Variables with the Greatest Impact on Website Development using Count Regression

STAT 823: Summer Class Project, 2021

Madhu Peduri
Mary Duncan

Department of Biostatistics and Data Science
University of Kansas, USA
July 30, 2021

Contents

List of Tables

List of Figures

Abstract

Modeling count variables is a common task in statistical regression. Due to its assumption of Normality, classical linear regression is a more limited model in dealing with such data. The Poisson regression model performs better with response variables that are discrete and are limited to non-negative values such as the counted number of occurrences of an event. In this report, we attempt to analyze one such outcome that represents the number of websites delivered by the management of a company. We explored the dataset with respect to each covariate and made transformations when necessary. We used a Poisson regression on the given dataset as the base model and created multiple more accurate models that better represent the data by transforming data and removing outliers.

Overdispersion is a common problem in a Poisson regression. In order to mediate this, we used both a Quasi-poisson and a Negative binomial model. We conclude by discussing different metrics to determine the best fitting model and which covariates most impact the number of websites delivered.

Introduction

Statistical Regression is the concept used to determine how a variable of interest, or a dependent variable, is affected by one or more independent variables. The outcome is an equation that can be used to make predictions based on data collected. While Linear Regression is a good tool for prediction analysis, it relies on the following four assumptions:

Linearity: Linear Regression assumes that the relation between dependent and independent variables is linear.

Homoscedasticity: According to this assumption, different samples have the same variance, even if they came from different populations. Variance of the residuals will be constant even with a change in the independent variables.

Collinearity: According to this assumption, observations are non-collinear or independent to each other.

Normality: This states that, the residuals between the actual and predicted values of the dependent variable follow normal distribution.

Normality is an understood assumption for linear regression. Often, the response variable of interest is categorical or discrete, not continuous. In this case, a linear regression model cannot produce normally distributed errors. An alternative is to use a Poisson regression model or one of its variants. These models have a number of advantages over an ordinary linear regression model, including a skew, discrete distribution, and the restriction of predicted values to non-negative numbers. A Poisson model is similar to an ordinary linear regression, with the following three assumptions:

Poisson distribution: Model assumes that the response variable follows a Poisson distribution, instead of a normal distribution.

$$P(y) = \frac{\mu e^{-\mu}}{y!} \quad y = 0, 1, \ldots$$

Log function: Models the natural log of the response variable, ln(Y), as a linear function of the coefficients.

$$log(y) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

E(Y)=Var(Y): This is another aspect of the first assumption. A characteristic of the Poisson distribution is that its mean is equal to its variance. In certain circumstances, it will be found that the observed variance is greater than the mean; this is known as overdispersion and indicates that the model is not appropriate.

The objective of this paper is to study the association between the backlog, team number, team experience, and process, with the number of websites delivered to customers per quarter and to determine which covariables most impact this outcome. This poses an opportunity to increase this outcome by optimizing and/or maximizing the most influential covariates, which in turn, may increase company productivity.

Primary Analysis Objectives

The primary analysis objective of this report is to perform a Poisson regression model on the 'Website delivered dataset'. Additionally, we will validate the model assumptions and look for any overdispersion. In the case of overdispersion, we will apply a Quasi-poisson and a Negative binomial regression model. Validate different metrics applicable and determine the model equation that best fits the data. In other words, determine which covariates most impact the outcome of the number of websites delivered to customers per quarter.

Materials and Methods

Data Sources

The dataset was obtained from the Management company that develops websites and was interested in determining which variables have the greatest impact on the number of websites developed and delivered to customers per quarter. Data were collected on website production output for 13 three-person website development teams, from January 2001 through August 2002. Each line of the dataset has 7 variables and one response variable. Below is the description of each variable,

Table 1: Dataset Features and Types

	Variable Name	Type
1	idnum: Identification number	Cardinal
2	delivered: Websites delivered	Discrete (Response Variable)
3	backlog: Backlog of orders	Continous
4	teamnum: Team number	Cardinal
5	experience: Team experience	Continous
6	change: Process change	Categorical
7	year: Year	Categorical-nominal
8	quarter: Quarter	Categorical-nominal

We can determine that, our response variable, Websites delivered, is a discrete count variable which makes it suitable for poisson regression. Out of 7 features, we have 2 cardinal, 2 Continous and 3 Categorical of type. All the categorical variables are of nominal type, that is, there is no scaling factor among different categories.

Statistical Analysis

Exploratory Data Analysis

Understanding Variables: We plot the distributions of our variables to gain insight into each of them. For Continuous variables, we use r-plots and for categorical, we use barplots.

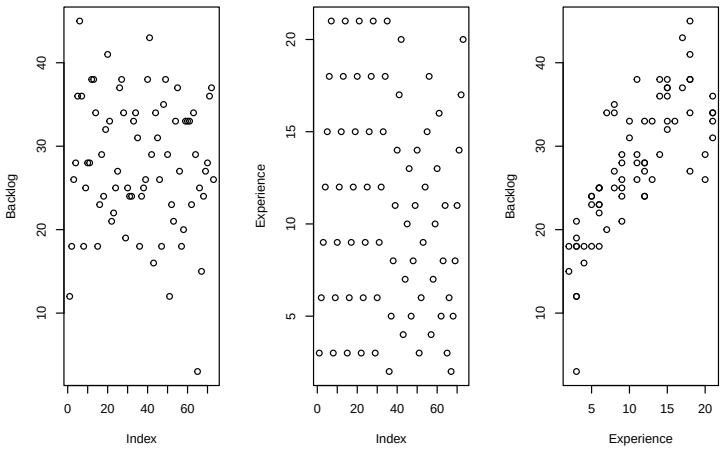

Figure 1: Continous Variables.

From Figure 1, it can be observed that both continuous variables Backlog and Experience have non-constant variance. Data is scattered across the plot suggesting no outliers. Some linearity exists between two variables.

