Chapter 4:
Fitting curves with the linear model

STATS 201/8

University of Auckland

Learning outcomes

In this chapter you will learn about:

- Identifying a curved relationship between x and y
- Fitting a quadratic curve using a linear model
- Relevant `R`-code.

Section 4.1
Identifying a curved relationship

New Example – Exam vs. assignment marks

We'll continue working with the STATS 20x data, but now we are interested to see if assignment mark is associated with exam mark.

Again, we are pretty sure we know what the answer is, but we need to formally confirm our suspicions. Also, we want to use assignment mark to help explain (i.e., make inference about) exam mark.

The variables of interest are:

`Exam`	the student's exam mark (out of 100)
`Assign`	the student's assignment mark (out of 20).

Once again, `Exam` is the (numeric) response variable, and now `Assign` is the (numeric) explanatory variable.

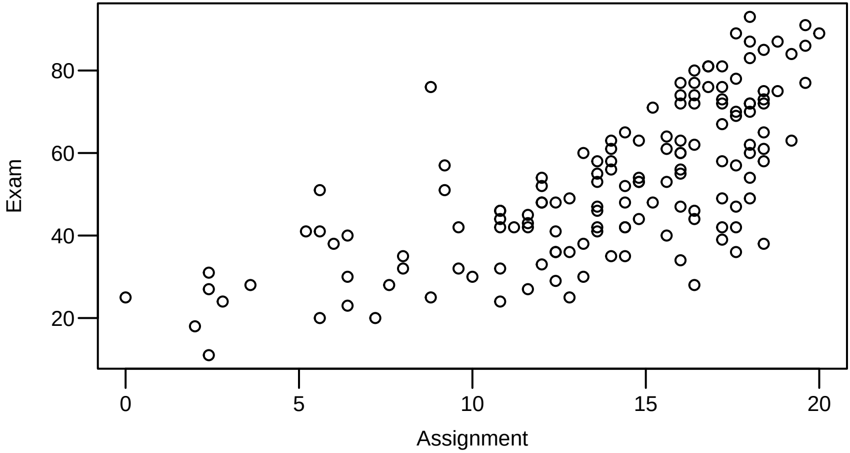
Exam vs. assignment marks. . .

Setting things up

```
> ## Load the s20x library into our R session
> library(s20x)
> ## Importing data into R
> Stats20x.df = read.table("Data/STATS20x.txt", header=T)
> ## Examine the data
> plot(Exam ~ Assign, data = Stats20x.df,xlab="Assignment")
```

Exam vs. assignment marks. . .

Scatterplot of the data



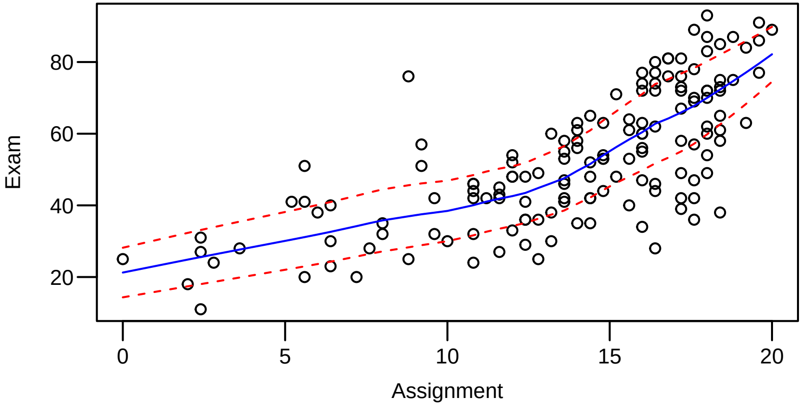
Hmmm, not quite a straight line – could be some curvature.

Maybe `trendscatter` will paint a clearer picture.

Exam vs. assignment marks. . .

