Lab 7 /Assignment 3

Task 1

We interpret this task in the following way: For two seperate intervals of time (before 2012, and 2012 onwards), make linear learning curves determined by the fitted values of a simple linear regression of the log base 2 of the cumulative solar capacity and solar price.

Note that we include 2012 in the 'post' data as we found that the dataset stops at 2014. An exclusion of 2012 would have affected the results substantially, as there would have been too few datapoints to get a proper result.

Seperating pre and post 2012 data.

```r
df_pre_2012 = pv_df %>%  filter(year < 2012)
df_post_2012 = pv_df %>% filter(year >= 2012)
```

Creating the variable cum_cap, which calculates the cumulative sum of nameplate.

```r
df_pre_2012 = df_pre_2012 %>% arrange(date) %>% mutate(
  cum_cap = cumsum(nameplate)
)

df_post_2012 = df_post_2012 %>% arrange(date) %>% mutate(
  cum_cap = cumsum(nameplate)
)
```

Removing zero values

```r
df_pre_2012 = df_pre_2012 %>% filter(cost_per_kw != 0)
df_post_2012 = df_post_2012 %>% filter(cost_per_kw != 0)
```

Creating the variables *log2_cum_cap* and *log2_cost_per_kw*

```r
df_pre_2012["log2_cum_cap"] = log2(df_pre_2012$cum_cap)
df_pre_2012["log2_cost_per_kw"] = log2(df_pre_2012$cost_per_kw)

df_post_2012["log2_cum_cap"] = log2(df_post_2012$cum_cap)
df_post_2012["log2_cost_per_kw"] = log2(df_post_2012$cost_per_kw)
```

Signficance tests show that the capacity has a statistical signifcant correlation with the solar price. As we will see later, plotting the regression with observed data confirms this.

```r
learning_mod_pre = lm(log2_cost_per_kw~log2_cum_cap, data = df_pre_2012)
summary(learning_mod_pre)
```

```
##
## Call:
## lm(formula = log2_cost_per_kw ~ log2_cum_cap, data = df_pre_2012)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2249 -0.2209 -0.0404  0.2050  3.7030
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  13.855151   0.018761  738.51   <2e-16 ***
## log2_cum_cap -0.062382   0.001121  -55.62   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.3449 on 45087 degrees of freedom
## Multiple R-squared:  0.06422,    Adjusted R-squared:  0.0642
## F-statistic:  3094 on 1 and 45087 DF,  p-value: < 2.2e-16
```

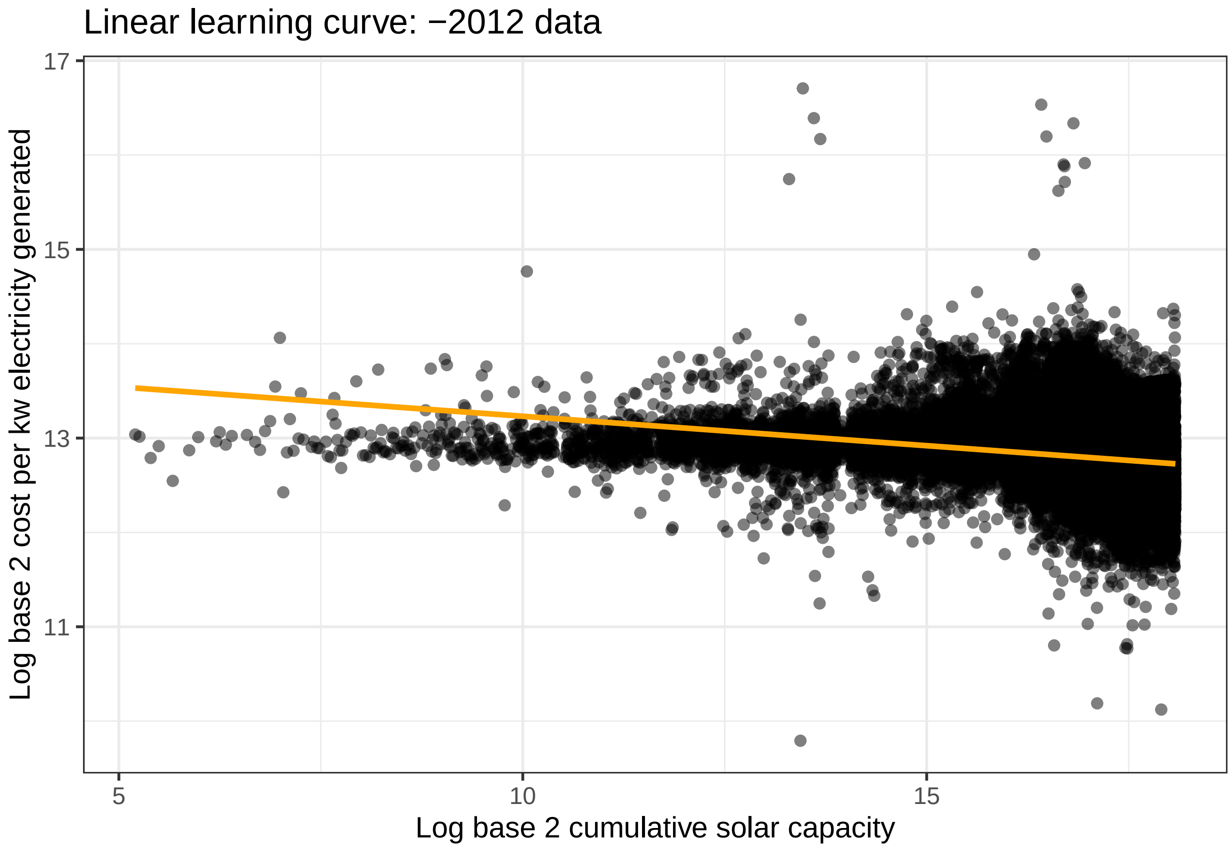
```
learning_mod_post = lm(log2_cost_per_kw~log2_cum_cap, data = df_post_2012)
summary(learning_mod_post)
```

```
##
## Call:
## lm(formula = log2_cost_per_kw ~ log2_cum_cap, data = df_post_2012)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2294 -0.1546  0.0079  0.1483  2.4925
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  13.7742858  0.0134914  1021.0   <2e-16 ***
## log2_cum_cap -0.0847891  0.0007834  -108.2   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.2983 on 61523 degrees of freedom
## Multiple R-squared:   0.16,  Adjusted R-squared:   0.16
## F-statistic: 1.172e+04 on 1 and 61523 DF,  p-value: < 2.2e-16
```

