**14) This problem focuses on the *collinearity* problem.**

1. Perform the following commands in R:

> set .seed (1)

> x1=runif (100)

> x2 =0.5\* x1+rnorm (100) /10

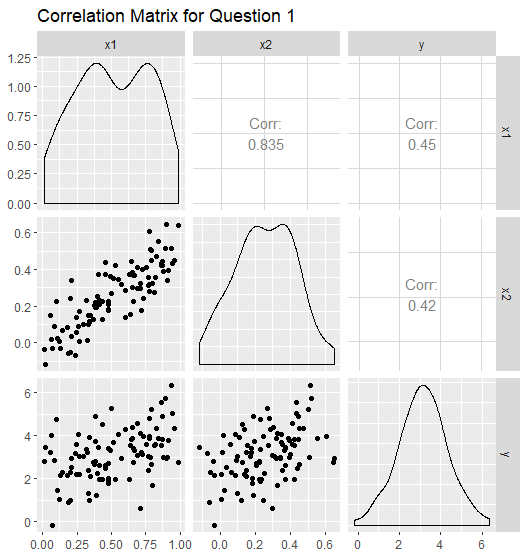
> y=2+2\* x1 +0.3\* x2+rnorm (100)

The last line corresponds to creating a linear model in which y is a function of x1 and x2. Write out the form the linear model. What are the regression coefficients?

The linear model has the form , where

The regression coefficients are

1. What is the correlation between x1 and x2? Create a scatterplot displaying the relationship between the variables.



From the above plot we can see the correlation between is 0.835. This can be described as a strong positive correlation.

1. Using this data, fit a least squares regression to predict y using x1 and x2. Describe the results obtained. What are ˆ *β*0, ˆ *β*1, and ˆ *β*2? How do these relate to the true *β*0, *β*1, and *β*2? Can you reject the null hypothesis *H*0 : *β*1 = 0? How about the null hypothesis *H*0 : *β*2 = 0?

lm(formula = y ~ x1 + x2, data = question\_1)

Residuals:

Min 1Q Median 3Q Max

-2.8311 -0.7273 -0.0537 0.6338 2.3359

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.1305 0.2319 9.188 7.61e-15 \*\*\*

x1 1.4396 0.7212 1.996 0.0487 \*

x2 1.0097 1.1337 0.891 0.3754

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

We have the regression equation with an . The predicted coefficients are . We can see that the predicted coefficients for and are higher than their true value and the predicted coefficient for is lower.

We reject the null hypothesis that at α = 0.05. However, we cannot reject the null hypothesis that at α = 0.05

1. Now fit a least squares regression to predict y using only x1. Comment on your results. Can you reject the null hypothesis *H*0 : *β*1 = 0?

lm(formula = y ~ x1, data = question\_1)

Residuals:

Min 1Q Median 3Q Max

-2.89495 -0.66874 -0.07785 0.59221 2.45560

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.1124 0.2307 9.155 8.27e-15 \*\*\*

x1 1.9759 0.3963 4.986 2.66e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Our regression equation with only is with an . This model has a slightly better fit compared to the model with both . We reject the null hypothesis that at α = 0.05.

1. Now fit a least squares regression to predict y using only x2. Comment on your results. Can you reject the null hypothesis *H*0 : *β*2 = 0?

lm(formula = y ~ x2, data = question\_1)

Residuals:

Min 1Q Median 3Q Max

-2.62687 -0.75156 -0.03598 0.72383 2.44890

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.3899 0.1949 12.26 < 2e-16 \*\*\*

x2 2.8996 0.6330 4.58 1.37e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Our regression equation with only is with an . This model has a worse fit compared to the model with both . We reject the null hypothesis that at α = 0.05.

1. Do the results obtained in (c)–(e) contradict each other? Explain your answer.

Our answers in parts c – e contradict each other. In part c we concluded that , but in part e we concluded that was a significant predictor. This occurred because of the high collinearity between . Their correlation value is 0.835 which indicates a strong positive correlation between the variables. This explains why we were able to remove either of the two variables and still have a significant predictor.

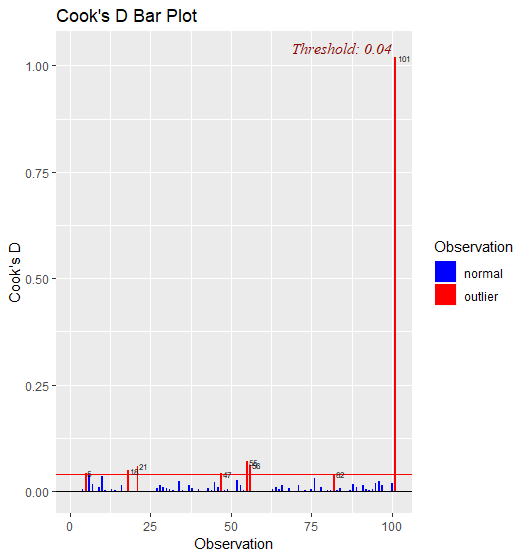
1. Now suppose we obtain one additional observation, which was unfortunately mis-measured.

> x1=c(x1 , 0.1)

> x2=c(x2 , 0.8)

> y=c(y,6)

Re-fit the linear models from (c) to (e) using this new data. What effect does this new observation have on the each of the models? In each model, is this observation an outlier? A high-leverage point? Both? Explain your answers.

lm(formula = y ~ x1 + x2, data = question\_1)

Residuals:

Min 1Q Median 3Q Max

-2.73348 -0.69318 -0.05263 0.66385 2.30619

Coefficients:

Estimate Std. Error t value Pr(>|t|)

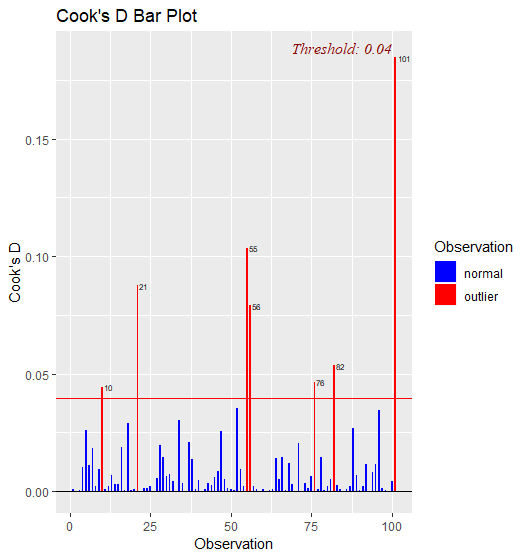
(Intercept) 2.2267 0.2314 9.624 7.91e-16 \*\*\*

x1 0.5394 0.5922 0.911 0.36458

x2 2.5146 0.8977 2.801 0.00614 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

lm(formula = y ~ x1, data = question\_1)

Residuals:

Min 1Q Median 3Q Max

-2.8897 -0.6556 -0.0909 0.5682 3.5665

Coefficients:

