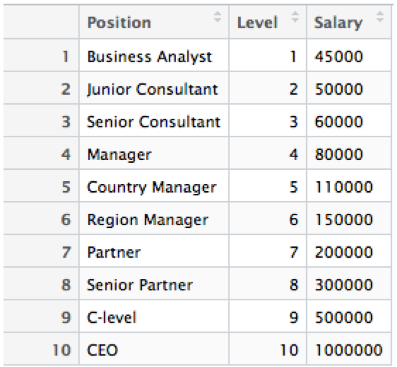
Polynomial Regression Vs Linear Regression

I desire to show that in some cases using the polynomial regression model is far more accurate than a linear regression model. To do this I am using the data from *Position_Salaries.csv* and have created a hypothetical situation.

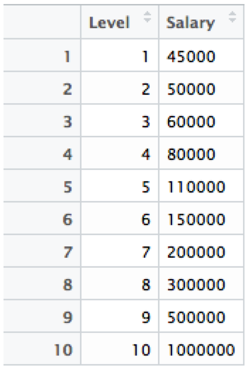
A new employee is hired and HR is wanting to know what they should pay this employee. This new employee says in his previous position he was being paid $160,000.00. HR knows he was between level 6 and 7. HR has a list of ten levels of which the average pay is given for each level. It takes 4 years to move up a level, and the new employee has 20 years of experience.

The machine learning model (polynomial regression) will be able to learn the correlations between *Level* and *Salary* and thus will be able to show the rate of growth for employee wages across the levels.

Our data set has only 10 rows. The *Position* column is removed as it is not needed.



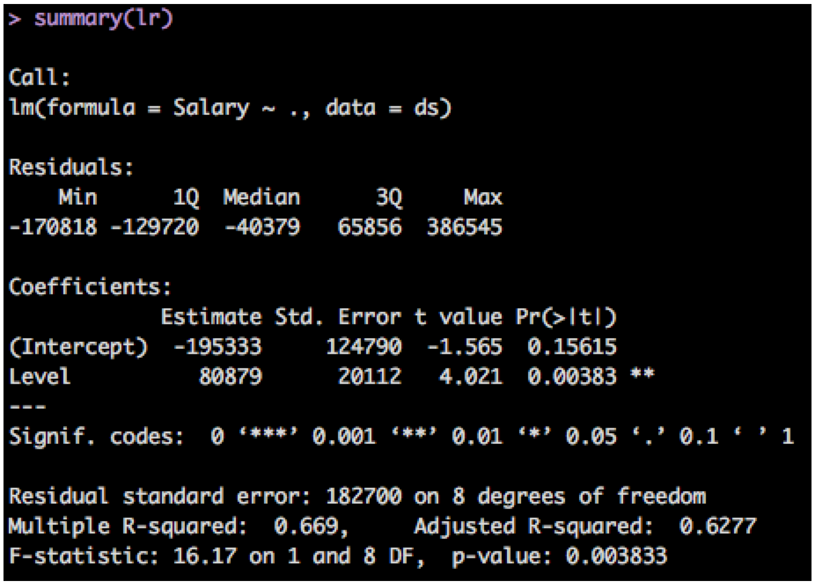
	Position	Level	Salary
1	Business Analyst	1	45000
2	Junior Consultant	2	50000
3	Senior Consultant	3	60000
4	Manager	4	80000
5	Country Manager	5	110000
6	Region Manager	6	150000
7	Partner	7	200000
8	Senior Partner	8	300000
9	C-level	9	500000
10	CEO	10	1000000



	Level	Salary
1	1	45000
2	2	50000
3	3	60000
4	4	80000
5	5	110000
6	6	150000
7	7	200000
8	8	300000
9	9	500000
10	10	1000000

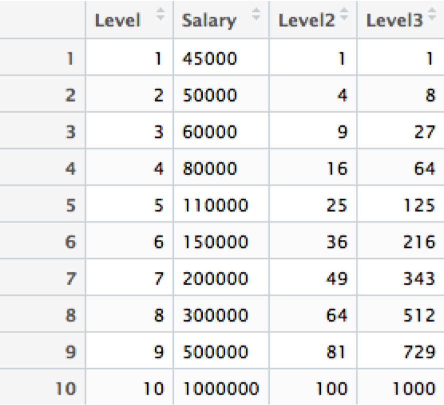
Because the dataset is so small it didn’t make sense to split it into a training and test set. We wanted to have the maximum amount of data in order to achieve the most accurate prediction.

