Hu_Anqi_HW4

Anqi Hu

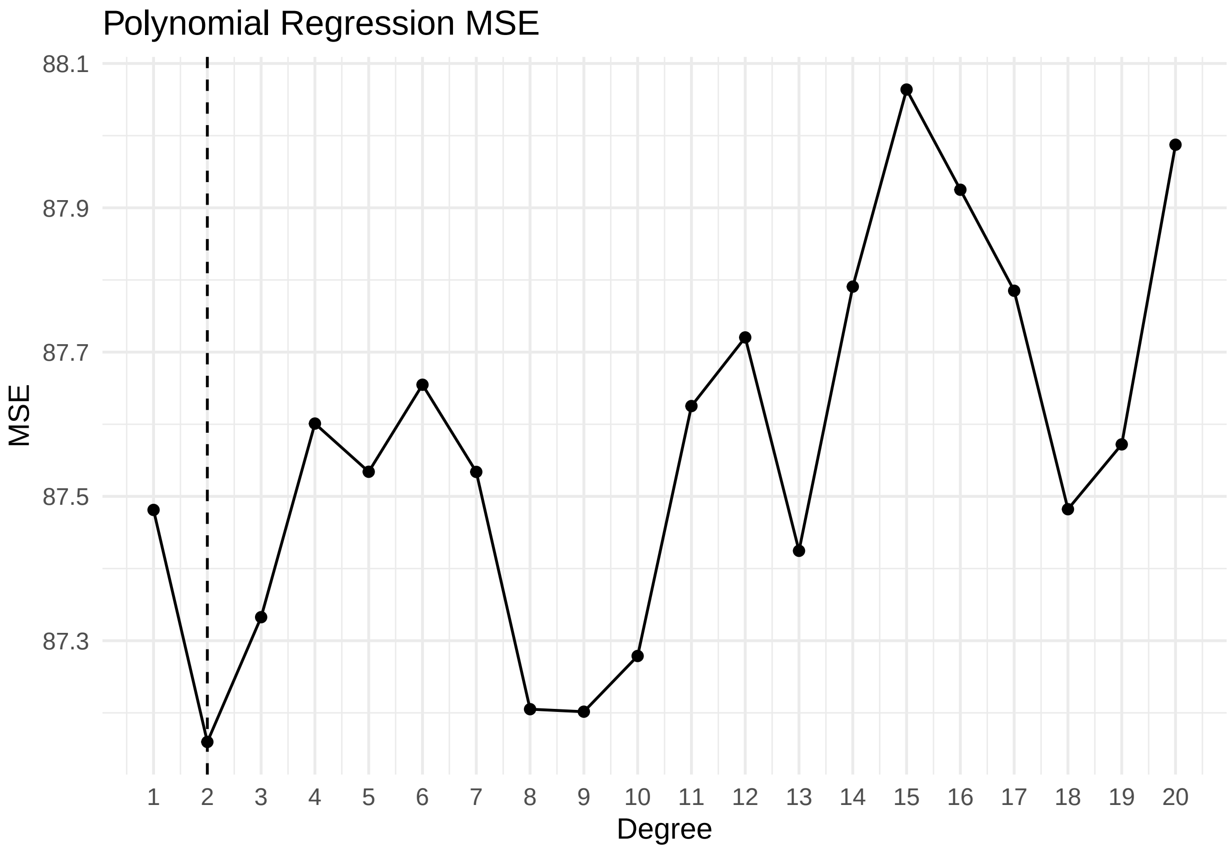
2/16/2020

Egalitarian and Income

1

```
set.seed(123)
theme_set(theme_minimal())
gss_train <- read_csv('data/gss_train.csv')
gss_test <- read_csv('data/gss_test.csv')
```

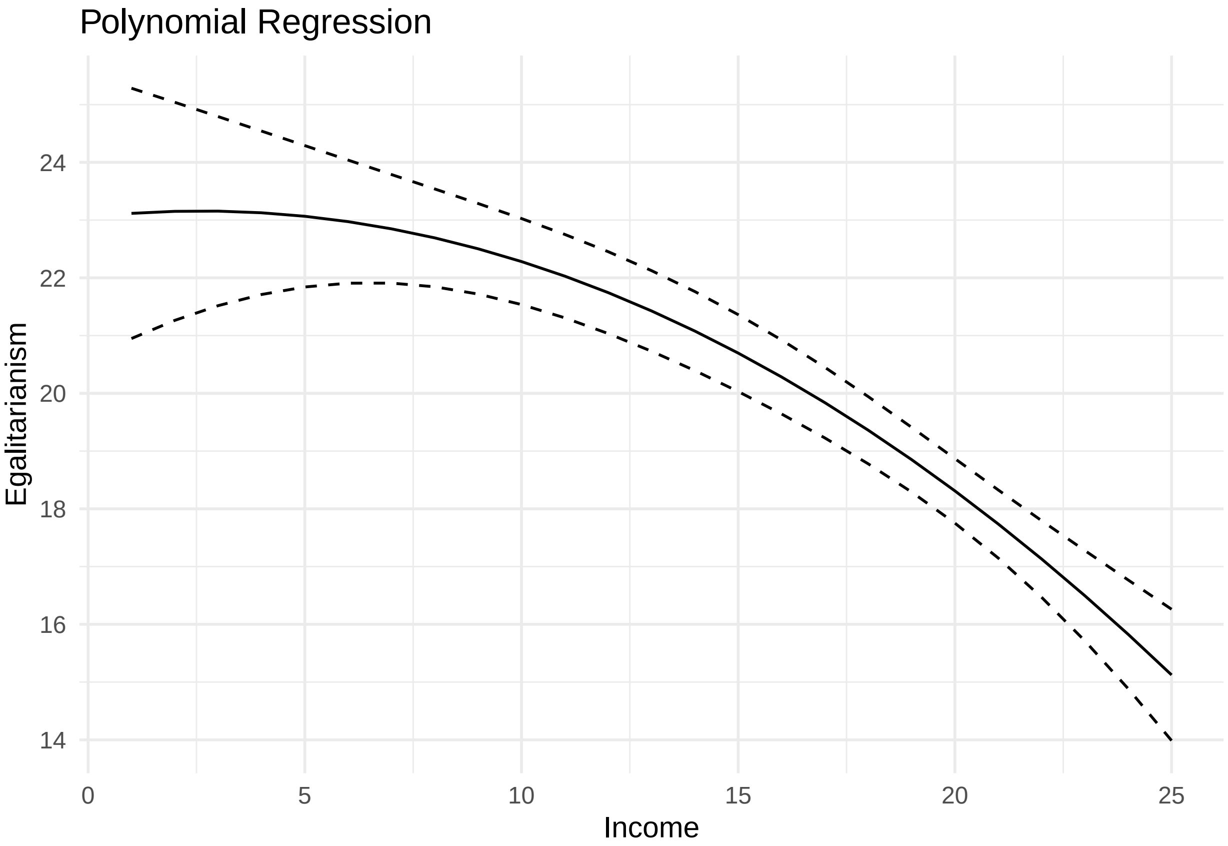
```
# CV to select the best d
errors = data.frame(degree = (1:20), mse=0)

for (i in 1:20) {
  poly <- glm(egalit_scale ~ poly(income06, degree = i, raw = TRUE), data = gss_train)
  errors[i, 2] = cv.glm(gss_train, poly, K = 10)$delta[1]
}

ggplot(errors) +
  geom_point(aes(x = degree, y = mse)) +
  geom_line(aes(x = degree, y = mse)) +
  scale_x_continuous(breaks=1:20) +
  geom_vline(xintercept = which.min(errors$mse), linetype = "dashed") +
  labs(title = "Polynomial Regression MSE",
       x = "Degree",
       y = "MSE")
```



