HW9

Shaojie Zhang (shaojiez)

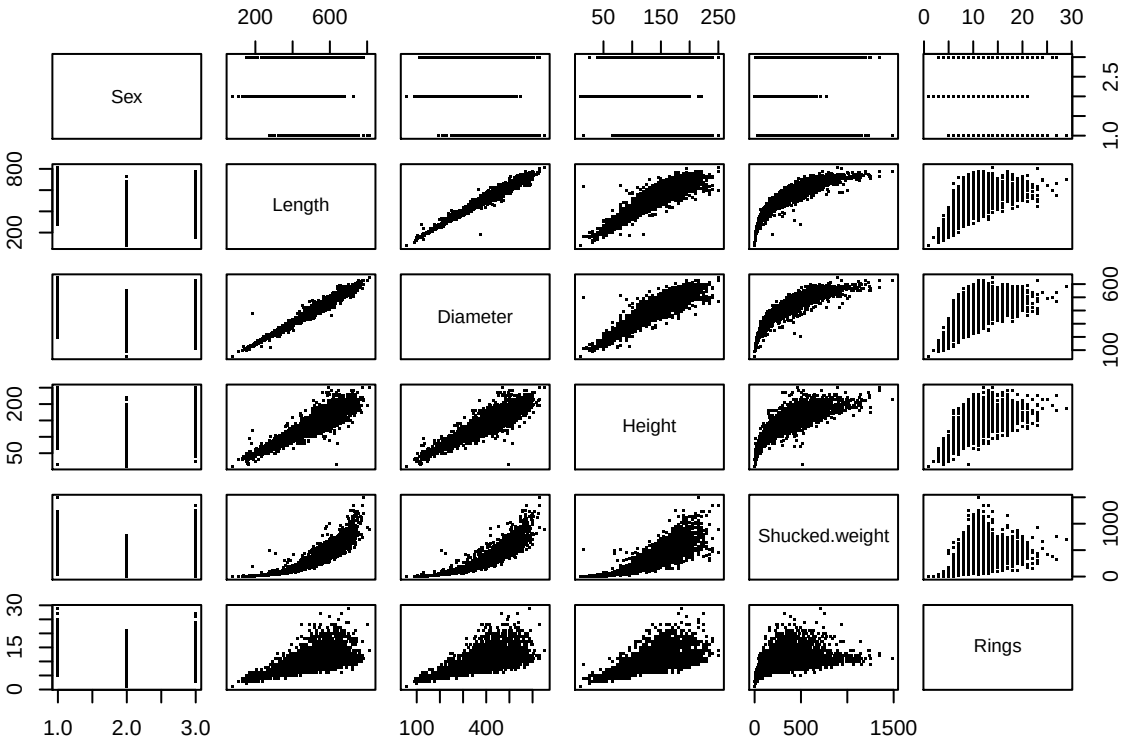
04/04/2019

Problem 1

```
library(gam)
```

```
## Warning: package 'gam' was built under R version 3.4.4
## Loading required package: splines
## Loading required package: foreach
## Warning: package 'foreach' was built under R version 3.4.3
## Loaded gam 1.16
```

```
abalone=read.csv("abalonemt.csv",header=T)

# exploratory analysis
pairs(abalone[,c(1,2,3,4,6,9)],pch=".")
```



Before we get into the actually problems, follow the instructions and do exploratory analysis and basic EDA.

