Workshop 4 - Non-Linear Effects (full answers)

JPS

Introduction

In this workshop we are going to practice the modelling of non-linear relationships between explanatory and outcome variables. To do so we are going to practice two methods, *polynomial regression* and *LOESS* curves. As usual, we have two exercises, where we will explore the relationship between both experience and age on salaries, paying special attention to potential non-linear effects.

Exercise 1: In this exercise we are going to work with the sample of academic salaries that we introduced in the bonus exercise last week; this time to explore the potential non-linear effect of experience (measured as years since obtaining a phd) on salaries. The specific research question to be addressed could be formulated as: Do salaries increase with the number of years since the PhD was obtained? Non-linear effects will be explored using polynomial regression.

Exercise 2: Here we will use the Labour Force Survey (LFS) to explore the relationship between age and salaries in the UK. Specifically we will seek to answer the following research question: Are salaries directly proportional to years of experience? To explore this we will use polynomial regression and also LOESS curves.

Exercise 1. Academic Salaries

Let's start by accessing the academic salaries data, which is available in the **car** library.

```r
library(car)
data(Salaries)
summary(Salaries)
```

In the bonus exercise last week I used this dataset to explore the gender gap in the academic sector. We found that 'sex' is not statistically significant after controlling for some other relevant factors. One of the factors we considered was 'yrs.since.phd'. Much like what we observed for the case of 'sex', the explanatory variable 'yrs.since.phd' appear to be correlated with 'salary' when using simple bivariate analyses.

```r
cor(Salaries$salary, Salaries$yrs.since.phd)
```

The correlation is positive, which make sense, higher experience/seniority provide higher salaries. Specifically, we can estimate that for every additional year after the phd was obtained academic salaries go up by $985.

```r
model1 = lm(salary~yrs.since.phd, data=Salaries)
summary(model1)
```

```
##
## Call:
## lm(formula = salary ~ yrs.since.phd, data = Salaries)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -84171 -19432  -2858  16086 102383
##
## Coefficients:
```

```
##                Estimate Std. Error t value Pr(>|t|)    
## (Intercept)     91718.7     2765.8  33.162   <2e-16 ***
## yrs.since.phd     985.3      107.4   9.177   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 27530 on 395 degrees of freedom
## Multiple R-squared:  0.1758, Adjusted R-squared:  0.1737 
## F-statistic: 84.23 on 1 and 395 DF,  p-value: < 2.2e-16
```

However, once we control for other variables like 'rank', the effect of 'yrs.since.phd' is not significant anymore.

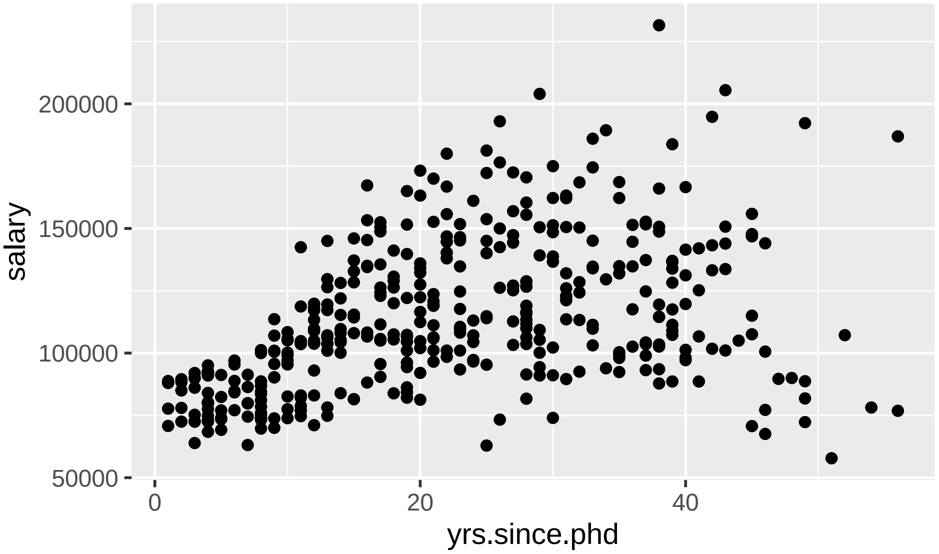
```r
model2 = lm(salary~rank+discipline+sex+yrs.since.phd, data=Salaries)
summary(model2)
```

```
## 
## Call:
## lm(formula = salary ~ rank + discipline + sex + yrs.since.phd, 
##     data = Salaries)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -67451 -13860  -1549  10716  97023 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   67884.32    4536.89  14.963  < 2e-16 ***
## rankAssocProf 13104.15    4167.31   3.145  0.00179 ** 
## rankProf      46032.55    4240.12  10.856  < 2e-16 ***
## disciplineB   13937.47    2346.53   5.940 6.32e-09 ***
## sexMale        4349.37    3875.39   1.122  0.26242    
## yrs.since.phd    61.01     127.01   0.480  0.63124    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 22660 on 391 degrees of freedom
## Multiple R-squared:  0.4472, Adjusted R-squared:  0.4401 
## F-statistic: 63.27 on 5 and 391 DF,  p-value: < 2.2e-16
```

It is possible that 'yrs.since.phd' does not have a significant effect on 'salary'. This is what we would have concluded based on our analysis. However, there is one specific assumption that we are violating, **linearity**, leading to a problem of misspecification in our model, which in turn has got the potential to bias our regression coefficients and measures of uncertainty.

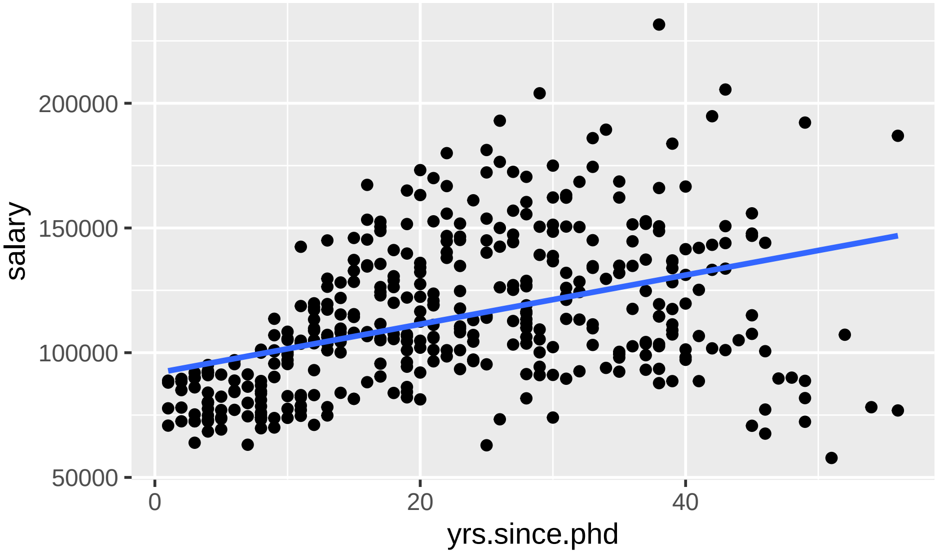
Let's look into this. As a rule of thumb, whenever you are interested in the relationship between continuous variables, always complement your exploratory analysis with scatter-plots so you can anticipate any potential non-linear relationships.

```r
library(ggplot2)
ggplot(Salaries, aes(x=yrs.since.phd, y=salary)) + geom_point()
```



