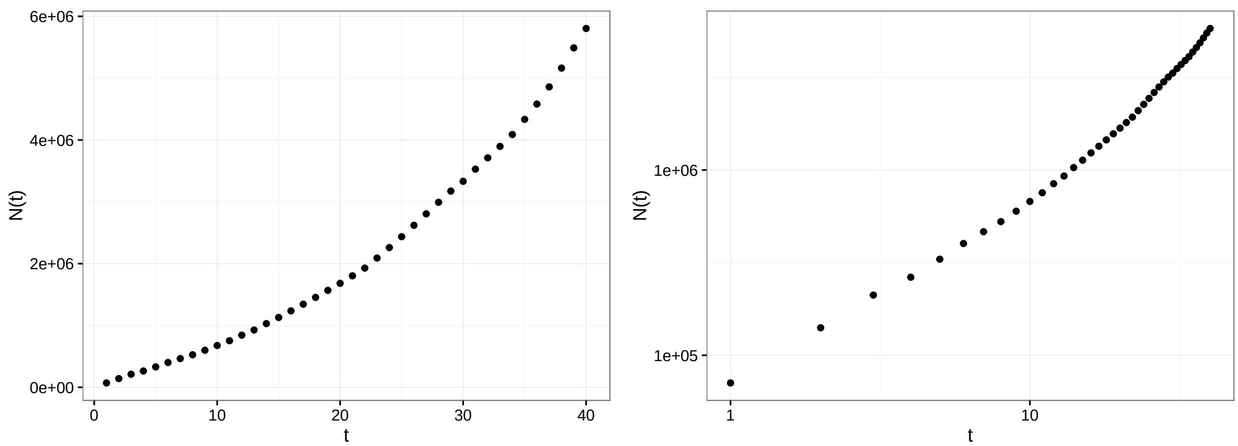
Curve Fitting Plot1 inset

```r
library(ggplot2)
library(gridExtra)
load("year_counts.RData")
```

As we saw in the reproducing valverde notebook the valverde suggests that the cumluative number of patents is a power law N(t) ~ t^theta. From our fit and including more modern data we can see that this fit doesn't seem appropriate. In this script we look at alternative distributions and see which best explains the distribution.

```r
gg1b <- ggplot(data = year_counts, aes(x = rel_year, y = cum_count)) +
    geom_point() +
    theme_bw() +
    labs(x = "t", y = "N(t)")

grid.arrange(gg1b, gg1b + scale_x_log10() + scale_y_log10(), ncol=2)
```




```r
attach(year_counts, warn.conflicts = FALSE)
exp.model <- lm(log(count) ~ rel_year)
summary(exp.model)
```

```
##
## 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
```

```r
pl.model <- lm(log(count) ~ log(rel_year))
summary(pl.model)
```

```
##
## 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
```

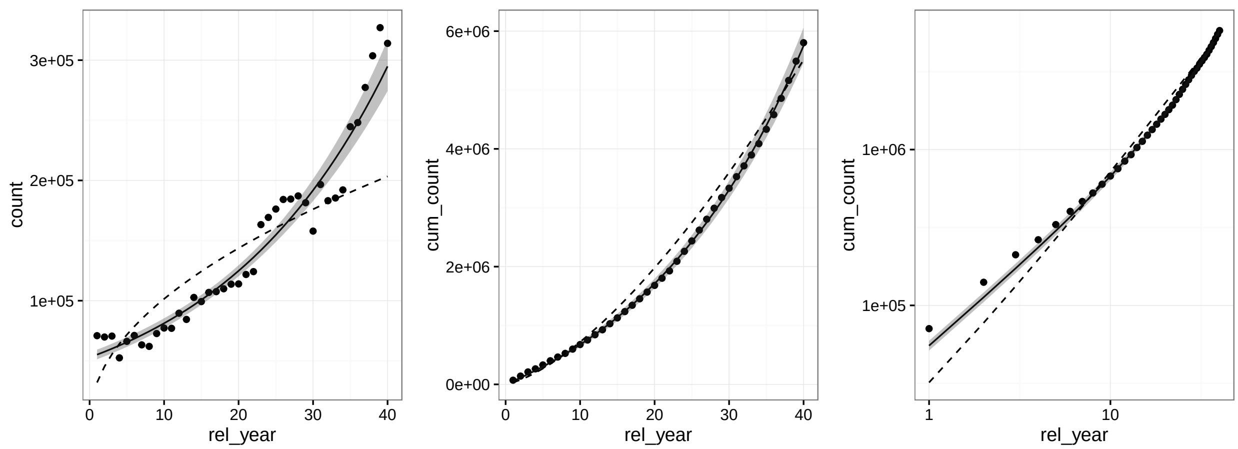
By fitting an exponential model we can see a much better fit, with lower p values and higher adjusted R squared.

```r
exp.pred <- exp(predict(exp.model, interval = "confidence", level = 0.95))
pl.pred <- exp(predict(pl.model, newdata = data.frame(rel_year = 1:40), interval = "confidence", level =

gg1 <- ggplot(data = year_counts, aes(x = rel_year, y = count)) +
    geom_point() +
    geom_line(aes(y = exp.pred[,"fit"])) +
    geom_ribbon(aes(ymax = exp.pred[,"upr"], ymin = exp.pred[,"lwr"]), alpha = 0.3) +
    theme_bw() +
    geom_line(aes(y = pl.pred[,"fit"], x = 1:40), linetype = "dashed")

gg <- ggplot(data = year_counts, aes(x = rel_year, y = cum_count)) +
    geom_point() +
    geom_line(aes(y = cumsum(exp.pred[,"fit"]))) +
    geom_ribbon(aes(ymax = cumsum(exp.pred[,"upr"]), ymin = cumsum(exp.pred[,"lwr"])), alpha = 0.3) +
    theme_bw() +
    geom_line(aes(y = cumsum(pl.pred[,"fit"])), linetype = "dashed")

grid.arrange(gg1, gg, gg + scale_x_log10() + scale_y_log10(), ncol = 3)
```



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

```
png("Figures/patentCountFit.png"); gg1; dev.off()
```

```
## pdf
##   2
```

```
png("Figures/patentCountFit_cum.png"); gg; dev.off()
```

```
## pdf
##   2
```

```
png("Figures/patentCountFit_cum_loglog.png"); gg + scale_x_log10() + scale_y_log10(); dev.off()
```

```
## pdf
##   2
```