Lab 4

Christopher Hainzl

2023-04-18

```
healthdata <- read_csv("lab4data(1).csv")
```

```
## Rows: 95 Columns: 8
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (8): subject, Weight, Height, Protein, Carbohydrates, Calcium, gender, C...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
healthdata
```

```
## # A tibble: 95 x 8
##    subject Weight Height Protein Carbohydrates Calcium gender Calories
##      <dbl>  <dbl>  <dbl>   <dbl>         <dbl>   <dbl>  <dbl>    <dbl>
##  1       1    180     74    84.3          266.    605       1    2441.
##  2       2    175     69    65.3          177.    735.      1    1288.
##  3       3    165     71    66.2          104.    569.      1    1340.
##  4       4    220     71    74.0          152.    991.      1    1747.
##  5       5    230     75    86.5          223.    901.      1    2293.
##  6       6    165     70    80.0          243.    805.      1    2138.
##  7       7    134     63    48.4          149.    744.      0    1478.
##  8       8    190     73    84.2          289.   1240.      1    2319.
##  9       9    185     72    93.9          250.   2123.      1    1652.
## 10      10    165     72    74.5          210.    954.      1    1758.
## # ... with 85 more rows
```

- Weight (lbs) = The measured weight, in pounds, of each subject.
- Height (in.) = The measured height, in inches, of each subject.
- Protein = The protein intake of each subject.
- Carbohydrates = The carbohydrate intake of each subject.
- Calcium = The calcium intake of each subject.
- Gender = The gender of each subject (1 = male, 0 = female)
- Calories = The amount of calories consumed by each subject.

The most appropriate way to begin this analysis would be to set up a linear regression model that predicts a subject's caloric intake based on weight, height, protein, carbohydrates, calcium, and gender. While setting up this model, the assumptions are made that the model's residuals will be normally distributed and that the variances are equal. We will also be working with a 95% confidence level.

```
# Model 1 - All predictors included
healthmodel1 <- lm(Calories~Weight+Height+Protein+Carbohydrates+Calcium+
gender, data = healthdata)
summary(healthmodel1)
```

```
##
## Call:
## lm(formula = Calories ~ Weight + Height + Protein + Carbohydrates +
##     Calcium + gender, data = healthdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -385.98 -122.97  -18.74  141.27  437.45
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -366.70646  449.70704  -0.815  0.41703
## Weight           1.25236    0.78046   1.605  0.11215
## Height           3.01398    7.30509   0.413  0.68091
## Protein         13.42889    1.41199   9.511 3.65e-15 ***
## Carbohydrates    4.59163    0.40044  11.466  < 2e-16 ***
## Calcium         -0.23660    0.06607  -3.581  0.00056 ***
## gender         -70.03331   61.90992  -1.131  0.26104
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 188.2 on 88 degrees of freedom
## Multiple R-squared:  0.9326, Adjusted R-squared:  0.928
## F-statistic:   203 on 6 and 88 DF,  p-value: < 2.2e-16
```

- For each unit increase in the subject's weight, their caloric intake increases by approximately 1.25 on average. (B1)
- For each unit increase in the subject's height, their caloric intake increases by approximately 3.01 on average. (B2)
- For each unit increase in the amount of protein consumed by the subject, their caloric intake increases by approximately 13.43 on average. (B3)
- For each unit increase in the amount of carbohydrates consumed by the subject, their caloric intake increases by approximately 4.59 on average. (B4)
- For each unit increase in the amount of calcium consumed by the subject, their caloric intake decreases by approximately 0.24 on average. (B5)
- For the subjects that are males, their caloric intake decreases by approximately 70.03 on average when compared to females. (B6)

However, based on the p-values associated with each of these predictors, the only ones that appear to be significant when working with a 95% confidence level are protein, carbohydrates, and calcium. The p-values associated with each of those predictors are all less than .05.

Therefore, we should try removing each of the non-significant predictors one at a time and determine if the relationship between caloric intake and each of the predictors changes.

```
# Model 2 - Remove weight
healthmodel2 <- lm(Calories~Height+Protein+Carbohydrates+Calcium+
gender, data = healthdata)
summary(healthmodel2)
```

```
##
## Call:
## lm(formula = Calories ~ Height + Protein + Carbohydrates + Calcium +
##     gender, data = healthdata)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -396.34 -133.23  -24.05  122.72  402.73
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -497.28127  446.17979  -1.115 0.268052
## Height           7.99944    6.66972   1.199 0.233569
## Protein         13.59677    1.42051   9.572 2.46e-15 ***
## Carbohydrates    4.58203    0.40393  11.344  < 2e-16 ***
## Calcium         -0.24603    0.06639  -3.706 0.000365 ***
## gender         -53.06202   61.53718  -0.862 0.390853
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 189.9 on 89 degrees of freedom
## Multiple R-squared:  0.9306, Adjusted R-squared:  0.9267
## F-statistic: 238.8 on 5 and 89 DF,  p-value: < 2.2e-16
```

After removing weight, the p-values for all the predictor variables changed, but the amounts of protein, carbohydrates, and calcium consumed by the subject still remain the most significant predictors of their caloric intake.

```
# Model 3 - Remove height
healthmodel3 <- lm(Calories~Protein+Carbohydrates+Calcium+
gender, data = healthdata)
summary(healthmodel3)
```

```
##
## Call:
## lm(formula = Calories ~ Protein + Carbohydrates + Calcium + gender,
##     data = healthdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -392.87 -130.75   -6.33  116.72  369.50
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    33.02920   59.91176   0.551 0.582796
## Protein        13.39071    1.41352   9.473 3.56e-15 ***
## Carbohydrates   4.63348    0.40262  11.508  < 2e-16 ***
## Calcium        -0.23674    0.06609  -3.582 0.000553 ***
## gender        -13.16963   51.89967  -0.254 0.800266
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 190.4 on 90 degrees of freedom
## Multiple R-squared:  0.9295, Adjusted R-squared:  0.9264
## F-statistic: 296.7 on 4 and 90 DF,  p-value: < 2.2e-16
```