What do you think? Does the relationship between 'salary' and 'yrs.since.phd' look linear? You can explore this visually by drawing the linear line of best fit and assessing whether some observations seem to be systematically over/under-estimated at different segments of the range of 'yrs.since.phd'.

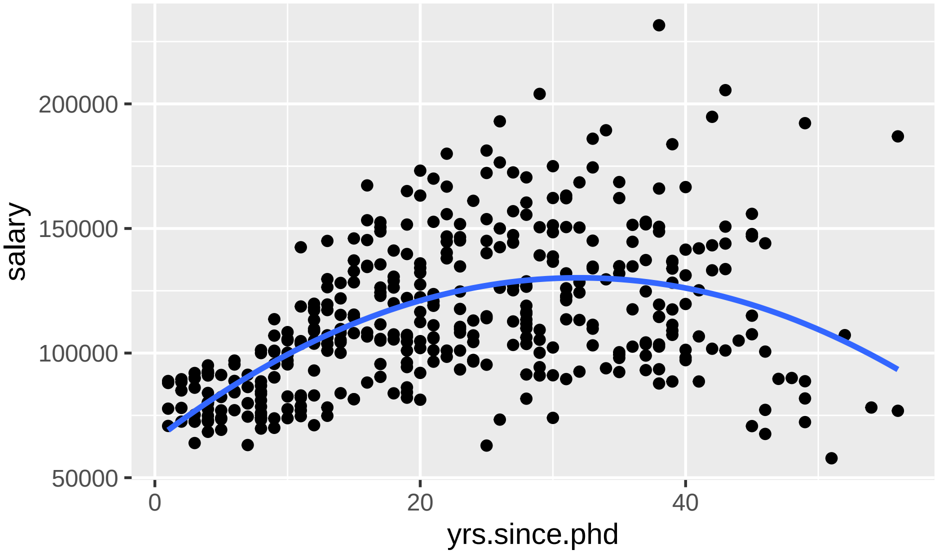
```
ggplot(Salaries, aes(x=yrs.since.phd, y=salary)) + geom_point() +
       stat_smooth(method="lm", formula=y~x, size = 1, se=FALSE) #This is an additional
```



```
#layer to your ggplot to draw a straight line of best fit.
```

It looks like our linear model overestimates the salaries of academics under 10 years of experience and those over 45 (or so), and perhaps tends to underestimate the salaries for those in between. This looks suspiciously like a quadratic (non-linear) relationship. To explore that we can request a quadratic line of best fit based on polynomial regression. Notice that the quadratic effect is introduced in the formula below within *I()*, this command specifies that ˆ is to be understood mathematically, i.e. as the sign for the exponent. This is necessary since within the *lm()* function ˆ is used to perform different programming functions.

```
ggplot(Salaries, aes(x=yrs.since.phd, y=salary)) + geom_point() +
                 stat_smooth(method="lm", formula=y~x+I(x^2), size=1, se=FALSE)
```



Ok, this looks much better. Clearly not a perfect fit to the data but much closer than before. Still, the line is rather horizontal and the curvature is not massive. In order to test whether such quadratic effect was certainly an improvement we can look at the significance of the quadratic term and at the typical measures of goodness of fit. If only the coefficient for X is significant but the coefficient for X^2 is not, we can take the effect as purely linear, if both are significant then we can claim the effect is quadratic. If only X^2 is significant, then this can be a lot trickier to interpret.

```
summary(model1)
```

```
##
## Call:
## lm(formula = salary ~ yrs.since.phd, data = Salaries)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -84171 -19432  -2858  16086 102383
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    91718.7     2765.8  33.162   <2e-16 ***
## yrs.since.phd    985.3      107.4   9.177   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27530 on 395 degrees of freedom
## Multiple R-squared:  0.1758, Adjusted R-squared:  0.1737
## F-statistic: 84.23 on 1 and 395 DF,  p-value: < 2.2e-16
```

```
model3 = lm(salary~yrs.since.phd + I(yrs.since.phd^2), data=Salaries)
summary(model3)
```

```
##
## Call:
## lm(formula = salary ~ yrs.since.phd + I(yrs.since.phd^2), data = Salaries)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -64228 -18329  -1535  14744 103649
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        65051.172   3949.715  16.470   <2e-16 ***
## yrs.since.phd       4075.903    364.819  11.172   <2e-16 ***
## I(yrs.since.phd^2)   -63.739      7.246  -8.797   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25200 on 394 degrees of freedom
## Multiple R-squared:  0.3111, Adjusted R-squared:  0.3076
## F-statistic: 88.95 on 2 and 394 DF,  p-value: < 2.2e-16
```

The difference between models once we account for the quadratic effect is remarkable. The *Adjusted R-squared* goes from 0.174 to 0.308. Notice that such improvement has not been achieved by adding new information, only by relaxing the linearity assumption invoked in ‘model 3’ using a quadratic effect from a variable that was already in the model, by simply going from $Y = \beta_0 + \beta_1 X + e$ to $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + e$. We can also see that the specific effect of ‘yrs.since.phd’ changes importantly. To interpret this we are going to proceed to specify a better model, where we control for other relevant factors, ‘Model 4’.

```r
summary(model2)
```

```
##
## Call:
## lm(formula = salary ~ rank + discipline + sex + yrs.since.phd,
##     data = Salaries)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -67451 -13860  -1549  10716  97023
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    67884.32    4536.89  14.963  < 2e-16 ***
## rankAssocProf  13104.15    4167.31   3.145  0.00179 **
## rankProf       46032.55    4240.12  10.856  < 2e-16 ***
## disciplineB    13937.47    2346.53   5.940 6.32e-09 ***
## sexMale         4349.37    3875.39   1.122  0.26242
## yrs.since.phd     61.01     127.01   0.480  0.63124
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22660 on 391 degrees of freedom
## Multiple R-squared:  0.4472, Adjusted R-squared:  0.4401
## F-statistic: 63.27 on 5 and 391 DF,  p-value: < 2.2e-16
```

```r
model4 = lm(salary~rank+discipline+sex+yrs.since.phd+I(yrs.since.phd^2), data=Salaries)
summary(model4)
```

```
##
## Call:
## lm(formula = salary ~ rank + discipline + sex + yrs.since.phd +
##     I(yrs.since.phd^2), data = Salaries)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max
## -62956 -13315  -1405   9831  96306
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        60378.762   5329.187  11.330  < 2e-16 ***
## rankAssocProf       5551.081   5033.353   1.103  0.27077
## rankProf           34100.878   6184.019   5.514 6.38e-08 ***
## disciplineB        14199.904   2331.061   6.092 2.68e-09 ***
## sexMale             5233.598   3860.948   1.356  0.17604
## yrs.since.phd       1512.625    565.503   2.675  0.00779 **
## I(yrs.since.phd^2)   -25.037      9.508  -2.633  0.00879 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22490 on 390 degrees of freedom
## Multiple R-squared:  0.4569, Adjusted R-squared:  0.4485
## F-statistic: 54.68 on 6 and 390 DF,  p-value: < 2.2e-16
```

