MG 226, June 14th, 2020

Assignment #02

Submitted by:- Kumar Prerak

Question 1

Consider the dataset in >birthwt and use it to develop a risk score for low weight births. Your score should be based on logistic regression and use variables judiciously. Balance health aspects of low weight prediction as well as potential management costs of false positives.

Ans)

Checking for variables in birthwt dataset

```
> library(MASS)
> names(birthwt)

 [1] "low"   "age"   "lwt"   "race"  "smoke" "ptl"   "ht"    "ui"    "ftv"
"bwt"
```

Convert factor variables

```
> birthwt$low <- factor(birthwt$low, levels = c(0,1),labels = c("No", "Yes"))
> birthwt$race <- factor(birthwt$race, levels = c(1:3), labels=c("white","black","o
ther"))
> birthwt$smoke <- factor(birthwt$smoke, levels = c(0,1), labels = c("No", "Yes"))
> birthwt$ht <- factor(birthwt$ht, levels = c(0,1), labels = c("No", "Yes"))
> birthwt$ui <- factor(birthwt$ui, levels = c(0,1),labels = c("No", "Yes"))
> birthwt$ptl <- factor(birthwt$ptl)
> birthwt$ftv <- factor(birthwt$ftv)
```

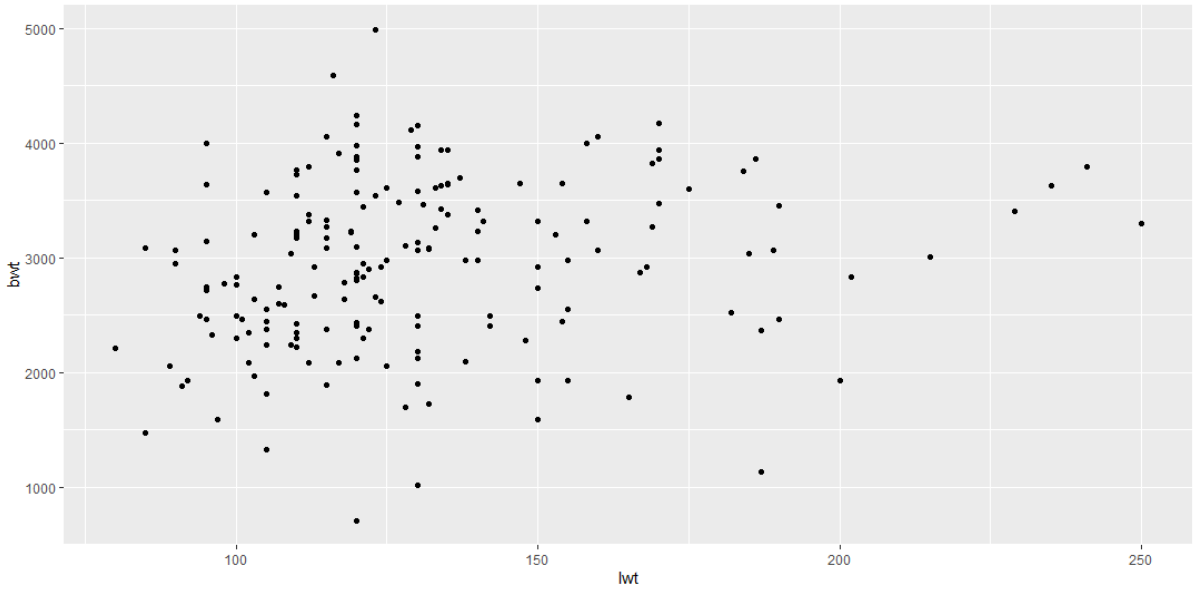
Exploratory Analysis:

View Summary

```
> summary(birthwt)
```

```
  low          age              lwt           race    smoke     ptl       ht        ui
No :130   Min.   :14.00   Min.   : 80.0   white:96   No :115   0:159   No :177   No :161
Yes: 59   1st Qu.:19.00   1st Qu.:110.0   black:26   Yes: 74   1: 24   Yes: 12   Yes: 28
          Median :23.00   Median :121.0   other:67             2:  5
          Mean   :23.24   Mean   :129.8                        3:  1
          3rd Qu.:26.00   3rd Qu.:140.0
          Max.   :45.00   Max.   :250.0
ftv          bwt
0:100   Min.   : 709
1: 47   1st Qu.:2414
2: 30   Median :2977
3:  7   Mean   :2945
4:  4   3rd Qu.:3487
6:  1   Max.   :4990
```

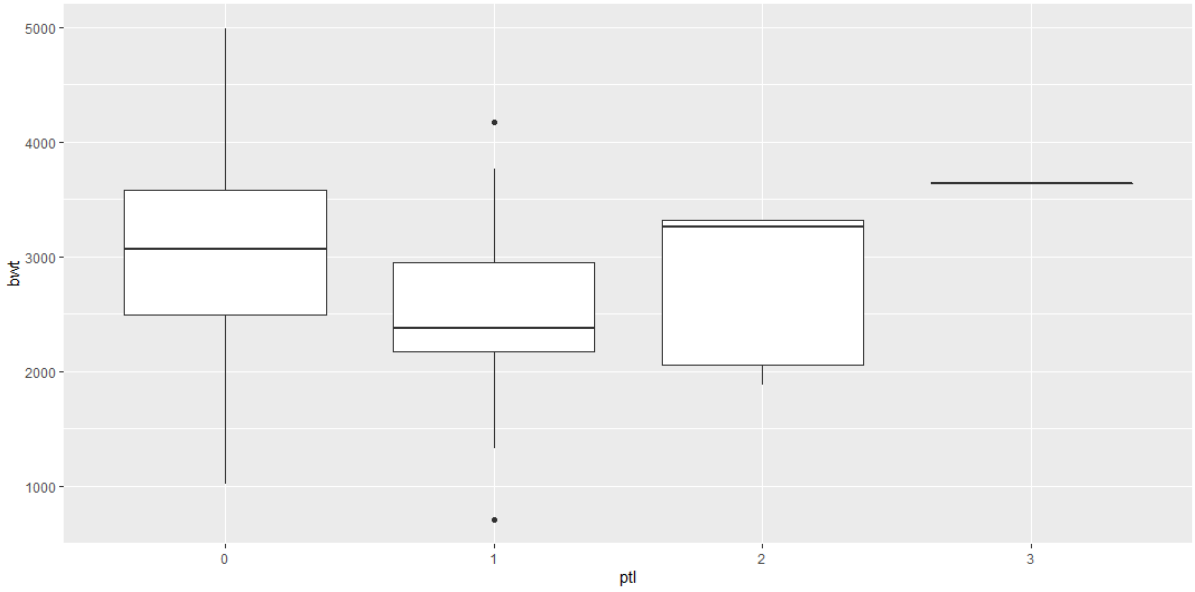
Scatter Plot of Mother's Weight (lwt) vs. Birth Weight (bwt)



Inference: There seems to be a positive correlation between mother's weight and birth weight.

