# Jin_Kweon_3032235207_HW2

*Jin Kweon*

*10/10/2017*

SIZE: 252 observations, 19 variables

DESCRIPTIVE ABSTRACT: Percentage of body fat, age, weight, height, and ten body circumference measurements (e.g., abdomen) are recorded for 252 men. Body fat, a measure of health, is estimated through an underwater weighing technique. Fitting body fat to the other measurements using multiple regression provides a convenient way of estimating body fat for men using only a scale and a measuring tape.

VARIABLE DESCRIPTIONS:

Columns

1 Case Number

2 Percent body fat using Brozek's equation, 457/Density - 414.2

3 Percent body fat using Siri's equation, 495/Density - 450

4 Density (gm/cm^3)

5 Age (yrs)

6 Weight (lbs)

7 Height (inches)

8 Adiposity index = Weight/Height^2 (kg/m^2)

9 Fat Free Weight = (1 - fraction of body fat) * Weight, using Brozek's formula (lbs)

10 Neck circumference (cm)

11 Chest circumference (cm)

12 Abdomen circumference (cm) "at the umbilicus and level with the iliac crest"

13 Hip circumference (cm)

14 Thigh circumference (cm)

15 Knee circumference (cm)

16 Ankle circumference (cm)

17 Extended biceps circumference (cm)

18 Forearm circumference (cm)

19 Wrist circumference (cm) "distal to the styloid processes"

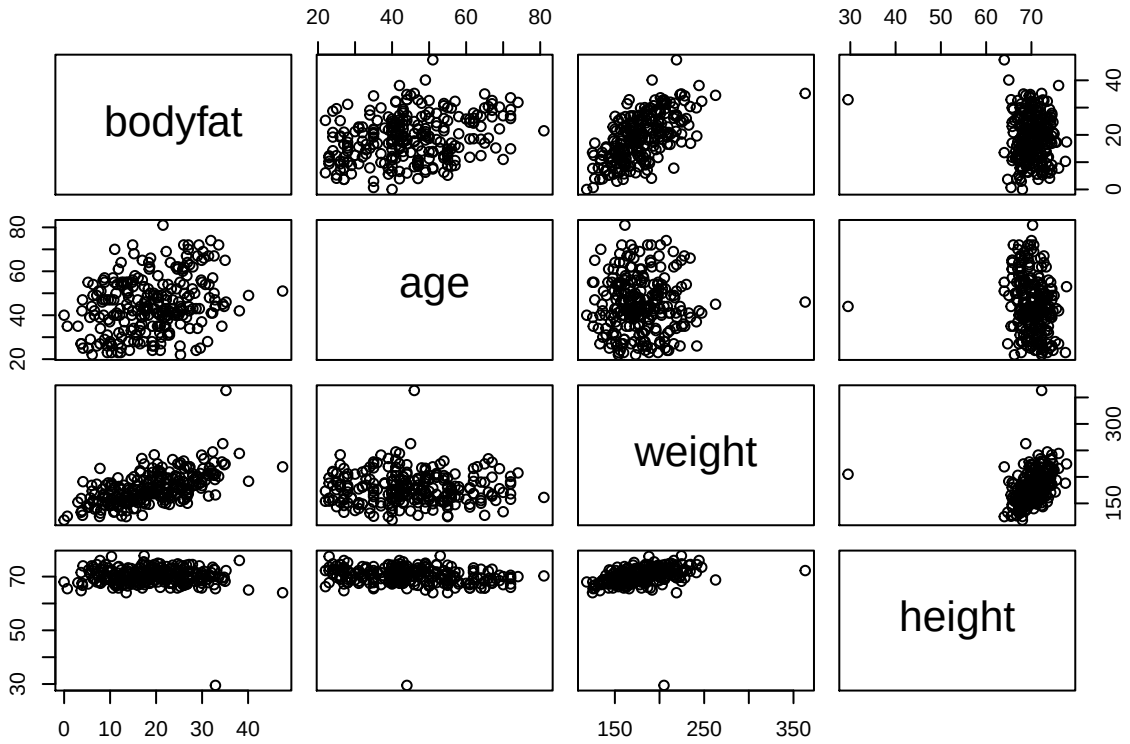
## Data import

```
data <- read.table("Data-HW2-Bodyfat.txt")
colnames(data) <- c("case_#", "Brozek_body_fat_%", "bodyfat", "density",
                    "age", "weight", "height", "adiposity", "fat_free", "neck", "chest",
                    "abdomen", "hip", "thigh", "knee", "ankle", "bicep", "forearm", "wrist")
dim(data)
```

