UC **SANTA BARBARA**

Final Project PSTAT 126: Regression Analysis

Regression Analysis on Body Fat Percentage in Men

Shaiyon Hariri - Section: Tuesday 3pm
Tamjid Islam - Section: Wednesday 8am
June 7, 2020

1. Introduction

In this project, we analyze a dataset that lists estimates of the percentage of body fat determined by underwater weighing and various body circumference measurements for 252 men. Accurate figures for body fat percentage are extremely inconvenient and costly to obtain, and a solid model to predict this metric from easily collected measurements could save a lot of time and money.

This data comes from the Carnegie Mellon University StatLib Dataset Archive, and contains 15 variables: Density determined from underwater weighing, Percent body fat, 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 circumference (cm), Forearm circumference (cm), and Wrist circumference (cm). We will use Percent body fat as the response, and attempt to determine which of the predictors influences this figure the most.

2. Questions of Interest

I. Which metric is most strongly correlated with a high body fat percentage?

II. Which metric is most strongly correlated with a low body fat percentage?

III. Is high weight correlated with a high body fat percentage?

3. Regression Method

The idea of a regression model is to better understand the relationship between a response variable and selected predictor variable(s). In order for linear regression to be effective, the model must follow certain criteria known as the "LINE" conditions. If one of these conditions aren't followed, we can adjust the data by creating a transformation on the response, predictors, or both.
The "LINE" conditions check for:

- Linearity: The estimated value for the response must be a linear function of the predictors
- Independence: The errors are independent from each other
- Normality: The errors are normally distributed
- Equal variance: The errors have equal variances

We will verify that these conditions are satisfied through a series of statistical tests with R. Once the most optimal model has been selected, and "LINE" conditions met, the results can be analyzed and interpreted to answer the questions of interest and extract further insight about body fat percentage in men.

4. Regression Analysis, Results and Interpretation

The first step is to determine which predictors we should use in our model. A quick correlation matrix shows us that Density and Percent body fat are almost perfectly correlated, so it wouldn't make sense to include that as a predictor, as it would heavily skew the regression coefficients. The dataset description also states that Density was used directly to derive the Percent body fat for each subject, reinforcing this observation.

```
> cor(bodyfat)
              Density     Bodyfat
Density    1.00000000 -0.98778240
Bodyfat   -0.98778240  1.00000000
```

Figure 1: first two variables of correlation matrix

To further explore the data, scatterplots could give us insight on the relationships between our variables. From a glance, there seems to be consistent trends in the data when compared to Percent body fat, which is a great sign. A noticeable outlier, however, is Height, being almost completely horizontal in the plots, indicating that the correlation between it and the rest of the data may be weak at best.

