Reihaneh Moghisi RMI 8300
Assignment 6 Apr 17 2023

Question 1

This question relates to the College data set.

(A) Split the data into a training set and a test set. Using out-of-state tuition as the response and the other variables as the predictors, perform forward step wise selection on the training set in order to identify a satisfactory model that uses just a subset of the predictors.

```
data(College)
set.seed(234)
College.split = initial_split(College, strata = "Outstate", prop = .7)
College.train = training(College.split)
College.test = testing (College.split)

ctrl <- trainControl(method = "repeatedcv",
                     number = 10,
                     repeats = 1,
                     selectionFunction = "oneSE")

model.fwd <- train(Outstate ~ .,
                       data = College.train,
                       method = "leapForward",
                       metric = "MSE",
                       maximize = F,
                       trControl = ctrl,
                       tuneGrid = data.frame(nvmax = 1:17))
```

```
## Warning in train.default(x, y, weights = w, ...): The metric "MSE" was not
in
## the result set. RMSE will be used instead.
```

```
model.fwd
```

```
## Linear Regression with Forward Selection
##
## 542 samples
##  17 predictor
##
## No pre-processing
```

```
## Resampling: Cross-Validated (10 fold, repeated 1 times)
## Summary of sample sizes: 487, 487, 488, 489, 489, 488, ...
## Resampling results across tuning parameters:
##
##   nvmax  RMSE      Rsquared   MAE
##    1     2821.222  0.5421468  2263.313
##    2     2432.681  0.6511000  1868.439
##    3     2226.759  0.7038217  1723.816
##    4     2081.243  0.7405207  1643.016
##    5     2025.848  0.7518705  1595.431
##    6     2015.033  0.7547385  1596.790
##    7     1993.910  0.7601704  1583.245
##    8     2009.634  0.7558769  1598.434
##    9     2029.106  0.7502363  1601.002
##   10     2034.763  0.7506942  1602.750
##   11     2037.931  0.7506845  1602.482
##   12     1995.502  0.7625791  1577.820
##   13     1997.814  0.7624667  1580.192
##   14     2007.636  0.7598428  1578.395
##   15     2007.645  0.7596646  1577.778
##   16     2001.546  0.7608492  1574.528
##   17     1996.867  0.7618954  1570.084
##
## RMSE was used to select the optimal model using  the one SE rule.
## The final value used for the model was nvmax = 5.
```

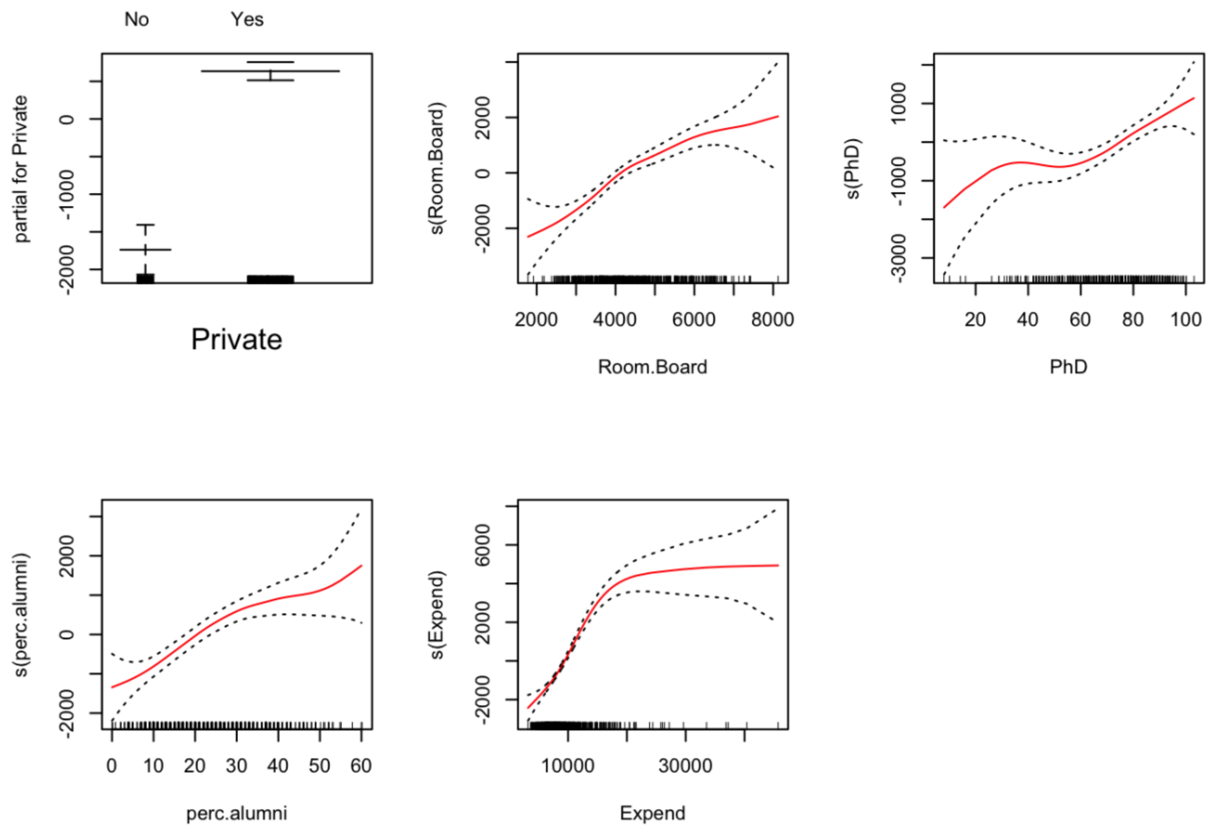
The forward step wise selection process identified a 5 variable solution to be a satisfactory model.