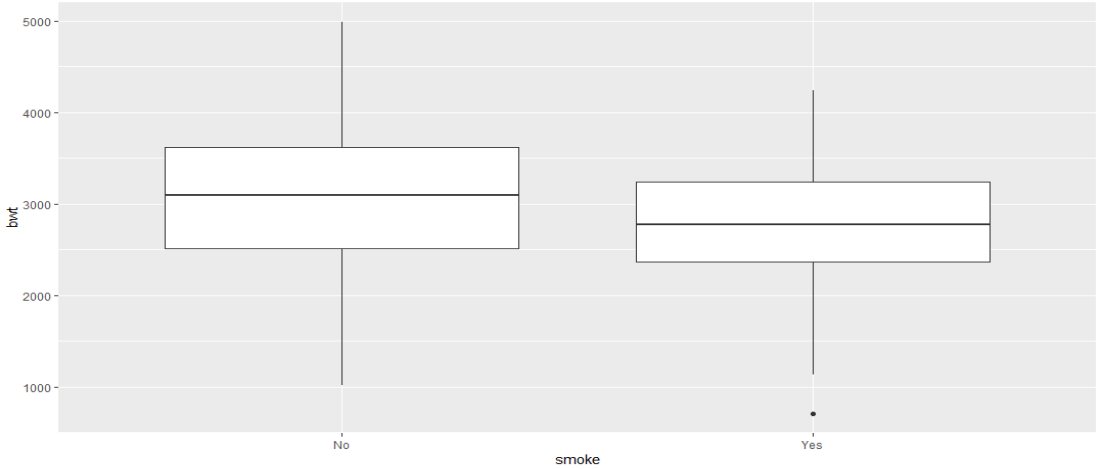
Boxplots of Categorical Variables in Relation to Birth Weight

1. ptl: number of previous premature labours



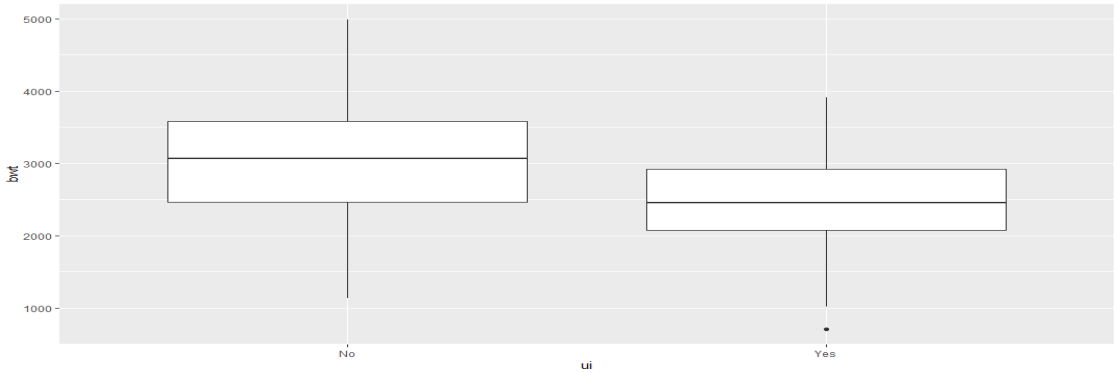
Inference: We observe that birthweight is more if number of previous premature labours is zero. Birthweight decreases when the number of previous premature labours increases to 1. We also observe that we did not have sufficient data for ptl being 2 and 3 (only 5 instances when ptl =2 and 1 instance when ptl=3) to infer any insight based on that.

2. Smoke/No Smoking



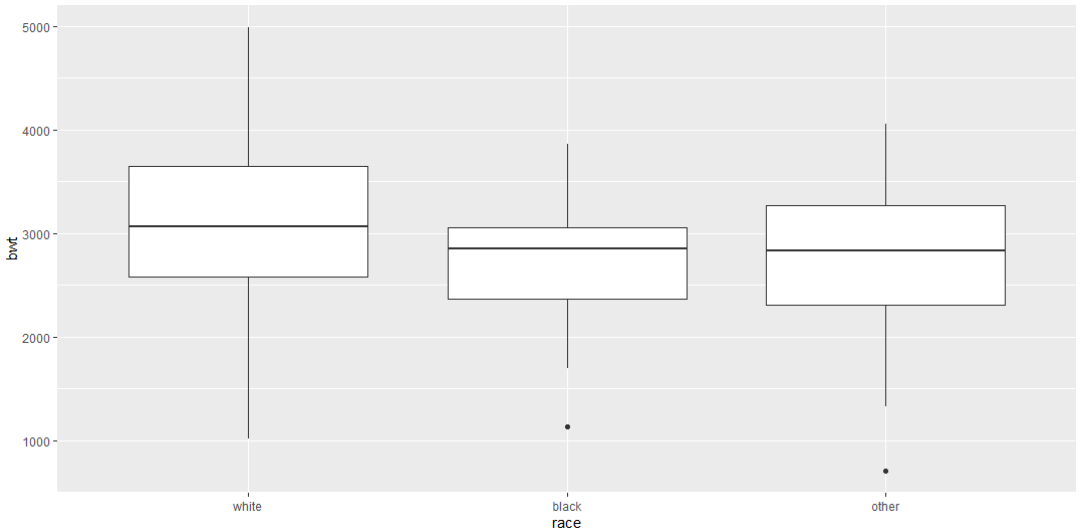
Inference: Birthweight is less for those children whose mother smoked during pregnancy.

3. ui: presence of uterine irritability



Inference: Presence of uterine irritability indicates lower birthweight.

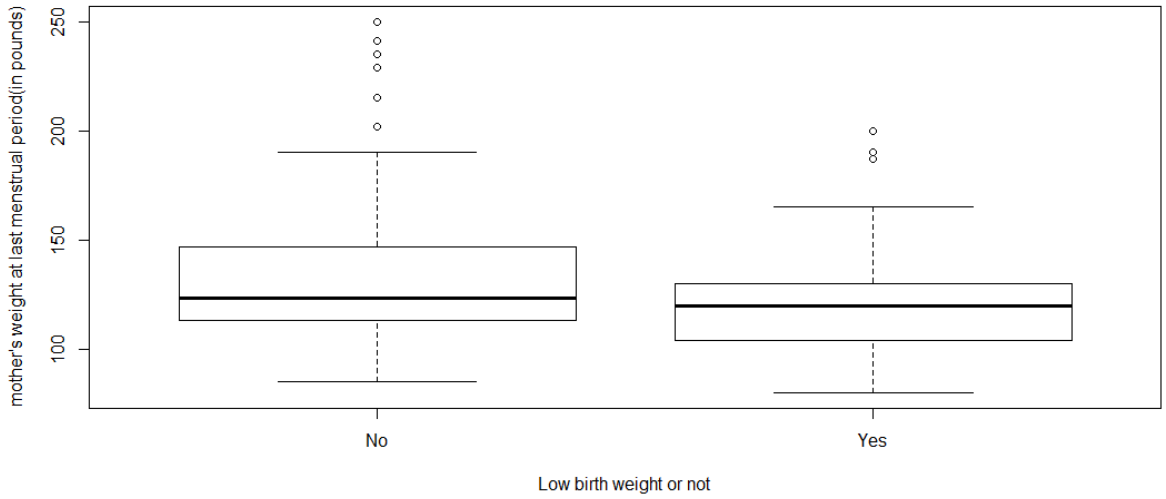
4. Race



Inference: Birthweight was maximum for white children, followed by other and least for black children. Although, factor levels of race appear to be much closer together in terms of data spread.