After fitting the data set to a linear regression model we achieve the following results (below).



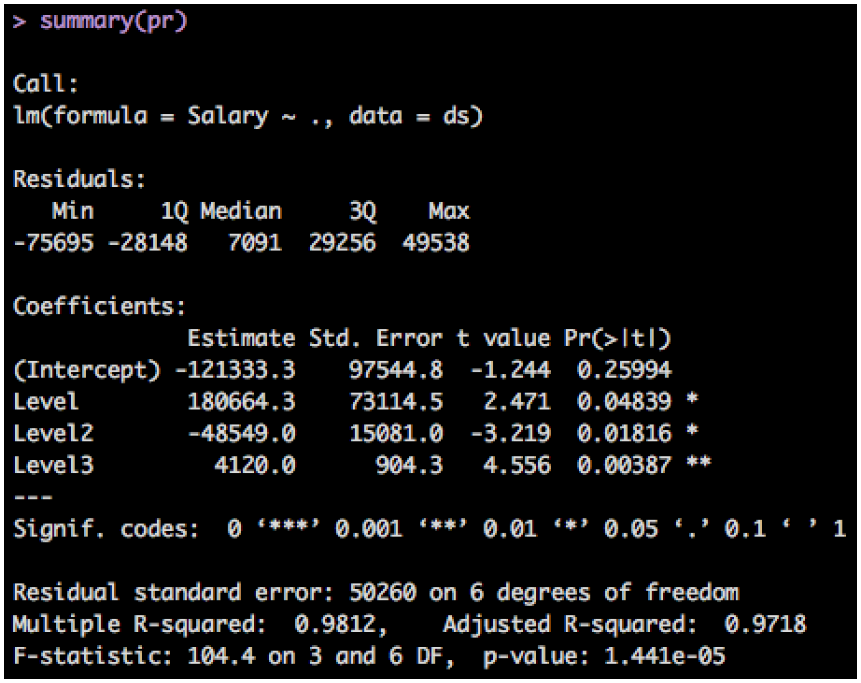
```
> summary(lr)

Call:
lm(formula = Salary ~ ., data = ds)

Residuals:
    Min      1Q  Median      3Q     Max 
-170818 -129720  -40379   65856  386545 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)  -195333     124790  -1.565  0.15615   
Level          80879      20112   4.021  0.00383 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 182700 on 8 degrees of freedom
Multiple R-squared:  0.669,	Adjusted R-squared:  0.6277 
F-statistic: 16.17 on 1 and 8 DF,  p-value: 0.003833
```

To obtain the polynomial regression model we must increase the independent variables by some factor. To do this I simply added several extra columns where the value had been squared and cubed.



	Level	Salary	Level2	Level3
1	1	45000	1	1
2	2	50000	4	8
3	3	60000	9	27
4	4	80000	16	64
5	5	110000	25	125
6	6	150000	36	216
7	7	200000	49	343
8	8	300000	64	512
9	9	500000	81	729
10	10	1000000	100	1000

Fitting this dataset to polynomial regression resulted in the following.

