Understanding how workforce, expense and enterprise value affect revenue

Puneet Jaiswal

California State University East Bay

Author Note

Puneet Jaiswal is pursuing MS in statistics from CSU East Bay.

This research is done as part of a course work STAT 6509 Theory Application Regression (Spring 2017) under supervision of Dr. Ayona Chattergee.

Correspondence concerning this article should be addressed to Puneet Jaiswal

Contact: pjaiswal3@horizon.csueastbay.edu

Abstract

This paper explores factors which can be used to predict revenue for a company. The purpose of this experiment is to build a predictive model using variables which affect revenue and then using that model, we can explain the variability of the data. For this experiment, I used data for US based companies only and ran the tests using finance data for public companies listed in NASDAQ and NYSE. In order to build the model, we used finance data for 109 public companies, data was collected from yahoo-finance on 05/11/2017 for companies having software defined business. After running correlation tests, we concluded that workforce (number of full time employees), overall expense and enterprise value are the most significant variables which affect the revenue. In the process of building the final predict model, we analyzed how each predictor is related to the response and why there is a need for log transformation among the variables. Finally we established the relationship between the predictors and response using linear regression on the log transformed data. We used step function to find the best model with most significant interaction terms and then ran various diagnostic tests to conclude that the predict model is good enough.

Keywords: linear regression with log transformation, revenue prediction model

INTRODUCTION

In this report, I collected very recent (collected on 05/11/17) finance data for software driven business based organization in US market (public companies listed in NASDAQ and NYSE) from yahoo-finance API. The companies selected for this experiment are very diverse, these are from different fields such as e-commerce, retail, aviation, software, hardware, finance etc. On a day I looked at top market movers (both losers and gainers) and chose around 100 tickers (NASDAQ/NYSE symbols), and then added 21 additional technology companies of my own interests and then ran a field extractor script to fetch data from yahoo-finance API. After running the script I got complete data for 109 companies (for remaining 12 companies few fields were missing). I decided to choose revenue as response because revenue is the final output any company makes. There were many options for predictor variables, few being workforce (total number of employees), expense, profit (revenue - expense), market capital, enterprise evaluation (market cap - total debt), stock volume and price, operating cash flow etc. Profit is not a good predictor as it is derived from revenue also I had to choose either among enterprise value and market capital since enterprise value is derived from market capital. Stock volume and price are again not good predictors as these are result of market capital.

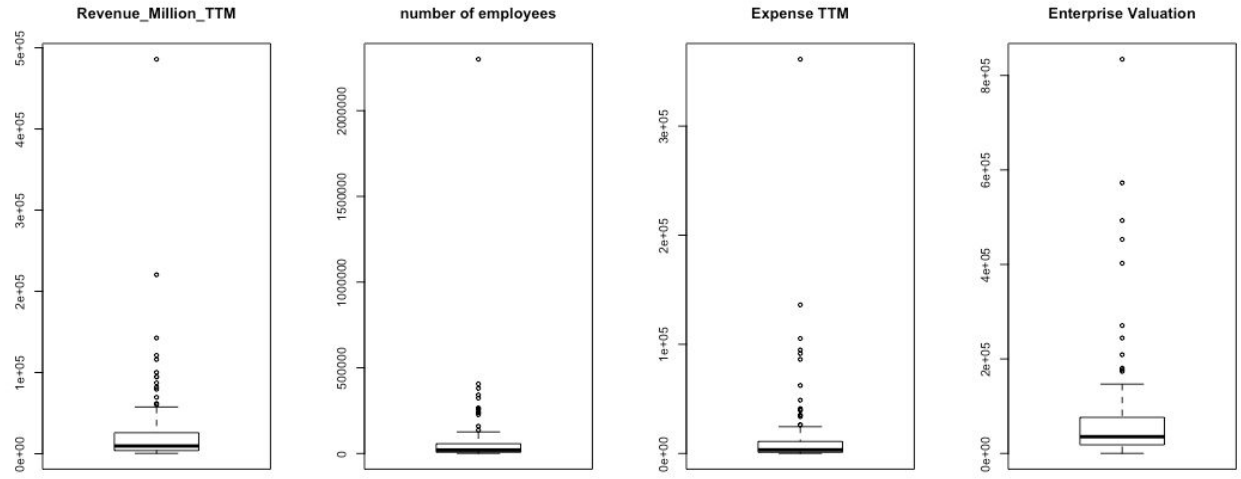
METHODOLOGY

It is important to understand how each of the response and predictors are distributed and find how response and predictors are related, how significant is the correlation between response with each predictor. After understanding the relationship, we then can build and model and then validate it by running various tests. In this paper, I have described each such steps and the output of it. I have used R to run these tests, building models, validation and prediction.

I started with boxplot to understand how each variable is distributed and how far outliers are from the median. Boxplot would give a better understanding of the data and help decide if there is a transformation required to better represent the data. Usually when data is collected for very diverse sample, oftentimes distribution is skewed and log transformation would may be a right choice to represent the data. After this analysis scatter plot help decide how relevant predictors are, and by looking at that, predictor selection is justified. After understanding the relationship between response and predictor it becomes easier to build a initial regression model. Once a regression model is build we check for its r-squared value and if it is significant enough, we try to find best model with most significant interaction terms. I used step function with null and full model (model with all predictors and interaction terms) to find best model (one that has lowest AIC value). Once I found the final model, I ran a few diagnostic tests to confirm the validity of the model, some essential tests are normality test for the residuals, variance constancy test and finally marginal plot to understand residual vs leverage, fitted vs observed.

