Predicting Research Citations

NYU Stern Program for Undergraduate Research (SPUR)
Faculty: Professor Panos Ipeirotis
Student Research: Rachel Lu

Content

1. Summary Statistics
2. Exploratory Graphs
3. Linear Regression Models
4. Logistical Regression Models
5. Conclusion

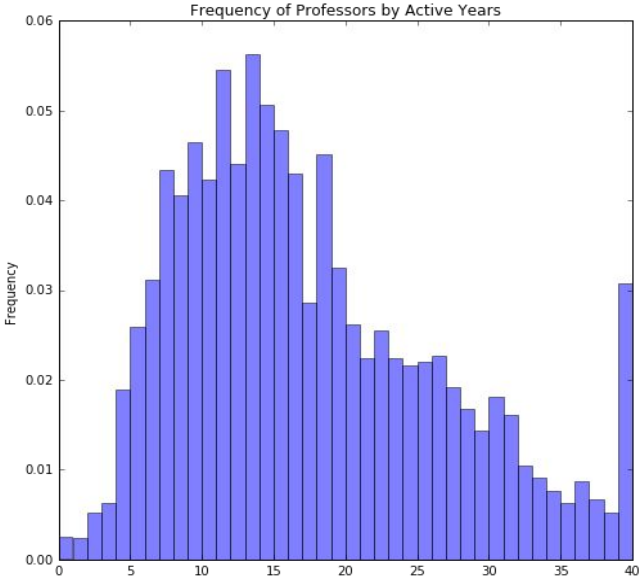
1. Summary Statistics

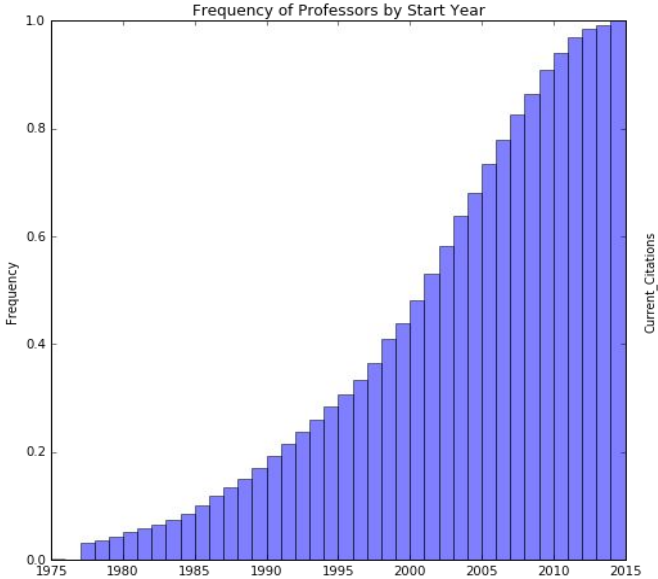
Data at a Glance

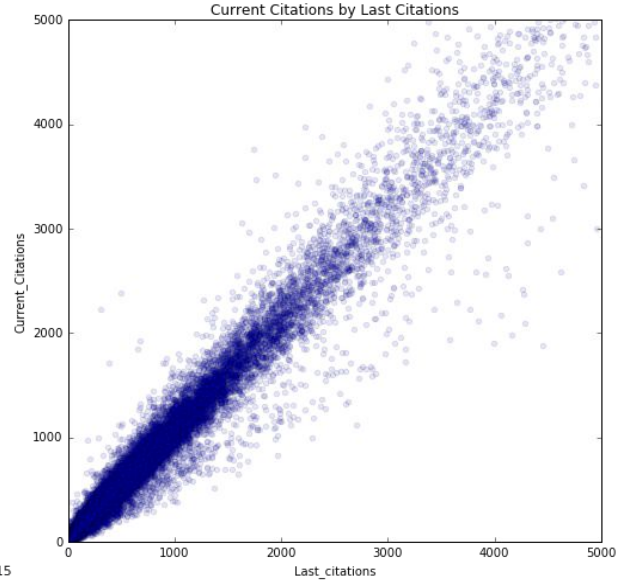
- 2861 Total Researchers
- Mean Years Active: 17 Years
- Mean Start Year: 1999
- Mean Number of Current Citations: 37

	Start	Years Active	Year	Current_Citations	Age	Last_citations	Last2_citations
count	2861.000000	2861.000000	2861.000000	2861.000000	2861	0	0
mean	1998.932192	17.039846	1998.932192	37.201678	0	NaN	NaN
std	9.050265	9.082517	9.050265	79.194195	0	NaN	NaN
min	1977.000000	0.000000	1977.000000	1.000000	0	NaN	NaN
25%	1993.000000	10.000000	1993.000000	4.000000	0	NaN	NaN
50%	2001.000000	15.000000	2001.000000	15.000000	0	NaN	NaN
75%	2006.000000	23.000000	2006.000000	39.000000	0	NaN	NaN
max	2015.000000	39.000000	2015.000000	1945.000000	0	NaN	NaN

