Why is survival modelling different from linear regression modelling?

We explore National Life Tables data from the Office of National Statistics (2018 – 2020). The full dataset is available at https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/lifeexpectancies/datasets/nationallifetablesunitedkingdomreferencetables and on QM Plus

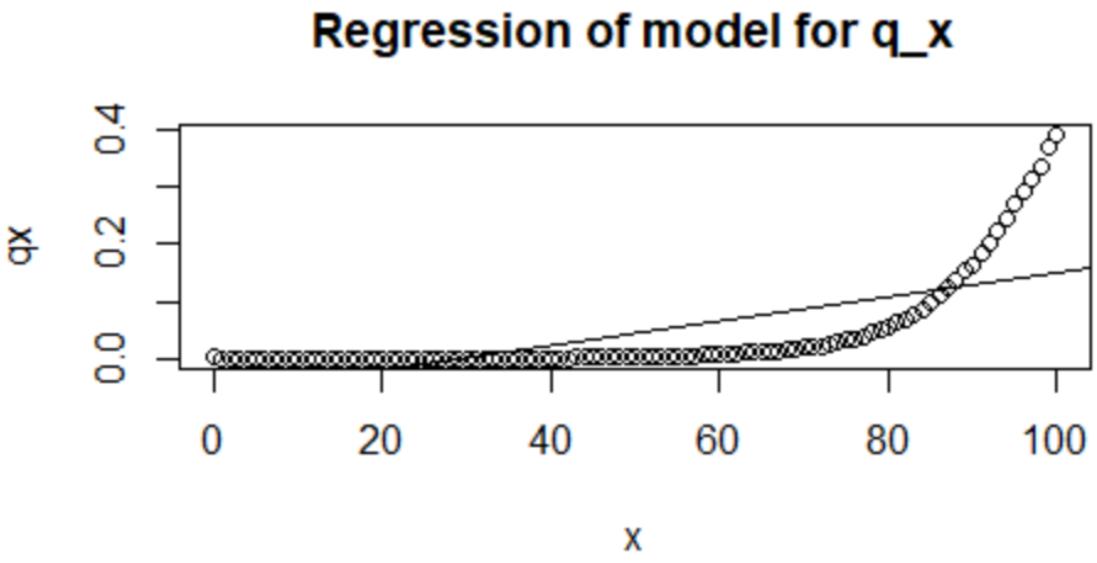
We import the data and calculate q_x

```
> NationalLifeTables2020 <- read.csv("…/NationalLifeTables2020.csv")
> x <- NationalLifeTables2020$age
> lx <- NationalLifeTables2020$lx
> dx <- NationalLifeTables2020$dx
> qx = dx / lx
```

First, we look for a simple linear regression model of q_x on age x

```
> model1 <- lm(qx ~ x)
> summary(model1)
Call:
lm(formula = qx ~ x)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.06815 -0.05007 -0.01187  0.03141  0.24061 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.060781   0.012617  -4.817 5.25e-06
x            0.002107   0.000218   9.666 5.85e-16

(Intercept) ***
x           ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06387 on 99 degrees of freedom
```

```
Multiple R-squared:  0.4855,	Adjusted R-squared:  0.4803 
F-statistic: 93.44 on 1 and 99 DF,  p-value: 5.85e-16
> plot(x,qx, main = "Regression of model for q_x")
> abline(model1)
```