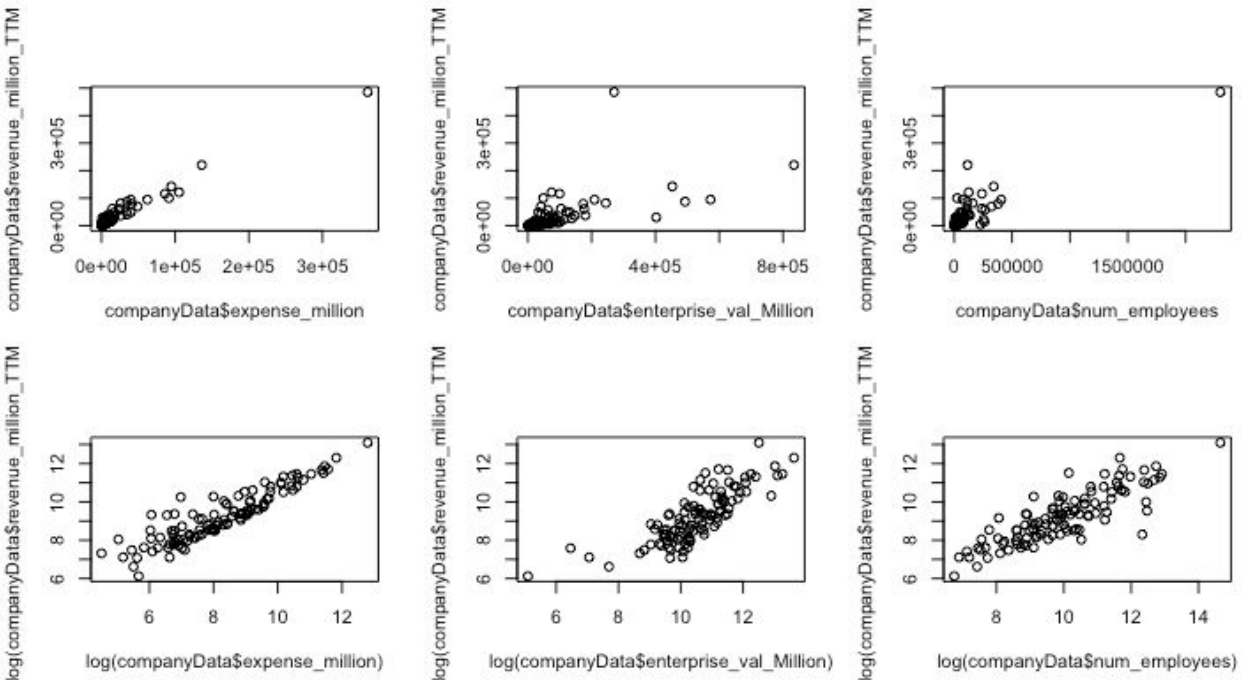
ANALYSIS AND PRESENTATION OF DATA

Before I start building the model, I would want to understand the data distribution and correlation of the response variable with various predictors. Box plot is one of the best tools to understand the data distribution. I plotted boxplot to demonstrate how each variable is distributed. Below is a summary of how these variables are distributed.

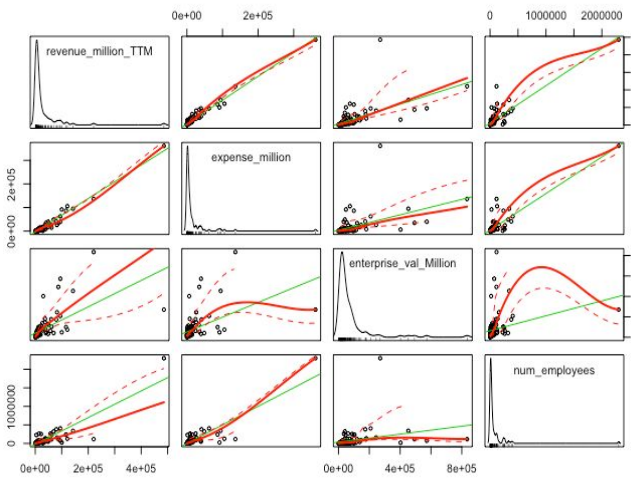
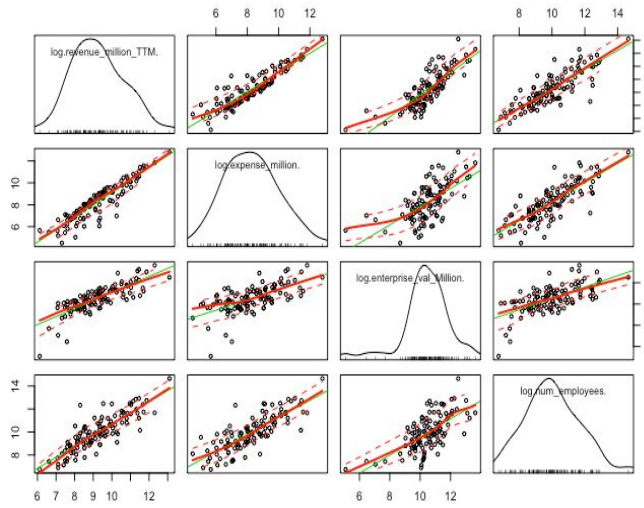


As we can see here, there are so many points which are far away from the median. This is usually observed when we sample from diverse real life dataset. In the plot we can see one point farthest from the median; that company is Walmart which has highest revenue (~480 B), full time employees (~2.3 M), expense and enterprise value. There is another company in the list which has minimum number of full time employees (851). To explain the variability better, I thought to transform the data using log function. After applying log transformation, the data distribution looks much better as we can see a lot fewer outliers in the boxplot.

Now as we see the log transformation made the data distribution more uniformly distributed, a response-predictor plot would explain it better that why we need log-transformed variables in the regression model. Here is the plot explaining the relationship between with and without the log transformation.



