**Here is the output of the R file:**

|  |
| --- |
| > #an example of cross validation  >  > if (FALSE)  + {"  + Measuring body fat is not so simple. Hydrostatic underwater weighing is a method of determining body composition (body fat to lean mass).  + Obtain a person's total body density by submerging the body underwater in a tank and measuring the displacement.  + Considered the gold standard for body composition assessment. More sophisticated methods may make underwater weighing obsolete in the near future.  + n=252 men - Brozek's equation was applied to each man's hydrostatic underwater weighing results to accurately estimate their percentage of body fat.  + "}  >  > library(faraway) #this command brings in a library of regression functions  > data(fat,package="faraway")  > #Can we predict body fat using only easy-to-record measurements?  >  > # The full model should include all the candidate predictor variables.  > head(fat$brozek,3L)  [1] 12.6 6.9 24.6  >  > #full model  > lmod <- lm(brozek ~ age + weight + height + neck + chest + abdom + hip + thigh + knee + ankle + biceps + forearm + wrist, data=fat)  > summary(lmod)  Call:  lm(formula = brozek ~ age + weight + height + neck + chest +  abdom + hip + thigh + knee + ankle + biceps + forearm + wrist,  data = fat)  Residuals:  Min 1Q Median 3Q Max  -10.264 -2.572 -0.097 2.898 9.327  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) -15.29255 16.06992 -0.952 0.34225  age 0.05679 0.02996 1.895 0.05929 .  weight -0.08031 0.04958 -1.620 0.10660  height -0.06460 0.08893 -0.726 0.46830  neck -0.43754 0.21533 -2.032 0.04327 \*  chest -0.02360 0.09184 -0.257 0.79740  abdom 0.88543 0.08008 11.057 < 2e-16 \*\*\*  hip -0.19842 0.13516 -1.468 0.14341  thigh 0.23190 0.13372 1.734 0.08418 .  knee -0.01168 0.22414 -0.052 0.95850  ankle 0.16354 0.20514 0.797 0.42614  biceps 0.15280 0.15851 0.964 0.33605  forearm 0.43049 0.18445 2.334 0.02044 \*  wrist -1.47654 0.49552 -2.980 0.00318 \*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 3.988 on 238 degrees of freedom  Multiple R-squared: 0.749, Adjusted R-squared: 0.7353  F-statistic: 54.63 on 13 and 238 DF, p-value: < 2.2e-16  >  > #install.packages("olsrr") #install olsrr if package has not been installed on your computer  > library(olsrr)  >  > #If details is set to TRUE, each step is displayed.  > both <- ols\_step\_both\_p(lmod,details=FALSE)  > both  Stepwise Selection Summary  --------------------------------------------------------------------------------------  Added/ Adj.  Step Variable Removed R-Square R-Square C(p) AIC RMSE  --------------------------------------------------------------------------------------  1 abdom addition 0.662 0.661 72.3570 1478.8012 4.5144  2 weight addition 0.719 0.716 20.6840 1434.5921 4.1272  3 wrist addition 0.728 0.724 14.3120 1428.5544 4.0700  4 forearm addition 0.735 0.731 9.1520 1423.4707 4.0214  --------------------------------------------------------------------------------------  >  > back <- ols\_step\_backward\_p(lmod,details=FALSE)  > back  Elimination Summary  --------------------------------------------------------------------------  Variable Adj.  Step Removed R-Square R-Square C(p) AIC RMSE  --------------------------------------------------------------------------  1 knee 0.749 0.7364 12.0027 1425.9187 3.9796  2 chest 0.7489 0.7374 10.0669 1423.9867 3.9719  3 height 0.7484 0.738 8.5494 1422.4968 3.9676  4 ankle 0.7477 0.7383 7.2537 1421.2398 3.9653  5 biceps 0.7467 0.7383 6.1913 1420.2254 3.9649  --------------------------------------------------------------------------  > ##-------------------------------------------------------------------------------------##  >  > #model chosen by sequential stepwise selection  > both <- lm(brozek ~ abdom + weight + wrist + forearm, data=fat)  > summary(both)  Call:  lm(formula = brozek ~ abdom + weight + wrist + forearm, data = fat)  Residuals:  Min 1Q Median 3Q Max  -9.8002 -2.8728 -0.1545 2.8980 8.3845  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) -31.29679 6.70886 -4.665 5.06e-06 \*\*\*  abdom 0.92137 0.05192 17.747 < 2e-16 \*\*\*  weight -0.12557 0.02292 -5.479 1.05e-07 \*\*\*  wrist -1.39177 0.40991 -3.395 0.000799 \*\*\*  forearm 0.44638 0.16822 2.654 0.008480 \*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 4.021 on 247 degrees of freedom  Multiple R-squared: 0.7351, Adjusted R-squared: 0.7308  F-statistic: 171.4 on 4 and 247 DF, p-value: < 2.2e-16  >  > #model chosen by backward elimination  > backward <- lm(brozek ~ age + weight + neck + abdom + hip + thigh + forearm + wrist, data=fat)  > summary(backward)  Call:  lm(formula = brozek ~ age + weight + neck + abdom + hip + thigh +  forearm + wrist, data = fat)  Residuals:  Min 1Q Median 3Q Max  -10.0574 -2.7411 -0.1912 2.6929 9.4977  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) -20.06213 10.84654 -1.850 0.06558 .  