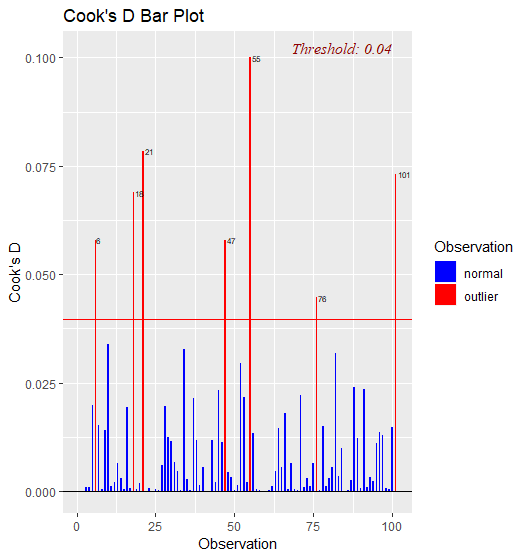
Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.2569 0.2390 9.445 1.78e-15 \*\*\*

x1 1.7657 0.4124 4.282 4.29e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

lm(formula = y ~ x2, data = question\_1)

Residuals:

Min 1Q Median 3Q Max

-2.64729 -0.71021 -0.06899 0.72699 2.38074

Coefficients:

Estimate Std. Error t value Pr(>|t|)

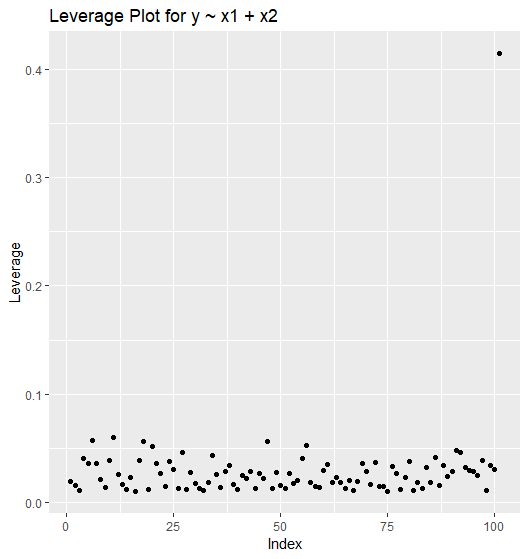
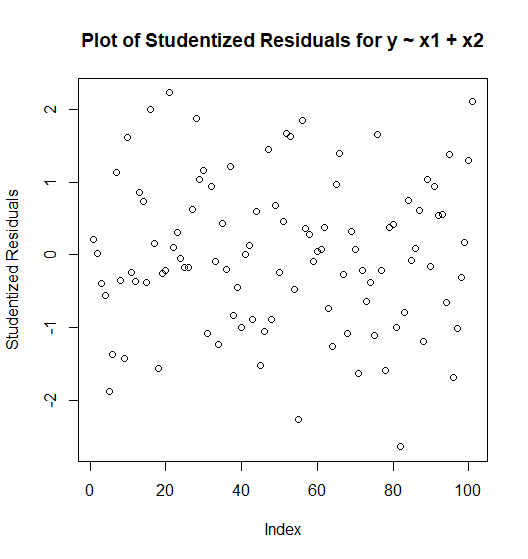
(Intercept) 2.3451 0.1912 12.264 < 2e-16 \*\*\*

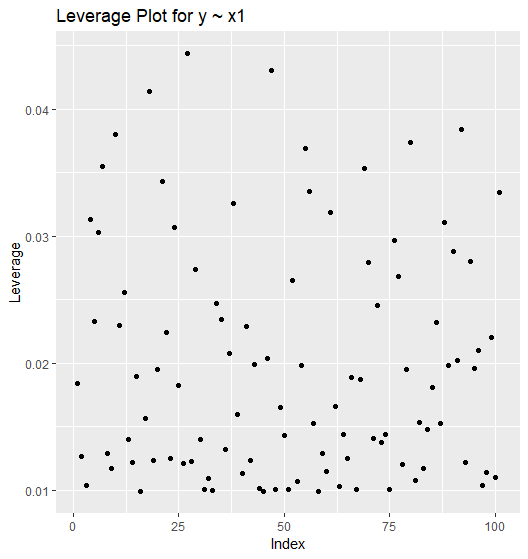
x2 3.1190 0.6040 5.164 1.25e-06 \*\*\*

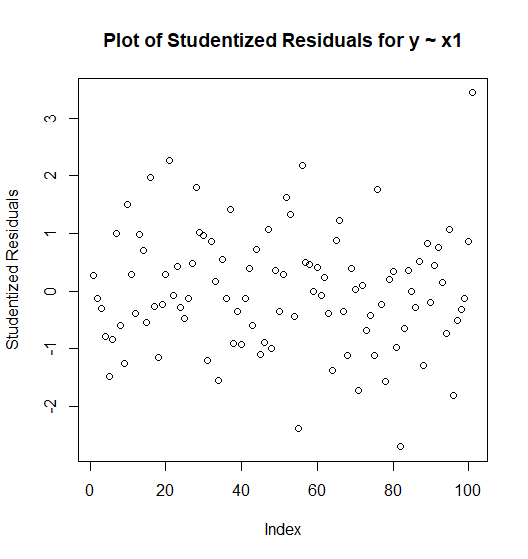
---

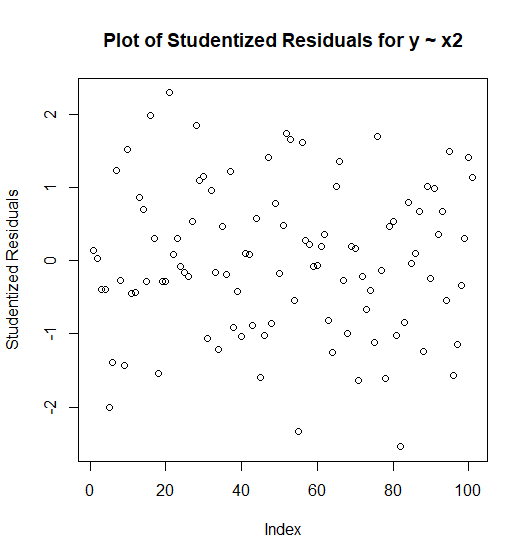
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

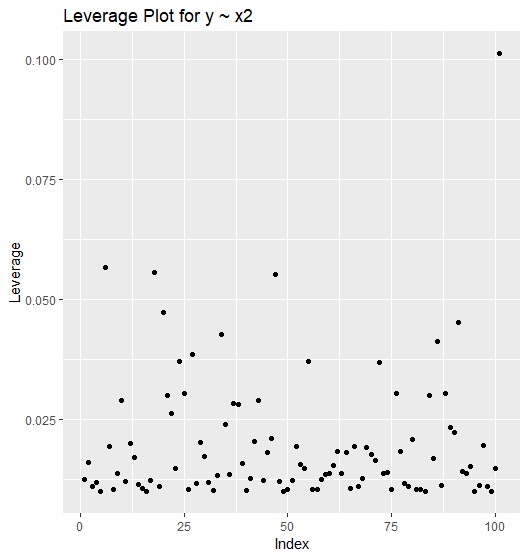
From the above R output and the Cooks Distance bar charts we can see that the new observation (101) is considered an outlier in all our models. Furthermore, the new observation caused to become insignificant in our first model and to become significant. This is the direct opposite to what we observed in part (c). In the below plots of studentized residuals we can also see that observation 101 is an outlier under this criterion for .