(B) Fit a GAM on the training data, using out-of-state tuition as the response and the features selected in the previous step as the predictors. Plot the results, and explain your findings.

The variables in the selected solution are:

```
coef(model.fwd$finalModel, id = 5)
##   (Intercept)    PrivateYes    Room.Board           PhD   perc.alumni
## -2567.2683333  2642.6569135     0.9689926    37.4598121    64.1658168
##        Expend
```

```
## 0.2808643
model.gam <- gam(Outstate ~ Private + s(Room.Board) + s(PhD) + s(perc.alumni)
+ s(Expend), data = College.train)


par(mfrow = c(2, 3))

plot(model.gam, se = T, col = "red")
```



The categorical variable "Private" indicates whether a university is privately or publicly funded. This model predominantly utilizes data from the public sector. The cost of room and board (Room.Board) appears to have a linear relationship, which is not surprising since out-of-state students require accommodation and meals, whether provided by the university or private sector. As out-of-state tuition increases, so does the proportion of faculty members with Ph.D.'s (PhD). The model implies that institutions with higher costs tend to have a larger percentage of alumni who donate (perc_alumni) to the school. The expenditure on instruction per student (Expend) reaches a plateau when out-of-state tuition is around $16,000, indicating that the university's decision on instructional expenses is not driven by out-of-state tuition.

(C) Evaluate the model obtained on the test set, and explain the results obtained.

The GAM MSE and R^2 for College.test is:

```
gam.pred=predict(model.fwd, College.test)
(gam.mse=mean((College.test$Outstate-gam.pred)^2))
## [1] 5023810
gam.tss = sum((College.test$Outstate- mean(College.test$Outstate))^2)
gam.rss = sum((gam.pred -College.test$Outstate)^2)
(1-gam.rss/gam.tss)
## [1] 0.6930349
```

The GAM model can explain 69.3% of the variance in the data.

(D) For which variables, if any, is there evidence of a non-linear relationship with the response?

```
summary(model.gam)
##
## Call: gam(formula = Outstate ~ Private + s(Room.Board) + s(PhD) + s(perc.a
lumni) +
##     s(Expend), data = College.train)
## Deviance Residuals:
##     Min      1Q  Median      3Q     Max
## -5440.1 -1136.9   103.7  1202.5  7010.9
##
## (Dispersion Parameter for gaussian family taken to be 3448194)
##
##     Null Deviance: 8712201403 on 541 degrees of freedom
## Residual Deviance: 1806853709 on 524 degrees of freedom
## AIC: 9716.746
##
## Number of Local Scoring Iterations: NA
##
## Anova for Parametric Effects
##                  Df     Sum Sq    Mean Sq F value    Pr(>F)
## Private           1 2257060436 2257060436  654.56 < 2.2e-16 ***
## s(Room.Board)     1 1880675047 1880675047  545.41 < 2.2e-16 ***
## s(PhD)            1  672093668  672093668  194.91 < 2.2e-16 ***
## s(perc.alumni)    1  409119952  409119952  118.65 < 2.2e-16 ***
## s(Expend)         1  757963449  757963449  219.81 < 2.2e-16 ***
```

```
## Residuals      524 1806853709    3448194
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Anova for Nonparametric Effects
##                Npar Df  Npar F    Pr(F)
## (Intercept)
## Private
## s(Room.Board)        3  2.7045  0.04481 *
## s(PhD)               3  2.3431  0.07225 .
## s(perc.alumni)       3  1.7452  0.15675
## s(Expend)            3 26.9990 2.22e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The Anova for Nonparametric Effects tests show that the variable Expend has a non-linear relationship and is significant at a significance value of $p<0.05$. Conversely, the Anova for Parametric Effects tests reveal that all variables are significant and have a non-linear relationship at the same significance value.

Question 2

The `Wage` Dataset (Various Non-Linear Methods)

Q: *The `Wage` data set contains a number of other features not explored in this chapter, such as marital status (`maritl`), job class (`jobclass`), and others. Explore the relationships between some of these other predictors and wage, and use non-linear fitting techniques in order to fit flexible models to the data. Create plots of the results obtained, and write a summary of your findings.*

We first look into different variables.

Marital status:

```
set.seed(1)
summary(Wage$maritl)
```

```
1. Never Married       2. Married        3. Widowed       4. Divorced      5. Se
parated
```

```
              648               2074                 19                204
55
```

Jobclass:

```
# table(Wage$maritl) the same with `summary`
summary(Wage$jobclass)
 1. Industrial 2. Information
         1544           1456
```

Wage:

```
par(mfrow = c(1, 2))
plot(Wage$maritl, Wage$wage)
plot(Wage$jobclass, Wage$wage)
```

`year`:

As the only other numeric variable in the dataset, it is limited to seven distinct values (2003 - 2009), which makes it unsuitable to employ splines to depict the relationship.