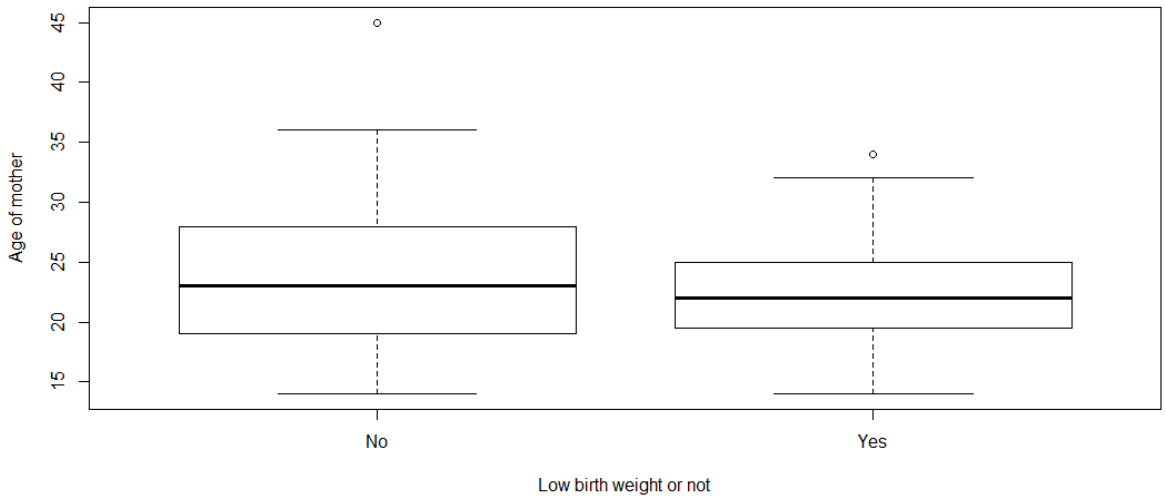
Boxplots of Numerical Variables in Relation to low Birth Weight variable

1. lwt: mother's weight in pounds at last menstrual period.



Inference: Lower mother's weight in pounds at last menstrual period indicates towards low birth weight of child. Although, the difference does not seem to be too significant.

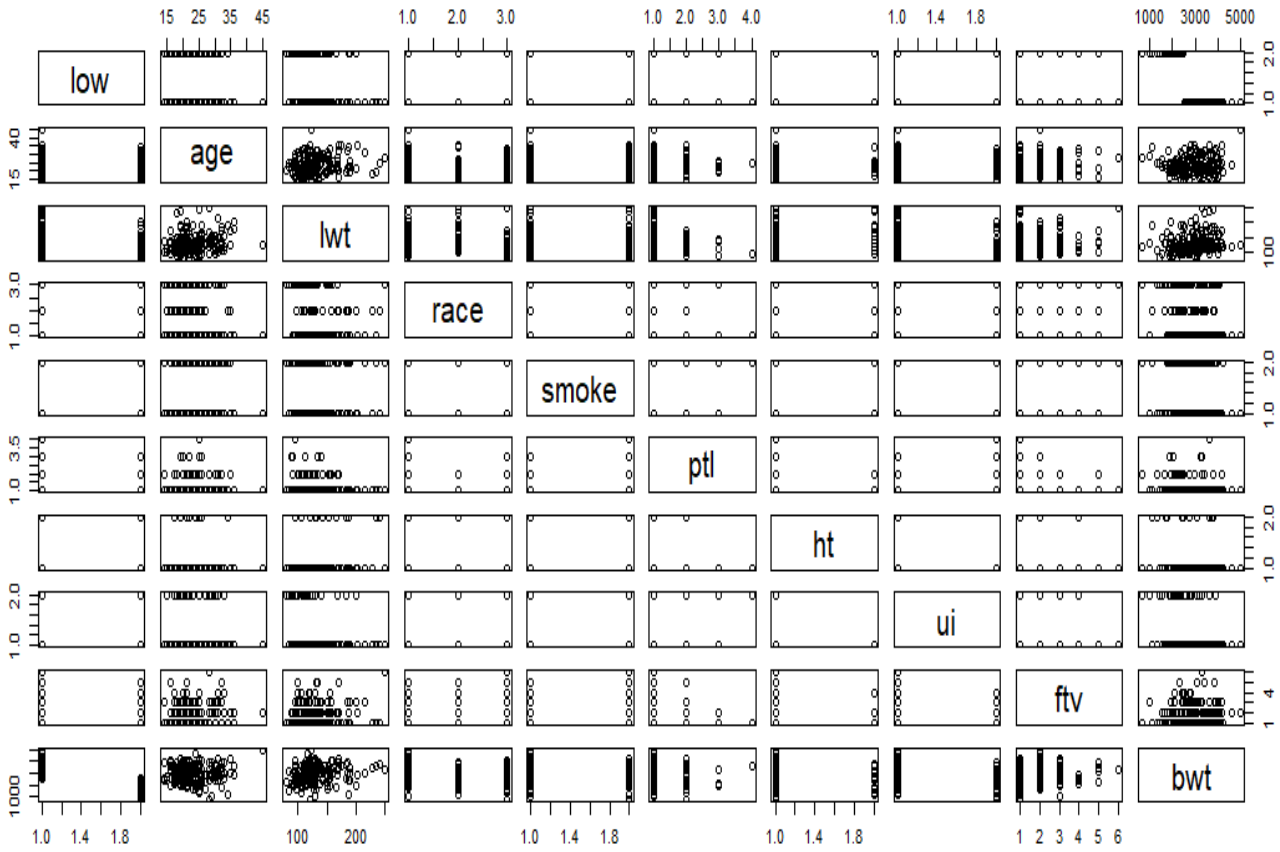
2. age: Age of mother



Inference: Average age of mothers whose children were not suffering from low birth weight was observed to be slightly more than those whose children at birth were observed to be suffering from low birth rate. The difference seems to be very less though.

Pairs plots and correlations of numerical variables:

```
> pairs(birthwt)
```




```
> cor(x=birthwt$age,y=birthwt$bwt)
[1] 0.09031781
> cor(x=birthwt$lwt,y=birthwt$bwt)
[1] 0.1857333
```

Inference: Correlation coefficients between numerical predictors do not appear to be too significant.

Model selection based on Logistic Regression

After running logistic regression model on different set of independent variables and checking if all the variables in the model are coming significant or not(p value<0.05), we try different models and check the summary for the same. We select the model in which all variables are significant with minimum AIC value. The final selected model is :

```
> glm(birthwt$low~birthwt$lwt+birthwt$smoke+birthwt$ht+birthwt$race+birthwt$ui,family="binomial")

Call:  glm(formula = birthwt$low ~ birthwt$lwt + birthwt$smoke + birthwt$ht +
    birthwt$race + birthwt$ui, family = "binomial")

Coefficients:
     (Intercept)       birthwt$lwt   birthwt$smokeYes      birthwt$htYes  birthwt$raceblack
         0.05628          -0.01673           1.03583           1.87142           1.32456
birthwt$raceother      birthwt$uiYes
         0.92620           0.90497

Degrees of Freedom: 188 Total (i.e. Null);  182 Residual
Null Deviance:      234.7
Residual Deviance: 204.2        AIC: 218.2
```