Scatterplot with trend line

```
> trendscatter(Exam ~ Assign, data = Stats20x.df,xlab="Assignment")
```



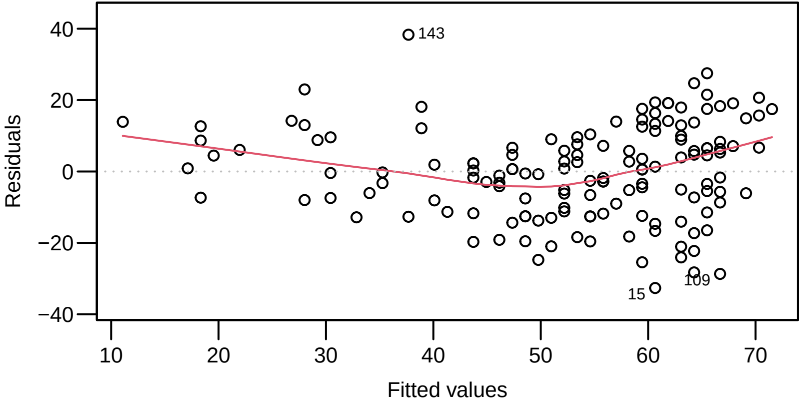
Sure looks like some curvature, but, at least the scatter looks fairly constant around this curve.

Exam vs. assignment marks...

Simple linear model

Let's fit a simple linear model to these data and see if it works out or not.

```
> examassign.fit=lm(Exam~ Assign,data = Stats20x.df)
> plot(examassign.fit,which=1)
```



Not surprisingly, we still have a curved relationship.

Exam vs. assignment marks. . .

Dealing with curvature

The assumption of identical distribution with expected value of 0 looks to be questionable here. There tend to be more negative residuals in the middle, but more positive residuals at the extremes of the fitted values.

Potential solution – add a quadratic (squared term) for `Assign`.

Section 4.2
Fitting a quadratic model

The quadratic curve

The standard notation for a quadratic curve is[1]

$$y = ax^2 + bx + c$$

Here we will use different notation: $\beta_0 \equiv c$, $\beta_1 = b$ and $\beta_2 = a$ and use the quadratic curve to describe the expected value of our dependent variable y. That is,

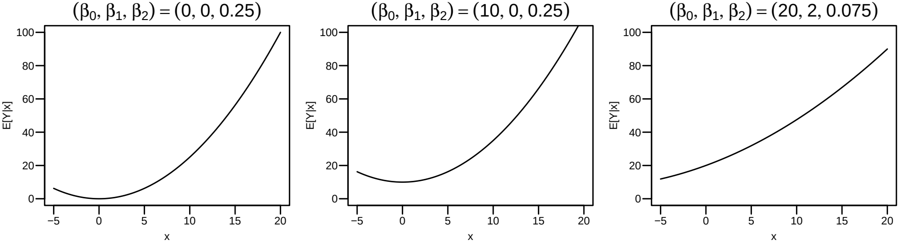
$$\mathrm{E}[Y|x] = \beta_0 + \beta_1 x + \beta_2 x^2$$

This is a linear model with explanatory terms x and x^2 – remember, the intercept β_0 is implicitly included.

[1] If you have done a bit of calculus, then you might recall that the roots (the values of x that give $y = 0$) of a quadratic are $\frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$.

The quadratic curve. . .

If $\beta_2 > 0$, then the quadratic has slope that increases with increasing x:

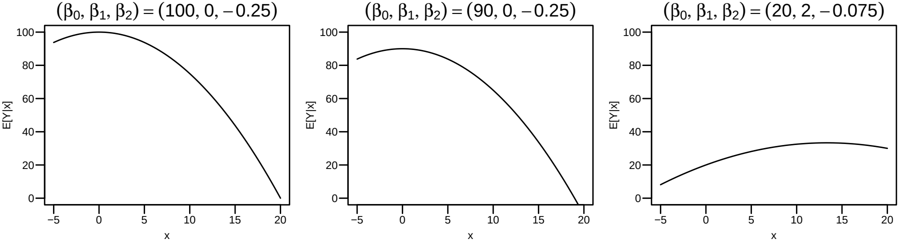


$(\beta_0, \beta_1, \beta_2) = (0, 0, 0.25)$

$(\beta_0, \beta_1, \beta_2) = (10, 0, 0.25)$

$(\beta_0, \beta_1, \beta_2) = (20, 2, 0.075)$

If $\beta_2 < 0$, then the quadratic has slope that decreases with increasing x:



$(\beta_0, \beta_1, \beta_2) = (100, 0, -0.25)$

$(\beta_0, \beta_1, \beta_2) = (90, 0, -0.25)$

$(\beta_0, \beta_1, \beta_2) = (20, 2, -0.075)$

How can a quadratic be a linear model?

(Non-examinable)

Throughout this course, when we use the term “linear model” we mean a model that is linear with respect to the β coefficients.

This mean that the derivative of the linear model with respect to any β coefficient is a constant.

The quadratic curve model

$$\mathrm{E}[Y|x] = \beta_0 + \beta_1 x + \beta_2 x^2$$

is a quadratic model for x.

The derivatives of this quadratic with respect to β_0, β_1 and β_2 are 1, x and x^2, respectively. These derivatives are all considered “constants” because they do not depend on any β coefficient.

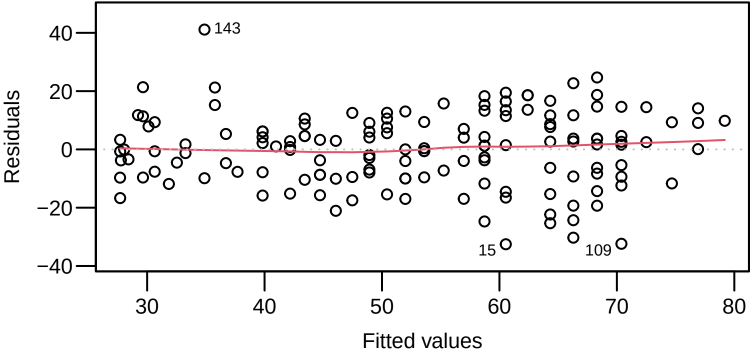
That is, the quadratic (in x) model is linear in β_0, β_1 and β_2.

Exam vs. assignment marks. . .

Adding a squared term

Add a squared term for `Assign` via `I(Assign^2)`, like this:[2]

```
> examassign.fit2=lm(Exam~ Assign + I(Assign^2), data = Stats20x.df)
> plot(examassign.fit2,which=1)
```



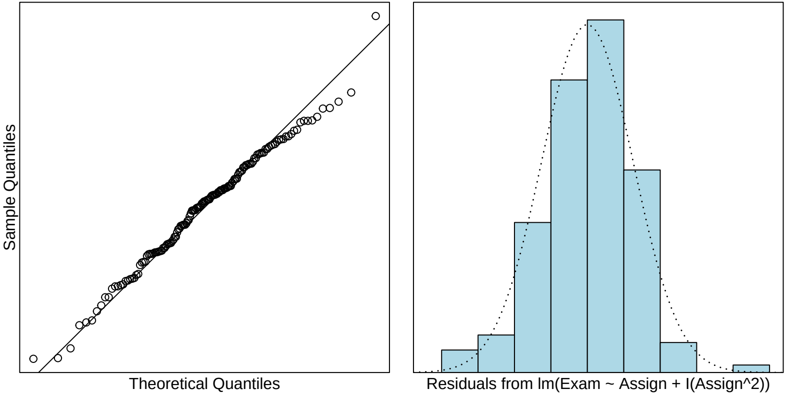
That is looking much better.

[2] **NOTE:** In the `lm` formula it is necessary to enclose the `Assign^2` term inside `I()` so that `lm` can make sense of it.

Exam vs. assignment marks. . .

