# Problem Set 7

*Prateek Kumar*

## Contents

## Problem 1: Predicting a categorical variable

```
> library(plotrix) #loading the library
> library(faraway)
>
> kanga_final <- kanga[!is.na(rowSums(kanga[,3:20])),] #filtered data
>
> sex_data <- as.numeric(kanga_final$sex) #converting gender into numeric
>
> #colnames(kanga_final)
>
> var_comb <- kanga_final$basilar.length+kanga_final$occipitonasal.length+kanga_final$palate
.length+kanga_final$palate.width+kanga_final$nasal.length+kanga_final$nasal.width+kanga_fina
l$squamosal.depth+kanga_final$lacrymal.width+kanga_final$zygomatic.width+kanga_final$orbital
.width+kanga_final$.rostral.width+kanga_final$occipital.depth+kanga_final$crest.width+kanga_
final$foramina.length+kanga_final$mandible.length+kanga_final$mandible.width+kanga_final$man
dible.depth+kanga_final$ramus.height
>
> model_kanga_kanga <- lm(sex_data ~ var_comb, data=kanga_final)
>
> summary_kanga<-summary(model_kanga_kanga)
> summary_kanga

Call:
lm(formula = sex_data ~ var_comb, data = kanga_final)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.6765 -0.4553 -0.2409  0.4424  0.9184 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.209e-01  4.608e-01  -0.262 0.793536    
var_comb     1.531e-04  4.427e-05   3.459 0.000764 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4787 on 114 degrees of freedom
Multiple R-squared:  0.09496,	Adjusted R-squared:  0.08702 
F-statistic: 11.96 on 1 and 114 DF,  p-value: 0.0007645

>
>
> plot(model_kanga$fit~kanga_final$sex, xlab="Gender",ylab =" Gender coefficient ")
> points(sex_data,model_kanga$fit )
> abline(1.5 ,0 , lwd =3)
> predictedgender <- model_kanga $ fit > 1.5
> sex_tab <- table (sex_data ,c("Female","Male")[( predictedgender +1) ])
> sex_tab

sex_data Female Male
       1     46   11
       2     21   23
> val <- (sex_tab[1,1]+sex_tab[2,2]) / (sex_tab[1,1]+sex_tab[1,2]+sex_tab[2,1]+sex_tab[2,2])
*100
>
> print(paste("Accuracy =", round(val,2),"%"))
[1] "Accuracy = 68.32 %"
```

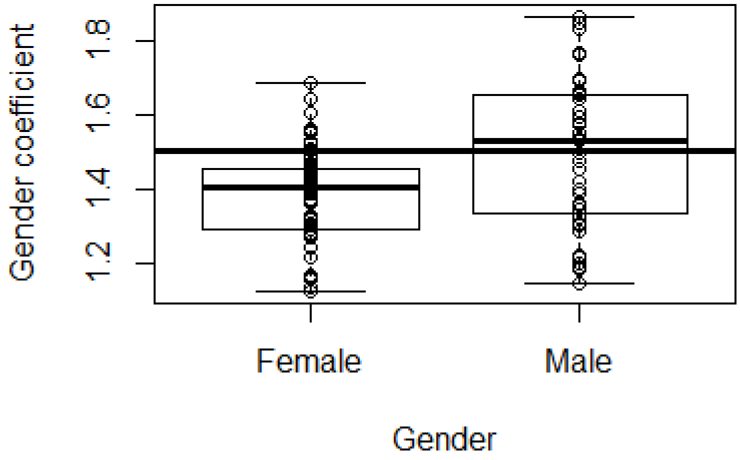


*Figure 1: Box plot of the sex column*

Here we got an accuracy of 68.32%.

When we have categorical predictors, there are several approaches we can take. There are better methods like logistic regression and discriminant analysis, but these are also versions of regression with certain transformations. But, if our categorical predictor is binary, then it fits into regression model easily.

If we have multiple levels, the regression model needs to do some underlying coding scheme to represent those levels as a combination of binary predictors. If we give lm a categorical predictor, it will code it for each of its levels with respect to the first level of a factor, so that the first level is equivalent to the intercept-only model.

The issues according to me when trying to predict categorical values are:

- We have to do additional work converting the categories to numeric
- Since the converted category is a factor we are not much accurate in predicting the value

As we see above we are just 68.32% accurate in predicting the values if we use linear regression in predicting categorical variables.