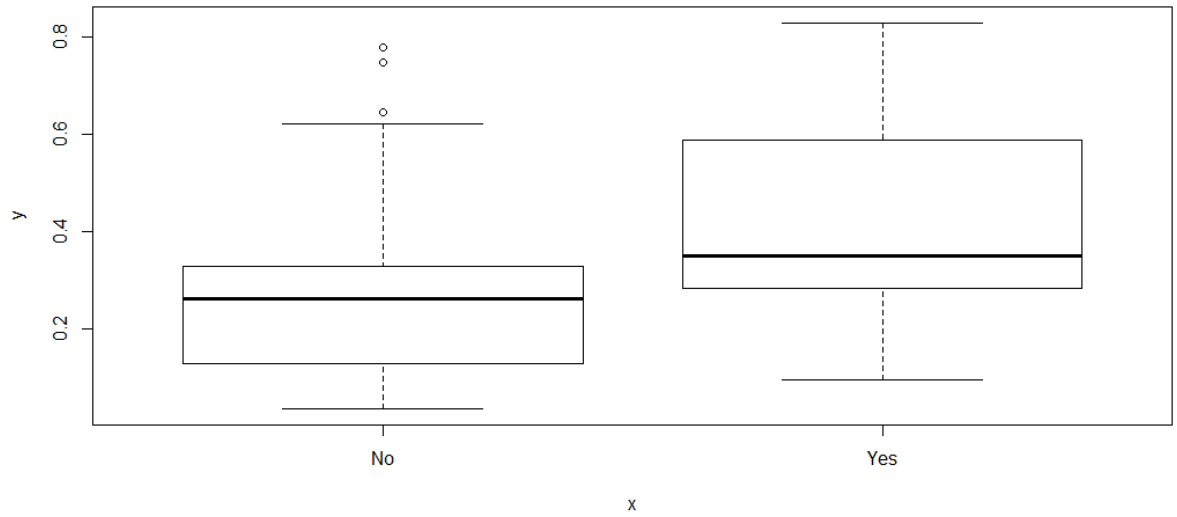
```
> summary(glm(birthwt$low~birthwt$lwt+birthwt$smoke+birthwt$ht+birthwt$race+birthwt$ui,family="binomial"))


Call:
glm(formula = birthwt$low ~ birthwt$lwt + birthwt$smoke + birthwt$ht +
    birthwt$race + birthwt$ui, family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7396  -0.8322  -0.5359   0.9873   2.1692

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       0.056276   0.937853   0.060  0.95215
birthwt$lwt      -0.016732   0.006803  -2.459  0.01392 *
birthwt$smokeYes  1.035831   0.392558   2.639  0.00832 **
birthwt$htYes     1.871416   0.690902   2.709  0.00676 **
birthwt$raceblack 1.324562   0.521464   2.540  0.01108 *
birthwt$raceother 0.926197   0.430386   2.152  0.03140 *
birthwt$uiYes     0.904974   0.447553   2.022  0.04317 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 204.22  on 182  degrees of freedom
AIC: 218.22

Number of Fisher Scoring iterations: 4
```

Then, we fit the above selected model which gives the probabilities (of low birth rate) applied to the dataset.

```
> model.low=glm(birthwt$low~birthwt$lwt+birthwt$smoke+birthwt$ht+birthwt$r
ace+birthwt$ui,family="binomial")

> model.low$fit
```

Setting the Classification threshold in the fitted model :

```
> plot(birthwt$low,model.low$fit)
```



Using the above boxplot, we can assign a cut-off below which we would consider "No low birth weight" and above which we would consider "Low birth weight being Yes". This cut-off point generally depends on the objective of modelling and risk appetite and varies from subject to subject . Here, our objective is to have a cut-off point such that the health aspects of low weight prediction as well as potential management costs of false positives are balanced.

```
> summary(birthwt$low)
 No Yes
130  59
```

Since, when no data is available, there is 31% chance of getting a low birth weight, so we see the result by first keeping the cut-off at 0.31

```
> birthwt$low.pred=ifelse(model.low$fit>0.31,1,0)
> table(birthwt$low,birthwt$low.pred)

      0  1
  No  87 43
  Yes 18 41
```

This gives
Sensitivity =41/59= 0.695, Specificity= 87/130 = 0.669

In order to increase the specificity, we need to increase the cut-off to 0.315 and check for sen sitivity and specificity.

```
> birthwt$low.pred=ifelse(model.low$fit>0.315,1,0)
> table(birthwt$low,birthwt$low.pred)

      0  1
  No  91 39
  Yes 18 41
```

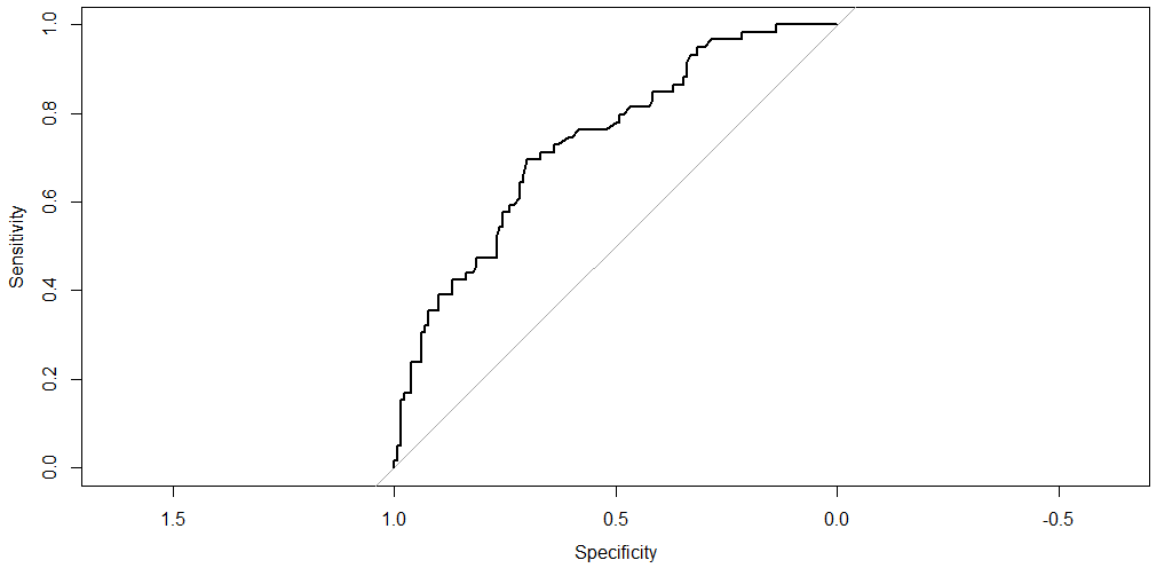
This gives
Sensitivity =41/59= 0.695, Specificity= 91/130 = 0.7

So, we keep the classification threshold to be **0.315**

Evaluation of the fitted Model

1. Plotting ROC-AUC Curve :

```
> library(pROC)
> roc(birthwt$low,model.low$fit)

Setting levels: control = No, case = Yes
Setting direction: controls < cases

Call:
roc.default(response = birthwt$low, predictor = model.low$fit)

Data: model.low$fit in 130 controls (birthwt$low No) < 59 cases (birthwt$l
ow Yes).
Area under the curve: 0.7351

> plot(roc(birthwt$low,model.low$fit))
```



AUC(Area under the curve) represents degree or measure of separability. It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s.

AUC for our model is 0.7351 which suggests that our model has quite good predictive power.