```
table(Wage$year)
##
## 2003 2004 2005 2006 2007 2008 2009
##  513  485  447  392  386  388  389
```

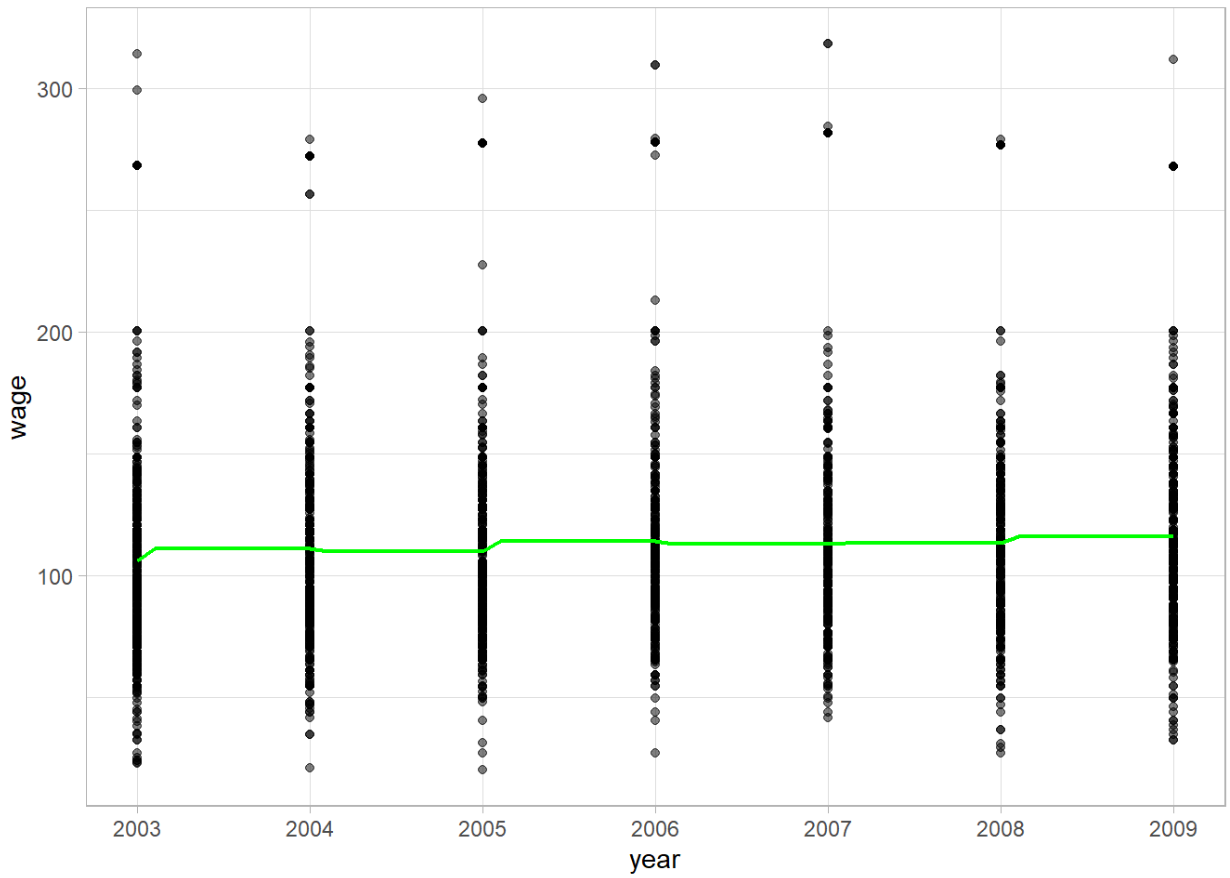
Because of the discrete nature of `year` I fit a step function with the following intervals:

```
table(cut(Wage$year, breaks = 2002:2009))
##
## (2002,2003] (2003,2004] (2004,2005] (2005,2006] (2006,2007] (2007,2008]
##         513         485         447         392         386         388
## (2008,2009]
##         389
```

I plot the relationship between `year` and `wage` (with the step function) below. It looks pretty uninformative:

```
year_step <- lm(wage ~ cut(year, breaks = 2002:2009), Wage)
fitted <- data.frame(year = seq(2003, 2009, 0.1),
```

