Stats 500: HW #6

Diana Liang

12/5/2019

# Part 1: Model Performance of fat data

## 1. Model A: Linear regression with all predictors

Here is the summary for the basic linear regression:

##   
## Call:  
## lm(formula = siri ~ ., data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.3285 -2.9442 -0.1046 2.9091 9.6650   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -19.82090 17.98296 -1.102 0.27162   
## age 0.06717 0.03409 1.970 0.05013 .   
## weight -0.09557 0.05561 -1.718 0.08718 .   
## height -0.04456 0.11226 -0.397 0.69183   
## adipos -0.04914 0.31640 -0.155 0.87673   
## neck -0.43798 0.24846 -1.763 0.07937 .   
## chest -0.08242 0.10944 -0.753 0.45219   
## abdom 1.03016 0.09780 10.533 < 2e-16 \*\*\*  
## hip -0.20410 0.15574 -1.311 0.19144   
## thigh 0.25359 0.15187 1.670 0.09644 .   
## knee 0.02971 0.26088 0.114 0.90944   
## ankle 0.15723 0.22680 0.693 0.48891   
## biceps 0.18965 0.18024 1.052 0.29391   
## forearm 0.46766 0.20384 2.294 0.02275 \*   
## wrist -1.74316 0.56008 -3.112 0.00211 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.324 on 212 degrees of freedom  
## Multiple R-squared: 0.7591, Adjusted R-squared: 0.7432   
## F-statistic: 47.71 on 14 and 212 DF, p-value: < 2.2e-16

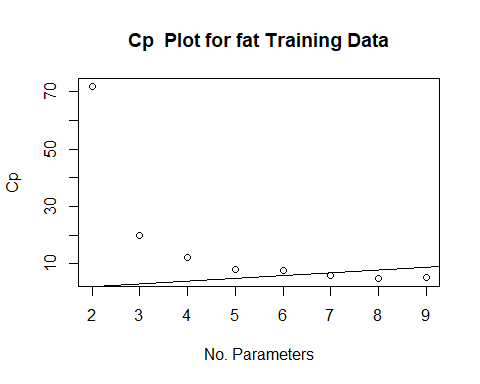
The RMSE for model A was **4.396**. This value will be used as a baseline to compare to that of other regression models.

## 2. Model B: Linear regression with Mallows Cp

First, the predictor variables need to be chosen by Mallows Cp.

Here are the possible combinations:

## Subset selection object  
## Call: regsubsets.formula(siri ~ ., train)  
## 14 Variables (and intercept)  
## Forced in Forced out  
## age FALSE FALSE  
## weight FALSE FALSE  
## height FALSE FALSE  
## adipos FALSE FALSE  
## neck FALSE FALSE  
## chest FALSE FALSE  
## abdom FALSE FALSE  
## hip FALSE FALSE  
## thigh FALSE FALSE  
## knee FALSE FALSE  
## ankle FALSE FALSE  
## biceps FALSE FALSE  
## forearm FALSE FALSE  
## wrist FALSE FALSE  
## 1 subsets of each size up to 8  
## Selection Algorithm: exhaustive  
## age weight height adipos neck chest abdom hip thigh knee ankle  
## 1 ( 1 ) " " " " " " " " " " " " "\*" " " " " " " " "   
## 2 ( 1 ) " " "\*" " " " " " " " " "\*" " " " " " " " "   
## 3 ( 1 ) " " "\*" " " " " " " " " "\*" " " " " " " " "   
## 4 ( 1 ) " " "\*" " " " " " " " " "\*" " " " " " " " "   
## 5 ( 1 ) " " "\*" " " " " "\*" " " "\*" " " " " " " " "   
## 6 ( 1 ) "\*" "\*" " " " " " " " " "\*" " " "\*" " " " "   
## 7 ( 1 ) "\*" "\*" " " " " "\*" " " "\*" " " "\*" " " " "   
## 8 ( 1 ) "\*" "\*" " " " " "\*" " " "\*" "\*" "\*" " " " "   
## biceps forearm wrist  
## 1 ( 1 ) " " " " " "   
## 2 ( 1 ) " " " " " "   
## 3 ( 1 ) " " " " "\*"   
## 4 ( 1 ) " " "\*" "\*"   
## 5 ( 1 ) " " "\*" "\*"   
## 6 ( 1 ) " " "\*" "\*"   
## 7 ( 1 ) " " "\*" "\*"   
## 8 ( 1 ) " " "\*" "\*"