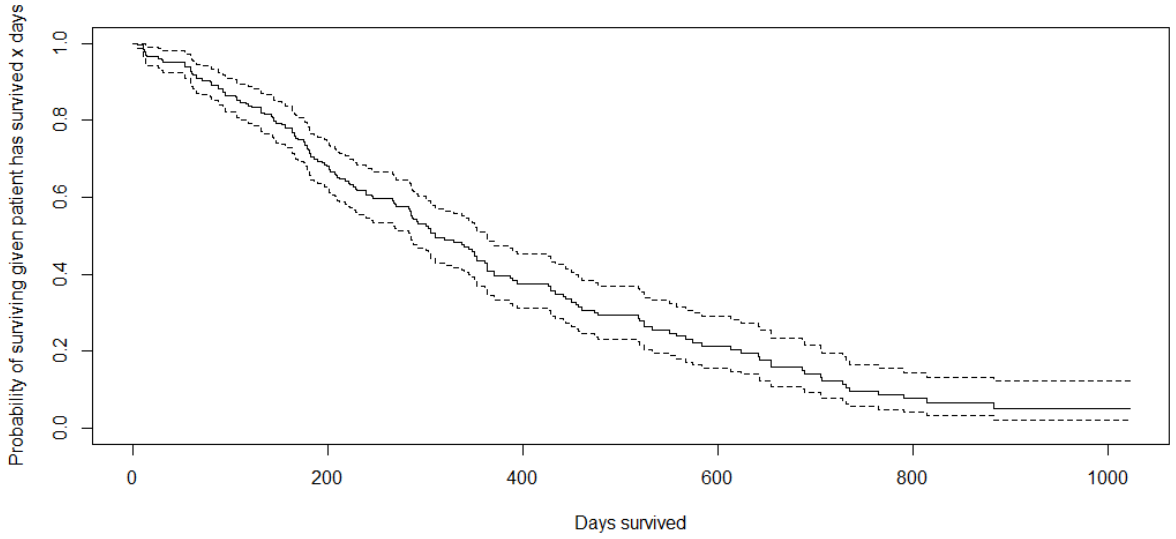
Question 2

Consider the dataset in >lung and estimate both Weibull and Cox models for the hazard function. Demonstrate that, for the two models, the signs of the regression coefficients are opposite for the same covariate. Give a mathematical justification for why this is the case.

Ans)

Getting the lung dataset inbuilt in survival library and plotting the graph

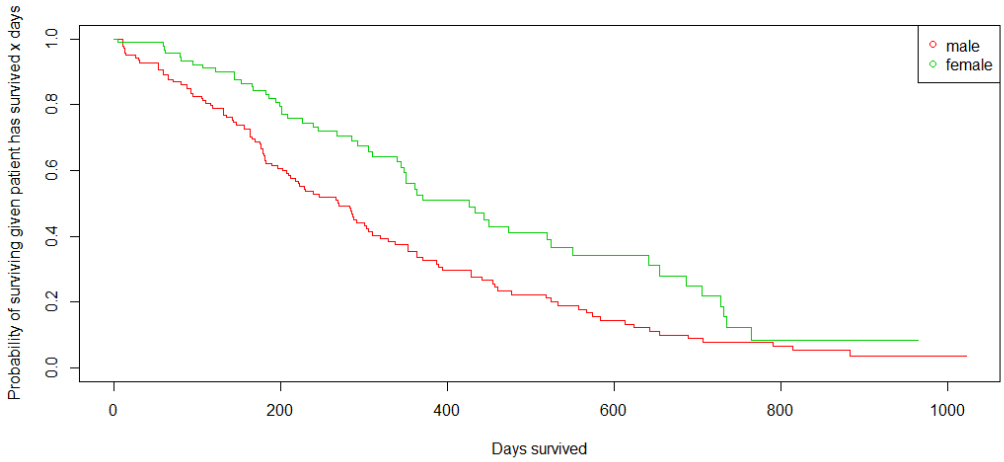
```
> library(survival)
> plot(Surv(time,status),xlab="Days survived",ylab="Probability of survivi
ng given patient has survived x days")
```



Exploratory Analysis

Plotting graphs to find which variables can be significant:

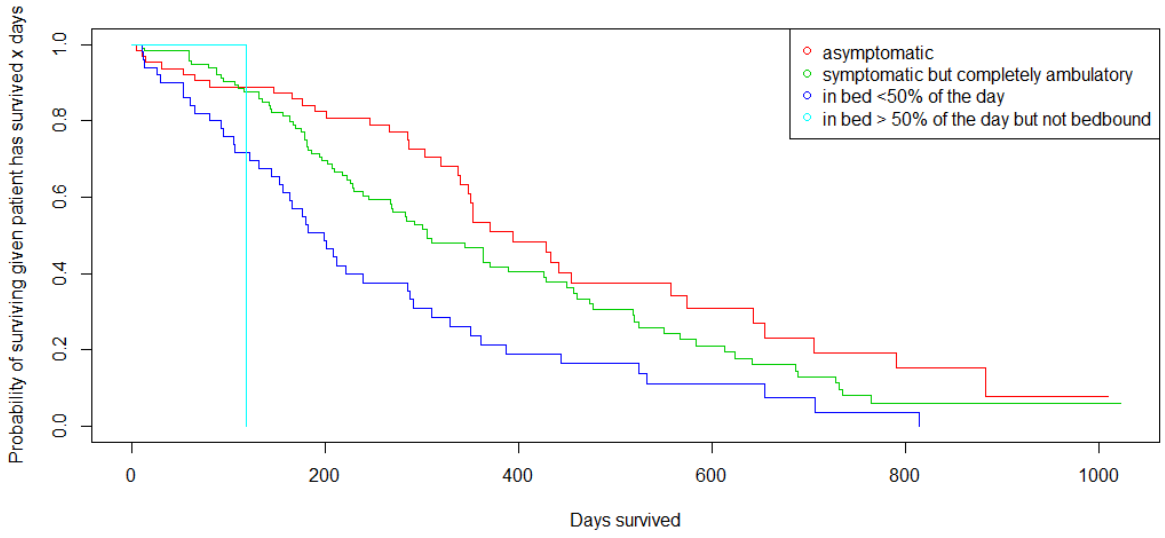
1. Sex(1:Male, 2:Female)



```
> survfit(Surv(lung$time,lung$status)~lung$sex)

Call: survfit(formula = Surv(lung$time, lung$status) ~ lung$sex)

              n events median 0.95LCL 0.95UCL
lung$sex=1 138    112    270     212     310
lung$sex=2  90     53    426     348     550
```

Inference: Survival probability for female is more than that of men.

2. ph.ecog: ECOG performance score

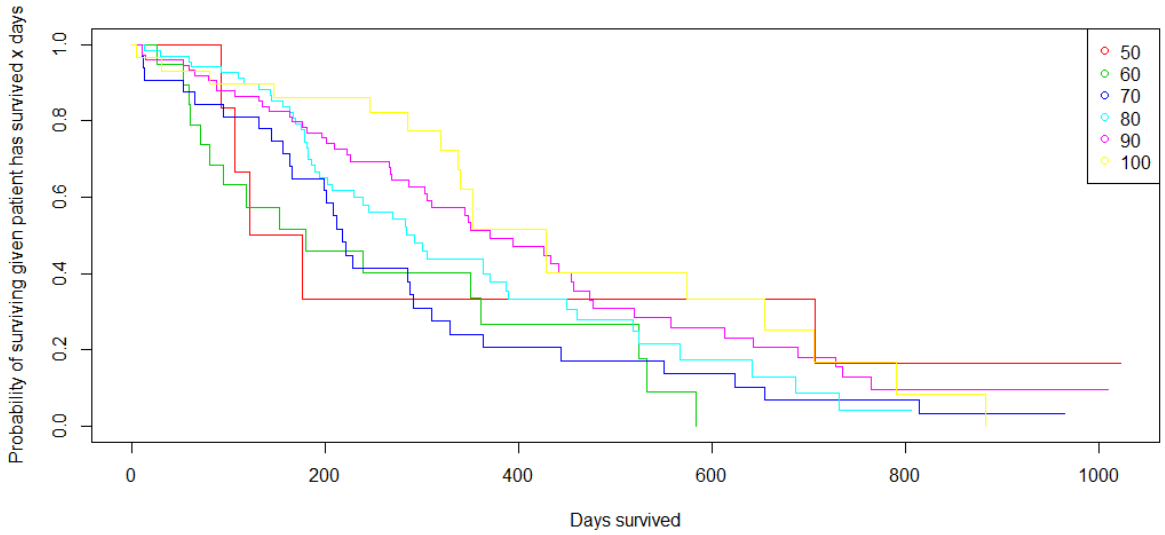



```
> survfit(Surv(lung$time,lung$status)~(lung$ph.ecog))
Call: survfit(formula = Surv(lung$time, lung$status) ~ (lung$ph.ecog))

   1 observation deleted due to missingness
                 n events median 0.95LCL 0.95UCL
lung$ph.ecog=0  63     37    394     348     574
lung$ph.ecog=1 113     82    306     268     429
lung$ph.ecog=2  50     44    199     156     288
lung$ph.ecog=3   1      1    118      NA      NA
```

