Question 3: Non-linearities

Lab 7 Handout Solutions

Statistics 139

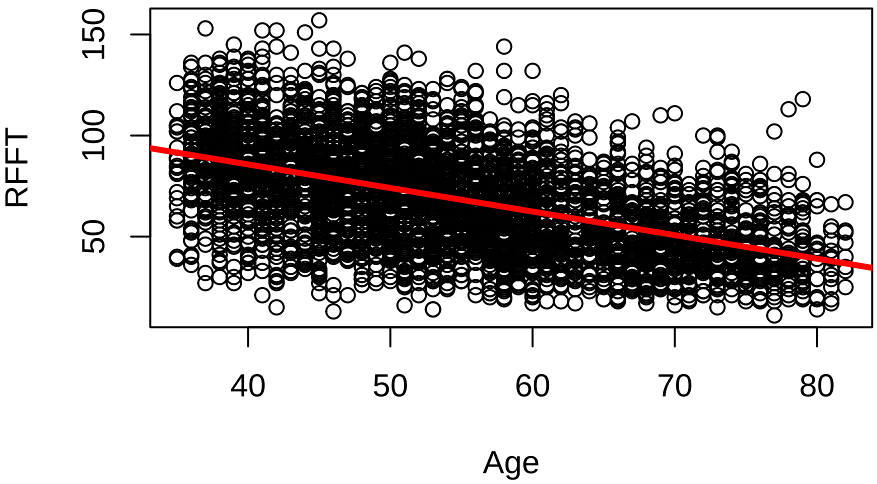
Question 3: Non-linearities

a) Fit a [linear] model to predict RFFT score from Age. Add the estimated line to the scatterplot and comment on the appropriateness of a simple linear model here.

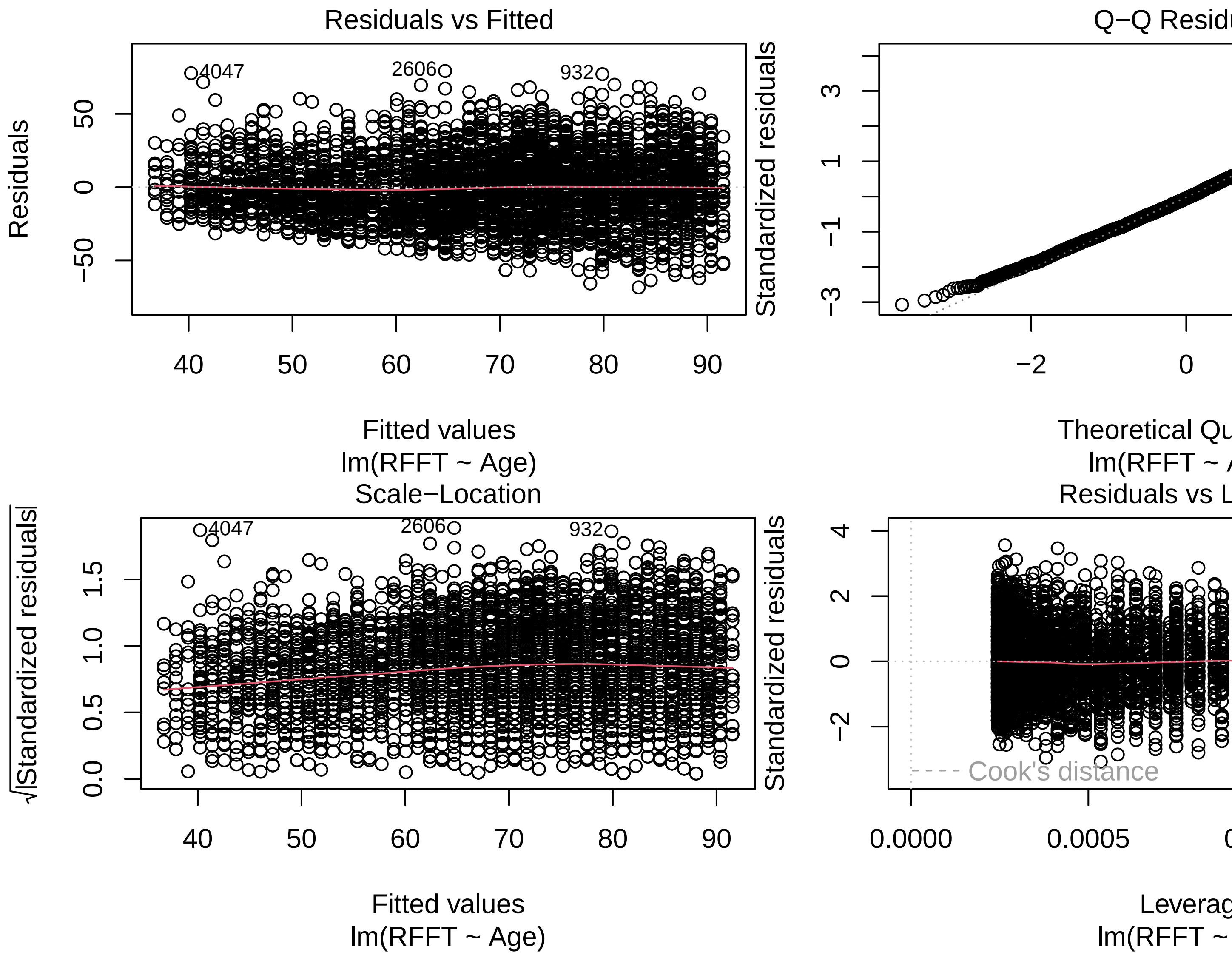
```
#load the data
prevend = read.csv("data/prevend.csv")

prevend$Education.Factor = factor(prevend$Education, levels = 0:3,
                                  labels = c("Primary", "LowSec", "HighSec", "Univ"))
```