From the above leverage plots, we can see that observation 101 has high leverage for . This explains why in the first fitting became significant.

**2) Referring to the Body fat data on *beachboard*, find the optimal model to predict percent body fat.**

**Summary of Findings:**

I started by performing model section based on VIF and P-Value. After 10 fittings I had a model where all the predictors were significant

Call:

lm(formula = percent\_body\_fat ~ . - weight\_lbs - abdomen\_circumference\_cm -

hip\_circumference\_cm - thigh\_circumference\_cm - neck\_circumference\_cm -

chest\_circumference\_cm - wrist\_circumference\_cm - biceps\_extended\_circumference\_cm -

knee\_circumference\_cm, data = Body\_Fat\_2)

Residuals:

Min 1Q Median 3Q Max

-14.957 -5.038 0.597 5.192 23.625

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -26.8446 10.6544 -2.520 0.01238 \*

age\_years 0.2078 0.0364 5.708 3.26e-08 \*\*\*

height\_inches -0.3814 0.1302 -2.929 0.00372 \*\*

ankle\_circumference\_cm 0.9849 0.2989 3.296 0.00113 \*\*

forearm\_circumference\_cm 1.4188 0.2481 5.719 3.09e-08 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.143 on 247 degrees of freedom

Multiple R-squared: 0.283, Adjusted R-squared: 0.2714

F-statistic: 24.37 on 4 and 247 DF, p-value: < 2.2e-16

age\_years height\_inches

1.035119 1.118799

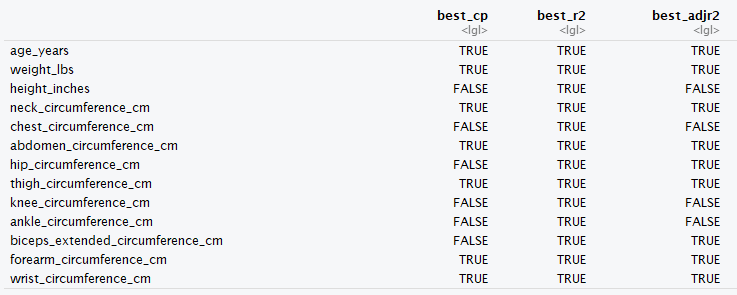
ankle\_circumference\_cm forearm\_circumference\_cm

1.261926 1.236157

Regression Equation

|  |  |  |
| --- | --- | --- |
| Percent body fat | = | -26.8 + 0.2078 Age (years) - 0.381 Height (inches) + 0.985 Ankle circumference (cm) + 1.419 Forearm circumference (cm) |

Next, I performed model selection using CP, and . The goal is to find models that have the lowest CP value and highest and . After running the code, we have the following models



Minimum CP

Regression Equation

|  |  |  |
| --- | --- | --- |
| Percent body fat | = | -33.26 + 0.0682 Age (years) - 0.1194 Weight (lbs) - 0.404 Neck circumference (cm) + 0.9179 Abdomen circumference (cm) + 0.222 Thigh circumference (cm) + 0.553 Forearm circumference (cm) - 1.532 Wrist circumference (cm) |

Model Summary

|  |  |  |  |
| --- | --- | --- | --- |
| S | R-sq | R-sq(adj) | R-sq(pred) |
| 4.29060 | 74.45% | 73.71% | 72.46% |

Max

Regression Equation

|  |  |  |
| --- | --- | --- |
| Percent body fat | = | -18.2 + 0.0621 Age (years) - 0.0884 Weight (lbs) - 0.0696 Height (inches) - 0.471 Neck circumference (cm) - 0.0239 Chest circumference (cm) + 0.9548 Abdomen circumference (cm) - 0.208 Hip circumference (cm) + 0.236 Thigh circumference (cm) + 0.015 Knee circumference (cm) + 0.174 Ankle circumference (cm) + 0.182 Biceps (extended) circumference + 0.452 Forearm circumference (cm) - 1.621 Wrist circumference (cm) |

Model Summary

|  |  |  |  |
| --- | --- | --- | --- |
| S | R-sq | R-sq(adj) | R-sq(pred) |
| 4.30529 | 74.90% | 73.53% | 71.45% |

Max

Regression Equation

|  |  |  |
| --- | --- | --- |
| Percent body fat | = | -23.3 + 0.0635 Age (years) - 0.0984 Weight (lbs) - 0.493 Neck circumference (cm) + 0.9493 Abdomen circumference (cm) - 0.183 Hip circumference (cm) + 0.265 Thigh circumference (cm) + 0.179 Biceps (extended) circumference + 0.451 Forearm circumference (cm) - 1.542 Wrist circumference (cm) |

Model Summary

|  |  |  |  |
| --- | --- | --- | --- |
| S | R-sq | R-sq(adj) | R-sq(pred) |
| 4.28075 | 74.77% | 73.84% | 72.44% |

Lastly, I performed model selection using Stepwise, Forward, and Backward selection with AIC and BIC

**Stepwise AIC model:**

lm(formula = percent\_body\_fat ~ age\_years + weight\_lbs + neck\_circumference\_cm +

abdomen\_circumference\_cm + hip\_circumference\_cm + thigh\_circumference\_cm +

forearm\_circumference\_cm + wrist\_circumference\_cm, data = Body\_Fat\_2)

Residuals:

Min 1Q Median 3Q Max

-10.9757 -2.9937 -0.1644 2.9766 10.2244

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -22.65637 11.71385 -1.934 0.05426 .

age\_years 0.06578 0.03078 2.137 0.03356 \*

weight\_lbs -0.08985 0.03991 -2.252 0.02524 \*

neck\_circumference\_cm -0.46656 0.22462 -2.077 0.03884 \*

abdomen\_circumference\_cm 0.94482 0.07193 13.134 < 2e-16 \*\*\*

hip\_circumference\_cm -0.19543 0.13847 -1.411 0.15940

thigh\_circumference\_cm 0.30239 0.12904 2.343 0.01992 \*

forearm\_circumference\_cm 0.51572 0.18631 2.768 0.00607 \*\*

wrist\_circumference\_cm -1.53665 0.50939 -3.017 0.00283 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.282 on 243 degrees of freedom

Multiple R-squared: 0.7466, Adjusted R-squared: 0.7382

F-statistic: 89.47 on 8 and 243 DF, p-value: < 2.2e-16

Regression Equation

|  |  |  |
| --- | --- | --- |
| Percent body fat | = | -22.7 + 0.0658 Age (years) - 0.0899 Weight (lbs) - 0.467 Neck circumference (cm) + 0.9448 Abdomen circumference (cm) - 0.195 Hip circumference (cm) + 0.302 Thigh circumference (cm) + 0.516 Forearm circumference (cm) - 1.537 Wrist circumference (cm) |