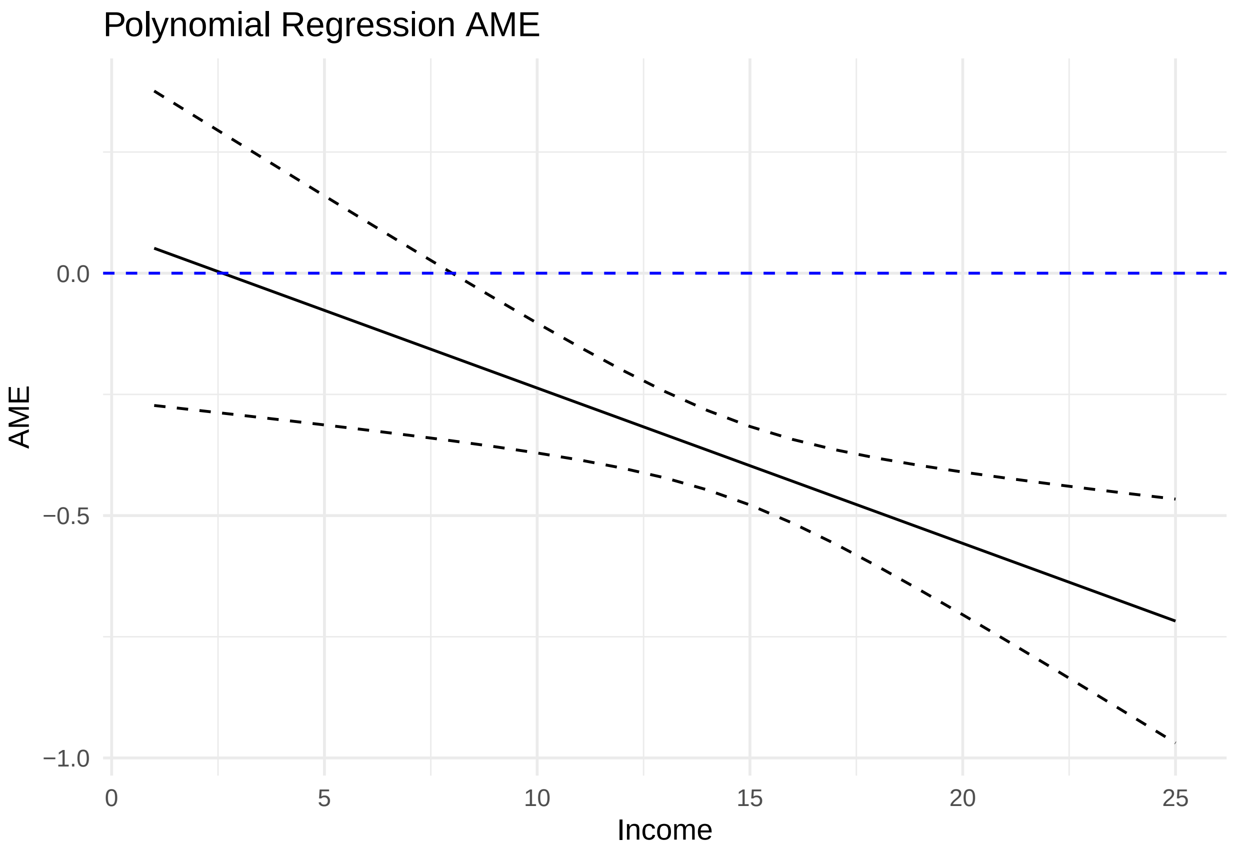
```
poly_reg <- lm(egalit_scale ~ income06 + I(income06^2), data = gss_train)
cp1 <- cplot(poly_reg, "income06", what = "prediction", draw = FALSE)
```

```
##    xvals    yvals    upper    lower
## 1      1 23.11631 25.28371 20.94890
## 2      2 23.15180 25.04001 21.26359
## 3      3 23.15524 24.79232 21.51817
## 4      4 23.12665 24.54195 21.71134
## 5      5 23.06600 24.29036 21.84165
## 6      6 22.97331 24.03899 21.90764
## 7      7 22.84858 23.78870 21.90846
## 8      8 22.69180 23.53894 21.84467
## 9      9 22.50298 23.28678 21.71917
## 10    10 22.28211 23.02670 21.53752
## 11    11 22.02920 22.75132 21.30708
## 12    12 21.74424 22.45297 21.03552
## 13    13 21.42724 22.12502 20.72946
## 14    14 21.07819 21.76262 20.39376
## 15    15 20.69710 21.36290 20.03130
## 16    16 20.28396 20.92501 19.64291
## 17    17 19.83878 20.45044 19.22712
## 18    18 19.36155 19.94356 18.77955
## 19    19 18.85228 19.41247 18.29210
## 20    20 18.31096 18.86919 17.75274
```

```
ggplot(cp1) +
  geom_line(aes(x = xvals, y = yvals)) +
  geom_line(aes(x = xvals, y = upper), linetype = "dashed") +
  geom_line(aes(x = xvals, y = lower), linetype = "dashed") +
  labs(title = "Polynomial Regression",
       x = "Income",
       y = "Egalitarianism")
```