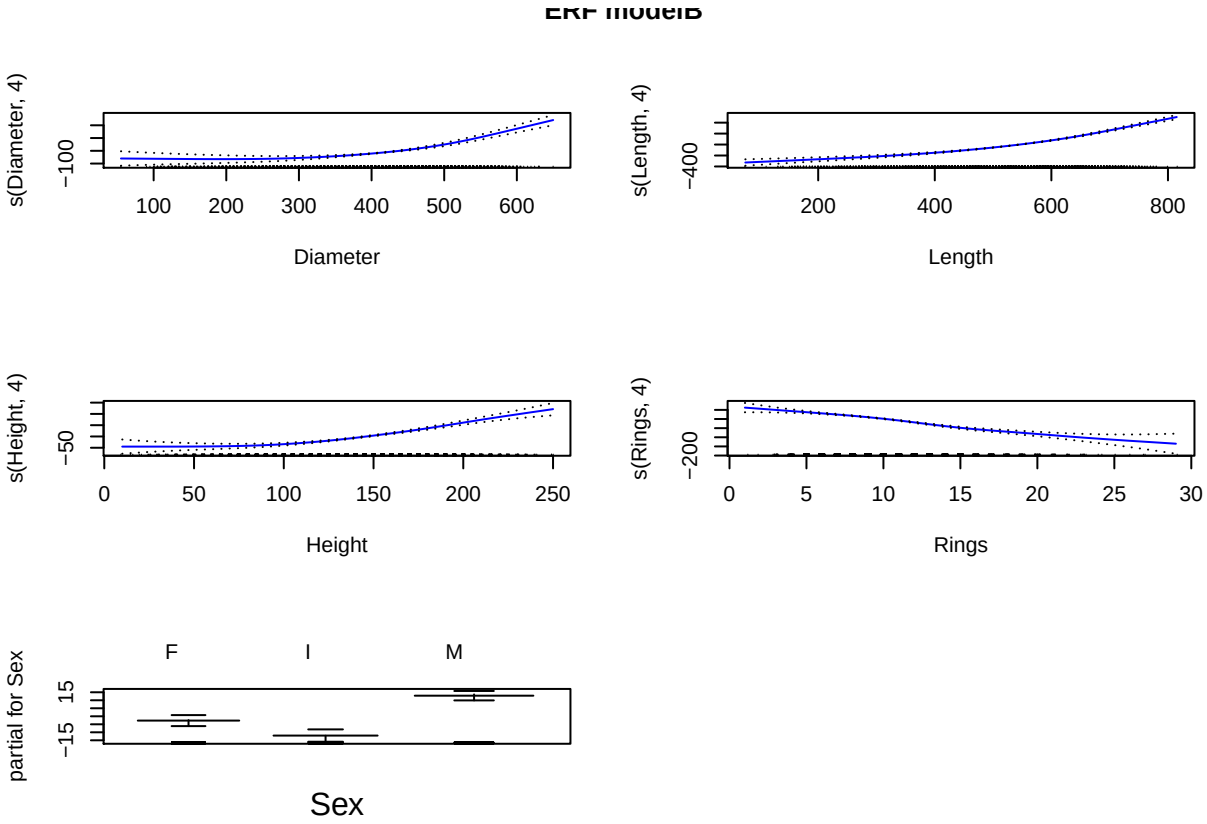
a

```
modelA = gam(Shucked.weight ~ Diameter+Length+Height+Rings+Sex, data=abalone)
summary(modelA)
```

```
##
## Call: gam(formula = Shucked.weight ~ Diameter + Length + Height + Rings +
##     Sex, data = abalone)
## Deviance Residuals:
##     Min      1Q  Median      3Q     Max
## -209.70  -59.30  -15.67   38.29  675.02
##
## (Dispersion Parameter for gaussian family taken to be 8410.622)
##
##     Null Deviance: 205066331 on 4172 degrees of freedom
## Residual Deviance: 35038652 on 4166 degrees of freedom
## AIC: 49563.9
##
## Number of Local Scoring Iterations: 2
##
## Anova for Parametric Effects
##             Df    Sum Sq   Mean Sq   F value    Pr(>F)
## Diameter     1 163650949 163650949 19457.651 < 2.2e-16 ***
## Length       1   2152783   2152783   255.960 < 2.2e-16 ***
## Height       1    643485    643485    76.509 < 2.2e-16 ***
## Rings        1   3165192   3165192   376.333 < 2.2e-16 ***
## Sex          2    415271    207636    24.687 2.195e-11 ***
## Residuals 4166  35038652      8411
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

b

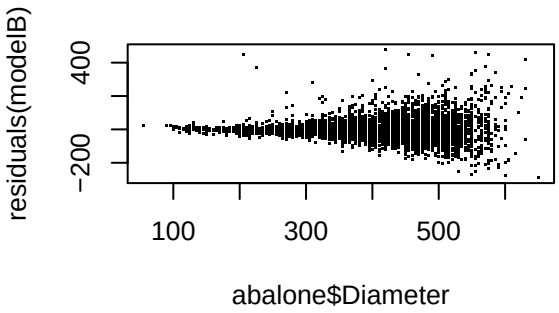
```r
modelB = gam(Shucked.weight ~ s(Diameter,4) + s(Length,4) + s(Height,4) + s(Rings,4) + Sex, data=abalone