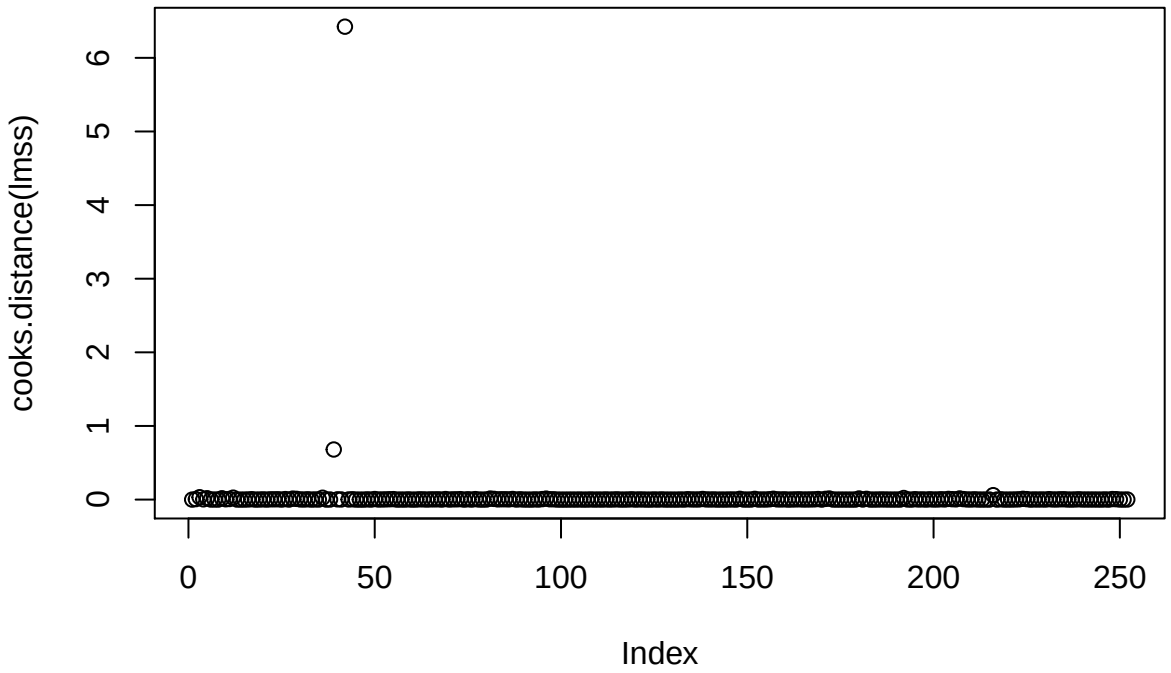
```
## [1] 252  19
```

## Remove outliers

```
plot(data[,c(3, 5, 6, 7)])
```



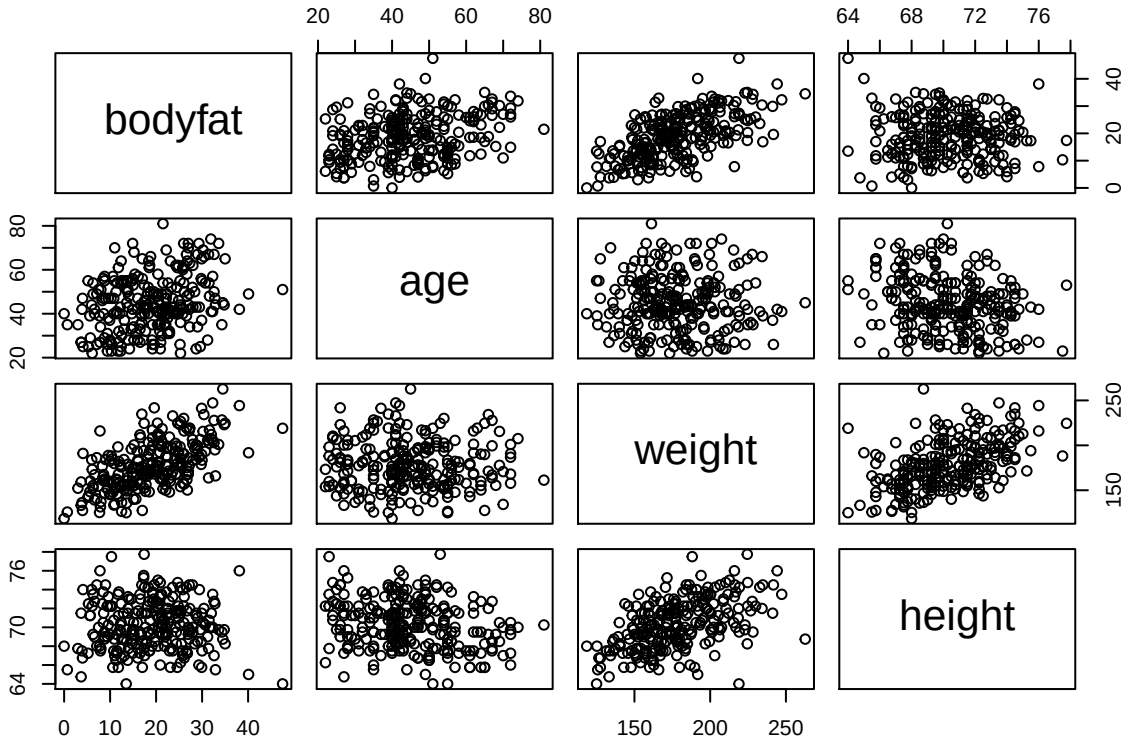

```
lmss <- lm(bodyfat ~ age + weight + height, data = data)
plot(cooks.distance(lmss)) #shows 39th and 42th observations are outliers and having high leverage.
```




```
outlier1 <- order(data[,6], decreasing = T)[1]
outlier2 <- order(data[,7])[1]
#These outliers show 39th and 42th!!!

newdata <- data[-c(outlier1, outlier2),]
dim(newdata)
```

```
## [1] 250  19
```

```
plot(newdata[,c(3, 5, 6, 7)])
```



*Comment:*

I plotted the same plots as on the slide 6 from the lecture note, and grasp some idea what I should remove. I renamed to "newdata" after I removed two outliers.

# Part a

```
partadata <- newdata[,c(3:7, 10:19)]
partadata <- partadata[,-2]

lin_fit <- lm(bodyfat ~., data = partadata)
summary(lin_fit)
```

```
##
## Call:
## lm(formula = bodyfat ~ ., data = partadata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.9900  -3.1244  -0.1674   3.0248   9.8648
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.68516   23.37412   0.072 0.942587
## age          0.07189    0.03217   2.234 0.026389 *
## weight      -0.01762    0.06714  -0.263 0.793153
## height      -0.24675    0.19114  -1.291 0.197989
## neck        -0.38682    0.23486  -1.647 0.100887
## chest       -0.11919    0.10825  -1.101 0.272004
## abdomen      0.90452    0.09140   9.897  < 2e-16 ***
## hip         -0.15878    0.14586  -1.089 0.277446
## thigh        0.17299    0.14683   1.178 0.239926
## knee        -0.04580    0.24560  -0.186 0.852230
## ankle        0.18502    0.21985   0.842 0.400862
## bicep        0.17968    0.17039   1.054 0.292732
## forearm      0.27605    0.20692   1.334 0.183454
## wrist       -1.80162    0.53304  -3.380 0.000848 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 4.255 on 236 degrees of freedom
## Multiple R-squared:  0.7505, Adjusted R-squared:  0.7368
## F-statistic: 54.61 on 13 and 236 DF,  p-value: < 2.2e-16
```

*Comment:*

I fitted a linear regression with lm() function.