## Problem 2: Selecting variables

```
> library(faraway)
>
> kanga_final <- kanga[!is.na(rowSums(kanga[,3:20])),]
> sex_data <- as.numeric(kanga_final$sex)
>
> lm1 <- lm(sex_data~kanga_final$basilar.length+kanga_final$occipitonasal.length+kanga_final
$palate.length+kanga_final$palate.width+kanga_final$nasal.length+kanga_final$nasal.width+kan
ga_final$squamosal.depth+kanga_final$lacrymal.width+kanga_final$zygomatic.width+kanga_final$
orbital.width+kanga_final$.rostral.width+kanga_final$occipital.depth+kanga_final$crest.width
+kanga_final$foramina.length+kanga_final$mandible.length+kanga_final$mandible.width+kanga_fi
nal$mandible.depth+kanga_final$ramus.height,data=kanga_final)
> summary(lm1)

Call:
lm(formula = sex_data ~ kanga_final$basilar.length + kanga_final$occipitonasal.length +
    kanga_final$palate.length + kanga_final$palate.width + kanga_final$nasal.length +
    kanga_final$nasal.width + kanga_final$squamosal.depth + kanga_final$lacrymal.width +
    kanga_final$zygomatic.width + kanga_final$orbital.width +
    kanga_final$.rostral.width + kanga_final$occipital.depth +
    kanga_final$crest.width + kanga_final$foramina.length + kanga_final$mandible.length +
    kanga_final$mandible.width + kanga_final$mandible.depth +
    kanga_final$ramus.height, data = kanga_final)

Residuals:
     Min       1Q   Median       3Q      Max
-0.75517 -0.32848 -0.07782  0.34893  0.95973

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                      -9.513e-01  9.556e-01  -0.996  0.32196
kanga_final$basilar.length        8.335e-04  2.512e-03   0.332  0.74079
kanga_final$occipitonasal.length  6.255e-03  2.061e-03   3.035  0.00309 **
kanga_final$palate.length         1.318e-03  2.566e-03   0.514  0.60867
kanga_final$palate.width          1.049e-04  1.720e-03   0.061  0.95152
kanga_final$nasal.length         -8.379e-03  1.988e-03  -4.214 5.63e-05 ***
kanga_final$nasal.width           7.677e-03  3.922e-03   1.957  0.05318 .
kanga_final$squamosal.depth       1.406e-03  2.812e-03   0.500  0.61828
kanga_final$lacrymal.width       -5.642e-03  3.946e-03  -1.430  0.15600
kanga_final$zygomatic.width       1.194e-03  2.546e-03   0.469  0.64005
kanga_final$orbital.width         4.580e-03  3.346e-03   1.369  0.17423
kanga_final$.rostral.width       -9.493e-05  3.002e-03  -0.032  0.97483
kanga_final$occipital.depth      -1.146e-03  2.405e-03  -0.477  0.63470
kanga_final$crest.width          -3.785e-03  1.956e-03  -1.935  0.05593 .
kanga_final$foramina.length      -2.360e-03  3.317e-03  -0.711  0.47854
kanga_final$mandible.length      -2.494e-03  2.244e-03  -1.111  0.26911
kanga_final$mandible.width        7.099e-03  6.909e-03   1.028  0.30670
kanga_final$mandible.depth       -1.346e-03  4.564e-03  -0.295  0.76877
kanga_final$ramus.height         -3.165e-03  2.595e-03  -1.219  0.22565
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4255 on 97 degrees of freedom
Multiple R-squared:  0.3914,   Adjusted R-squared:  0.2785
F-statistic: 3.466 on 18 and 97 DF,  p-value: 3.849e-05

> val1 <- drop1(lm1,test="F")
> val1
Single term deletions
```

```
Model:
sex_data ~ kanga_final$basilar.length + kanga_final$occipitonasal.length + 
    kanga_final$palate.length + kanga_final$palate.width + kanga_final$nasal.length + 
    kanga_final$nasal.width + kanga_final$squamosal.depth + kanga_final$lacrymal.width + 
    kanga_final$zygomatic.width + kanga_final$orbital.width + 
    kanga_final$.rostral.width + kanga_final$occipital.depth + 
    kanga_final$crest.width + kanga_final$foramina.length + kanga_final$mandible.length + 
    kanga_final$mandible.width + kanga_final$mandible.depth + 
    kanga_final$ramus.height
                                 Df Sum of Sq    RSS     AIC F value    Pr(>F)    
<none>                                        17.565 -180.97                      
kanga_final$basilar.length        1    0.0199 17.585 -182.84  0.1101  0.740786    
kanga_final$occipitonasal.length  1    1.6676 19.232 -172.45  9.2094  0.003091 ** 
kanga_final$palate.length         1    0.0478 17.612 -182.66  0.2638  0.608672    
kanga_final$palate.width          1    0.0007 17.565 -182.97  0.0037  0.951522    
kanga_final$nasal.length          1    3.2156 20.780 -163.47 17.7578 5.627e-05 ***
kanga_final$nasal.width           1    0.6937 18.258 -178.48  3.8312  0.053181 .  
kanga_final$squamosal.depth       1    0.0453 17.610 -182.68  0.2499  0.618282    
kanga_final$lacrymal.width        1    0.3702 17.935 -180.55  2.0442  0.155998    
kanga_final$zygomatic.width       1    0.0398 17.605 -182.71  0.2201  0.640047    
kanga_final$orbital.width         1    0.3393 17.904 -180.75  1.8735  0.174234    
kanga_final$.rostral.width        1    0.0002 17.565 -182.97  0.0010  0.974835    
kanga_final$occipital.depth       1    0.0411 17.606 -182.70  0.2272  0.634699    
kanga_final$crest.width           1    0.6779 18.242 -178.58  3.7435  0.055925 .  
kanga_final$foramina.length       1    0.0916 17.656 -182.37  0.5061  0.478540    
kanga_final$mandible.length       1    0.2237 17.788 -181.51  1.2354  0.269111    
kanga_final$mandible.width        1    0.1912 17.756 -181.72  1.0559  0.306697    
kanga_final$mandible.depth        1    0.0157 17.580 -182.87  0.0869  0.768766    
kanga_final$ramus.height          1    0.2693 17.834 -181.21  1.4869  0.225649    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> max(val1$`Pr(>F)`, na.rm = T)
[1] 0.9748348
```

