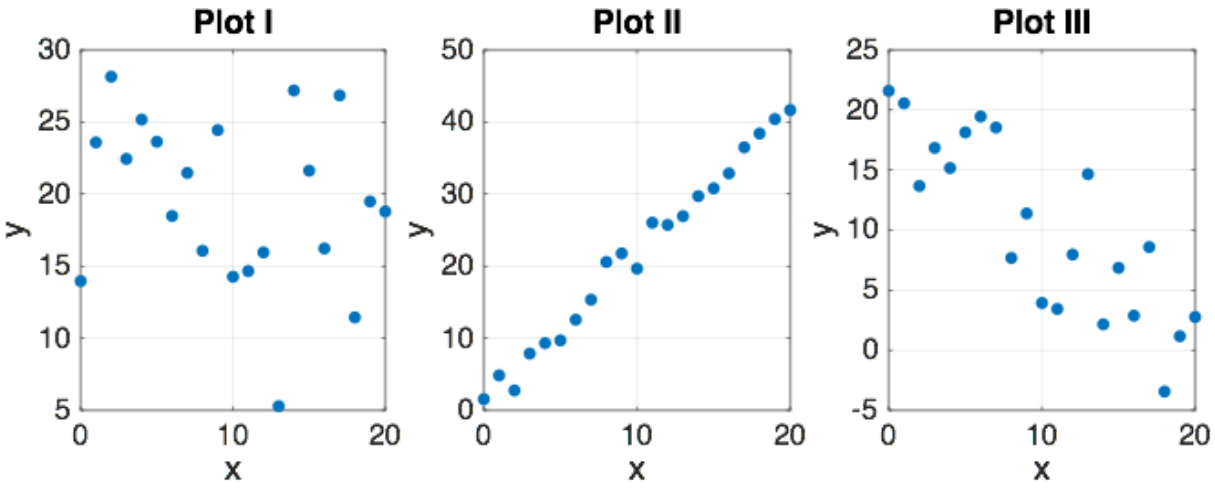
Cogs 109: Modeling and Data Analysis

Homework 2

Due Friday 10/13 in class



1. For each of the three data sets plotted above (I, II and III), answer the following:

   a. (3 points) Does the data show a positive or negative correlation between x and y?

   No correlation, positive, and negative, respectively.

   b. (3 points) Which function (equation) best describes each data set?

      i. $f(x) = 1 + 2x + \epsilon$

      Plot II

      ii. $f(x) = 20 + \epsilon$

      Plot I

      iii. $f(x) = 20 - x + \epsilon$

      Plot III

   c. (3 points) Which regression table corresponds to each plot?

i.

	Estimate	SE	tStat	pValue
(Intercept)	1.2478	0.61327	2.0347	0.056077
x1	2.0417	0.052459	38.92	1.3891e-19

Plot II

ii.

	Estimate	SE	tStat	pValue
(Intercept)	20.273	1.7491	11.591	4.6458e-10
x1	-1.0082	0.14962	-6.7383	1.9406e-06

Plot III

iii.

	Estimate	SE	tStat	pValue

(Intercept)	21.808	2.4438	8.9236	3.1883e-08
x1	-0.2323	0.20905	-1.1112	0.28033

Plot I

2. (A: 3 points, B: 1 point, C: 2 points) ISLR chapter 3, problem 3 (page 120)

   A) For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.

   B) 137.1

   C) False. We use the p-value to determine whether a regression coefficient is significant or not.

3. (2 points for each part) ISLR chapter 3, problem 4 (pages 120-121)

   A) The cubic model would have a lower RSS because more flexible models always have a lower RSS in the training data.

   B) If the true model is linear, then the cubic model would have a higher RSS in our testing data because of overfitting.

   C) Again, the cubic model will have a lower RSS because more flexible models always have a lower RSS in the training data.

   D) Because we do not know the true model for the data, we cannot say.