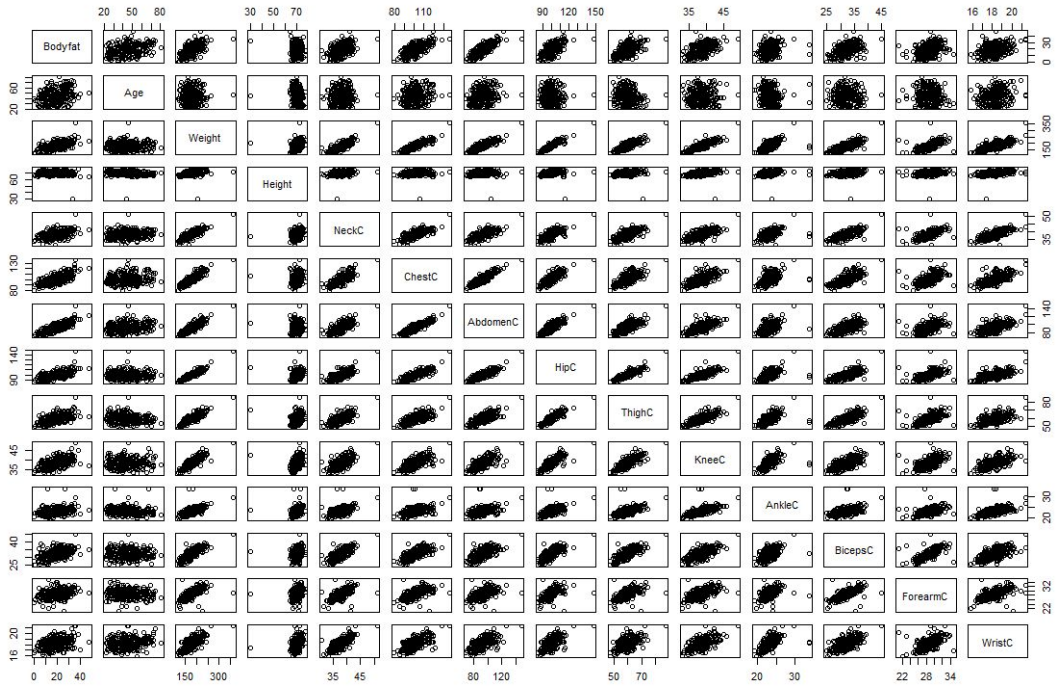

Figure 2: matrix of scatterplots

Next, we will use stepwise regression to find the best combination of predictors for our model. First the AIC method will be used, then subsequently the partial F test method.

```
Step:  AIC=738.98
Bodyfat ~ AbdomenC + Weight + WristC + ForearmC + NeckC + Age +
    ThighC + HipC

           Df Sum of Sq    RSS    AIC
<none>                  4437.7 738.98
- HipC      1     36.13 4473.9 739.02
+ BicepsC   1     23.02 4414.7 739.68
+ Height    1     12.94 4424.8 740.25
+ AnkleC    1     10.56 4427.2 740.39
+ ChestC    1      0.21 4437.5 740.97
+ KneeC     1      0.02 4437.7 740.98
- NeckC     1     76.94 4514.7 741.30
- Age       1     81.73 4519.5 741.57
- Weight    1     90.36 4528.1 742.04
- ThighC    1     93.99 4531.7 742.25
- ForearmC  1    134.46 4572.2 744.48
- WristC    1    170.31 4608.0 746.44
- AbdomenC  1   3143.18 7580.9 871.39

Call:
lm(formula = Bodyfat ~ AbdomenC + Weight + WristC + ForearmC +
    NeckC + Age + ThighC + HipC, data = bodyfat)
```

Figure 3: stepwise regression (AIC) results

The AIC method results indicate that Height, Chest circumference, Knee circumference, Ankle circumference, and Biceps circumference do not provide significant enough predictive power to our model and should be disregarded. Some caveats of stepwise regression are that there may be multiple equally performing "optimal models", and the procedure doesn't take into account non-statistical

knowledge about the predictors. Thus, we can't definitively say that this is the best model, but it's a great estimate, and close enough for quality analysis.

Now, we will use the partial F test method. This process starts with a model with no coefficients, and iteratively adds the predictor with the largest F test value to the model (as long as its p-value is above the chosen significance level). For this method, we chose a significance level α of 0.15, which is relatively large and allows the maximum number of relevant predictors to be allowed into the model.

The model chosen by stepwise regression with the partial F test method ended up excluding Height, Chest circumference, Hip circumference, Knee circumference, Ankle circumference, and Biceps circumference. The main and only difference between the AIC and partial F test model is that the AIC model did not exclude Hip circumference, while the partial F test model did. From our findings so far, the set of predictors found by the AIC method are better, as the regression model using them has a higher adjusted R squared score than the alternative found by the partial F test method, meaning it explains more of the variance in the data, and has better predictive power.

The last tool for model selection we will use is the leaps package. The leaps function performs an exhaustive search for the best subsets of the predictors for predicting the response, using an efficient branch-and-bound algorithm.

```
Selection Algorithm: exhaustive
         Age Weight Height NeckC ChestC AbdomenC HipC ThighC KneeC AnkleC BicepsC ForearmC WristC
1  ( 1 ) " " " "    " "    " "   " "    "*"      " "  " "    " "   " "    " "    " "      " "
2  ( 1 ) " " "*"    " "    " "   " "    "*"      " "  " "    " "   " "    " "    " "      " "
3  ( 1 ) " " "*"    " "    " "   " "    "*"      " "  " "    " "   " "    " "    " "      "*"
4  ( 1 ) " " "*"    " "    " "   " "    "*"      " "  " "    " "   " "    " "    "*"      "*"
5  ( 1 ) " " "*"    " "    "*"   " "    "*"      " "  " "    " "   " "    " "    "*"      "*"
6  ( 1 ) "*" "*"    " "    " "   " "    "*"      " "  "*"    " "   " "    " "    "*"      "*"
7  ( 1 ) "*" "*"    " "    "*"   " "    "*"      " "  "*"    " "   " "    " "    "*"      "*"
8  ( 1 ) "*" "*"    " "    "*"   " "    "*"      "*"  "*"    " "   " "    " "    "*"      "*"
> summary.mod$adjr2
[1] 0.6542717 0.7120745 0.7204548 0.7264415 0.7282402 0.7301009 0.7325661 0.7336296
```

