HW4

Andrew Shao

2024-10-21

(a)

In the `fat` dataset there are 252 observations and 18 variables, with each observation corresponding to the measurements for a singular man. The response variable in this analysis is `brozek`, a measure of body fat using Brozek's equation. There are 17 other potential predictor variables. `siri` is a measure of body fat using Siri's equation. `density`, `age`, `weight`, and `height` are what their names suggest. `adipos` is the adiposity index while `free` is the fat free weight using Brozek's formula. The remaining variables `neck`, `chest`, `abdom`, `hip`, `thigh`, `knee`, `ankle`, `biceps`, `forearm`, and `wrist`, are the circumference measurements for the named body part. All 18 of these variables are numeric. `brozek` appears to be normally distributed. The pairs of variables with a correlation coefficient of at least 0.9 are `brozek` and `siri`, `hip` and `weight`, `chest` and `adipos`, `abdom` and `adipos`, `chest` and `abdom`, and `hip` and `thigh`. Fat percentage shows strong correlations with several measurements.

(b)

The fitted model is as follows:

$$\begin{aligned} brozek =& 12.152 + 0.888 \cdot siri - 9.846 \cdot density - 0.001 \cdot age + 0.008 \cdot weight - 0.001 \cdot height - 0.015 \cdot adipos \\ & - 0.010 \cdot free + 0.001 \cdot neck + 0.002 \cdot chest + 0.001 \cdot abdom - 0.004 \cdot hip + 0.016 \cdot thigh \\ & - 0.025 \cdot knee + 0.003 \cdot ankle - 0.015 \cdot biceps + 0.015 \cdot forearm + 0.033 \cdot wrist \end{aligned}$$

Residual standard error is 0.1706 with 234 degrees of freedom. Multiple R-squared and adjusted R-squared are both 0.9995. The F-statistic p-value is extremely small, less than $2.2 \cdot 10^{-16}$. The variables that are significant at the 0.05 level are `siri`, `density`, `weight`, `free`, `thigh`, `knee`, and `biceps`. This model is questionable because it includes many variables which are already highly correlated, like `siri` and `density` where `density` is directly used to calculate siri‘ but it does seem to have good predictive power as it has very high R-squared values and an extremely small F-statistic p-value. There are many non-significant variables which could be removed to possibly simplify the model.

(c)

The fitted model is as follows:

$$\begin{aligned} brozek =& -10.546 + 0.005 \cdot age + 0.335 \cdot weight + 0.047 \cdot height - 0.0447 \cdot adipos - 0.5523 \cdot free \\ & + 0.020 \cdot neck + 0.111 \cdot chest + 0.130 \cdot abdom - 0.004 \cdot hip + 0.183 \cdot thigh + 0.082 \cdot knee \\ & + 0.127 \cdot ankle + 0.099 \cdot biceps + 0.216 \cdot forearm + 0.139 \cdot wrist \end{aligned}$$

Residual standard error is 1.384 with 236 degrees of freedom. Multiple R-squared is 0.97 and adjusted R-squared is 0.9681. The F-statistic p-value is extremely small still, less than $2.2 \cdot 10^{-16}$. The variables that are significant at the 0.05 level are `weight`, `adipos`, `free`, `chest`, `abdom`, `thigh`, and `forearm`. This model still has very high predictive power with higher R-squared values but also much higher residual standard error, suggesting decreased precision. Additionally there are still many non-significant variables which could be removed to possibly simplify the model.

(d)

The fitted model is as follows:

$$brozek = -6.397 + 0.346 \cdot weight - 0.530 \cdot adipos - 0.522 \cdot free + 0.113 \cdot chest + 0.179 \cdot thigh + 0.137 \cdot ankle + 0.103 \cdot biceps + 0.220 \cdot forearm + 0.225 \cdot wrist$$

Residual standard error is 1.378 with 241 degrees of freedom. Multiple R-squared is 0.9363 and adjusted R-squared is 0.9684. The F-statistic p-value is extremely small still, less than $2.2 \cdot 10^{-16}$. The variables that are significant at the 0.05 level are `weight`, `adipos`, `free`, `chest`, `abdom`, `thigh`, `ankle`, and `forearm`. This model is much simpler with almost all of the variables significant while maintaining predictive power with quite large R-squared values.