Descriptive Data



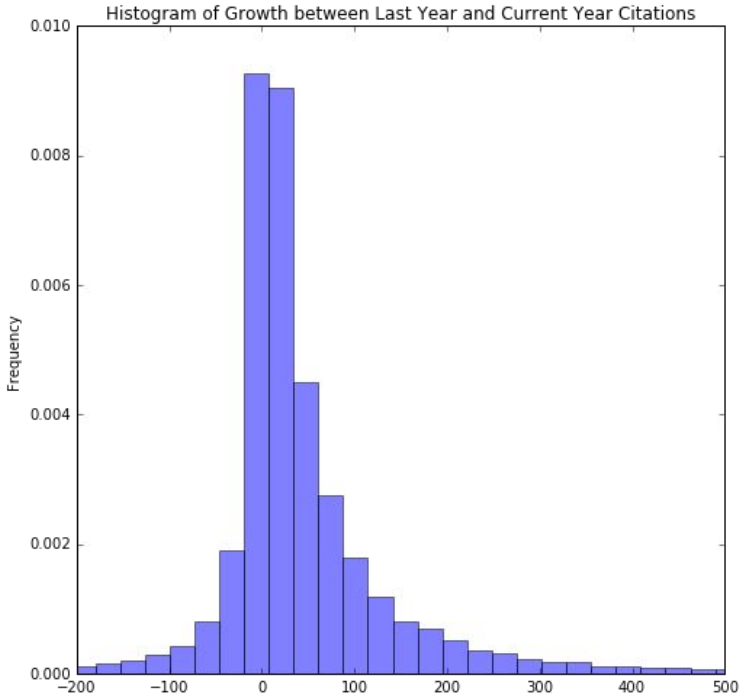


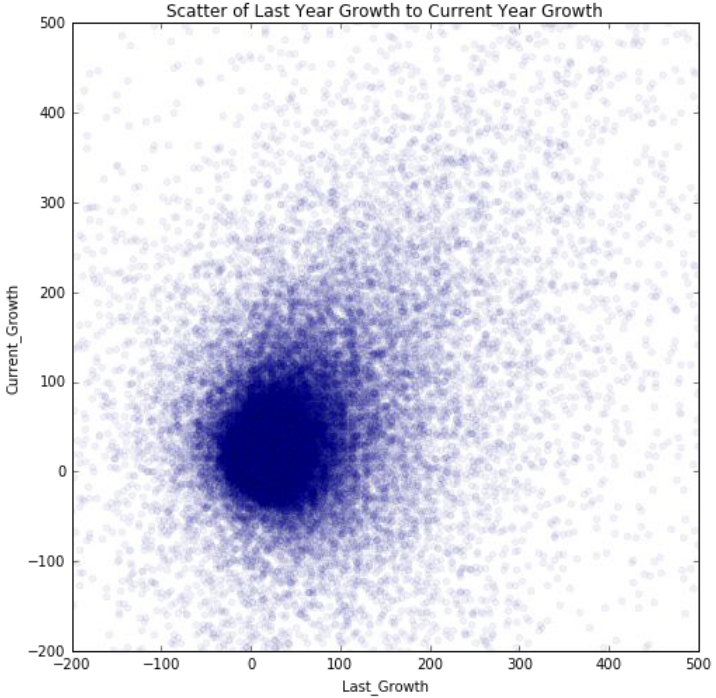


Frequency of Professors by Active Years is skewed slightly to the left, indicating that most researchers in the data have been active for 8 to 18 years.

2. Exploratory Graphs

Citation Growth





Frequency of Professors by Active Years is skewed slightly to the left, indicating that most researchers in the data have been active for 8 to 18 years.