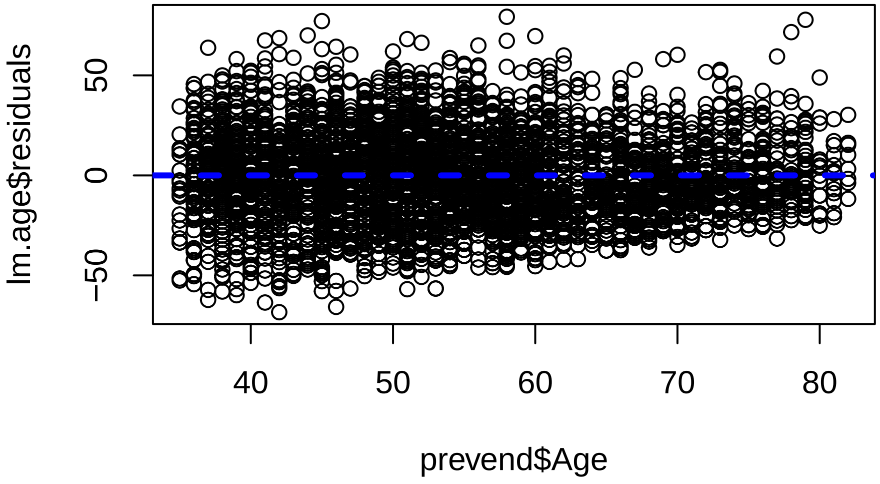
```
lm.age = lm(RFFT ~ Age, data = prevend)
plot(RFFT ~ Age, data = prevend)
abline(lm.age,col="red",lwd=3)
```



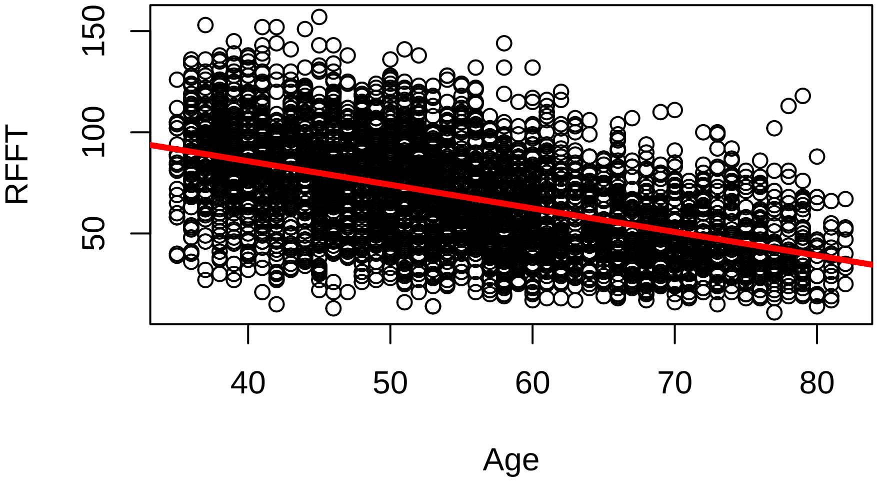
```
plot(lm.age, which=c(1))
plot(lm.age)
```



```r
plot(y=lm.age$residuals, prevend$Age)
abline(h=0, col="blue", lty=2, lwd=3)
```



```r
lm.age = lm(RFFT ~ Age, data = prevend)
plot(RFFT ~ Age, data = prevend)
abline(lm.age,col="red",lwd=3)
```