Figure 4: leaps function results

For the figure above, the last row has the highest adjusted R squared, and we can see that it contains the exact same combination of predictors found by the AIC method stepwise regression. This reinforces its status as the "optimal model", thus we will use this model going forward.

Arguably the most important step in regression analysis is the interpretation of the residuals. Residuals must be distributed normally and have constant and equal variance for the model to be interpretable. We can make sure these conditions are met with plots.

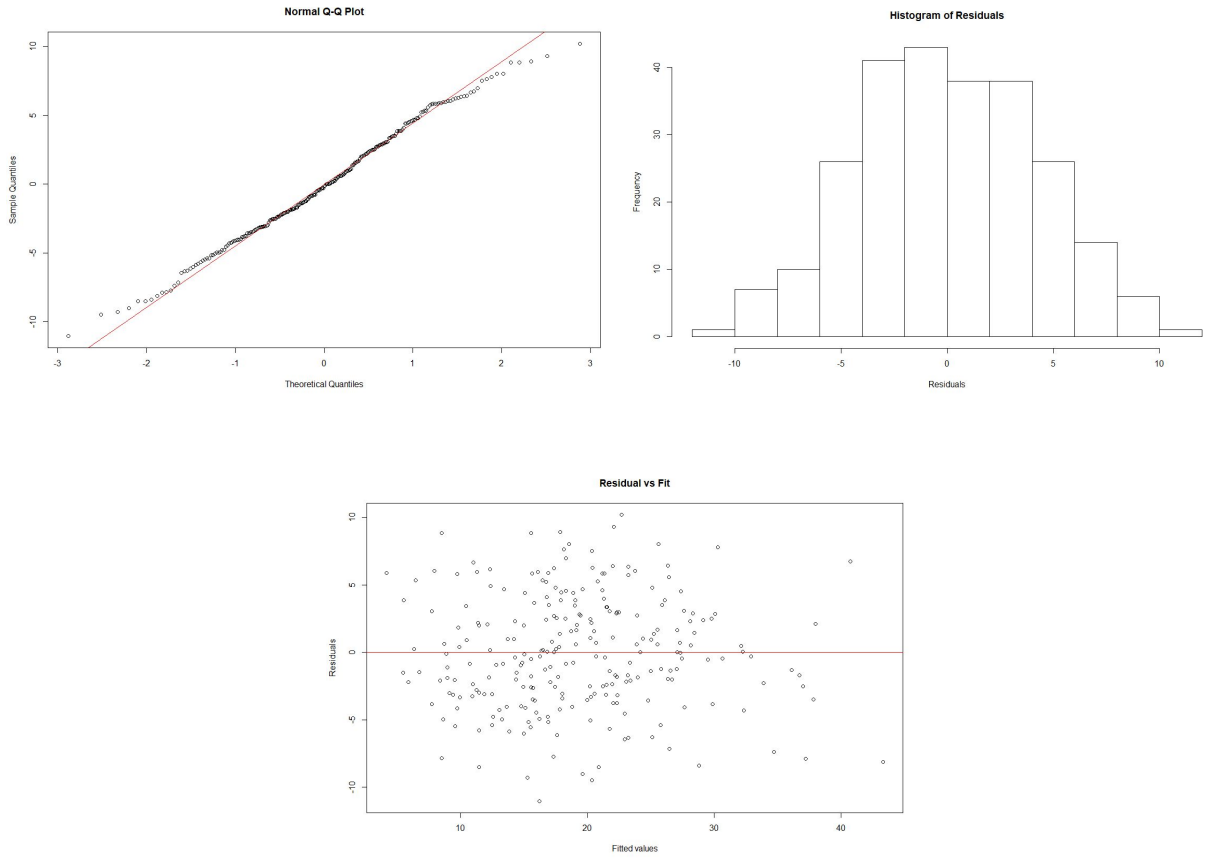

Figure 5: residual plots

The points hug the reference line in the Q-Q plot, and the histogram looks almost perfectly normal, meaning it's safe to say that our residuals are normally distributed. Additionally, in the Residual vs Fit plot, the points are nicely scattered around the reference line, and there is no trend in the variance. Therefore, the assumption that the residuals have equal variance holds true as well. With these plots, we have confirmed that the "LINE" conditions have been met.

Figuring out which transformation(s) one should apply on the data is crucial for building the best model possible. While exploring the data, we found that the response Percent body fat contains a singular 0 value, disabling the boxcox function from working. Thus, we removed the problematic row, and ran boxcox on the reduced model.

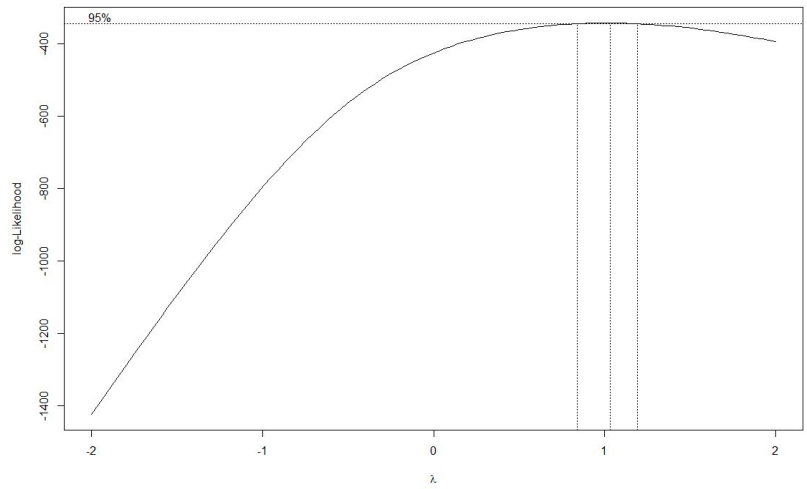

Figure 6: boxcox results

The value of λ was found to be roughly 1.03, indicating that no transformation is needed.

To further reinforce that the new/reduced model is better than the full model (all the original predictors included), we will conduct an F-test of overall significance.