3. Linear Regression Models

Initial Linear Regression Tests

formula = Current_Citations ~ Last_citations

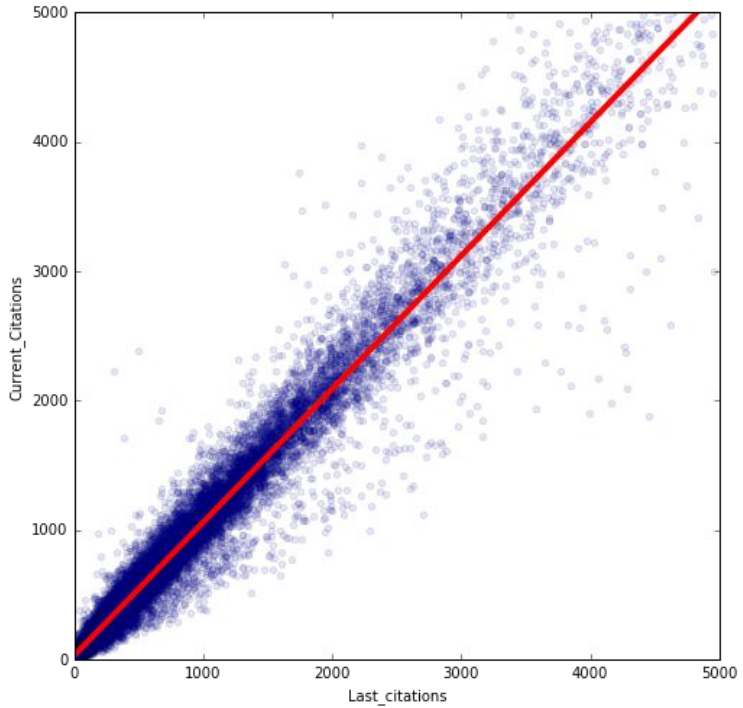
OLS Regression Results

Dep. Variable:	Current_Citations	**R-squared:**	0.974
Model:	OLS	**Adj. R-squared:**	0.974
Method:	Least Squares	**F-statistic:**	1.791e+06
Date:	Sun, 08 May 2016	**Prob (F-statistic):**	0.00
Time:	21:45:45	**Log-Likelihood:**	-3.2477e+05
No. Observations:	48374	**AIC:**	6.495e+05
Df Residuals:	48372	**BIC:**	6.496e+05
Df Model:	1		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	28.1291	0.996	28.243	0.000	26.177 30.081
Last_citations	1.0311	0.001	1338.16	0.000	1.030 1.033

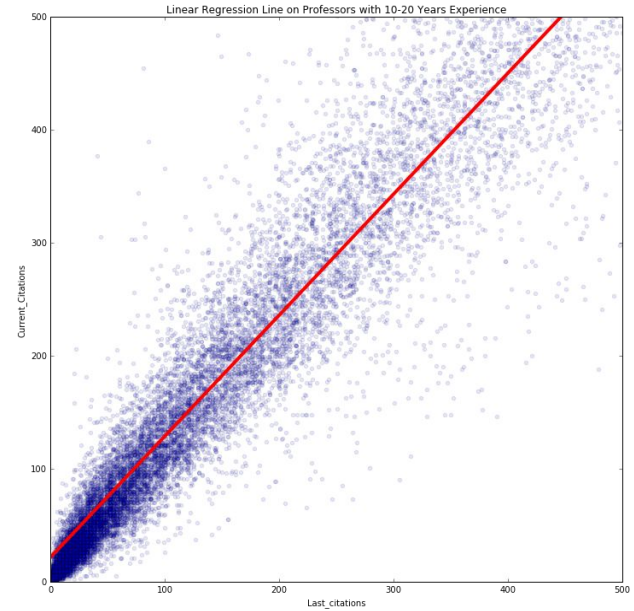
Omnibus:	72255.503	**Durbin-Watson:**	1.460
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	257235124.322
Skew:	-8.381	**Prob(JB):**	0.00
Kurtosis:	359.850	**Cond. No.**	1.42e+03

- Last_citations looks statistically significant in determining the variable Current_Citations
- 0.974 R-squared value is another promising sign