In the above plot relationship between response and predictor is shown without and with the log-transformation. When we apply log-transformation, the relationship comes out to be linear. We can confirm the same by running scatterplot.

Scatter plot without log transformation	Scatter plot with log transformation
	

From the scatterplot comparison it is clear that after applying log transformation all predictors and response distributions became close to normal with better linearity and since the relationships are very much linear, I decided to build linear regression model.

BUILDING THE REGRESSION MODEL

With the scatterplot and correlation tests it is clear that revenue can be predicted by using number of employees, expense and enterprise value predictors. In order to find best model, I ran step function with null and full models as parameter. Here full model has all three predictor terms and all possible interactions terms as well.

<table>
<tr><td>Null model</td><td>

```
lm(formula = log(revenue_million_TTM) ~ 1, data = companyData)
```

</td></tr>
<tr><td>No interaction model</td><td>

```
lm(formula = log(revenue_million_TTM) ~ log(num_employees) +
    log(expense_million) + log(enterprise_val_Million), data = companyData)

Residuals:
     Min       1Q   Median       3Q      Max
-0.95702 -0.28240 -0.07003  0.22016  1.36275

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 0.25821    0.31142   0.829 0.408917
log(num_employees)          0.15968    0.04115   3.881 0.000182 ***
log(expense_million)        0.44628    0.03739  11.936  < 2e-16 ***
log(enterprise_val_Million) 0.35325    0.03840   9.198 3.85e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3751 on 105 degrees of freedom
Multiple R-squared:  0.925,     Adjusted R-squared:  0.9229
F-statistic: 431.9 on 3 and 105 DF,  p-value: < 2.2e-16
```

</td></tr>
<tr><td>Full model</td><td>

```
lm(formula = log(revenue_million_TTM) ~ log(num_employees) +
    log(expense_million) + log(enterprise_val_Million) + log(num_employees) *
    log(expense_million) + log(num_employees) * log(enterprise_val_Million) +
    log(expense_million) * log(enterprise_val_Million), data = companyData)

Residuals:
     Min       1Q   Median       3Q      Max
-0.91418 -0.25782 -0.06887  0.18816  1.33401

Coefficients:
                                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                                      3.49784    1.48065   2.362   0.0201 *
log(num_employees)                              -0.55098    0.37912  -1.453   0.1492
log(expense_million)                             0.67038    0.37731   1.777   0.0786 .
log(enterprise_val_Million)                      0.20326    0.19091   1.065   0.2895
log(num_employees):log(expense_million)          0.01821    0.01701   1.070   0.2869
log(num_employees):log(enterprise_val_Million)   0.05264    0.03580   1.470   0.1445
log(expense_million):log(enterprise_val_Million) -0.04027    0.03573  -1.127   0.2624
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3689 on 102 degrees of freedom
Multiple R-squared:  0.9295,    Adjusted R-squared:  0.9254
F-statistic: 224.3 on 6 and 102 DF,  p-value: < 2.2e-16
```

</td></tr>
</table>

In the table above note that in the model with no interaction terms, all the coefficients are significant as respective p-values are smaller than 0.05 however F-statistic is high enough to consider the inclusion of the interaction terms. Using null and full model, after running the step function, I found below model to have lowest AIC value.

```
Step:  AIC=-212.51
log(revenue_million_TTM) ~ log(expense_million) + log(enterprise_val_Million) +
    log(num_employees) + log(enterprise_val_Million):log(num_employees)

                                                Df Sum of Sq    RSS     AIC
<none>                                                       14.154 -212.51
+ log(expense_million):log(enterprise_val_Million)  1  0.113571 14.040 -211.39
+ log(num_employees):log(expense_million)           1  0.096686 14.057 -211.26

Call:
lm(formula = log(revenue_million_TTM) ~ log(expense_million) +
    log(enterprise_val_Million) + log(num_employees) + log(enterprise_val_Million):log(num_employees),
    data = companyData)

Coefficients:
                                    (Intercept)                            log(expense_million)
                                        3.32910                                         0.42654
                     log(enterprise_val_Million)                              log(num_employees)
                                        0.07223                                        -0.17184
log(enterprise_val_Million):log(num_employees)
                                        0.03160
```

Final model:

```
lm(formula = log(revenue_million_TTM) ~ log(expense_million) +
    log(enterprise_val_Million) + log(num_employees) + log(enterprise_val_Million) *
    log(num_employees), data = companyData)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9645 -0.2654 -0.0817  0.1972  1.3569

Coefficients:
                                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                                 3.32910    1.47095   2.263   0.0257 *
log(expense_million)                        0.42654    0.03792  11.249   <2e-16 ***
log(enterprise_val_Million)                 0.07223    0.13697   0.527   0.5991
log(num_employees)                         -0.17184    0.16050  -1.071   0.2868
log(enterprise_val_Million):log(num_employees)  0.03160    0.01480   2.134   0.0352 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3689 on 104 degrees of freedom
Multiple R-squared:  0.9282,    Adjusted R-squared:  0.9254
F-statistic:   336 on 4 and 104 DF,  p-value: < 2.2e-16
```

The final model has r-squared value 0.93 which means that 93% of the variability can be explained by this model which is good enough to confirm accuracy of prediction. Also residual standard error is 0.37 which is pretty small to confirm the quality. Now to confirm the validity of model, constancy of variance and normality of residuals should be tested.

