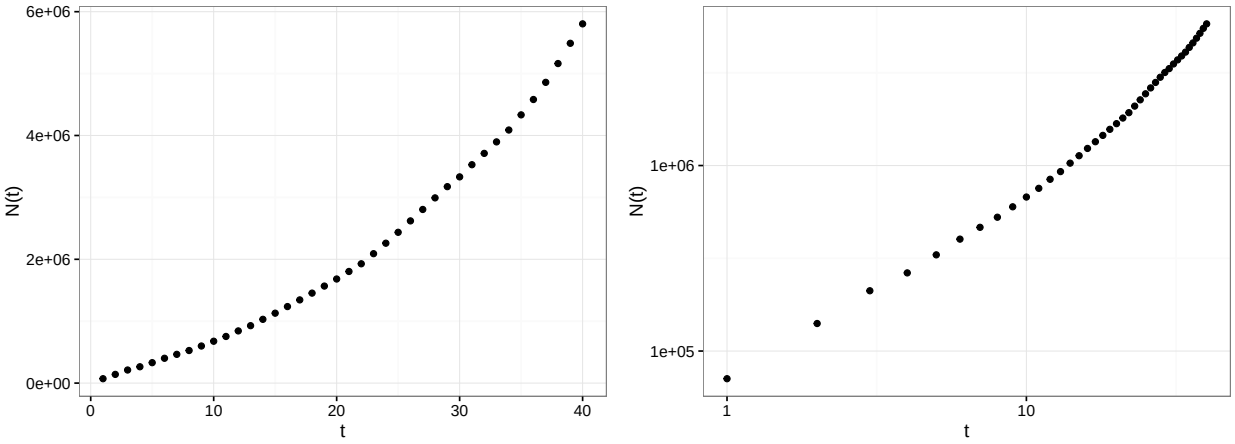
Curve Fitting Plot1 inset

In this notebook we explore the claim that the distribution of patent numbers over time follows a power law distribution.

As we saw in the reproducing valverde notebook the valverde suggests that the cumluative number of patents is a power law $N(t) \approx t^{\theta}$. From our fit and including more modern data we can see that this fit doesn't seem appropriate. The two plots below show the cdf of patents each year on a linear and log-log scale respectively.



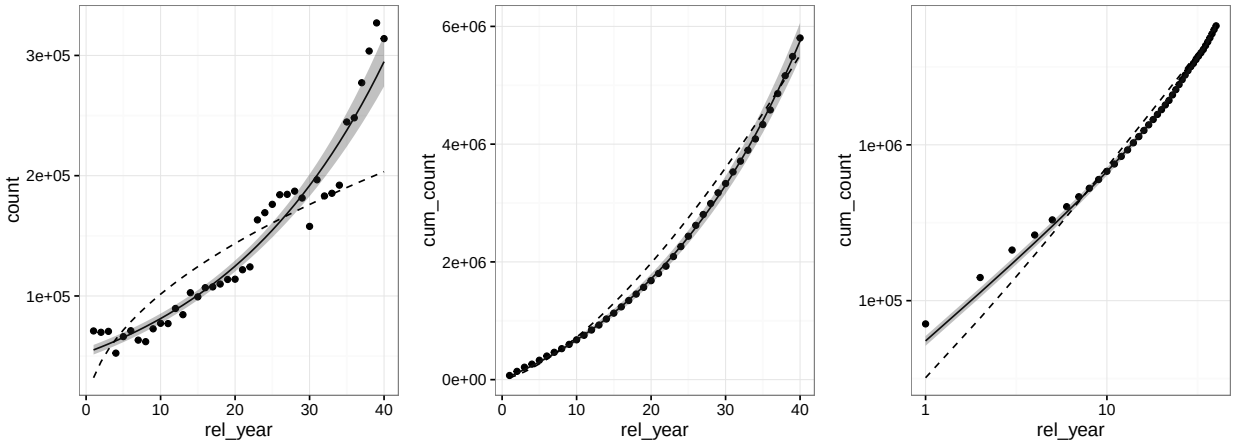
By transforming the data in different ways we can fit standard linear models to predict power law and exponential functions respectively: The exponential function clearly performs better with low p value and higher R squared.

```
##
## Call:
## lm(formula = log(count) ~ rel_year)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.194812 -0.090126 -0.006605  0.073276  0.251412
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.875091   0.037011  293.83   <2e-16 ***
## rel_year     0.042974   0.001573   27.32   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1149 on 38 degrees of freedom
## Multiple R-squared:  0.9515, Adjusted R-squared:  0.9503
## F-statistic: 746.2 on 1 and 38 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = log(count) ~ log(rel_year))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.38077 -0.20980 -0.03792  0.10411  0.79591
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   10.37357    0.14559  71.254  < 2e-16 ***
## log(rel_year)  0.50126    0.05038   9.949 3.93e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2748 on 38 degrees of freedom
## Multiple R-squared:  0.7226, Adjusted R-squared:  0.7153
## F-statistic: 98.99 on 1 and 38 DF,  p-value: 3.93e-12
```

We can visualise these fits onto the data. This clearly shows how the exponential function fits but more importantly when translated to non-cdf how the exponential function still nicely fits the data but the power law looks very odd and is only fitting to one cluster of data, it almost looks like a textbook example of high bias.



Questions

- Should I try to fit other distributions
- Is this enough analysis of this or should I do something more thorough? What would I do?
- Plots In report should I have plots like this? If so I will make them