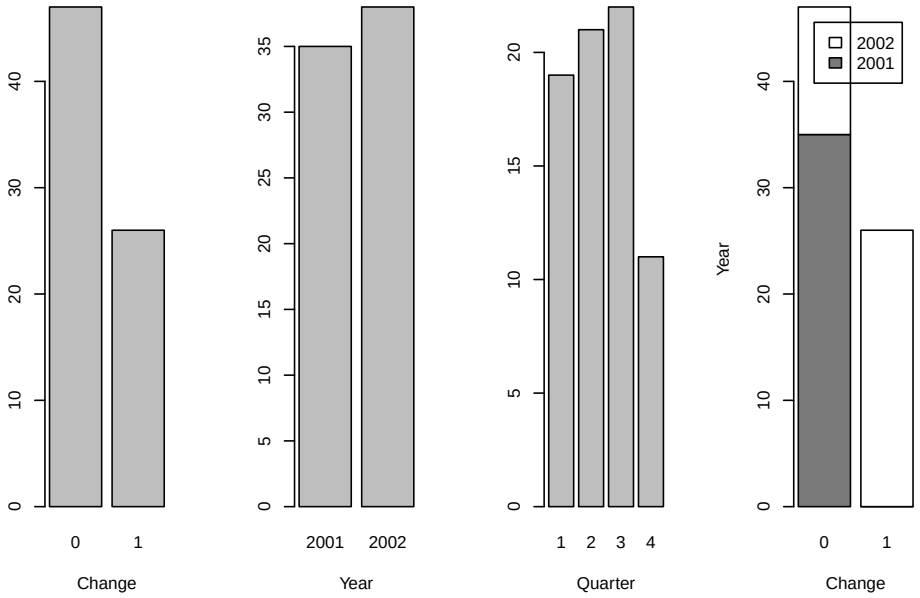

Figure 2: Categorical Variables.

From Figure 2, we observe no significant imbalance between different categories of Year or Quarter and little imbalance in the Change variable. Using this plot, we know the categories

involved, which can be used to encode them if necessary. As we know our categorical variables are nominal, we can use one-hot encoding.

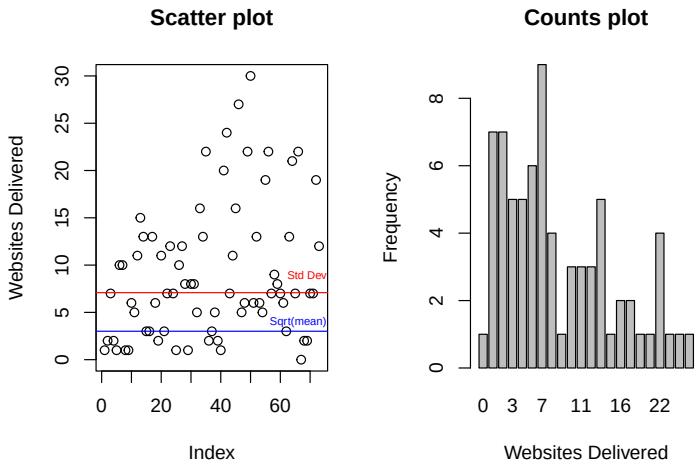

Figure 3: Response Variable.

From Figure 3, we can see a poisson distribution (for a given expected average) in the counts plot. From the scatter plot, it can be determined that our response variable has non-constant variance and its $\sqrt{mean} = 3.007$(~mean) is not equal to its standard deviation(~variance)=7.084. This difference suggests the presence of overdispersion.

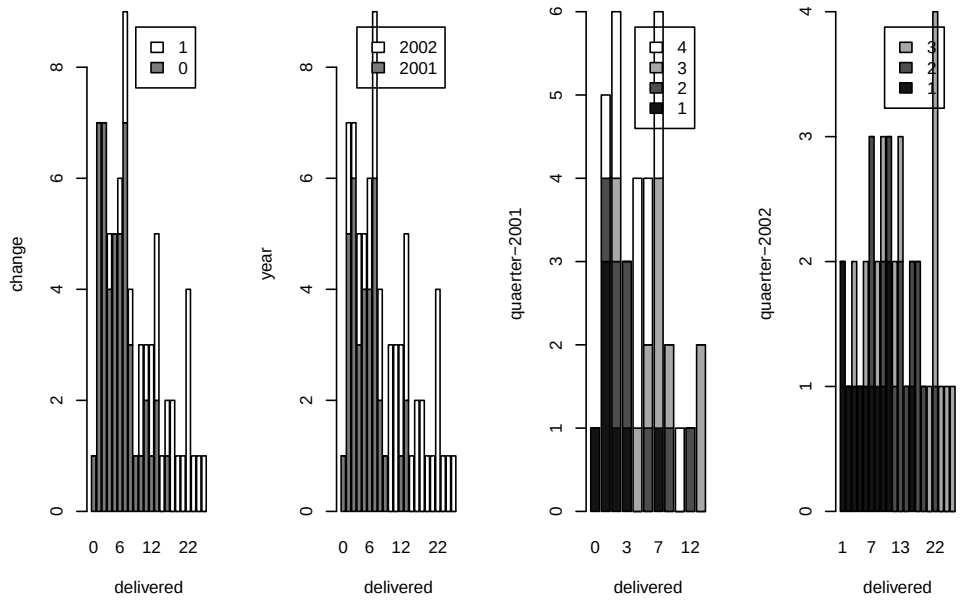

Figure 4: Relation between Categorical and Response Variable.

From Figure 4, We can establish the relation between different categorical variables and the response variable. We plotted the distribution of the response variable 'delivered' highlighting the contriubution of each category for a given categorical variable.

Change: We can see approximately equal contribution from both categories, 0 and 1, towards the response variable.

Year: We have two categories 2001 and 2002 which contribute equally toward the response variable.

Quarter: Each year 2001 and 2002 have been divided across 4 quaters. We plotted the contribution of each quater per year towards the response variable. For year 2001, we have websites delivered in all quarters and for 2002, we do not have any websites delivered in quarter 4. But we believe imbalance created by this will not be significant.

From above observations, we can determine that, all categories are nominal without any sacling among themselves and contributed towards the response variable. This would suggest the division of categories in to individual features can contribute to model better than combined features.