$$H_0 : \text{The fit of the reduced model and full model are equal.}$$
$$H_1 : \text{The fit of the reduced model is significantly worse than the full model.}$$

Using ANOVA, the F-test yields an extremely high p-value of 0.7658, meaning that we fail to reject the null hypothesis that the fit of the reduced and full model are equal. A good rule of thumb for predictive modeling is that a simple model is always preferable to an equal performing and more complicated one. As we just found that the reduced model does not have a significantly worse fit than the full model, we will use the reduced model as the final model.

$$Bodyfat = -21.78 + 0.94 * AbdomenC - 0.09 * Weight - 1.55 * WristC + 0.51 * ForearmC - 0.46 * NeckC + 0.07 * Age + 0.29 * ThighC - 0.19 * HipC + \epsilon$$

Figure 7: final regression equation (rounded to hundreths)

5. Conclusion

We conducted this analysis with the intent of solving the questions of interest proposed earlier in this paper. The regression coefficients in Figure 7 give us strong clues to how the predictors influence body fat percentage in men.

I. Which metric is most strongly correlated with a high body fat percentage?

Abdomen circumference has the strongest positive linear relationship with Percent body fat. A bigger belly/abdomen is generally known to be associated with more body fat. High mass individuals that exercise store dense muscle tissues in areas other than their abdomen, and having visible abdominal muscles is a prestigious indicator that an individual has a low body fat percentage. To our eyes an enlarged abdominal area is important to determining high body fat, and this observation holds up statistically as well.