par(mfrow = c(3,2))
plot.Gam(modelB, scale = 0, se = TRUE, col = "blue", lwd = 1)
title(main = "ERF modelB", outer = T)
```

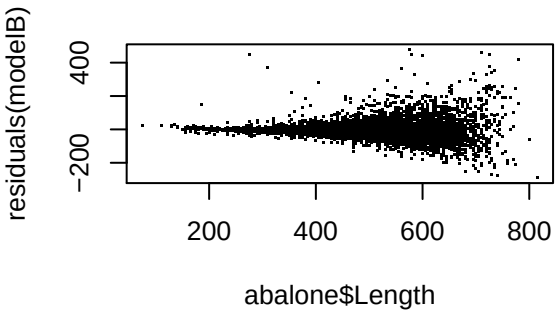


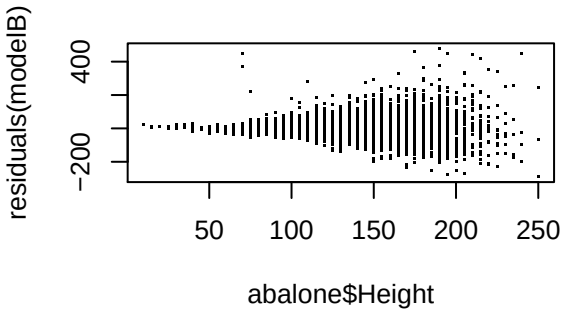
From the plot, we see that for Diameter, Length and Height, they all have a positive impact on Shucked weight. Ring parameter has a negative impact on shucked weight. As for Sex, the infants has the lowest shucked weight while the males have the highest shucked weight.

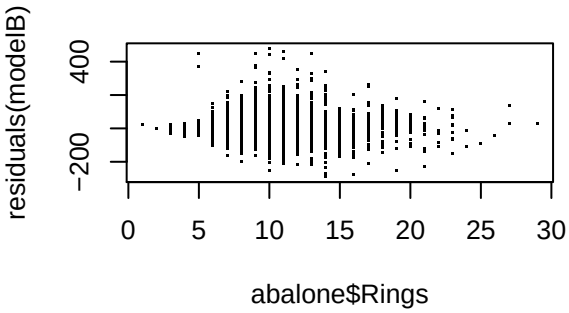
Now we want to look at the standard error by looking at the residual plots.

```
par(mfrow=c(2,2))
plot(abalone$Diameter,residuals(modelB),pch=".")
plot(abalone$Length,residuals(modelB),pch=".")
plot(abalone$Height,residuals(modelB),pch=".")
plot(abalone$Rings,residuals(modelB),pch=".")
```

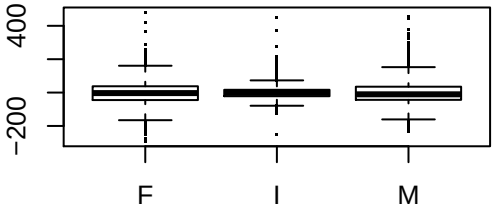









```
plot(abalone$Sex,residuals(modelB),pch=".")
```



We see that the standard error for Diameter and length are quite similar while the standard error for height is a little higher while the standard error for rings is the highest.

c

```
# Diameter
modD1 = gam(Shucked.weight ~ Diameter + s(Length,4) + s(Height,4) + s(Rings,4)+ Sex, data = abalone)
modD2 = gam(Shucked.weight ~ s(Length,4) + s(Height,4) + s(Rings,4) + Sex, data = abalone)
anova(modD2, modD1, modelB, test="F")
```

