final model. Therefore, we are able to retrieve the truth model with 2000 observations.

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$