Table 2: Dataset summary

	ï..idnum	delivered	backlog	teamnum	experience	change	year	quarter
Min	1.00	0.00	3.00	1.00	2.00	0.00	2001.00	1.00
1st Qu	19.00	3.00	23.00	3.00	6.00	0.00	2001.00	1.00
Median	37.00	7.00	28.00	6.00	11.00	0.00	2002.00	2.00
Mean	37.00	9.04	27.82	6.29	10.85	0.36	2001.52	2.34
3rd Qu	55.00	13.00	34.00	9.00	15.00	1.00	2002.00	3.00
Max	73.00	30.00	45.00	13.00	21.00	1.00	2002.00	4.00

The observations from the plots can be identified from the summary of the dataset as well. We do not see significant differences between mean and media of continous variables suggesting no outliers. No significant skewness among the features.

Data Transformation: We perform below changes to the given dataset according to the observations made from data analysis.

- We have two cardinal variables which acts as indexes. We believe these features do not contriubte to the model.
- We have three categorical variables, Change, Year and quarter. We encode them such that each category will be fabricated as a separate feature. For example, Variable 'Change' has 0 and 1 as categories. We fabricate two features 'Change_0', which will have 1 where 'Change=0' and 'Change_1', which will have 1, where 'Change=1'.
- We observed some skewness (not significant) in the continous variables background and experience. We will try using log transformation while building the model.

Model Assumptions

Below statistical parameters are used to determine a good model:

Table 3: Features after transformation

	1	2	3	4
1	delivered	backlog	teamnum	experience
2	change_0	change_1	year_01	year_02
3	quarter_1	quarter_2	quarter_3	quarter_4

Deviance: This measures the unexplained variance by the model. Low deviance is good.

Chisquare: This measures how much the fitted values differ from the expected values. Lesser the chisquare more good the model is.

R-squared: This measures the percentage of variance, of the response variable, measured by using the independent features collectively. Higher R-squared is preferred.

AIC and BIC: These are probabilistic model selection parameters. These represent how bad a moder performed on the training dataset and its complexity. Model with low values are preferred.

Primary Objective Analysis

Univariate Analysis:

First we perform univariate analysis and determine if any feature has enough relation to explain the variance of the response variable.

Using Backlog feature: From Figure 5, we can observe that regression model's fit is good from Table 4, we can observe the standard error is low for the feature.

Table 4: Model summary with Backlog feature

	Estimate	Std. Error	z value	Pr(>|z|)	RR	2.5 %	97.5 %
(Intercept)	1.11	0.16	6.89	0.00	1.11	0.79	1.42
backlog	0.04	0.01	7.29	0.00	0.04	0.03	0.05

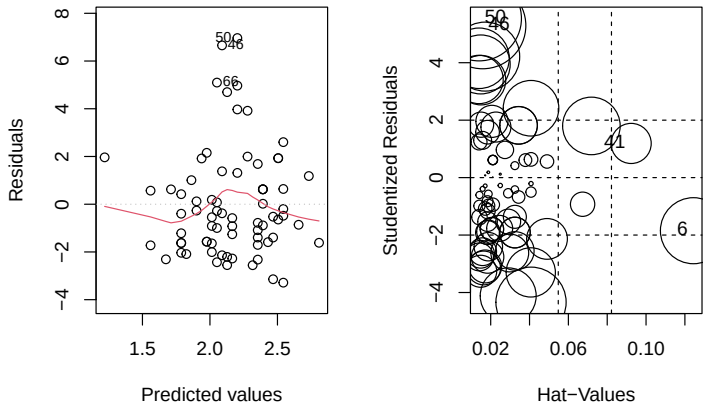

Figure 5: Regression plots using backlog feature.

Using Experience feature: From Figure 6, we can observe that regression model's fit is good from Table 5, we can observe the standard error is low for the feature.

Table 5: Model summary with Backlog feature

	Estimate	Std. Error	z value	Pr(>|z|)	RR	2.5 %	97.5 %
(Intercept)	1.48	0.10	15.39	0.00	1.48	1.29	1.67
experience	0.06	0.01	8.81	0.00	0.06	0.05	0.07

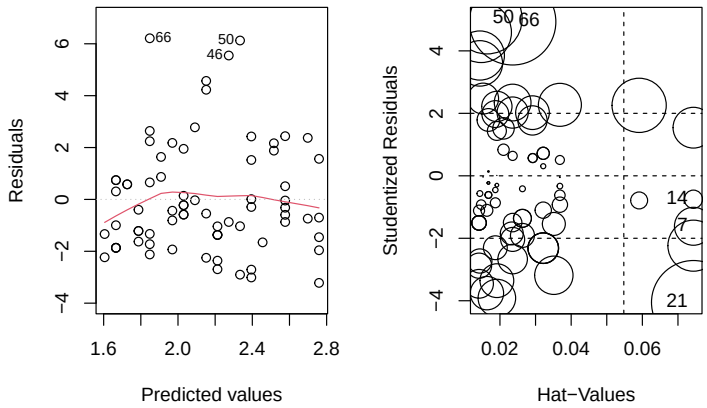

Figure 6: Regression plots using backlog feature.

Below Table 6 shows the consolidated statistics for above univariate analysis. We can determine that the metrics Deviance, AIC and BIC are very high suggesting the both the

models failed at explaining the variance in the response variable. R-squared values are too low suggesting that individual features do not explain the pattern in the response variable.

Table 6: Statistics for univariate Analysis

	Deviance	Dof	Chisq	R2	AIC	BIC	Gof.Dev	Gof.Pearson
Backlog	335.70	71.00	0.00	0.08	608.95	613.54	335.70	354.75
Experience	312.58	71.00	0.00	0.12	585.83	590.41	312.60	321.87

Poisson Regression Models:

Base Model: Figure 7 and Table 7 shows the summary of the base model. In this model we used all the features without transformation. We fabricated a multiplicative feature from (backlog * experience). We can observe that residuals has constant variance and regression line hovers around its mean suggesting a good fit. We can see less P-values from the summary. We can also observe some outliers from Influence plot.