Normality check of the quadratic model

```
> normcheck(examassign.fit2)
```

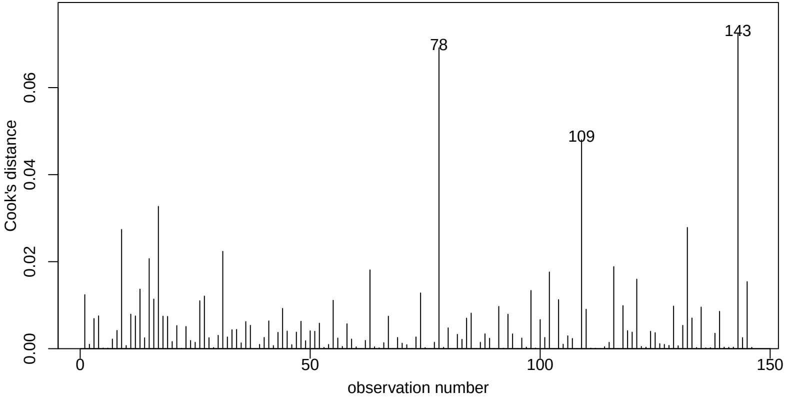


Looking good. There is one potential outlier. Let us check if it is influential.

Exam vs. assignment marks...

Influence check of the quadratic model

```
> cooks20x(examassign.fit2)
```



No high influence points.

Exam vs. assignment marks. . .

The fitted models

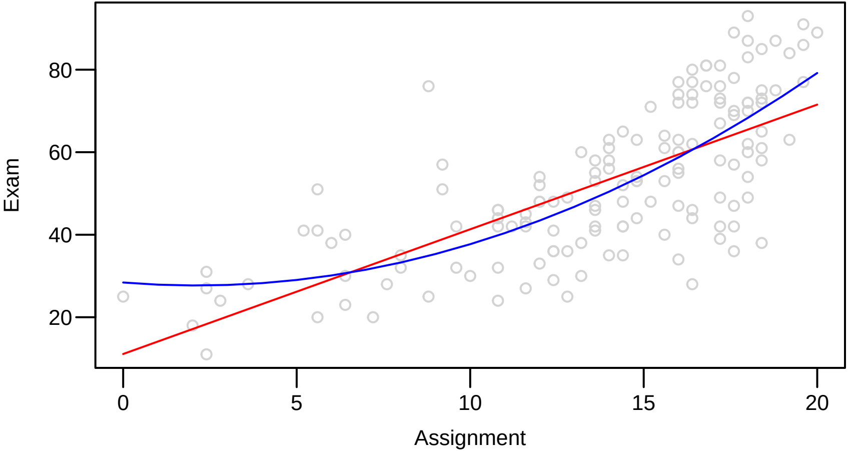
We have fitted a quadratic to see if we can add the ‘curviness’ in the relationship between test score and exam mark into our model.

Let us compare the two models visually – model 1 (linear) in red and model 2 (quadratic) in blue.

```
> plot(Exam˜ Assign, data = Stats20x.df, xlab="Assignment")
> x=0:20 #Assignment values at which to predict exam mark
> ## Plot model 1
> lines(x, predict(examassign.fit,data.frame(Assign=x)), col="red")
> ## Plot model 2
> lines(x, predict(examassign.fit2,data.frame(Assign=x)), col="blue")
```

Exam vs. assignment marks. . .

The fitted models. . .



Exam vs. assignment marks. . .