```
                    wage = predict(year_step, data.frame(year = seq(2003, 20
09, 0.1))))



ggplot(Wage, aes(x = year, y = wage)) +
  geom_point(alpha = 0.5) +
  geom_line(data = fitted,
            aes(x = year, y = wage), size = 0.8, col = "green") +
  scale_x_continuous(breaks = 2003:2009, minor_breaks = NULL)
```



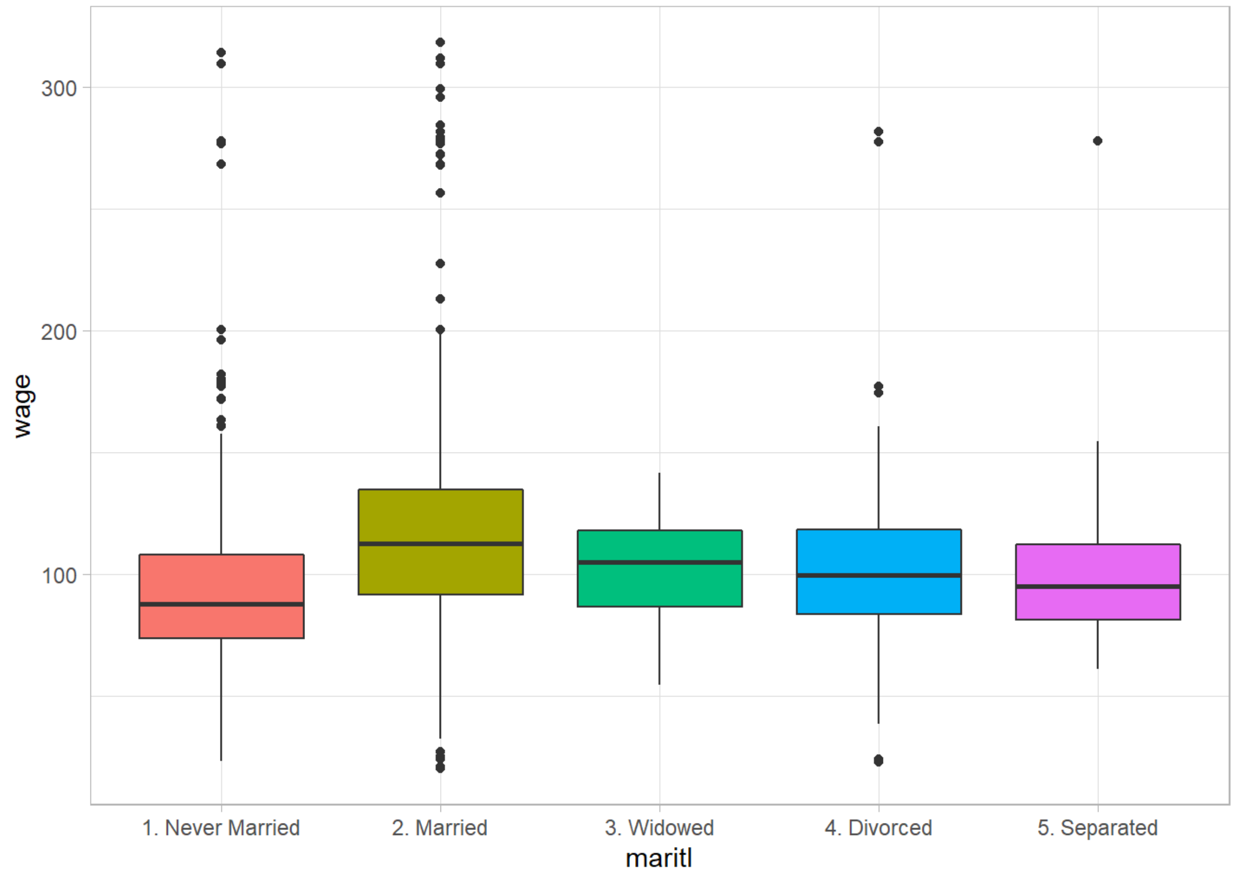
maritl:

Note that the volume for `3. Widowed` in particular is pretty low:

```
##
## 1. Never Married       2. Married       3. Widowed       4. Divorced
```