Inference: As ECOG performance score increases, survival probability decreases.

3. ph.karno: Karnofsky performance score rated by physician

```
> survfit(Surv(lung$time,lung$status)~(lung$ph.karno))

Call: survfit(formula = Surv(lung$time, lung$status) ~ (lung$ph.karno))
1 observation deleted due to missingness
                   n events median 0.95LCL 0.95UCL
lung$ph.karno=50   6      5    150     107      NA
lung$ph.karno=60  19     16    180      95      NA
lung$ph.karno=70  32     29    218     199     310
lung$ph.karno=80  67     47    293     230     390
lung$ph.karno=90  74     49    371     306     473
lung$ph.karno=100 29     18    428     340       N
```



Inference: As Karnofsky performance score rated by physician increases, survival probability increases.

4. pat.karno: Karnofsky performance score as rated by patient

```
> survfit(Surv(lung$time,lung$status)~(lung$pat.karno))

Call: survfit(formula = Surv(lung$time, lung$status) ~ (lung$pat.karno))

   3 observations deleted due to missingness
                     n events median 0.95LCL 0.95UCL
lung$pat.karno=30    2      1  156.0     156      NA
lung$pat.karno=40    2      1   93.0      93      NA
lung$pat.karno=50    4      4   93.5      12      NA
lung$pat.karno=60   30     27  197.0     163     288
lung$pat.karno=70   41     31  267.0     176     519
lung$pat.karno=80   51     39  348.0     226     520
lung$pat.karno=90   60     38  426.0     286     473
lung$pat.karno=100  35     21  371.0     310      NA
```

Inference: There was no monotonic relation observed between Karnofsky performance score as rated by patient and survival probability.

5. age

```
> summary(survreg(Surv(lung$time,lung$status)~(lung$age)))

Call:
survreg(formula = Surv(lung$time, lung$status) ~ (lung$age))
               Value Std. Error     z        p
(Intercept)  6.8871     0.4466 15.42 < 2e-16
lung$age    -0.0136     0.0070 -1.94   0.052
Log(scale)  -0.2761     0.0624 -4.43 9.6e-06

Scale= 0.759

Weibull distribution
Loglik(model)= -1151.9   Loglik(intercept only)= -1153.9
        Chisq= 3.91 on 1 degrees of freedom, p= 0.048
```

```
Number of Newton-Raphson Iterations: 5
n= 228
```

Inference: Here, we just fail to reject the null hypothesis that coefficient of age is 0 which Signifies that the age variable is not too significant as a predictor.

6. meal.calorie: Calories consumed at meal

```
> summary(survreg(Surv(lung$time,lung$status)~(lung$meal.cal)))

Call:
survreg(formula = Surv(lung$time, lung$status) ~ (lung$meal.cal))
                  Value Std. Error     z       p
(Intercept)    5.91e+00   1.79e-01 33.10 < 2e-16
lung$meal.cal  9.07e-05   1.77e-04  0.51 0.60865
Log(scale)    -2.55e-01   6.98e-02 -3.66 0.00026

Scale= 0.775

Weibull distribution
Loglik(model)= -932.1   Loglik(intercept only)= -932.2
        Chisq= 0.27 on 1 degrees of freedom, p= 0.61
Number of Newton-Raphson Iterations: 5
n=181 (47 observations deleted due to missingness)
```

Inference: Calories consumed at meal is not a significant predictor.

7. wt.loss: Weight loss in last six months

```
> summary(survreg(Surv(lung$time,lung$status)~(lung$wt.loss)))

Call:
survreg(formula = Surv(lung$time, lung$status) ~ (lung$wt.loss))
                 Value Std. Error     z      p
(Intercept)   6.085141   0.075736 80.35 <2e-16
lung$wt.loss -0.000958   0.004513 -0.21   0.83
Log(scale)   -0.294795   0.065103 -4.53  6e-06

Scale= 0.745

Weibull distribution
Loglik(model)= -1069   Loglik(intercept only)= -1069.1
        Chisq= 0.04 on 1 degrees of freedom, p= 0.83
Number of Newton-Raphson Iterations: 5
n=214 (14 observations deleted due to missingness)
```

Inference: Weight loss in last six months is not a significant predictor.

Weibull Regression Model for hazard function:

After trying few models based on above EDA , we select the final model based on keeping only significant predictors in the model as following:

```
> survreg(Surv(lung$time,lung$status)~lung$sex+lung$ph.ecog)
```

```
Call:
survreg(formula = Surv(lung$time, lung$status) ~ lung$sex + lung$ph.ecog)

Coefficients:
 (Intercept)     lung$sex lung$ph.ecog
   5.8195907    0.4013684   -0.3557319

Scale= 0.7310495

Loglik(model)= -1133.1   Loglik(intercept only)= -1147.4
        Chisq= 28.73 on 2 degrees of freedom, p= 5.76e-07
n=227 (1 observation deleted due to missingness)
```

```
> summary(survreg(Surv(lung$time,lung$status)~lung$sex+lung$ph.ecog))

Call:
survreg(formula = Surv(lung$time, lung$status) ~ lung$sex + lung$ph.ecog)

                Value Std. Error     z       p
(Intercept)    5.8196     0.1902 30.60 < 2e-16
lung$sex       0.4014     0.1237  3.24  0.0012
lung$ph.ecog -0.3557     0.0826 -4.31 1.7e-05
Log(scale)   -0.3133     0.0613 -5.11 3.3e-07

Scale= 0.731

Weibull distribution
Loglik(model)= -1133.1   Loglik(intercept only)= -1147.4
        Chisq= 28.73 on 2 degrees of freedom, p= 5.8e-07
Number of Newton-Raphson Iterations: 5
n=227 (1 observation deleted due to missingness)
```

Therefore,

$\dot{\eta}$=exp(5.8196+0.4014(sex)-0.3557(ph.ecog))

