Regression and its Applications

Multiple Linear Regression

Albert C. Yang, M.D., Ph.D.

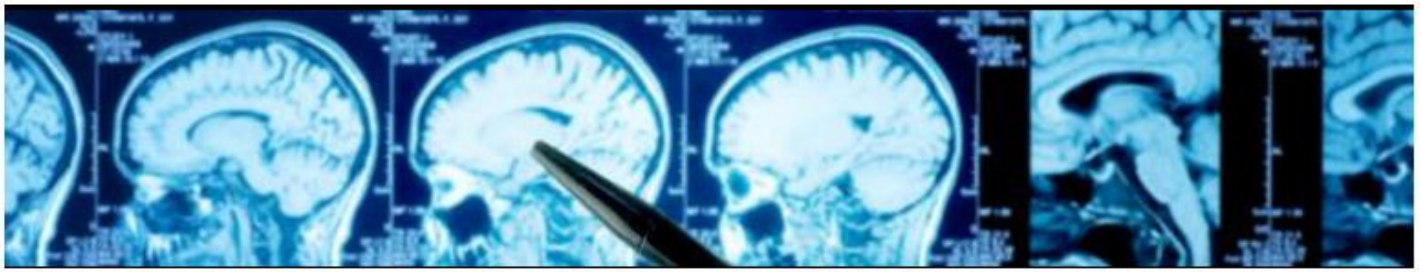
Institutes of Brain Science/Digital Medicine Center

National Yang-Ming University

Apr 9, 2020

accyang@gmail.com

Laboratory of Precision Psychiatry



Intelligent Healthcare and its Applications

1. Introduction to intelligent healthcare
   Workshop
   Chest X-Ray DICOM Files

2. Clinically Driven Artificial Intelligence - Workshop
   Chest X-Ray DICOM Files
   Case Presentation 1 Material

3. An Overview of Machine Learning Methods
   Workshop

4. Regression and its Applications
   Workshop
   Medical Insurance Data

Generalized Machine Learning Workflow

- Divide data into training and testing subset
- Model training data
- Evaluate trained model in training data
- Use trained model to predict response in testing data
- Evaluate model performance in testing data

Dataset

https://www.kaggle.com/mirichoi0218/insurance

Medical Insurance Dataset

	A	B	C	D	E	F	G
1	age	sex	bmi	children	smoker	region	charges
2	19	female	27.9	0	yes	southwest	16884.92
3	18	male	33.77	1	no	southeast	1725.552
4	28	male	33	3	no	southeast	4449.462
5	33	male	22.705	0	no	northwest	21984.47
6	32	male	28.88	0	no	northwest	3866.855
7	31	female	25.74	0	no	southeast	3756.622
8	46	female	33.44	1	no	southeast	8240.59
9	37	female	27.74	3	no	northwest	7281.506
10	37	male	29.83	2	no	northeast	6406.411
11	60	female	25.84	0	no	northwest	28923.14
12	25	male	26.22	0	no	northeast	2721.321
13	62	female	26.29	0	yes	southeast	27808.73
14	23	male	34.4	0	no	southwest	1826.843
15	56	female	39.82	0	no	southeast	11090.72
16	27	male	42.13	0	yes	southeast	39611.76
17	19	male	24.6	1	no	southwest	1837.237
18	52	female	30.78	1	no	northeast	10797.34
19	23	male	23.845	0	no	northeast	2395.172
20	56	male	40.3	0	no	southwest	10602.39
21	30	male	35.3	0	yes	southwest	36837.47
22	60	female	36.005	0	no	northeast	13228.85

Code Review

record_computation_workshop04.m

This file can be opened as a Live Script. For more information, see Creating Live Scripts.

```
%% Read Data into Matlab
[num,txt,raw] = xlsread('insurance.csv');
age=num(:,1);
insurance=num(:,7);

%% Plot relationship between Age and Insurance Claims
plot(age,insurance,'.');
xlabel('age');
ylabel('insurance');

%% Fitting simple linear regression model
model1 = fitlm(age,insurance)

%% Visualization of regression results
ypred = predict(model1,age);

plot(age,insurance,'.');
hold on
plot(age,ypred,'ro');
xlabel('age');
ylabel('insurance');

%% Model evaluation
model1.RMSE;

%% Divide data into training and testing subset
test_index = zeros(length(insurance),1);
test_sample = randsample(length(insurance),fix(length(insurance)*0.3));
test_index(test_sample) = 1;
train_index = ~test_index;

data = [age insurance];
train_data = data(train_index==1,:);
test_data = data(test_index==1,:);

%% Fitting training data
model2 = fitlm(train_data(:,1),train_data(:,2))

%% Predict Response in Testing Data
ypred = predict(model2,test_data(:,1));

RMSE_test = sqrt(mean((ypred-test_data(:,2)).^2))
RMSE_train = model2.RMSE
```

Include More Predictors in the Regression Model

- **Continuous variables**

bmi = num(:,3);

Include More Predictors in the Regression Model

- **Ordinal variables**

children = num(:,4);

Include More Predictors in the Regression Model

- **Binary variables**

```
smoker_str = txt(2:end,5);
smoker = cellfun(@(x)(strcmp(x,'yes')), smoker_str);
sex_str = txt(2:end,2);
sex = cellfun(@(x)(strcmp(x,'male')), sex_str);
```