We here get the max p-value as 0.9748 which is of squamosal.depth so we will now calculate the regression without that column

```
> lm2 <- lm(sex_data~kanga_final$basilar.length+kanga_final$occipitonasal.length+kanga_final
$palate.length+kanga_final$palate.width+kanga_final$nasal.length+kanga_final$nasal.width+kan
ga_final$lacrymal.width+kanga_final$zygomatic.width+kanga_final$orbital.width+kanga_final$.r
ostral.width+kanga_final$occipital.depth+kanga_final$crest.width+kanga_final$foramina.length
+kanga_final$mandible.length+kanga_final$mandible.width+kanga_final$mandible.depth+kanga_fin
al$ramus.height,data=kanga_final)
> summary(lm2)

Call:
lm(formula = sex_data ~ kanga_final$basilar.length + kanga_final$occipitonasal.length + 
    kanga_final$palate.length + kanga_final$palate.width + kanga_final$nasal.length + 
    kanga_final$nasal.width + kanga_final$lacrymal.width + kanga_final$zygomatic.width + 
    kanga_final$orbital.width + kanga_final$.rostral.width + 
    kanga_final$occipital.depth + kanga_final$crest.width + kanga_final$foramina.length + 
    kanga_final$mandible.length + kanga_final$mandible.width + 
    kanga_final$mandible.depth + kanga_final$ramus.height, data = kanga_final)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.74498 -0.33636 -0.06895  0.33845  0.97017 

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)                       -1.035e+00  9.373e-01  -1.104  0.27228
kanga_final$basilar.length         9.864e-04  2.484e-03   0.397  0.69217
kanga_final$occipitonasal.length  6.202e-03  2.050e-03   3.025  0.00318 **
kanga_final$palate.length          1.223e-03  2.550e-03   0.480  0.63244
kanga_final$palate.width           6.999e-05  1.712e-03   0.041  0.96748
kanga_final$nasal.length          -8.401e-03  1.980e-03  -4.242 5.02e-05 ***
kanga_final$nasal.width            8.053e-03  3.835e-03   2.100  0.03829 *
kanga_final$lacrymal.width        -5.715e-03  3.928e-03  -1.455  0.14889
kanga_final$zygomatic.width        1.486e-03  2.469e-03   0.602  0.54850
kanga_final$orbital.width          4.407e-03  3.316e-03   1.329  0.18686
kanga_final$.rostral.width        -7.140e-06  2.985e-03  -0.002  0.99810
kanga_final$occipital.depth       -1.089e-03  2.393e-03  -0.455  0.65004
kanga_final$crest.width           -3.741e-03  1.947e-03  -1.922  0.05755 .
kanga_final$foramina.length       -2.121e-03  3.270e-03  -0.649  0.51809
kanga_final$mandible.length       -2.569e-03  2.230e-03  -1.152  0.25216
kanga_final$mandible.width         6.482e-03  6.771e-03   0.957  0.34078
kanga_final$mandible.depth        -1.230e-03  4.541e-03  -0.271  0.78702
kanga_final$ramus.height          -3.009e-03  2.567e-03  -1.172  0.24387
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4239 on 98 degrees of freedom
Multiple R-squared:  0.3899,   Adjusted R-squared:  0.284
F-statistic: 3.683 on 17 and 98 DF,  p-value: 2.119e-05

> val1 <- drop1(lm2,test="F")
> val1
Single term deletions

Model:
sex_data ~ kanga_final$basilar.length + kanga_final$occipitonasal.length +
    kanga_final$palate.length + kanga_final$palate.width + kanga_final$nasal.length +
    kanga_final$nasal.width + kanga_final$lacrymal.width + kanga_final$zygomatic.width +
    kanga_final$orbital.width + kanga_final$.rostral.width +
    kanga_final$occipital.depth + kanga_final$crest.width + kanga_final$foramina.length +
    kanga_final$mandible.length + kanga_final$mandible.width +
    kanga_final$mandible.depth + kanga_final$ramus.height
                                 Df Sum of Sq    RSS     AIC F value    Pr(>F)
<none>                                        17.610 -182.68
kanga_final$basilar.length        1    0.0283 17.638 -184.49  0.1577  0.692172
kanga_final$occipitonasal.length  1    1.6441 19.254 -174.32  9.1497  0.003177 **
kanga_final$palate.length         1    0.0414 17.651 -184.40  0.2302  0.632438
kanga_final$palate.width          1    0.0003 17.610 -184.67  0.0017  0.967479
kanga_final$nasal.length          1    3.2342 20.844 -165.12 17.9985 5.019e-05 ***
kanga_final$nasal.width           1    0.7925 18.402 -179.57  4.4104  0.038289 *
kanga_final$lacrymal.width        1    0.3804 17.990 -182.20  2.1168  0.148890
kanga_final$zygomatic.width       1    0.0651 17.675 -184.25  0.3625  0.548499
kanga_final$orbital.width         1    0.3175 17.927 -182.60  1.7668  0.186865
kanga_final$.rostral.width        1    0.0000 17.610 -184.68  0.0000  0.998096
kanga_final$occipital.depth       1    0.0372 17.647 -184.43  0.2071  0.650043
kanga_final$crest.width           1    0.6636 18.273 -180.38  3.6929  0.057548 .
kanga_final$foramina.length       1    0.0756 17.686 -184.18  0.4207  0.518093
kanga_final$mandible.length       1    0.2384 17.848 -183.12  1.3269  0.252163
kanga_final$mandible.width        1    0.1647 17.774 -183.59  0.9164  0.340784
kanga_final$mandible.depth        1    0.0132 17.623 -184.59  0.0734  0.787022
kanga_final$ramus.height          1    0.2470 17.857 -183.06  1.3746  0.243872
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> max(val1$`Pr(>F)`, na.rm = T)
[1] 0.9980963
>
```

```
> anova(lm1,lm2)
Analysis of Variance Table

Model 1: sex_data ~ kanga_final$basilar.length + kanga_final$occipitonasal.length + 
    kanga_final$palate.length + kanga_final$palate.width + kanga_final$nasal.length + 
    kanga_final$nasal.width + kanga_final$squamosal.depth + kanga_final$lacrymal.width + 
    kanga_final$zygomatic.width + kanga_final$orbital.width + 
    kanga_final$.rostral.width + kanga_final$occipital.depth + 
    kanga_final$crest.width + kanga_final$foramina.length + kanga_final$mandible.length + 
    kanga_final$mandible.width + kanga_final$mandible.depth + 
    kanga_final$ramus.height
Model 2: sex_data ~ kanga_final$basilar.length + kanga_final$occipitonasal.length + 
    kanga_final$palate.length + kanga_final$palate.width + kanga_final$nasal.length + 
    kanga_final$nasal.width + kanga_final$lacrymal.width + kanga_final$zygomatic.width + 
    kanga_final$orbital.width + kanga_final$.rostral.width + 
    kanga_final$occipital.depth + kanga_final$crest.width + kanga_final$foramina.length + 
    kanga_final$mandible.length + kanga_final$mandible.width + 
    kanga_final$mandible.depth + kanga_final$ramus.height
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     97 17.565                           
2     98 17.610 -1  -0.04525 0.2499 0.6183
```

We now have the max p-value as 0.998 which is of palate.width so we will now remove that column. Previously as we removed the squamosal.depth value we can see the anova result of the two models and we cannot see a much of a difference so we can go with the simpler model.