```
## Analysis of Deviance Table
##
## Model 1: Shucked.weight ~ s(Length, 4) + s(Height, 4) + s(Rings, 4) +
##     Sex
## Model 2: Shucked.weight ~ Diameter + s(Length, 4) + s(Height, 4) + s(Rings,
##     4) + Sex
## Model 3: Shucked.weight ~ s(Diameter, 4) + s(Length, 4) + s(Height, 4) +
##     s(Rings, 4) + Sex
##   Resid. Df Resid. Dev Df Deviance      F    Pr(>F)
## 1      4158   20506971
## 2      4157   20058680  1   448290 93.510 < 2.2e-16 ***
## 3      4154   19914324  3   144357 10.037 1.376e-06 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We see that for Diameter, the two reduced models are rejected because of the small p value. So we keep the original model for Diameter.

```
#Length
modL1 = gam(Shucked.weight ~ s(Diameter,4) + Length+ s(Height,4) + s(Rings,4) + Sex, data = abalone)
modL2 = gam(Shucked.weight ~ s(Diameter,4) + s(Height,4) + s(Rings,4) + Sex, data = abalone)
anova(modL2, modL1, modelB, test="F")
```

```
## Analysis of Deviance Table
##
## Model 1: Shucked.weight ~ s(Diameter, 4) + s(Height, 4) + s(Rings, 4) +
##     Sex
## Model 2: Shucked.weight ~ s(Diameter, 4) + Length + s(Height, 4) + s(Rings,
##     4) + Sex
## Model 3: Shucked.weight ~ s(Diameter, 4) + s(Length, 4) + s(Height, 4) +
##     s(Rings, 4) + Sex
##   Resid. Df Resid. Dev Df Deviance       F    Pr(>F)
## 1      4158   23336036
## 2      4157   20520511  1  2815525 587.301 < 2.2e-16 ***
## 3      4154   19914324  3   606187  42.149 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For Length, it's the same as Diameter, we should keep the original model.

```
# Height
modH1 = gam(Shucked.weight ~ s(Diameter,4) + s(Length,4)+ Height + s(Rings,4) + Sex, data = abalone)
modH2 = gam(Shucked.weight ~ s(Diameter,4) + s(Length,4)+ s(Rings,4) + Sex, data = abalone)
anova(modH2, modH1, modelB, test="F")
```