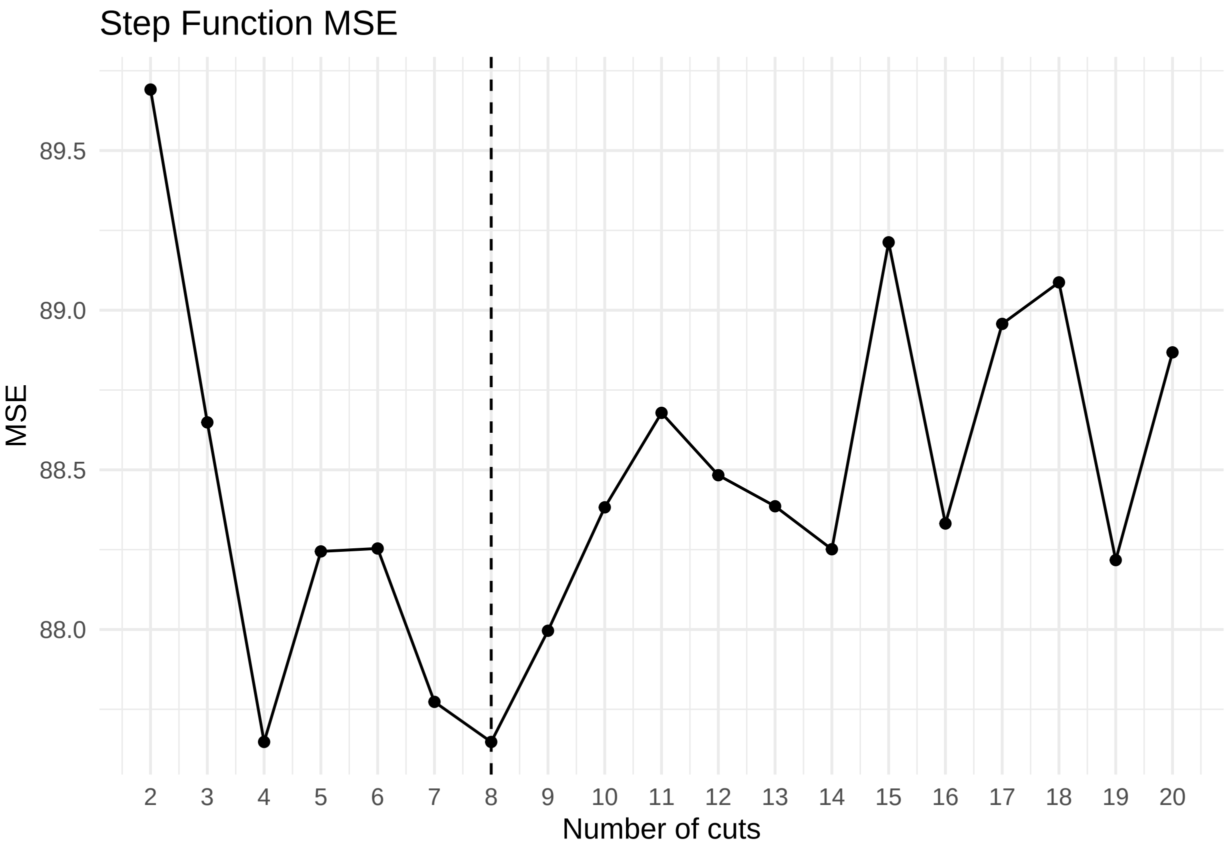
```
cp2 <- cplot(poly_reg, "income06", what = "effect", draw = FALSE)

ggplot(cp2) +
  geom_line(aes(x = xvals, y = yvals)) +
  geom_line(aes(x = xvals, y = upper), linetype = "dashed") +
  geom_line(aes(x = xvals, y = lower), linetype = "dashed") +
  geom_hline(yintercept = 0, linetype = "dashed", color = 'blue') +
  labs(title = "Polynomial Regression AME",
       x = "Income",
       y = "AME")
```