Table 7: Model summary for base model

	Estimate	Std. Error	z value	Pr(>|z|)	RR	2.5 %	97.5 %
(Intercept)	-1114.50	425.42	-2.62	0.01	-1114.50	-1946.84	-277.40
backlog	0.05	0.01	3.77	0.00	0.05	0.02	0.08
experience	0.08	0.04	2.08	0.04	0.08	0.01	0.16
change	0.75	0.16	4.60	0.00	0.75	0.43	1.07
year	0.56	0.21	2.62	0.01	0.56	0.14	0.97
quarter	0.17	0.08	2.25	0.02	0.17	0.02	0.32
backlog:experience	-0.00	0.00	-3.20	0.00	-0.00	-0.01	-0.00

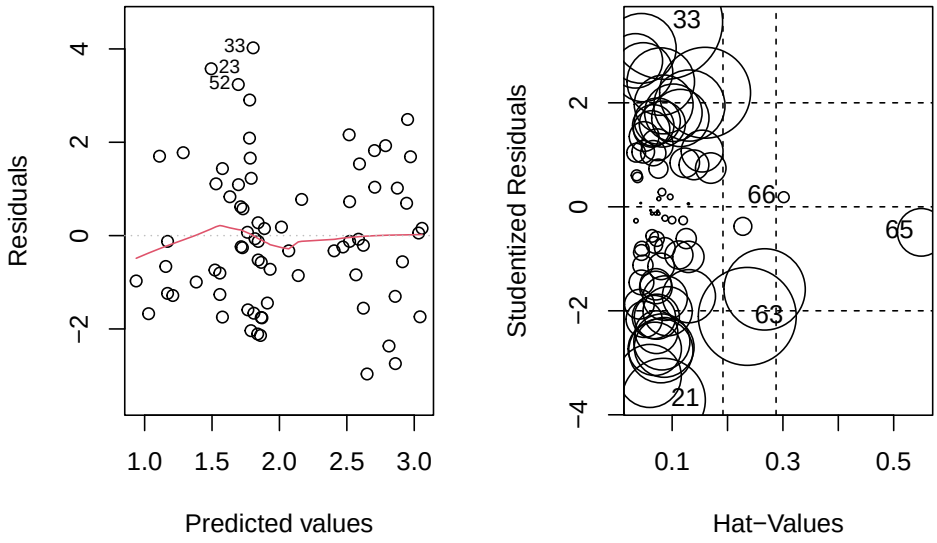

Figure 7: Regression plots for base model.

Model with Encoded features: Figure 8, Table 8 provided the summary for this model. We can observe a good regression line and higher P-values and standard errors compared to base model. We can also observe some outliers from Influence plot.

Table 8: Model summary with encoded features

	Estimate	Std. Error	z value	Pr(>|z|)	RR	2.5 %	97.5 %
(Intercept)	1.78	0.58	3.05	0.00	1.78	0.61	2.89
backlog	0.04	0.01	3.19	0.00	0.04	0.02	0.07
experience	0.05	0.04	1.38	0.17	0.05	-0.02	0.13
change_1	0.11	0.32	0.34	0.74	0.11	-0.55	0.71
year_01	-1.00	0.32	-3.18	0.00	-1.00	-1.65	-0.41
quarter_1	-0.76	0.32	-2.41	0.02	-0.76	-1.41	-0.16
quarter_2	0.00	0.20	0.02	0.98	0.00	-0.38	0.39
quarter_3	0.27	0.17	1.57	0.12	0.27	-0.06	0.61
backlog:experience	-0.00	0.00	-2.48	0.01	-0.00	-0.01	-0.00

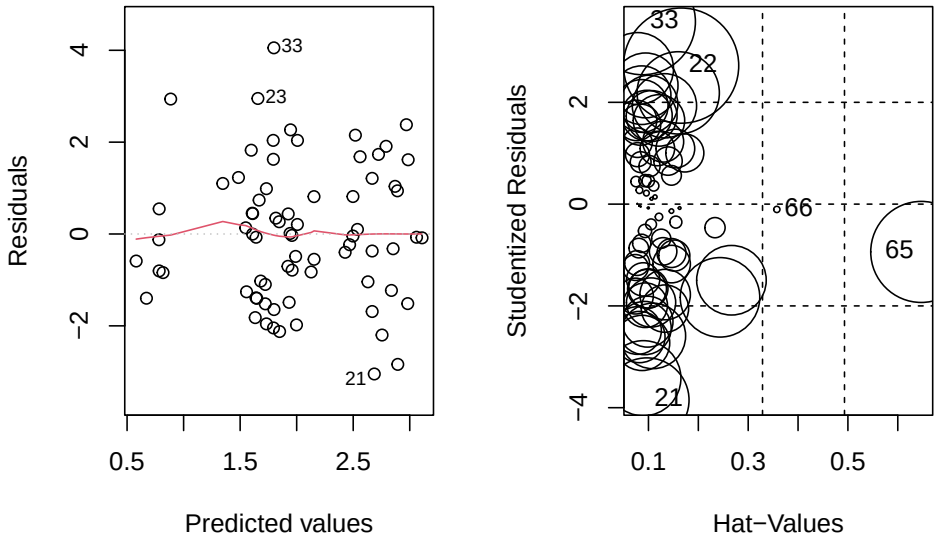

Figure 8: Regression plots for model with encoded features.

Without outliers: Figure 9 and Table 9 provide the summary for this model. We built this model on the subset of the data formed by removing the outliers identified in earlier models. We can see the regression line fits better compared to above two models. We have P-values less than the previous models. We further analyze the statistical metrics for these models.