```
## Analysis of Deviance Table
##
## Model 1: Shucked.weight ~ s(Diameter, 4) + s(Length, 4) + s(Rings, 4) +
##     Sex
## Model 2: Shucked.weight ~ s(Diameter, 4) + s(Length, 4) + Height + s(Rings,
##     4) + Sex
## Model 3: Shucked.weight ~ s(Diameter, 4) + s(Length, 4) + s(Height, 4) +
##     s(Rings, 4) + Sex
##   Resid. Df Resid. Dev     Df Deviance       F Pr(>F)
## 1      4158   20847556
## 2      4157   19968986 1.0000   878570 183.264 <2e-16 ***
## 3      4154   19914324 2.9998    54663   3.801 0.0098 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Different for Height because the p value is not small anymore, so we might just use linear for height.

```
# Rings
modR1 = gam(Shucked.weight ~ s(Diameter,4) + s(Length,4) + s(Height,4) + Rings + Sex, data = abalone)
modR2 = gam(Shucked.weight ~ s(Diameter,4) + s(Length,4) + s(Height,4) + Sex, data = abalone)
anova(modR2, modR1, modelB, test="F")
```

```
## Analysis of Deviance Table
##
## Model 1: Shucked.weight ~ s(Diameter, 4) + s(Length, 4) + s(Height, 4) +
```

```
##     Sex
## Model 2: Shucked.weight ~ s(Diameter, 4) + s(Length, 4) + s(Height, 4) +
##     Rings + Sex
## Model 3: Shucked.weight ~ s(Diameter, 4) + s(Length, 4) + s(Height, 4) +
##     s(Rings, 4) + Sex
##   Resid. Df Resid. Dev     Df Deviance        F  Pr(>F)
## 1      4158   21879500
## 2      4157   19967091 1.0000  1912409 398.9163 < 2e-16 ***
## 3      4154   19914324 3.0001    52767   3.6689 0.01176 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For rings, we see that it's the same as Height.

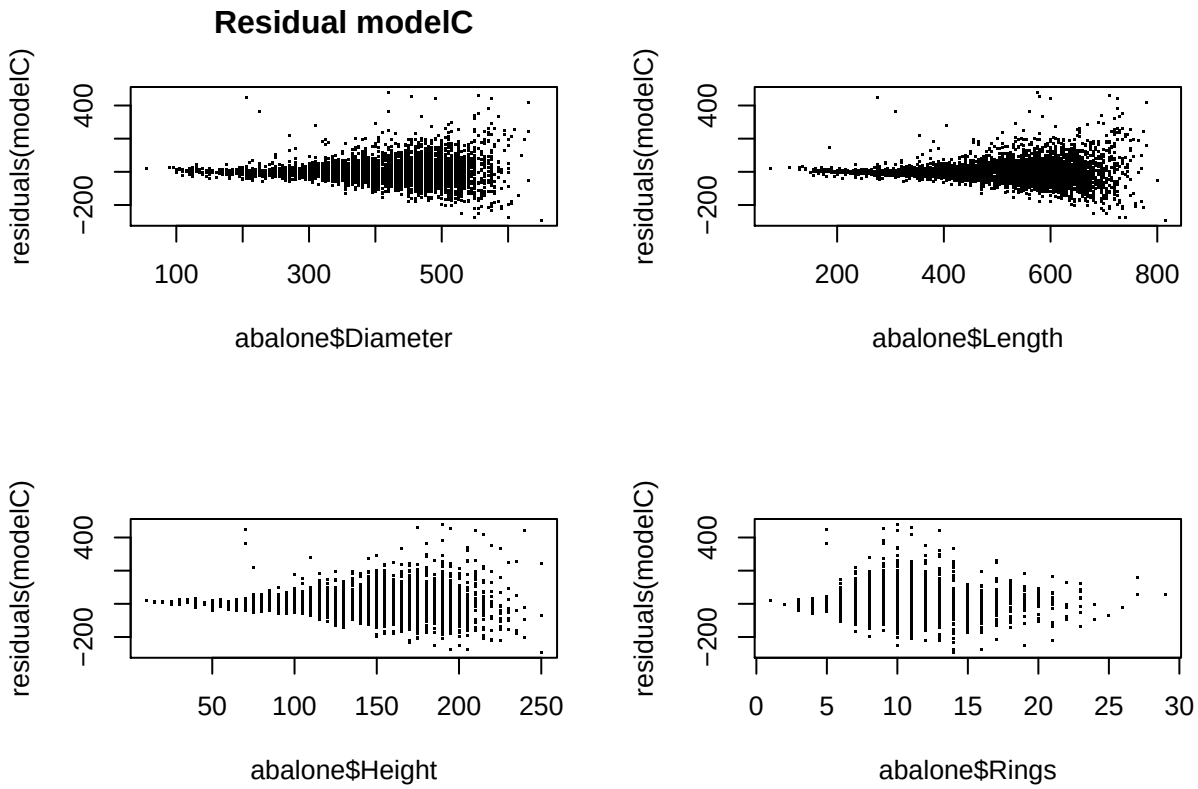
```
# Sex
modS = gam(Shucked.weight ~ s(Diameter,4)+ s(Length,4) + s(Height,4) + s(Rings,4), data = abalone)
anova(modS, modelB, test="F")
```

```
## Analysis of Deviance Table
##
## Model 1: Shucked.weight ~ s(Diameter, 4) + s(Length, 4) + s(Height, 4) +
##     s(Rings, 4)
## Model 2: Shucked.weight ~ s(Diameter, 4) + s(Length, 4) + s(Height, 4) +
##     s(Rings, 4) + Sex
##   Resid. Df Resid. Dev Df Deviance      F    Pr(>F)
## 1      4156   20270501
## 2      4154   19914324  2   356178 37.148 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