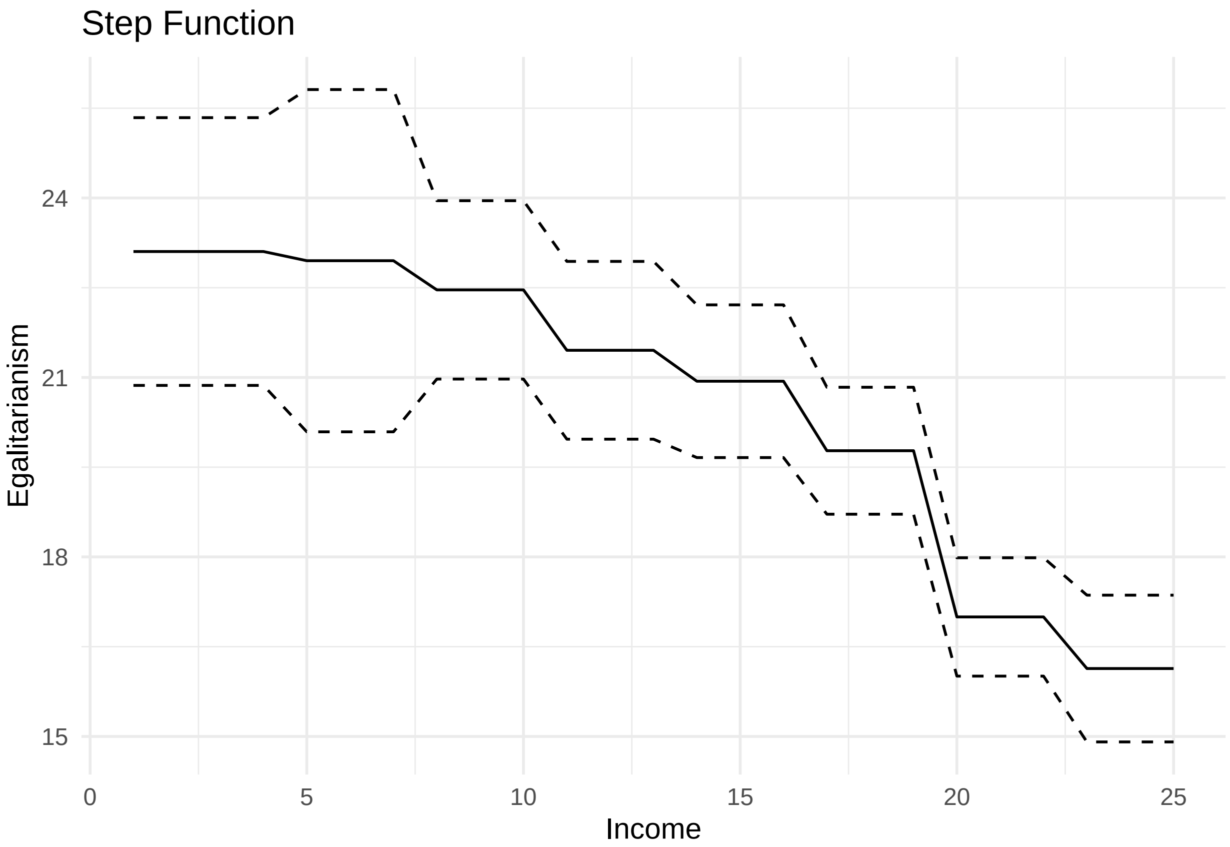
In terms of MSE, the value reaches minimum when d=2. As income increases, the average marginal effect decreases. However, the 95% confidence interval shrinks and then expands as income increases.

2

```r
errors = rep(0, 19)

for (i in 2:20) {
  gss_train$cuts <- cut_interval(gss_train$income06, i)

  step <- glm(egalit_scale ~ cuts, data = gss_train)
  errors[i-1] <- cv.glm(gss_train, step, K = 10)$delta[1]
}

tibble(n = 2:20, err = errors) %>%
  ggplot(aes(x = n, y = err)) +
  geom_line() +
  geom_point() +
  scale_x_continuous(breaks=1:20) +
  geom_vline(xintercept = which.min(errors) + 1, linetype = "dashed") +
  labs(title = "Step Function MSE",
       x = "Number of cuts",
       y = "MSE")
```



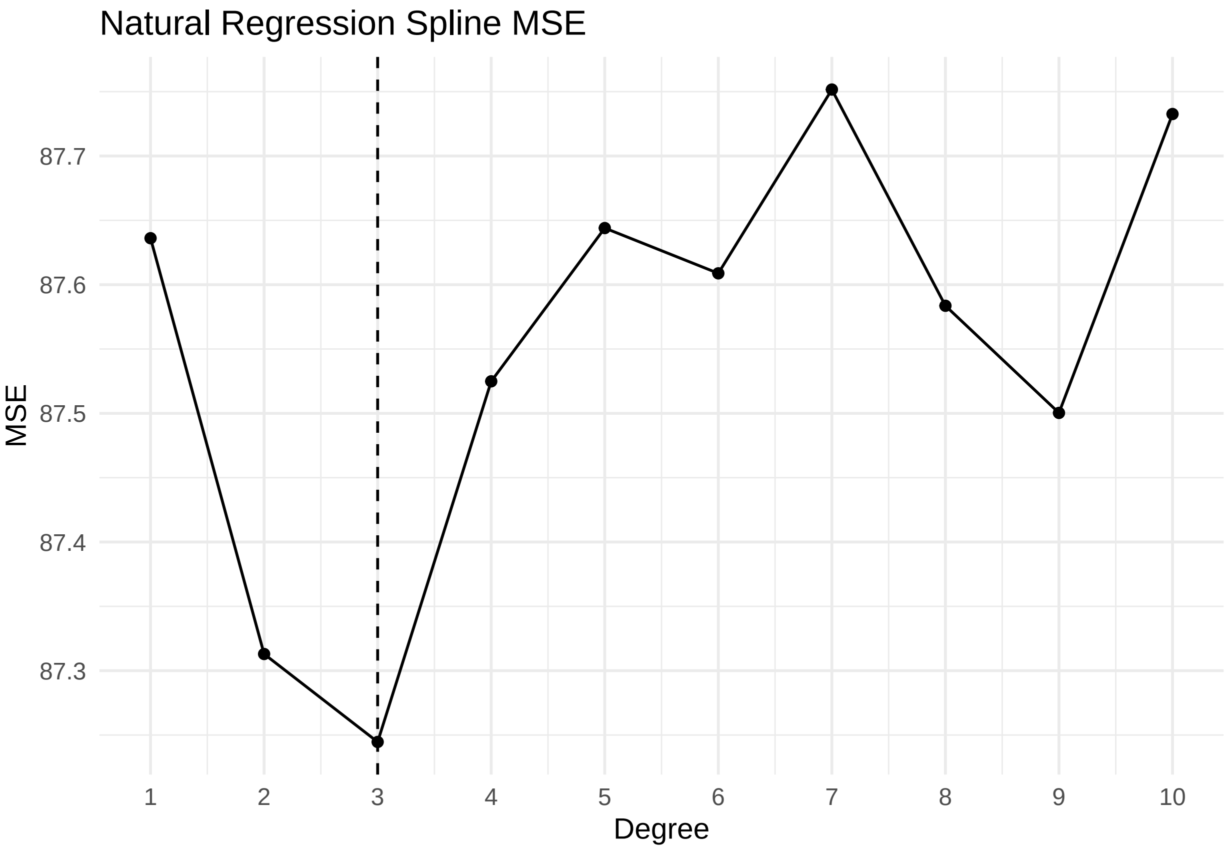
```r
step <- lm(egalit_scale ~ cut_interval(income06, which.min(errors) + 1), data = gss_train)

step %>%
  prediction %>%
  ggplot() +
  geom_line(aes(x = income06, y = fitted)) +
  geom_line(aes(x = income06, y = fitted + 1.96*se.fitted), linetype = "dashed") +
  geom_line(aes(x = income06, y = fitted - 1.96*se.fitted), linetype = "dashed") +
  labs(title = "Step Function",
       x = "Income",
       y = "Egalitarianism")
```