Include More Predictors in the Regression Model

- **Categorical (nominal) variables**

region	code	region1	region2	region3	region4
southwest	0	1	0	0	0
southeast	1	0	1	0	0
southeast	1	0	1	0	0
northwest	2	0	0	1	0
northwest	2	0	0	0	0
southeast	1	0	1	0	0
southeast	1	0	1	0	0
northwest	2	0	0	0	0
northeast	3	0	0	0	1
northwest	2	0	0	1	0
northeast	3	0	0	0	1
southeast	1	0	1	0	0
southwest	0	1	0	0	0
southeast	1	0	1	0	0
southeast	1	0	1	0	0

There is no intrinsic order in categorical variables

Include More Predictors in the Regression Model

- **Categorical (nominal) variables**

```
region_str = txt(2:end,6);
sw = cellfun(@(x)(strcmp(x,'southwest')), region_str);
se = cellfun(@(x)(strcmp(x,'southeast')), region_str);
nw = cellfun(@(x)(strcmp(x,'northwest')), region_str);
ne = cellfun(@(x)(strcmp(x,'northeast')), region_str);
```

Combine All Predictors and Response into a Data Matrix

- data = [age sex bmi children smoker sw se nw ne insurance];

Apply Generalized Machine Learning Workflow to New Data

- data = [age sex bmi children smoker sw se nw ne insurance];

Apply Generalized Machin Learning Workflow to New Data

- data = [age sex bmi children smoker sw se nw ne insurance];

- **Divide data into training and testing subset**

```
test_index = zeros(length(insurance),1);
test_sample = randsample(length(insurance),fix(length(insurance)*0.3));
test_index(test_sample) = 1;
train_index = ~test_index;
train_data = data(train_index==1,:);
test_data = data(test_index==1,:);
```

- **Fit multiple linear regression to training data**

```
model3 = fitlm(train_data(:,1:8),train_data(:,10))
```

- **Predict Response in Testing Data**

```
ypred = predict(model3,test_data(:,1:8));
```

- **Evaluate the model**

```
RMSE_test = sqrt(mean((ypred-test_data(:,10)).^2))
RMSE_train = model3.RMSE
```

```
RMSE_test =

   6.0028e+03

RMSE_train =

   6.1219e+03
```

Multiple Linear Regression Results

Command Window

New to MATLAB? See resources for Getting Started.

```
>> model3 = fitlm(train_data(:,1:8),train_data(:,10))

model3 =


Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8

Estimated Coefficients:
```

	Estimate	SE	tStat	pValue
(Intercept)	-11011	1190.5	-9.2488	1.5199e-19
x1	254.86	14.297	17.826	2.358e-61
x2	101.11	401.31	0.25194	0.80114
x3	305.85	34.407	8.8894	3.1445e-18
x4	560.77	164.3	3.4131	0.00067007
x5	24672	499.26	49.417	4.0344e-262
x6	-910.51	573.79	-1.5868	0.11289
x7	-1005.4	572.6	-1.7558	0.079447
x8	-43.97	572.36	-0.076821	0.93878

```
Number of observations: 937, Error degrees of freedom: 928
Root Mean Squared Error: 6.12e+03
R-squared: 0.756,  Adjusted R-Squared: 0.754
F-statistic vs. constant model: 360, p-value = 1.96e-278
```

Apply Generalized Machin Learning Workflow to New Data (Improved)

- data = table(age,sex,bmi,children,smoker,region_str,insurance,
'VariableNames',{'age','sex','bmi','children','smoker','region','insurance'});

- **Divide data into training and testing subset**

```
test_index = zeros(length(insurance),1);
test_sample = randsample(length(insurance),fix(length(insurance)*0.3));
test_index(test_sample) = 1;
train_index = ~test_index;
train_data = data(train_index==1,:);
test_data = data(test_index==1,:);
```

- **Fit multiple linear regression to training data**

```
model4 = fitlm(train_data,'ResponseVar','insurance')
```

- **Predict Response in Testing Data**

```
ypred = predict(model4,test_data);
```

- **Evaluate the model**

```
RMSE_test = sqrt(mean((ypred-test_data.insurance).^2))
RMSE_train = model4.RMSE
```

```
RMSE_test =

   5.8911e+03
```

```
RMSE_train =

   6.1424e+03
```

Multiple Linear Regression Results (Improved Interpretability)

Command Window

New to MATLAB? See resources for Getting Started.

```
Linear regression model:
    insurance ~ 1 + age + sex + bmi + children + smoker + region

Estimated Coefficients:
```

	Estimate	SE	tStat	pValue
(Intercept)	-13548	1226.6	-11.046	9.7922e-27
age	257.89	14.347	17.976	3.2052e-62
sex_1	-191.6	403.37	-0.47498	0.63491
bmi	362.13	34.459	10.509	1.7469e-24
children	425.12	167.35	2.5402	0.01124
smoker_1	23391	503.51	46.456	2.2141e-244
region_southeast	-56.098	577.61	-0.097122	0.92265
region_northwest	735.37	577.89	1.2725	0.20351
region_northeast	1251.3	579.77	2.1583	0.031163

```
Number of observations: 937, Error degrees of freedom: 928
Root Mean Squared Error: 6.14e+03
R-squared: 0.743,  Adjusted R-Squared: 0.741
F-statistic vs. constant model: 335, p-value = 1.35e-267
```