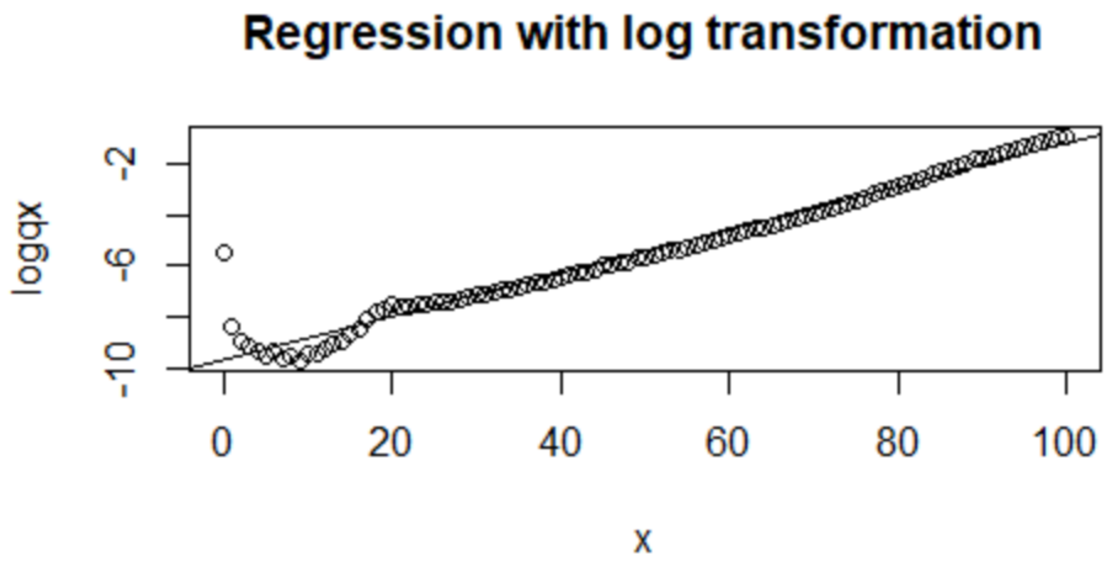
We immediately see a number of problems with a simple linear regression model:

- low R^2 of 48.6%
- a non-linear relationship in age
- negative fitted values for a probability (q_x) at young ages do not make sense
- fitted values that systematically underestimate then overestimate then underestimate

The obvious next step is to try a log transformation of the response variable

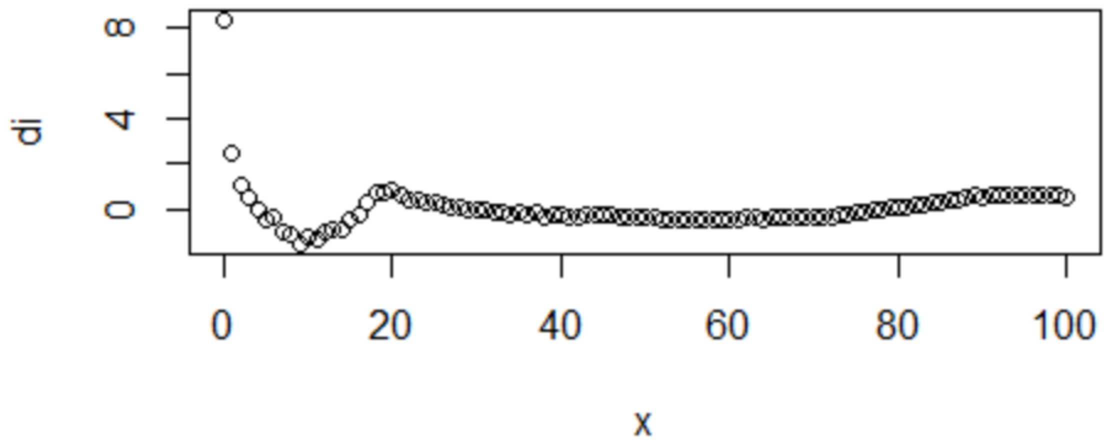
```
> logqx = log(qx)
> model2 <- lm(logqx ~ x)
> summary(model2)

Call:
lm(formula = logqx ~ x)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.7932 -0.2013 -0.1122  0.1694  4.2178 
```