4. **UPDATED:** In this problem, we will simulate a dataset and use multiple linear regression to investigate it. Imagine we conduct a survey of N=100 students and ask them how much time per week they spend on work (x_1) and how much time on play (x_2). We also ask them about their overall level of satisfaction (y), which we take to be the outcome. Download the dataset HW2.csv from the course website, which contains these data.

   a. (3 points) Make a scatter plot showing y vs. x_1. Comment on the relationship between these variables: do they appear correlated (positively or negatively)? Is their relationship linear or non-linear?

   b. (4 points) Fit a simple linear regression of y vs. x_1. In MATLAB, you could use the function `regress` or `fitlm`. Report the estimated intercept and slope, and make a plot showing the data points together with the regression line. Is there a statistically significant effect of x_1 on y? NOTE: The Matlab function `regress` sdf

   c. (1 point) What is the 95% confidence interval for the slope of x_1?

   d. (2 points) Now fit a multiple linear regression with x_1 and x_2 as independent variables. Report a table with the regression results (similar to Table 3.9 on page 88 in ISLR). Which parameters have a statistically significant effect?

   e. (2 points) Make a scatter plot showing y vs. $\hat{y}$, the predicted value of y.

   f. (3 points) Create a categorical variable with 3 levels called WorkType, where WorkType="Idle" for $x_1 < 10$, WorkType="Diligent" for $10 \leq x_1 < 30$, and WorkType="Workaholic" for $x_1 \geq 30$. Fit a linear regression of y against WorkType and x_2 , and report the regression table.

g. (2 point) In part (f) you should have obtained two different coefficients for WorkType corresponding to different “levels” of this categorical variable. What is your interpretation of the term corresponding to WorkType=Workaholic?

HW2Solutions

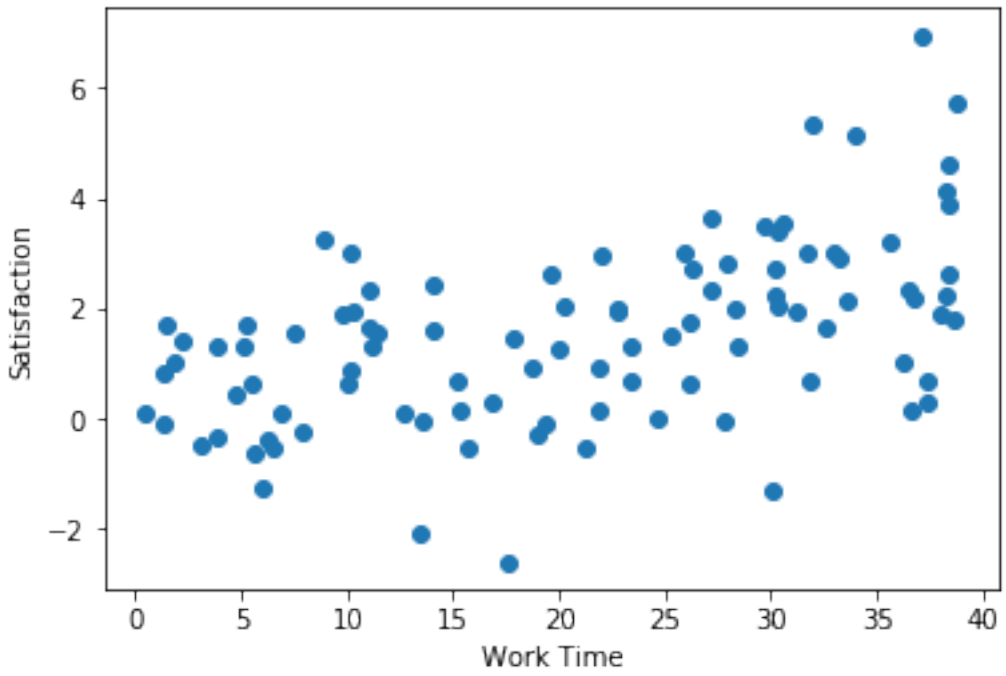
October 10, 2018

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import statsmodels.api as sm
from patsy import dmatrices
```

In [8]:
```
df = pd.read_csv('HW2.csv', header=0, names=["x1", "x2", "y"])

