Prediction: Multiple Linear Regression
Part 1: Estimation of the Regression Equation and Prediction

Aram Balagyozyan

Department of Operations and Information Management
Kania School of Management
The
University of Scranton

July 31, 2020

1. Introduction
2. Explanatory vs. Predictive Modeling
3. Estimation of the Regression Equation and Prediction

Some Preliminaries

- Consider the following *linear* model that describes the relationship between the numerical variable that you are trying to predict (Y) and variables that can help you to predict it (X_i):

$$Y = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \epsilon$$

- In the equation above
  - Y is the *outcome* variable (aka *response*, *target*, or *dependent* variable)
  - X_1, $X_2, \ldots$, X_k are the *predictors* (aka *independent variables*, *input* variables, *regressors*, or *covariates*)
  - β_0, $\beta_1, \ldots, \beta_p$ are the *coefficients* and
  - ϵ is the *noise* or *unexplained* part.
- Regression modeling means not only estimating the coefficients given data on Y and Xs but also choosing which predictors to include and in what form.
- Choosing the right form depends on domain knowledge, data availability, and needed predictive power.

Some Preliminaries

Multiple linear regression is applicable to numerous predictive modeling situations. Examples are

- Predicting the price and quality of wine based on weather conditions and geographical information.
- Predicting customer activity on credit cards from their demographics and historical activity patterns.
- Predicting expenditures on vacation travel based on historical frequent flyer data.
- Predicting staffing requirements at help desks based on historical data and product and sales information.
- Predicting sales from cross-selling of products from historical information
- Predicting the impact of discounts on sales in retail outlets.

Explanatory vs. Predictive Modeling

- Two popular but different objectives behind fitting a regression model are:
  1. Explaining or quantifying the average effect of inputs on an outcome (explanatory or descriptive task, respectively).
  2. Predicting the outcome value for new records, given their input values (predictive task).

Explanatory Modeling

- In the first scenario, using a sample of observations we attempt to capture the *average* impact of the regressors on the outcome variable in a larger population.
- In this case, we are interested in generating statements such as "a unit increase in service speed (X_1) is associated with an average increase of 5 points in customer satisfaction (Y), all other factors (X_2, X_3,...,X_p) being equal."
- If X_1 is known to cause Y, then such a statement indicates actionable policy changes — this is called *explanatory* modeling.
- When the causal structure is unknown, then this model quantifies the degree of association between the inputs and outcome variable, and the approach is called *descriptive* modeling.

Predictive Modeling

- When predicting new individual records is the objective, we are not interested in the coefficients themselves, nor in the "average" effect, but rather in the predictions that this model can generate for new records.
- As an example for this case, we might be interested in using the regression model to predict customer satisfaction for each new customer of interest.
- Both explanatory and predictive modeling philosophies involve using a dataset to fit a model (i.e., to estimate coefficients), checking model validity, assessing its performance, and comparing to other models. However, the modeling steps and performance assessment differ in the two cases, usually leading to different final models. Therefore, the choice of model is closely tied to whether the goal is explanatory or predictive.

Explanatory vs. Predictive Modeling

Below is the summary of the main differences between an explanatory and predictive regression modeling:

Explanatory.Models	Predictive.Models
1. The objective is to fit the data closely.	1. The objective is to predict new variables accurately.
2. The entire dataset is used for estimating the best-fit model.	2. The data are typically split into a training set and validation set.
3. Performance is measured by how closely the data fit the model and how strong the average relationship is.	3. Performance is measured by by predictive accuracy.
4. The focus is on the coefficients.	4. The focus is on the predictions.

Explanatory vs. Predictive Modeling

- For the reasons above, it is extremely important to know the goal of the analysis before beginning the modeling process.
- A good predictive model can have a looser fit to the data on which it is based, and a good explanatory model can have low prediction accuracy.
- In the remainder of this module, we focus on predictive models because these are more popular in data mining and because most statistics textbooks focus on explanatory modeling.

Estimation of the Regression Equation and Prediction

- To predict the value of the outcome variable for a record with predictor values x_1, x_2,...,x_p, we use the *estimated* regression equation:

  $$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_p x_p$$

  where
  - $\hat{Y}$ is the predicted value of the outcome variable
  - $\hat{\beta}_0$, $\hat{\beta}_1$,...,$\hat{\beta}_p$ are the estimated coefficients.
- We estimate the coefficients of the regression formula from the data on Y and x_1, x_2,...,x_p using a method called *ordinary least squares* (OLS).
- This method finds values $\hat{\beta}_0$, $\hat{\beta}_1$,...,$\hat{\beta}_p$ that minimize the sum of squared deviations between the actual outcome values Y and their predicted values based on the model ($\hat{Y}$).