Constancy of variance test:

After running ncvTest (test for checking non constancy of variance)

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.5780414    Df = 1     p = 0.4470811
```

Null Hypothesis H_0: variance is nonconstant.

Alternate Hypothesis H_a: variance is constant.

Since p-value is bigger than 0.05, we can reject the null hypothesis. So we can confirm constancy of variance.

Residual normality test:

After running shapiro-wilk normality test

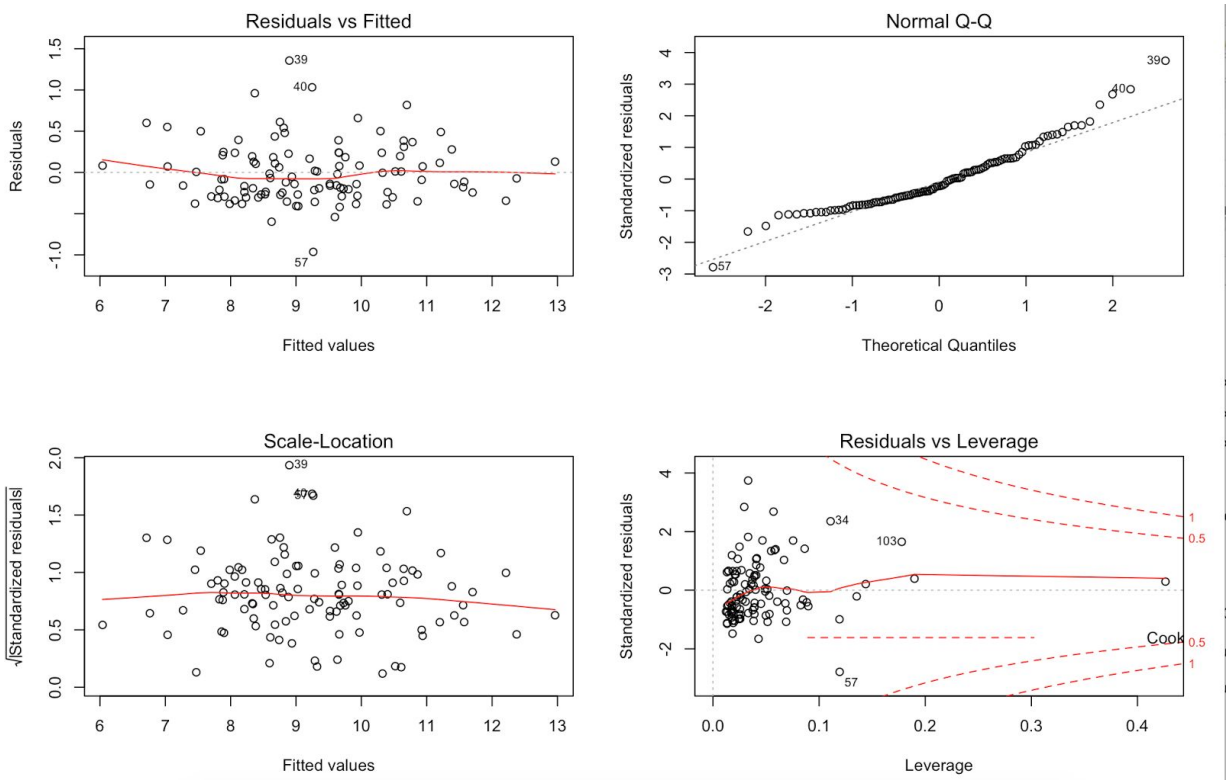
```
        Shapiro-Wilk normality test