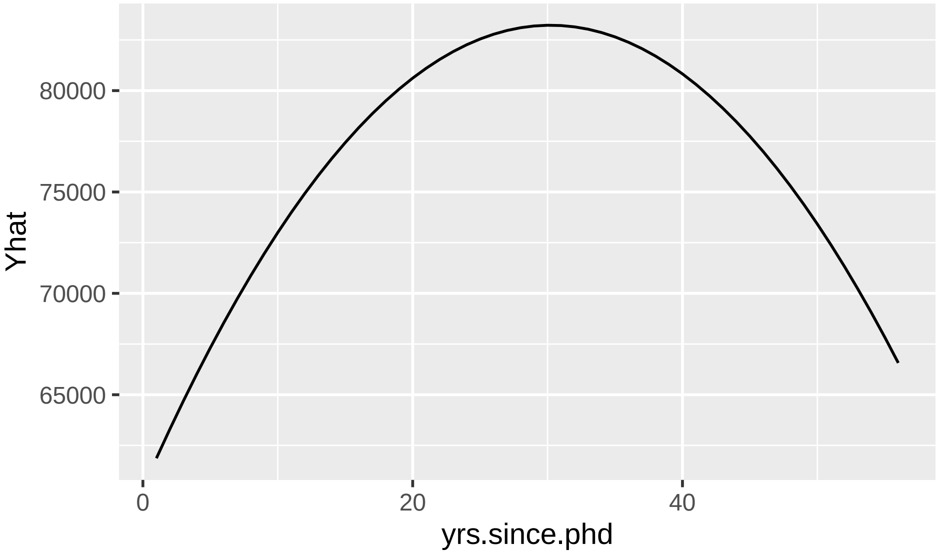
In 'model 2' the effect of 'yrs.since.phd' was not significant. This has changed now that we are specifying a non-linear effect. Ok, but what is now the specific effect of 'yrs.since.phd'? When we have a quadratic effect we cannot interpret it as simply as we do for linear effects. We cannot say the coefficient represents 'by how much Y changes for a one unit change in X, holding all other explanatory variables constant' since the effect of x on Y will now vary across the range of values of X. Instead, what we can do is describe such effect 'visually' using plots, 'qualitatively' using words, and 'quantitatively' using reference categories. For all of those options it helps to predict the outcome ($\hat{Y}$) first, using the non-linear effects obtained in our model.

Taken estimates of 'salary' based on our model to be defined as:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1\text{AssocProf} + \hat{\beta}_2\text{rankProf} + \hat{\beta}_3\text{disciplineB} + \hat{\beta}_4\text{Male} + \hat{\beta}_5\text{years.since.phd} + \hat{\beta}_6\text{years.since.phd}^2.$$

We can use $\hat{\beta}_0 + \hat{\beta}_5\text{years.since.phd} + \hat{\beta}_6\text{years.since.phd}$ to predict salaries for a specific *reference category*, i.e. that specific case defined when all explanatory variables included in the model equal 0. In our case, the reference category will be a female assistant professor working in a discipline A. We proceed to predict salaries based on our model for the range of 'yrs.since.phd' values contemplated in our sample (1 to 56).

```
yrs.since.phd = 1:56  #The range of yrs.since.phd in our dataset.
Yhat = model4$coefficients[1] + model4$coefficients[6]*yrs.since.phd +
       model4$coefficients[7]*yrs.since.phd^2  #The estimated salary for different
       #levels of yrs.since.phd.
pred = as.data.frame(cbind(yrs.since.phd, Yhat)) #Combining yrs.since.phd and predicted
       #salaries into the same dataset so we can plot them together.
ggplot(pred, aes(x=yrs.since.phd, y=Yhat)) + geom_line()
```



Based on the above plot we can describe the effect of ‘yrs.since.phd’ as positive, increasing at a diminishing rate for every additional year up to roughly 30 years, when the highest value is reached, after that point the effect becomes negative at an increasing rate for every additional year. If we want to be more concrete and provide the specific effect of ‘yrs.since.phd’ we will need to refer to a specific range of values. For example, we could estimate that, holding everything else constant, salary increases by roughly $5450 when the number of years since obtaining the phd goes from 1 to 5 years.

```
Yhat5 = model4$coefficients[1] + model4$coefficients[6]*5 + model4$coefficients[7]*5^2
Yhat1 = model4$coefficients[1] + model4$coefficients[6]*1 + model4$coefficients[7]*1^2
Yhat5 - Yhat1
```

```
## (Intercept)
##    5449.603
```

```
#You can also do the following so the output looks a bit tidier.
Yhat5_1 = Yhat5 - Yhat1
names(Yhat5_1) = "Difference in salary 1 to 5 years after PhD"
Yhat5_1
```

```
## Difference in salary 1 to 5 years after PhD
##                                    5449.603
```

Now, many academics never retire and keep working until their death bed. Question: Can you use our model4 to predict the salary for someone like John Bannister Goodenough (2019 Nobel Prize in Chemistry), who obtained his phd at the age of 30 and in 2019 was 97 years old?

```
Yhat67 = model4$coefficients[1] + model4$coefficients[6]*67 + model4$coefficients[7]*67^2
Yhat67
```

```
## (Intercept)
##     49332.1
```

Does this prediction make sense? An academic with 67 years of experience is earning less than someone who just got their phd? This example illustrates one of the main issues affecting parametric regression (which can be loosely understood as regression models where effects of explanatory variables are pre-established based on a given functional form, be that linear, quadratic, log-linear, etc.). Extrapolating beyond the range of a given sample should be done carefully. We need to use common sense and consider whether the cases to be predicted could realistically be derived from the sample we are using. This is particularly problematic when using quadratic functions, which can increase/decrease at an accelerated rate as we approach their tails, so it can be quite misleading to extrapolate beyond the sample range.

