Stepwise Linear Regression

Veerasak Kritsanapraphan

15/1/2020

## Stepwise Linear Regression

### Import packages necessary first.

library(MASS)  
library(plyr)  
library(ggplot2)  
library(knitr)

### Prepare data

# Assign better variable names  
colnames(birthwt) <- c("birthwt.below.2500", "mother.age", "mother.weight",   
 "race", "mother.smokes", "previous.prem.labor", "hypertension", "uterine.irr",   
 "physician.visits", "birthwt.grams")  
  
# Assign better labels to categorical variables  
birthwt <- transform(birthwt,   
 race = as.factor(mapvalues(race, c(1, 2, 3),   
 c("white","black", "other"))),  
 mother.smokes = as.factor(mapvalues(mother.smokes,   
 c(0,1), c("no", "yes"))),  
 hypertension = as.factor(mapvalues(hypertension,   
 c(0,1), c("no", "yes"))),  
 uterine.irr = as.factor(mapvalues(uterine.irr,   
 c(0,1), c("no", "yes"))),  
 birthwt.below.2500 = as.factor(mapvalues(birthwt.below.2500,  
 c(0,1), c("no", "yes")))  
 )

### Run Linear regression

formula = birthwt.grams ~ mother.age + mother.weight + physician.visits + mother.smokes + uterine.irr + previous.prem.labor + hypertension  
fullmod = lm(formula, data = birthwt)  
summary(fullmod)

##   
## Call:  
## lm(formula = formula, data = birthwt)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1689.4 -467.6 -7.6 464.5 1750.6   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2508.467 294.477 8.518 6.15e-15 \*\*\*  
## mother.age 4.745 9.721 0.488 0.626019   
## mother.weight 4.272 1.725 2.477 0.014172 \*   
## physician.visits -8.012 48.113 -0.167 0.867923   
## mother.smokesyes -228.486 102.506 -2.229 0.027044 \*   
## uterine.irryes -527.097 143.887 -3.663 0.000327 \*\*\*  
## previous.prem.labor -71.709 105.515 -0.680 0.497619   
## hypertensionyes -642.048 209.323 -3.067 0.002492 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 674.2 on 181 degrees of freedom  
## Multiple R-squared: 0.1769, Adjusted R-squared: 0.1451   
## F-statistic: 5.559 on 7 and 181 DF, p-value: 8.177e-06

### No independent variables

nothing <- lm(birthwt.grams ~ 1, data = birthwt)  
summary(nothing)

##   
## Call:  
## lm(formula = birthwt.grams ~ 1, data = birthwt)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2235.59 -530.59 32.41 542.41 2045.41   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2944.59 53.04 55.51 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 729.2 on 188 degrees of freedom

### Stepwise Linear Regression (“Backward”)

backwards = step(fullmod) # Backwards selection is the default

## Start: AIC=2469.96  
## birthwt.grams ~ mother.age + mother.weight + physician.visits +   
## mother.smokes + uterine.irr + previous.prem.labor + hypertension  
##   
## Df Sum of Sq RSS AIC  
## - physician.visits 1 12607 82293521 2468.0  
## - mother.age 1 108335 82389248 2468.2  
## - previous.prem.labor 1 209962 82490875 2468.4  
## <none> 82280914 2470.0  
## - mother.smokes 1 2258613 84539526 2473.1  
## - mother.weight 1 2788878 85069791 2474.3  
## - hypertension 1 4276845 86557758 2477.5  
## - uterine.irr 1 6100389 88381303 2481.5  
##   
## Step: AIC=2467.99  
## birthwt.grams ~ mother.age + mother.weight + mother.smokes +   
## uterine.irr + previous.prem.labor + hypertension  
##   
## Df Sum of Sq RSS AIC  
## - mother.age 1 98286 82391807 2466.2  
## - previous.prem.labor 1 206657 82500178 2466.5  
## <none> 82293521 2468.0  
## - mother.smokes 1 2257167 84550688 2471.1  
## - mother.weight 1 2783045 85076566 2472.3  
## - hypertension 1 4273665 86567186 2475.6  
## - uterine.irr 1 6089392 88382912 2479.5  
##   
## Step: AIC=2466.21  
## birthwt.grams ~ mother.weight + mother.smokes + uterine.irr +   
## previous.prem.labor + hypertension  
##   
## Df Sum of Sq RSS AIC  
## - previous.prem.labor 1 175821 82567628 2464.6  
## <none> 82391807 2466.2  
## - mother.smokes 1 2314920 84706727 2469.4  
## - mother.weight 1 3103917 85495724 2471.2  
## - hypertension 1 4383189 86774996 2474.0  
## - uterine.irr 1 6244871 88636678 2478.0  
##   
## Step: AIC=2464.62  
## birthwt.grams ~ mother.weight + mother.smokes + uterine.irr +   
## hypertension  
##   
## Df Sum of Sq RSS AIC  
## <none> 82567628 2464.6  
## - mother.smokes 1 2623742 85191369 2468.5  
## - mother.weight 1 3306358 85873986 2470.0  
## - hypertension 1 4441221 87008848 2472.5  
## - uterine.irr 1 6979875 89547503 2477.9

formula(backwards)

## birthwt.grams ~ mother.weight + mother.smokes + uterine.irr +   
## hypertension

summary(backwards)