# Part b

```
summary(lin_fit)
```

```
##
## Call:
## lm(formula = bodyfat ~ ., data = partadata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.9900  -3.1244  -0.1674   3.0248   9.8648
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.68516   23.37412   0.072 0.942587
## age          0.07189    0.03217   2.234 0.026389 *
## weight      -0.01762    0.06714  -0.263 0.793153
## height      -0.24675    0.19114  -1.291 0.197989
## neck        -0.38682    0.23486  -1.647 0.100887
## chest       -0.11919    0.10825  -1.101 0.272004
## abdomen      0.90452    0.09140   9.897  < 2e-16 ***
## hip         -0.15878    0.14586  -1.089 0.277446
## thigh        0.17299    0.14683   1.178 0.239926
## knee        -0.04580    0.24560  -0.186 0.852230
## ankle        0.18502    0.21985   0.842 0.400862
## bicep        0.17968    0.17039   1.054 0.292732
## forearm      0.27605    0.20692   1.334 0.183454
## wrist       -1.80162    0.53304  -3.380 0.000848 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.255 on 236 degrees of freedom
## Multiple R-squared:  0.7505, Adjusted R-squared:  0.7368
```

```
## F-statistic: 54.61 on 13 and 236 DF,  p-value: < 2.2e-16
```

*Comment:*

First, estimated coefficient from OLS for age is 0.07189. So, I can tell that age has positive relationship with bodyfat. And, when we get aged 1-year, bodyfat will increase 0.07189 (in bodyfat variable unit).

Second, if I say the null hypothesis is that the coefficient of age equals to zero, p-value will be 0.026389 (from t-test: $\frac{\hat{\beta_{age}} - \beta_{age}}{s.e.(\beta_{age})} \sim t_{n-p-1}$).

Third, with the signficance level of 0.05, I would reject the null, as p-value is smaller than 0.05. So, it means that coefficient of age should not be zero. (we just reject the null: $\beta_{age} = 0$)

# Part c

```
summary(lin_fit)
```

```
##
## Call:
## lm(formula = bodyfat ~ ., data = partadata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.9900  -3.1244  -0.1674   3.0248   9.8648
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.68516   23.37412   0.072 0.942587
## age          0.07189    0.03217   2.234 0.026389 *
## weight      -0.01762    0.06714  -0.263 0.793153
## height      -0.24675    0.19114  -1.291 0.197989
## neck        -0.38682    0.23486  -1.647 0.100887
## chest       -0.11919    0.10825  -1.101 0.272004
## abdomen      0.90452    0.09140   9.897  < 2e-16 ***
## hip         -0.15878    0.14586  -1.089 0.277446
## thigh        0.17299    0.14683   1.178 0.239926
## knee        -0.04580    0.24560  -0.186 0.852230
## ankle        0.18502    0.21985   0.842 0.400862
## bicep        0.17968    0.17039   1.054 0.292732
## forearm      0.27605    0.20692   1.334 0.183454
## wrist       -1.80162    0.53304  -3.380 0.000848 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.255 on 236 degrees of freedom
## Multiple R-squared:  0.7505, Adjusted R-squared:  0.7368
## F-statistic: 54.61 on 13 and 236 DF,  p-value: < 2.2e-16
```

*Comment:*

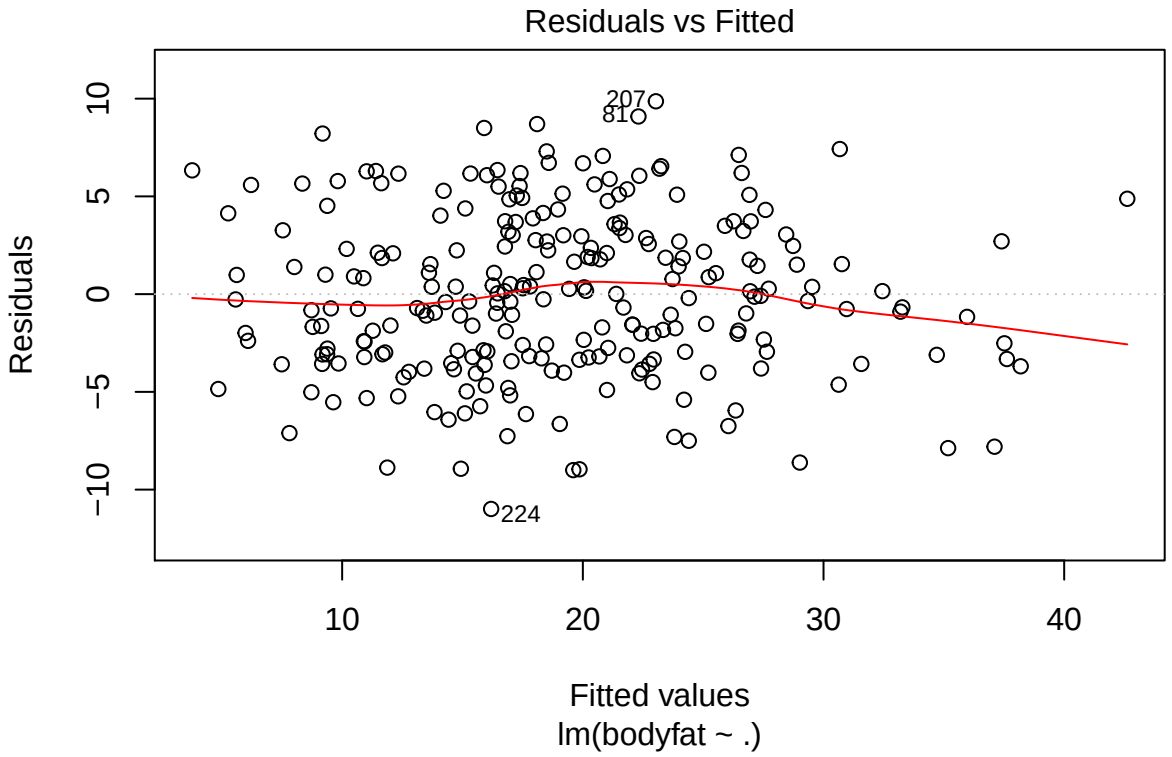
First, estimated coefficient from OLS for abdomen is 0.90452. So, I can tell that abdomen has strong positive relationship with bodyfat. And, when we get abdomen circumference increased 1 cm, bodyfat will increase 0.90452 (in bodyfat variable unit).