Theoretically, it does not make sense to see negative salaries, or even salaries decreasing at an accelerated rate past 30 yeas since obtaining a phd, as suggested by our model. I would expect that such reduction in salaries flattens out at one point, so salaries stay positive. We could explore that hypothesis using polynomial regression, in particular we could include a cubic term to allow for a second point of inflection in the relationship between 'yrs.since.phd' and 'salary'. Question: Would you know how to expand Model 4 to do so? Hint: you just need to include one more polynomial term for 'yrs.since.phd'. As for the quadratic term, you can use the *I()* function for the new cubic term to be added.

```
model5 = lm(salary~rank+discipline+sex+yrs.since.phd+I(yrs.since.phd^2)+
            I(yrs.since.phd^3), data=Salaries)
summary(model5)
```

```
## 
## Call:
## lm(formula = salary ~ rank + discipline + sex + yrs.since.phd + 
##     I(yrs.since.phd^2) + I(yrs.since.phd^3), data = Salaries)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -62707 -13369  -1298   9525  94275 
## 
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)    
## (Intercept)        68805.1532  7050.1899   9.759  < 2e-16 ***
## rankAssocProf      11484.4941  5985.9782   1.919   0.0558 .  
## rankProf           40640.9614  7138.0574   5.694 2.45e-08 ***
## disciplineB        14171.6497  2324.2491   6.097 2.60e-09 ***
## sexMale             4703.7794  3860.5898   1.218   0.2238    
## yrs.since.phd       -519.1300  1251.4977  -0.415   0.6785    
## I(yrs.since.phd^2)    56.2123    45.6749   1.231   0.2192    
## I(yrs.since.phd^3)    -0.9567     0.5261  -1.818   0.0698 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 22430 on 389 degrees of freedom
## Multiple R-squared:  0.4615, Adjusted R-squared:  0.4518 
## F-statistic: 47.62 on 7 and 389 DF,  p-value: < 2.2e-16
```

We can see that the new cubic term is not significant, so we would fall back to model 4 as our best model. However, as we saw last week, this sample of academic salaries is quite small, and specific to one particular American college. We are now going to proceed to assess whether using the LFS (a bigger and more generalisable sample) we can figure out the concrete form of the relationship between age and salaries, for which we might be able to do better than assuming they are linearly or quadratically related.

Exercise 2. UK Salaries

Download the LFS file from Minerva into a folder in your computer.

```
load("lfs.rda")
```

This is a huge dataset, so, as we did last week, we are going to start by trimming it down. We will keep a small set of variables, 8 in total, and discard the other 755. The variables to be used are 'SEX', 'AGE', 'SC10MMJ' (major occupation group), 'EMPMON' (number of months continuously employed), 'TRVTME' (usual home to work travel time in minutes), 'TTUSHR' (total usual hours worked including overtime), 'QUAL_1' (degree level qualification), 'GRSSWK' (gross weekly pay in main job).

```r
vars = c("SEX","AGE","SC10MMJ","EMPMON","TRVTME","TTUSHR","QUAL_1","GRSSWK")
lfs = lfs[vars]
summary(lfs)
```

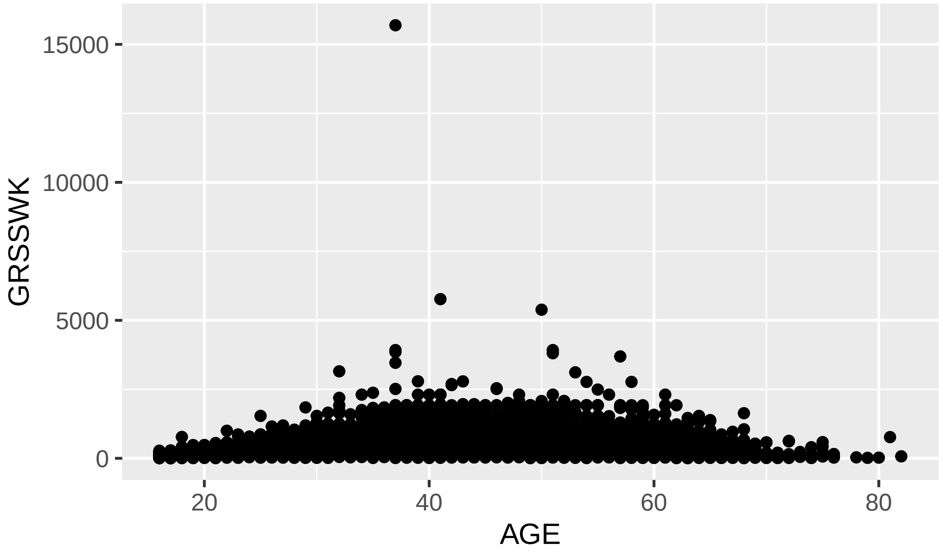
```
##               SEX              AGE
##  Does not apply:    0   Min.   : 0.00
##  No answer     :    0   1st Qu.:18.00
##  Male          :49392   Median :39.00
##  Female        :52667   Mean   :38.17
##                         3rd Qu.:57.00
##                         Max.   :99.00
##
##                                              SC10MMJ          EMPMON
##  Does not apply                                  :54664   Min.   : -9.00
##  Professional Occupations                        : 9021   1st Qu.: -9.00
##  Associate Professional and Technical Occupations: 6406   Median : -9.00
##  Administrative and Secretarial Occupations      : 5498   Mean   : 45.01
##  Elementary Occupations                          : 5219   3rd Qu.: 62.00
##  Skilled Trades Occupations                      : 5203   Max.   :792.00
##  (Other)                                         :16048
##      TRVTME            TTUSHR                 QUAL_1          GRSSWK
##  Min.   : -9.000   Min.   :-9.00   Does not apply:    0   Min.   :   -9.00
##  1st Qu.: -9.000   1st Qu.:-9.00   No answer     :    0   1st Qu.:   -9.00
##  Median : -9.000   Median :-9.00   No            :85959   Median :   -9.00
##  Mean   :  3.058   Mean   :11.42   Yes           :16100   Mean   :   44.29
##  3rd Qu.: 10.000   3rd Qu.:38.00                          3rd Qu.:   -9.00
##  Max.   :180.000   Max.   :97.00                          Max.   :15692.00
##
```