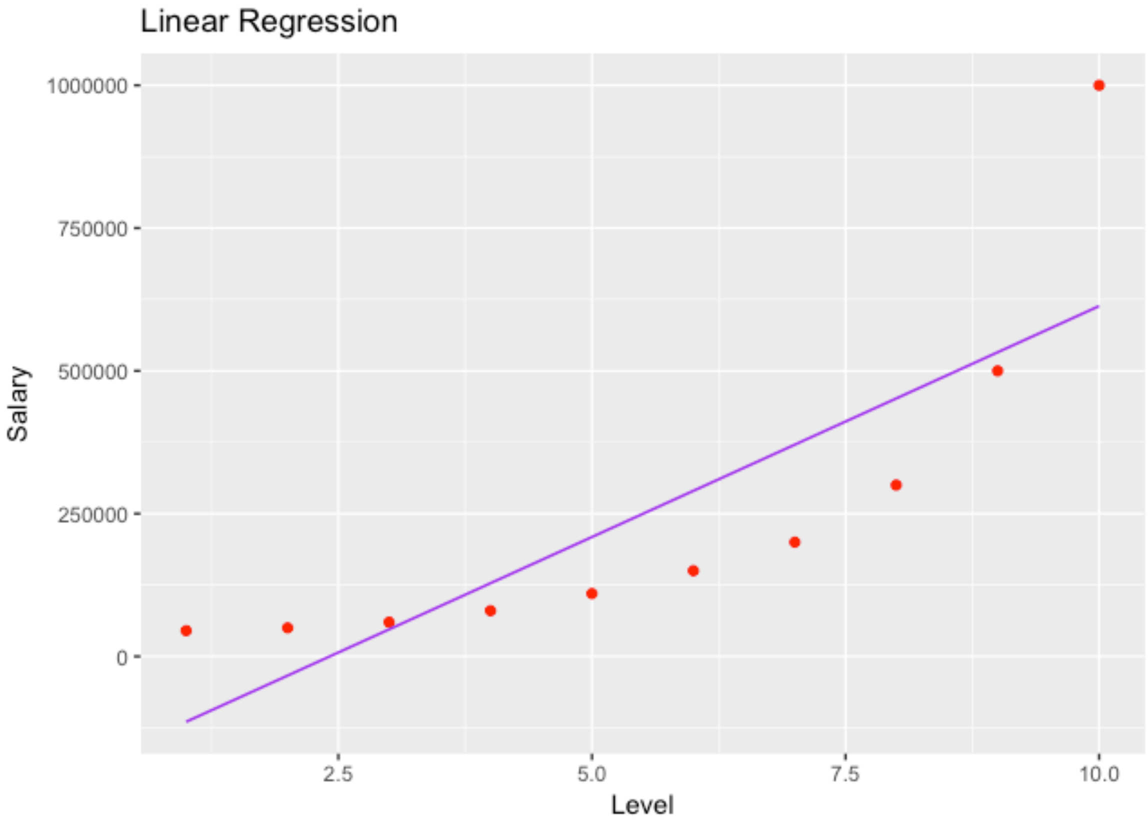


```
> summary(pr)

Call:
lm(formula = Salary ~ ., data = ds)

Residuals:
   Min     1Q Median     3Q    Max 
-75695 -28148   7091  29256  49538 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)   
(Intercept) -121333.3    97544.8  -1.244  0.25994   
Level        180664.3    73114.5   2.471  0.04839 * 
Level2       -48549.0    15081.0  -3.219  0.01816 * 
Level3         4120.0      904.3   4.556  0.00387 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 50260 on 6 degrees of freedom
Multiple R-squared:  0.9812,	Adjusted R-squared:  0.9718 
F-statistic: 104.4 on 3 and 6 DF,  p-value: 1.441e-05
```

Visualizing the data graphically

We created graphs in order to better visualize the difference between linear and polynomial regression