data:  finalModel$residuals
W = 0.94224, p-value = 0.000135
```

Null Hypothesis H_0: residuals are normally distributed.

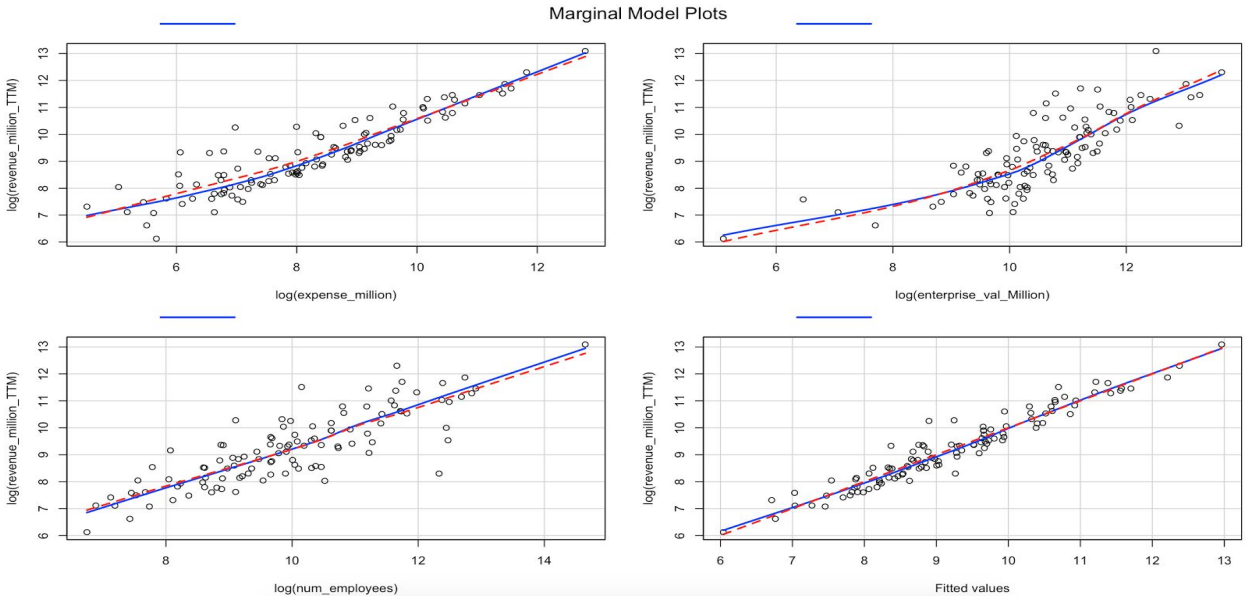
Alternate Hypothesis H_a: residuals are not normally distributed.

By looking at the output of shapiro-wilk test, p-value is smaller than 0.05 so we can reject the null hypothesis, however W value is close to 0.95, we can say that residual distribution is close to normal.



Marginal plots:

Now since the model is already validated, we can compare the fitted values with observed values by running marginal plot. Here is the result of marginal plot.



In the above observation we can see that fitted values are very close to observed values, which confirms the efficiency of the regression model.

CONCLUSIONS, RECOMMENDATIONS FOR PRACTICE

This model can be used to forecast the growth of an enterprise, given a change in workforce and expense, revenue can be predicted. There can be few usage of this model, such as:

1. Finding how efficiently an enterprise is running
2. Projection enterprise growth with respect to workforce and expense.
3. By looking at revenue, an organization can decide whether or not to add or reduce workforce/expense.
4. How an enterprise is performing as compared to its competitors.

Below table explains fitted and observed values for a few companies. Last column in this table is efficiency factor defined as unit change between observed and predicted revenue.

Company	#Employees	enterprise value (market cap - debts)	expense	Observed Revenue	Predicted Reven	(E-F)/E
Oracle	136000	180430	7860	37430	36980	0.0120
Boeing	150500	112620	79130	92910	81871	0.118
Splunk	2700	8300	191	949	1231	-0.297
HortonWorks	1080	435.78	86.8	199.09	334	-0.677
Box Inc	1495	2220	112.13	398.61	615	-0.542

In the above table I selected a few enterprises and tried to predict their revenues using the number of employees, expense and enterprise value. It is observed that well established companies have higher revenue observed than predicted, however newer companies are yet to reach that point.

Below is another example, where I selected companies in retail commerce domain. These all are

doing great it seems as these are having higher observed revenue than predicted, however there is one thing to notice, Walmart is biggest of all among these in terms of revenue, employees and market cap, others are smaller but more efficient as they are competing with a retail giant. The trend here is, smaller the company, more efficient it is.

Company	#Employees	enterprise value (market cap - debts)	expense	Observed Revenue	Predicted Reveni	(E-F)/E
Walmart	2300000	270603	361255	485872	426432.9	0.122
Costco	218000	78070	118350	121200	88286.13	0.271
Target	323000	30080	66534	69320	47282.87	0.317
Sears	140000	834	22556	21050	5284.194	0.748

APPENDIX

Yahoo Finance API Url

https://query1.finance.yahoo.com/v10/finance/quoteSummary/MSFT?region=US&modules=defaultKeyStatistics,assetProfile,financialData

2017 Company finance data

https://github.com/puneetjaiswal/regression-analysis/blob/master/2017_us_company_data.txt

R Code for the experiments

https://github.com/puneetjaiswal/regression-analysis/blob/master/AnalyzeCompanyRevenueEmployee.R

Slides for the presentation

https://docs.google.com/presentation/d/1ZyRH9jbsTHpBZOJ_66_xY45mZspve1hZBF-MnsieFBE/edit?usp=sharing

DATA

Symbol	Company	num_employees	revenue_million_TTM		profit_Million_TTM		expense_million	enterprise_val_Million
AAL	American Airlines Group Inc	124300	40369	30300	10069	40792.08448		
AAPL	Apple Inc	116000	220456	84263	136193	834284.6177		
ADBE	Adobe Systems Inc	15706	6152	5034	1118	64658.8416		
ADI	Analog Devices Inc	10000	3636	2227	1409	21742.50598		
ADP	Automatic Data Processing Inc		57000	12213	4828	7385	42161.00045	
ADSK	Autodesk Inc	9000	2031	1690	341	20662.03034		
AKAM	Akamai Technologies Inc	6672	2381	1531	850	8424.699904		
ALXN	Alexion Pharmaceuticals Inc	3121	3252	2826	426	28550.21158		
AMAT	Applied Materials Inc	15600	11845	4511	7334	46129.0537		
AMGN	Amgen Inc	19200	22927	18829	4098	113416.6917		
AMZN	Amazon.com Inc	341400	142572	47722	94850	453007.147		
ATVI	Activision Blizzard Inc	9400	6879	4214	2665	43161.67782		
AVGO	Broadcom Ltd	15700	15608	5940	9668	102776.8484		
BIDU	Baidu Inc	45887	10390	5075	5315	59738.79194		
BIIB	Biogen Inc	7400	11532	9970	1562	57543.07789		
BMRN	Biomarin Pharmaceutical Inc	2293	1183	907	276	15596.58086		
CA	CA Inc	11000	4032	3186	846	12652.88602		
CELG	Celgene Corp	7132	11677	10791	886	98748.12109		
CERN	Cerner Corp	24400	4827	4018	809	21413.5296		
CHKP	Check Point Software Technologies Ltd		4281	1772	1539	233	15987.53587	
CHTR	Charter Communications Inc	91500	36636	10348	26288	146769.4776		
CTRP	Ctrip.Com International Ltd	37000	3066	2086	980	29753.67168		
CTAS	Cintas Corp	35000	5143	2130	3013	13903.52179		
CSCO	Cisco Systems Inc	71959	48569	30960	17609	131489.2186		