We should also trim down the number of cases in the dataset by removing those with missing data in any of the variables. Often missing cases are coded as non-sensical negative values since this allows to differentiate for different types of missingness (e.g. -9 for those who were not asked the question, -8 for those who declined to provide an answer, etc.), which is something that cannot be done if all missing cases are set as NA.

```r
lfs = lfs[which(lfs$GRSSWK>-1),]
lfs = lfs[which(lfs$SC10MMJ!="Does not apply"),]
lfs = lfs[which(lfs$EMPMON>-1),]
lfs = lfs[which(lfs$TTUSHR>-1),]
table(lfs$TRVTME, useNA="ifany")
#There are 687 cases left with missing information for TRVTME.
#Since there are only 19 who report taking 0 minutes to travel to work,
#here I am just going to assume that those with TRVTME=-9 represent people working from home.
lfs$TRVTME = ifelse(lfs$TRVTME==-9, 0, lfs$TRVTME)
```

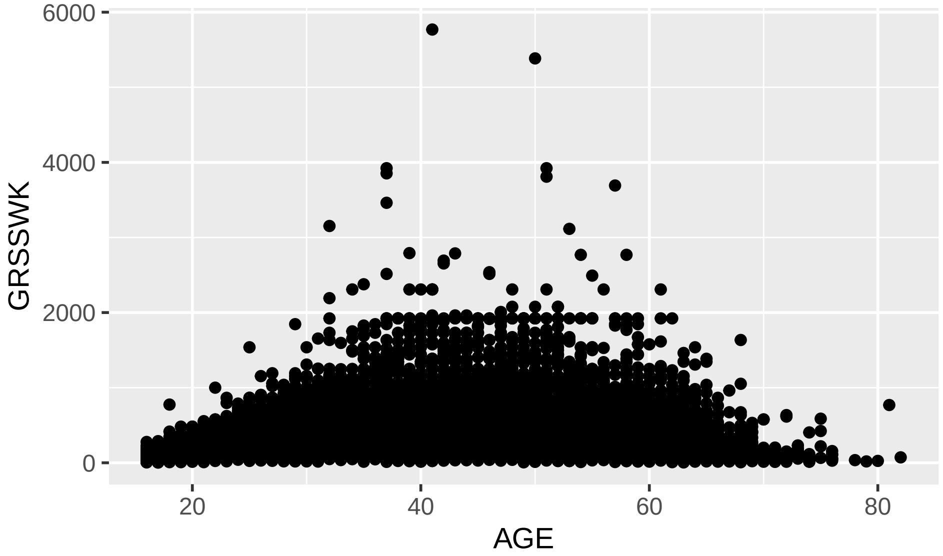
Ok, let's now get deeper into the exploratory analysis by looking at potential non-linear relationships between 'GRSSWK', 'AGE' and any other continuous variables in our sample.

```r
ggplot(lfs, aes(x=AGE, y=GRSSWK)) + geom_point()
```



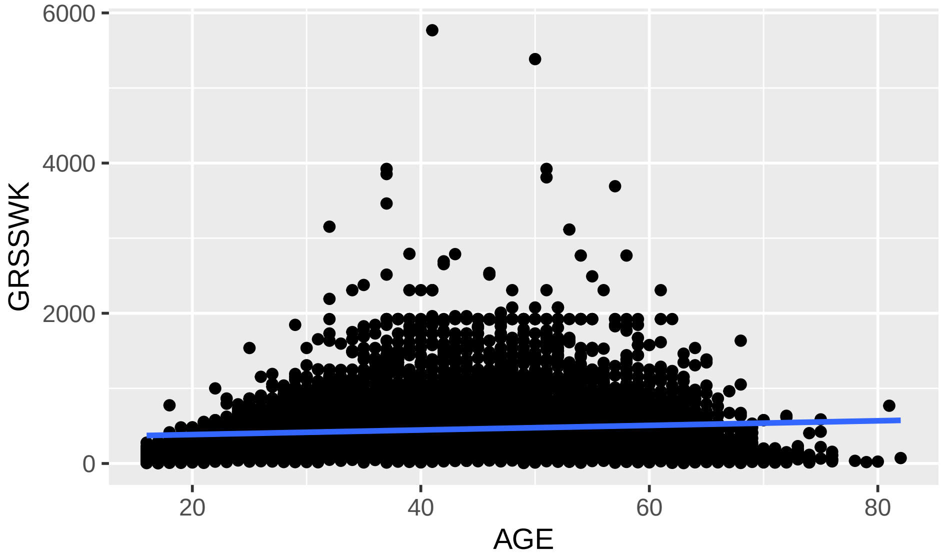
There is one outlier above £15,000 that is pulling the y-axis up. Question: Can you get rid of that case and rerun the ggplot so we can visualise the relationship between 'GRSSWK' and 'AGE' better? Hint: You can see how we did that for negative cases of 'GRSSWK' above.

```
lfs = lfs[which(lfs$GRSSWK<10000),]
ggplot(lfs, aes(x=AGE, y=GRSSWK)) + geom_point()
```

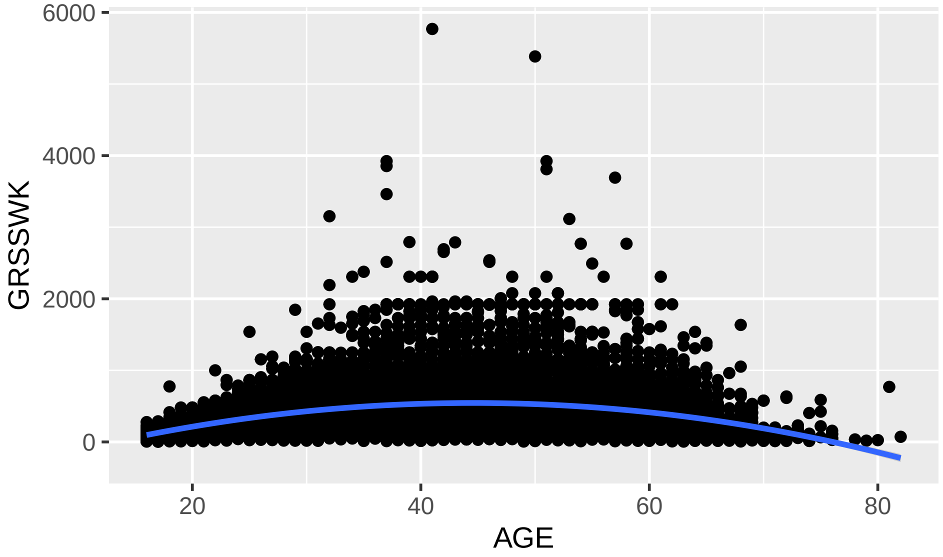


Removing the outlier allows compressing the y-axis in the plot, which in turn helps us assess whether the relationship between age and salary could be non-linear (possibly quadratic). Question: To have a better idea of whether that is the case, can you add a straight and a quadratic line of best fit to the ggplot? Hint1: To do so you can add the *stats_smooth()* function that we used in Exercise 1. Hint2: to request a linear or a quadratic function you need to change the *formula* option.

```
ggplot(lfs, aes(x=AGE, y=GRSSWK)) + geom_point() +
  stat_smooth(method="lm", formula=y~x, size = 1)
```