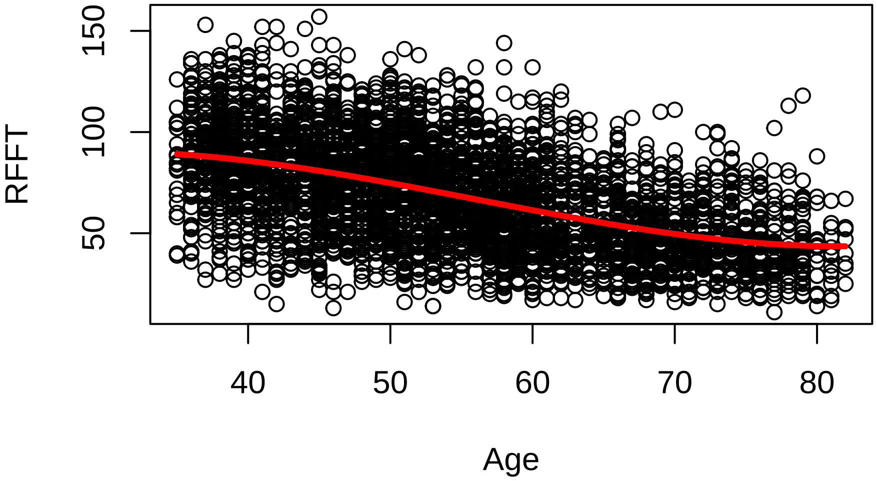
The estimated line suggests that a linear relationship may not be the best way to model $\mu_{Y|X}$: there appears to be a more gradual relationship (*plateau*) at both ends of the range of X.