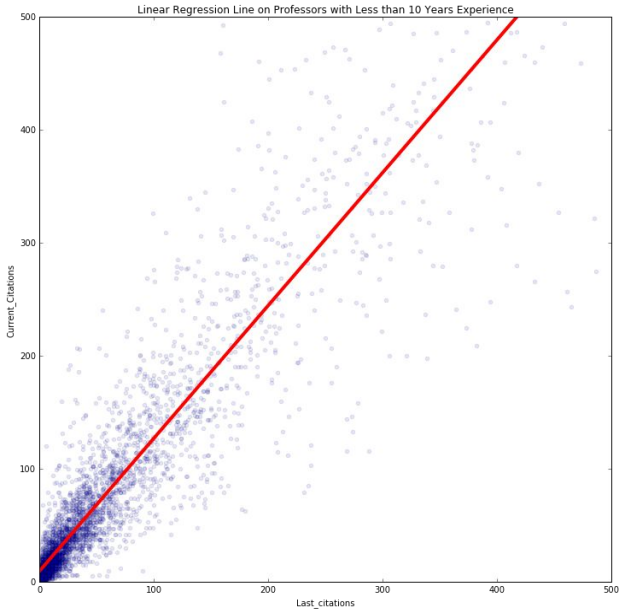


Initial Linear Regression Plots

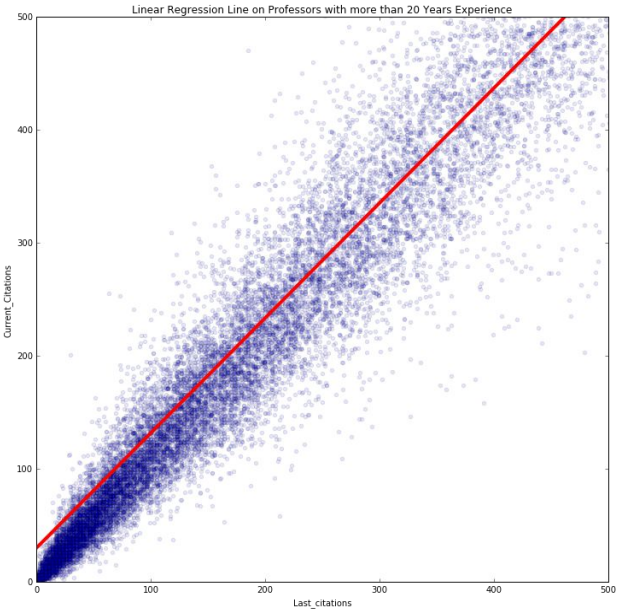
Less than 10 Years Active



Between 10-20 Years Active



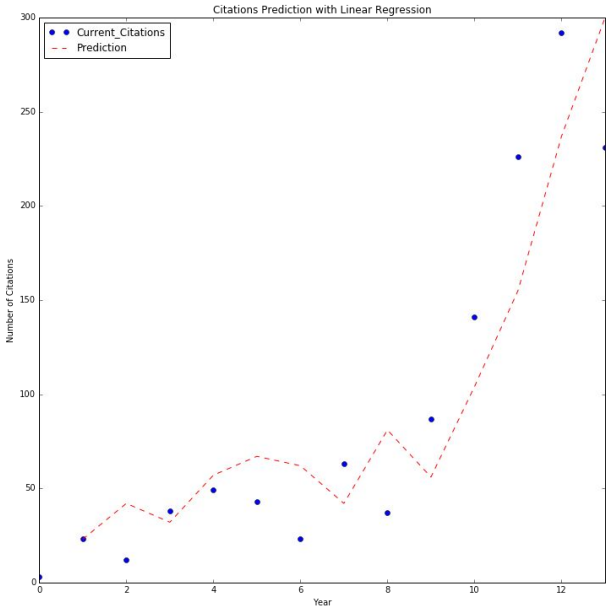
More than 20 Years Active



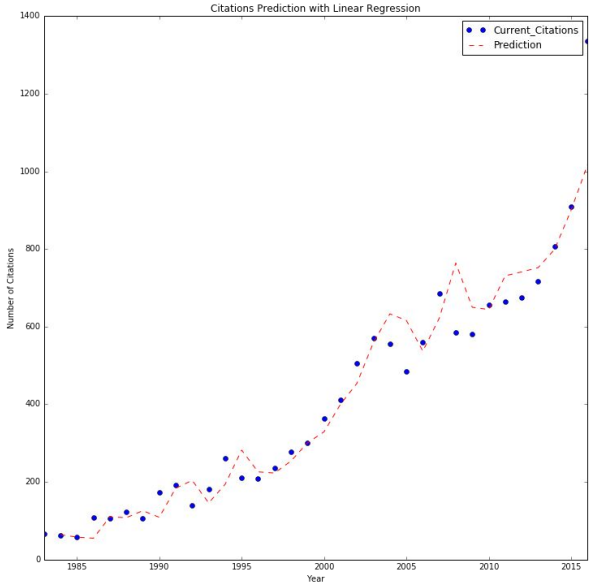
There is lot of variability in comparing current citations to last citations, but there is still a visible upward trend.