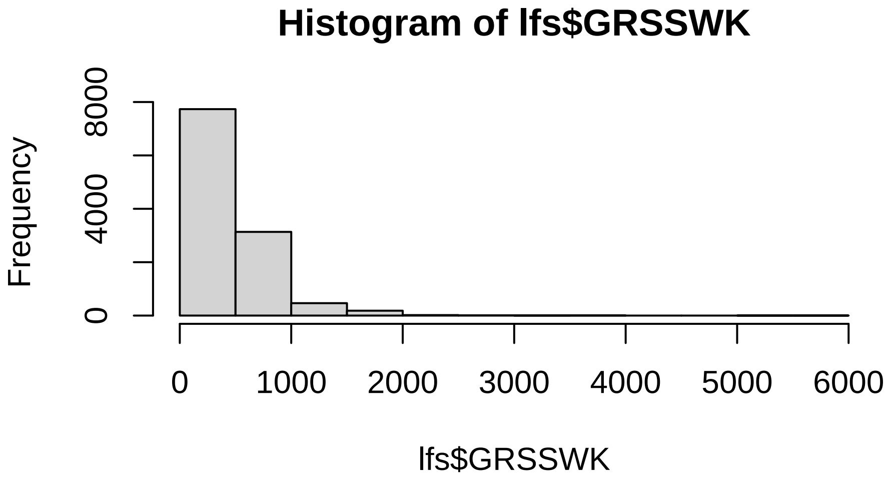
```
ggplot(lfs, aes(x=AGE, y=GRSSWK)) + geom_point() +
  stat_smooth(method="lm", formula=y~x+I(x^2), size = 1)
```



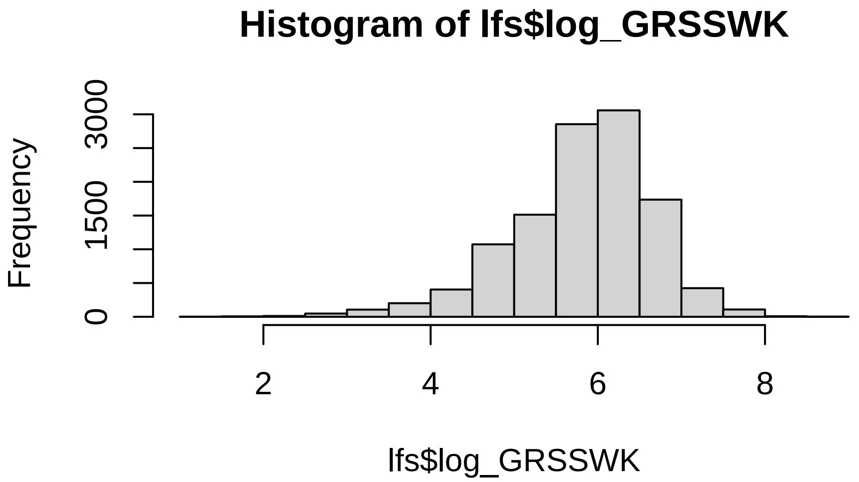
It seems that the quadratic model offers a better fit than the linear model, although it is not perfectly clear. This is because the sample size is so huge that it is difficult to visualise where most points lay. In addition, we can see how the quadratic model might introduce the non-sensical estimations of negative salaries that we observed for the academic sector data, perhaps a second point of inflexion at the tail of the age distribution might help. We should keep in mind all this information that we are gathering from our exploratory analysis to inform our modelling strategy.

Let's do that now, model the effect of age on salaries, which we can do while while controlling for other relevant factors (potential confounders). To do so we have to take a look at the distribution of 'GRSSWK', our outcome variable. We learnt last week how this distribution is right-skewed, which is why we should first log-transform it, so the model's residuals will be roughly normally distributed.

```r
hist(lfs$GRSSWK)
```



```r
lfs$log_GRSSWK = log(lfs$GRSSWK)
hist(lfs$log_GRSSWK)
```



Question: Can you test whether including a quadratic effect of age on log-salaries provides a better fit of the data than simply representing that relationship linearly? Hint: Estimate a model for 'log_GRSSWK' including 'AGE', but also potential confounders that you can control for. Then estimate another model where you introduce a quadratic term for 'AGE', to do so you can use *I()*, as we did in Exercise 1. Question: Would you also add a cubic term? Hint: To test this you can expand the quadratic model by including a cubic term using *I()* again. In order to assess which is the better model you can check the adjusted R^2, and whether the additional terms added to the model are statistically significant and strong enough to be considered consequential. Remember that if in doubt the more parsimonious model should be preferred.

We specify a first model including only linear effects, which we take as a benchmark.

```
model1 = lm(lfs$log_GRSSWK ~ SEX+QUAL_1+AGE+EMPMON+TRVTME+TTUSHR+SC10MMJ, data=lfs)
summary(model1)
```

All the variables included are significant (the only exception being ‘Associate Professional and Technical Occupations’, one of the categories for ‘SC10MMJ’). Let’s now extend the model considering the non-linear effect observed in the exploratory analysis.

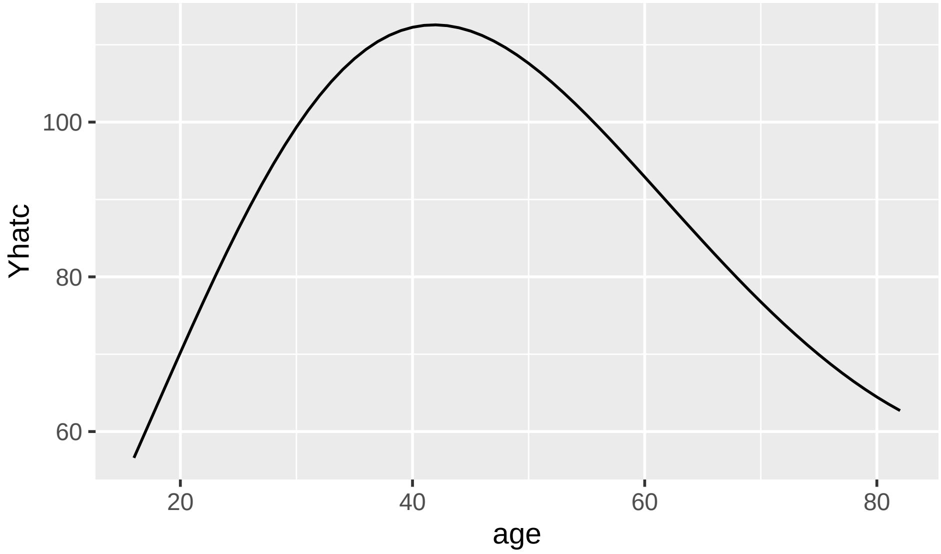
```
model2 = lm(lfs$log_GRSSWK ~ SEX+QUAL_1+AGE+I(AGE^2)+EMPMON+TRVTME+TTUSHR+SC10MMJ,
            data=lfs)
summary(model2)
```