β= 1/scale = 1.367896

Cox Regression Model:

```
> coxph(Surv(lung$time,lung$status)~lung$sex+lung$ph.ecog)

Call:
coxph(formula = Surv(lung$time, lung$status) ~ lung$sex + lung$ph.ecog)

                 coef  exp(coef) se(coef)    z        p
lung$sex     -0.5530    0.5752   0.1676 -3.300 0.000967
lung$ph.ecog  0.4875    1.6282   0.1122  4.344  1.4e-05

Likelihood ratio test=29.05  on 2 df, p=4.91e-07
n= 227, number of events= 164
   (1 observation deleted due to missingness)
```

```
> summary(coxph(Surv(lung$time,lung$status)~lung$sex+lung$ph.ecog))
Call:
coxph(formula = Surv(lung$time, lung$status) ~ lung$sex + lung$ph.ecog)

  n= 227, number of events= 164
   (1 observation deleted due to missingness)
```

```
                coef exp(coef) se(coef)      z Pr(>|z|)
lung$sex     -0.5530    0.5752   0.1676 -3.300 0.000967 ***
lung$ph.ecog  0.4875    1.6282   0.1122  4.344  1.4e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

             exp(coef) exp(-coef) lower .95 upper .95
lung$sex        0.5752     1.7384    0.4142    0.7989
lung$ph.ecog    1.6282     0.6142    1.3067    2.0288

Concordance= 0.642  (se = 0.025 )
Likelihood ratio test= 29.05  on 2 df,   p=5e-07
Wald test            = 28.96  on 2 df,   p=5e-07
Score (logrank) test = 29.41  on 2 df,   p=4e-07
```

Clearly, the signs of the regression coefficients are opposite for the same covariates in Weibull regression model and cox regression model.

Covariate	Weibull Coefficient	Cox Coefficient
Sex	0.4014	-0.5530
Ph.ecog	-0.3557	0.4875

Mathematical justification of observing opposite signs in coefficients :

Although the Weibull hazard model can be written in both a proportional hazard form and an accelerated failure time(AFT) form, we have used Survreg function which uses the AFT form. In an AFT model, we model the characteristic failure time (time to failure). Positive coefficients are good (longer time to death).

While Cox hazard model is written in proportional hazard form. Coxph uses proportion hazard form(PH). In a PH model, we model the death rate or the risk . Positive coefficients are bad (higher death rate).

Question 3

Consider the dataset in >Boston and identify predictive variables for median house value. Use both a p-value approach as well as a lasso approach. Do they give the same or similar set of variables? Do you think there is an effect due to correlations between predictor variables?

Ans)

i) p-value approach:

Let us consider all the variables in the model and run a linear regression to predict the median house value and then check the p-value of each predictors to identify which among them are significant.

```
> lm(medv~.,data=Boston)

Call:
lm(formula = medv ~ ., data = Boston)

Coefficients:
(Intercept)         crim           zn        indus         chas           n
ox           rm
  3.646e+01   -1.080e-01    4.642e-02    2.056e-02    2.687e+00   -1.777e+
01    3.810e+00
        age          dis          rad          tax      ptratio        bla
ck        lstat
  6.922e-04   -1.476e+00    3.060e-01   -1.233e-02   -9.527e-01    9.312e-
03   -5.248e-01
```

```
> summary(lm(medv~.,data=Boston))

Call:
lm(formula = medv ~ ., data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-15.595  -2.730  -0.518   1.777  26.199

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
crim        -1.080e-01  3.286e-02  -3.287 0.001087 **
zn           4.642e-02  1.373e-02   3.382 0.000778 ***
indus        2.056e-02  6.150e-02   0.334 0.738288
chas         2.687e+00  8.616e-01   3.118 0.001925 **
nox         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
rm           3.810e+00  4.179e-01   9.116  < 2e-16 ***
age          6.922e-04  1.321e-02   0.052 0.958229
dis         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
rad          3.060e-01  6.635e-02   4.613 5.07e-06 ***
tax         -1.233e-02  3.760e-03  -3.280 0.001112 **
ptratio     -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
black        9.312e-03  2.686e-03   3.467 0.000573 ***
lstat       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.745 on 492 degrees of freedom
Multiple R-squared:  0.7406,    Adjusted R-squared:  0.7338
F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

We observe that p-value for indus and age are quite high suggesting that these variables are n ot significant in predicting the median house value in the presence of other independent variables.
So, we remove indus and age from our model and run the linear regression model again:

```
> lm(medv~crim+zn+chas+nox+rm+dis+rad+tax+ptratio+black+lstat,data=Boston)

Call:
lm(formula = medv ~ crim + zn + chas + nox + rm + dis + rad +
    tax + ptratio + black + lstat, data = Boston)

Coefficients:
(Intercept)         crim           zn         chas          nox
rm          dis
  36.341145    -0.108413     0.045845     2.718716   -17.376023     3.8015
79    -1.492711
        rad          tax      ptratio        black        lstat
   0.299608    -0.011778    -0.946525     0.009291    -0.522553

> summary(lm(medv~crim+zn+chas+nox+rm+dis+rad+tax+ptratio+black+lstat,data
=Boston))

Call:
lm(formula = medv ~ crim + zn + chas + nox + rm + dis + rad +
    tax + ptratio + black + lstat, data = Boston)

Residuals:
     Min       1Q   Median       3Q      Max
-15.5984  -2.7386  -0.5046   1.7273  26.2373

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  36.341145   5.067492   7.171 2.73e-12 ***
crim         -0.108413   0.032779  -3.307 0.001010 **
zn            0.045845   0.013523   3.390 0.000754 ***
chas          2.718716   0.854240   3.183 0.001551 **
nox         -17.376023   3.535243  -4.915 1.21e-06 ***
rm            3.801579   0.406316   9.356  < 2e-16 ***
dis          -1.492711   0.185731  -8.037 6.84e-15 ***
rad           0.299608   0.063402   4.726 3.00e-06 ***
tax          -0.011778   0.003372  -3.493 0.000521 ***
ptratio      -0.946525   0.129066  -7.334 9.24e-13 ***
black         0.009291   0.002674   3.475 0.000557 ***
lstat        -0.522553   0.047424 -11.019  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.736 on 494 degrees of freedom
Multiple R-squared:  0.7406,    Adjusted R-squared:  0.7348
F-statistic: 128.2 on 11 and 494 DF,  p-value: < 2.2e-16
```

Inference: Here**,** we observe that all the predictor variables in the model are significant and around 74% of total variability in the median house value can be explained by the predictors in this model.

ii) LASSO approach:

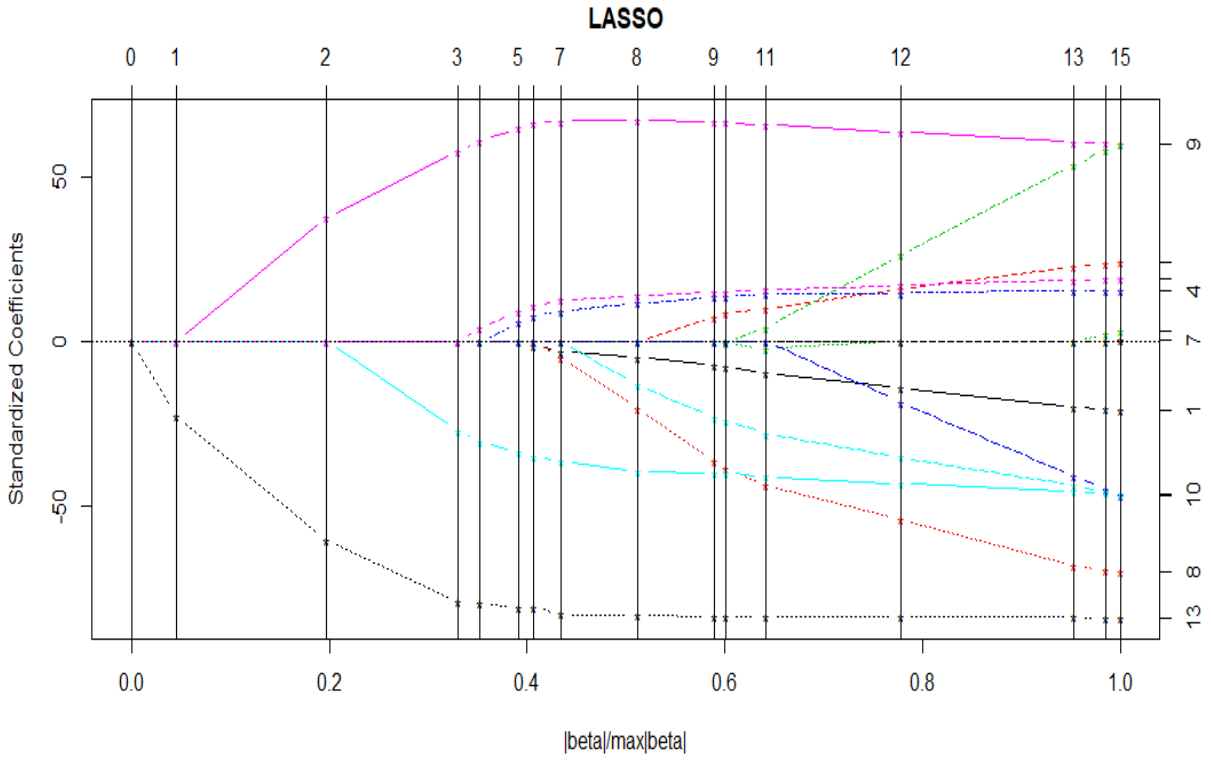
```
> lars(as.matrix(Boston[,-14]),y=Boston[,14])