```
##            648          2074           19          204
##      5. Separated
##             55
```

It appears that married workers (for those in this dataset - male workers in the Mid-Atlantic region) earn more.



Since the number of cases is relatively low, it would be reasonable to merge the "3. Widowed" category with another category, such as "4. Divorced," to form a combined category called "6. Previously Married" before proceeding with modeling.

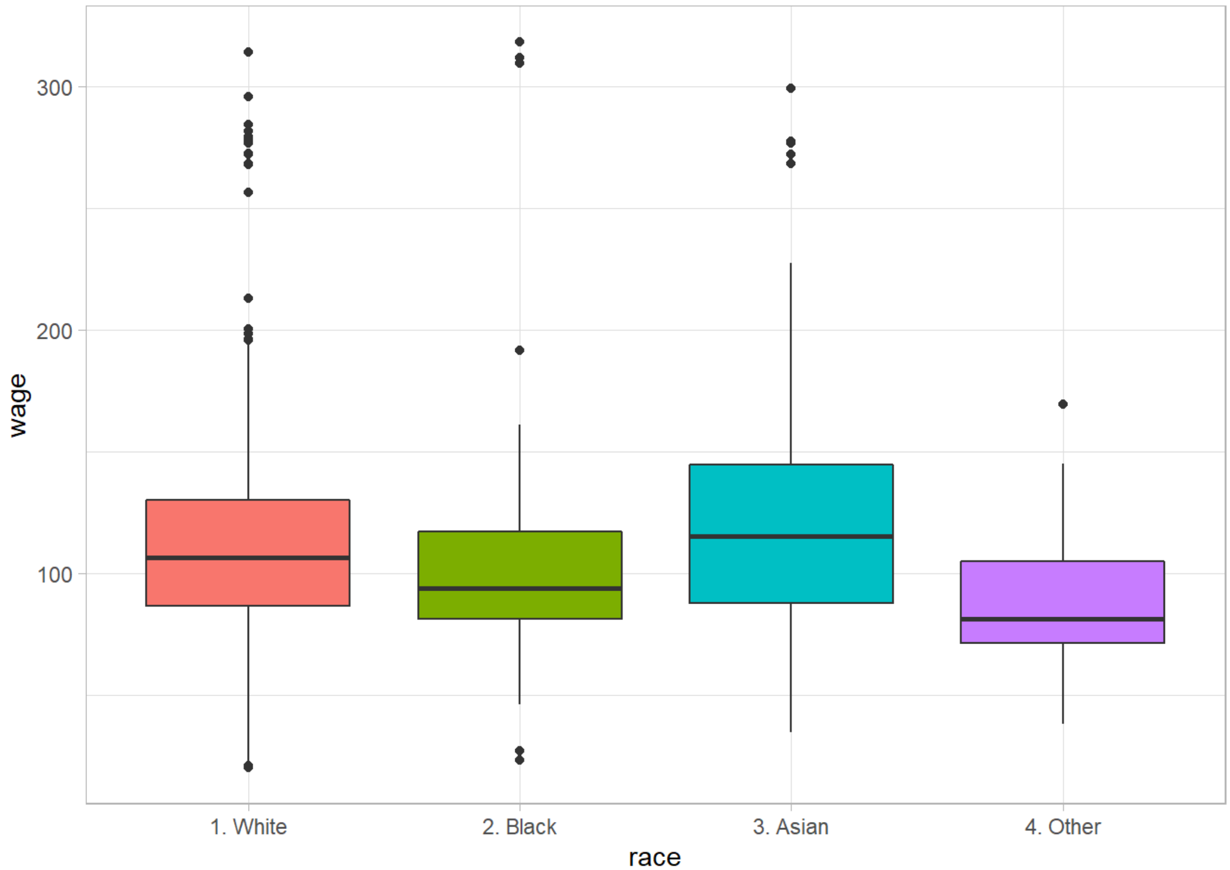
race:

Note the low volumes again for the `4. Other` category:

```
##
## 1. White 2. Black 3. Asian 4. Other
##     2480      293      190       37
```

According to this dataset, Asian men earn the highest, followed by white and black men. To prepare for modeling, it might be worth considering merging the "4. Other" category with either

"2. Black" (the category closest to it in terms of the response) or "3. Asian" (the second-smallest category).

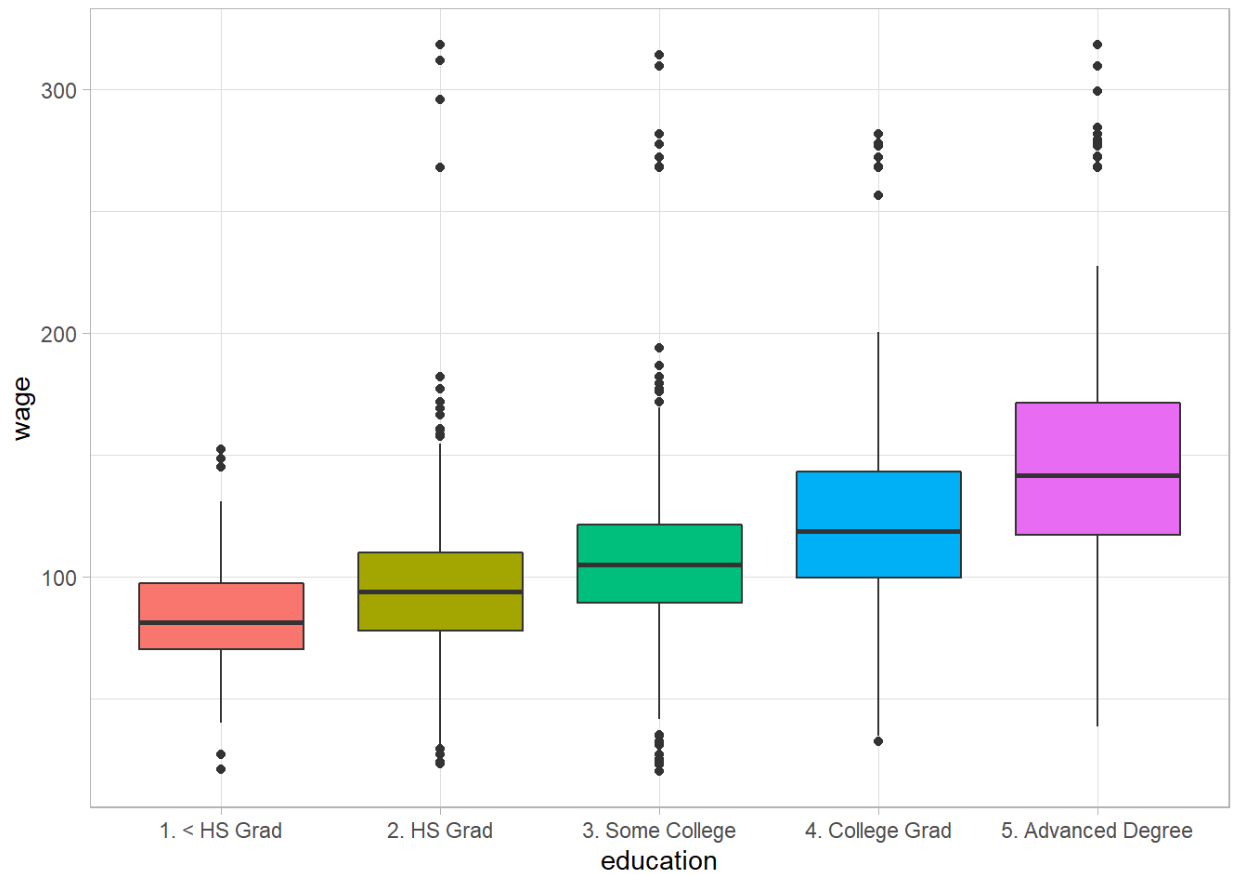


education:

Again, very balanced:

```
##
##        1. < HS Grad          2. HS Grad    3. Some College    4. College Gra
d
##                 268                 971                650                68
5
## 5. Advanced Degree
##                 426
```

This is an ordinal categorical variable, and we can see a clear positive relationship between education-level and wage.



`region:`

It is not surprising that this variable's distribution is as follows, considering the dataset only represents a sample of male workers from the Mid-Atlantic region.

```
##
##         1. New England    2. Middle Atlantic 3. East North Central
##                      0                  3000                     0
## 4. West North Central      5. South Atlantic 6. East South Central
##                      0                     0                     0
## 7. West South Central            8. Mountain            9. Pacific
##                      0                     0                     0
```

The variable has no variance and so obviously won't be used in modelling.

`jobclass:`

This variable is much more balanced:

```
##
##   1. Industrial 2. Information
##           1544           1456
```

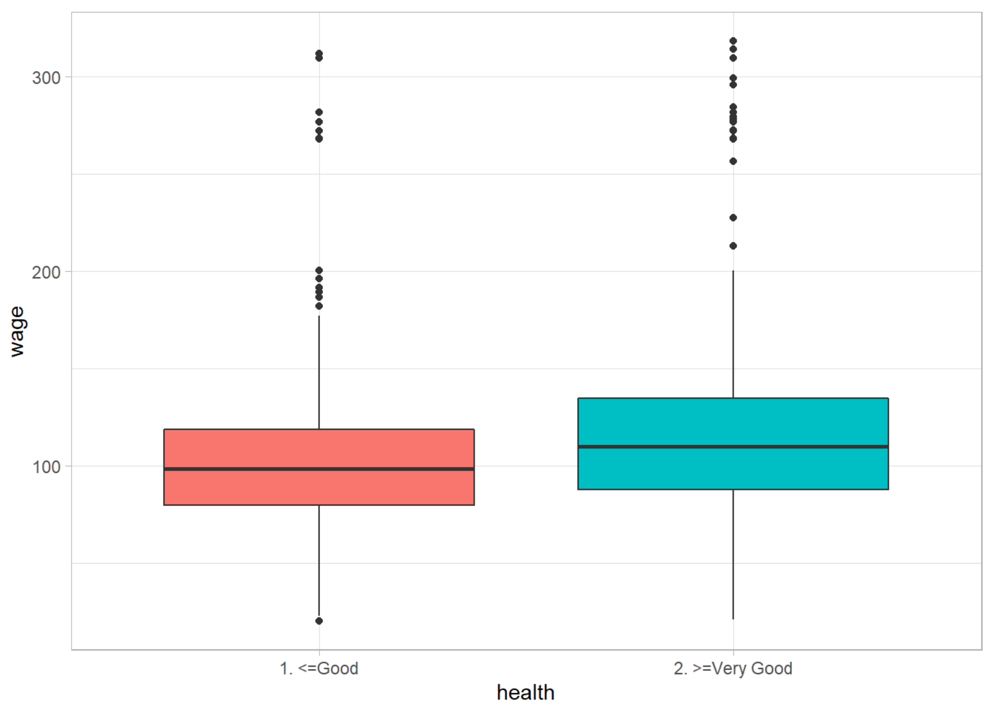
We see that those with `Information` as their type of job earn more than those with `Industrial` jobs.

health:

This variable is also very balanced, with most workers falling in the `2. Very Good` category for `health`:

```
##
##       1. <=Good 2. >=Very Good
##             858           2142
```

We see that those with better health tend to earn more.

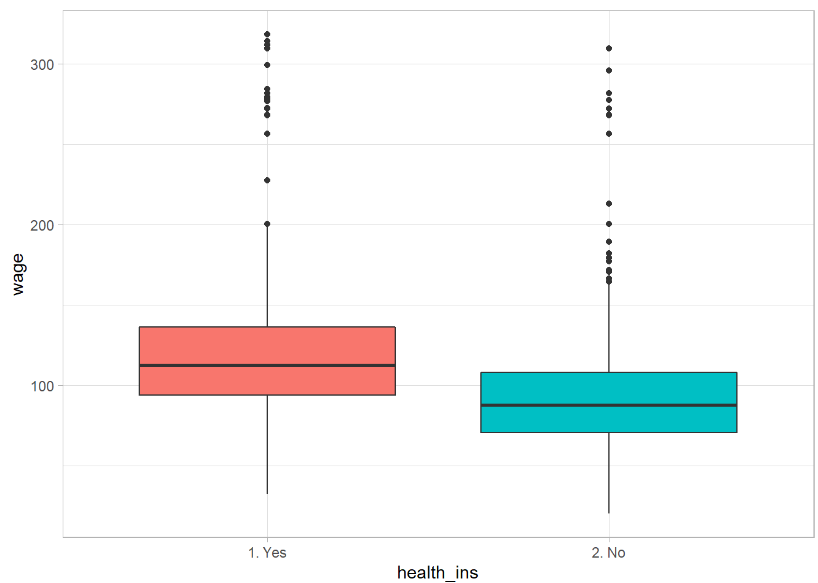


health_ins:

Most workers don't have health insurance:

```
##
## 1. Yes  2. No
##    2083    917
```

Workers with health insurance are, on average, higher paid:



Fitting the model:

We try to Fit wage on multiple predictors with GAM using year: and from there we see that model fit4 fits the data best. Below is the results and plots of prediction accordingly.

```
library(gam)
fit1 <- gam(wage ~ lo(year, span = 0.7) + s(age, 5) + education, data = Wage)
fit2 <- gam(wage ~ lo(year, span = 0.7) + s(age, 5) + education + jobclass, d
ata = Wage)
fit3 <- gam(wage ~ lo(year, span = 0.7) + s(age, 5) + education + maritl, dat
a = Wage)
fit4 <- gam(wage ~ lo(year, span = 0.7) + s(age, 5) + education + jobclass +
maritl, data = Wage)
anova(fit1, fit2, fit3, fit4)
```

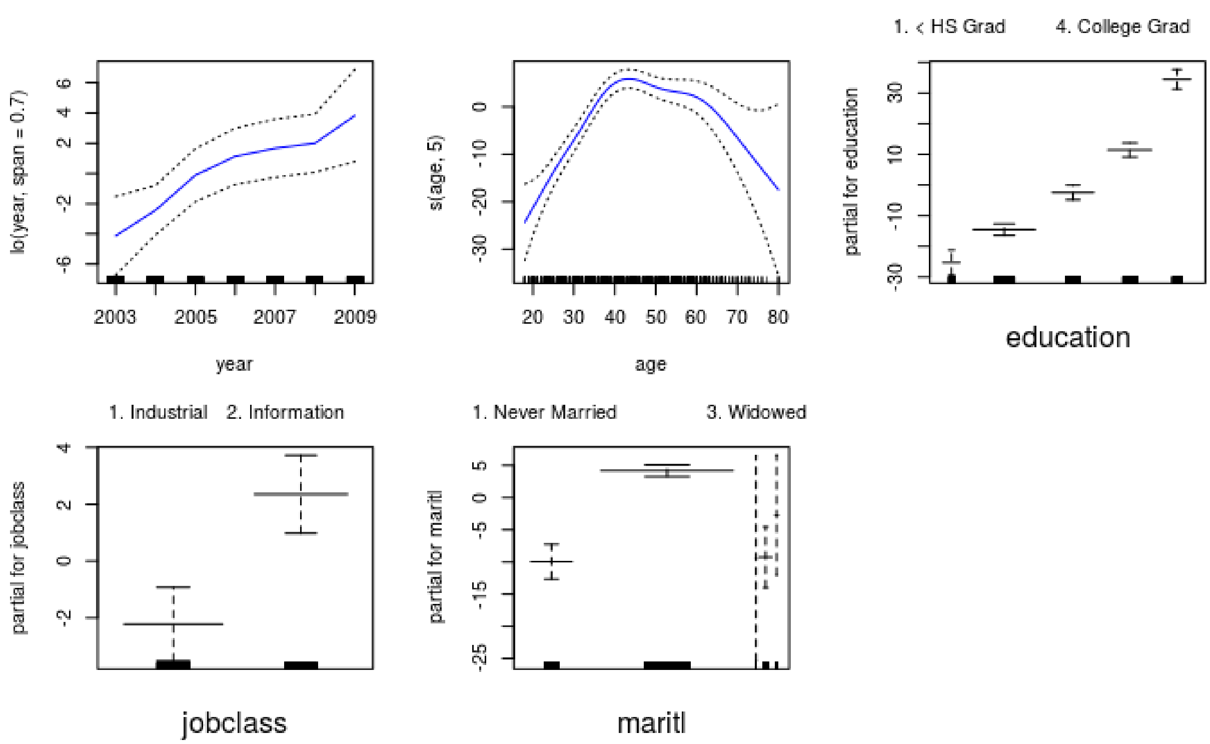
```
Analysis of Deviance Table