Prediction with Linear Regression

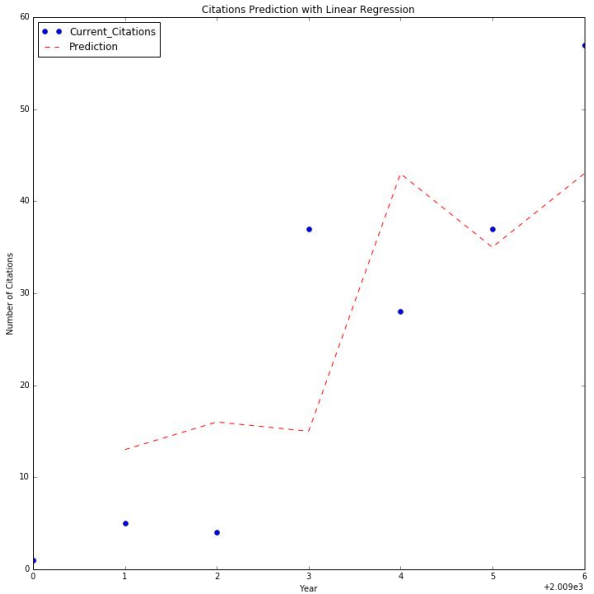
Paulo Blikstein: 13 Years Active



Robert Anderson: 33 Years Active



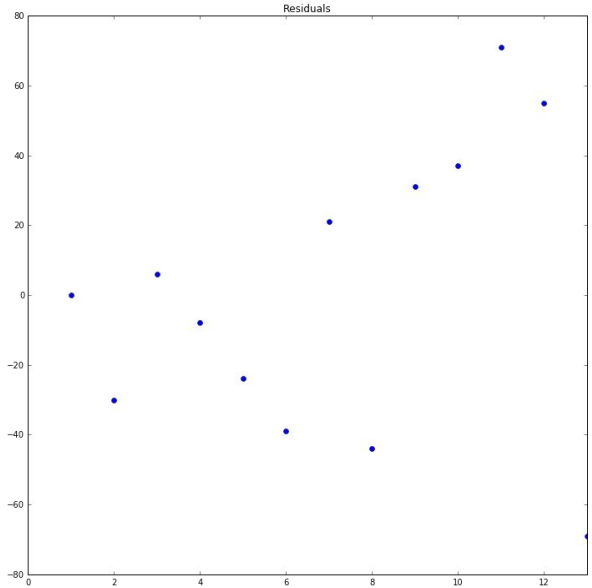
Yiping Zhu: 6 Years Active



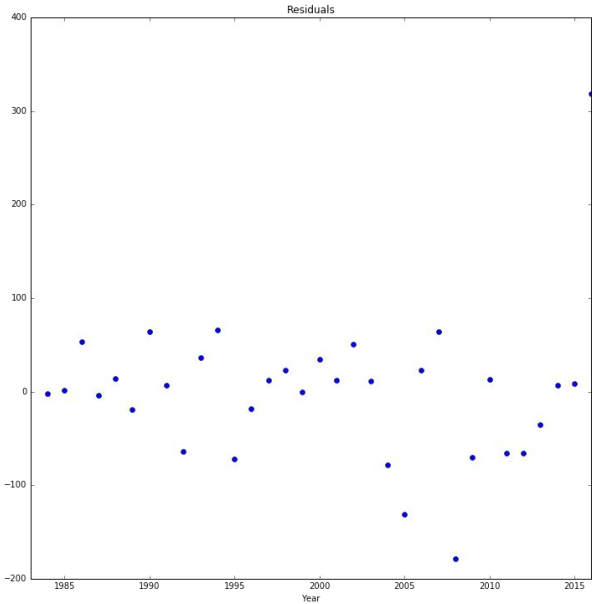
The linear regression model's fitted values are plotted along with the current citations. The OLS results revealed the highest r-squared value for Robert Anderson and lowest r-squared value with Yiping Zhu. The P-value of of Yiping Zhu was 0.128, suggesting that Last_citations is not statistically significant in determining Current_Citations, likely because 6 years is not enough data for an accurate prediction.

Linear Regression Residuals

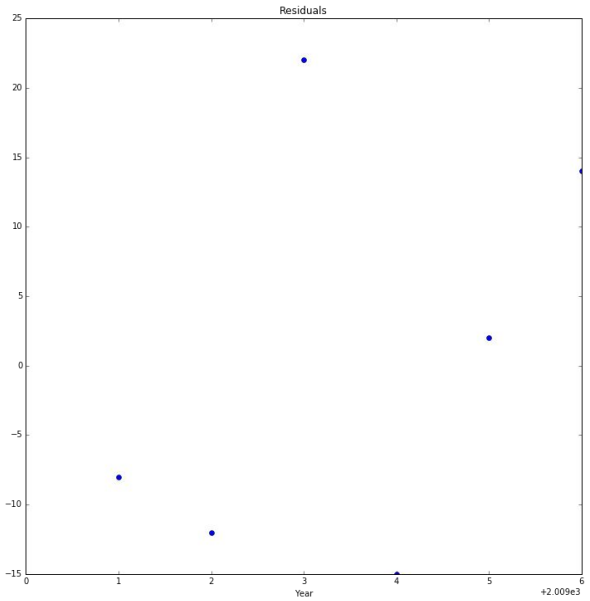
Paulo Blikstein: 13 Years Active



Robert Anderson: 33 Years Active



Yiping Zhu: 6 Years Active



The residual plots appear to look random for Blikstein and Anderson, but difficult to really make a call visually for Zhu. There may be a pattern in the residuals for Zhu, but it is difficult with only 6 years of data.