Model Summary

|  |  |  |  |
| --- | --- | --- | --- |
| S | R-sq | R-sq(adj) | R-sq(pred) |
| 4.28190 | 74.66% | 73.82% | 72.52% |

**Stepwise BIC Model:**

lm(formula = percent\_body\_fat ~ weight\_lbs + abdomen\_circumference\_cm +

forearm\_circumference\_cm + wrist\_circumference\_cm, data = Body\_Fat\_2)

Residuals:

Min 1Q Median 3Q Max

-10.5626 -3.1235 -0.1461 3.1313 9.0867

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -34.85407 7.24500 -4.811 2.62e-06 \*\*\*

weight\_lbs -0.13563 0.02475 -5.480 1.05e-07 \*\*\*

abdomen\_circumference\_cm 0.99575 0.05607 17.760 < 2e-16 \*\*\*

forearm\_circumference\_cm 0.47293 0.18166 2.603 0.009790 \*\*

wrist\_circumference\_cm -1.50556 0.44267 -3.401 0.000783 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.343 on 247 degrees of freedom

Multiple R-squared: 0.735, Adjusted R-squared: 0.7307

F-statistic: 171.3 on 4 and 247 DF, p-value: < 2.2e-16

Regression Equation

|  |  |  |
| --- | --- | --- |
| Percent body fat | = | -34.85 - 0.1356 Weight (lbs) + 0.9958 Abdomen circumference (cm) + 0.473 Forearm circumference (cm) - 1.506 Wrist circumference (cm) |

Model Summary

|  |  |  |  |
| --- | --- | --- | --- |
| S | R-sq | R-sq(adj) | R-sq(pred) |
| 4.34272 | 73.50% | 73.07% | 72.08% |

**Forward AIC Model:**

lm(formula = percent\_body\_fat ~ abdomen\_circumference\_cm + weight\_lbs +

wrist\_circumference\_cm + forearm\_circumference\_cm + neck\_circumference\_cm +

age\_years + thigh\_circumference\_cm + hip\_circumference\_cm,

data = Body\_Fat\_2)

Residuals:

Min 1Q Median 3Q Max

-10.9757 -2.9937 -0.1644 2.9766 10.2244

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -22.65637 11.71385 -1.934 0.05426 .

abdomen\_circumference\_cm 0.94482 0.07193 13.134 < 2e-16 \*\*\*

weight\_lbs -0.08985 0.03991 -2.252 0.02524 \*

wrist\_circumference\_cm -1.53665 0.50939 -3.017 0.00283 \*\*

forearm\_circumference\_cm 0.51572 0.18631 2.768 0.00607 \*\*

neck\_circumference\_cm -0.46656 0.22462 -2.077 0.03884 \*

age\_years 0.06578 0.03078 2.137 0.03356 \*

thigh\_circumference\_cm 0.30239 0.12904 2.343 0.01992 \*

hip\_circumference\_cm -0.19543 0.13847 -1.411 0.15940

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.282 on 243 degrees of freedom

Multiple R-squared: 0.7466, Adjusted R-squared: 0.7382

F-statistic: 89.47 on 8 and 243 DF, p-value: < 2.2e-16

Regression Equation

|  |  |  |
| --- | --- | --- |
| Percent body fat | = | -22.7 - 0.0899 Weight (lbs) + 0.9448 Abdomen circumference (cm) + 0.516 Forearm circumference (cm) - 1.537 Wrist circumference (cm) - 0.467 Neck circumference (cm) + 0.0658 Age (years) + 0.302 Thigh circumference (cm) - 0.195 Hip circumference (cm) |

Model Summary

|  |  |  |  |
| --- | --- | --- | --- |
| S | R-sq | R-sq(adj) | R-sq(pred) |
| 4.28190 | 74.66% | 73.82% | 72.52% |

**Forward BIC Model:**

lm(formula = percent\_body\_fat ~ abdomen\_circumference\_cm + weight\_lbs +

wrist\_circumference\_cm + forearm\_circumference\_cm + neck\_circumference\_cm +

age\_years + thigh\_circumference\_cm + hip\_circumference\_cm,

data = Body\_Fat\_2)

Residuals:

Min 1Q Median 3Q Max

-10.9757 -2.9937 -0.1644 2.9766 10.2244

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -22.65637 11.71385 -1.934 0.05426 .

abdomen\_circumference\_cm 0.94482 0.07193 13.134 < 2e-16 \*\*\*

weight\_lbs -0.08985 0.03991 -2.252 0.02524 \*

wrist\_circumference\_cm -1.53665 0.50939 -3.017 0.00283 \*\*

forearm\_circumference\_cm 0.51572 0.18631 2.768 0.00607 \*\*

neck\_circumference\_cm -0.46656 0.22462 -2.077 0.03884 \*

age\_years 0.06578 0.03078 2.137 0.03356 \*

thigh\_circumference\_cm 0.30239 0.12904 2.343 0.01992 \*

hip\_circumference\_cm -0.19543 0.13847 -1.411 0.15940

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.282 on 243 degrees of freedom

Multiple R-squared: 0.7466, Adjusted R-squared: 0.7382

F-statistic: 89.47 on 8 and 243 DF, p-value: < 2.2e-16

0.06578 0.30239

hip\_circumference\_cm

-0.19543

Regression Equation

|  |  |  |
| --- | --- | --- |
| Percent body fat | = | -22.7 - 0.0899 Weight (lbs) + 0.9448 Abdomen circumference (cm) + 0.516 Forearm circumference (cm) - 1.537 Wrist circumference (cm) - 0.467 Neck circumference (cm) + 0.0658 Age (years) + 0.302 Thigh circumference (cm) - 0.195 Hip circumference (cm) |

Model Summary

|  |  |  |  |
| --- | --- | --- | --- |
| S | R-sq | R-sq(adj) | R-sq(pred) |
| 4.28190 | 74.66% | 73.82% | 72.52% |

**Backward AIC Model:**

lm(formula = percent\_body\_fat ~ age\_years + weight\_lbs + neck\_circumference\_cm +

abdomen\_circumference\_cm + hip\_circumference\_cm + thigh\_circumference\_cm +

forearm\_circumference\_cm + wrist\_circumference\_cm, data = Body\_Fat\_2)

Residuals:

Min 1Q Median 3Q Max

-10.9757 -2.9937 -0.1644 2.9766 10.2244

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -22.65637 11.71385 -1.934 0.05426 .

age\_years 0.06578 0.03078 2.137 0.03356 \*

weight\_lbs -0.08985 0.03991 -2.252 0.02524 \*

neck\_circumference\_cm -0.46656 0.22462 -2.077 0.03884 \*

abdomen\_circumference\_cm 0.94482 0.07193 13.134 < 2e-16 \*\*\*

hip\_circumference\_cm -0.19543 0.13847 -1.411 0.15940

thigh\_circumference\_cm 0.30239 0.12904 2.343 0.01992 \*

forearm\_circumference\_cm 0.51572 0.18631 2.768 0.00607 \*\*

wrist\_circumference\_cm -1.53665 0.50939 -3.017 0.00283 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.282 on 243 degrees of freedom