Model 1: wage ~ lo(year, span = 0.7) + s(age, 5) + education
```

```
Model 2: wage ~ lo(year, span = 0.7) + s(age, 5) + education + jobclass
Model 3: wage ~ lo(year, span = 0.7) + s(age, 5) + education + maritl
Model 4: wage ~ lo(year, span = 0.7) + s(age, 5) + education + jobclass +
    maritl
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1    2987.1    3691855
2    2986.1    3679689  1    12166 0.0014637 **
3    2983.1    3597526  3    82163  9.53e-15 ***
4    2982.1    3583675  1    13852 0.0006862 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

>>> model *fit4* fits the best.

Plot the model:

```
par(mfrow = c(2, 3))
plot(fit4, se = T, col = "blue")
```



I also combined some categories and employed flexible basis functions such as cubic and step-function for age and year to build a linear regression model. The results are seen below.

```
##
## Call:
## lm(formula = wage ~ poly(age, 3, raw = T) + cut(year, breaks = 2002:2009)
+
##     maritl + race + education + jobclass + health + health_ins,
##     data = Wage_cleaned)
##
## Residuals:
##     Min      1Q Median      3Q     Max
## -101.4  -18.6   -3.3   13.8  209.2
##
## Coefficients:
##                                              Estimate Std. Error t value Pr(
>|t|)
## (Intercept)                                 3.187e+00  1.972e+01   0.162 0.8
71596
## poly(age, 3, raw = T)1                      3.592e+00  1.409e+00   2.550 0.0
10832
## poly(age, 3, raw = T)2                     -4.883e-02  3.190e-02  -1.530 0.1
26009
## poly(age, 3, raw = T)3                      1.634e-04  2.321e-04   0.704 0.4
81443
## cut(year, breaks = 2002:2009)(2003,2004]  2.603e+00  2.145e+00   1.214 0.2
24950
## cut(year, breaks = 2002:2009)(2004,2005]  3.122e+00  2.190e+00   1.426 0.1
54054
## cut(year, breaks = 2002:2009)(2005,2006]  7.629e+00  2.269e+00   3.362 0.0
00782
## cut(year, breaks = 2002:2009)(2006,2007]  5.201e+00  2.281e+00   2.280 0.0
22658
## cut(year, breaks = 2002:2009)(2007,2008]  6.408e+00  2.274e+00   2.818 0.0
04866
## cut(year, breaks = 2002:2009)(2008,2009]  8.393e+00  2.275e+00   3.689 0.0
00229
## maritl2. Married                            1.337e+01  1.811e+00   7.385 1.9
8e-13
## maritl5. Separated                          7.160e+00  4.857e+00   1.474 0.1
40521
```

```
## marit16. Previously Married                1.629e-01  2.845e+00   0.057 0.9
54348
## race3. Asian                              -2.683e+00  2.587e+00  -1.037 0.2
99772
## race5. Black & Other                      -4.856e+00  2.022e+00  -2.402 0.0
16374
## education2. HS Grad                        7.564e+00  2.351e+00   3.218 0.0
01306
## education3. Some College                   1.799e+01  2.499e+00   7.200 7.5
9e-13
## education4. College Grad                   3.064e+01  2.531e+00  12.107  <
2e-16
## education5. Advanced Degree                5.315e+01  2.793e+00  19.031  <
2e-16
## jobclass2. Information                     3.455e+00  1.317e+00   2.623 0.0
08750
## health2. >=Very Good                       6.310e+00  1.412e+00   4.469 8.1
4e-06
## health_ins2. No                           -1.634e+01  1.404e+00 -11.635  <
2e-16
##
## (Intercept)
## poly(age, 3, raw = T)1                    *
## poly(age, 3, raw = T)2
## poly(age, 3, raw = T)3
## cut(year, breaks = 2002:2009)(2003,2004]
## cut(year, breaks = 2002:2009)(2004,2005]
## cut(year, breaks = 2002:2009)(2005,2006] ***
## cut(year, breaks = 2002:2009)(2006,2007] *
## cut(year, breaks = 2002:2009)(2007,2008] **
## cut(year, breaks = 2002:2009)(2008,2009] ***
## marit12. Married                          ***
## marit15. Separated
## marit16. Previously Married
## race3. Asian
## race5. Black & Other                      *
## education2. HS Grad                       **
```

```
## education3. Some College                  ***
## education4. College Grad                  ***
## education5. Advanced Degree               ***
## jobclass2. Information                    **
## health2. >=Very Good                      ***
## health_ins2. No                           ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.75 on 2978 degrees of freedom
## Multiple R-squared:  0.3503, Adjusted R-squared:  0.3457
## F-statistic: 76.44 on 21 and 2978 DF,  p-value: < 2.2e-16
```