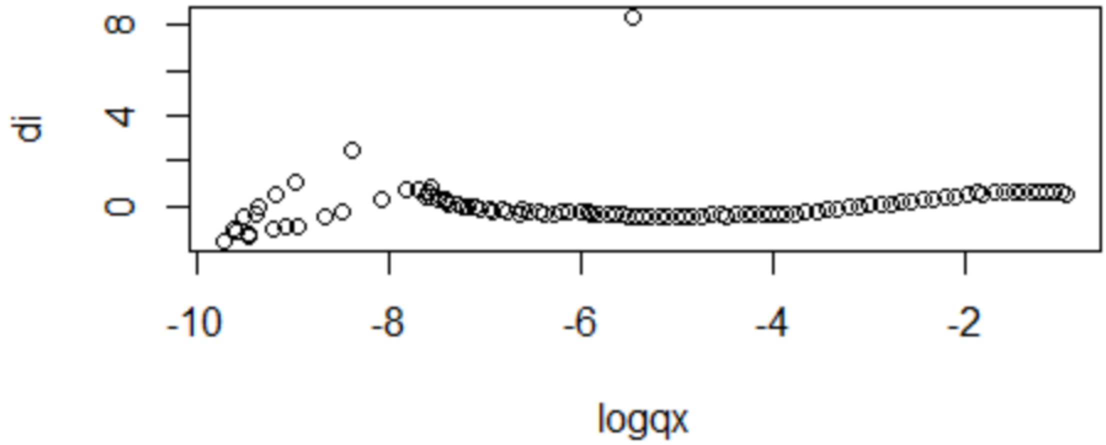
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.684780   0.101551  -95.37   <2e-16
x            0.084657   0.001755   48.25   <2e-16

(Intercept) ***
x           ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5141 on 99 degrees of freedom
Multiple R-squared:  0.9592,     Adjusted R-squared:  0.9588
F-statistic:  2328 on 1 and 99 DF,  p-value: < 2.2e-16
> plot(x,logqx, main = "Regression with log transformation")
> abline(model2)
```

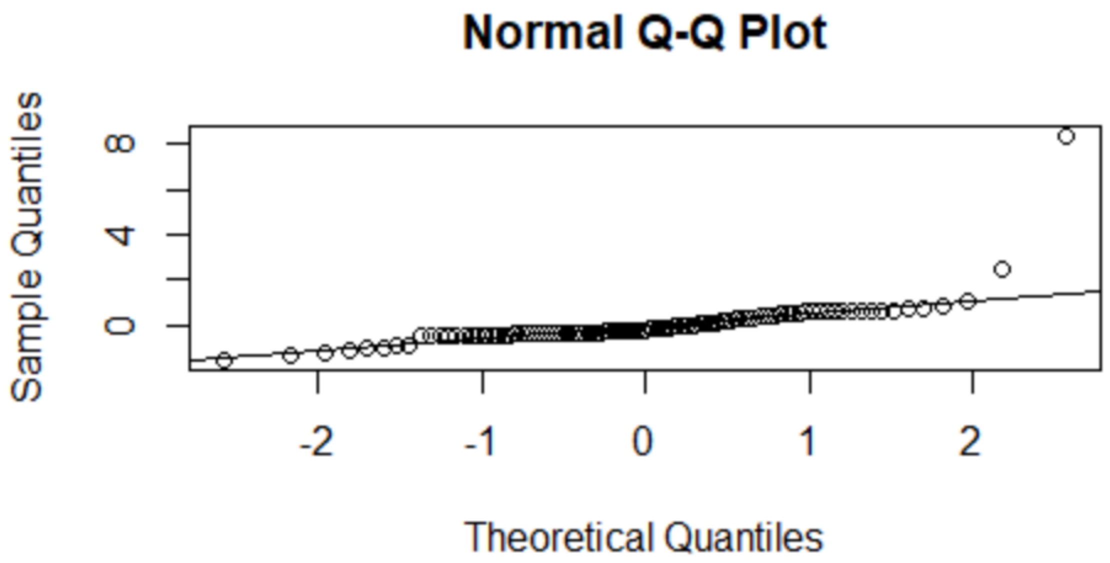


With R^2 now 96% this model looks more promising – so we will investigate its properties by calculating standardised residuals and examining their plots.

```
> di = rstandard(model2)
> plot(x,di)
> plot(logqx,di)
> qqnorm(di)
> qqline(di)
```







The residual plots cause us to doubt whether a regression model in log(q_x) is suitable:

- the pattern of standardised residuals is not random suggesting some elements of the relationship at least are not linear
- the residuals at the lower ages remain very high
- there is evidence that the distribution of residuals is not normal
- there remains some systematic underestimation of mortality at higher ages

These, combined with R^2 of 96% which whilst high in many modelling contexts is not large enough in many of the capital management or medical statistics contexts of survival modelling, suggest that we need to look beyond linear regression for our survival and mortality modelling in this module.