Eventually we will proceed with the above steps until the anova values difference suffices.

```
> lm11 <- lm(sex_data~kanga_final$occipitonasal.length+kanga_final$nasal.length+kanga_final$
nasal.width+kanga_final$lacrymal.width+kanga_final$orbital.width+kanga_final$crest.width+kan
ga_final$mandible.length+kanga_final$mandible.width,data=kanga_final)
> summary(lm11)

Call:
lm(formula = sex_data ~ kanga_final$occipitonasal.length + kanga_final$nasal.length + 
    kanga_final$nasal.width + kanga_final$lacrymal.width + kanga_final$orbital.width + 
    kanga_final$crest.width + kanga_final$mandible.length + kanga_final$mandible.width, 
    data = kanga_final)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.75456 -0.35555 -0.03254  0.33614  0.99543 

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)                      -1.065243   0.798200  -1.335  0.18485    
kanga_final$occipitonasal.length  0.006907   0.001578   4.379 2.79e-05 ***
kanga_final$nasal.length         -0.008612   0.001719  -5.009 2.17e-06 ***
kanga_final$nasal.width           0.009225   0.003324   2.775  0.00651 ** 
kanga_final$lacrymal.width       -0.007401   0.003295  -2.246  0.02673 *  
kanga_final$orbital.width         0.004362   0.003113   1.401  0.16399    
kanga_final$crest.width          -0.003648   0.001683  -2.167  0.03244 *  
kanga_final$mandible.length      -0.002199   0.001222  -1.799  0.07487 .  
kanga_final$mandible.width        0.005487   0.005829   0.941  0.34871    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4155 on 107 degrees of freedom
Multiple R-squared:  0.3599,	Adjusted R-squared:  0.312 
F-statistic:  7.52 on 8 and 107 DF,  p-value: 6.319e-08
```

```
> val1 <- drop1(lm11,test="F")
> val1
Single term deletions

Model:
sex_data ~ kanga_final$occipitonasal.length + kanga_final$nasal.length + 
    kanga_final$nasal.width + kanga_final$lacrymal.width + kanga_final$orbital.width + 
    kanga_final$crest.width + kanga_final$mandible.length + kanga_final$mandible.width
                                 Df Sum of Sq    RSS     AIC F value    Pr(>F)    
<none>                                        18.475 -195.11                      
kanga_final$occipitonasal.length  1    3.3103 21.785 -177.99 19.1722 2.792e-05 ***
kanga_final$nasal.length          1    4.3317 22.806 -172.68 25.0877 2.174e-06 ***
kanga_final$nasal.width           1    1.3299 19.805 -189.05  7.7024   0.00651 ** 
kanga_final$lacrymal.width        1    0.8713 19.346 -191.77  5.0465   0.02673 *  
kanga_final$orbital.width         1    0.3391 18.814 -195.00  1.9639   0.16399    
kanga_final$crest.width           1    0.8109 19.286 -192.13  4.6967   0.03244 *  
kanga_final$mandible.length       1    0.5587 19.033 -193.66  3.2357   0.07487 .  
kanga_final$mandible.width        1    0.1530 18.628 -196.16  0.8859   0.34871    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> max(val1$`Pr(>F)`, na.rm = T)
[1] 0.3487097
> 
> anova(lm10,lm11)
Analysis of Variance Table

Model 1: sex_data ~ kanga_final$occipitonasal.length + kanga_final$palate.length + 
    kanga_final$nasal.length + kanga_final$nasal.width + kanga_final$lacrymal.width + 
    kanga_final$orbital.width + kanga_final$crest.width + kanga_final$mandible.length + 
    kanga_final$mandible.width
Model 2: sex_data ~ kanga_final$occipitonasal.length + kanga_final$nasal.length + 
    kanga_final$nasal.width + kanga_final$lacrymal.width + kanga_final$orbital.width + 
    kanga_final$crest.width + kanga_final$mandible.length + kanga_final$mandible.width
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    106 18.162                           
2    107 18.475 -1   -0.3125 1.8238 0.1797
> anova(lm1,lm11)
Analysis of Variance Table

Model 1: sex_data ~ kanga_final$basilar.length + kanga_final$occipitonasal.length + 
    kanga_final$palate.length + kanga_final$palate.width + kanga_final$nasal.length + 
    kanga_final$nasal.width + kanga_final$squamosal.depth + kanga_final$lacrymal.width + 
    kanga_final$zygomatic.width + kanga_final$orbital.width + 
    kanga_final$.rostral.width + kanga_final$occipital.depth + 
    kanga_final$crest.width + kanga_final$foramina.length + kanga_final$mandible.length + 
    kanga_final$mandible.width + kanga_final$mandible.depth + 
    kanga_final$ramus.height
Model 2: sex_data ~ kanga_final$occipitonasal.length + kanga_final$nasal.length + 
    kanga_final$nasal.width + kanga_final$lacrymal.width + kanga_final$orbital.width + 
    kanga_final$crest.width + kanga_final$mandible.length + kanga_final$mandible.width
  Res.Df    RSS  Df Sum of Sq      F Pr(>F)
1     97 17.565                            
2    107 18.475 -10  -0.91013 0.5026 0.8844
```

From the drop1 function above we see that any other smaller model will fit worse so we would prefer model lm11. Also on checking the $R^2$ values we see that the values are pretty close to each other.

```
> data.frame(model=paste("lm",1:16,sep=""),
+            rbind(extractAIC(lm1),
+                  extractAIC(lm2),
+                  extractAIC(lm3),
+                  extractAIC(lm4),
+                  extractAIC(lm5),
+                  extractAIC(lm6),
+                  extractAIC(lm7),
+                  extractAIC(lm8),
+                  extractAIC(lm9),
+                  extractAIC(lm10),
+                  extractAIC(lm11),
+                  extractAIC(lm12),
+                  extractAIC(lm13),
+                  extractAIC(lm14),
+                  extractAIC(lm15),
+                  extractAIC(lm16)))
   model X1        X2
1    lm1 19 -180.9736
2    lm2 18 -182.6751
3    lm3 17 -184.6731
4    lm4 16 -186.4875
5    lm5 15 -188.4354
6    lm6 14 -190.4252
7    lm7 13 -192.0918
8    lm8 12 -193.4879
9    lm9 11 -194.8720
10  lm10 10 -195.0924
11  lm11  9 -195.1135
12  lm12  8 -193.6576
13  lm13  7 -195.6379
14  lm14  6 -193.0703
15  lm15  5 -193.0494
16  lm16  4 -182.8879
```

Checking the above AIC values, we can see that the values gets more negative till model lm11 and then increases so we are correct from the above result.

```
> library(BayesFactor)
>
> kanga_final$sex <- as.numeric(kanga_final$sex)
>
> bmodel <- regressionBF(sex~occipitonasal.length+palate.length+nasal.length+nasal.width+lac
rymal.width+orbital.width+crest.width+mandible.length+mandible.width,data=kanga_final)
  |=========================================================================================
==========================| 100%
>
> plot(head(bmodel))
> head(bmodel)
Bayes factor analysis
--------------
[1] occipitonasal.length + palate.length + nasal.length + nasal.width + lacrymal.width + man
dible.length                 : 732790.6 ±0%
[2] occipitonasal.length + nasal.length + nasal.width + lacrymal.width + mandible.length
: 565770.7 ±0%
[3] occipitonasal.length + nasal.length + nasal.width + lacrymal.width + crest.width + mandi
ble.length                   : 546628   ±0%
```

```
[4] occipitonasal.length + palate.length + nasal.length + nasal.width + lacrymal.width + cre
st.width + mandible.length : 500787.9 ±0%
[5] occipitonasal.length + nasal.length + nasal.width + lacrymal.width + orbital.width + cre
st.width                   : 467917   ±0%
[6] occipitonasal.length + nasal.length + nasal.width + lacrymal.width + orbital.width + cre
st.width + mandible.length : 405456.5 ±0%

Against denominator:
  Intercept only
---
Bayes factor type: BFlinearModel, JZS
```

We now run the bayesfactor regression above on the simplest model.