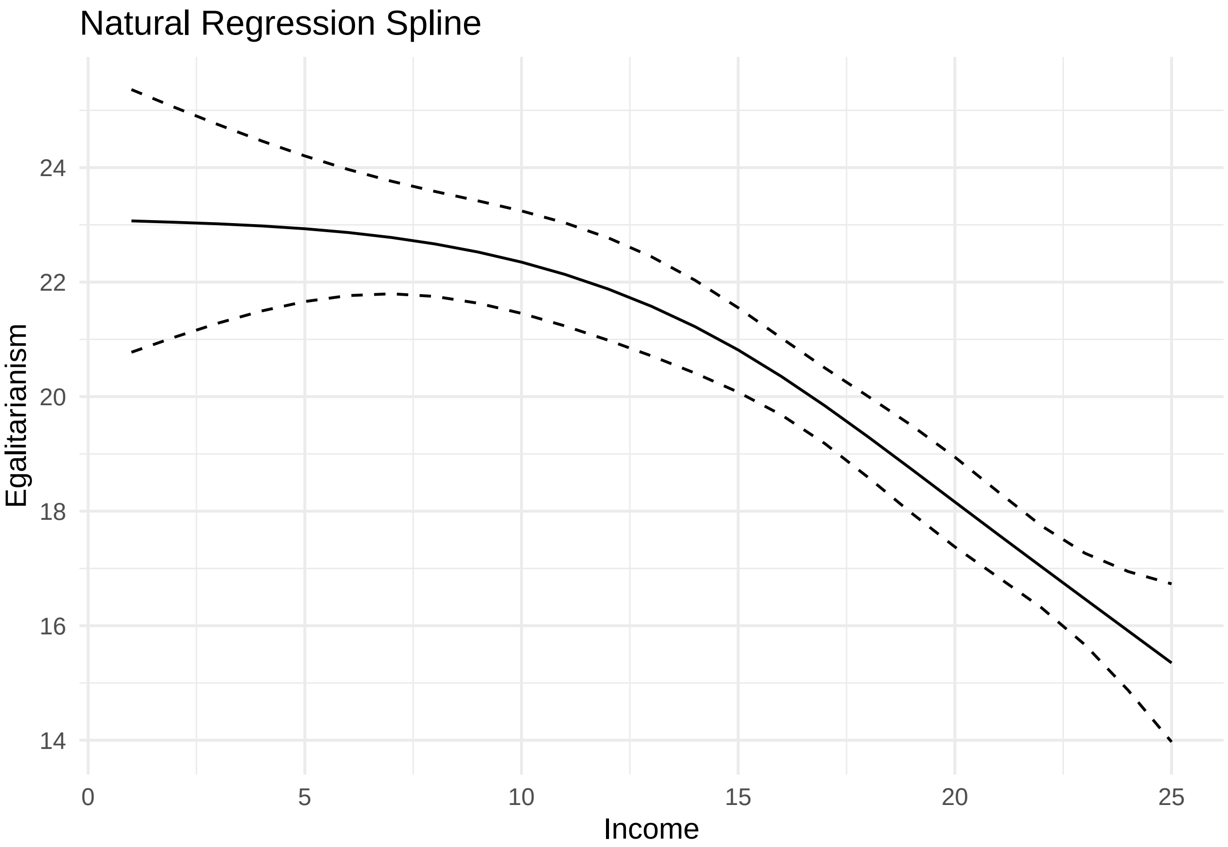
In terms of MSE, the optimal number of cuts(knots) is 8. We can see that in step function, as income increases, predicted egalitarianism decreases.

3

```
errors = data.frame(degree = (1:10), mse=0)

for (i in 1:10) {
  ns <- glm(egalit_scale ~ ns(income06, df = i), data = gss_train)
  errors[i, 2] = cv.glm(gss_train, ns, K = 10)$delta[1]
}

# MSE natural spline
ggplot(errors) +
  geom_line(aes(x = degree, y = mse)) +
  geom_point(aes(x = degree, y = mse)) +
  geom_vline(xintercept = which.min(errors$mse), linetype = "dashed") +
  scale_x_continuous(breaks=1:10) +
  labs(title = "Natural Regression Spline MSE",
       x = "Degree",
       y = "MSE")
```



```r
# optimal model
ns <- glm(egalit_scale ~ ns(income06, df = which.min(errors$mse)), data = gss_train)

ns %>%
  prediction %>%
  ggplot() +
  geom_line(aes(x = income06, y = fitted)) +
  geom_line(aes(x = income06, y = fitted + 1.96*se.fitted), linetype = "dashed") +
  geom_line(aes(x = income06, y = fitted - 1.96*se.fitted), linetype = "dashed") +
  labs(title = "Natural Regression Spline",
       x = "Income",
       y = "Egalitarianism")
```



In terms of MSE, it is the lowest when the degree is 3. The general trend of the regression shows that as income increases, predicted egalitarianism decreases.