Both ‘AGE’ and ‘AGEˆ2’ are significant, and the adjusted R^2 goes from 0.701 to 0.722. So, we can conclude that the quadratic effect provides a better fit. However, as we saw in the previous exercise, it does not really make sense to think of negative salaries, or even an accelerated rate of decrease in the effect of age. Adding a cubic effect would solve that problem by adding another point of inflection, let’s see if that is the case.

```
model3 = lm(lfs$log_GRSSWK ~ SEX+QUAL_1+AGE+I(AGE^2)+AGE+I(AGE^3)+EMPMON+TRVTME+
                             TTUSHR+SC10MMJ, data=lfs)
summary(model3)
```

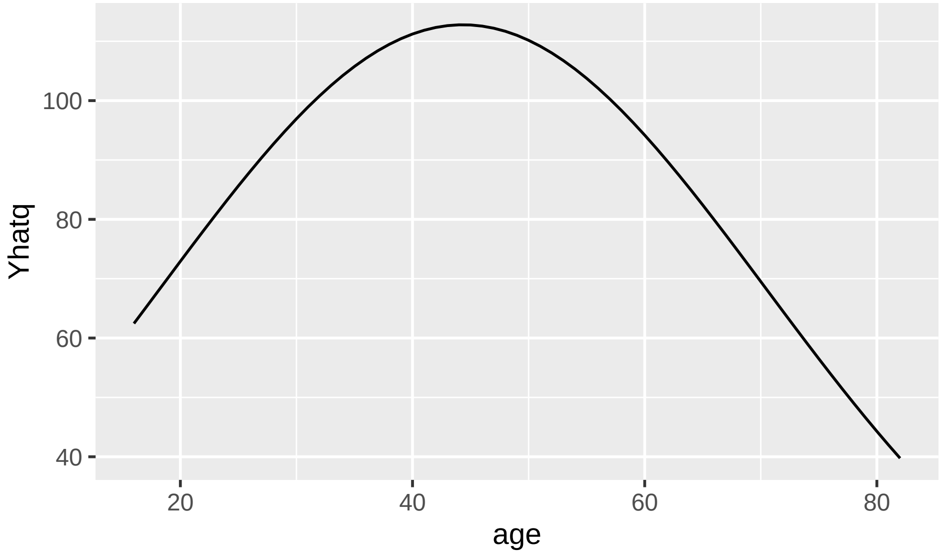
Again, the model is improved although this time only marginally so. The three terms in the polynomial function for ‘AGE’ are significant, but the adjusted R^2 only goes from 0.722 to 0.723. If facing this situation, and if my research goal was not centered around the estimation of the effect of age on salaries (say for example, if my interest was on the gender gap), I would probably drop the cubic term and keep the simpler quadratic effect. However, if the research question was to ascertain the effect of age, I would definitely keep the cubic term as it provides a more accurate estimate. I would also proceed to report the age effect visually as we did in the previous exercise

```
range(lfs$AGE)
age = 16:82
model3$coefficients
Yhatc = exp(model3$coefficients[1] + model3$coefficients[4]*age +
            model3$coefficients[5]*age^2 + model3$coefficients[6]*age^3)
pred = as.data.frame(cbind(age, Yhatc))
ggplot(pred, aes(x=age, y=Yhatc)) + geom_line()
```



The cubic part of the effect is not massive, but sufficient to provide a more realistic representation of the overall effect of age on salaries, particularly at the end of the age range. Compare this to the estimations based on the quadratic effect.

```
model2$coefficients
Yhatq = exp(model2$coefficients[1] + model2$coefficients[4]*age +
            model2$coefficients[5]*age^2)
pred = as.data.frame(cbind(age, Yhatq))
ggplot(pred, aes(x=age, y=Yhatq)) + geom_line()
```

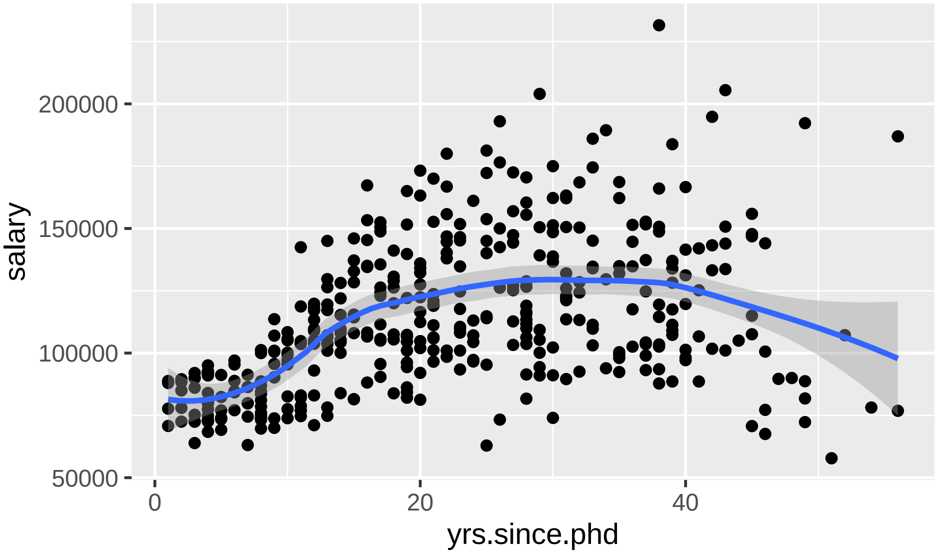


We have seen that the effect of age on log-salaries is not linear, but we have only tested two non-linear functions (quadratic and cubic), how do we know one of this is the line of best fit? We could consider other non-linear functions based for example on a log transformation of 'AGE', or even based on the use of exponential functions on a different base, e.g. 0.5 or 1.5. We could try some of these different functions one by one, or we could also rely on a more computationally intensive, data-driven (non-parametric) approach, based on LOESS.