After splitting and calculating the log base 2 of the cumulative solar capacity and solar cost, we plot the fitted values of the linear regression with the observed values.

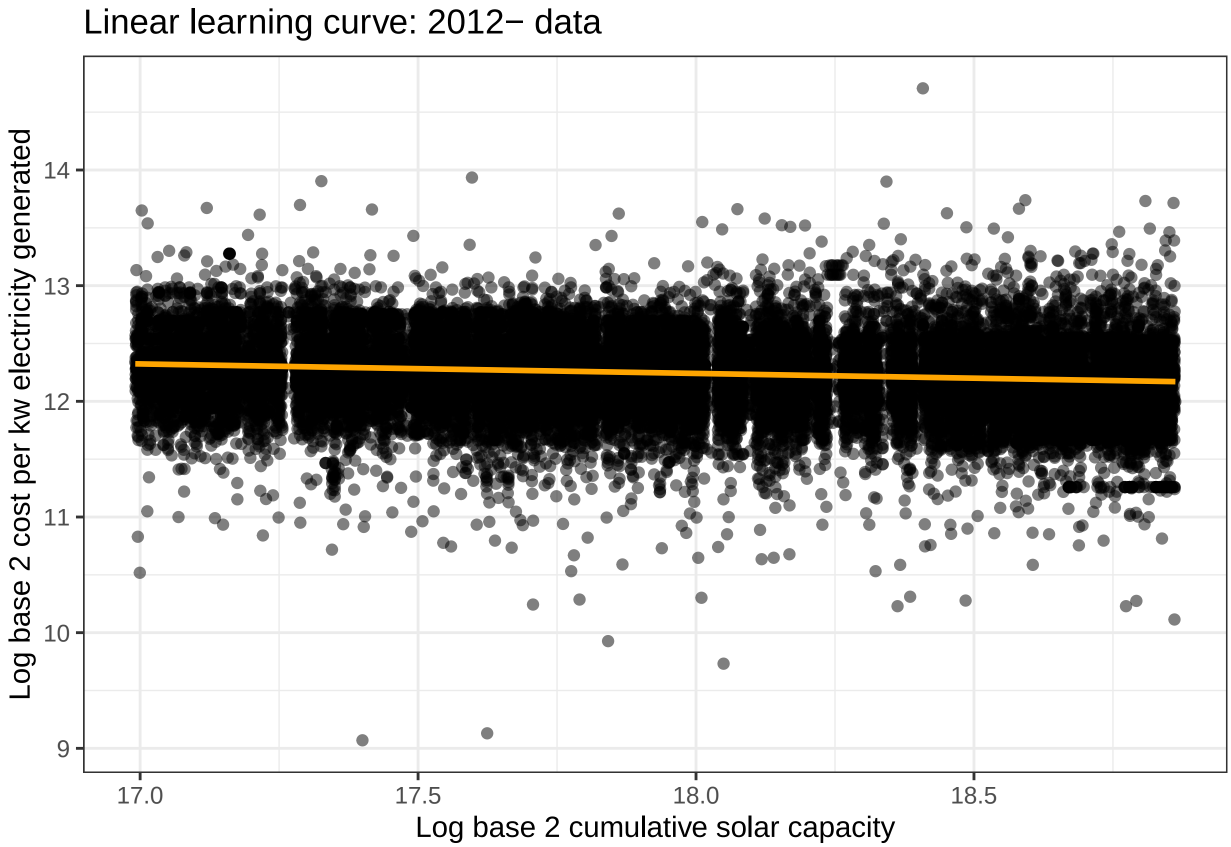
```
df_pre_2012  %>% filter(year < 2012)  %>%  ggplot(aes(x = log2_cum_cap, y = log2_cost_per_kw)) +
  geom_point(alpha=.5, color = "black") +
  geom_smooth(method = "lm", color = "orange") +
  labs(title = "Linear learning curve: -2012 data",
       x     = "Log base 2 cumulative solar capacity",
       y     = "Log base 2 cost per kw electricity generated") +
  theme_bw()
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



```
df_post_2012  %>% filter(year > 2012) %>%  ggplot(aes(x = log2_cum_cap, y = log2_cost_per_kw)) +
  geom_point(alpha=.5, color = "black") +
  geom_smooth(method = "lm", color = "orange") +
  labs(title = "Linear learning curve: 2012- data",
       x     = "Log base 2 cumulative solar capacity",
       y     = "Log base 2 cost per kw electricity generated") +
  theme_bw()
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



By using simple linear regression we avoid overfitting, as higher-order polynomials might incurr. While, our estimates does not overfit, we do in all likelihood suffer from signifcant bias as we assume a linear relationship between the cumulative solar capacity and solar price. In reality, a linear relationship is very unlikely and as a result a linear model will give poor predictive performance.

We do not however, capture large amount of the information present within the data. As we can see from the plots above, there is signifcant variation and this variation will only be accounted for to a small extent.

Task 2