Second, if I say the null hypothesis is that the coefficient of age equals to zero, p-value is less than 2e-16 (from t-test: $\frac{\hat{\beta_{abdomen}} - \beta_{abdomen}}{s.e.(\beta_{abdomen})} \sim t_{n-p-1}$).

Third, with the signficance level of 0.05, I would reject the null, as p-value is smaller than 0.05. So, it means that coefficient of abdomen should not be zero. (we just reject the null: $\beta_{abdomen} = 0$)

# Part d

```
plot(lin_fit, which = 1)
```




*Comment:*

I simply used the function embedded in R, to draw a residual plot. The red line in the diagram shows the loess fitted curve/line.

The red fitted line (and the dots) form a constant horizontal band along the residual = 0, so I can say this model is linear. Also, the variance of residual seems quite constant (although it slightly goes down at the end), meaning error is homoscedasticitic. Last but not least, there seems to be no pattern in this plot. (implying this model is linear)

In conlcusion, I will *not* say there is any significant violation of the key assumptions of linear model.

All these came from the key assumptions:

1. $y|X$ is normally distribued
2. error is normally distributed with mean of 0 and variance of $\sigma^2$ and is indepent with X (it implies that $E(y|X) = X\beta$)
3. var$(y|X)$ is constant with $\sigma^2$.
4. Error is i.i.d. (identical independent distributed) normally distributed (thus $y|X$ is normally distributed).
5. We usually say n > p.

## Part e

```
summary(lin_fit) #full model
```

```
##
## Call:
## lm(formula = bodyfat ~ ., data = partadata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.9900  -3.1244  -0.1674   3.0248   9.8648
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.68516   23.37412   0.072 0.942587
## age          0.07189    0.03217   2.234 0.026389 *
## weight      -0.01762    0.06714  -0.263 0.793153
## height      -0.24675    0.19114  -1.291 0.197989
## neck        -0.38682    0.23486  -1.647 0.100887
## chest       -0.11919    0.10825  -1.101 0.272004
## abdomen      0.90452    0.09140   9.897  < 2e-16 ***
## hip         -0.15878    0.14586  -1.089 0.277446
## thigh        0.17299    0.14683   1.178 0.239926
## knee        -0.04580    0.24560  -0.186 0.852230
## ankle        0.18502    0.21985   0.842 0.400862
## bicep        0.17968    0.17039   1.054 0.292732
## forearm      0.27605    0.20692   1.334 0.183454
## wrist       -1.80162    0.53304  -3.380 0.000848 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.255 on 236 degrees of freedom
## Multiple R-squared:  0.7505, Adjusted R-squared:  0.7368
## F-statistic: 54.61 on 13 and 236 DF,  p-value: < 2.2e-16
```

```
partedata <- partadata[,1:4]

lin_small_fit <- lm(bodyfat ~., data = partedata)

summary(lin_small_fit) #small model
```

```
##
## Call:
## lm(formula = bodyfat ~ ., data = partedata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.9223  -3.9437   0.0327   3.9586  12.8856
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 57.27217   10.39897   5.507 9.15e-08 ***
## age          0.13732    0.02806   4.895 1.78e-06 ***
## weight       0.25366    0.01483  17.110  < 2e-16 ***
## height      -1.27416    0.15801  -8.064 3.24e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.382 on 246 degrees of freedom
## Multiple R-squared:  0.584,  Adjusted R-squared:  0.579
## F-statistic: 115.1 on 3 and 246 DF,  p-value: < 2.2e-16
```

*Comment:*

lin_fit is the summary of the big model, and lin_small_fit is the summary of the small model. We decided to use adjusted $R^2$ instead of multiple/normal $R^2$, since $R^2$ has tendency to decrease (or sometimes stays the same) as taking out the explanatory variables.

Adjusted $R^2$ is significantly decreases, and it is telling me that the proportion of variance in the response variable can be better explained by the linear model in full model.

To check the sgnificant predictors for both models, I will check the p-value of t-statistics for every coefficient for both models. (Significance means to check whether that coefficient is significant in the model; thus, having $H_0 : coef = 0$. When p-value is less than the significant level, then I will reject the null, implying that coefficient is significnat.) In full model, we had 3 significant predictors (p-value less than 0.05) out of 13 explanatory variables, but in small model, all of 3 predictors are sinificant. It is telling me that I cannot take out any more variable from the small model (less flexible...), but I can take some of unsignificant predictors from the full model. Also, some of the variables that were insignificant in full model such as weight and height are significant in small model. So, I will say the full model in this case is a bit more preferrable.

F-statistics (null is all coefficients of explanatory variables are zero) in both summary are telling me that linear model is pretty good.

# Part f

```r
#fitcompare <- lm(bodyfat ~ age + weight + height, data = partadata)
#summary(fitcompare)


#The other way to do it.
rss1 <- sum((partadata$bodyfat - lin_fit$fitted.values)^2)
rss0 <- sum((partadata$bodyfat - lin_small_fit$fitted.values)^2)
```

```r
f <- ((rss0 - rss1) / 10) / (rss1 / (nrow(partadata) - 13 - 1))
f
```

```
## [1] 15.74523
```

```r
pf(f, 10, nrow(partadata) - 13 - 1, lower.tail = F)
```

```
## [1] 1.517719e-21
```

```r
#The other way to do it.
L <- cbind(rep(0,10), rep(0,10), rep(0,10), rep(0,10),
                  c(1, 0, 0, 0, 0, 0, 0, 0, 0, 0), c(0, 1, 0, 0, 0, 0, 0, 0, 0, 0),
                  c(0, 0, 1, 0, 0, 0, 0, 0, 0, 0), c(0, 0, 0, 1, 0, 0, 0, 0, 0, 0),
                  c(0, 0, 0, 0, 1, 0, 0, 0, 0, 0), c(0, 0, 0, 0, 0, 1, 0, 0, 0, 0),
                  c(0, 0, 0, 0, 0, 0, 1, 0, 0, 0), c(0, 0, 0, 0, 0, 0, 0, 1, 0, 0),
                  c(0, 0, 0, 0, 0, 0, 0, 0, 1, 0), c(0, 0, 0, 0, 0, 0, 0, 0, 0, 1))
beta_hat <- as.matrix(lin_fit$coefficients)
X <- as.matrix(cbind(1, lin_fit$model))[,-2]

Fstat2 <- as.numeric(t(L %*% beta_hat) %*% solve(L %*% solve(t(X) %*% X) %*% t(L)) %*% (L %*% beta_hat)

Fstat2
```

```
## [1] 15.74523
```

```r
pf(Fstat2, 10, nrow(partadata) - 13 - 1, lower.tail = F)
```

```
## [1] 1.517719e-21
```

*Comment:*

I will use F-statistics (our null is the coefficients of the ten body circumference measurements are zero = reduced model is preferred) and get p-value to compare those two models.