Multiple R-squared: 0.7466, Adjusted R-squared: 0.7382

F-statistic: 89.47 on 8 and 243 DF, p-value: < 2.2e-16

Regression Equation

|  |  |  |
| --- | --- | --- |
| Percent body fat | = | -22.7 - 0.0899 Weight (lbs) + 0.9448 Abdomen circumference (cm) + 0.516 Forearm circumference (cm) - 1.537 Wrist circumference (cm) - 0.467 Neck circumference (cm) + 0.0658 Age (years) + 0.302 Thigh circumference (cm) - 0.195 Hip circumference (cm) |

Model Summary

|  |  |  |  |
| --- | --- | --- | --- |
| S | R-sq | R-sq(adj) | R-sq(pred) |
| 4.28190 | 74.66% | 73.82% | 72.52% |

**Backward BIC Model:**

lm(formula = percent\_body\_fat ~ weight\_lbs + abdomen\_circumference\_cm +

forearm\_circumference\_cm + wrist\_circumference\_cm, data = Body\_Fat\_2)

Residuals:

Min 1Q Median 3Q Max

-10.5626 -3.1235 -0.1461 3.1313 9.0867

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -34.85407 7.24500 -4.811 2.62e-06 \*\*\*

weight\_lbs -0.13563 0.02475 -5.480 1.05e-07 \*\*\*

abdomen\_circumference\_cm 0.99575 0.05607 17.760 < 2e-16 \*\*\*

forearm\_circumference\_cm 0.47293 0.18166 2.603 0.009790 \*\*

wrist\_circumference\_cm -1.50556 0.44267 -3.401 0.000783 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.343 on 247 degrees of freedom

Multiple R-squared: 0.735, Adjusted R-squared: 0.7307

F-statistic: 171.3 on 4 and 247 DF, p-value: < 2.2e-16

Regression Equation

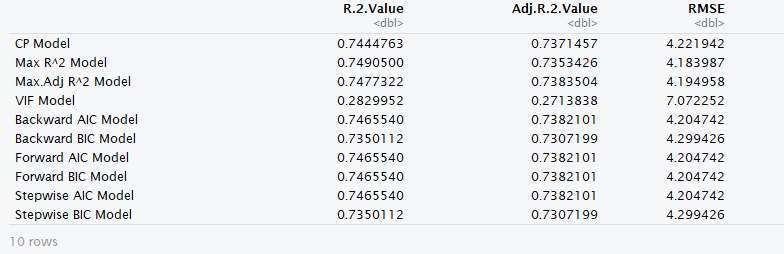
|  |  |  |
| --- | --- | --- |
| Percent body fat | = | -34.85 - 0.1356 Weight (lbs) + 0.9958 Abdomen circumference (cm) + 0.473 Forearm circumference (cm) - 1.506 Wrist circumference (cm) |

Model Summary

|  |  |  |  |
| --- | --- | --- | --- |
| S | R-sq | R-sq(adj) | R-sq(pred) |
| 4.34272 | 73.50% | 73.07% | 72.08% |

**Comparison of Models**

To best compare the models, I created a dataframe with the most pertinent information. I compared the . Generally, the best mode will be the one with the lowest RMSE and the highest .



From the above dataframe we can see that the model with the highest value is the Max model, similarly the model with the highest is the Max . These are expected results because these models were selected to have the highest and . We will exclude the model with the highest since it includes all the predictors and will overfit the data.

**Model Checking**

The Max model has insignificant predictors.

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -23.30499 11.72660 -1.987 0.04801 \*

age\_years 0.06348 0.03084 2.058 0.04064 \*

weight\_lbs -0.09843 0.04070 -2.418 0.01634 \*

neck\_circumference\_cm -0.49330 0.22596 -2.183 0.02999 \*

abdomen\_circumference\_cm 0.94926 0.07204 13.177 < 2e-16 \*\*\*

hip\_circumference\_cm -0.18287 0.13893 -1.316 0.18934

thigh\_circumference\_cm 0.26538 0.13362 1.986 0.04816 \*

biceps\_extended\_circumference\_cm 0.17889 0.16827 1.063 0.28878

forearm\_circumference\_cm 0.45150 0.19581 2.306 0.02197 \*

wrist\_circumference\_cm -1.54208 0.50928 -3.028 0.00273 \*\*

I want to see if these predictors can be removed to create a smaller model. I found that if I remove the 2 insignificant predictors we introduce more insignificant predictors and it negatively affects our RMSE and the .

RMSE = 4.221942 = 0.7371457

**Conclusion**

In conclusion I would recommend we use the Max model

|  |  |  |
| --- | --- | --- |
| Percent body fat | = | -23.3 + 0.0635 Age (years) - 0.0984 Weight (lbs) - 0.493 Neck circumference (cm) + 0.9493 Abdomen circumference (cm) - 0.183 Hip circumference (cm) + 0.265 Thigh circumference (cm) + 0.179 Biceps (extended) circumference + 0.451 Forearm circumference (cm) - 1.542 Wrist circumference (cm) |

This model has the highest value with the lowest RMSE.

**Code for Question 1**

# Load Packages

library("ggplot2")

library("GGally")

library("olsrr")

library("MASS")

# Setup data

set.seed(1)

x1 = runif(100)

x2 = 0.5 \* x1 + rnorm(100)/10

y = 2 + 2\*x1 + 0.3 \* x2 + rnorm(100)

# Find Correlation Between X1 and X2

question\_1 = data.frame("x1" = c(x1), "x2" = c(x2), "y" = c(y))

ggcorr(question\_1, palette = "RdBu", label = TRUE, label\_size = 3, label\_round = 2, hjust = 1)

cor(question\_1)

ggpairs(question\_1, title = "Correlation Matrix for Question 1")

# Fit regression model to predict y using x1 and x2

model\_1 = lm(y ~ x1 + x2, data = question\_1)

summary(model\_1)

# Fit regression model to predict y using x1

model\_2 = lm(y ~ x1, data = question\_1)

summary(model\_2)

# Fit regression model to predict y using x2

model\_3 = lm(y ~ x2, data = question\_1)

summary(model\_3)

# Add new observations

x1=c(x1 , 0.1)

x2=c(x2 , 0.8)

y=c(y,6)

question\_1 = data.frame("x1" = c(x1), "x2" = c(x2), "y" = c(y))

# Fit new regression model to predict y using x1 and x2

model\_4 = lm(y ~ x1 + x2, data = question\_1)

summary(model\_4)

# Check for outliers

ols\_plot\_cooksd\_bar(model\_4)

leverage\_4 = (hatvalues(model\_4))

ggplot(data = NULL, aes(x = seq(1, length(leverage\_4)), y = leverage\_4)) + geom\_point() + labs(title="Leverage Plot for y ~ x1 + x2") + xlab("Index") + ylab("Leverage")

sresid\_4 = studres(model\_4)

plot(sresid\_4, ylab = "Studentized Residuals", main = ("Plot of Studentized Residuals for y ~ x1 + x2"))