(e)

Model #3 has slightly lower multiple R-squared but slightly higher adjusted R-squared. I prefer model #3 because it is much simpler as it has less predictors while maintaining comparable R-squared values. Additionally, the ANOVA F-test p value is very high which suggests not much difference between the two models.

(f)

The training error for model #2 is 1.740 and the training error for model #3 is 1.786. The testing error for model #2 is 0.0096 and 0.0037 for model #3.

(g)

The mean training error for model 2 is 1.6757 which is less than the error for model 3 which is 1.7701. The mean testing error for model 2 is 0.2593 which is higher than the mean testing error for model 3 which is 0.0531. The p-value from the tests are both significant, indicating the difference in performance is significant. I prefer model #3 like with part (e) because it is both simpler but performs better in terms of testing error indicating better generalization/less overfitting.

Appendix

(a)

```
## Warning: package 'faraway' was built under R version 4.3.3
```

```
##   brozek siri density age weight height adipos  free neck chest abdom   hip
## 1   12.6 12.3  1.0708  23 154.25  67.75   23.7 134.9 36.2  93.1  85.2  94.5
## 2    6.9  6.1  1.0853  22 173.25  72.25   23.4 161.3 38.5  93.6  83.0  98.7
## 3   24.6 25.3  1.0414  22 154.00  66.25   24.7 116.0 34.0  95.8  87.9  99.2
## 4   10.9 10.4  1.0751  26 184.75  72.25   24.9 164.7 37.4 101.8  86.4 101.2
## 5   27.8 28.7  1.0340  24 184.25  71.25   25.6 133.1 34.4  97.3 100.0 101.9
## 6   20.6 20.9  1.0502  24 210.25  74.75   26.5 167.0 39.0 104.5  94.4 107.8
##   thigh knee ankle biceps forearm wrist
## 1  59.0 37.3  21.9   32.0    27.4  17.1
## 2  58.7 37.3  23.4   30.5    28.9  18.2
## 3  59.6 38.9  24.0   28.8    25.2  16.6
## 4  60.1 37.3  22.8   32.4    29.4  18.2
## 5  63.2 42.2  24.0   32.2    27.7  17.7
## 6  66.0 42.0  25.6   35.7    30.6  18.8
```

```
## [1] 252  18
```