And, our null $H_0$ is $\beta_{neck} = \beta_{chest} = \beta_{abdomen} = \beta_{hip} = \beta_{thigh} = \beta_{knee} = \beta_{ankle} = \beta_{bicep} = \beta_{forearm} = \beta_{wrist} = 0$.

Then, $F_{q,\ n-p-1} \sim \frac{(RSS(m)\ -\ RSS(M))/q}{RSS(M)/n-p-1}$, where RSS(m) is the RSS for the small model (= reduced model), RSS(M) is the RSS for the big model (= full model), q is the number of variables dropped (or number of constraints), and p is the number of explanatory variables from the full model.

Or, I can use formula $\frac{(L\hat{\beta}-c)^\top [L(X^\top X)^{-1}L^\top]^{-1}(L\hat{\beta}-c)/q}{\text{RSS}(M)/(n-p-1)}$.

**F statistic value:**

The F statistic value is approimately 15.745.

**Degree of freedom:**

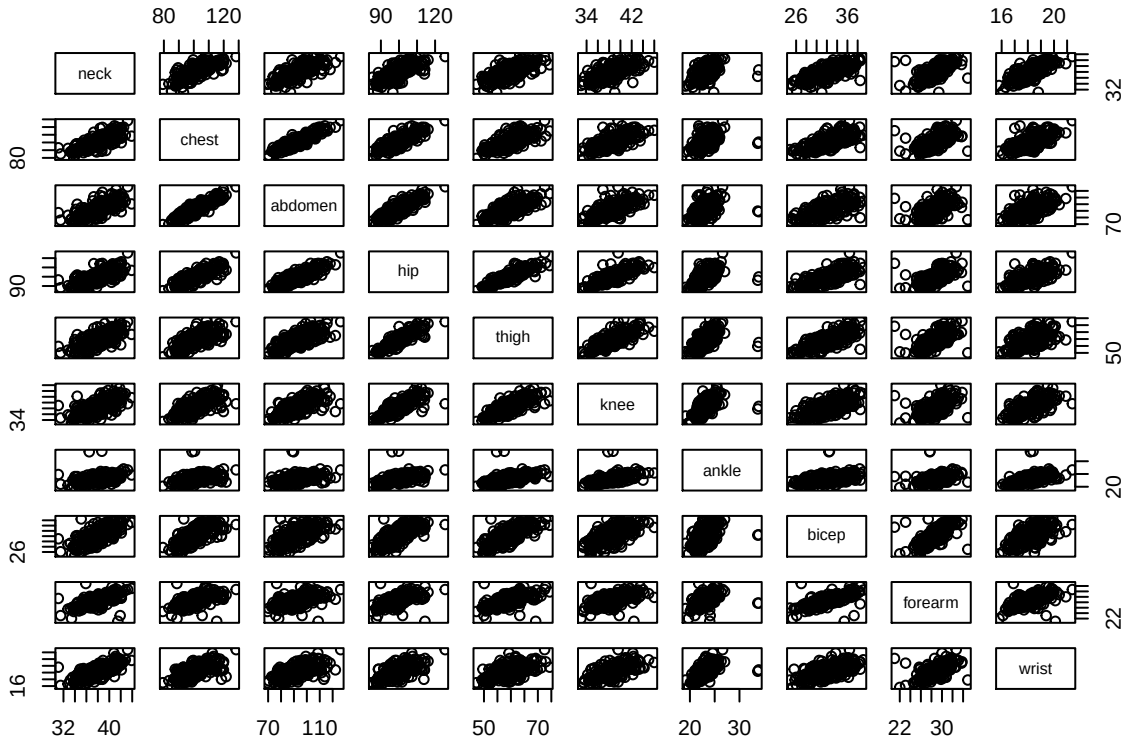
Degree of freedom will be 10 and 236.

**p value:**

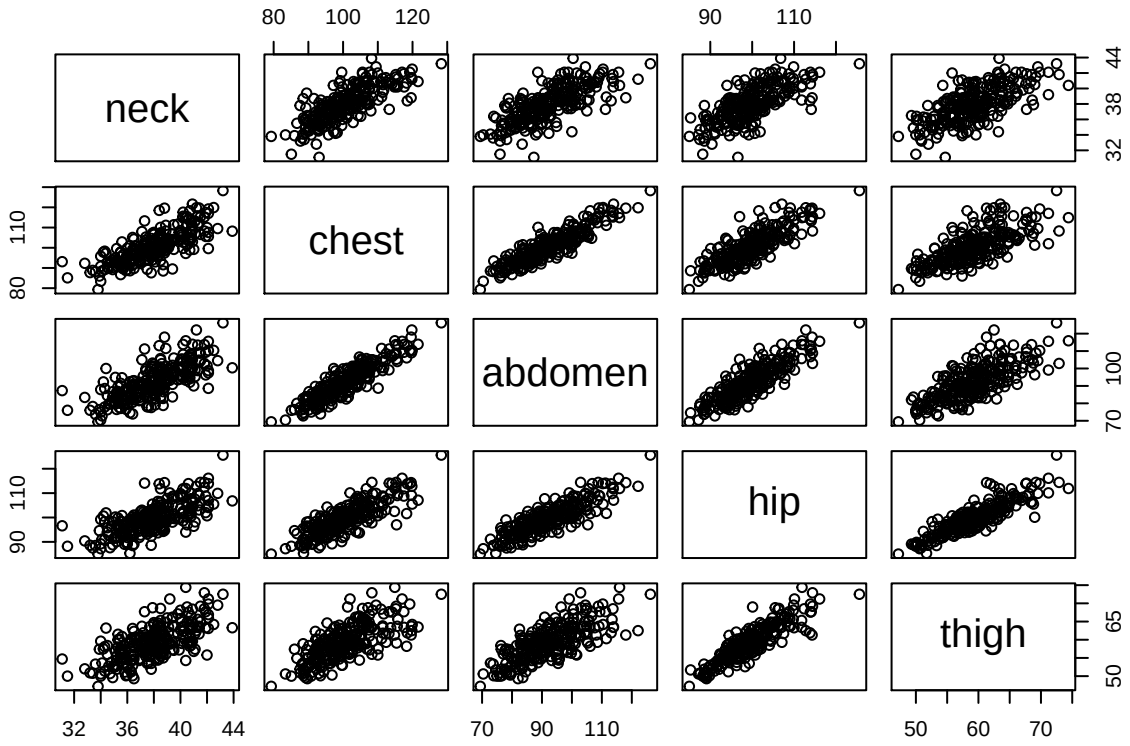
The P-value is around approximately 0, so I would reject the null. It implies that it is better for me to use full model.

