Regression Analysis
Poisson Regression

Nicoleta Serban, Ph.D.
Professor
School of Industrial and Systems Engineering

Predicting Demand for Rental Bikes: Poisson Regression

Georgia Tech

1

About This Lesson

Georgia Tech

2

Predicting Demand for Rental Bikes

Bike sharing systems are of great interest due to their important role in traffic management.

Dataset: Historical data for years 2011-2012 for the bike sharing system in Washington D.C.

Data Source: UCI Machine Learning Repository

***Acknowledgement**: This example was prepared with support from students in the Masters of Analytics program, including Naman Arora, Puneeth Banisetti, Mani Chandana Chalasani, Joseph (Mike) Tritchler and Kevin West*

Georgia Tech

3

Response & Predicting Variables

The response variable is:

Y (*Cnt*): Total bikes rented by both casual & registered users together

The qualitative predicting variables are:

Season: Season which the observation is made (1 = Winter, 2 = Spring, 3 = Summer, 4 = Fall)
Yr: Year on which the observation is made
Mnth: Month on which the observation is made
Hr: Day on which the observation is made (0 through 23)
Holiday: Indictor of a public holiday or not (1 = public holiday, 0 = not a public holiday)
Weekday: Day of week (0 through 6)
Weathersit: Weather condition (1 = Clear, Few clouds, Partly cloudy, Partly cloudy, 2 = Mist & Cloudy, Mist & Broken clouds, Mist & Few clouds, Mist, 3 = Snow, Rain, Thunderstorm & Scattered clouds, Ice Pallets & Fog)

The quantitative predicting variables are:

Temp: Normalized temperature in Celsius
Atemp: Normalized feeling temperature in Celsius
Hum: Normalized humidity
Windspeed: Normalized wind speed

Georgia Tech

4

Poisson Regression Analysis in R

```
# Applying multiple linear regression model
model1 = glm(cnt ~ ., data=train, family='poisson')
summary(model1)
```

	Estimate	Std. Error	z-value	Pr(>|z|)
(Intercept)	2.93659	0.007629	384.941	<2e-16
season2	0.265486	0.004129	64.298	<2e-16
season3	0.255689	0.00473	54.059	<2e-16
season4	0.448706	0.004582	97.918	<2e-16
yr1	0.4684	0.001289	363.518	<2e-16
mnth2	0.115282	0.004247	27.143	<2e-16
mnth3	0.235149	0.004422	53.179	<2e-16
mnth4	0.210302	0.005857	35.909	<2e-16
mnth5	0.271895	0.006138	44.295	<2e-16
mnth6	0.2239	0.006247	35.84	<2e-16

⋮

Deviance Residuals:

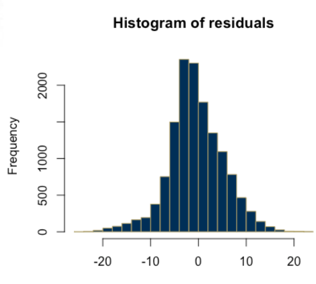
Min	1Q	Median	3Q	Max
-24.6089	-3.7805	-0.8685	3.0436	22.6553

In the full output there are 51 predictor rows in addition to the intercept.

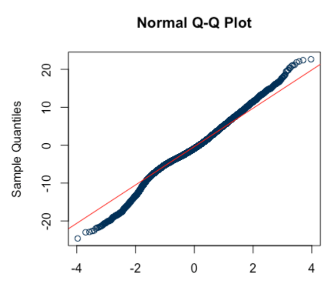
- All predicting variables are statistically significantly explaining the variability in the response (all p-values are small)
- Inflated statistical significance is also an issue in Poisson regression when the sample size is large

Goodness of Fit

```
## Checking normality
# histogram
hist(resids1,
    nclass=20,
    col=gtblue,
    border=techgold,
    main="Histogram of residuals")
```




```
# q-q plot
qqnorm(resids1,
    col="gtblue")
qqline(resids1,
    col="red")
```




```
# GOF Test
with(model1, cbind(res.deviance = deviance, df = df.residual,
        p = pchisq(deviance, df.residual, lower.tail=FALSE)))
```

res.deviance	df	p
1458653.4	13851	0

Prediction

Read New Data (Test Data)
```
test=data[-picked,]
test <- test[-c(1,2,9,15,16)]
```

Prepare the test data the same as the training data
Convert the numerical categorical variables to predictors in the test data
```
test$season = as.factor(test$season)
test$yr = as.factor(test$yr)
test$mnth = as.factor(test$mnth)
test$hr = as.factor(test$hr)
test$holiday = as.factor(test$holiday)
test$weekday = as.factor(test$weekday)
test$weathersit = as.factor(test$weathersit)
```

Build a prediction for model1 with the test data
Specify whether a confidence or prediction interval
```
pred = predict(model1, test, interval = 'prediction')
```

7

Prediction Accuracy: Model 1

Save Predictions to compare with observed data
```
test.pred1 <- predict(model1, test, type='response')
```

Mean Squared Prediction Error (MSPE)
```
mean((test.pred1-test$cnt)^2)
[1] 8060.083
```

Mean Absolute Prediction Error (MAE)
```
mean(abs(test.pred1-test$cnt))
[1] 59.96461
```

Mean Absolute Percentage Error (MAPE)
```
mean(abs(test.pred1-test$cnt)/test$cnt)
[1] 0.8214892
```

Precision Measure (PM)
```
sum((test.pred1-test$cnt)^2)/sum((test$cnt-mean(test$cnt))^2)
[1] 0.2425596
```

Accuracy Measures

MSPE = $\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \overset{*}{Y}_i\right)^2$

MAE = $\frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \overset{*}{Y}_i\right|$

MAPE = $\frac{1}{n}\sum_{i=1}^{n}\frac{\left|Y_i-\overset{*}{Y}_i\right|}{Y_i}$

PM = $\frac{\sum_{i=1}^{n}\left(Y_i-\overset{*}{Y}_i\right)^2}{\sum_{i=1}^{n}(Y_i-\bar{Y}_i)^2}$

Prediction Accuracy
MSPE = 8060.08
MAE = 59.96
MAPE = 0.82
PM = 0.243

8

Model Comparison

Model	MSPE	MAE	MAPE	PM
Full MLR	10304.95	74.52	2.72	0.310
MLR Transformed	8955.41	62.69	0.80	0.271
Poisson Reg	8060.08	59.96	0.82	0.243

- **The Poisson regression models outperform the multi-variable linear regression models in terms of predictive power across most prediction measures except MAPE.**
- **While the GOF test rejects the null of good fit, the deviance residuals seem approximately normally distributed.**

9

Summary

10