

```
%{
-----
      Tenure-track professor salary
      Multiple linear regression exercise
      Sam Qian
-----

I downloaded a dataset containing profiles of 52 tenure-track professors from a small college in 1985.
The data includes:
1) sex
2) academic rank
3) years in current rank
4) highest degree
5) number of years since highest degree was achieved
6) salary (in year 1985 dollars)

I created a multiple linear regression model to predict salary from the
other five variables.

Dataset from http://data.princeton.edu/wws509/datasets/#salary
%}
clear all;
clc;

%{
Import data

Note categorical variables in the raw file were already re-coded as follows:
sex: male=0; female=1
rank: assistant_prof=1; associate_prof=2; full_prof=3;
highest degree: masters=0; doctorate=1
%}
raw = importdata('salary.raw');
disp('Sample data from 5 individuals');
disp('sex, rank, years in rank, highest degree, years since degree, salary');
disp(raw(1:5,:))
```

```
Sample data from 5 individuals
(sex, rank, years in rank, highest degree, years since degree, salary)
      0      3      25      1      35      36350
      0      3      13      1      22      35350
      0      3      10      1      23      28200
      1      3      7      1      27      26775
      0      3      19      0      30      33696
```

```
% Calculating weight(w) for optimal model fit
X = raw(:,1:5);
X(:,6) = 1; % constant term
X_t = transpose(X);
y = raw(:,6);

w = inv(X_t*X)*(X_t*y);

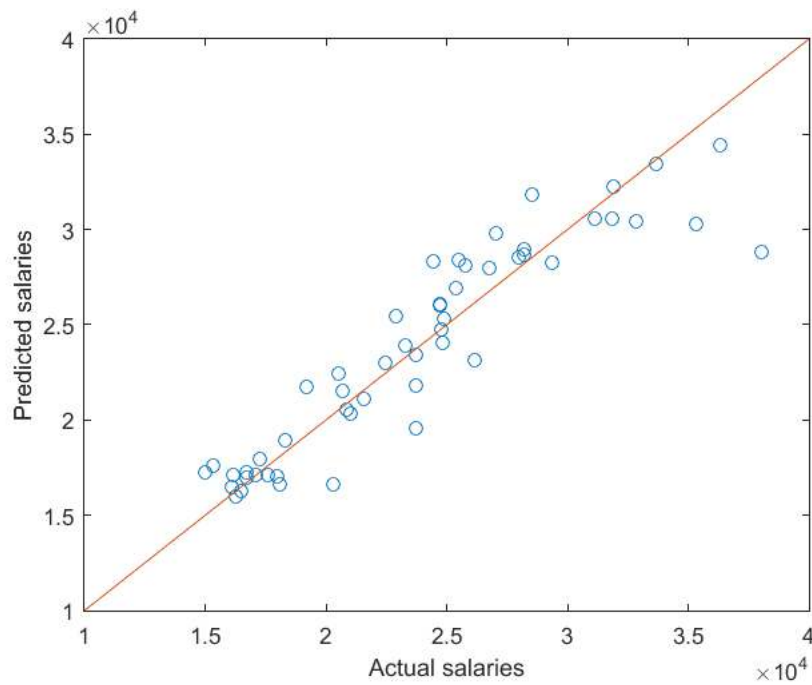
% Calculate predicted salary from weights
for i=1:52
    y_pred(i) = w(1)*X(i,1)+w(2)*X(i,2)+w(3)*X(i,3)+w(4)*X(i,4)+w(5)*X(i,5)+w(6)*X(i,6);
end

y_pred=round(y_pred');

% Visualizing actual vs. predicted salaries
plot(y, y_pred,'o')
xlim([10000 40000])
ylim([10000 40000])

xlabel('Actual salaries');
ylabel('Predicted salaries');

hold on;
xyline=[0,40000];
plot(xyline,xyline);
```



```
% Max Error of model
e = y-y_pred;
MaxE = max(e);
disp(['Max error of model is ', num2str(MaxE)]);
```

Max error of model is 9194

```
% Using the model to predict my PI's salary
input=[0, 2, 4, 1, 15, 1];

PI_salary = 0;
for i = 1:6;
    PI_salary = PI_salary + w(i)*input(i);
end

% Adjusting for inflation since 1985 and correcting for faculty salary difference between Columbia and small colleges (chose Albright College at random) (various Go
PI_salary_adjusted = round(PI_salary * 2.17 * 2.64);

disp(['My regression model predicts my PI's salary is $', num2str(PI_salary_adjusted)]);
```

My regression model predicts my PI's salary is \$121740