# Part g

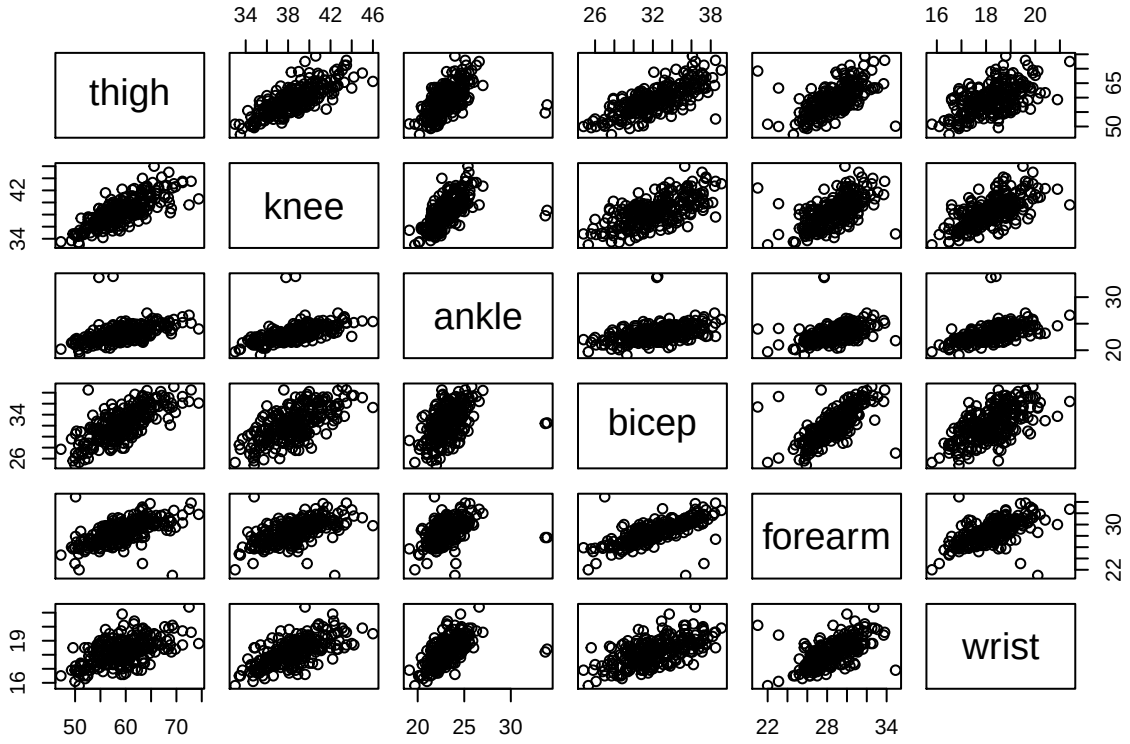
```
plot(partadata[,5:ncol(partadata)])
```




```
plot(partadata[,5:9])
```




```
plot(partadata[,9:ncol(partadata)])
```




```
#Extra
head(order(partadata$ankle, decreasing = T), 2)
```

```
## [1] 31 84
```

```r
head(sort(partadata$ankle, decreasing = T), 2)
```

```
## [1] 33.9 33.7
```

*Comment:*

Most of them have positive relationship with each variable, which makes sense to me. I can see two outlier/unusual leverage/influential observations (need to investigate further what that points are) on the ankle variable. (31th: 33.9 and 84th: 33.7)