Assumptions the Underlie the OLS Estimator

- Predictions based on the linear regression equation above and OLS are the best predictions possible in the sense that they will be equal to the true values on average (they will be *unbiased*) and will result in the smallest mean squared error compared to any unbiased estimates *IF* the following assumptions hold:
  1. The noise ϵ (or equivalently, Y) follows a normal distribution.
  2. The choice of predictors and their form is correct (*linearity*).
  3. The records are independent of each other.
  4. The variability in the outcome values for a given set of predictors is the same regardless of the values of the predictors (*homoskedasticity*).
- Even if the assumptions above are violated, it is still possible that the resulting predictions are sufficiently accurate and precise for the purpose they are intended for.
- The key is to evaluate predictive performance of the model, which is the main priority. Satisfying assumptions is of secondary interest and residual analysis can give clues to potential improved models to examine.

Regression Analysis in Action: Predicting the Price of Used Toyota Corolla Cars

A large Toyota car dealership offers buyers of new Toyota cars the option to trade-in their used car. The dealer then sells the used cars for a small profit. To ensure a reasonable profit, the dealer needs to be able to predict the price that the dealership will get for the used cars. For that reason, data were collected on all previous sales of used Toyota Corollas at the dealership. The data include the sales price and other information on the car, such as its age, mileage, fuel type, and engine size. A description of each of these variables is given in the table below.

Regression Analysis in Action: Predicting the Price of Used Toyota Corolla Cars

Variable	Description
Price	Offer price in Euros
Age_08_04	Age in months as of August 2004
Kilometers	Accumulated kilometers on odometer
Fuel_Type	Type Fuel type (Petrol, Diesel, CNG)
HP	Horsepower
Met_Color	Metallic color? (Yes = 1, No = 0)
Automatic	Automatic (Yes = 1, No = 0)
CC	Cylinder volume in cubic centimeters
Doors	Number of doors
Quarterly_Tax	Quarterly road tax in Euros
Weight	Weight in kilograms

Regression Analysis in Action: Predicting the Price of Used Toyota Corolla Cars

- A small sample of the dataset is shown below.

```
car.df <- read.csv("ToyotaCorolla.csv")
car.df <- car.df[1:1000, ]
selected.var <- c(3, 4, 7, 8, 9, 10, 12, 13, 14, 17, 18)
head(car.df[,selected.var],15)
```

Price	Age_08_04	KM	Fuel_Type	HP	Met_Color	Automatic	CC	Doors	Quarterly_Tax	Weight
13500	23	46986	Diesel	90	1	0	2000	3	210	1165
13750	23	72937	Diesel	90	1	0	2000	3	210	1165
13950	24	41711	Diesel	90	1	0	2000	3	210	1165
14950	26	48000	Diesel	90	0	0	2000	3	210	1165
13750	30	38500	Diesel	90	0	0	2000	3	210	1170
12950	32	61000	Diesel	90	0	0	2000	3	210	1170
16900	27	94612	Diesel	90	1	0	2000	3	210	1245
18600	30	75889	Diesel	90	1	0	2000	3	210	1245
21500	27	19700	Petrol	192	0	0	1800	3	100	1185
12950	23	71138	Diesel	69	0	0	1900	3	185	1105
20950	25	31461	Petrol	192	0	0	1800	3	100	1185
19950	22	43610	Petrol	192	0	0	1800	3	100	1185
19600	25	32189	Petrol	192	0	0	1800	3	100	1185
21500	31	23000	Petrol	192	1	0	1800	3	100	1185
22500	32	34131	Petrol	192	1	0	1800	3	100	1185

Regression Analysis in Action: Predicting the Price of Used Toyota Corolla Cars

- Just for the sake of this exercise, we've limited the dataset to 1000 cars only.
- Next, we need to partition the dataset into training (60%) and validation (40%) sets:

```
set.seed(1)#set seed for reproducing the partition
train.index <- sample(c(1:1000), 600)
train.df <- car.df[train.index, selected.var]
valid.df <- car.df[-train.index, selected.var]
```

Regression Analysis in Action: Predicting the Price of Used Toyota Corolla Cars

- Next, we estimate the multiple regression coefficients using the training set and the `lm()` function:

```
car.lm <- lm(Price ~ ., data = train.df)
```

- The `Price ~ .` expression inside the `lm()` command instructs R to form a linear model (lm) with `Price` as an outcome (dependent) variable and all other variables in the `train.df` dataset as predictors.
- A dot (`.`) after ~ instructs R to include all the remaining columns in the `train.df` dataset as predictors.