print(df.head())
```

```
        x1       x2       y
0  36.2320  31.7710  1.0401
1   5.0795  12.4490  1.3170
2  36.5350  21.1410  2.3423
3  25.2940   6.6259  1.5134
4   3.9016  24.0790  1.3138
```

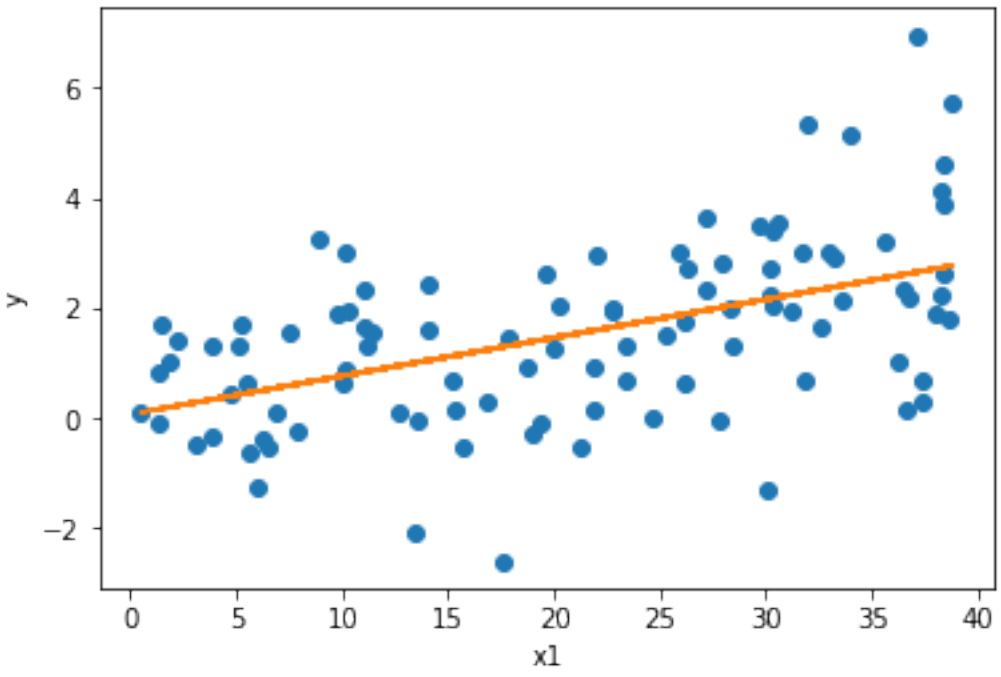
In [17]:
```
# 4 (a)
fig, ax = plt.subplots()
ax.scatter(df.x1, df.y)
ax.set_xlabel('Work Time')
ax.set_ylabel('Satisfaction')
plt.show()
```



In [10]:
```
# 4 (b) (c)
### do linear regression
# setup input data
y, X = dmatrices('y ~ x1', data=df, return_type='dataframe')
# print(y.head())
# print(X.head())
# describe model
mod = sm.OLS(y, X)
# fit model
res = mod.fit()
# look at results
print(res.summary())
yhat = np.dot(X.values, res.params.values)
fig, ax = plt.subplots()
ax.scatter(df.x1, df.y)
ax.plot(df.x1, yhat, color='C1')
ax.set_xlabel('x1')
ax.set_ylabel('y')
plt.show()
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.256
Model:                            OLS   Adj. R-squared:                  0.248
Method:                 Least Squares   F-statistic:                     33.34
```

```
Date:                Wed, 10 Oct 2018   Prob (F-statistic):           9.32e-08
Time:                        20:19:41   Log-Likelihood:                -174.00
No. Observations:                  99   AIC:                             352.0
Df Residuals:                      97   BIC:                             357.2
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.0737      0.290      0.254      0.800      -0.503       0.650
x1             0.0696      0.012      5.774      0.000       0.046       0.093
==============================================================================
Omnibus:                        1.686   Durbin-Watson:                   2.230
Prob(Omnibus):                  0.430   Jarque-Bera (JB):                1.202
Skew:                           0.004   Prob(JB):                        0.548
Kurtosis:                       3.540   Cond. No.                         49.2
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