CTXS	Citrix Systems Inc	9600	3422	2859	563	13210.01574		
CMCSA	Comcast Corp	159000	82076	55940	26136	244342.2761		
COST	Costco Wholesale Corp	126000	121193	15818	105375	74280.07526		
CSX	CSX Corp	26628	11320	8287	3033	58185.90822		
CTSH	Cognizant Technology Solutions Corp	261200	13831	5379	8452	34778.05056		
DISCA	Discovery Communications Inc		7000	6549	4065	2484	23106.40026	
DISH	DISH Network Corp	16000	14947	3948	10999	39146.17651		
DLTR	Dollar Tree Inc	55300	20719	6395	14324	24707.31776		
EBAY	eBay Inc	12600	9059	6972	2087	39141.3719		
ESRX	Express Scripts Holding Co	25600	100150	8620	91530	48367.04666		
EXPE	Expedia Inc	20075	9058	7177	1881	20927.55149		
FAST	Fastenal Co	19822	4023	1965	2058	13241.79046		
FB	Facebook	18770	30287	23849	6438	402536.4644		
FISV	Fiserv Inc	23000	5568	2546	3022	29785.62458		
FOXA	21st Century Fox Class A	21500	28398	27326	1072	67171.56147		
GILD	Gilead Sciences Inc	9000	29101	26129	2972	81674.96909		
GOOGL	Alphabet Class A	73992	94764	55134	39630	572733.5875		
HAS	Hasbro Inc	5400	5038	3115	1923	12841.87341		
HSIC	Henry Schein Inc	21000	11781	3234	8547	15354.02086		
HOLX	Hologic Inc	5333	2894	1564	1330	14366.37798		
ILMN	Illumina Inc	5500	2424	1667	757	25812.04173		
INCY	Incyte Corp	980	1226	1048	178	23371.53434		
INTC	Intel Corp	106000	60480	36191	24289	176520.2002		
INTU	Intuit Inc	7900	4851	3964	887	32967.04102		
ISRG	Intuitive Surgical Inc		3755	2784	1890	894	29663.06611	
JBHT	J.B. Hunt Transport Services Inc		22190	6655	2638	4017	10407.74963	
JD	JD.com Inc	122405	40967	5673	35294	54815.98566		
KLAC	KLA-Tencor Corp	5580	3460	1821	1639	15968.33382		
KHC	Kraft Heinz Co	41000	26281	9586	16695	138148.2824		
LBTYA	Liberty Global PLC	41000	19850	15402	4448	73723.54355		
LBTYK	Liberty Global PLC	41000	19850	15403	4447	73064.60365		
LRCX	Lam Research Corp	8600	7214	2618	4596	20999.99539		
MAR	Marriott International Inc	226500	4020	2626	1394	46689.54624		

MAT	Mattel Inc	32000	5322	2554	2768	9609.783296	
MDLZ	Mondelez International Inc	90000	25881	10128	15753	84743.64314	
MNST	Monster Beverage Corp	1910	3111	1942	1169	26401.12845	
MSFT	Microsoft Corp	114000	87247	52540	34707	492965.4292	
MU	Micron Technology Inc	31400	14732	2505	12227	40671.09683	
MXIM	Maxim Integrated Products Inc	7213	2259	1244	1015	11753.93894	
MYL	Mylan NV	35000	11605	4697	6908	35346.7351	
NCLH	Norwegian Cruise Line Holdings Ltd	30000	4947	2024	2923	18023.70662	
NFLX	Netflix Inc	3200	9509	2801	6708	70353.89542	
NTES	NetEase Inc	15948	6369	3117	3252	31227.136	
NVDA	NVIDIA Corp	10299	7542	4063	3479	70884.99917	
ORLY	O Reilly Automotive Inc	74580	8653	4509	4144	25272.76851	
PAYX	Paychex Inc	13900	3106	2952	154	20196.9152	
PCAR	PACCAR Inc	23000	16971	3090	13881	28132.56909	
PCLN	The Priceline Group	20500	11014	10315	699	92037.24288	
PYPL	PayPal Holdings Inc	18100	11272	10842	430	50774.54438	
QCOM	Qualcomm Inc	30500	23242	13805	9437	82741.23162	
REGN	Regeneron Pharmaceuticals Inc	5505	4978	4561	417	45897.59693	
ROST	Ross Stores Inc	78600	12866	3693	9173	24848.21606	
SBAC	SBA Communications Corp	1241	1656	1212	444	24026.53798	
STX	Seagate Technology PLC	45500	11018	2615	8403	14805.73338	
SHPG	Shire PLC	23906	13259	7580	5679	76365.03757	
SIRI	Sirius XM Holdings Inc	2402	5110	2511	2599	28289.39059	
SWKS	Skyworks Solutions Inc	7300	3353	1666	1687	17527.16493	
SBUX	Starbucks Corp	254000	21976	12805	9171	88849.58822	
TMUS	T-Mobile US Inc	50000	38193	20692	17501	80037.30637	
TRIP	TripAdvisor Inc	3327	1500	1409	91	5913.636864	
TSCO	Tractor Supply Co	13000	6875	2325	4550	8416.03072	
TSLA	Tesla Inc	17782	8549	1599	6950	56546.83648	
TXN	Texas Instruments Inc	29865	13763	8240	5523	80413.43386	
ULTA	Ulta Beauty Inc	11600	4854	1747	3107	18219.07763	
VOD	Vodafone Group PLC	111684	50579	16842	33737	120917.0739	
VRSK	Verisk Analytics Inc	6148	2005	1281	724	15071.81158	
VRTX	Vertex Pharmaceuticals Inc	2150	2018	1491	527	27872.19251	
WBA	Walgreens Boots Alliance Inc	240000	116081	29874	86207	99511.30419	
WDC	Western Digital Corp	72878	17745	3435	14310	33529.61843	
XRAY	Dentsply Sirona Inc	15700	3873	2001	1872	15592.07014	
YHOO	Yahoo Inc	8800	5409	2450	2959	42265.91539	
IBM	International Business Machines Corp.	380300	79389	38294	41095	173626.196	
HPQ	Hewlett-Packard Co.	49000	48675	8998	39677	33079.5008	
GPRO	Go pro	1327	1220	462	758	1164.543488	
SQ	Square	1853	1791	577	1214	6828.079104	
TWTR	Twitter	3583	2483	1598	885	11115.96442	
FUEL	Rocket Fuel	851	456	167	289	162.9464	
ZNGA	Zynga	1681	748	502	246	2203.907584	
FIT	Fitbit	1722	1963	846	1117	640.404288	
TGT	Target Corp	323000	69494	20623	48871	41218.98189	
HD	The home depot	406000	94594	32313	62281	208958.8163	
UPS	United Parcel Service INC	237040	61802	47084	14718	103917.527	
FDX	Fedex	266000	57570	33038	24532	62610.83341	
ORCL	Oracle Corp	136000	37428	29568	7860	180434.1412	
WMT	Wal-Mart Stores Inc.	2300000	485872	124617	361255	270603.976704	

R Code