Egalitarian and Everything

4

a

```
convert <- function(column) {
  return(as.numeric(as.factor(column)))
}

train <- as.data.frame(sapply(gss_train, convert))
test <- as.data.frame(sapply(gss_test, convert))

# Train control
tc <- trainControl(method = "cv", number = 10)

lm <- train(egalit_scale ~ ., data = train, method = "lm",
            metric = "RMSE", trControl = tc)

lm.pred <- predict(lm, newdata = test)
lm.mse <- mse(preds = lm.pred, actuals = test$egalit_scale)
```

b

```
enet <- train(egalit_scale ~ ., data = train, method = "glmnet",
              metric = "RMSE", trControl = tc, tuneLength = 10)

# alpha = 0.6 and lambda = 0.3219274.

enet.pred <- predict(enet, newdata = test, s = 0.3219274)

enet.mse <- mse(preds = enet.pred, actuals = test$egalit_scale)
```

c

```
# pre-processing
pcr <- train(egalit_scale ~ ., data = train, method = "pcr",
             metric = "RMSE", trControl = tc,
             preProcess = c("zv", "center", "scale"))

pcr.pred <- predict(pcr, newdata = test)

pcr.mse <- mse(preds = pcr.pred, actuals = test$egalit_scale)
```

d

```
pls <- train(egalit_scale ~ ., data = train, method = "pls",
             metric = "RMSE", trControl = tc,
             preProcess = c("zv", "center", "scale"))

pls.pred <- predict(pls, newdata = test)

pls.mse <- mse(preds = pls.pred, actuals = test$egalit_scale)
```

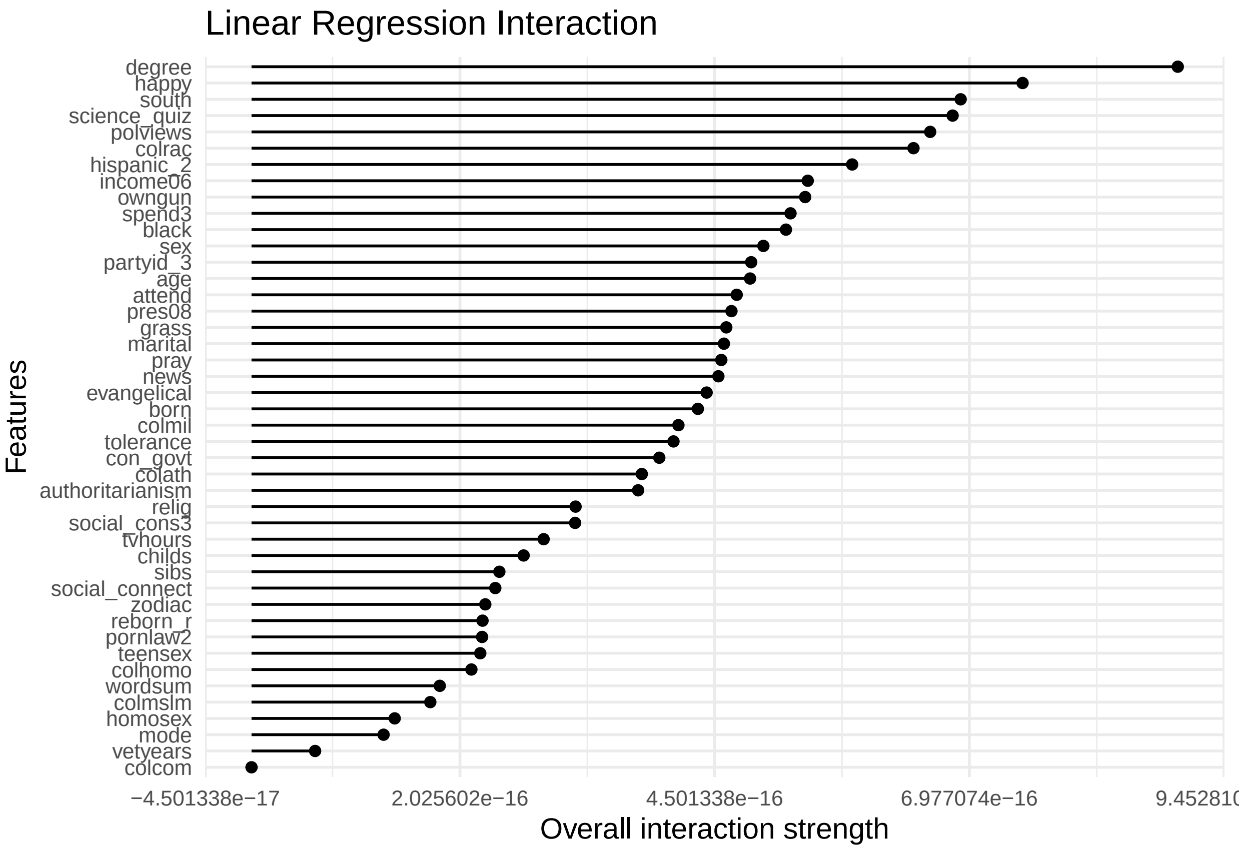
```
tibble(lm.mse, enet.mse, pcr.mse, pls.mse)
```

```
## # A tibble: 1 x 4
##   lm.mse enet.mse pcr.mse pls.mse
##    <dbl>    <dbl>   <dbl>   <dbl>
## 1   64.1     62.8    63.9    62.7
```