# Fit new regression model to predict y using x1

model\_5 = lm(y ~ x1, data = question\_1)

summary(model\_5)

# Check for outliers

ols\_plot\_cooksd\_bar(model\_5)

leverage\_5 = (hatvalues(model\_5))

ggplot(data = NULL, aes(x = seq(1, length(leverage\_5)), y = leverage\_5)) + geom\_point() + labs(title="Leverage Plot for y ~ x1") + xlab("Index") + ylab("Leverage")

sresid\_5 = studres(model\_5)

plot(sresid\_5, ylab = "Studentized Residuals", main = ("Plot of Studentized Residuals for y ~ x1"))

# Fit new regression model to predict y using x2

model\_6 = lm(y ~ x2, data = question\_1)

summary(model\_6)

# Check for outliers

ols\_plot\_cooksd\_bar(model\_6)

leverage\_6 = (hatvalues(model\_6))

ggplot(data = NULL, aes(x = seq(1, length(leverage\_6)), y = leverage\_6)) + geom\_point() + labs(title="Leverage Plot for y ~ x2") + xlab("Index") + ylab("Leverage")

sresid\_6 = studres(model\_6)

plot(sresid\_6, ylab = "Studentized Residuals",main = ("Plot of Studentized Residuals for y ~ x2"))

**Question 2**

---

title: 'STAT 595 HW 2 Question 2'

output: word\_document

---

```{r setup, include=FALSE}

knitr::opts\_chunk$set(echo = TRUE)

```

Referring to the Body fat data, find the optimal model to predict percent body fat using: Age (years), Weight (lbs), Height (inches), Neck circumference (cm), Chest circumference (cm), Abdomen circumference (cm), Hip circumference (cm), Thigh circumference (cm), Knee circumference (cm), Ankle circumference (cm), Biceps (extended) circumference (cm), Forearm circumference (cm), Wrist circumference (cm)

```{r}

# Load Packages

library("readr")

library("ggplot2")

library("GGally")

library("dplyr")

library("scales")

library("car")

library("janitor")

library("leaps")

library("MASS")

library("Metrics")

# Fix clashes

select = dplyr::select

```

```{r}

# Import Dataset

library(readr)

Body\_Fat = read\_csv("Body Fat.csv")

head(Body\_Fat)

ls(Body\_Fat)

# Change variable names to have no spaces or parenthesis

Body\_Fat\_2 = Body\_Fat%>%clean\_names()

head(Body\_Fat\_2)

ls(Body\_Fat\_2)

```

```{r}

# Descriptive statistics

summary(Body\_Fat\_2)

```

```{r}

# Check for missing values

sum(is.na(Body\_Fat\_2))

```

```{r}

# Check how data is categorized

str(Body\_Fat\_2)

```

Our first step is to see how correlated these variables are

```{r}

ggcorr(Body\_Fat\_2, palette = "RdBu", label = TRUE, label\_size = 3, label\_round = 2, hjust = 1)

```

From the above graph and the output we can see that many of them are highly correlated. This leads us to suspect that there might be multicollinearity between our variables. We will check the VIF of the linear combinaton

```{r}

# First model fitting for VIF

vif\_model\_1 = lm(percent\_body\_fat~ ., data = Body\_Fat\_2)

summary(vif\_model\_1)

vif(vif\_model\_1)

```

We see very few signficant predictors and 3 of our predictors have a VIF higher than 10. We will start to remove high values of VIF and refit the model to see of we can adjust for multicollinearity

```{r}

# second model fitting for VIF. Remove Weight\_lbs

vif\_model\_2 = lm(percent\_body\_fat~ .-weight\_lbs, data = Body\_Fat\_2)

summary(vif\_model\_2)

vif(vif\_model\_2)

```

From the second fitting we can see that the number of significant variables has increased and the number of variables with a VIF > 10 is now 2

```{r}

# third model fitting for VIF. -weight\_lbs -abdomen\_circumference\_cm

vif\_model\_3 = lm(percent\_body\_fat~ .-weight\_lbs -abdomen\_circumference\_cm, data = Body\_Fat\_2)

summary(vif\_model\_3)

vif(vif\_model\_3)

```

```{r}

# 4th Fitting. Remove -weight\_lbs -abdomen\_circumference\_cm -hip\_circumference\_cm

vif\_model\_4 = lm(percent\_body\_fat~ .-weight\_lbs -abdomen\_circumference\_cm -hip\_circumference\_cm , data = Body\_Fat\_2)

summary(vif\_model\_4)

vif(vif\_model\_4)

```

```{r}

# 5th Fitting. Remove -weight\_lbs -abdomen\_circumference\_cm -hip\_circumference\_cm -thigh\_circumference\_cm

vif\_model\_5 = lm(percent\_body\_fat~ .-weight\_lbs -abdomen\_circumference\_cm -hip\_circumference\_cm -thigh\_circumference\_cm, data = Body\_Fat\_2)

summary(vif\_model\_5)

vif(vif\_model\_5)

```

```{r}

#6th Fitting. Remove -weight\_lbs -abdomen\_circumference\_cm -hip\_circumference\_cm -thigh\_circumference\_cm -neck\_circumference\_cm

vif\_model\_6 = lm(percent\_body\_fat~ .-weight\_lbs -abdomen\_circumference\_cm -hip\_circumference\_cm -thigh\_circumference\_cm -neck\_circumference\_cm, data = Body\_Fat\_2)

summary(vif\_model\_6)

vif(vif\_model\_6)

```

```{r}

#7th Fitting. Remove -weight\_lbs -abdomen\_circumference\_cm -hip\_circumference\_cm -thigh\_circumference\_cm -neck\_circumference\_cm -chest\_circumference\_cm

vif\_model\_7 = lm(percent\_body\_fat~ .-weight\_lbs -abdomen\_circumference\_cm -hip\_circumference\_cm -thigh\_circumference\_cm -neck\_circumference\_cm -chest\_circumference\_cm, data = Body\_Fat\_2)

summary(vif\_model\_7)

vif(vif\_model\_7)

```

```{r}

# 8th Fitting. Remove -weight\_lbs -abdomen\_circumference\_cm -hip\_circumference\_cm -thigh\_circumference\_cm -neck\_circumference\_cm -chest\_circumference\_cm -wrist\_circumference\_cm

vif\_model\_8 = lm(percent\_body\_fat~ .-weight\_lbs -abdomen\_circumference\_cm -hip\_circumference\_cm -thigh\_circumference\_cm -neck\_circumference\_cm -chest\_circumference\_cm -wrist\_circumference\_cm, data = Body\_Fat\_2)

summary(vif\_model\_8)

vif(vif\_model\_8)

```

```{r}

#9th Fitting. Remove -weight\_lbs -abdomen\_circumference\_cm -hip\_circumference\_cm -thigh\_circumference\_cm -neck\_circumference\_cm -chest\_circumference\_cm -wrist\_circumference\_cm -biceps\_extended\_circumference\_cm

vif\_model\_9 = lm(percent\_body\_fat~ .-weight\_lbs -abdomen\_circumference\_cm -hip\_circumference\_cm -thigh\_circumference\_cm -neck\_circumference\_cm -chest\_circumference\_cm -wrist\_circumference\_cm -biceps\_extended\_circumference\_cm, data = Body\_Fat\_2)

summary(vif\_model\_9)

vif(vif\_model\_9)