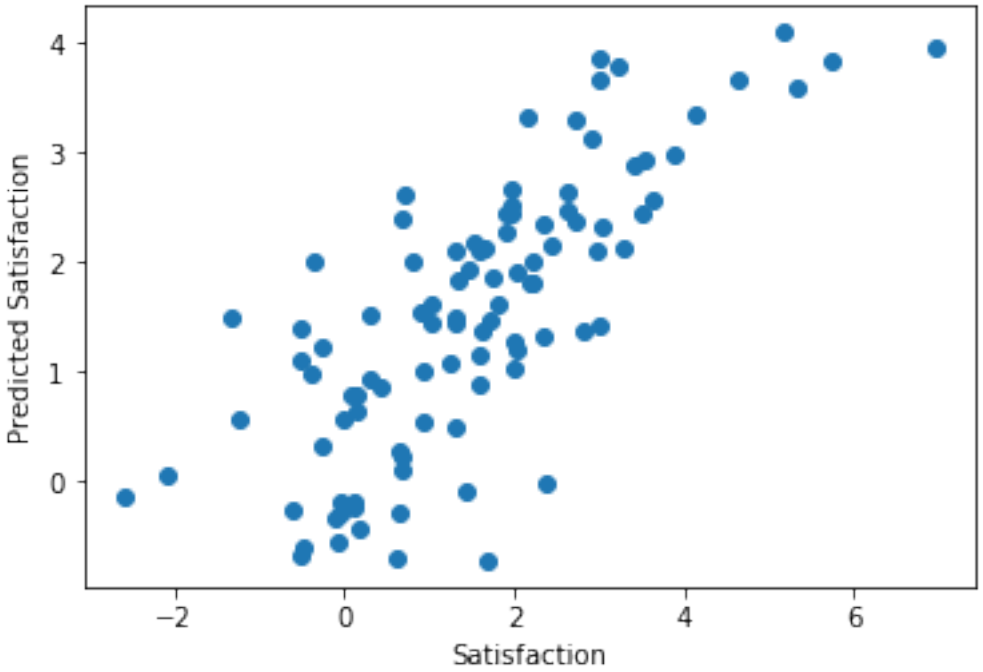



```
In [11]: # 4 (d)
         ### do multi-linear regression
         3
```

```
# setup input data
y, X = dmatrices('y ~ x1 + x2', data=df, return_type='dataframe')
# print(y.head())
# print(X.head())
# describe model
mod = sm.OLS(y, X)
# fit model
res = mod.fit()
# look at results
print(res.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.568
Model:                            OLS   Adj. R-squared:                  0.559
Method:                 Least Squares   F-statistic:                     63.01
Date:                Wed, 10 Oct 2018   Prob (F-statistic):           3.32e-18
Time:                        20:19:54   Log-Likelihood:                -147.12
No. Observations:                  99   AIC:                             300.2
Df Residuals:                      96   BIC:                             308.0
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      1.8557      0.309      6.009      0.000       1.243       2.469
x1             0.0564      0.009      6.021      0.000       0.038       0.075
x2            -0.0800      0.010     -8.321      0.000      -0.099      -0.061
==============================================================================
Omnibus:                        1.820   Durbin-Watson:                   2.211
Prob(Omnibus):                  0.403   Jarque-Bera (JB):                1.312
Skew:                           0.093   Prob(JB):                        0.519
Kurtosis:                       3.532   Cond. No.                         85.4
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

In [16]:
```
# 4 (e)
yhat = res.predict(X.values)
4

fig, ax = plt.subplots()
ax.plot(y, yhat, 'o')
ax.set_xlabel('Satisfaction')
ax.set_ylabel('Predicted Satisfaction')
plt.show()
```



In [14]:
```
# 4(f)
WorkType = []
for item in df.x1:
    if item < 10:
        WorkType.append('Idle')
    elif 10 <= item < 30:
        WorkType.append('Diligent')
    elif item >=30:
        WorkType.append('Workaholic')

print(WorkType)
df['WorkType'] = WorkType
print(df.head())
```

```
['Workaholic', 'Idle', 'Workaholic', 'Diligent', 'Idle', 'Diligent', 'Diligent', 'Workaholic',
        x1       x2       y    WorkType
0  36.2320  31.7710  1.0401  Workaholic
1   5.0795  12.4490  1.3170        Idle
2  36.5350  21.1410  2.3423  Workaholic
3  25.2940   6.6259  1.5134    Diligent
4   3.9016  24.0790  1.3138        Idle
```

In [15]:
```
# 4(f)
```

