Regression Analysis

Multiple Linear Regression

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

Predicting Demand for Rental Bikes: Prediction, Interpretation

1

About This Lesson

2

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')

Apply similar R code for the other two models.

3

Prediction (cont'd)

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')

Prediction Output

	Fit	lwr	upr
6	-104.3303581	-3.038988e+02	95.238132
9	239.0013629	3.941481e+01	438.587917
30	-82.5358710	-2.822639e+02	117.192193
35	58.5579012	-1.410152e+02	258.130976
38	22.5421861	-1.770914e+02	222.175777
44	102.8402463	-9.671724e+01	302.397729
47	-40.1522581	-2.396963e+02	159.391774
48	-69.0241889	-2.685984e+02	130.549996
63	334.4570824	1.349013e+02	534.012852
65	176.2306906	-2.336174e+01	375.823119
68	-31.2412576	-2.308027e+02	168.320195
69	-45.1215422	-2.446761e+02	154.433034
78	69.0246421	-1.305309e+02	268.580201
82	99.6552263	-9.989334e+01	299.203794
85	176.4458539	-2.309072e+01	375.982429
87	289.1456026	8.960119e+01	488.690014

Apply similar R code for the other two models.

4

Prediction Accuracy

Prediction Error Measures

- Compare observed response Y_i to the predicted Y_i^*
- Mean squared prediction error (MSPE) $= \frac{1}{n}\sum_{i=1}^{n}(Y_i - Y_i*)^2$
- Mean absolute prediction errors (MAE) $= \frac{1}{n}\sum_{i=1}^{n}|Y_i - Y_i*|$
- Mean absolute percentage error (MAPE) $= \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i - Y_i*|}{Y_i}$
- Precision error (PM) $= \frac{\sum_{i=1}^{n}(Y_i - Y_i*)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y}_i)^2}$
- Confidence Interval error (CIM) $= \frac{1}{n}\sum_{i=1}^{n} I(Y_i* \notin CI)$

Georgia Tech

5

Prediction Error Measure Insights

Mean squared prediction error (MSPE)

- Appropriate for linear regression model prediction but depends on scale and it is sensitive to outliers

Mean absolute prediction errors (MAE)

- Not appropriate for linear regression model prediction and depends on scale but robust to outliers

Mean absolute percentage error (MAPE)

- Not appropriate for linear regression model prediction but it does not depend on scale and robust to outliers

Precision error (PM)

- Appropriate for linear regression model and does not depend on scale

Confidence Interval error (CIM)

Georgia Tech

6

Prediction Error Measure Insights

Mean squared prediction error (MSPE)

- Appropriate for linear regression model prediction but depends on scale and it is sensitive to outliers

Mean absolute prediction errors (MAE)

- Not appropriate for linear regression model prediction and depends on scale but robust to outliers

Mean absolute percentage error (MAPE)

- Not appropriate for linear regression model prediction but it does not depend on scale and robust to outliers

Precision error (PM)

- Appropriate for linear regression model and does not depend on scale

Confidence Interval error (CIM)

While MAE and MAPE are commonly used to evaluate prediction error, I recommend using the precision measure.

-- Regression models are estimated using by minimizing sum of least squares hence the accuracy error shall be best of squared differences not absolute differences

7

Prediction Accuracy: Model 1

```
## Save Predictions to compare with observed data
pred1 <- predict(model1, test, interval = 'prediction')
test.pred1 <- pred1[,1]
test.lwr1 <- pred1[,2]
test.upr1 <- pred1[,3]

# Mean Squared Prediction Error (MSPE)
mean((test.pred1-test$cnt)^2)
[1] 10304.95
# Mean Absolute Prediction Error (MAE)
mean(abs(test.pred1-test$cnt))
[1] 74.52024
# Mean Absolute Percentage Error (MAPE)
mean(abs(test.pred1-test$cnt)/test$cnt)
[1] 2.724609
# Precision Measure (PM)
sum((test.pred1-test$cnt)^2)/sum((test$cnt-mean(test$cnt))^2)
[1] 0.3101164
# CI Measure (CIM)
sum(test$cnt<test.lwr1)+sum(test$cnt>test.upr1)/nrow(test)
[1] 0.06904488
```

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 = 10304

MAE = 74.52

MAPE = 2.72

PM = 0.31

CIM = 0.069

8

Prediction Accuracy: Model 3

```
## Save Predictions to compare with observed data
pred3 <- predict(model3, test_red, interval = 'prediction')
test.pred3 <- pred3[,1]
test.lwr3 <- pred3[,2]
test.upr3 <- pred3[,3]

# Mean Squared Prediction Error (MSPE)
mean((test.pred3-test$cnt)^2)
 [1] 11271.78
# Mean Absolute Prediction Error (MAE)
mean(abs(test.pred3-test$cnt))
[1] 78.67701
# Mean Absolute Percentage Error (MAPE)
mean(abs(test.pred3-test$cnt)/test$cnt)
[1] 0.5172032
# Precision Measure (PM)
sum((test.pred3-test$cnt)^2)/sum((test$cnt-mean(test$cnt))^2)
[1] 0.316168
# CI Measure (CIM)
sum(test$cnt<test.lwr3)+sum(test$cnt>test.upr3)/nrow(test)
[1] 0.060984
```

Prediction Accuracy
MSPE = 11271
MAE = 78.67
MAPE = 0.517
PM = 0.361
CIM = 0.061

Georgia Tech

9

Model Comparison

Model	MSPE	Precision.Measure	Adjusted.R.Squared	R squared
Full MLR	10304.95	0.310	0.684	0.685
MLR (sqrt transformation)	8955.41	0.271	0.784	0.785
MLR (sqrt transformation-no low demand data)	11271.78	0.362	0.656	0.658

- **The model with the square-root transformation outperforms the other models in terms of predictive power as reflected in the Precision Measure and R squared.**
- **The non-constant variance assumption is violated across all models.**

Georgia Tech

10

Summary

Georgia Tech

11