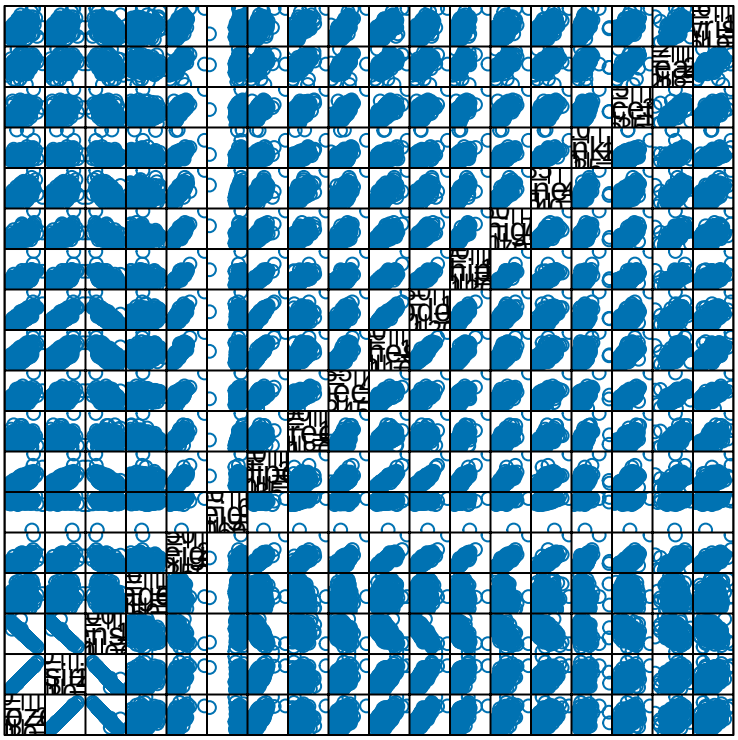
```
##      brozek           siri          density           age       
##  Min.   : 0.00   Min.   : 0.00   Min.   :0.995   Min.   :22.00  
```

```
##  1st Qu.:12.80   1st Qu.:12.47   1st Qu.:1.041   1st Qu.:35.75
##  Median :19.00   Median :19.20   Median :1.055   Median :43.00
##  Mean   :18.94   Mean   :19.15   Mean   :1.056   Mean   :44.88
##  3rd Qu.:24.60   3rd Qu.:25.30   3rd Qu.:1.070   3rd Qu.:54.00
##  Max.   :45.10   Max.   :47.50   Max.   :1.109   Max.   :81.00
##      weight          height          adipos           free
##  Min.   :118.5   Min.   :29.50   Min.   :18.10   Min.   :105.9
##  1st Qu.:159.0   1st Qu.:68.25   1st Qu.:23.10   1st Qu.:131.3
##  Median :176.5   Median :70.00   Median :25.05   Median :141.6
##  Mean   :178.9   Mean   :70.15   Mean   :25.44   Mean   :143.7
##  3rd Qu.:197.0   3rd Qu.:72.25   3rd Qu.:27.32   3rd Qu.:153.9
##  Max.   :363.1   Max.   :77.75   Max.   :48.90   Max.   :240.5
##       neck            chest            abdom             hip
##  Min.   :31.10   Min.   : 79.30   Min.   : 69.40   Min.   : 85.0
##  1st Qu.:36.40   1st Qu.: 94.35   1st Qu.: 84.58   1st Qu.: 95.5
##  Median :38.00   Median : 99.65   Median : 90.95   Median : 99.3
##  Mean   :37.99   Mean   :100.82   Mean   : 92.56   Mean   : 99.9
##  3rd Qu.:39.42   3rd Qu.:105.38   3rd Qu.: 99.33   3rd Qu.:103.5
##  Max.   :51.20   Max.   :136.20   Max.   :148.10   Max.   :147.7
##      thigh            knee            ankle           biceps         forearm
##  Min.   :47.20   Min.   :33.00   Min.   :19.1   Min.   :24.80   Min.   :21.00
##  1st Qu.:56.00   1st Qu.:36.98   1st Qu.:22.0   1st Qu.:30.20   1st Qu.:27.30
##  Median :59.00   Median :38.50   Median :22.8   Median :32.05   Median :28.70
##  Mean   :59.41   Mean   :38.59   Mean   :23.1   Mean   :32.27   Mean   :28.66
##  3rd Qu.:62.35   3rd Qu.:39.92   3rd Qu.:24.0   3rd Qu.:34.33   3rd Qu.:30.00
##  Max.   :87.30   Max.   :49.10   Max.   :33.9   Max.   :45.00   Max.   :34.90
##      wrist
##  Min.   :15.80
##  1st Qu.:17.60
##  Median :18.30
##  Mean   :18.23
##  3rd Qu.:18.80
##  Max.   :21.40
```


```
## Warning: package 'lattice' was built under R version 4.3.3
```

```
##
## Attaching package: 'lattice'
```

```
## The following object is masked from 'package:faraway':
##
##     melanoma
```