```

```{r}

# 10th Fitting. Remove -weight\_lbs -abdomen\_circumference\_cm -hip\_circumference\_cm -thigh\_circumference\_cm -neck\_circumference\_cm -chest\_circumference\_cm -wrist\_circumference\_cm -biceps\_extended\_circumference\_cm -knee\_circumference\_cm

vif\_model\_10 = lm(percent\_body\_fat~ .-weight\_lbs -abdomen\_circumference\_cm -hip\_circumference\_cm -thigh\_circumference\_cm -neck\_circumference\_cm -chest\_circumference\_cm -wrist\_circumference\_cm -biceps\_extended\_circumference\_cm -knee\_circumference\_cm, data = Body\_Fat\_2)

summary(vif\_model\_10)

vif(vif\_model\_10)

```

```{r}

# Check correlation and calculate R^2, ADJ R^2, and RMSE of model

Lowest\_vif = Body\_Fat\_2 %>% select(-weight\_lbs, -abdomen\_circumference\_cm, -hip\_circumference\_cm, -thigh\_circumference\_cm, -neck\_circumference\_cm, -chest\_circumference\_cm, -wrist\_circumference\_cm, -biceps\_extended\_circumference\_cm, -knee\_circumference\_cm)

ggcorr(Lowest\_vif, palette = "RdBu", label = TRUE, hjust = 0.7)

RSS\_VIF = c(crossprod(vif\_model\_10$residuals))

MSE\_VIF = RSS\_VIF/length(vif\_model\_10$residuals)

RMSE\_VIF = sqrt(MSE\_VIF)

R2\_VIF = summary(lm(formula = percent\_body\_fat ~ . - weight\_lbs -

abdomen\_circumference\_cm -

hip\_circumference\_cm - thigh\_circumference\_cm - neck\_circumference\_cm -

chest\_circumference\_cm - wrist\_circumference\_cm -

biceps\_extended\_circumference\_cm -

knee\_circumference\_cm, data = Body\_Fat\_2))$r.squared

ajdR2\_VIF = summary(lm(formula = percent\_body\_fat ~ . - weight\_lbs -

abdomen\_circumference\_cm -

hip\_circumference\_cm - thigh\_circumference\_cm - neck\_circumference\_cm -

chest\_circumference\_cm - wrist\_circumference\_cm -

biceps\_extended\_circumference\_cm -

knee\_circumference\_cm, data = Body\_Fat\_2))$adj.r.squared

```

We will now perform model selection using CP, R^2 and Adj R^2

Using leaps we will look for the model with the minimum CP, highest Adj R^2 and Highest R^2. NOTE: R^2 will usually chose the largest possible model and is generally not a good indication.

```{r}

Body\_Fat\_2 %>% select(c(percent\_body\_fat, everything())) # Order variables to make leaps easier

Body\_Fat\_2.mat = as.matrix(Body\_Fat\_2) # Create matrix to work with leaps

cp\_model = leaps(x = Body\_Fat\_2.mat[,2:14], y = Body\_Fat\_2.mat[,1], names = names(Body\_Fat\_2)[2:14], method = "Cp")

r2\_model = leaps(x = Body\_Fat\_2.mat[,2:14], y = Body\_Fat\_2.mat[,1], names = names(Body\_Fat\_2)[2:14], method = "r2")

adjr2\_model = leaps(x = Body\_Fat\_2.mat[,2:14], y = Body\_Fat\_2.mat[,1], names = names(Body\_Fat\_2)[2:14], method = "adjr2")

# Create dataframe of overall CP, R^2, and ADJ R^2

options(warn = -1)

overall\_cp\_r2\_adjr2 = data.frame(cp\_model$which, Cp = cp\_model$Cp, R2 = r2\_model$r2, AdjR2 = adjr2\_model$adjr2)

# Select best CP, R^2 and AdjR^2

best\_cp = cp\_model$which[which((cp\_model$Cp == min(cp\_model$Cp))),]

best\_adjr2 = adjr2\_model$which[which((adjr2\_model$adjr2 == max(adjr2\_model$adjr2))),]

best\_r2 = r2\_model$which[which((r2\_model$r2 == max(r2\_model$r2))),]

# Create dataframe of model

options(warn = -1)

model\_1 = data.frame(best\_cp, best\_r2, best\_adjr2)

# View Model

model\_1

```

```{r}

# Calculate the RMSE, R^2, and ADJR^2 for each model

cp\_model =

lm(formula = percent\_body\_fat ~ . -height\_inches -chest\_circumference\_cm -hip\_circumference\_cm -knee\_circumference\_cm -ankle\_circumference\_cm -biceps\_extended\_circumference\_cm, data = Body\_Fat\_2)

RSS\_CP = c(crossprod(cp\_model$residuals))

MSE\_CP = RSS\_CP/length(cp\_model$residuals)

RMSE\_CP = sqrt(MSE\_CP)

R2\_CP = summary(cp\_model)$r.squared

ajdR2\_CP = summary(cp\_model)$adj.r.squared

max\_r2\_model = lm(formula = percent\_body\_fat ~ ., data = Body\_Fat\_2)

RSS\_Max\_R2 = c(crossprod(max\_r2\_model$residuals))

MSE\_Max\_R2 = RSS\_Max\_R2/length(max\_r2\_model$residuals)

RMSE\_Max\_R2 = sqrt(MSE\_Max\_R2)

R2\_Max\_R2 = summary(max\_r2\_model)$r.squared

ajdR2\_Max\_R2 = summary(max\_r2\_model)$adj.r.squared

max\_adj\_r2\_model = lm(formula = percent\_body\_fat ~ . -height\_inches -chest\_circumference\_cm -knee\_circumference\_cm -ankle\_circumference\_cm, data = Body\_Fat\_2)

RSS\_Max\_Adj\_R2 = c(crossprod(max\_adj\_r2\_model$residuals))

MSE\_Max\_Adj\_R2 = RSS\_Max\_Adj\_R2/length(max\_adj\_r2\_model$residuals)

RMSE\_Max\_Adj\_R2 = sqrt(MSE\_Max\_Adj\_R2)

R2\_Max\_Adj\_R2 = summary(max\_adj\_r2\_model)$r.squared

ajdR2\_Max\_Adj\_R2 = summary(max\_adj\_r2\_model)$adj.r.squared

```

We will now do model selection using Forward, Backward, and Stepwise selection with AIC and BIC

```{r}

# Full model for stepwise selection with another for directional

full.model = lm(percent\_body\_fat ~ ., data = Body\_Fat\_2)

null.model = lm(percent\_body\_fat ~ 1, data = Body\_Fat\_2)

n = nrow(Body\_Fat\_2)

