ST300: Regression and Generalized Linear Models

Candidate Number: 17613

Summary

The aim of this coursework was to explain how the height of an individual is related to other body measurements of the individual.

There were originally 24 possible covariates. Variable selection was first carried out to choose variables that have a significant relationship with height out of the 24 possible covariates.

Regression diagnostic was then performed to check if the linear regression assumptions are satisfied. Outliers and influential points are also dealt with. No transformations were required as the relationship between height and the covariates appears to be linear.

The final model consists of 11 covariates which have significant relationship with height.

Introduction

The initial data set consisted of 21 body dimension measurements, together with age, weight, height and gender on 507 individuals. The 247 men and 260 women were primarily individuals in their twenties and thirties, with a scattering of older men and women, all exercising several hours a week.

Methods

First, stepwise selection with BIC criterion was carried out using the following commands:

```
> m1 = lm(height~., data=body)
> step(m1,direction='both', k=log(507))
```

The model selected using the procedure has the following 11 covariates: biacromial diameter, pelvic breadth, knee diameter, ankle diameter, chest girth, waist girth, thigh girth, forearm girth, calf girth, weight, and gender.

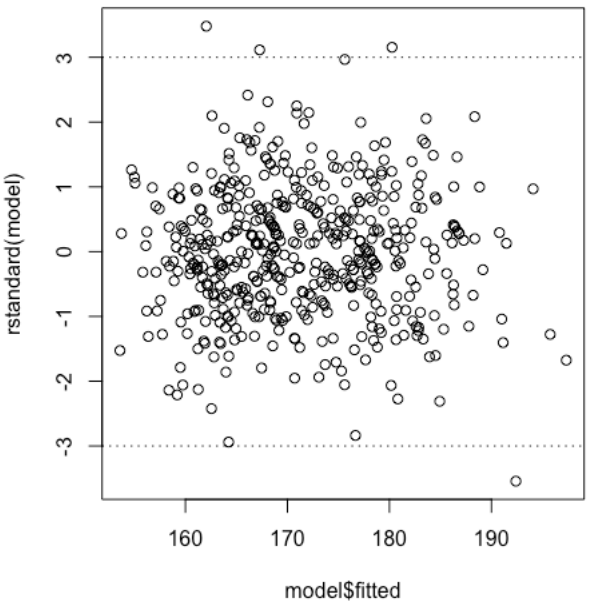
The summary of the model is as follow:

```
Call:
lm(formula = height ~ biacromial + pelvic.breadth + knee.diam +
    ankle.diam + chest.girth + waist.girth + thigh.girth + forearm.girth +
    calf.girth + weight + gender)

Residuals:
     Min       1Q   Median       3Q      Max
-14.5952  -2.8076   0.0673   2.7831  14.4521

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    197.10068    7.44898  26.460  < 2e-16 ***
biacromial       0.62369    0.11608   5.373 1.20e-07 ***
pelvic.breadth   0.34291    0.11081   3.095 0.002083 **
knee.diam       -0.79070    0.25429  -3.109 0.001982 **
ankle.diam       1.01628    0.26961   3.769 0.000183 ***
chest.girth     -0.28497    0.05883  -4.844 1.70e-06 ***
waist.girth     -0.60816    0.04844 -12.555  < 2e-16 ***
thigh.girth     -0.58886    0.07904  -7.451 4.17e-13 ***
forearm.girth   -0.84896    0.19436  -4.368 1.53e-05 ***
calf.girth      -0.48432    0.11943  -4.055 5.82e-05 ***
weight           1.22771    0.06392  19.207  < 2e-16 ***
gender           5.28764    0.93742   5.641 2.85e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.226 on 495 degrees of freedom
Multiple R-squared:  0.8026,    Adjusted R-squared:  0.7982
F-statistic: 182.9 on 11 and 495 DF,  p-value: < 2.2e-16
```

From the output, the F-statistic is significant at 1% level. Hence the model shows a significant relationship between height and the covariates. All of the variables are significant at a 1% level.

Regression diagnostic was then carried out. A standardized residual plot was plotted to check for outliers and high leverage points.

```
>plot(model$fitted,rstandard(model)) ; abline(3,0,lty=3) ;
  abline(-3,0,lty=3)
```



From the plot, there were 7 outliers with their absolute value of standardised residuals exceeding or very close to 3.

There were also data points with high fitted values indicating high leverage points.