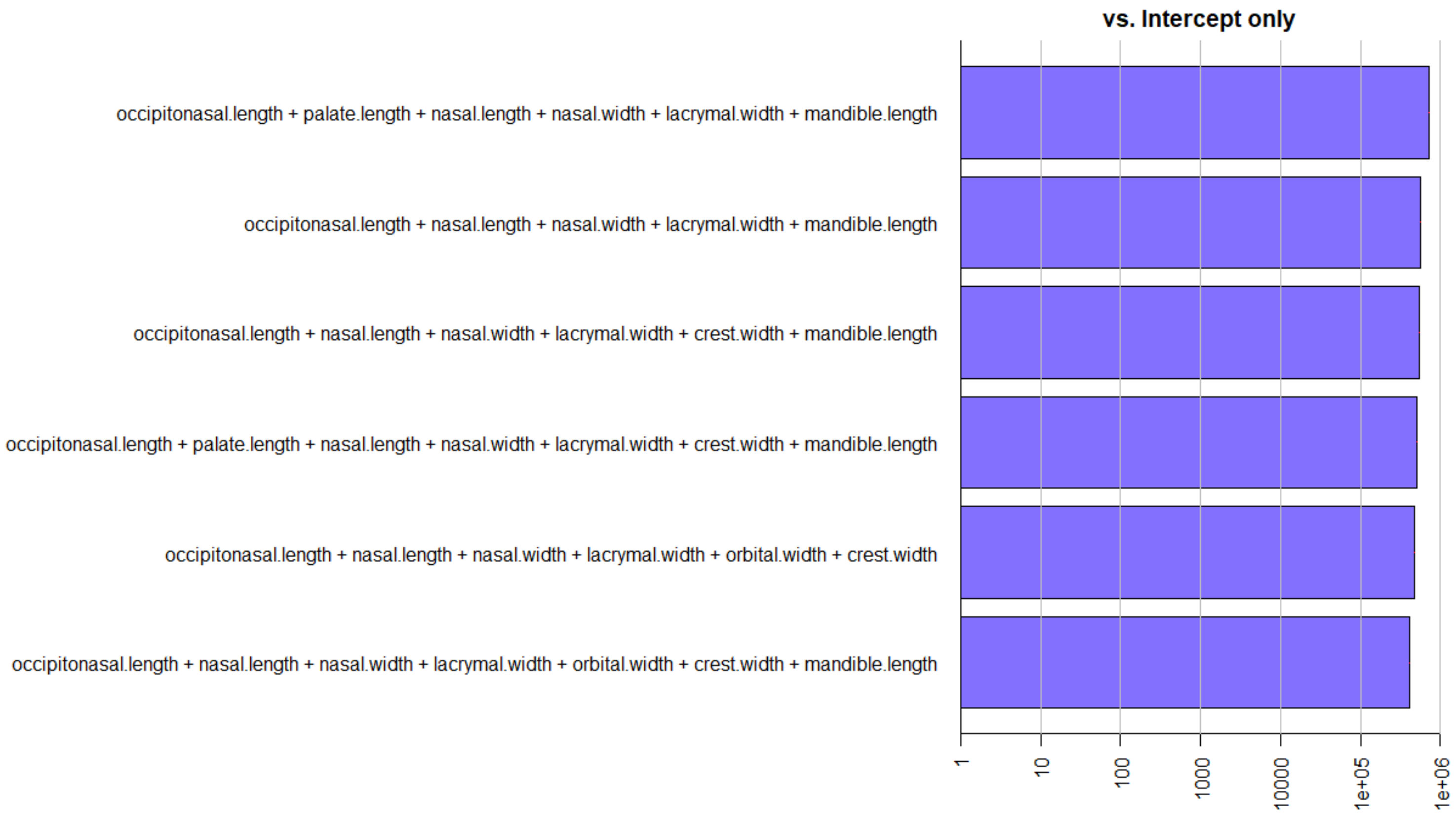


*Figure 2: Bayesfactor regression plot*

```
> gsmall <- step(lm1,direction="both", k=log(nrow(dat)))
```

```
Step:  AIC=-181.52
sex_data ~ kanga_final$occipitonasal.length + kanga_final$nasal.length + 
    kanga_final$ramus.height

                                   Df Sum of Sq    RSS     AIC
<none>                                          19.968 -181.52
+ kanga_final$crest.width           1    0.4029 19.565 -178.24
+ kanga_final$nasal.width           1    0.3462 19.622 -177.90
+ kanga_final$mandible.width        1    0.2182 19.750 -177.15
+ kanga_final$palate.length         1    0.0865 19.881 -176.38
+ kanga_final$orbital.width         1    0.0852 19.883 -176.37
+ kanga_final$basilar.length        1    0.0546 19.913 -176.19
+ kanga_final$occipital.depth       1    0.0490 19.919 -176.16
+ kanga_final$foramina.length       1    0.0466 19.921 -176.14
+ kanga_final$mandible.length       1    0.0316 19.936 -176.06
+ kanga_final$squamosal.depth       1    0.0306 19.937 -176.05
+ kanga_final$palate.width          1    0.0270 19.941 -176.03
+ kanga_final$lacrymal.width        1    0.0240 19.944 -176.01
+ kanga_final$.rostral.width        1    0.0075 19.960 -175.91
+ kanga_final$zygomatic.width       1    0.0032 19.965 -175.89
+ kanga_final$mandible.depth        1    0.0001 19.968 -175.87
- kanga_final$ramus.height          1    3.1815 23.149 -170.01
- kanga_final$nasal.length          1    5.2036 25.171 -160.30
- kanga_final$occipitonasal.length  1    7.0312 26.999 -152.17
```