```
### do linear regression with categorical variables
# setup input data
y, X = dmatrices('y ~ WorkType + x2', data=df, return_type='dataframe')
print(y.head())
print(X.head())
# describe model
mod = sm.OLS(y, X)
# fit model
res = mod.fit()
# look at results
print(res.summary())
```

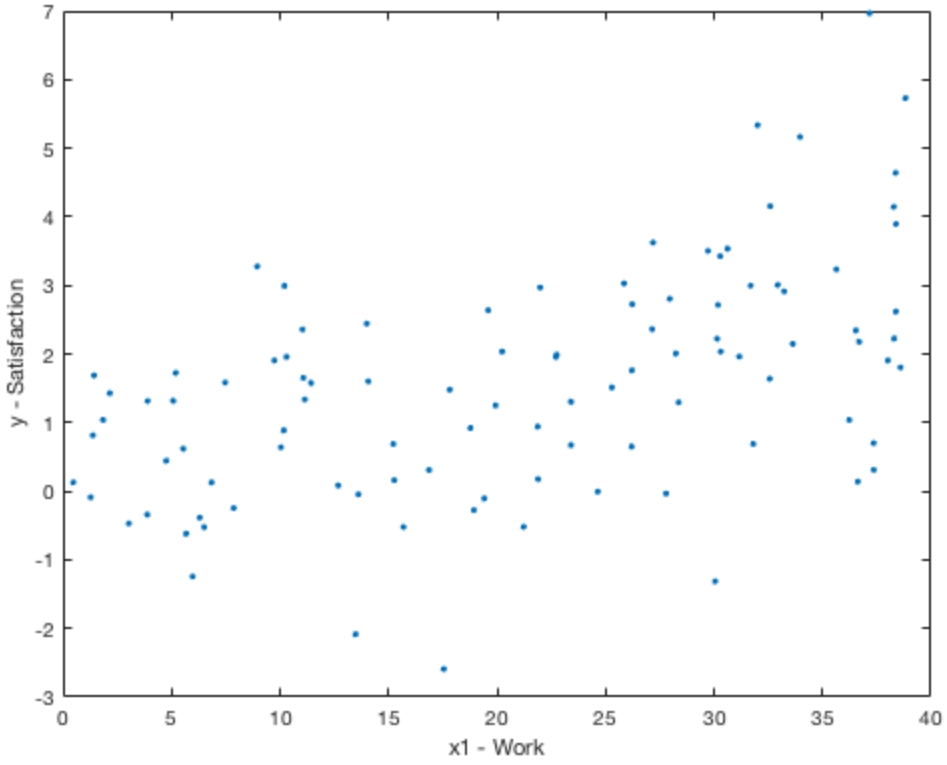
```
        y
0  1.0401
1  1.3170
2  2.3423
3  1.5134
4  1.3138
   Intercept  WorkType[T.Idle]  WorkType[T.Workaholic]       x2
0        1.0               0.0                     1.0  31.7710
1        1.0               1.0                     0.0  12.4490
2        1.0               0.0                     1.0  21.1410
3        1.0               0.0                     0.0   6.6259
4        1.0               1.0                     0.0  24.0790
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.579
Model:                            OLS   Adj. R-squared:                  0.565
Method:                 Least Squares   F-statistic:                     43.51
Date:                Wed, 10 Oct 2018   Prob (F-statistic):           8.77e-18
Time:                        20:21:17   Log-Likelihood:                -145.83
No. Observations:                  99   AIC:                             299.7
Df Residuals:                      95   BIC:                             310.0
Df Model:                           3
Covariance Type:            nonrobust
================================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                     2.7269      0.234     11.651      0.000       2.262       3.192
WorkType[T.Idle]             -0.2284      0.283     -0.808      0.421      -0.789       0.333
WorkType[T.Workaholic]        1.3862      0.250      5.534      0.000       0.889       1.883
x2                           -0.0837      0.010     -8.769      0.000      -0.103      -0.065
==============================================================================
Omnibus:                        1.532   Durbin-Watson:                   2.025
Prob(Omnibus):                  0.465   Jarque-Bera (JB):                0.997
Skew:                          -0.106   Prob(JB):                        0.607
Kurtosis:                       3.443   Cond. No.                         68.7
==============================================================================
```

```
Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Table of Contents

Problem 4a

```
clear
clf
data = readtable('HW2.csv');
data.Properties.VariableNames = {'x1','x2','y'};

figure(1)
plot(data.x1,data.y,'.')
xlabel('x1 - Work')
ylabel('y - Satisfaction')
```



4b

```
p = fitlm(data,'y~x1')
yhat_x1 = predict(p,data);
```

```
p =


Linear regression model:
    y ~ 1 + x1

Estimated Coefficients:
                   Estimate       SE        tStat       pValue
                   ________    _________    _______    __________

    (Intercept)    0.060392     0.29112    0.20745       0.83609
    x1             0.071053    0.012028     5.9072    5.0376e-08


Number of observations: 100, Error degrees of freedom: 98
Root Mean Squared Error: 1.42
R-squared: 0.263,  Adjusted R-Squared 0.255
F-statistic vs. constant model: 34.9, p-value = 5.04e-08
```

4c

```
confidence = coefCI(p)
```

```
confidence =

   -0.5173    0.6381
    0.0472    0.0949
```