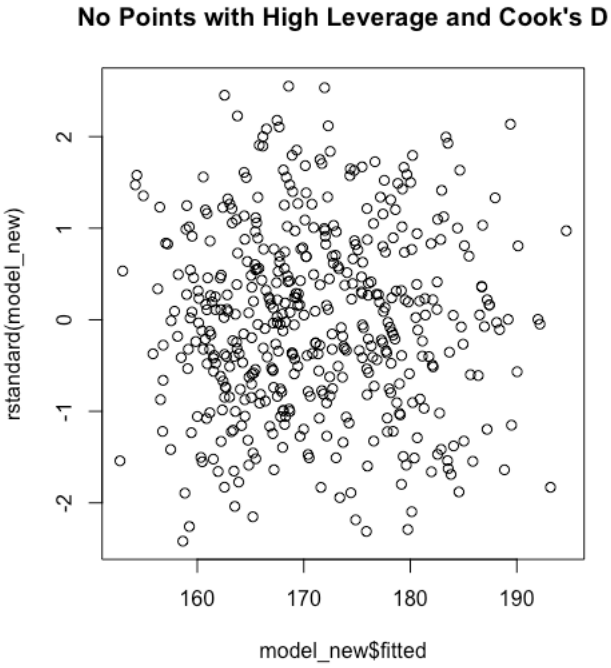
```
> sum(hatvalues(model)>2*12/507 | cooks.distance(model)>4/507)
[1] 45
```

There were 45 data points with high leverage or Cook's distance.

These data points were removed using the following lines:

```
#removing points with high leverage and Cook's Disance
index = 1:507
I = index[hatvalues(model)>2*12/507]
J = index[cooks.distance(model)>4/507]
K = sort(union(I,J))
length(K)
model_new = lm(height[-K] ~ biacromial[-K] + pelvic.breadth[-K] + knee.diam[-K] +
                 ankle.diam[-K] + chest.girth[-K] + waist.girth[-K] +
                 thigh.girth[-K] + forearm.girth[-K] + calf.girth[-K] +
                 weight[-K] + gender[-K])
```

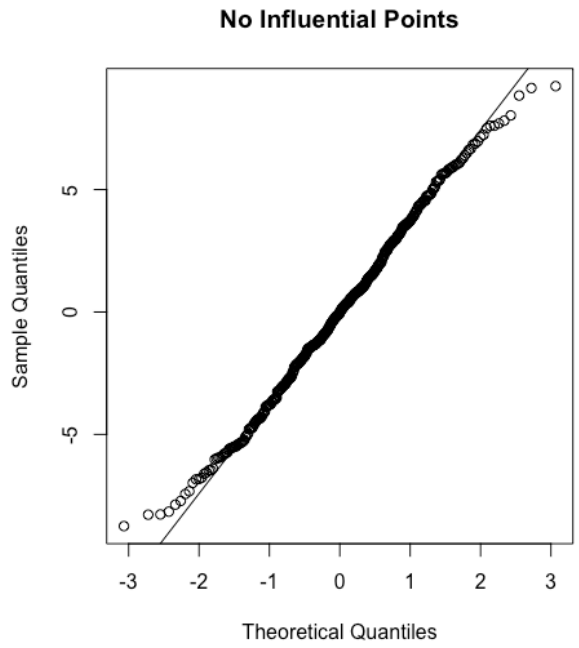
The standardized residual plot was then plotted:



From the plot, we can see that with the influential points deleted, the plot is fine, without apparent pattern, the cloud of points centered around 0, and with approximately constant variance throughout, with no outliers.

The model appeared to be linear hence no transformation of the variables was needed.

Normality of the model was then checked using Q-Q plot.



Most points lie on the straight line with some points at tails lying outside the straight line. Thus, the normality assumption was met.

```
> cbind(BIC(model),BIC(model_new))
          [,1]     [,2]
[1,] 2969.126 2572.385
> cbind(summary(model)$adj.r.squared, summary(model_new)$adj.r.squared)
           [,1]      [,2]
[1,] 0.7981675 0.8421628
```

After deleting the influential points, the model appeared to have a better fit. The BIC was reduced, and the adjusted R squared was increased.

Results and conclusion

These are the coefficients of the variables in the model:

```
> coef(model_new)
       (Intercept)     biacromial[-K] pelvic.breadth[-K]      knee.diam[-K]
       205.9880889          0.4814908          0.3256312         -0.7208400
    ankle.diam[-K]    chest.girth[-K]    waist.girth[-K]    thigh.girth[-K]
         0.7922470         -0.2627171         -0.6808899         -0.6741001
  forearm.girth[-K]     calf.girth[-K]         weight[-K]         gender[-K]
        -0.9639849         -0.4847556          1.3828323          4.6176846
```

The final model is as follow:

```
Height = (205.9880889 + 0.4814908*biacromial + 0.3256312*pelvic.breadth
        - 0.7208400*knee.diam + 0.7922470*ankle.diam - 0.2627171*chest.girth
        - 0.6808899*waist.girth - 0.6741001*thigh.girth - 0.9639849*forearm.girth
        - 0.4847556*calf.girth + 1.3828323*weight + 4.6176846*gender)
```

The interpretation is as follow:
On average, holding all the other covariates constant, a unit increase in each of the covariate will increase height by the coefficient.

For example, holding other variables unchanged, a unit increase (in kg) in weight will increase height by 1.3828323 cm on average.

The model also shows that holding other variables fixed, on average, a male is taller than a female by 4.6176846 cm.

Keep in mind that from the data collected, the height ranges between 147.2 and 198.1 cm. Hence, to avoid extrapolation, the model should have a good reflection of the relationship between height and the other variables of individuals of height between 147.2 and 198.1 cm.