We get the above output from the step function. We see that only 3 columns suffice.

```
> summary(gsmall)

Call:
lm(formula = sex_data ~ kanga_final$occipitonasal.length + kanga_final$nasal.length + 
    kanga_final$ramus.height, data = kanga_final)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.8014 -0.3422 -0.1057  0.3716  1.1174 

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)                      -1.741081   0.488745  -3.562 0.000541 ***
kanga_final$occipitonasal.length  0.007990   0.001272   6.280 6.63e-09 ***
kanga_final$nasal.length         -0.008596   0.001591  -5.403 3.74e-07 ***
kanga_final$ramus.height         -0.005061   0.001198  -4.224 4.90e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4222 on 112 degrees of freedom
Multiple R-squared:  0.3082,	Adjusted R-squared:  0.2896 
F-statistic: 16.63 on 3 and 112 DF,  p-value: 5.275e-09
```

## Problem 3: Predicting missing data

```
> library(faraway)
>
> kanga <- faraway::kanga
> old1 <- kanga$palate.width
>
> lm1 <- lm(kanga$palate.width~kanga$basilar.length+kanga$occipitonasal.length+kanga$palate.
length+kanga$nasal.length+kanga$nasal.width+kanga$squamosal.depth+kanga$lacrymal.width+kanga
$zygomatic.width+kanga$orbital.width+kanga$.rostral.width+kanga$occipital.depth+kanga$crest.
width+kanga$foramina.length+kanga$mandible.length+kanga$mandible.width+kanga$mandible.depth+
kanga$ramus.height,data=kanga)
> summary(lm1)

Call:
lm(formula = kanga$palate.width ~ kanga$basilar.length + kanga$occipitonasal.length +
    kanga$palate.length + kanga$nasal.length + kanga$nasal.width +
    kanga$squamosal.depth + kanga$lacrymal.width + kanga$zygomatic.width +
    kanga$orbital.width + kanga$.rostral.width + kanga$occipital.depth +
    kanga$crest.width + kanga$foramina.length + kanga$mandible.length +
    kanga$mandible.width + kanga$mandible.depth + kanga$ramus.height,
    data = kanga)

Residuals:
    Min      1Q  Median      3Q     Max
-31.106  -9.908  -0.329  10.341  41.875

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                21.23209   37.36968   0.568  0.57146
kanga$basilar.length       -0.20908    0.09784  -2.137  0.03555 *
kanga$occipitonasal.length  0.01861    0.08087   0.230  0.81855
kanga$palate.length         0.08261    0.10150   0.814  0.41804
kanga$nasal.length         -0.06187    0.07823  -0.791  0.43126
kanga$nasal.width           0.43610    0.16204   2.691  0.00861 **
kanga$squamosal.depth       0.03677    0.15653   0.235  0.81488
kanga$lacrymal.width       -0.17336    0.15435  -1.123  0.26461
kanga$zygomatic.width       0.27819    0.09960   2.793  0.00648 **
kanga$orbital.width        -0.02816    0.12986  -0.217  0.82885
kanga$.rostral.width        0.02716    0.11667   0.233  0.81647
kanga$occipital.depth       0.08008    0.09979   0.802  0.42460
kanga$crest.width          -0.15513    0.07712  -2.012  0.04750 *
kanga$foramina.length       0.03151    0.12858   0.245  0.80704
kanga$mandible.length       0.18689    0.09483   1.971  0.05209 .
kanga$mandible.width       -0.10714    0.29959  -0.358  0.72153
kanga$mandible.depth       -0.19189    0.18726  -1.025  0.30848
kanga$ramus.height         -0.03194    0.10046  -0.318  0.75133
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.84 on 83 degrees of freedom
  (47 observations deleted due to missingness)
Multiple R-squared:  0.7763,   Adjusted R-squared:  0.7305
F-statistic: 16.94 on 17 and 83 DF,  p-value: < 2.2e-16

>
> #orbital.width|occipitonasal.length|foramina.length
>
> gsmall <- step(lm1,direction="both", k=log(nrow(dat)))
Start:  AIC=639.79
kanga$palate.width ~ kanga$basilar.length + kanga$occipitonasal.length +
    kanga$palate.length + kanga$nasal.length + kanga$nasal.width +
    kanga$squamosal.depth + kanga$lacrymal.width + kanga$zygomatic.width +
    kanga$orbital.width + kanga$.rostral.width + kanga$occipital.depth +
    kanga$crest.width + kanga$foramina.length + kanga$mandible.length +
    kanga$mandible.width + kanga$mandible.depth + kanga$ramus.height

                             Df Sum of Sq   RSS    AIC
- kanga$orbital.width         1     11.79 20829 634.20
- kanga$occipitonasal.length  1     13.28 20831 634.21
- kanga$.rostral.width        1     13.60 20831 634.21
- kanga$squamosal.depth       1     13.84 20831 634.21
- kanga$foramina.length       1     15.06 20833 634.22
- kanga$ramus.height          1     25.35 20843 634.27
- kanga$mandible.width        1     32.08 20850 634.30
- kanga$nasal.length          1    156.89 20974 634.90
- kanga$occipital.depth       1    161.49 20979 634.92
- kanga$palate.length         1    166.14 20984 634.95
- kanga$mandible.depth        1    263.36 21081 635.41
- kanga$lacrymal.width        1    316.39 21134 635.67