age 0.05922 0.02850 2.078 0.03876 \*  weight -0.08414 0.03695 -2.277 0.02366 \*  neck -0.43189 0.20799 -2.077 0.03889 \*  abdom 0.87721 0.06661 13.170 < 2e-16 \*\*\*  hip -0.18641 0.12821 -1.454 0.14727  thigh 0.28644 0.11949 2.397 0.01727 \*  forearm 0.48255 0.17251 2.797 0.00557 \*\*  wrist -1.40487 0.47167 -2.978 0.00319 \*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 3.965 on 243 degrees of freedom  Multiple R-squared: 0.7467, Adjusted R-squared: 0.7383  F-statistic: 89.53 on 8 and 243 DF, p-value: < 2.2e-16  >  > ##-------------------------------------------------------------------------------------##  >  > #Start the cross-validation to choose between the model chosen by sequential stepwise selection  > # and the model chosen by backward elimination  >  > #install.packages("caret") #install this package if needed  > require(caret)  Loading required package: caret  Loading required package: lattice  Attaching package: ‘lattice’  The following object is masked from ‘package:faraway’:  melanoma  Loading required package: ggplot2  >  > # Define training control  > set.seed(13245)  > train.control <- trainControl(method = "cv", number = 10)  > # Train the model  > model\_both <- train(brozek ~ abdom + weight + wrist + forearm,data = fat, method = "lm",  + trControl = train.control)  > # Summarize the results  > print(model\_both)  Linear Regression  252 samples  4 predictor  No pre-processing  Resampling: Cross-Validated (10 fold)  Summary of sample sizes: 226, 227, 227, 227, 227, 226, ...  Resampling results:  RMSE Rsquared MAE  4.092029 0.7409491 3.378596  Tuning parameter 'intercept' was held constant at a value of TRUE  > ##  >  > # Define training control  > set.seed(14235)  > train.control <- trainControl(method = "cv", number = 10)  > # Train the model  > model\_backward <- train(brozek ~ age + weight + neck + abdom + hip + thigh + forearm + wrist,data = fat, method = "lm",  + trControl = train.control)  > # Summarize the results  > print(model\_backward)  Linear Regression  252 samples  8 predictor  No pre-processing  Resampling: Cross-Validated (10 fold)  Summary of sample sizes: 226, 227, 225, 227, 228, 226, ...  Resampling results:  RMSE Rsquared MAE  4.004975 0.7334675 3.30121  Tuning parameter 'intercept' was held constant at a value of TRUE  > ##  > summary(both)  Call:  lm(formula = brozek ~ abdom + weight + wrist + forearm, data = fat)  Residuals:  Min 1Q Median 3Q Max  -9.8002 -2.8728 -0.1545 2.8980 8.3845  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) -31.29679 6.70886 -4.665 5.06e-06 \*\*\*  abdom 0.92137 0.05192 17.747 < 2e-16 \*\*\*  weight -0.12557 0.02292 -5.479 1.05e-07 \*\*\*  wrist -1.39177 0.40991 -3.395 0.000799 \*\*\*  forearm 0.44638 0.16822 2.654 0.008480 \*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 4.021 on 247 degrees of freedom  Multiple R-squared: 0.7351, Adjusted R-squared: 0.7308  F-statistic: 171.4 on 4 and 247 DF, p-value: < 2.2e-16  > summary(backward)  Call:  lm(formula = brozek ~ age + weight + neck + abdom + hip + thigh +  forearm + wrist, data = fat)  Residuals:  Min 1Q Median 3Q Max  -10.0574 -2.7411 -0.1912 2.6929 9.4977  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) -20.06213 10.84654 -1.850 0.06558 .  age 0.05922 0.02850 2.078 0.03876 \*  weight -0.08414 0.03695 -2.277 0.02366 \*  neck -0.43189 0.20799 -2.077 0.03889 \*  abdom 0.87721 0.06661 13.170 < 2e-16 \*\*\*  hip -0.18641 0.12821 -1.454 0.14727  thigh 0.28644 0.11949 2.397 0.01727 \*  forearm 0.48255 0.17251 2.797 0.00557 \*\*  wrist -1.40487 0.47167 -2.978 0.00319 \*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 3.965 on 243 degrees of freedom  Multiple R-squared: 0.7467, Adjusted R-squared: 0.7383  F-statistic: 89.53 on 8 and 243 DF, p-value: < 2.2e-16  >  > ####End of 10-fold Cross Validation#### |
|  |
| |  | | --- | | > | |

**We will choose a model that has largest R-squared values and the lowest RMSE and MAE values. Looking at the 10-fold regression summaries, we would prefer a model with all 8 variables as the value of RMSE and MAE are lower. Looking at the summary of backward and stepwise regression, we can confirm this analogy by looking at higher values of R-squared and lesser error values.**