Table 9: Model summary without outliers

	Estimate	Std. Error	z value	Pr(>|z|)	RR	2.5 %	97.5 %
(Intercept)	2.42	0.85	2.86	0.00	2.42	0.72	4.05
backlog	0.03	0.02	1.30	0.19	0.03	-0.01	0.07
experience	-0.01	0.05	-0.14	0.89	-0.01	-0.10	0.09
change_1	0.23	0.34	0.67	0.50	0.23	-0.47	0.88
year_01	-1.08	0.38	-2.87	0.00	-1.08	-1.83	-0.35
quarter_1	-0.97	0.34	-2.86	0.00	-0.97	-1.66	-0.32
quarter_2	-0.16	0.21	-0.79	0.43	-0.16	-0.56	0.24
quarter_3	0.29	0.17	1.69	0.09	0.29	-0.04	0.63
backlog:experience	-0.00	0.00	-0.88	0.38	-0.00	-0.00	0.00

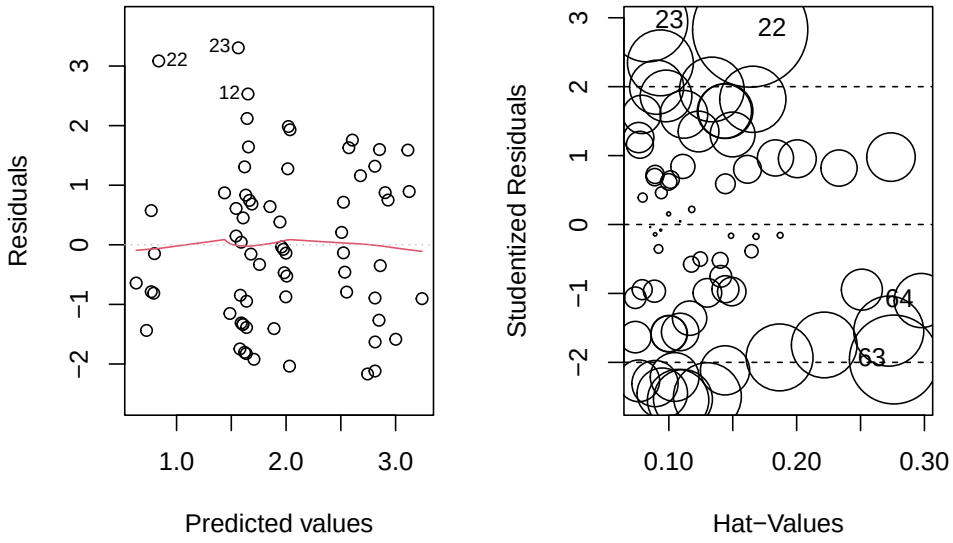

Figure 9: Regression plots for model without outliers

Quasi-Poisson: In our EDA, we established that our response variable has overdispersion. Figure 10 and Table 10 provided the summary of this model. We can see P-values similar to the model without outliers and a simiar fit for regression line in residual plots.

Table 10: Summary for Quasi-Poisson Model

	Estimate	Std. Error	t value	Pr(>|t|)	RR	2.5 %	97.5 %
(Intercept)	2.42	1.21	2.01	0.05	2.42	-0.03	4.71
backlog	0.03	0.03	0.92	0.36	0.03	-0.03	0.09
experience	-0.01	0.07	-0.10	0.92	-0.01	-0.14	0.13
change_1	0.23	0.49	0.47	0.64	0.23	-0.78	1.15
year_01	-1.08	0.53	-2.02	0.05	-1.08	-2.16	-0.05
quarter_1	-0.97	0.48	-2.01	0.05	-0.97	-1.96	-0.05
quarter_2	-0.16	0.29	-0.56	0.58	-0.16	-0.73	0.42
quarter_3	0.29	0.24	1.19	0.24	0.29	-0.18	0.78
backlog:experience	-0.00	0.00	-0.62	0.54	-0.00	-0.01	0.00

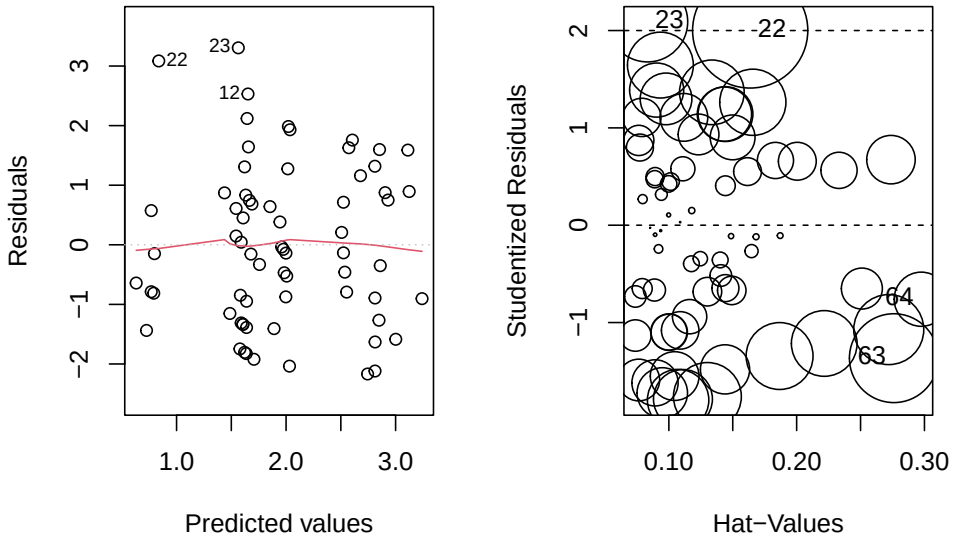

Figure 10: Regression plots for Quasi-Poisson Model

Negative Binomial: This model is also used to deal with overdispersion in the response variable. Figure 11 and Table 11 provided the summary for this model. We can see better P-values compared to previous models and a better fitted regression line against its residuals. We further analyze the metrics for this model to evaluate.