Regression Analysis in Action: Predicting the Price of Used Toyota Corolla Cars

- You could include only a select few columns as regressors by explicitly specifying them: e.g `Price ~ Age_08_04 + KM`
- Finally note that `Fuel_Type` has 3 categories: *Petrol*, *Diesel*, and *CNG*. We therefore have 2 dummy variables in the model: `Fuel_Type_Petrol` (0/1), and `Fuel_TypeDiesel` (0/1); the third, for `CNG` (0/1), is redundant given the information on the first two dummies. Technically, including the redundant dummy would cause the regression to fail, since the redundant dummy will be a perfect linear combination of the other two;
- However, R's `lm()` routine automatically handles this issue.

Regression Analysis in Action: Predicting the Price of Used Toyota Corolla Cars

- To print the results of the `lm()` command (above), I often rely on the **stargazer** package

```
# to avoid scientific notations
options(scipen = 999)
# for nicer results use stargazer
library(stargazer)
stargazer(car.lm,header=FALSE, type="text")
```

Regression Analysis in Action: Predicting the Price of Used Toyota Corolla Cars

	Dependent variable:
	Price
Age_08_04	-133.272^{***} (4.902)
KM	-0.021^{***} (0.002)
Fuel_TypeDiesel	896.206 (603.164)
Fuel_TypePetrol	$2{,}191.368^{***}$ (575.629)
HP	37.258^{***} (5.233)
Met_Color	51.315 (123.395)
Automatic	63.568 (262.282)
CC	0.011 (0.098)
Doors	−55.700 (63.966)
Quarterly_Tax	13.080^{***} (2.608)
Weight	16.220^{***} (1.527)
Constant	$-4{,}754.380^{***}$ (1,661.720)
Observations	600
R^2	0.870
Adjusted R^2	0.868
Residual Std. Error	1,392.116 (df = 588)
F Statistic	358.719^{***} (df = 11; 588)
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

- The numbers in the second column are the estimated coefficients while the numbers in parentheses are the standard errors. Stars reflect statistical significance.

Regression Analysis in Action: Predicting the Price of Used Toyota Corolla Cars

- We can now use the estimated model to make predictions about individual Toyota Corollas based on their age, mileage, and so on.

```
library(forecast) # for various accuracy measures
car.lm.pred<-predict(car.lm, valid.df)
options(scipen=999, digits = 0)
residuals <- valid.df$Price - car.lm.pred
fcast<-data.frame("Predicted" = car.lm.pred,
                  "Actual" = valid.df$Price,
                  "Residual" = residuals)
head(fcast,4)
```

```
##   Predicted Actual Residual
## 2     16447  13750    -2697
## 7     16757  16900      143
## 8     16750  18600     1850
## 9     20959  21500      541
```

Regression Analysis in Action: Predicting the Price of Used Toyota Corolla Cars

- We can also obtain the overall measures of predictive accuracy

```r
options(scipen=999, digits = 3)
accuracy(car.lm.pred, valid.df$Price)
```

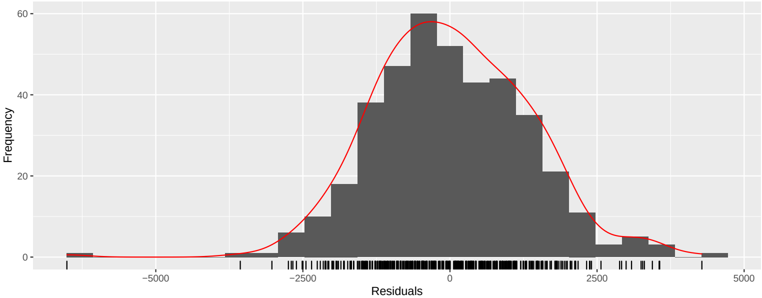
```
##              ME RMSE  MAE   MPE MAPE
## Test set 19.6 1325 1049 -0.75 9.35
```

- Note that the mean error (ME) is $19.6 and the root mean squared error (RMSE) is $1321.
- The ME being very small (compared to the value of a car), implies that by and large positive residuals (under-prediction) and negative residuals (over-prediction) are of about same magnitude and frequency. Thus they cancel each other out.
- However, the RMSE indicates that the predictor makes and average mistake (positive or negative) of $1321. With tight profit margins in the car sales industry, this may be of significant concern.

Regression Analysis in Action: Predicting the Price of Used Toyota Corolla Cars

- We can also plot the residuals

```
library(ggplot2)
residuals<-data.frame(resid=residuals)
ggplot(residuals,aes(x = resid))+
geom_histogram(aes(y=..count..), bins = 25)+
xlab("Residuals")+ylab("Frequency")+
geom_density(aes(y=500*..count..),color="red")+
geom_rug()
```



Regression Analysis in Action: Predicting the Price of Used Toyota Corolla Cars

- A histogram of the residuals shows that most of the errors are between -$2500 and +$2500. This error magnitude might be small relative to the car price, but should be taken into account when considering the profit.
- Another observation of interest is the large positive and negative residuals (under-predictions and over-prediction respectively), which may or may not be a concern, depending on the application.