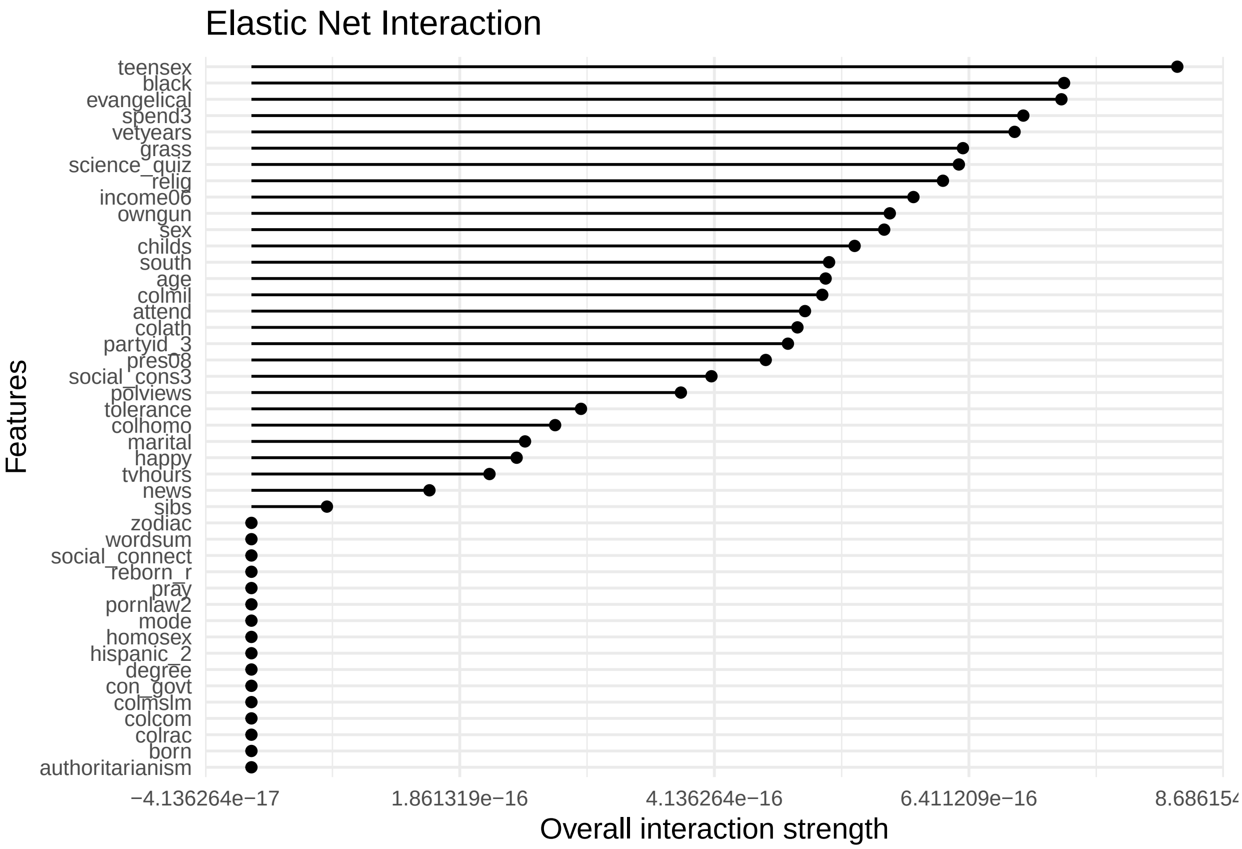
Comparing the MSE, elastic net has the lowest value, while principal component regression has the highest value. However, all four models do not have a large variance with regard to MSE.

5

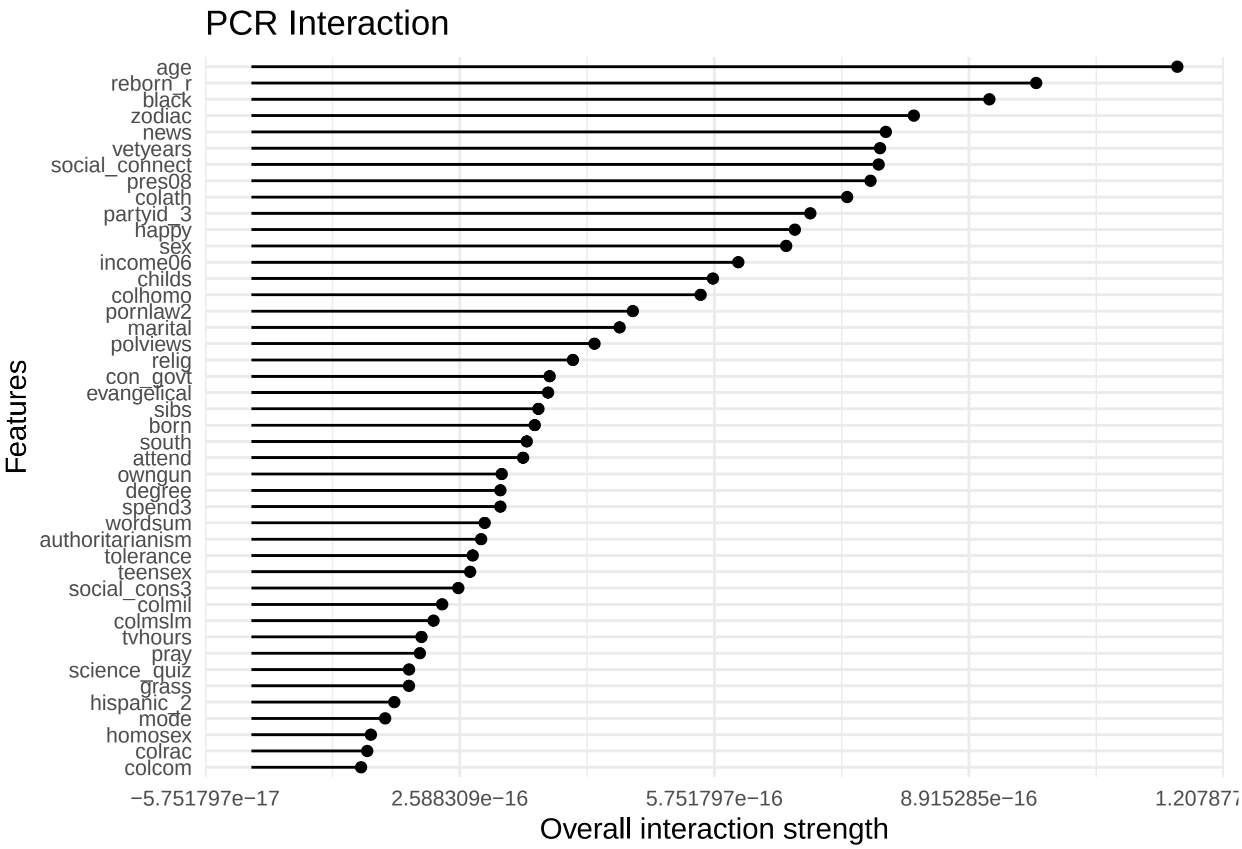
```r
features <- as.data.frame(select(train, -egalit_scale))
response <- as.numeric(train$egalit_scale)

predictor.lm <- Predictor$new(
  model = lm,
  data = features,
  y = response
)

predictor.enet <- Predictor$new(
  model = enet,
  data = features,
  y = response
)

predictor.pcr <- Predictor$new(
  model = pcr,
  data = features,
  y = response
)

predictor.pls <- Predictor$new(
  model = pls,
  data = features,
  y = response
)

# interaction plots
interact.lm <- Interaction$new(predictor.lm) %>%
  plot() +
  ggtitle("Linear Regression Interaction") +
  theme(axis.text = element_text(size = 8))

interact.enet  <- Interaction$new(predictor.enet) %>%
  plot() +
  ggtitle("Elastic Net Interaction") +
  theme(axis.text = element_text(size = 8))

interact.pcr <- Interaction$new(predictor.pcr) %>%
  plot() +
  ggtitle("PCR Interaction") +
  theme(axis.text = element_text(size = 8))

interact.pls <- Interaction$new(predictor.pls) %>%
  plot() +
  ggtitle("PLS Interaction") +
  theme(axis.text = element_text(size = 8))

interact.lm
```