Table 11: Summary for Negative-Binomial Model

	Estimate	Std. Error	z value	Pr(>|z|)	RR	2.5 %	97.5 %
(Intercept)	2.49	1.05	2.38	0.02	2.49	0.43	4.50
backlog	0.03	0.03	1.05	0.30	0.03	-0.02	0.08
experience	-0.02	0.07	-0.30	0.77	-0.02	-0.15	0.11
change_1	0.23	0.40	0.59	0.56	0.23	-0.56	1.00
year_01	-1.09	0.44	-2.48	0.01	-1.09	-1.97	-0.24
quarter_1	-1.01	0.39	-2.55	0.01	-1.01	-1.79	-0.26
quarter_2	-0.20	0.26	-0.75	0.45	-0.20	-0.71	0.32
quarter_3	0.29	0.21	1.36	0.17	0.29	-0.12	0.70
backlog:experience	-0.00	0.00	-0.52	0.60	-0.00	-0.00	0.00

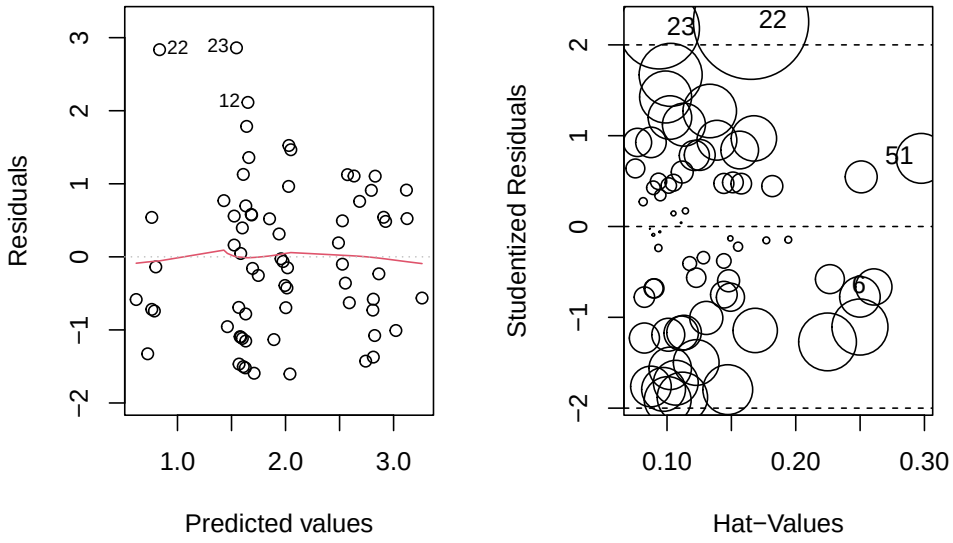

Figure 11: Regression plots for Negative Binomial Model

Secondary Objective Analysis

We attempted multiple models to determine that Negative-Binomial model performed better compared to other Models. This is because of the presence of overdispersion in response variable. We further apply step function to this Model to determine the best set of features that has greatest impact on it.

Start: AIC=380.24 delivered ~ backlog + experience + change_0 + change_1 + year_01 + year_02 + quarter_1 + quarter_2 + quarter_3 + quarter_4 + backlog:experience

Step: AIC=380.24 delivered ~ backlog + experience + change_0 + change_1 + year_01 + year_02 + quarter_1 + quarter_2 + quarter_3 + backlog:experience

Step: AIC=380.24 delivered ~ backlog + experience + change_0 + change_1 + year_01 + quarter_1 + quarter_2 + quarter_3 + backlog:experience

Step: AIC=380.24 delivered ~ backlog + experience + change_1 + year_01 + quarter_1 + quarter_2 + quarter_3 + backlog:experience

`Df Deviance AIC`

- backlog:experience 1 78.6 379
- change_1 1 78.7 379
- quarter_2 1 78.9 379
- quarter_3 1 80.2 380 78.4 380
- year_01 1 84.7 385
- quarter_1 1 85.3 385

Step: AIC=378.5 delivered ~ backlog + experience + change_1 + year_01 + quarter_1 + quarter_2 + quarter_3

```
            Df Deviance AIC
```

- change_1 1 78.3 377
- quarter_2 1 79.0 377
- backlog 1 79.2 378
- quarter_3 1 80.0 378 78.1 379
- experience 1 83.7 382
- year_01 1 87.9 386
- quarter_1 1 89.0 387

Step: AIC=376.69 delivered ~ backlog + experience + year_01 + quarter_1 + quarter_2 + quarter_3

```
            Df Deviance AIC
```

- quarter_2 1 79.0 375
- backlog 1 79.2 376 78.3 377
- quarter_3 1 80.5 377
- experience 1 83.7 380
- quarter_1 1 97.0 393
- year_01 1 113.0 409

Step: AIC=375.42 delivered ~ backlog + experience + year_01 + quarter_1 + quarter_3

```
            Df Deviance AIC
```

- backlog 1 79.0 374 78.5 375
- experience 1 83.1 378
- quarter_3 1 87.2 382
- quarter_1 1 113.1 408
- year_01 1 114.9 410

Step: AIC=373.95 delivered ~ experience + year_01 + quarter_1 + quarter_3

```
            Df Deviance AIC
```

78.8 374 - experience 1 83.0 376 - quarter_3 1 87.5 381 - quarter_1 1 114.0 407 - year_01 1 129.6 423

Call: glm.nb(formula = delivered ~ experience + year_01 + quarter_1 + quarter_3, data = wdf[-c(21, 33, 61, 65, 66),], init.theta = 11.43972939, link = log)

Coefficients: (Intercept) experience year_01 quarter_1
3.2040 -0.0356 -1.3263 -1.0140
quarter_3
0.3815

Degrees of Freedom: 67 Total (i.e. Null); 63 Residual Null Deviance: 213 Residual Deviance: 78.8 AIC: 376

Results

We have performed multiple models as part of primary analysis and computed aforementioned Goodness of fit metrics to evaluate the best model. Table 12 is the result of that analysis. We can see that Negative Binomial model has lowest Deviance, AIC, BIC and Pearson Goodness of fit metrics. This suggests that Negative binomial model is suited well for this dataset explaining the variance of the response variable better than other models.

Table 12: Goodness of Fit Metrics

	Deviance	Dof	Chisq	R2	AIC	BIC	Gof.Dev	Gof.Pearson
Base Poisson	175.83	66.00	0.00	0.33	459.09	475.12	175.83	168.78
Encoded for Categorical	165.58	64.00	0.00	0.34	452.83	473.45	165.57	156.88
Without Outliers	123.25	59.00	0.00	0.39	390.49	410.47	123.25	118.97
Quasi-poisson	123.25	59.00	0.00	0.39			123.25	118.97
Negative Binomial	78.36	59.00	0.05	0.16	382.24	404.43	78.36	74.41

We used Step method to deduce the set of features that has greatest impact on the model. Below Table 13 is the result of the secondary analysis that gives below equation for prediction.