Scatter Plot Matrix

```
##          brozek  siri density   age weight height adipos  free  neck chest abdom
## brozek     1.00  1.00   -0.99  0.29   0.61  -0.09   0.73  0.02  0.49  0.70  0.81
## siri       1.00  1.00   -0.99  0.29   0.61  -0.09   0.73  0.02  0.49  0.70  0.81
## density   -0.99 -0.99    1.00 -0.28  -0.59   0.10  -0.71 -0.01 -0.47 -0.68 -0.80
## age        0.29  0.29   -0.28  1.00  -0.01  -0.17   0.12 -0.24  0.11  0.18  0.23
## weight     0.61  0.61   -0.59 -0.01   1.00   0.31   0.89  0.79  0.83  0.89  0.89
## height    -0.09 -0.09    0.10 -0.17   0.31   1.00  -0.02  0.49  0.25  0.13  0.09
## adipos     0.73  0.73   -0.71  0.12   0.89  -0.02   1.00  0.55  0.78  0.91  0.92
## free       0.02  0.02   -0.01 -0.24   0.79   0.49   0.55  1.00  0.68  0.59  0.50
## neck       0.49  0.49   -0.47  0.11   0.83   0.25   0.78  0.68  1.00  0.78  0.75
## chest      0.70  0.70   -0.68  0.18   0.89   0.13   0.91  0.59  0.78  1.00  0.92
## abdom      0.81  0.81   -0.80  0.23   0.89   0.09   0.92  0.50  0.75  0.92  1.00
## hip        0.63  0.63   -0.61 -0.05   0.94   0.17   0.88  0.70  0.73  0.83  0.87
## thigh      0.56  0.56   -0.55 -0.20   0.87   0.15   0.81  0.68  0.70  0.73  0.77
## knee       0.51  0.51   -0.50  0.02   0.85   0.29   0.71  0.70  0.67  0.72  0.74
## ankle      0.27  0.27   -0.26 -0.11   0.61   0.26   0.50  0.58  0.48  0.48  0.45
## biceps     0.49  0.49   -0.49 -0.04   0.80   0.21   0.75  0.65  0.73  0.73  0.68
## forearm    0.36  0.36   -0.35 -0.09   0.63   0.23   0.56  0.55  0.62  0.58  0.50
## wrist      0.35  0.35   -0.33  0.21   0.73   0.32   0.63  0.67  0.74  0.66  0.62
##             hip thigh  knee ankle biceps forearm wrist
## brozek     0.63  0.56  0.51  0.27   0.49    0.36  0.35
## siri       0.63  0.56  0.51  0.27   0.49    0.36  0.35
## density   -0.61 -0.55 -0.50 -0.26  -0.49   -0.35 -0.33
## age       -0.05 -0.20  0.02 -0.11  -0.04   -0.09  0.21
## weight     0.94  0.87  0.85  0.61   0.80    0.63  0.73
## height     0.17  0.15  0.29  0.26   0.21    0.23  0.32
```

```
## adipos   0.88  0.81  0.71  0.50   0.75    0.56  0.63
## free     0.70  0.68  0.70  0.58   0.65    0.55  0.67
## neck     0.73  0.70  0.67  0.48   0.73    0.62  0.74
## chest    0.83  0.73  0.72  0.48   0.73    0.58  0.66
## abdom    0.87  0.77  0.74  0.45   0.68    0.50  0.62
## hip      1.00  0.90  0.82  0.56   0.74    0.55  0.63
## thigh    0.90  1.00  0.80  0.54   0.76    0.57  0.56
## knee     0.82  0.80  1.00  0.61   0.68    0.56  0.66
## ankle    0.56  0.54  0.61  1.00   0.48    0.42  0.57
## biceps   0.74  0.76  0.68  0.48   1.00    0.68  0.63
## forearm  0.55  0.57  0.56  0.42   0.68    1.00  0.59
## wrist    0.63  0.56  0.66  0.57   0.63    0.59  1.00
```