The other main notable positive correlations with Percent body fat are Forearm circumference, Thigh circumference, and Age, albeit to a lesser extent than Abdomen circumference. While the coefficient for Age is relatively small, the variable's variance is much larger, so it still has a significant impact. The model values 10 years of Age the same as approximately two inches of Thigh circumference when concerning a 1% increase in body fat percentage.

```
Range of Age: 22 81                          Variance: 159.3508
Range of Forearm circumference: 21 34.9      Variance: 4.033202
Range of Thigh circumference: 49.3 87.3      Variance: 27.07393
```

Figure 8: range and variance of other notable positive correlations

II. Which metric is most strongly correlated with a low body fat percentage?

Wrist circumference is the largest coefficient in the model and has the strongest negative linear relationship with Percent body fat. According to a study by A. Öztürk et al. (2017), "WrC [wrist circumference] is a simple, easy-to-detect anthropometric index which is not subject to measurement errors. Additionally, WrC can be used both to decide about frame size and to determine metabolic risks related to obesity".[1] Wrist circumference's reliable relationship with frame size gives us context on why the coefficient is so large in the model, and suggests that a person with larger wrists should have a larger frame that can hold more weight, lowering expected body fat percentage.

Hip circumference and Neck circumference are two other variables with a negative linear relationship to Percent body fat, as is Weight, surprisingly enough.

III. Is high weight correlated with a high body fat percentage?

From the coefficients of the final model, Weight is shown to have a negative linear relationship with Percent body fat. This might seem counterintuitive, as we tend to associate a higher weight with more body fat, but there are reasons this occurs. An individual's weight measures all of their mass, not just body fat. Muscles, bones, organs, and other matter make up the majority of a person's weight.

A muscular person could be heavier than someone with a high body fat percentage at the same height, as muscle tissue is more dense than fat, and in this case weight would be a terrible predictor if a positive linear relationship was assumed. The mean of Percent body fat in the data is 19.2% (including all age groups), significantly under the national average of 22.9% for boys ages 16-19 to 30.9% for men ages 60-79.[2] This means the sample of men in this dataset are in better shape than the average American male, further contributing to the negative linear relationship observed by the model between Weight and Percent body fat due to situations like the one described at the beginning of this paragraph.

From this analysis, we have gained some insight on what affects body fat percentage in men. Further exploration of this topic could include collecting additional data, adding more metrics, or doing a similar analysis but for women instead.

6. Appendix

1. Öztürk, A., Çiçek, B., Mazıcıoğlu, M. M., Zararsız, G., & Kurtoğlu, S. (2017). Wrist Circumference and Frame Size Percentiles in 6-17-Year-Old Turkish Children and Adolescents in Kayseri. Journal of Clinical Research in Pediatric Endocrinology, 329–336. https://doi.org/10.4274/jcrpe.4265

2. "QuickStats: Mean Percentage Body Fat,* by Age Group and Sex --- National Health and Nutrition Examination Survey, United States, 1999--2004†". cdc.gov.