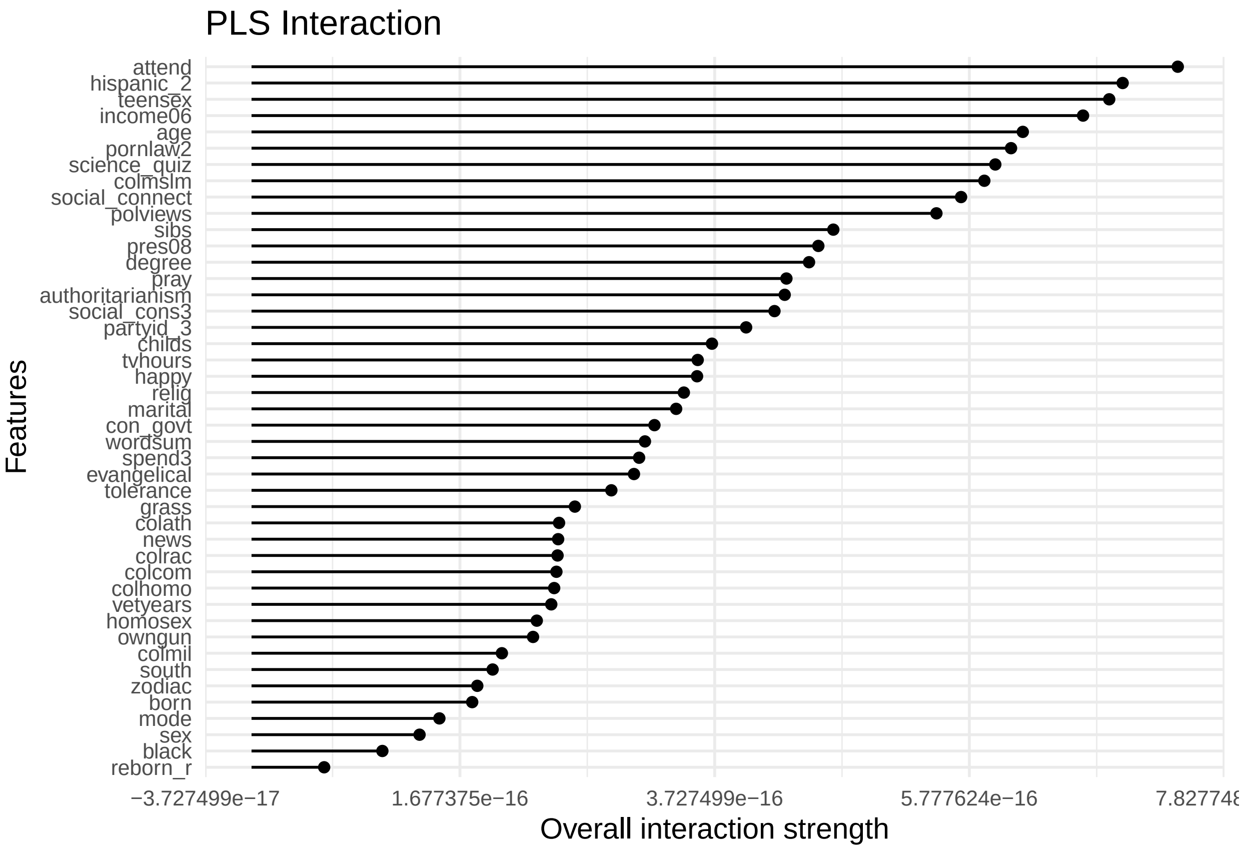
```
interact.enet
```



```
interact.pcr
```



```
interact.pls
```



To evaluate feature importance, we use interaction plots for each model here. In linear regression, the top five features are degree, happy, aouth science_quiz and polviews. In elastic net, the top five features are teensex, black, evangelical, spend3 and vetyears. In principal component regression, the top five are age, reborn_r, black, zodiac and news. Finally in PLS, the top five are attend, hispanic_2, teensex, income06 and age. Overall, these four non-linear models do not share many features that are important across the board. However, teensex, age and black all appear twice in the four models, indicating that they are slightly more important than others.