(b)

```
##
## Call:
## lm(formula = brozek ~ ., data = fat)
##
## Coefficients:
## (Intercept)         siri      density          age       weight       height
##  12.1524013    0.8884085   -9.8456305   -0.0005268    0.0084855   -0.0005459
##      adipos         free         neck        chest        abdom          hip
##  -0.0153248   -0.0097388    0.0005002    0.0021454    0.0014464   -0.0044514
##       thigh         knee        ankle       biceps      forearm        wrist
##   0.0156926   -0.0252126    0.0027790   -0.0147134    0.0149983    0.0326518
```

```
##
## Call:
## lm(formula = brozek ~ ., data = fat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.11191 -0.04847  0.00277  0.04625  1.47542
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.1524013  4.1718589   2.913  0.00393 **
## siri         0.8884085  0.0111341  79.792  < 2e-16 ***
## density     -9.8456305  3.7471770  -2.627  0.00917 **
## age         -0.0005268  0.0012935  -0.407  0.68421
## weight       0.0084855  0.0036200   2.344  0.01991 *
## height      -0.0005459  0.0044439  -0.123  0.90234
## adipos      -0.0153248  0.0124778  -1.228  0.22062
## free        -0.0097388  0.0044270  -2.200  0.02880 *
## neck         0.0005002  0.0094279   0.053  0.95773
## chest        0.0021454  0.0043013   0.499  0.61840
## abdom        0.0014464  0.0044217   0.327  0.74388
## hip         -0.0044514  0.0058941  -0.755  0.45087
## thigh        0.0156926  0.0059507   2.637  0.00892 **
## knee        -0.0252126  0.0098531  -2.559  0.01113 *
## ankle        0.0027790  0.0089580   0.310  0.75667
## biceps      -0.0147134  0.0069201  -2.126  0.03454 *
## forearm      0.0149983  0.0080832   1.855  0.06478 .
## wrist        0.0326518  0.0218000   1.498  0.13554
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1706 on 234 degrees of freedom
## Multiple R-squared:  0.9995, Adjusted R-squared:  0.9995
## F-statistic: 3.046e+04 on 17 and 234 DF,  p-value: < 2.2e-16
```

(c)

```
##
## Call:
## lm(formula = brozek ~ . - siri - density, data = fat)
##
## Coefficients:
## (Intercept)          age       weight       height       adipos         free
##  -10.545815     0.004945     0.335454     0.046754    -0.447348    -0.522555
##        neck        chest        abdom          hip        thigh         knee
##    0.019628     0.110709     0.130007    -0.003863     0.182879     0.081735
##       ankle       biceps      forearm        wrist
##    0.126725     0.099114     0.216383     0.139184
```

```
##
## Call:
## lm(formula = brozek ~ . - siri - density, data = fat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7404 -0.6200  0.1792  0.8169  6.1104
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -10.545815   5.578700  -1.890 0.059934 .
## age           0.004945   0.010471   0.472 0.637153
## weight        0.335454   0.019953  16.813  < 2e-16 ***
## height        0.046754   0.035895   1.302 0.194017
## adipos       -0.447348   0.097010  -4.611 6.56e-06 ***
## free         -0.522555   0.012524 -41.724  < 2e-16 ***
## neck          0.019628   0.076435   0.257 0.797563
## chest         0.110709   0.034030   3.253 0.001308 **
## abdom         0.130007   0.034607   3.757 0.000217 ***
## hip          -0.003863   0.047700  -0.081 0.935518
## thigh         0.182879   0.046948   3.895 0.000128 ***
## knee          0.081735   0.079594   1.027 0.305520
## ankle         0.126725   0.071873   1.763 0.079163 .
## biceps        0.099114   0.055421   1.788 0.074997 .
## forearm       0.216383   0.064210   3.370 0.000878 ***
## wrist         0.139184   0.176288   0.790 0.430594
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.384 on 236 degrees of freedom
## Multiple R-squared:   0.97,  Adjusted R-squared:  0.9681
## F-statistic: 509.4 on 15 and 236 DF,  p-value: < 2.2e-16
```