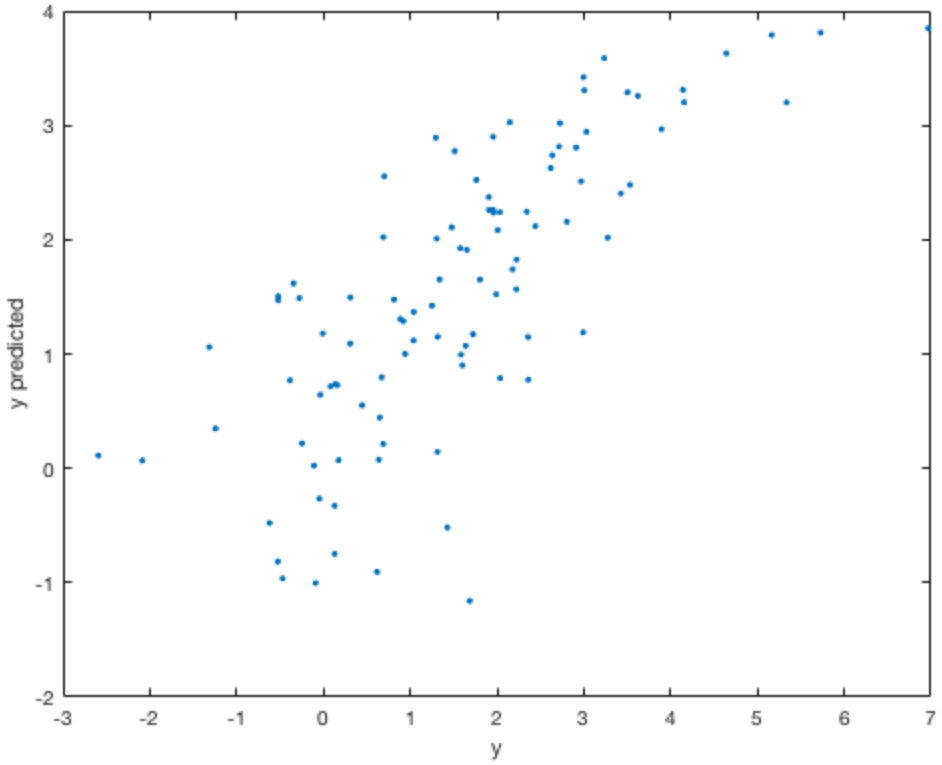
4d

```
p = fitlm(data,'y~x1+x2')
yhat_x1_x2 = predict(p,data);
```

```
p =


Linear regression model:
    y ~ 1 + x1 + x2

Estimated Coefficients:
                   Estimate        SE         tStat       pValue
                   ________    _________    _______    __________

    (Intercept)       1.8659      0.30828     6.0528     2.683e-08
    x1              0.057066    0.0093256     6.1193     1.987e-08
    x2             -0.080764    0.0095621    -8.4462    2.9961e-13


Number of observations: 100, Error degrees of freedom: 97
Root Mean Squared Error: 1.08
R-squared: 0.575,  Adjusted R-Squared 0.566
F-statistic vs. constant model: 65.6, p-value = 9.39e-19
```

4e

```
figure(2); clf
plot(data.y, yhat_x1_x2, '.')
hold on
%plot([0,6],[0,6],'k-')
xlabel('y')
ylabel('y predicted')
```



4f

```
N = size(data,1);
data.WorkType = repmat({'Idle'},N,1);
data.WorkType(data.x1>=10 & data.x1<30) = {'Diligent'};
data.WorkType(data.x1>=30) = {'Workaholic'};
p = fitlm(data,'y~1+WorkType+x2')
```

```
p =


Linear regression model:
    y ~ 1 + x2 + WorkType

Estimated Coefficients:
                        Estimate        SE          tStat       pValue
                        ________    _________    _______
 __________

    (Intercept)           4.1396      0.24919     16.612
 5.4727e-30
    x2                  -0.084184    0.0094664   -8.8929
 3.5502e-14
```

```
    WorkType_Idle          -1.6302      0.30179    -5.4017
 4.7861e-07
    WorkType_Diligent       -1.404      0.24728    -5.6777
 1.4511e-07


Number of observations: 100, Error degrees of freedom: 96
Root Mean Squared Error: 1.07
R-squared: 0.588,  Adjusted R-Squared 0.575
F-statistic vs. constant model: 45.7, p-value = 1.94e-18
```

Published with MATLAB® R2018b