```
# Company - employee - revenue analysis
# Revenue (response), #Employee (predictor)

setwd("/opt/Code/regression-analysis/")

companyData <- read.table("2017_us_company_data.txt", header = TRUE, sep='\t')

par(mfrow=c(1,2))

par(mfrow=c(1,4))
# Revenue
boxplot(companyData$revenue_million_TTM,main="Revenue_Million_TTM")
boxplot(log(companyData$revenue_million_TTM),main="Log(Revenue_Million_TTM)")

# Num_emp
boxplot(companyData$num_employees,main="number of employees")
boxplot(log(companyData$num_employees),main="Log(number of employees)")

# Expense
boxplot(companyData$expense_million,main="Expense TTM")
boxplot(log(companyData$expense_million),main="Log (Expense TTM)")

# Enterprise Value
boxplot(companyData$enterprise_val_Million, main="Enterprise Valuation")
boxplot(log(companyData$enterprise_val_Million), main="Log (Enterprise Valuation)")

par(mfrow=c(2,3))
plot(x=companyData$expense_million, y=companyData$revenue_million_TTM)
plot(x=companyData$enterprise_val_Million, y=companyData$revenue_million_TTM)
plot(x=companyData$num_employees, y=companyData$revenue_million_TTM)

plot(x=log(companyData$expense_million), y=log(companyData$revenue_million_TTM))
plot(x=log(companyData$enterprise_val_Million), y=log(companyData$revenue_million_TTM))
plot(x=log(companyData$num_employees), y=log(companyData$revenue_million_TTM))

cor.test(x=companyData$expense_million, y=companyData$revenue_million_TTM)
cor.test(x=log(companyData$expense_million), y=log(companyData$revenue_million_TTM))

cor.test(x=companyData$enterprise_val_Million, y=companyData$revenue_million_TTM)
cor.test(x=companyData$enterprise_val_Million, y=companyData$revenue_million_TTM)

cor.test(x=companyData$num_employees, y=companyData$revenue_million_TTM)
cor.test(x=companyData$num_employees, y=companyData$revenue_million_TTM)

par(mfrow=c(1,2))
revenue_lm <- lm(revenue_million_TTM ~ num_employees, companyData)
summary(revenue_lm)
abline(revenue_lm)


# Using log transformation to reduce the residual st. error
boxplot(log(companyData$num_employees))
boxplot(log(companyData$revenue_million_TTM))
cor.test(log(companyData$num_employees),log(companyData$revenue_million_TTM))
plot(log(companyData$num_employees),log(companyData$revenue_million_TTM))
revenue_lm2 <- lm(log(revenue_million_TTM) ~ log(num_employees), companyData)
abline(revenue_lm2)
summary(revenue_lm2)
```

```
# including expanse in model
boxplot(log(companyData$expense_million))
boxplot(companyData$expense_million)
revenue_lm3 <- lm(log(revenue_million_TTM) ~ log(num_employees)+log(expense_million), companyData)
summary(revenue_lm3)

par(mfrow=c(2,2))
plot(revenue_lm3)

# interaction term
revenue_lm4 <- lm(log(revenue_million_TTM) ~ log(num_employees)+log(expense_million) + log(num_employees) *
log(expense_million), companyData)
summary(revenue_lm4)
plot(revenue_lm4)
# Interaction term has reduced the F-statistic value but no significan improvement in R-squared value

# include company valuation also in the model
boxplot(companyData$enterprise_val_Million)
boxplot(log(companyData$enterprise_val_Million))
revenue_lm5 <- lm(log(revenue_million_TTM) ~ log(num_employees)+log(expense_million)+log(enterprise_val_Million),
companyData)
summary(revenue_lm5)
plot(revenue_lm5)


# more interaction terms
revenue_lm6 <- lm(log(revenue_million_TTM) ~ log(num_employees)+log(expense_million)+log(enterprise_val_Million)+
log(num_employees) * log(expense_million)+ log(num_employees) * log(enterprise_val_Million), companyData)
summary(revenue_lm6)
plot(revenue_lm6)


# No Log transform
revenue_lm7 <- lm(revenue_million_TTM ~ num_employees+expense_million+enterprise_val_Million, companyData)
summary(revenue_lm7)
plot(revenue_lm7)
hist(revenue_lm7$residuals)

# No Log transform, with interaction terms
revenue_lm8 <- lm(revenue_million_TTM ~
num_employees+expense_million+enterprise_val_Million+(num_employees*enterprise_val_Million), companyData)
summary(revenue_lm8)
plot(revenue_lm)

plot(companyData$num_employees,revenue_lm8$fitted.values)
plot(companyData$num_employees,companyData$revenue_million_TTM)