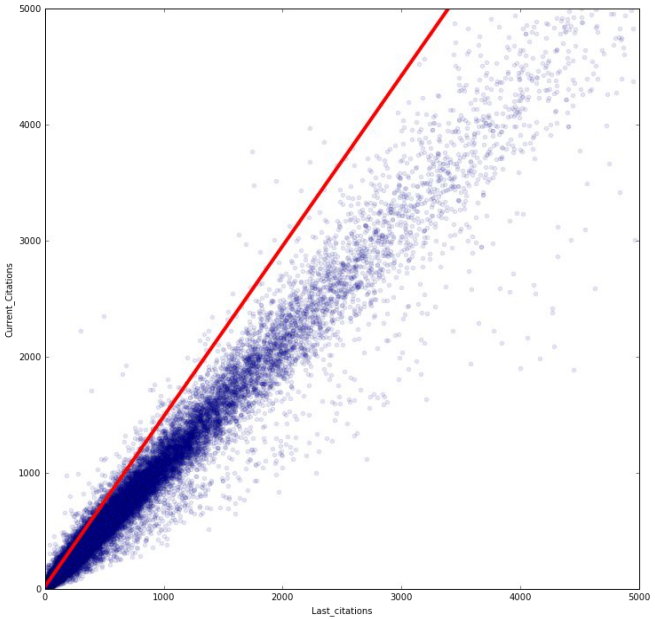
Testing Multiple Variables

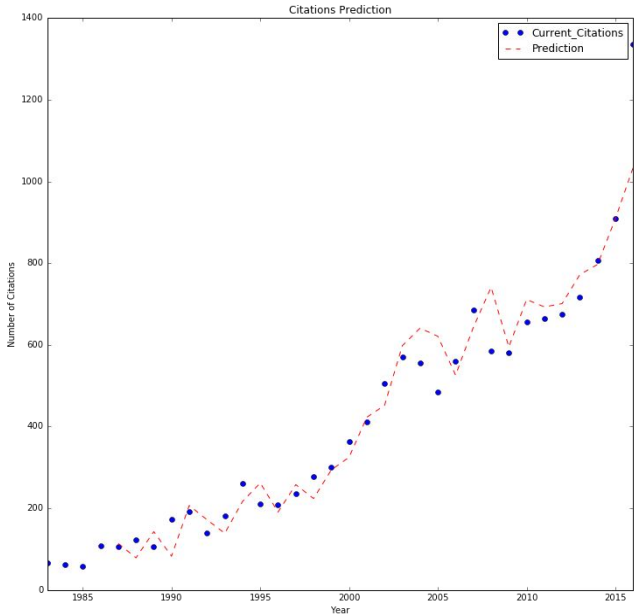
Adding additional variables (more years of citations) to tests its correlation with Current_Citations **did not improve the model**, even though the r-squared value went up slightly and all variables seem statistically significant. The plot in the middle is the linear regression line against entire dataset, and the plot to the far right is the predictions for one professor, Robert Anderson.

Dep. Variable:	Current_Citations	**R-squared:**	0.980
Model:	OLS	**Adj. R-squared:**	0.980
Method:	Least Squares	**F-statistic:**	3.549e+05
Date:	Sun, 08 May 2016	**Prob (F-statistic):**	0.00
Time:	23:37:44	**Log-Likelihood:**	-2.4796e+05
No. Observations:	37038	**AIC:**	4.959e+05
Df Residuals:	37032	**BIC:**	4.960e+05
Df Model:	5		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	17.1569	1.151	14.903	0.000	4.900 19.413
Last_citations	1.4691	0.008	176.09	0.000	1.453 1.485
Last2_citations	0.0415	0.014	2.882	0.004	0.013 0.070
Last3_citations	-0.4935	0.015	-32.13	0.000	-0.524 -0.463
Last4_citations	-0.1015	0.017	-6.039	0.000	-0.134 -0.069
Last5_citations	0.0279	0.012	2.380	0.017	0.005 0.051

Omnibus:	50105.449	**Durbin-Watson:**	1.996
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	80223489.753
Skew:	-7.037	**Prob(JB):**	0.00
Kurtosis:	230.564	**Cond. No.**	3.20e+03





formula = Current_Citations ~ Last_citations + Last2_citations + Last3_citations + Last4_citations + Last5_citations

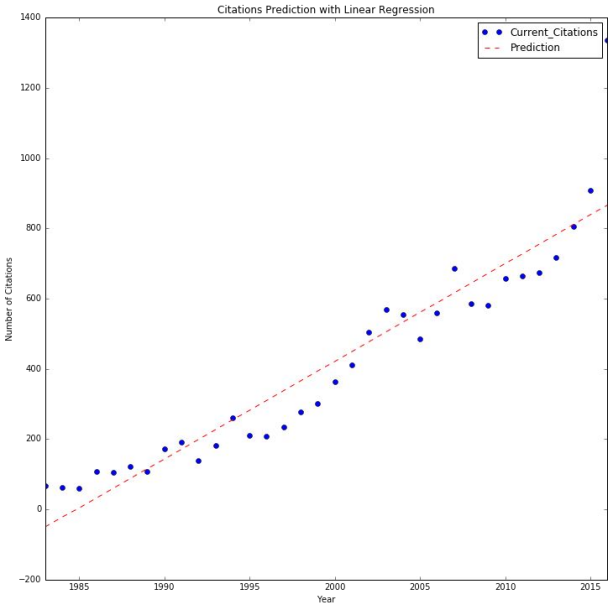
Testing Multiple Variables

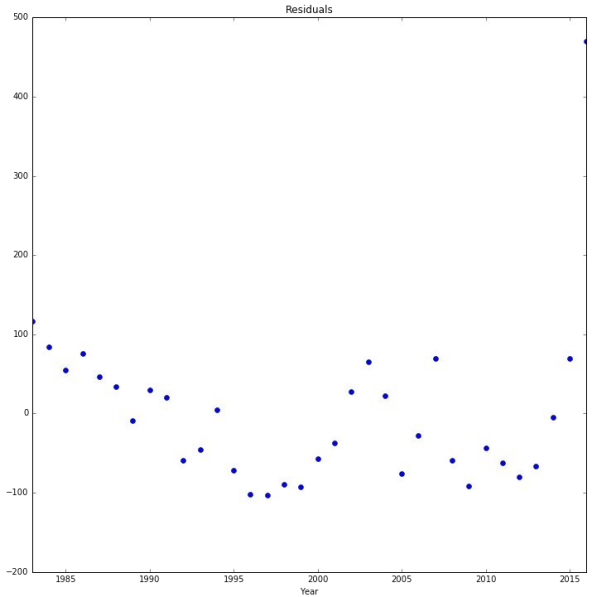
'Years_Active' and 'Age' variables is **not an improvement** for determining Current_Citations though statistically it seems so. Graphically, the prediction does not move dynamically as prediction line did for the single variable test. Furthermore, the residuals look like they are following a pattern. The model is tested on Professor Robert Anderson.