```

```{r}

# Stepwise AIC

Stepwise\_AIC = stepAIC(full.model, direction = "both", trace = FALSE)

summary(Stepwise\_AIC)

# Calculate RMSE, R^2, ADJ R^2

RSS\_Step\_AIC = c(crossprod(Stepwise\_AIC$residuals))

MSE\_Step\_AIC = RSS\_Step\_AIC/length(Stepwise\_AIC$residuals)

RMSE\_Step\_AIC = sqrt(MSE\_Step\_AIC)

R2\_Step\_AIC = summary(Stepwise\_AIC)$r.squared

ajdR2\_Step\_AIC = summary(Stepwise\_AIC)$adj.r.squared

```

```{r}

# Stepwise BIC

Stepwise\_BIC = stepAIC(full.model, direction = "both", trace = FALSE, k = log(n))

summary(Stepwise\_BIC)

# Calculate RMSE, R^2, ADJ R^2

RSS\_Step\_BIC = c(crossprod(Stepwise\_BIC$residuals))

MSE\_Step\_BIC = RSS\_Step\_BIC/length(Stepwise\_BIC$residuals)

RMSE\_Step\_BIC = sqrt(MSE\_Step\_BIC)

R2\_Step\_BIC = summary(Stepwise\_BIC)$r.squared

ajdR2\_Step\_BIC = summary(Stepwise\_BIC)$adj.r.squared

```

```{r}

# Forward AIC

Forward\_AIC = stepAIC(null.model,direction="forward",scope=list(upper=full.model,lower=null.model), trace = FALSE)

summary(Forward\_AIC)

RSS\_Forward\_AIC = c(crossprod(Forward\_AIC$residuals))

MSE\_Forward\_AIC = RSS\_Forward\_AIC/length(Forward\_AIC$residuals)

RMSE\_Forward\_AIC = sqrt(MSE\_Forward\_AIC)

R2\_Forward\_AIC = summary(Forward\_AIC)$r.squared

ajdR2\_Forward\_AIC = summary(Forward\_AIC)$adj.r.squared

```

```{r}

# Forward BIC

Forward\_BIC = stepAIC(null.model,direction="forward",scope=list(upper=full.model,lower=null.model, k=log(n)), trace = FALSE)

summary(Forward\_BIC)

RSS\_Forward\_BIC = c(crossprod(Forward\_BIC$residuals))

MSE\_Forward\_BIC = RSS\_Forward\_BIC/length(Forward\_BIC$residuals)

RMSE\_Forward\_BIC = sqrt(MSE\_Forward\_BIC)

R2\_Forward\_BIC = summary(Forward\_BIC)$r.squared

ajdR2\_Forward\_BIC = summary(Forward\_BIC)$adj.r.squared

```

```{r}

# Backward AIC

Backward\_AIC = stepAIC(full.model,direction="backward", trace = FALSE)

summary(Backward\_AIC)

RSS\_Backward\_AIC = c(crossprod(Backward\_AIC$residuals))

MSE\_Backward\_AIC = RSS\_Backward\_AIC/length(Backward\_AIC$residuals)

RMSE\_Backward\_AIC = sqrt(MSE\_Backward\_AIC)

R2\_Backward\_AIC = summary(Backward\_AIC)$r.squared

ajdR2\_Backward\_AIC = summary(Backward\_AIC)$adj.r.squared

```

```{r}

# Backward BIC

Backward\_BIC = stepAIC(full.model,direction="backward", trace = FALSE, k = log(n))

summary(Backward\_BIC)

RSS\_Backward\_BIC = c(crossprod(Backward\_BIC$residuals))

MSE\_Backward\_BIC = RSS\_Backward\_BIC/length(Backward\_BIC$residuals)

RMSE\_Backward\_BIC = sqrt(MSE\_Backward\_BIC)

R2\_Backward\_BIC = summary(Backward\_BIC)$r.squared

ajdR2\_Backward\_BIC = summary(Backward\_BIC)$adj.r.squared

```

```{r}

# Model Comparison Dataframe

Model\_Comparison = data.frame("R^2 Value" = c(R2\_CP, R2\_Max\_R2, R2\_Max\_Adj\_R2, R2\_VIF, R2\_Backward\_AIC, R2\_Backward\_BIC, R2\_Forward\_AIC, R2\_Forward\_BIC, R2\_Step\_AIC, R2\_Step\_BIC), "Adj R^2 Value" = c(ajdR2\_CP, ajdR2\_Max\_R2, ajdR2\_Max\_Adj\_R2, ajdR2\_VIF, ajdR2\_Backward\_AIC, ajdR2\_Backward\_BIC, ajdR2\_Forward\_AIC, ajdR2\_Forward\_BIC, ajdR2\_Step\_AIC, ajdR2\_Step\_BIC), "RMSE" = c(RMSE\_CP, RMSE\_Max\_R2, RMSE\_Max\_Adj\_R2, RMSE\_VIF, RMSE\_Backward\_AIC, RMSE\_Backward\_BIC, RMSE\_Forward\_AIC, RMSE\_Forward\_BIC, RMSE\_Step\_AIC, RMSE\_Step\_BIC))

rownames(Model\_Comparison) = c("CP Model", "Max R^2 Model", "Max.Adj R^2 Model", "VIF Model", "Backward AIC Model", "Backward BIC Model", "Forward AIC Model", "Forward BIC Model", "Stepwise AIC Model", "Stepwise BIC Model")

Model\_Comparison

```

```{r}

summary(max\_adj\_r2\_model)

```

Note that 2 of the above predictors are not significant. hip\_circumference\_cm and biceps\_extended\_circumference\_cm. We will remove biceps\_extended\_circumference\_cm first since it has the largest P-Vaue and check the new model

```{r}

check = lm(formula = percent\_body\_fat ~ . - height\_inches - chest\_circumference\_cm -

knee\_circumference\_cm - ankle\_circumference\_cm -biceps\_extended\_circumference\_cm , data = Body\_Fat\_2)

summary(check)

RSS\_check = c(crossprod(check$residuals))

MSE\_check = RSS\_check/length(check$residuals)

RMSE\_check = sqrt(MSE\_check)

R2\_checkC = summary(check)$r.squared

ajdR2\_check = summary(check)$adj.r.squared

```

This gives us a RMSE of 4.2047

We will remove hip\_circumference\_cm and check the RMSE

```{r}

check\_2 = lm(formula = percent\_body\_fat ~ . - height\_inches - chest\_circumference\_cm -

knee\_circumference\_cm - ankle\_circumference\_cm -biceps\_extended\_circumference\_cm -hip\_circumference\_cm, data = Body\_Fat\_2)

summary(check\_2)

RSS\_check\_2 = c(crossprod(check\_2$residuals))

MSE\_check\_2 = RSS\_check\_2/length(check\_2$residuals)

RMSE\_check\_2 = sqrt(MSE\_check\_2)

R2\_check\_2 = summary(check\_2)$r.squared

ajdR2\_check\_2 = summary(check\_2)$adj.r.squared

```

Note that if we remove the insignificant predictors for the Max\_Adj\_R2 model we create models that have more insignificant predictors and a lower Adjusted R^2. In conclusion I would recommend that we do not remove the insignificant predictors and keep the model as is.