Call:
lars(x = as.matrix(Boston[, -14]), y = Boston[, 14])
R-squared: 0.741
Sequence of LASSO moves:
     lstat rm ptratio black chas crim dis nox zn indus rad tax indus indus age
Var     13  6      11    12    4    1   8   5  2     3   9  10    -3     3   7
Step     1  2       3     4    5    6   7   8  9    10  11  12    13    14  15
```

The above function tells us which variable is included in the model at what step starting from no predictor variable in the model initially with a very high value of penalty term.

Here, we observe that indus variable gets included in the model in 10^{th} step but then gets dropped in 13^{th} step because of inclusion of other predictors which makes indus insignificant in their presence. Then again indus gets included in the second last step followed by age indicating that age and indus are two least significant predictors for median house value.

Let us plot it to see the inclusion of variables at each step graphically:

```
> plot(lars(as.matrix(Boston[,-14]),y=Boston[,14]))
```



Deciding which variables to keep in the model:

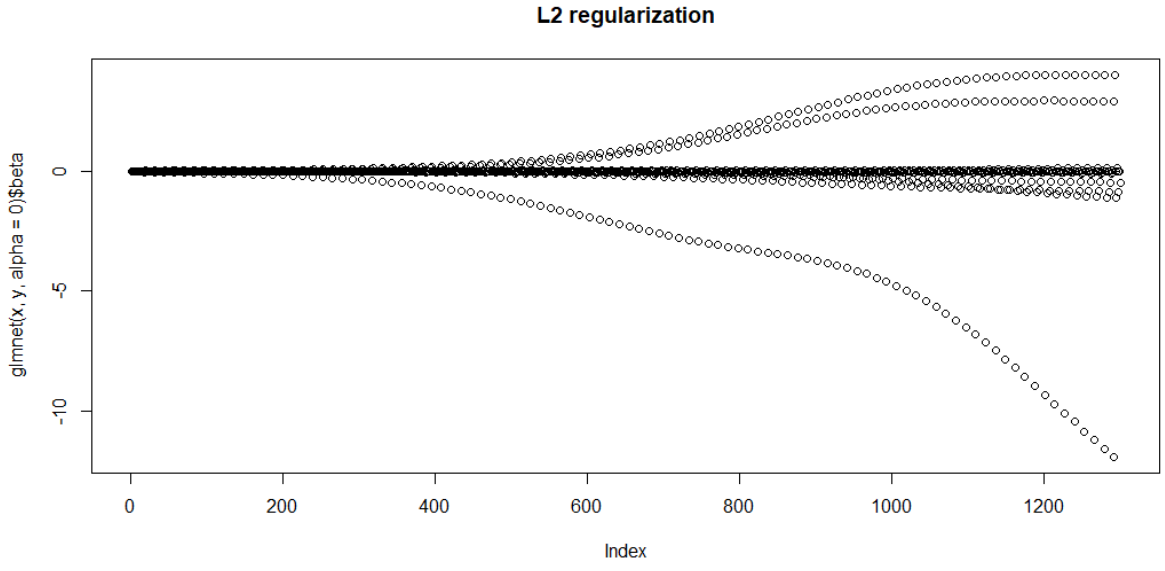
```
> summary(lars(as.matrix(Boston[,-14]),y=Boston[,14]))
LARS/LASSO
Call: lars(x = as.matrix(Boston[, -14]), y = Boston[, 14])

   Df   Rss       Cp
0   1 42716 1392.997
1   2 36326 1111.195
2   3 21335  447.485
3   4 14960  166.356
4   5 14402  143.588
5   6 13667  112.931
6   7 13449  105.281
7   8 13117   92.515
8   9 12423   63.717
9  10 11950   44.700
10 11 11899   44.446
11 12 11730   38.934
12 13 11317   22.590
13 12 11086   10.341
14 13 11080   12.032
15 14 11079   14.000
```

Using this summary table, we observe that as more and more variables keep on getting included in the model, residual sum of squares keep on decreasing. To decide which model to select, we can use the Mallow's Cp value. Lower Mallow's Cp value indicates towards better model. Here, the lowest Cp value is 10.341 which is arrived at the third last step. After that, the Cp value increases. Hence, we will stop including variable at that step which means the last two variables "indus" and "age" will not be present in our final model.

So, we observe that both p-value approach and LASSO approach gave the same set of variables.

Checking for multicollinearity using Ridge trace:

```
> library(glmnet)
> x=as.matrix(Boston[,-14])
> y=Boston[,14]
> plot(glmnet(x,y,alpha=0)$beta,main="L2 regularization")
```



We know that in a multicollinearity situation, before the coefficient shrinks to zero, it changes sign. It swings to the other side of the zero and comes back. So, if the data has a lot of multicollinearity among predictor variables, the above ridge coefficient plot would be very unstable. In multicollinear data, on the right side is the standard least squares fit and the penalty is increasing, and the numbers are very unstable. But in this dataset, we do not observe any such instance in the above ridge regularization coefficient plot, and we can infer that **multicollinearity among predictor variables is not there in this dataset.**

We can also verify this using Variance Inflation factor (VIF) quantifies the extent of correlation between one predictor and the other predictors in a model.

VIF= $1/(1-R^2)$

VIF>10 suggests there is multicollinearity among those predictor variables.

```
> library(car)
> model<-lm(medv~.,data=Boston)
> vif(model)
    crim       zn    indus     chas      nox       rm      age      dis      rad      tax
1.792192 2.298758 3.991596 1.073995 4.393720 1.933744 3.100826 3.955945 7.484496 9.008554
 ptratio    black    lstat
1.799084 1.348521 2.941491
```

We can see that vif for none of the predictor variables is more than 10. Vif for tax and rad are high which indicates that there is some correlation between rad and tax but it is not a problem since the value is less than 10. So, we can conclude that **there is no multicollinearity effect between predictor variables in this dataset.**