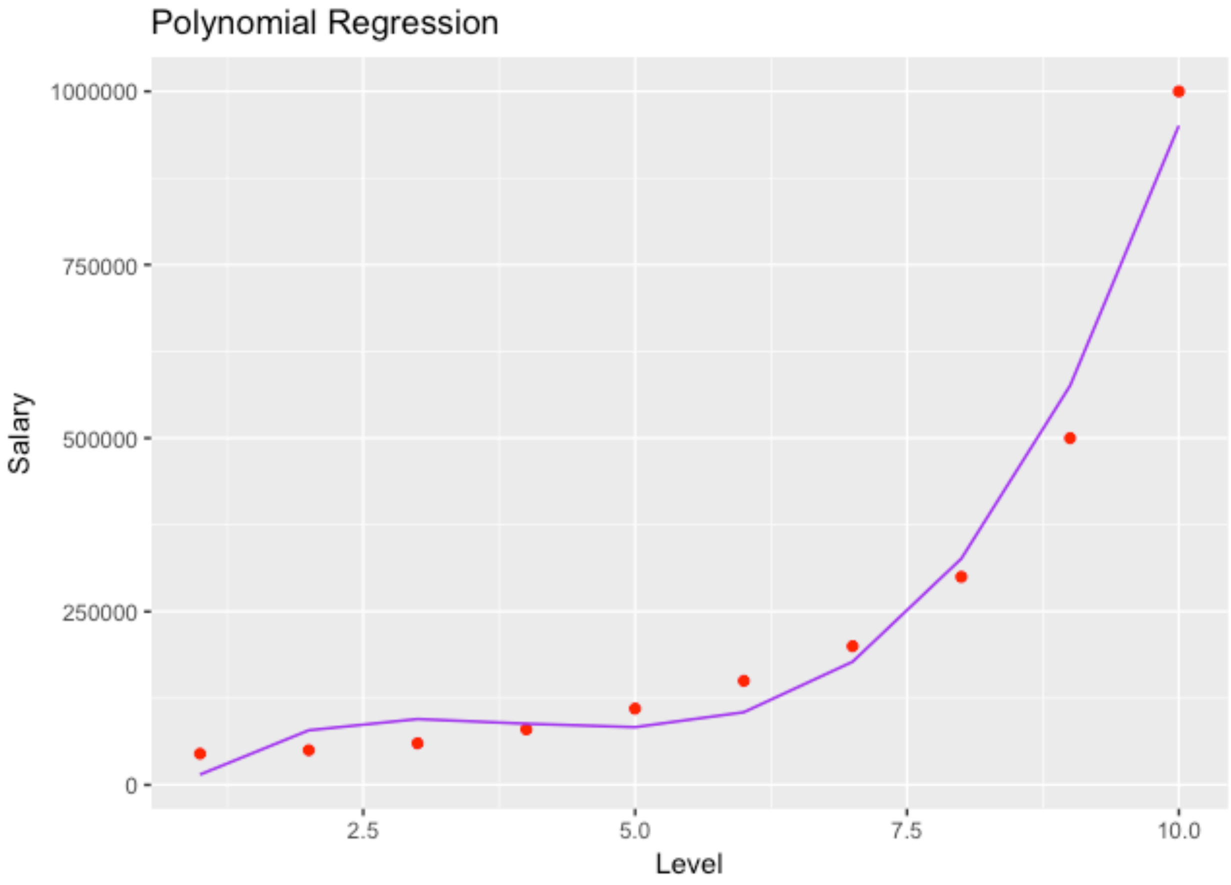
Linear Regression



The red dots are the actual salary while the purple line is the predicted salary using the linear regression model of machine learning.

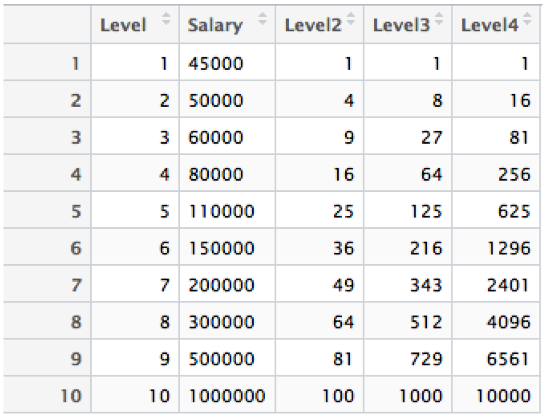
Clearly this model is not at all accurate for this type of data. If we were to take this model and apply it to our hypothetical situation the new employee would be receiving around $300,000.00 annually.

<u>Polynomial Regression</u>



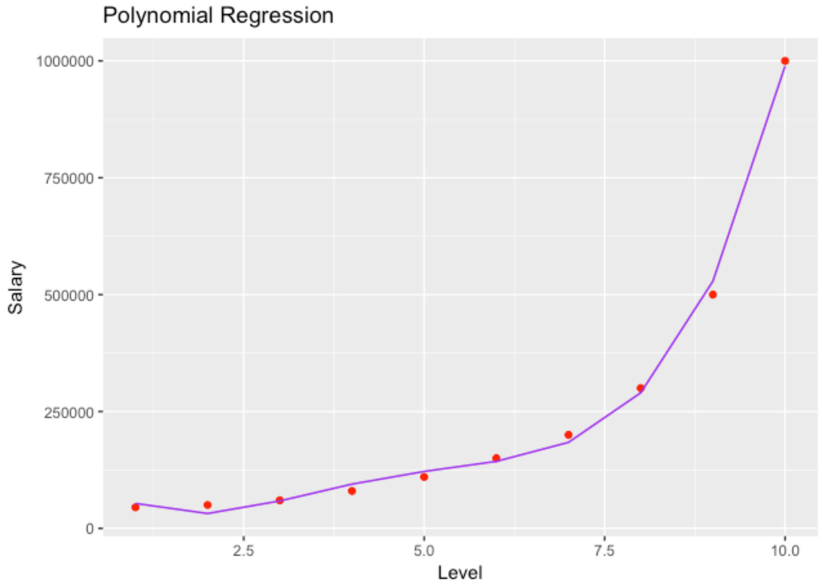
The red dots are the actual salary while the purple line is the predicted salary using the polynomial regression model of machine learning.

As we can see from this model the prediction is much more accurate. To make it even more accurate we can add another two degrees to the table.