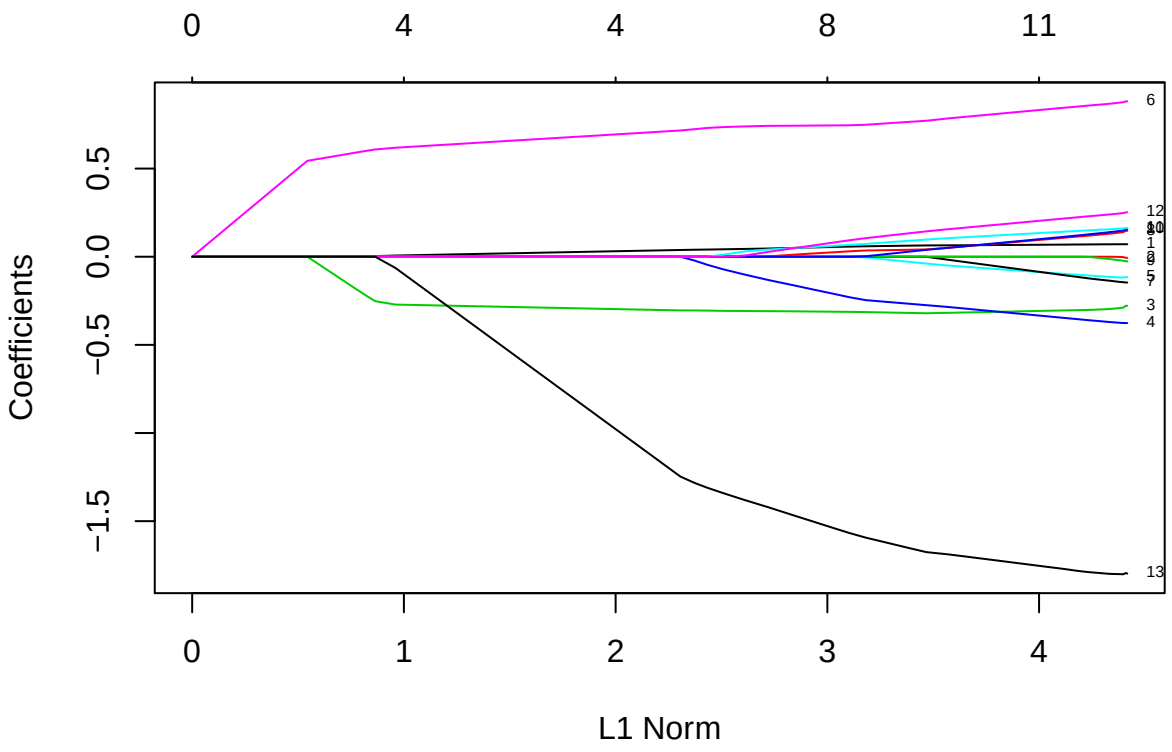
# Part h

When alpha is equal to 1, I fit a lasso model. (0: ridge & 1: lasso)

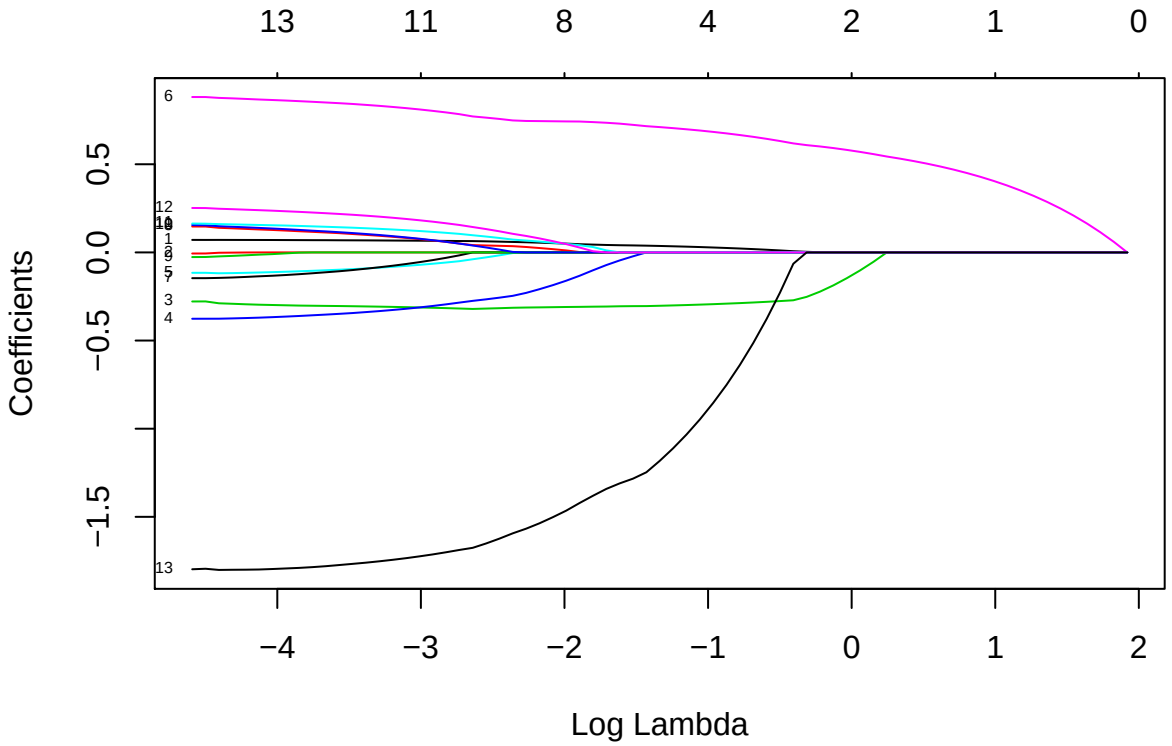
```r
lasso <- glmnet(as.matrix(partadata[,-1]), as.matrix(partadata[,1]), alpha = 1)

summary(lasso)
```

```
##           Length Class     Mode
## a0         71    -none-    numeric
## beta      923    dgCMatrix S4
## df         71    -none-    numeric
## dim         2    -none-    numeric
## lambda     71    -none-    numeric
## dev.ratio  71    -none-    numeric
## nulldev     1    -none-    numeric
## npasses     1    -none-    numeric
## jerr        1    -none-    numeric
## offset      1    -none-    logical
## call        4    -none-    call
## nobs        1    -none-    numeric
```