Dep. Variable:	Current_Citations	**R-squared:**	0.958
Model:	OLS	**Adj. R-squared:**	0.955
Method:	Least Squares	**F-statistic:**	365.7
Date:	Mon, 09 May 2016	**Prob (F-statistic):**	9.10e-23
Time:	17:00:18	**Log-Likelihood:**	-205.72
No. Observations:	34	**AIC:**	415.4
Df Residuals:	32	**BIC:**	418.5
Df Model:	2		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	-0.0465	0.033	-1.42	0.163	0.113 0.020
Years_Active	-1.5338	1.075	-1.42	0.163	3.724 0.656
Age	27.8031	1.850	15.02	0.000	24.035 31.572

Omnibus:	43.124	**Durbin-Watson:**	0.679
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	185.778
Skew:	2.716	**Prob(JB):**	4.56e-41
Kurtosis:	13.081	**Cond. No.**	nan





formula = Current_Citations ~ Years_Active + Age

Testing Growth in Citations

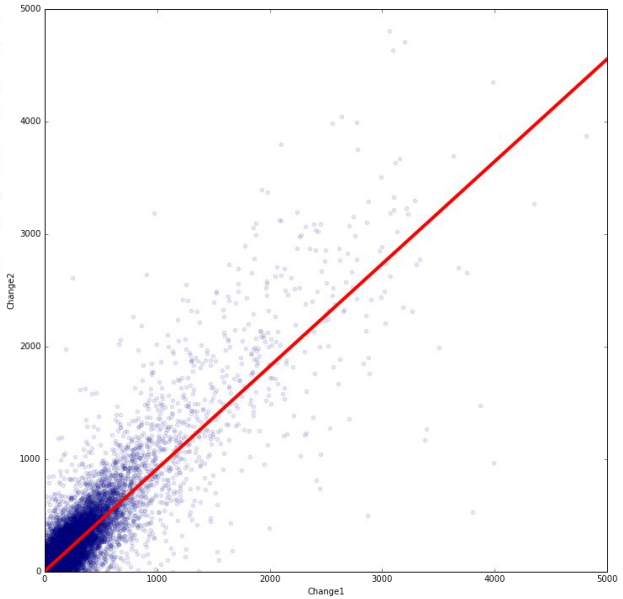
Now we use the *change in citations* to see if the last year's change has any effect on the current change in citations, 'Change1' being the change in citations from Last2_citations and Last_citations, and 'Change2' being the change in citations from Last_citations and Current_Citations

Dep. Variable:	Change2	**R-squared:**	0.597
Model:	OLS	**Adj. R-squared:**	0.597
Method:	Least Squares	**F-statistic:**	6.309e+04
Date:	Mon, 09 May 2016	**Prob (F-statistic):**	0.00
Time:	17:31:51	**Log-Likelihood:**	-2.8638e+05
No. Observations:	42594	**AIC:**	5.728e+05
Df Residuals:	42592	**BIC:**	5.728e+05
Df Model:	1		

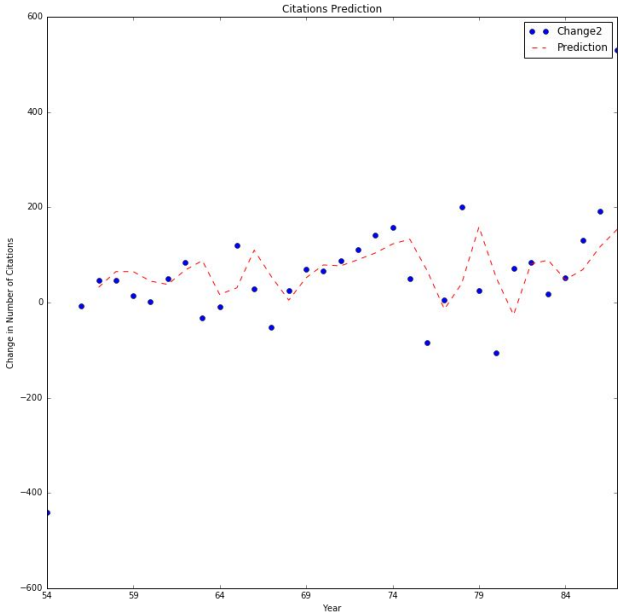
	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	1.6040	1.061	1.512	0.130	-0.475 3.683
Change1	0.9116	0.004	251.161	0.000	.904 0.919

Omnibus:	67572.061	**Durbin-Watson:**	1.692
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	175688281.108
Skew:	-9.698	**Prob(JB):**	0.00
Kurtosis:	317.033	**Cond. No.**	318.

formula = Change2 ~ Change1



Plot of linear regression of change in citations for entire data set



Predicting current change in citations for Professor Robert Anderson

4. Logistical Regression Models*

*Logistic Regressions Run in R

Initial Logistical Regression Tests

Logistic regression requires a binomial dependent variable, so the variable 'inc_dec' was created based on whether or not the Current_Citations is an increases or decrease from Last_citations, where 1 is for increase or 0 is for decrease.

```
Call:
glm(formula = inc_dec ~ Last_citations, family = binomial, data = professors)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6344  -1.6107   0.7819   0.7844   0.9960

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     1.031e+00  1.128e-02  91.392  < 2e-16 ***
Last_citations -2.658e-05  8.346e-06  -3.185  0.00145 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 56109  on 48450  degrees of freedom
Residual deviance: 56099  on 48449  degrees of freedom
  (3018 observations deleted due to missingness)
AIC: 56103

Number of Fisher Scoring iterations: 4
```