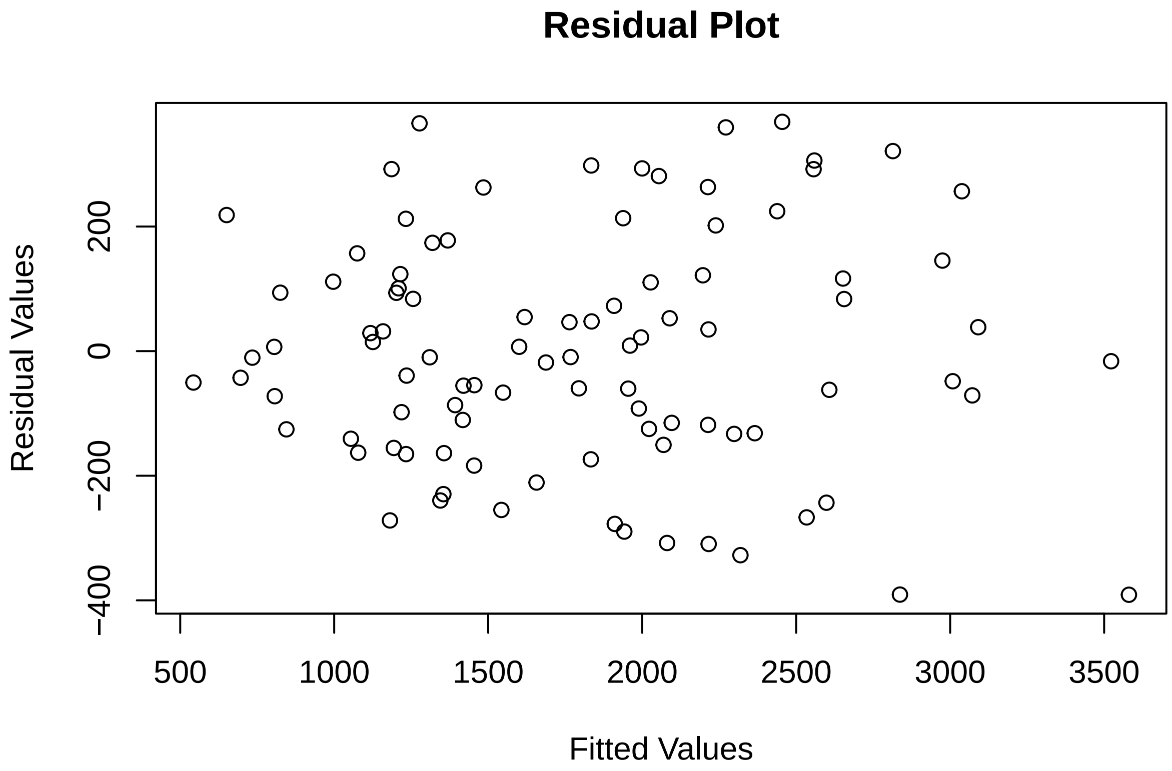
After removing height from our model, protein, carbohydrates, and calcium still remain the most significant predictors when working with a 95% confidence level.

```
# Model 4 - Remove gender
healthmodel4 <- lm(Calories~Protein+Carbohydrates+Calcium, data = healthdata)
summary(healthmodel4)
```

```
##
## Call:
## lm(formula = Calories ~ Protein + Carbohydrates + Calcium, data = healthdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -391.05 -128.63   -9.87  119.16  368.04
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   27.09131   54.86862   0.494 0.622673
## Protein       13.30226    1.36280   9.761 8.07e-16 ***
## Carbohydrates  4.62531    0.39926  11.585  < 2e-16 ***
## Calcium       -0.23224    0.06334  -3.667 0.000413 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 189.4 on 91 degrees of freedom
## Multiple R-squared:  0.9295, Adjusted R-squared:  0.9271
## F-statistic: 399.7 on 3 and 91 DF,  p-value: < 2.2e-16
```

Now that all the non-significant predictors have been removed, we can determine how well our final model fits the original data and if our model assumptions are met.

```
# Residual Plot - test to see if variances are equal
plot(healthmodel4$residuals ~ predict(healthmodel4),
     main = "Residual Plot", xlab = "Fitted Values",
     ylab = "Residual Values")
```



```
# Use shapiro.test() function to see if residuals are normally distributed
shapiro.test(healthmodel4$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  healthmodel4$residuals
## W = 0.98258, p-value = 0.2389
```

Our final, simplified model helps us draw the following conclusions:

Starting at a value of approximately 27.09:

- For each unit increase in how much protein each subject consumes, their average caloric intake will increase by approximately 13.30.
- For each unit increase in how much carbohydrates the subject consumes, their average caloric intake will increase by approximately 4.63.
- For each unit increase in how much calcium each subject consumes, their average caloric intake will decrease by approximately 0.23.
- Weight, height, and gender do not help with predicting the caloric intake of a subject.

Since the p-value associated with the Shapiro-Wilk test is greater than .05, and the graph labeled Residual Plot does not show any particular pattern (it looks like a random cloud centered at 0), this means that our model assumptions are satisfied. And because the value of R-squared (0.9295) is close to 1, this means that our final model fits the data very well and that a linear model was appropriate to use for this analysis.

Therefore, we can also be certain that all of the conclusions which were drawn using our final model are correct.