The Mallows Cp method suggests 7 variables: age, weight, neck, abdom, thigh, forearm, and wrist. Since this selection has a Cp below the *p+1* line, as shown in the plot, that selection will be used as the predictor variables for linear regression.

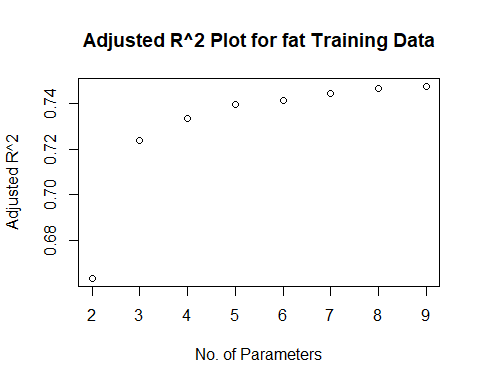
Here is that summary:

##   
## Call:  
## lm(formula = siri ~ age + weight + neck + abdom + thigh + forearm +   
## wrist, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.172 -3.125 -0.264 3.089 9.315   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -33.79207 9.43053 -3.583 0.000418 \*\*\*  
## age 0.07180 0.03200 2.243 0.025871 \*   
## weight -0.12792 0.03548 -3.606 0.000385 \*\*\*  
## neck -0.39624 0.23121 -1.714 0.087978 .   
## abdom 0.94869 0.07430 12.768 < 2e-16 \*\*\*  
## thigh 0.24222 0.11828 2.048 0.041776 \*   
## forearm 0.53976 0.18906 2.855 0.004718 \*\*   
## wrist -1.63732 0.53368 -3.068 0.002427 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.294 on 219 degrees of freedom  
## Multiple R-squared: 0.7546, Adjusted R-squared: 0.7467   
## F-statistic: 96.18 on 7 and 219 DF, p-value: < 2.2e-16

The RMSE for model B was **4.342**, which is incredibly similar to model A’s 4.396. So this new model performs just as well as the regular linear regression model.

## 3. Model C: Linear regression with Adjusted

The predictor variables need to be chosen by adjusted . The same combinations will be used as for the Mallows Cp.



As shown in the above plot, adjusted suggests 8 variables: age, weight, neck, abdom, hip, thigh, forearm, and wrist. These will be used as predictor variables in a linear regression model.

##   
## Call:  
## lm(formula = siri ~ age + weight + neck + abdom + hip + thigh +   
## forearm + wrist, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.2181 -2.8832 -0.1985 2.8211 9.8197   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -23.71280 12.11193 -1.958 0.05153 .   
## age 0.07011 0.03197 2.193 0.02938 \*   
## weight -0.09992 0.04126 -2.422 0.01625 \*   
## neck -0.46280 0.23623 -1.959 0.05138 .   
## abdom 0.97661 0.07712 12.664 < 2e-16 \*\*\*  
## hip -0.19051 0.14403 -1.323 0.18732   
## thigh 0.32262 0.13281 2.429 0.01594 \*   
## forearm 0.50778 0.19028 2.669 0.00819 \*\*   
## wrist -1.63149 0.53279 -3.062 0.00247 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.287 on 218 degrees of freedom  
## Multiple R-squared: 0.7565, Adjusted R-squared: 0.7476   
## F-statistic: 84.66 on 8 and 218 DF, p-value: < 2.2e-16

The RMSE for model C was **4.327**, which is again very similar to model A’s 4.396. So again this new model performs just as well as the last two.