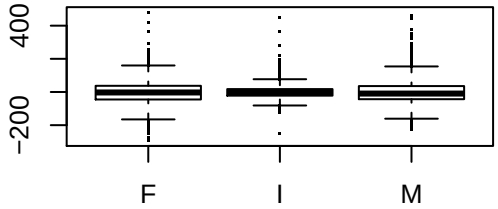
For Sex. Reject null, so we keep Sex.

d

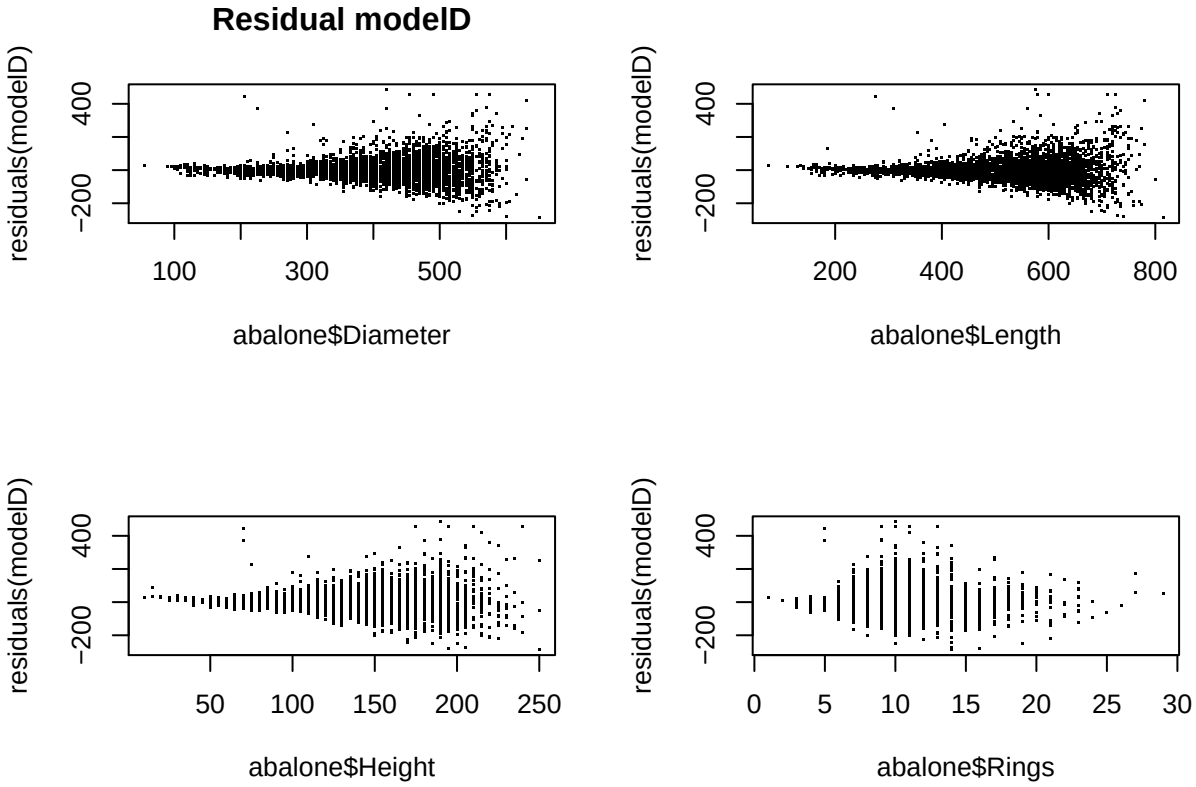
```
# Model C
modelC=gam(Shucked.weight~s(Diameter,4) + s(Length,4) + s(Height,4) + Rings + Sex, data=abalone)

par(mfrow=c(2,2))
plot(abalone$Diameter,residuals(modelC),pch=".")
title(main="Residual modelC")
plot(abalone$Length,residuals(modelC),pch=".")
plot(abalone$Height,residuals(modelC),pch=".")
plot(abalone$Rings,residuals(modelC),pch=".")
```