LOESS can be used for one or a short number (normally no more than four) explanatory variables, and for that reason is normally used in exploratory analysis. A simple way to use it is through the *stat_smooth* function from *ggplot* that we have already employed. Only now we will substitute the specification of a given function (above we specified a quadratic function using *ggplot*), for a span value. Remember that LOESS fits multiple regressions throughout different *local neighborhoods*. The size of the neighborhood can be controlled using the *span* argument, which ranges from 0 to 1. The greater the value of span the smoother the fitted curve will be, i.e. the closer to a linear model.

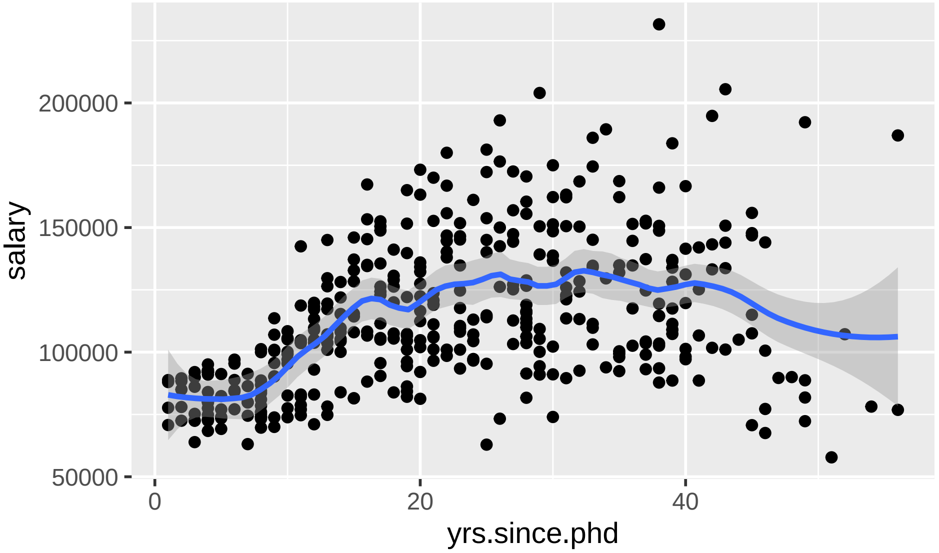
Deciding on the right level of span can be tricky. Remember that there is a trade-off between precision and accuracy. The higher the span the bigger the sample size to be used in each neighbourhood, hence the higher the level or precision (i.e, the smaller the standard errors). But at the same time, the higher the span the smoother the line of best fit will be, hence, the harder it will be to detect changes in the relationship between X and Y. Let's go back to the academic salaries data (based on a much smaller sample than the LFS), to illustrate this point. Let's obtain a first LOESS curve with a span of 0.5.

```
ggplot(Salaries, aes(x=yrs.since.phd, y=salary)) + geom_point() +
                 stat_smooth(method="loess", span=0.5, size = 1)
```

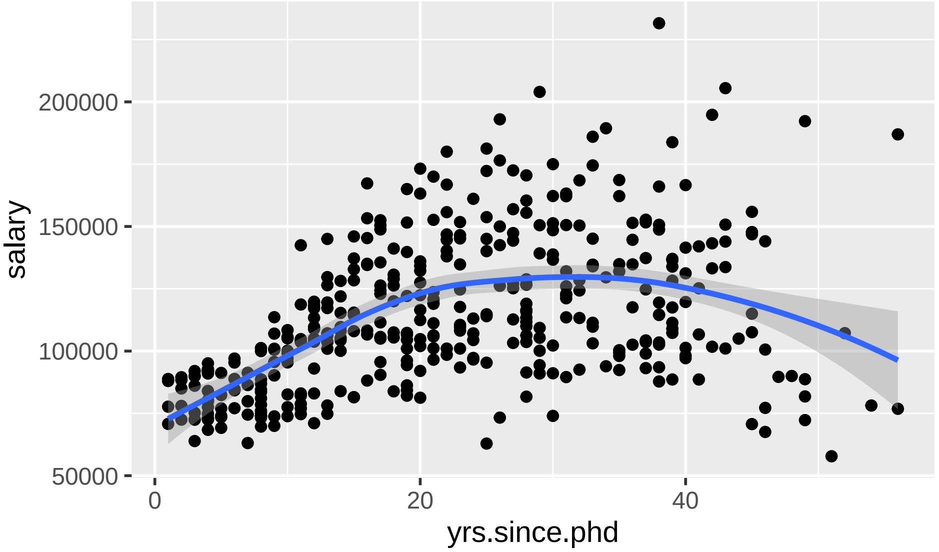


Notice how, unlike what we established before, the effect does not seem to be entirely quadratic. Notice as well how the confidence interval varies across the range of X. That is reflecting the number of observations available at each of the neighbourhoods defined by the level of span. This feature matters, polynomial regression will be wrongly assuming that standard errors and confidence intervals will remain uniform across the whole sample. LOESS offers a more realistic depiction of the level of uncertainty in areas where we do not have too many cases, which in our case would be translated in being a lot more cautious about making claims and predictions regarding the effect of age on salaries near the right-end of the range of age. Let's now use a span of 0.25 and 0.75 to assess visually the trade-off between precision and accuracy present in LOESS models.

```
ggplot(Salaries, aes(x=yrs.since.phd, y=salary)) + geom_point() +
                 stat_smooth(method="loess", span=0.25, size = 1)
```



```
ggplot(Salaries, aes(x=yrs.since.phd, y=salary)) + geom_point() +
                 stat_smooth(method="loess", span=0.75, size = 1)
```



The former identifies a series of bumps along the x-axis but these are probably not that informative. In addition, the level of precision is higher in the latter. So, I would probably go with a span from 0.5 to 0.75.

Now, this is a great exploratory tool, but we have seen repeatedly throughout the course how (when using observational data, i.e. non-experimental data) the relationship between two variables is often confounded by third factors. To embed a 'LOESS-type' function for one or a number of explanatory variables while controlling for other relevant variables we need to use generalised additive models (GAMs). In addition, these models can also be used when the outcome variables is binary, a count variable or for any other non-normal distribution. To learn more about GAMs have a look at this tutorial from Anish Singh.

Preparation for next week's workshop

Next week we will see how to analyse time-series. Specifically, we will learn how to explore time-series and estimate ARIMA models. We will do so step by step first, using a dataset on bike sharing, and then we will move on to automatise the modelling process using a data-driven approach, *auto.arima*, which will be employed to test whether the sentencing guidelines have increased sentence severity in England and Wales. There is a lot of new material that we will be covering here, which will make the practical a bit longer than usual. As always, take a look at the lecture and see if you can follow the instructions in the practical, as far as you can reach. See you next week.