##   
## Call:  
## lm(formula = birthwt.grams ~ mother.weight + mother.smokes +   
## uterine.irr + hypertension, data = birthwt)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1665.02 -452.62 11.16 473.64 1858.65   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2577.096 226.794 11.363 < 2e-16 \*\*\*  
## mother.weight 4.506 1.660 2.714 0.007270 \*\*   
## mother.smokesyes -242.113 100.127 -2.418 0.016579 \*   
## uterine.irryes -549.878 139.424 -3.944 0.000114 \*\*\*  
## hypertensionyes -649.098 206.327 -3.146 0.001931 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 669.9 on 184 degrees of freedom  
## Multiple R-squared: 0.1741, Adjusted R-squared: 0.1561   
## F-statistic: 9.695 on 4 and 184 DF, p-value: 3.886e-07

### Stepwise Linear Regression (“Forward”)

forwards <- step(fullmod, direction="forward")

## Start: AIC=2469.96  
## birthwt.grams ~ mother.age + mother.weight + physician.visits +   
## mother.smokes + uterine.irr + previous.prem.labor + hypertension

formula(forwards)

## birthwt.grams ~ mother.age + mother.weight + physician.visits +   
## mother.smokes + uterine.irr + previous.prem.labor + hypertension

summary(forwards)

##   
## Call:  
## lm(formula = birthwt.grams ~ mother.age + mother.weight + physician.visits +   
## mother.smokes + uterine.irr + previous.prem.labor + hypertension,   
## data = birthwt)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1689.4 -467.6 -7.6 464.5 1750.6   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2508.467 294.477 8.518 6.15e-15 \*\*\*  
## mother.age 4.745 9.721 0.488 0.626019   
## mother.weight 4.272 1.725 2.477 0.014172 \*   
## physician.visits -8.012 48.113 -0.167 0.867923   
## mother.smokesyes -228.486 102.506 -2.229 0.027044 \*   
## uterine.irryes -527.097 143.887 -3.663 0.000327 \*\*\*  
## previous.prem.labor -71.709 105.515 -0.680 0.497619   
## hypertensionyes -642.048 209.323 -3.067 0.002492 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 674.2 on 181 degrees of freedom  
## Multiple R-squared: 0.1769, Adjusted R-squared: 0.1451   
## F-statistic: 5.559 on 7 and 181 DF, p-value: 8.177e-06

### Stepwise Linear Regression (“Both”)

both <- step(fullmod, direction="both")

## Start: AIC=2469.96  
## birthwt.grams ~ mother.age + mother.weight + physician.visits +   
## mother.smokes + uterine.irr + previous.prem.labor + hypertension  
##   
## Df Sum of Sq RSS AIC  
## - physician.visits 1 12607 82293521 2468.0  
## - mother.age 1 108335 82389248 2468.2  
## - previous.prem.labor 1 209962 82490875 2468.4  
## <none> 82280914 2470.0  
## - mother.smokes 1 2258613 84539526 2473.1  
## - mother.weight 1 2788878 85069791 2474.3  
## - hypertension 1 4276845 86557758 2477.5  
## - uterine.irr 1 6100389 88381303 2481.5  
##   
## Step: AIC=2467.99  
## birthwt.grams ~ mother.age + mother.weight + mother.smokes +   
## uterine.irr + previous.prem.labor + hypertension  
##   
## Df Sum of Sq RSS AIC  
## - mother.age 1 98286 82391807 2466.2  
## - previous.prem.labor 1 206657 82500178 2466.5  
## <none> 82293521 2468.0  
## + physician.visits 1 12607 82280914 2470.0  
## - mother.smokes 1 2257167 84550688 2471.1  
## - mother.weight 1 2783045 85076566 2472.3  
## - hypertension 1 4273665 86567186 2475.6  
## - uterine.irr 1 6089392 88382912 2479.5  
##   
## Step: AIC=2466.21  
## birthwt.grams ~ mother.weight + mother.smokes + uterine.irr +   
## previous.prem.labor + hypertension  
##   
## Df Sum of Sq RSS AIC  
## - previous.prem.labor 1 175821 82567628 2464.6  
## <none> 82391807 2466.2  
## + mother.age 1 98286 82293521 2468.0  
## + physician.visits 1 2559 82389248 2468.2  
## - mother.smokes 1 2314920 84706727 2469.4  
## - mother.weight 1 3103917 85495724 2471.2  
## - hypertension 1 4383189 86774996 2474.0  
## - uterine.irr 1 6244871 88636678 2478.0  
##   
## Step: AIC=2464.62  
## birthwt.grams ~ mother.weight + mother.smokes + uterine.irr +   
## hypertension  
##   
## Df Sum of Sq RSS AIC  
## <none> 82567628 2464.6  
## + previous.prem.labor 1 175821 82391807 2466.2  
## + mother.age 1 67449 82500178 2466.5  
## + physician.visits 1 2121 82565507 2466.6  
## - mother.smokes 1 2623742 85191369 2468.5  
## - mother.weight 1 3306358 85873986 2470.0  
## - hypertension 1 4441221 87008848 2472.5  
## - uterine.irr 1 6979875 89547503 2477.9

formula(both)

## birthwt.grams ~ mother.weight + mother.smokes + uterine.irr +   
## hypertension

summary(both)

##   
## Call:  
## lm(formula = birthwt.grams ~ mother.weight + mother.smokes +   
## uterine.irr + hypertension, data = birthwt)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1665.02 -452.62 11.16 473.64 1858.65   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2577.096 226.794 11.363 < 2e-16 \*\*\*  
## mother.weight 4.506 1.660 2.714 0.007270 \*\*   
## mother.smokesyes -242.113 100.127 -2.418 0.016579 \*   
## uterine.irryes -549.878 139.424 -3.944 0.000114 \*\*\*  
## hypertensionyes -649.098 206.327 -3.146 0.001931 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 669.9 on 184 degrees of freedom  
## Multiple R-squared: 0.1741, Adjusted R-squared: 0.1561   
## F-statistic: 9.695 on 4 and 184 DF, p-value: 3.886e-07