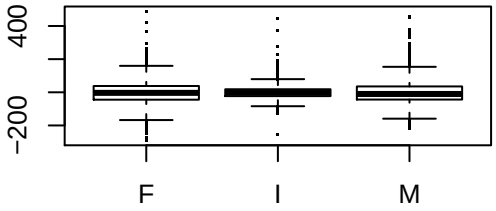

```
plot(abalone$Sex,residuals(modelC),pch=".")
```




```
# model D
modelD=gam(Shucked.weight~s(Diameter,4) + s(Length,4) + Height + Rings + Sex, data=abalone)

par(mfrow=c(2,2))
plot(abalone$Diameter,residuals(modelD),pch=".")
title(main="Residual modelD")
plot(abalone$Length,residuals(modelD),pch=".")
plot(abalone$Height,residuals(modelD),pch=".")
plot(abalone$Rings,residuals(modelD),pch=".")
```




```
plot(abalone$Sex,residuals(modelD),pch=".")
```



From the residual plots of modelC and D, we discover that the standard error is kind of similar for that of modelB. Next we do the 5-fold CV.

```
n = nrow(abalone)
K = 5
folds = rep(1:K, length = n)

set.seed(0)
Fivefolds = sample(folds,replace=F)
EA=NULL
EB=NULL
EC=NULL
ED=NULL
EE=NULL

for (k in 1:K) {
  train = abalone[Fivefolds != k, ]
  test = abalone[Fivefolds == k, ]

  modA=gam(Shucked.weight~Diameter + Length + Height + Rings + Sex, data=train)
  modB=gam(Shucked.weight~s(Diameter,4) + s(Length,4) + s(Height,4) + s(Rings,4) + Sex, data=train)
  modC=gam(Shucked.weight~s(Diameter,4) + s(Length,4) + s(Height,4) + Rings + Sex, data=train)
```

```
  modD=gam(Shucked.weight~s(Diameter,4) + s(Length,4) + Height + Rings + Sex, data=train)
  modE=gam(Shucked.weight~s(Diameter,4)+s(Length,4)+Sex+Height*Sex+Rings*Sex, data=train)

  EA[k] = mean((test$Shucked.weight - predict(modA, newdata = test))^2)
  EB[k] = mean((test$Shucked.weight - predict(modB, newdata = test))^2)
  EC[k] = mean((test$Shucked.weight - predict(modC, newdata = test))^2)
  ED[k] = mean((test$Shucked.weight - predict(modD, newdata = test))^2)
  EE[k] = mean((test$Shucked.weight - predict(modE, newdata = test))^2)
}
```

After doing the 5-fold CV, we want to look at the estimated error for all the four models.

```
mean(EA)
```

```
## [1] 8430.175
```

```
mean(EB)
```

```
## [1] 4823.362
```

```
mean(EC)
```

```
## [1] 4832.612
```

```
mean(ED)
```

```
## [1] 4826.763
```

Now calculate their respective SE.

```
sd(EA)/sqrt(5)
```

```
## [1] 472.1257
```

```
sd(EB)/sqrt(5)
```

```
## [1] 271.0511
```

```
sd(EC)/sqrt(5)
```

```
## [1] 270.1847
```

```
sd(ED)/sqrt(5)
```

```
## [1] 278.9036
```

We observe that the largest estimated error and largest standard errror appear at model1.