(d)

```
##
## Call:
## lm(formula = brozek ~ weight + adipos + free + chest + abdom +
##     thigh + ankle + biceps + forearm + wrist, data = fat)
##
## Coefficients:
## (Intercept)       weight       adipos         free        chest        abdom
##     -6.3966       0.3459      -0.5300      -0.5219       0.1128       0.1408
##       thigh        ankle       biceps      forearm        wrist
##      0.1790       0.1373       0.1029       0.2197       0.2249
```

```
##
## Call:
## lm(formula = brozek ~ weight + adipos + free + chest + abdom +
##     thigh + ankle + biceps + forearm + wrist, data = fat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7109 -0.6190  0.2100  0.7803  6.2817
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.39657    3.45450  -1.852 0.065298 .
## weight       0.34594    0.01646  21.012  < 2e-16 ***
## adipos      -0.52996    0.07817  -6.779 9.22e-11 ***
## free        -0.52192    0.01220 -42.794  < 2e-16 ***
## chest        0.11283    0.03291   3.428 0.000715 ***
## abdom        0.14078    0.03198   4.402 1.61e-05 ***
## thigh        0.17897    0.03903   4.586 7.27e-06 ***
## ankle        0.13727    0.06931   1.981 0.048778 *
## biceps       0.10288    0.05445   1.890 0.060019 .
## forearm      0.21967    0.06227   3.528 0.000502 ***
## wrist        0.22493    0.15380   1.463 0.144898
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.378 on 241 degrees of freedom
## Multiple R-squared:  0.9696, Adjusted R-squared:  0.9684
## F-statistic: 769.7 on 10 and 241 DF,  p-value: < 2.2e-16
```

(e)

```
## [1] 0.9700398 0.9696405
```

```
## [1] 0.9681356 0.9683808
```

```
## Analysis of Variance Table
##
## Model 1: brozek ~ (siri + density + age + weight + height + adipos + free +
##     neck + chest + abdom + hip + thigh + knee + ankle + biceps +
##     forearm + wrist) - siri - density
## Model 2: brozek ~ weight + adipos + free + chest + abdom + thigh + ankle +
##     biceps + forearm + wrist
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
```

```
## 1    236 451.77
## 2    241 457.79 -5   -6.0214 0.6291 0.6777
```

(f)

```
## [1] 1.740593 1.786454
```

```
## [1] 0.009638842 0.003714445
```

(g)

```
## [1] 100   2
```

```
##   Mod2   Mod3
## 1.6757 1.7701
```

```
##   Mod2   Mod3
## 0.0339 0.0265
```

```
## [1] 100   2
```

```
##   Mod2   Mod3
## 0.2593 0.0531
```

```
##   Mod2   Mod3
## 0.2237 0.0036
```

```
##
##  Paired t-test
##
## data:  TrainErr[, 1] and TrainErr[, 2]
## t = -7.5875, df = 99, p-value = 1.816e-11
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -0.11909555 -0.06971834
## sample estimates:
## mean difference
##     -0.09440695
```

```
##
##  Paired t-test
##
## data:  TestErr[, 1] and TestErr[, 2]
## t = 4.3324, df = 99, p-value = 3.542e-05
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  0.1117797 0.3006840
## sample estimates:
## mean difference
##       0.2062318
```

```
##   Mod2   Mod3
## 1.6757 1.7701
```

```
##   Mod2   Mod3
## 0.2593 0.0531
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
```

```
## data:  TrainErr[, 1] and TrainErr[, 2]
## V = 0, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  TestErr[, 1] and TestErr[, 2]
## V = 3725, p-value = 3.719e-05
## alternative hypothesis: true location shift is not equal to 0
```