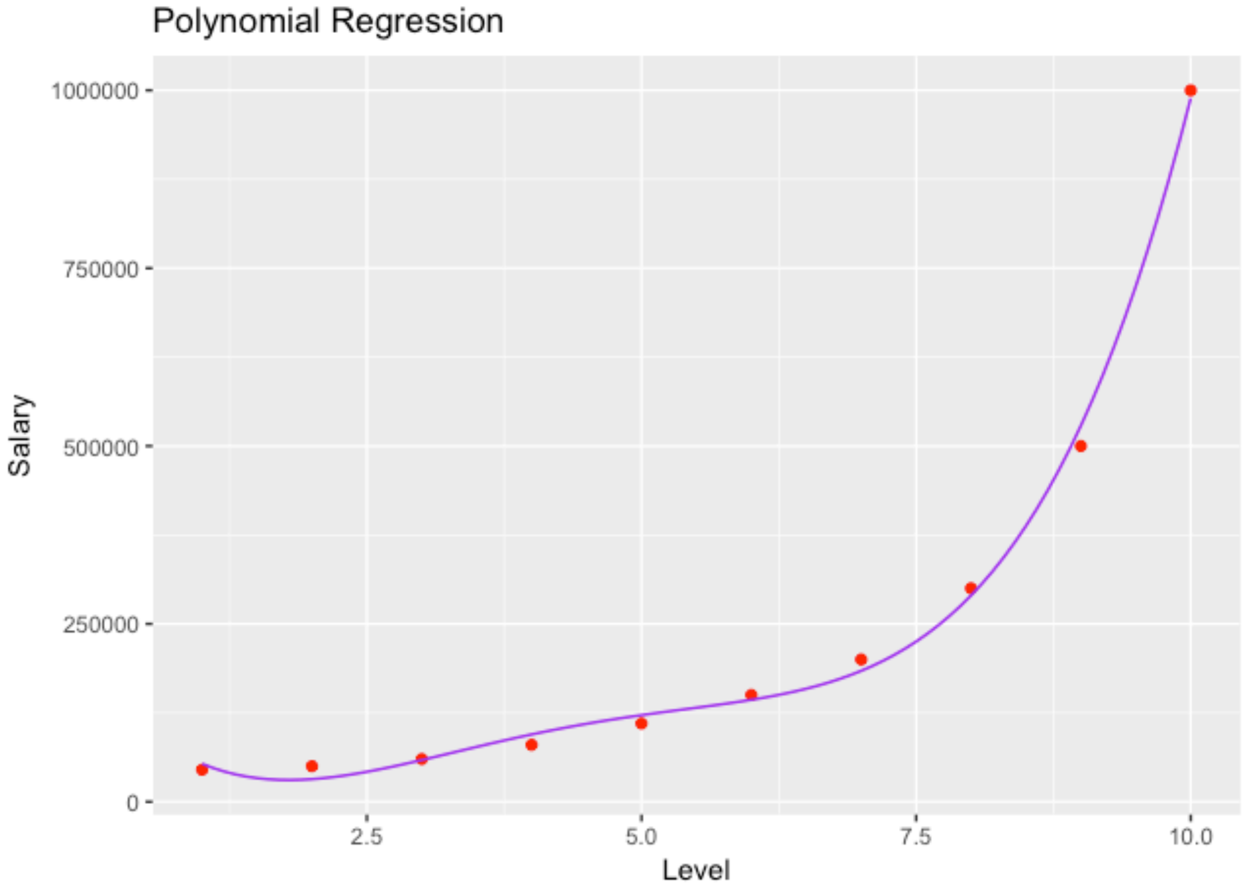


	Level	Salary	Level2	Level3	Level4
1	1	45000	1	1	1
2	2	50000	4	8	16
3	3	60000	9	27	81
4	4	80000	16	64	256
5	5	110000	25	125	625
6	6	150000	36	216	1296
7	7	200000	49	343	2401
8	8	300000	64	512	4096
9	9	500000	81	729	6561
10	10	1000000	100	1000	10000

This results in a graph that is even more accurate.



After smoothing out the lines



Validating the Model

The graph gives a good visual representation, but to show an exact representation we will predict a specific value. If we assume the new employee has a level of 6.5 we return the following results.

Linear Regression: 330379
Polynomial Regression: 158862

Therefore, we can conclude that the new employee's salary should be around $160,000.00 annually, but more importantly that in this case we have shown polynomial regression is far more accurate than linear regression.