Logistic Regression on entire dataset

- Last_citations is statistically significant in whether or not there is an increase or decrease in citations

Initial Logistical Regression Tests

```
glm(formula = inc_dec ~ Last_citations + Last2_citations + Last3_citations +
    Last4_citations + Last5_citations, family = binomial, data = professors)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.3564  -1.4177   0.7759   0.8394   4.2803

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      0.8342978  0.0136891  60.946  < 2e-16 ***
Last_citations   0.0004408  0.0001218   3.620 0.000295 ***
Last2_citations  0.0027556  0.0002006  13.735  < 2e-16 ***
Last3_citations -0.0014918  0.0002127  -7.013 2.33e-12 ***
Last4_citations -0.0010225  0.0002187  -4.675 2.95e-06 ***
Last5_citations -0.0012030  0.0001622  -7.419 1.18e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 44273  on 37037  degrees of freedom
Residual deviance: 42871  on 37032  degrees of freedom
  (14431 observations deleted due to missingness)
AIC: 42883
Number of Fisher Scoring iterations: 5
```

Running additional regressions indicate that all 5 years of citations (Last_citations to Last5_citations) are all statistically significant in determining inc_dec.

Prediction with Logistic Model

Method of Predicting Increase or Decrease:

- Build the Logistic Model

```
fit = glm(formula = inc_dec ~ Last_citations + Last2_citations + Last3_citations +
    Last4_citations + Last5_citations, family = binomial, data = professors)
```

- Create predicted probabilities based on model
  - The predictions returns the predicted probabilities of the response variable

```
citations_predict= predict(fit, type="response")
```

- Set a threshold for the probability to determine if there is a 1 or 0
  - The threshold was set manually using a threshold that yields the highest accuracy

```
citations_predict.f = as.numeric(citations_predict > .24)
```

- Calculate accuracy of the model
  - The accuracy is calculated by taking the mean of

```
accurate = as.numeric(citations_predict.f == professors$inc_dec)
accuracy = mean(accurate)
```

Results of Prediction

Logistic Regression on Entire Data Set

Five Variables - Last_citations to Last5_citations
- Accuracy: 73.28%

Three Variables - Last_citations, Years Active, Age
- Accuracy: 73.42%

Benchmarks:
- Accuracy of all 0s: 26.58%
- Accuracy of random 1s and 0s: 49.95%
- Accuracy of all 1s: 73.42%

Conclusion: logistic regression model is as only as accurate at predicting whether or not there is an increase in Current_Citations as just assuming there is an increase every year

Logistic Regression on Researchers of 10-20 Years

Five Variables - Last_citations to Last5_citations
- Accuracy: 77.46%

Three Variables - Last_citations, Years Active, Age
- Accuracy: 77.63%

Benchmarks:
- Accuracy of all 0s: 26.58%
- Accuracy of random 1s and 0s: 49.95%
- Accuracy of all 1s: 77.63%

Testing Growth in Citations

```
Call:
glm(formula = inc_dec ~ Change1, family = binomial, data = professors)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2062  -1.5686   0.7903   0.8104   1.1517

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 9.358e-01  1.138e-02   82.25   <2e-16 ***
Change1     1.281e-03  9.397e-05   13.63   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 53029  on 45534  degrees of freedom
Residual deviance: 52804  on 45533  degrees of freedom
  (5934 observations deleted due to missingness)
AIC: 52808

Number of Fisher Scoring iterations: 4
```

Since the binomial variable ‘inc_dec’ is already an indication of change, we test ‘Change1’ (change between Last_citations and Last2_citations).

Adding another year of change in citations - ‘Change3’ (change between Last2_citations and Last3_citations did not improve the model.

The accuracy of predicting ‘inc_dec’ is the same as the accuracy of the model with ‘Last_citaitons’, ‘Years.Active’, and ‘Age’ at 73.42% when applied to the entire dataset.

5. Conclusion

Takeaways

- Linear regression with one variable - ‘Last_citations’ - is best model for predicting ‘Current_Citations’
  - Visually, the prediction looks pretty accurate in detecting which direction citations will move (increase or decrease), but with a slight lag
  - The model works best with researchers with longer active years because there is more data to build on
  - Adding multiple variables did not significantly improve the model or predictions
- Linear regression is also pretty good and predicting the direction of change in citations, but doesn’t quite capture the magnitude of change
- Logistical regression is able to determine whether or not citations will increase or decrease with over 70% accuracy, but that accuracy is not any better than just assuming all citations will increase

Improvements

- Remove NaN values in the ‘Last_citations’ (and later years)
- **Test non-linear regression, probably a completely different model or equation**:
  - Linear regression model can only go so far now
  - Citations do not follow a linear, logarithmic, or exponential growth pattern
- Test other variables, such as number of courses research teaches, whether or not professor is tenured or not, etc.