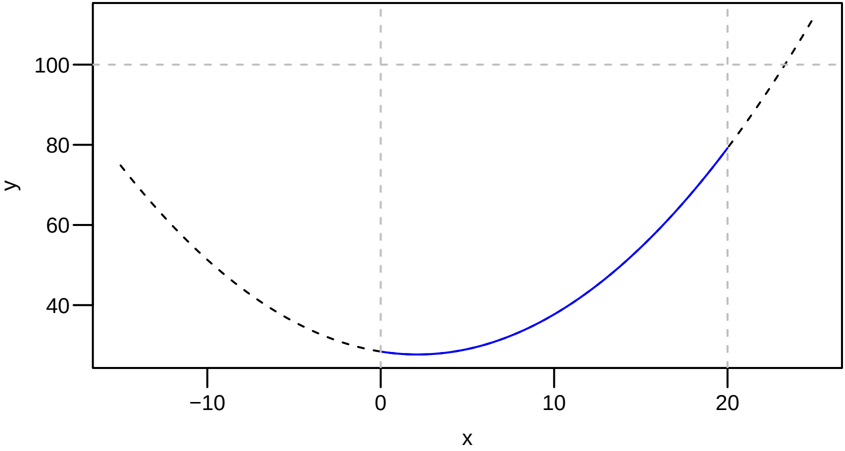
The fitted quadratic model

To plot the quadratic over a wider range of `x` (`=Assign`) values we can use the following code:

```
> x=seq(-15, 25,by =.10) #Sequence of from -15 to 25, in steps of 0.1
>
> y=predict(examassign.fit2,newdata = data.frame(Assign=x))
> plot(y~x, type="l",lty=2)
>
> ## The bits we want,  0<=x<=20 - N.B. Here & (ampersand) = AND
> lines(x[x>=0&x<=20],y[x>=0&x<=20],col="blue")
>
> ## The range of assign & exam respectively
> abline(v=range(Stats20x.df$Assign),lty=2, col="grey")
> abline(h=c(0,100),lty=2, col="grey")
```

Exam vs. assignment marks. . .

The fitted quadratic model. . .



Exam vs. assignment marks. . .

Comparison of straight line and quadratic models

```
> summary(examassign.fit)   ## Straight line model
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.0874     3.5954   3.084  0.00245 **
Assign        3.0222     0.2478  12.195  < 2e-16 ***
---
Residual standard error: 13.15 on 144 degrees of freedom
Multiple R-squared:  0.508,Adjusted R-squared:  0.5046
F-statistic: 148.7 on 1 and 144 DF,  p-value: < 2.2e-16
```

```
> summary(examassign.fit2)  ## Model with quadratic term
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.41396    5.99081   4.743 5.05e-06 ***
Assign      -0.68172    1.07242  -0.636 0.525999
I(Assign^2)  0.16102    0.04545   3.542 0.000536 ***
---
Residual standard error: 12.65 on 143 degrees of freedom
Multiple R-squared:  0.5477,Adjusted R-squared:  0.5414
F-statistic: 86.59 on 2 and 143 DF,  p-value: < 2.2e-16
```

Exam vs. assignment marks. . .

Comparison of straight line and quadratic models. . .

The small *P*-value $(= 0.000536)$ for testing $H_0 : \beta_2 = 0$ tells us that that the quadratic term is statistically significant. Our model went from:

$Exam_i = \beta_0 + \beta_1 \times Assign_i + \varepsilon_i$ where $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ to

$Exam_i = \beta_0 + \beta_1 \times Assign_i + \beta_2 \times Assign_i^2 + \varepsilon_i$ where $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.

Note that the coefficient $\beta_2 > 0$ associated with the `I(Assign)^2` term results in an improvement in expected exam score that 'accelerates' as `Assign` increases.

We could consider removing the non-significant 'straight line' `Assign` term. **What would you do?**

We have done a better job of modelling this data by adding this extra term, and the R^2 explained another 4% of the total variation.

MORAL: If it looks like a curve then fit a curve – provided the scatter about the curve is constant (**EOV**).

Section 4.3
Relevant R-code

Most of the R-code you need for this chapter

If you suspect the relationship between your x and y variables follows a curve rather than a straight line (as revealed in the plot of residuals vs fitted values), and the scatter remains constant around this curve, then fit a quadratic:

```
> examassign.fit2=lm(Exam~ Assign + I(Assign^2), data = Stats20x.df)
> #Check the residual plot again - hopefully the curvature is gone.
> plot(examassign.fit2,which=1)
```

NOTE: If the null hypothesis $H_0 : \beta_2 = 0$ is **not** rejected (i.e, P-value> 0.05), then our preferred model would be the simple linear regression model.