e

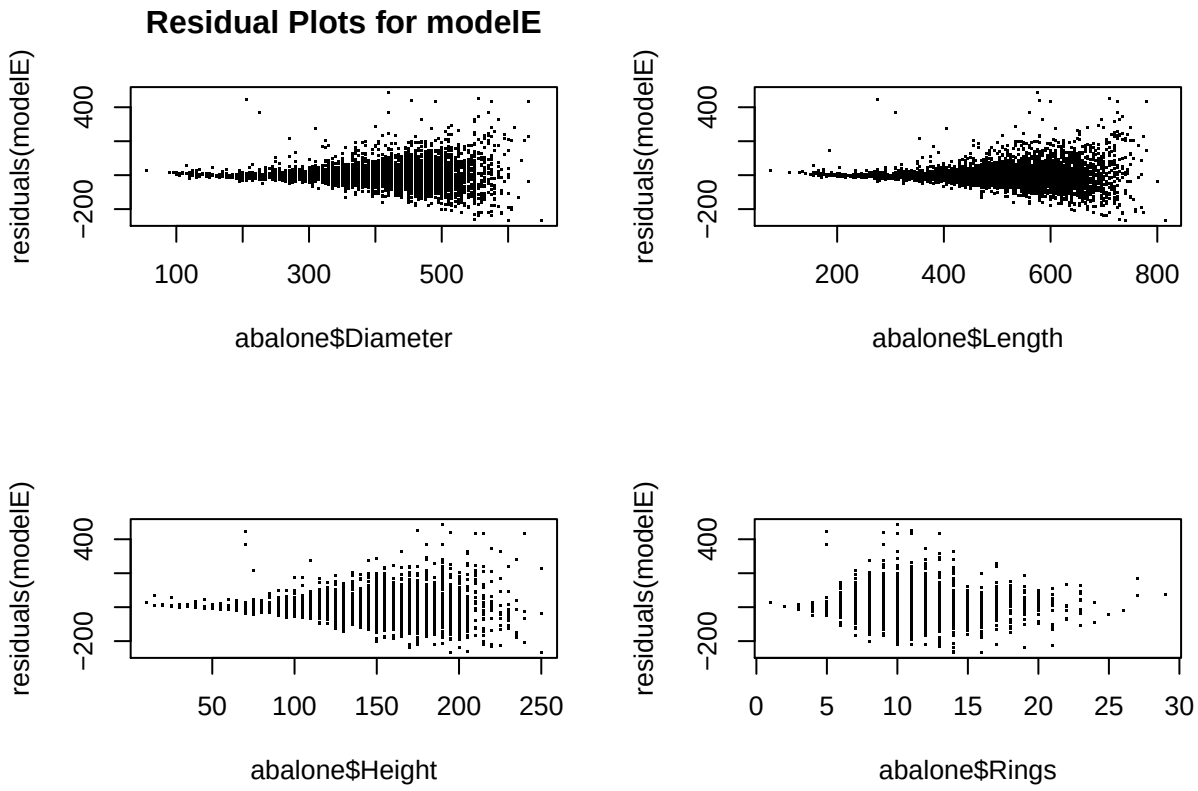
The two prediction errors are very close, so drop the non-linear effect of Rings can be justified.

f

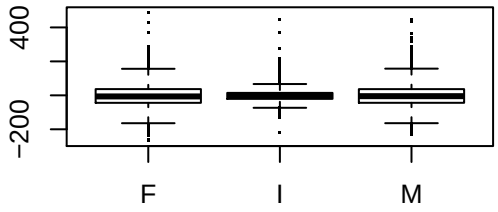
The two prediction errors are very close, and so are the SE. So drop the non-linear effect of Heights can be justified.

g

```
modelE=gam(Shucked.weight~s(Diameter,4) + s(Length,4) + Height*Sex + Rings*Sex, data=abalone)
par(mfrow=c(2,2))
plot(abalone$Diameter,residuals(modelE),pch=".")
title(main="Residual Plots for modelE")
plot(abalone$Length,residuals(modelE),pch=".")
plot(abalone$Height,residuals(modelE),pch=".")
plot(abalone$Rings,residuals(modelE),pch=".")
```




```
plot(abalone$Sex,residuals(modelE),pch=".")
```



Add the modelE in part D

```
mean(EE)
```

```
## [1] 4760.583
```

```
sd(EE)/sqrt(5)
```

```
## [1] 265.9818
```

We see that ModelE is the best.

h

ModelA has 7 (4+3) df, B has 19 (4*4+3) df, C has 16 df, D has 13 df, E has 17 (4+4+3+3+3) df. And to choose the best model, we need to consider the df-error trade-off.