- kanga$mandible.length       1    974.11 21792 638.76
- kanga$crest.width           1   1014.98 21832 638.95
- kanga$basilar.length        1   1145.34 21963 639.55
<none>                                    20817 639.79
- kanga$nasal.width           1   1816.55 22634 642.59
- kanga$zygomatic.width       1   1956.76 22774 643.22

Step:  AIC=634.2
kanga$palate.width ~ kanga$basilar.length + kanga$occipitonasal.length +
    kanga$palate.length + kanga$nasal.length + kanga$nasal.width +
    kanga$squamosal.depth + kanga$lacrymal.width + kanga$zygomatic.width +
    kanga$.rostral.width + kanga$occipital.depth + kanga$crest.width +
    kanga$foramina.length + kanga$mandible.length + kanga$mandible.width +
    kanga$mandible.depth + kanga$ramus.height

                             Df Sum of Sq   RSS    AIC
- kanga$occipitonasal.length  1      8.81 20838 628.60
- kanga$foramina.length       1     11.11 20840 628.61
- kanga$.rostral.width        1     16.58 20846 628.64
- kanga$squamosal.depth       1     19.85 20849 628.65
- kanga$ramus.height          1     24.87 20854 628.68
- kanga$mandible.width        1     29.06 20858 628.70
- kanga$nasal.length          1    147.08 20976 629.27
- kanga$occipital.depth       1    160.47 20990 629.33
- kanga$palate.length         1    171.56 21001 629.38
- kanga$mandible.depth        1    280.89 21110 629.91
- kanga$lacrymal.width        1    361.68 21191 630.29
- kanga$mandible.length       1   1036.13 21865 633.46
- kanga$basilar.length        1   1193.21 22022 634.18
<none>                                    20829 634.20
- kanga$crest.width           1   1205.91 22035 634.24
- kanga$nasal.width           1   1811.45 22641 636.98
- kanga$zygomatic.width       1   1956.70 22786 637.62
+ kanga$orbital.width         1     11.79 20817 639.79

Step:  AIC=628.6
kanga$palate.width ~ kanga$basilar.length + kanga$palate.length +
    kanga$nasal.length + kanga$nasal.width + kanga$squamosal.depth +
    kanga$lacrymal.width + kanga$zygomatic.width + kanga$.rostral.width +
    kanga$occipital.depth + kanga$crest.width + kanga$foramina.length +
    kanga$mandible.length + kanga$mandible.width + kanga$mandible.depth +
    kanga$ramus.height
```

```
                             Df Sum of Sq   RSS    AIC
- kanga$foramina.length       1      9.59 20848 623.00
- kanga$.rostral.width        1     12.76 20851 623.01
- kanga$squamosal.depth       1     17.90 20856 623.04
- kanga$ramus.height          1     25.76 20864 623.08
- kanga$mandible.width        1     32.10 20870 623.11
- kanga$palate.length         1    165.36 21003 623.75
- kanga$occipital.depth       1    183.35 21021 623.84
- kanga$nasal.length          1    185.01 21023 623.85
- kanga$mandible.depth        1    293.98 21132 624.37
- kanga$lacrymal.width        1    353.13 21191 624.65
- kanga$mandible.length       1   1027.46 21866 627.81
<none>                                    20838 628.60
- kanga$crest.width           1   1235.15 22073 628.77
- kanga$basilar.length        1   1619.58 22458 630.51
- kanga$nasal.width           1   1846.75 22685 631.53
- kanga$zygomatic.width       1   1980.95 22819 632.12
+ kanga$occipitonasal.length  1      8.81 20829 634.20
+ kanga$orbital.width         1      7.31 20831 634.21

Step:  AIC=623
kanga$palate.width ~ kanga$basilar.length + kanga$palate.length +
    kanga$nasal.length + kanga$nasal.width + kanga$squamosal.depth +
    kanga$lacrymal.width + kanga$zygomatic.width + kanga$.rostral.width +
    kanga$occipital.depth + kanga$crest.width + kanga$mandible.length +
    kanga$mandible.width + kanga$mandible.depth + kanga$ramus.height

                             Df Sum of Sq   RSS    AIC
- kanga$.rostral.width        1     12.76 20860 617.42
- kanga$squamosal.depth       1     20.64 20868 617.45
- kanga$ramus.height          1     25.44 20873 617.48
- kanga$mandible.width        1     33.20 20881 617.51
- kanga$nasal.length          1    179.92 21028 618.22
- kanga$palate.length         1    182.19 21030 618.23
- kanga$occipital.depth       1    184.72 21032 618.24
- kanga$mandible.depth        1    307.57 21155 618.83
- kanga$lacrymal.width        1    343.91 21192 619.01
- kanga$mandible.length       1   1043.82 21891 622.29
<none>                                    20848 623.00
- kanga$crest.width           1   1225.95 22074 623.12
- kanga$basilar.length        1   1735.49 22583 625.43
- kanga$nasal.width           1   1846.84 22695 625.93
- kanga$zygomatic.width       1   1989.41 22837 626.56
+ kanga$foramina.length       1      9.59 20838 628.60
+ kanga$occipitonasal.length  1      7.29 20840 628.61
+ kanga$orbital.width         1      4.83 20843 628.62
```