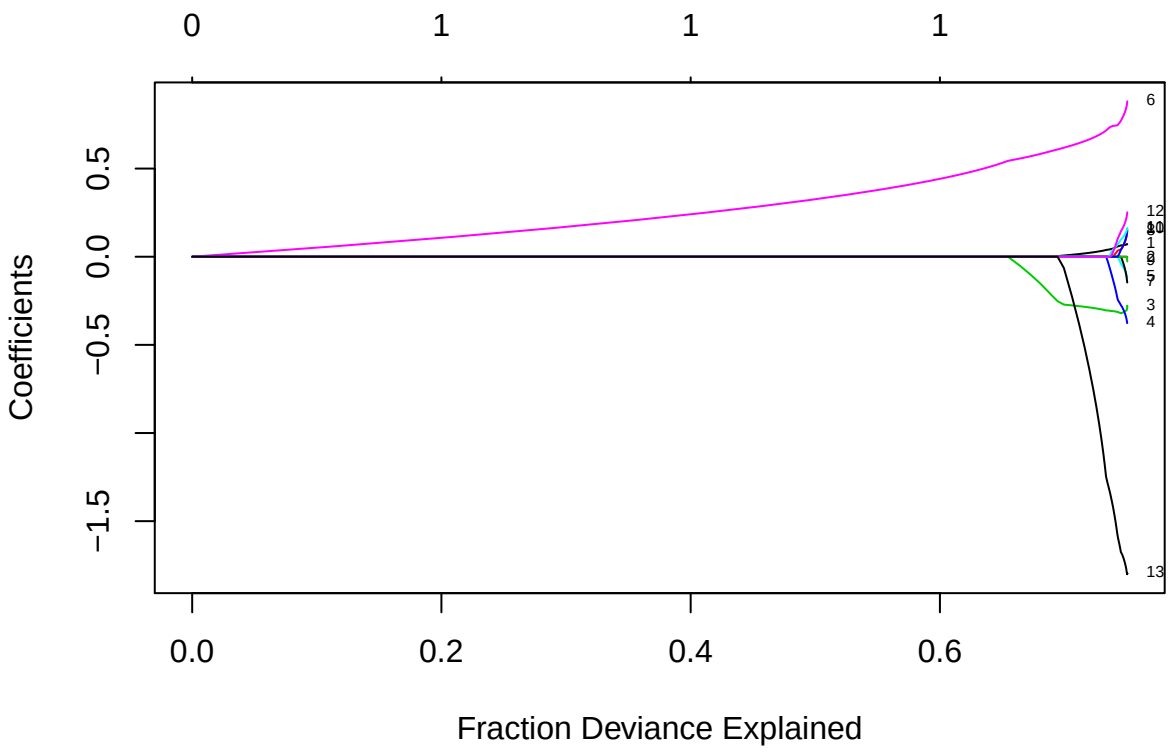
```r
plot(lasso, label = T)
```



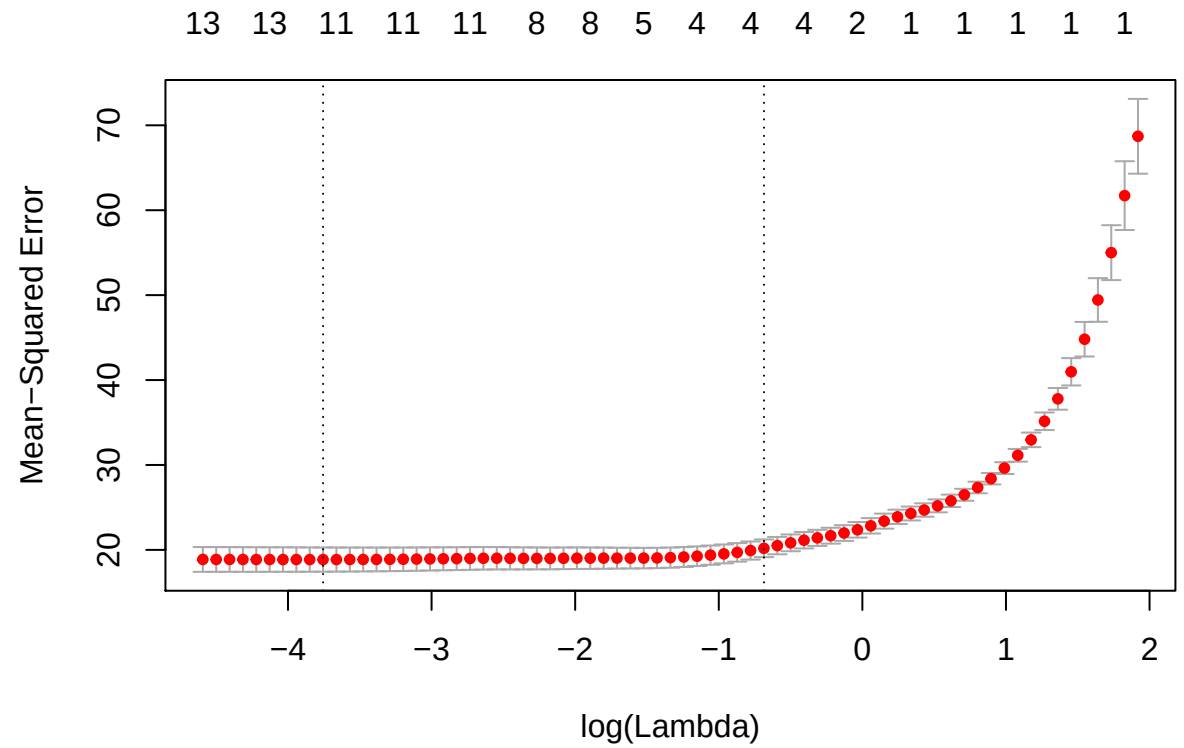

```
plot(lasso, xvar = "lambda", label = T)
```




```
plot(lasso, xvar = "dev", label = T)
```




```
cvlasso <- cv.glmnet(as.matrix(partadata[,-1]), as.matrix(partadata[,1]), alpha = 1)
plot(cvlasso)
```




```
coef(cvlasso, s = "lambda.min")
```

```
## 14 x 1 sparse Matrix of class "dgCMatrix"
##                        1
```

```
## (Intercept)  5.72221561
## age          0.06943563
## weight       .
## height      -0.30255151
## neck        -0.35671017
## chest       -0.10438214
## abdomen      0.85422807
## hip         -0.11837181
## thigh        0.11665366
## knee         .
## ankle        0.12308108
## bicep        0.14768149
## forearm      0.22624259
## wrist       -1.78517222
```

*Comment:*

When log of lambda is approximately 2.5, all the coefficients are zero. The deviance is equivalent to $R^2$ in this regression. I found much of $R^2$ is explained in shrunk coefficients. At the end (after 0.7 of fraction deviance explained), coefficients relatively grow large. This implies we might overfit at the end.

Also, coefficient extractor picked the 4 non-zero coefficient (the second vertical line on the cross validation plot - one standard error of the smallest MSE). But, if I want to get the smallest MSE, we need to have to regularize/shrink to 11 non-zero coefficients (shrink weight and knee coefficients to be exactly zero). It is possible for us to have the coefficients exactly be zero, as this is a lasso model (L1 norm), but ridge model (L2 norm) is not possible to shrink to exactly zero. When $\lambda$ is zero, we will have the estimator of coefficients be OLS estimators.

Lasso estimators prevent us to have overfitting/multi-collinearity problems