# antilog function used to reverse predicted response value
antilog<-function(lx,base)
{
  lbx<-lx/log(exp(1),base=base)
  result<-exp(lbx)
  result
}

plot(revenue_lm5$fitted.values,log(companyData$Revenue_Million_TTM))
abline(0,1)
```

```
######## Running AIC to determine best model ############
nullModel <- lm(log(revenue_million_TTM) ~ 1, companyData)
summary(nullModel)

noInteractionModel <- lm(log(revenue_million_TTM) ~
log(num_employees)+log(expense_million)+log(enterprise_val_Million), companyData)
summary(noInteractionModel)
fullModel0 <- lm(log(revenue_million_TTM) ~ log(num_employees)+log(expense_million)+log(enterprise_val_Million),
companyData)
summary(fullModel0)

fullModel <- lm(log(revenue_million_TTM) ~ log(num_employees)+log(expense_million)+log(enterprise_val_Million)+
log(num_employees) * log(expense_million)+ log(num_employees) * log(enterprise_val_Million)+log(expense_million) *
log(enterprise_val_Million), companyData)
summary(fullModel)

step(nullModel, scope=list(lower=nullModel, upper=fullModel), direction="forward")

finalModel <-  lm(log(revenue_million_TTM) ~ log(expense_million) + log(enterprise_val_Million) + log(num_employees) +
log(enterprise_val_Million)*log(num_employees), data=companyData)
summary(finalModel)
plot(finalModel)
hist(finalModel$residuals)

shapiro.test(finalModel$residuals)


install.packages("car")
library(car)

scatterplotMatrix(~revenue_million_TTM + expense_million + enterprise_val_Million + num_employees, data=companyData)
scatterplotMatrix(~log(revenue_million_TTM) + log(expense_million) + log(enterprise_val_Million) + log(num_employees),
data=companyData)

marginalModelPlots(finalModel)


qqPlot(finalModel)
ncvTest(finalModel)
leveragePlots(finalModel)
residualPlots(finalModel)
influencePlot(finalModel, id.n=2)


# Now lets predict revenue for Oracle Corporation (ORCL) - NYSE
# num_emp = 136000, enterprise_val = 180.43B, profit =     29.57B, Observed revenue = 37.43B, expanse = revenue - profit =
7.86B
newdata = list(num_employees=136000, expense_million= 7860, enterprise_val_Million=180430)
logRevenue = predict(finalModel,newdata, interval = "confidence", level = 0.95)
logRevenue
antilog(logRevenue, exp(1))

# Now lets predict revenue for The Boeing Company (BA) - NYSE
#num_emp = 150500, enterprise_val = 112.62B, Observed revenue = 92.91B, profit = 13.78B, expense = revenue - profit =
79.13B
newdata = list(num_employees=150500, expense_million= 79130, enterprise_val_Million=112620)
logRevenue = predict(finalModel,newdata, interval = "confidence", level = 0.95)
logRevenue
antilog(logRevenue, exp(1))
```

```
########## Under performers #############
# Now lets predict revenue for HORTONWORKS, INC. (HDP) - Nasdaq
# num_emp = 1080, enterprise_val = 435.78M, profit =112.29M, Observed revenue = 199.09M, expanse = revenue - profit =
86.8 M
newdata = list(num_employees=1080, expense_million= 86.8, enterprise_val_Million=435.78)
logRevenue = predict(finalModel,newdata, interval = "confidence", level = 0.95)
logRevenue
antilog(logRevenue, exp(1))

# Now lets predict revenue for BOX INC (BOX) - NYSE
# num_emp = 1495, enterprise_val = 2.22B, profit =          286.48M, Observed revenue = 398.61M, expanse = revenue - profit =
112.13M
newdata = list(num_employees=1495, expense_million= 112.13, enterprise_val_Million=2220)
logRevenue = predict(finalModel,newdata, interval = "confidence", level = 0.95)
logRevenue
antilog(logRevenue, exp(1))

# Now lets predict revenue for Splunk Inc. (SPLK) - NasdaqGS
# num_emp = 2700, enterprise_val = 8.3B, Observed revenue = 949.96M, profit = 758.9M, expanse = revenue - profit =
191.06M
newdata = list(num_employees=2700, expense_million= 191.06, enterprise_val_Million=8300)
logRevenue = predict(finalModel,newdata, interval = "confidence", level = 0.95)
logRevenue
antilog(logRevenue, exp(1))


# Company          #Employees          enterprise value      expense   Revenue
# Walmart          2300000  270603   361255    485872
# Costco  218000   78070    118350   121200
# Target  323000   30080    66534    69320
# Sears   140000   834      22556    21050
newdata = list(num_employees=2300000, expense_million= 361255, enterprise_val_Million=270603)
logRevenue = predict(finalModel,newdata, interval = "confidence", level = 0.95)
logRevenue
antilog(logRevenue, exp(1))

newdata = list(num_employees=218000, expense_million= 118350, enterprise_val_Million=78070)
logRevenue = predict(finalModel,newdata, interval = "confidence", level = 0.95)
logRevenue
antilog(logRevenue, exp(1))

newdata = list(num_employees=323000, expense_million= 66534, enterprise_val_Million=30080)
logRevenue = predict(finalModel,newdata, interval = "confidence", level = 0.95)
logRevenue
antilog(logRevenue, exp(1))

newdata = list(num_employees=140000, expense_million= 22556, enterprise_val_Million=834)
logRevenue = predict(finalModel,newdata, interval = "confidence", level = 0.95)
logRevenue
antilog(logRevenue, exp(1))
```