$log(Delivered) =$
$2.49 + 0.03 * backlog - 0.02 * experience - 1.09 * year01 - 1.01 * quarter1 + 0.29 * quarter3$

Table 13: Important Features

	Features	Coefficients
1	backlog	0.03
2	experience	-0.02
3	year_01	-1.09
4	quarter_1	-1.01
5	quarter_3	0.29

All of the statistical analyses in this document will be performed using R version 4.1.0 (2021-05-18) R (R Core Team, 2018). R packages used will be maintained using the packrat dependency management system.

Discussion and Conclusion

Features that were gathered by the process of counting the occurances of an event exhibit Poisson distribution. Our response variable is the count of websites delivered. To find the pattern and predict a count variable, poisson regression and its variants are used. However, a poisson model assume that the average expected value of the variable is equal to its Variance. In our case Variance is higher which is called as overdispersion. We have computed multiple models to determine that Negative binomial works better for overdispersion. Our goodness of fit metrics were not ideal due to the fact that the dataset do not possess sufficient pattern to be modeled. We further used step method to determine the best set of features that had greatest impact on the model.

Appendix: R-code

This area is where you can include the code that was used to do the analysis.

```
rm(list = ls(all=TRUE))
# load packages
library(knitr)
library(formatR)
library(stargazer)
library(xtable)
library(graphics)
library(ggplot2)
library(MASS)
library(Hmisc)
library(epiDisplay)
library(vcd)
library(mnormt)
library(MASS)
library(car)
library(ggpubr)
library(PairedData)
library(lmtest)
library(faraway)
library(leaps)
library(psych)
library(Matrix)
library(dobson)
library(jtools)
library(Rcpp)
library(pscl)
library(effects)

v1 <- c('idnum: Identification number','delivered: Websites delivered','backlog: Backlog of orc
v2 <- c('Cardinal','Discrete (Response Variable)','Continous','Cardinal','Continous','Categoric
td <- data.frame(cbind(v1,v2))
names(td) <- c('Variable Name','Type')
row.names(td) <- NULL
xtable(td,label = 'tab:tab1',caption = 'Dataset Features and Types')

wd <- read.csv("Data\\website.csv")

par(mfrow = c(1, 3))
plot(wd$backlog,ylab='Backlog')
plot(wd$experience,ylab='Experience')
plot(wd$backlog~wd$experience,xlab='Experience',ylab='Backlog')

par(mfrow = c(1, 4))
barplot(table(wd$change),xlab='Change')
```

```
barplot(table(wd$year),xlab='Year')
barplot(table(wd$quarter),xlab='Quarter')
counts <- table(wd$year,wd$change)
barplot(counts,col=c('gray48','gray100'),legend = rownames(counts),xlab='Change',ylab='Year')

par(mfrow = c(1, 2))
plot(wd$delivered, xlab='Index', ylab='Websites Delivered',main='Scatter plot')
abline(h=c(sqrt(mean(wd$delivered)),sd(wd$delivered)),col=c('blue','red'))
text(69,9.0,'Std Dev',cex=0.6,col='red')
text(66,4.0,'Sqrt(mean)',cex=0.6,col='blue')
barplot(table(wd$delivered), xlab='Websites Delivered', ylab='Frequency',main='Counts plot')

par(mfrow = c(1, 4))
counts <- table(wd$change,wd$delivered)
barplot(counts,col=c('gray48','gray100'),legend = rownames(counts),xlab='delivered',ylab='chang

counts <- table(wd$year,wd$delivered)
barplot(counts,col=c('gray48','gray100'),legend = rownames(counts),xlab='delivered',ylab='year

wd2001 <- wd[wd$year == '2001',]
counts <- table(wd2001$quarter,wd2001$delivered)
barplot(counts,col=c('gray8','gray30','gray66','gray100'),legend = rownames(counts),xlab='deliv

wd2002 <- wd[wd$year == '2002',]
counts <- table(wd2002$quarter,wd2002$delivered)
barplot(counts,col=c('gray8','gray30','gray66','gray100'),legend = rownames(counts),xlab='deliv

wdfm <- as.data.frame(sapply(wd,function(x) cbind(summary(x))))
row.names(wdfm) <- c('Min','1st Qu','Median','Mean','3rd Qu','Max')
xtable(wdfm,label = 'tab:tab2',caption = 'Dataset summary')

# Data transformation
names(wd)[1] = "id"

wdf <- data.frame(
  delivered = as.numeric(wd$delivered),
  backlog = as.numeric(wd$backlog),
  teamnum = as.numeric(wd$teamnum),
  experience = as.numeric(wd$experience),
  change_0 = as.numeric(wd$change),
  change_1 = as.numeric(wd$change),
  year_01 = as.numeric(wd$year),
  year_02 = as.numeric(wd$year),
  quarter_1 = as.numeric(wd$quarter),
  quarter_2 = as.numeric(wd$quarter),
  quarter_3 = as.numeric(wd$quarter),
  quarter_4 = as.numeric(wd$quarter)
```

```
)

wdf$change_0[wdf$change_0 == 0] <- 1
wdf$change_0[wdf$change_0 == 1] <- 0
wdf$change_1[wdf$change_1 == 0] <- 0
wdf$change_1[wdf$change_1 == 1] <- 1

wdf$year_01[wdf$year_01 == 2001] <- 1
wdf$year_01[wdf$year_01 == 2002] <- 0
wdf$year_02[wdf$year_02 == 2001] <- 0
wdf$year_02[wdf$year_02 == 2002] <- 1

wdf$quarter_1[wdf$quarter_1 != 1] <- 0
wdf$quarter_1[wdf$quarter_1 == 1] <- 1