## 4. Model D: Ridge Regression

Ridge regression requires standardized predictors and a for the penalty term.

CV is the most common tool to determine , and it returned = 0.6941839. Now this parameter can used for the ridge regression model.

Here is a list of the estimates:

## 14 x 1 sparse Matrix of class "dgCMatrix"  
## s0  
## age 1.468942871  
## weight -0.399341332  
## height -0.357054409  
## adipos 1.039787023  
## neck -0.860939776  
## chest 0.590447855  
## abdom 5.729511922  
## hip -0.111379579  
## thigh 0.973207899  
## knee 0.007818152  
## ankle -0.098947700  
## biceps 0.179696762  
## forearm 0.630775596  
## wrist -1.729168120

The RMSE for model D was **3.595**, which is smaller than all the previous RMSE’s: 4.396 for model A, 4.342 for model B, and 4.327 for model C.

# Part 2: Binomial Regression Model for pima

All values set at 0 are removed before modeling, so here is the summary for that binomial model:

##   
## Call:  
## glm(formula = cbind(test, 1 - test) ~ pregnant + glucose + diastolic +   
## bmi + diabetes + age, family = binomial, data = final)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.8459 -0.7067 -0.3827 0.7018 2.4302   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -9.354750 0.915697 -10.216 < 2e-16 \*\*\*  
## pregnant 0.130695 0.037880 3.450 0.00056 \*\*\*  
## glucose 0.035337 0.003900 9.061 < 2e-16 \*\*\*  
## diastolic -0.008673 0.009422 -0.920 0.35734   
## bmi 0.098547 0.017768 5.546 2.92e-08 \*\*\*  
## diabetes 1.020669 0.336136 3.036 0.00239 \*\*   
## age 0.016642 0.010553 1.577 0.11478   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 807.12 on 624 degrees of freedom  
## Residual deviance: 577.80 on 618 degrees of freedom  
## AIC: 591.8  
##   
## Number of Fisher Scoring iterations: 5

## 1. Deviance

Deviance cannot be used on binary data, which **test** is. So deviance should not be used as a measure of goodness-of-fit.

## 2. Odds Ratio for BMI

The summary of **bmi** will provide a quick view of the quartiles:

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 18.20 27.40 32.00 32.02 35.90 57.30

The ratio of odds for third quartile of **bmi** - first quartile of **bmi** is 2.311, which is the ratio of odds ratio of the third quartile BMI over the odds ratio of the first quartile BMI.

## 3. Confound Controlling for Diastolic Blood Pressure

Here is the model summary with **diastolic** as a predictor varible:

##   
## Call:  
## glm(formula = cbind(test, 1 - test) ~ diastolic, family = binomial,   
## data = final)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.6036 -0.9375 -0.7971 1.3226 2.0717   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.994720 0.536796 -5.579 2.42e-08 \*\*\*  
## diastolic 0.032434 0.007211 4.498 6.87e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 807.12 on 624 degrees of freedom  
## Residual deviance: 785.65 on 623 degrees of freedom  
## AIC: 789.65  
##   
## Number of Fisher Scoring iterations: 4

The coefficient is positive in this model but is negative in the original model. While this may seem like a contradiction, this new model does not control for other predictor variables that may be confounding the relationship between **test** and **diastolic**. Having more predictor variables lessens confounding by controlling for those predictor variables, providing a more accurate relationship.

## 4. Predict Probability

The probability can be found through the predicted odds ratio. The predicted odds ratio is the exponent of the predicted value. All of which is shown here:

pred <- exp(coef[1]+ coef[2]\*preg +coef[3]\*glu + coef[4]\*dias + coef[5]\*bmI + coef[6]\*diab + coef[7]\*agE)  
  
prob <- pred/(1+pred)

The probabiliy of testing positive (given **pregnant** = 1, **glucose** = 100, **diastolic** = 70, **bmi** = 25, **diabetes** = 0.6, and **age** = 30) is **0.062**.