We can see the reasonable values while predicting palate width. Here we used the step function in both-directions it will hence find the best simpler model.

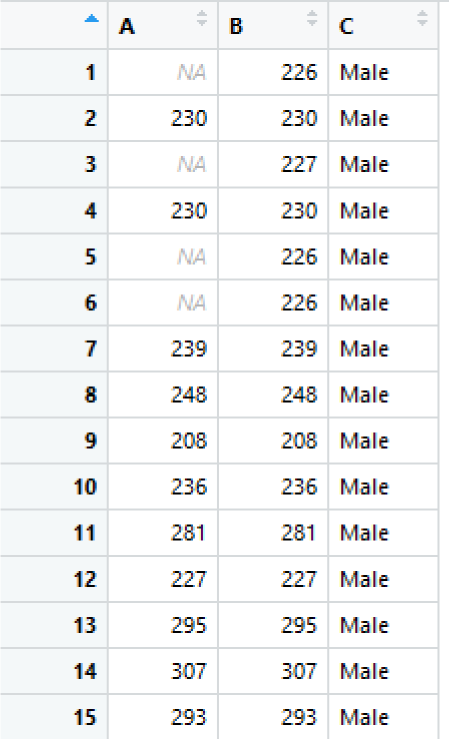
```
> summary(gsmall)

Call:
lm(formula = sex_data ~ kanga_final$occipitonasal.length + kanga_final$nasal.length +
    kanga_final$ramus.height, data = kanga_final)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8014 -0.3422 -0.1057  0.3716  1.1174
```

```
Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     -1.741081   0.488745  -3.562 0.000541 ***
kanga_final$occipitonasal.length  0.007990   0.001272   6.280 6.63e-09 ***
kanga_final$nasal.length        -0.008596   0.001591  -5.403 3.74e-07 ***
kanga_final$ramus.height        -0.005061   0.001198  -4.224 4.90e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4222 on 112 degrees of freedom
Multiple R-squared:  0.3082,	Adjusted R-squared:  0.2896 
F-statistic: 16.63 on 3 and 112 DF,  p-value: 5.275e-09

> 
> lm_f <- lm(kanga$palate.width~kanga$basilar.length+kanga$palate.length+kanga$nasal.length+
kanga$nasal.width+kanga$squamosal.depth+kanga$lacrymal.width+kanga$zygomatic.width+kanga$.ro
stral.width+kanga$occipital.depth+kanga$crest.width+kanga$mandible.length+kanga$mandible.wid
th+kanga$mandible.depth+kanga$ramus.height,data=kanga)
> summary(lm_f)

Call:
lm(formula = kanga$palate.width ~ kanga$basilar.length + kanga$palate.length + 
    kanga$nasal.length + kanga$nasal.width + kanga$squamosal.depth + 
    kanga$lacrymal.width + kanga$zygomatic.width + kanga$.rostral.width + 
    kanga$occipital.depth + kanga$crest.width + kanga$mandible.length + 
    kanga$mandible.width + kanga$mandible.depth + kanga$ramus.height, 
    data = kanga)

Residuals:
    Min      1Q  Median      3Q     Max 
-31.104  -9.954  -0.342  11.105  41.732 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)   
(Intercept)           24.47201   32.20713   0.760  0.44943   
kanga$basilar.length  -0.20425    0.07633  -2.676  0.00893 **
kanga$palate.length    0.08488    0.09791   0.867  0.38840   
kanga$nasal.length    -0.04867    0.05649  -0.862  0.39135   
kanga$nasal.width      0.43801    0.15869   2.760  0.00706 **
kanga$squamosal.depth  0.04375    0.14992   0.292  0.77113   
kanga$lacrymal.width  -0.16995    0.14268  -1.191  0.23690   
kanga$zygomatic.width  0.27617    0.09640   2.865  0.00524 **
kanga$.rostral.width   0.02569    0.11198   0.229  0.81907   
kanga$occipital.depth  0.08388    0.09609   0.873  0.38513   
kanga$crest.width     -0.16017    0.07122  -2.249  0.02708 * 
kanga$mandible.length  0.18953    0.09133   2.075  0.04097 * 
kanga$mandible.width  -0.10816    0.29224  -0.370  0.71222   
kanga$mandible.depth  -0.20383    0.18096  -1.126  0.26313   
kanga$ramus.height    -0.03198    0.09870  -0.324  0.74675   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.57 on 86 degrees of freedom
  (47 observations deleted due to missingness)
Multiple R-squared:  0.776,	Adjusted R-squared:  0.7395 
F-statistic: 21.28 on 14 and 86 DF,  p-value: < 2.2e-16
```

We get the adjusted R^2 value as 0.7395

```
> #checking missing values
> missing <- kanga[is.na(kanga$palate.width),]
> #View(missing)
>
> newpred <- round(predict(lm_f,missing))
>
> #inputting data
> kanga$palate.width[is.na(kanga$palate.width)] <- newpred
> new_pw <- kanga$palate.width
> #################### Predicting sex ####################
>
> kanga_new <- do.call(rbind, Map(data.frame, A=old1, B=new_pw, C=kanga$sex))
```

We here predicted the palate width value and inserted it to the palate.width column.



|  | A | B | C |
|---|---|---|---|
| 1 | NA | 226 | Male |
| 2 | 230 | 230 | Male |
| 3 | NA | 227 | Male |
| 4 | 230 | 230 | Male |
| 5 | NA | 226 | Male |
| 6 | NA | 226 | Male |
| 7 | 239 | 239 | Male |
| 8 | 248 | 248 | Male |
| 9 | 208 | 208 | Male |
| 10 | 236 | 236 | Male |
| 11 | 281 | 281 | Male |
| 12 | 227 | 227 | Male |
| 13 | 295 | 295 | Male |
| 14 | 307 | 307 | Male |
| 15 | 293 | 293 | Male |

*Figure 3: We can see the predicted values of palate.width and their sex values*

```
> kanga_final <- kanga
>
> sex_data <- as.numeric(kanga_final$sex)
>
> var_comb <- kanga_final$basilar.length+kanga_final$occipitonasal.length+kanga_final$palate
.length+kanga_final$palate.width+kanga_final$nasal.length+kanga_final$nasal.width+kanga_fina
l$squamosal.depth+kanga_final$lacrymal.width+kanga_final$zygomatic.width+kanga_final$orbital
.width+kanga_final$.rostral.width+kanga_final$occipital.depth+kanga_final$crest.width+kanga_
final$foramina.length+kanga_final$mandible.length+kanga_final$mandible.width+kanga_final$man
dible.depth+kanga_final$ramus.height
>
> model_kanga_kanga <- lm(sex_data ~ var_comb, data=kanga_final)
>
> summary_kanga<-summary(model_kanga_kanga)
> summary_kanga

Call:
lm(formula = sex_data ~ var_comb, data = kanga_final)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.6765 -0.4553 -0.2409  0.4424  0.9184 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.209e-01  4.608e-01  -0.262 0.793536    
var_comb     1.531e-04  4.427e-05   3.459 0.000764 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4787 on 114 degrees of freedom
  (32 observations deleted due to missingness)
Multiple R-squared:  0.09496,	Adjusted R-squared:  0.08702 
F-statistic: 11.96 on 1 and 114 DF,  p-value: 0.0007645

> 
> plot(model_kanga$fit~kanga_final$sex, xlab="Gender",ylab =" Gender coefficient ")
> points(sex_data,model_kanga$fit )
> abline(1.5 ,0 , lwd =3)
> predictedgender <- model_kanga $ fit > 1.5
> sex_tab <- table (sex_data ,c("Female","Male")[( predictedgender +1) ])
> sex_tab
        
sex_data Female Male
       1     46   11
       2     21   23
> val <- (sex_tab[1,1]+sex_tab[2,2]) / (sex_tab[1,1]+sex_tab[1,2]+sex_tab[2,1]+sex_tab[2,2])
*100
> 
> print(paste("Accuracy =", round(val,2),"%"))
[1] "Accuracy = 68.32 %"
```

Now we finally predict the sex values and we can see from the above result that our accuracy is 68.32% which means after predicting the NA values of palate width we get the same accuracy as removing all the NA values as we did in Q1.