Dataset URL: http://lib.stat.cmu.edu/datasets/bodyfat

Code

```
# Load in the data
bodyfat <- read.csv("bodyfat.csv", header=FALSE)
colnames(bodyfat) <- c("Density", "Bodyfat", "Age", "Weight", "Height", "NeckC", "ChestC",
"AbdomenC", "HipC", "ThighC", "KneeC", "AnkleC", "BicepsC", "ForearmC", "WristC")

# Correlation matrix
cor(bodyfat)

# Remove density
bodyfat <- subset(bodyfat, select = -c(Density))
```

```
# Scatterplot matrix
pairs(bodyfat)

# Remove row containg 0 value
bodyfat <- bodyfat[-which(bodyfat$Bodyfat == 0),]

model <- lm(Bodyfat ~ Age + Weight + Height +
          NeckC + ChestC + AbdomenC + HipC +
          ThighC + KneeC + AnkleC + BicepsC + ForearmC + WristC, data=bodyfat)

# Stepwise regression (AIC)
reduced <- lm(Bodyfat ~ 1, data = bodyfat)
step(reduced, scope = list(lower = reduced, upper = model))

newModel <- lm(formula = Bodyfat ~ AbdomenC + Weight + WristC + ForearmC +
                    NeckC + Age + ThighC + HipC, data = bodyfat)

# Stepwise Regression (Partial F test) (alpha set at 0.15)
x1 <- bodyfat$Age
x2 <- bodyfat$Weight
x3 <- bodyfat$Height
x4 <- bodyfat$NeckC
x5 <- bodyfat$ChestC
x6 <- bodyfat$AbdomenC
x7 <- bodyfat$HipC
x8 <- bodyfat$ThighC
x9 <- bodyfat$KneeC
x10 <- bodyfat$AnkleC
x11 <- bodyfat$BicepsC
x12 <- bodyfat$ForearmC
x13 <- bodyfat$WristC
y <- bodyfat$Bodyfat
mod0 <- lm(y ~ 1)
add1(mod0, ~ . + x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10 + x11 + x12 + x13, test = 'F')
mod1 <- update(mod0, ~. + x6)
add1(mod1, ~. + x1 + x2 + x3 + x4 + x5 + x7 + x8 + x9 + x10 + x11 + x12 + x13, test = 'F')
mod2 <-update(mod1, ~. + x2)
summary(mod2)
add1(mod2, ~. + x1 + x3 + x4 + x5 + x7 + x8 + x9 + x10 + x11 + x12 + x13, test = 'F')
mod3 <-update(mod2, ~. + x13)
summary(mod3)
add1(mod3, ~. + x1 + x3 + x4 + x5 + x7 + x8 + x9 + x10 + x11 + x12, test = 'F')
mod4 <- update(mod3, ~. + x12)
summary(mod4)
add1(mod4, ~. + x1 + x3 + x4 + x5 + x7 + x8 + x9 + x10 + x11, test = 'F')
mod5 <- update(mod4, ~. + x4)
summary(mod5)
add1(mod5, ~. + x1 + x3 + x5 + x7 + x8 + x9 + x10 + x11, test = 'F' )
mod6 <- update(mod5, ~. + x11)
```

```
summary(mod6)
add1(mod6, ~. + x1 + x3 + x5 + x7 + x8 + x9 + x10, test = 'F')
mod7 <- update(mod6, ~. + x1)
summary(mod7)
add1(mod7, ~. + x3 + x5 + x7 + x8 + x9 + x10, test = 'F')
mod8 <- update(mod7, ~. + x8)
summary(mod8)
# BicepC p-value is too high, higher than alpha = 0.15, remove it
mod8 <- update(mod8, ~. - x11)
summary(mod8)
add1(mod8, ~. + x3 + x5 + x7 + x9 + x10, test = 'F')
newModel2 <- lm(Bodyfat ~ x6 + x2 + x13 + x12 + x4 + x1 + x8)

# We will chose newModel as our final model, as it has a higher adjusted R squared.
summary(newModel)
summary(newModel2)

# Leaps Function (results agree with AIC model)
library(leaps)
mod <- regsubsets(subset(bodyfat, select=-c(Bodyfat)), bodyfat$Bodyfat)
summary.mod <- summary(mod)
summary.mod$which
summary.mod
summary.mod$adjr2

# Q-Q Plot
qqnorm(newModel$residuals)
qqline(newModel$residuals, col="red")

# Histogram
hist(newModel$residuals, main="Histogram of Residuals", xlab="Residuals")

# Residual vs Fit
plot(x=newModel$fitted.values, y=newModel$residuals, main="Residual vs Fit", xlab="Fitted values",
ylab="Residuals")
abline(0,0, col="red")

# Boxcox
library(MASS)
bc <- boxcox(newModel)
lambda <- bc$x[which(bc$y == max(bc$y))]
lambda

# F-test of overall significance
anova(model, newModel)

# Extract equation of final model
coefficients <- newModel$coefficients
(eqn <- paste("Bodyfat =", paste(round(coefficients[1],2), paste(round(coefficients[-1],2),
names(coefficients[-1]), sep=" * ", collapse=" + "), sep=" + "), "+ e"))
```

```
# Figure 8 code
cat("Range of Age:", range(bodyfat$Age), "       Variance:", var(bodyfat$Age))
cat("Range of Forearm circumference:", range(bodyfat$ForearmC), "    Variance:",
var(bodyfat$ForearmC))
cat("Range of Thigh circumference:", range(bodyfat$ThighC), "    Variance:", var(bodyfat$ThighC))
```