b) Fit a model to predict RFFT score from a cubic model (3rd-order polynomial function) of Age. Interpret the estimates of this model, create a visual to illustrate the relationship of RFFT score with Age based on this model, and formally test with this model is preferred to handling age simply as a linear effect.

```r
lm.age.cubic = lm(RFFT ~ poly(Age,3,raw=T), data = prevend)
summary(lm.age.cubic)
```

```
##
## Call:
## lm(formula = RFFT ~ poly(Age, 3, raw = T), data = prevend)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -69.003 -15.619  -0.968  14.946  79.914
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            23.6263306 36.8926492   0.640  0.52194
## poly(Age, 3, raw = T)1  5.0430139  2.0377951   2.475  0.01337 *
## poly(Age, 3, raw = T)2 -0.1145177  0.0365720  -3.131  0.00175 **
## poly(Age, 3, raw = T)3  0.0006827  0.0002135   3.198  0.00139 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.23 on 4091 degrees of freedom
## Multiple R-squared:  0.2715, Adjusted R-squared:  0.2709
## F-statistic: 508.2 on 3 and 4091 DF,  p-value: < 2.2e-16
```

```
x = min(prevend$Age):max(prevend$Age)
yhat = predict(lm.age.cubic,new = data.frame(Age=x))
plot(RFFT~Age,data=prevend)
lines(yhat~x,col="red",lwd=3)
```



```
#abline(lm.age,col="blue",lwd=3)

anova(lm.age,lm.age.cubic)
```

```
## Analysis of Variance Table
##
## Model 1: RFFT ~ Age
## Model 2: RFFT ~ poly(Age, 3, raw = T)
##   Res.Df     RSS Df Sum of Sq     F   Pr(>F)
## 1   4093 2027392
## 2   4091 2021958  2    5433.8 5.497 0.004129 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The extra-sums-of-squares (ESS) F-test ($F = 5.497$, $p = 0.0041$) rejects the null hypothesis and shows a significant improvement in R^2 when including the extra two terms in the model (the quadratic and cubic terms associations, combined). The estimates themselves are difficult to interpret individually, and instead should be interpreted in tandem (the 3 estimates together). The slope is estimated to be positive at `Age` of zero since the linear term is positive, but this slope becomes less and less, and eventually negative, at ages starting around 30 years of age. This relationship becomes most negative around an age of 55 (as evidenced in the plot) and flattens out again at older ages. More mathematically:

$$\hat{f}'(X) = \hat{\beta}_1 + 2\hat{\beta}_2 X + 3\hat{\beta}_3 X^2 = 5.043 + 2 \cdot -0.1145 \cdot X + 3 \cdot 0.0006827 \cdot X^2 \equiv 0 \implies X = 30.154,\ 81.657$$

$$\hat{f}''(X) = 2\hat{\beta}_2 + 6\hat{\beta}_3 X \equiv 0 \implies X = \frac{-2 \cdot 0.1145}{6 \cdot 0.0006827} = 55.91$$

c) What are the implications of using a cubic model here? Why does it make sense mathematically based on the resulting plot?

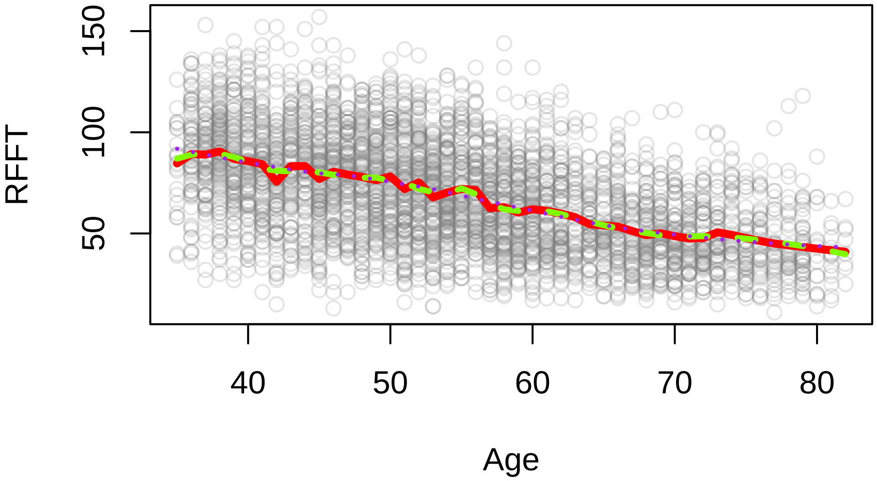
The cubic model allows for a relationship that tapers off on both sides, and as an odd function, will result in a slope that is either both negative or both positive at both ends of the range of X (depending on the range of X within the resulting estimated function for $\mu_{Y|X}$). A simple quadratic function would not allow for this.

d) Fit a loess model to predict RFFT score from Age. It is up to you to choose a well-suited value of `span` (include a visual to support your choice).

```
lo1.age = loess(RFFT ~ Age, data = prevend,span=0.1)
lo2.age = loess(RFFT ~ Age, data = prevend,span=0.2)
lo3.age = loess(RFFT ~ Age, data = prevend,span=0.5)

x = min(prevend$Age):max(prevend$Age)
yhat1 = predict(lo1.age,new = data.frame(Age=x))
yhat2 = predict(lo2.age,new = data.frame(Age=x))
yhat3 = predict(lo3.age,new = data.frame(Age=x))

plot(RFFT~Age,data=prevend,col=rgb(0.5,0.5,0.5,0.2))
lines(yhat1~x,col="red",lwd=4)
lines(yhat2~x,col="chartreuse",lwd=3,lty=2)
lines(yhat3~x,col="purple",lwd=2,lty=3)
```



The above plot fits the LOESS curve using three choices of `span`: 0.1, 0.2, 0.5. The smaller choices are too overfit, while 0.5 is plenty smooth enough (no need to go higher than that, though the default 0.75 would be very similar). It would be interesting to see how each would perform on a left-out test set.