wdf$quarter_2[wdf$quarter_2 != 2] <- 0
wdf$quarter_2[wdf$quarter_2 == 2] <- 1
wdf$quarter_3[wdf$quarter_3 != 3] <- 0
wdf$quarter_3[wdf$quarter_3 == 3] <- 1
wdf$quarter_4[wdf$quarter_4 != 4] <- 0
wdf$quarter_4[wdf$quarter_4 == 4] <- 1

newcol <- matrix(names(wdf), nrow = 3, byrow = TRUE)
xtable(newcol,label = 'tab:tab3',caption = 'Features after transformation')

modelstat <- function(mod,dtf){
  # Summaries
  mods1 <- summary(mod)
  mods2 <- summ(mod, confint = TRUE, digits = 3, ci.width = 0.95)

  # Attributes
  mods2atr <- attributes(mods2)

  # Goodness of fit
  devresd <- round(residuals(mod, type = "deviance"), 3)
  devgof <- sum(devresd^2)

  pearson.resid <- round(resid(mod, type = "pearson"), 3)
  prsgof <- sum(pearson.resid^2)

  # Residual plots
  stddevres <- rstandard(mod)
  stdpearsons <- rstandard(mod, type = "pearson")

  #poisson gof
  chiq <- pchisq(deviance(mod), df.residual(mod), lower = F)

  c(mods1$deviance,mods1$df.residual,chiq,mods2atr$rsqmc,mods2atr$aic,mods2atr$bic,devgof,prsgo
```

```
}
modelplot <- function(mod){
  # plots
  par(mfrow = c(1, 2))
  plot(mod, which = c(1), main = "", caption = "")
  ip <- influencePlot(mod)
}
modelsumm <- function(mod,lb,cp){
  # stat
  xtable(cbind( na.omit(summary(mod)$coef) ,RR= na.omit(coef(mod)), na.omit(confint(mod)) ),lab
}

glm1 <- glm(delivered ~ backlog, family = poisson(link = "log"),data = wdf)
glm1stat <- modelstat(glm1,wdf)
instatdf <- data.frame(rbind(glm1stat))
rownames(instatdf)[1] <- 'Backlog'
names(instatdf) <- c('Deviance','Dof','Chisq','R2','AIC','BIC','Gof-Dev','Gof-Pearson')
modelsumm(glm1,'tab:tab4',"Model summary with Backlog feature")

modelplot(glm1)

glm2 <- glm(delivered ~ experience, family = poisson(link = "log"),data = wdf)
glm2stat <- modelstat(glm2,wdf)
instatdf <- data.frame(rbind(instatdf,glm2stat))
rownames(instatdf)[2] <- 'Experience'
modelsumm(glm2,'tab:tab5',"Model summary with Backlog feature")

modelplot(glm2)

xtable(instatdf,label = 'tab:tab6', caption = 'Statistics for univariate Analysis')

rmv <- c("id","delivered","teamnum")
cn <- names(wd)
cn <- setdiff(cn,rmv)
cn <- c(cn,"backlog:experience")
fm <- as.formula(paste("delivered", paste(cn, collapse = "+"), sep = "~"))
glm3 <- glm(fm,family = poisson(link = "log"), data = wd)
glm3stat <- modelstat(glm3,wdf)
statdf <- data.frame(rbind(glm3stat))
rownames(statdf)[1] <- 'Base Poisson'
names(statdf) <- c('Deviance','Dof','Chisq','R2','AIC','BIC','Gof-Dev','Gof-Pearson')
modelsumm(glm3,'tab:tab7',"Model summary for base model")

modelplot(glm3)

rmv <- c("delivered","teamnum")
cn <- names(wdf)
```

```
cn <- setdiff(cn,rmv)
cn <- c(cn,"backlog:experience")
fm <- as.formula(paste("delivered", paste(cn, collapse = "+"), sep = "~"))
glm4 <- glm(fm,family = poisson(link = "log"), data = wdf)
glm4stat <- modelstat(glm4,wdf)
statdf <- data.frame(rbind(statdf,glm4stat))
rownames(statdf)[2] <- 'Encoded for Categorical'
modelsumm(glm4,'tab:tab8',"Model summary with encoded features")

modelplot(glm4)

rmv <- c("delivered","teamnum")
cn <- names(wdf)
cn <- setdiff(cn,rmv)
cn <- c(cn,"backlog:experience")
fm <- as.formula(paste("delivered", paste(cn, collapse = "+"), sep = "~"))
glm5 <- glm(fm,family = poisson(link = "log"), data = wdf[-c(21,33,61, 65,66),])
glm5stat <- modelstat(glm5,wdf[-c(21,33,61, 65,66),])
statdf <- data.frame(rbind(statdf,glm5stat))
rownames(statdf)[3] <- 'Without Outliers'
modelsumm(glm5,'tab:tab9',"Model summary without outliers")

modelplot(glm5)

rmv <- c("delivered","teamnum")
cn <- names(wdf)
cn <- setdiff(cn,rmv)
cn <- c(cn,"backlog:experience")
fm <- as.formula(paste("delivered", paste(cn, collapse = "+"), sep = "~"))
glm6 <- glm(fm,family = quasipoisson(link = "log"), data = wdf[-c(21,33,61, 65,66),])
glm6stat <- modelstat(glm6,wdf[-c(21,33,61, 65,66),])
statdf <- data.frame(rbind(statdf,glm6stat))
rownames(statdf)[4] <- 'Quasi-poisson'
modelsumm(glm6,'tab:tab10',"Summary for Quasi-Poisson Model")

modelplot(glm6)

rmv <- c("delivered","teamnum")
cn <- names(wdf)
cn <- setdiff(cn,rmv)
cn <- c(cn,"backlog:experience")
fm <- as.formula(paste("delivered", paste(cn, collapse = "+"), sep = "~"))
glm7 <- glm.nb(fm, data = wdf[-c(21,33,61, 65,66),])
glm7stat <- modelstat(glm7,wdf[-c(21,33,61, 65,66),])
statdf <- data.frame(rbind(statdf,glm7stat))
rownames(statdf)[5] <- 'Negative Binomial'
modelsumm(glm7,'tab:tab11',"Summary for Negative-Binomial Model")
```

```
modelplot(glm7)

step(glm7)

xtable(statdf,label = 'tab:tab12', caption = 'Goodness of Fit Metrics')
xtable(data.frame(Features = c('backlog','experience','year_01','quarter_1','quarter_3'),Coeffi
```

Bibliography

R Core Team (2018), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, URL https://www.R-project.org/.