BIM 105
Probability and Statistics for Biomedical Engineers

David M. Rocke

Department of Biomedical Engineering

1. Ozone (O_3) is a major component of air pollution in many cities. Atmospheric ozone levels are influenced by many factors, including weather. In one study, the mean percent relative humidity (x) and the mean ozone levels (y) were measured for 120 days in a western city. Mean ozone levels were measured in ppb. The following output (from MATLAB) describes the fit of a linear model to these data. Assume that the required assumptions are satisfied sufficiently well that we can do the usual analysis.

```
>> fitlm(Humidity,Ozone)
Linear regression model:
    y ~ 1 + x1
Estimated Coefficients:
                   Estimate        SE            tStat     pValue
    (Intercept)    29.7240        2.051         14.490     0.0000
    x1             -0.1268      0.03825         -3.315     0.0006

Number of observations: 120, Error degrees of freedom: 118
Root Mean Squared Error: 6.26
R-squared: 0.960,  Adjusted R-Squared 0.890
```

a. What is the slope of the least-squares line?

b. Find a 95% confidence interval for the slope.

c. Perform a test of the null hypothesis that the slope is greater than or equal to –0.1. What is the P-value?

a) The slope of the regression line is the predictor coefficient –0.1268

b) The standard error of the slope is 0.03825. We have $n = 120$, so we can use a z interval based on $z_{0.025} = 1.960$ (or we could use $t_{118,\ 0.025} = 1.98$) so
$-0.1268 \pm (1.960)(0.03825)$
-0.1268 ± 0.07497
$(-0.2018, -0.0518)$

c) The test statistic is $(-0.1268 - (-0.1)/0.03825 = -0.701$, so $p = 2(0.2416) = 0.4832$, don't reject the null

2. The article "Computation of Equilibrium Oxidation and Reduction Potentials for Reversible and Dissociative Electron-Transfer Reactions in Solution" (P. Winget, C. Cramer, and D. Truhlar, Theoretical Chemistry Accounts, 2004:217–227) presents several models for estimating aqueous one-electron potentials. The data set `HW7-2.csv` presents true potentials, measured experimentally in volts relative to the normal hydrogen electrode, for phenol and 23 substituted phenols, along with the corresponding value from the Austin model. Although the model values are not close to the true values, it is thought that they follow a linear model $y = \beta_0 + \beta_1 x + \varepsilon$, where y is the true value and x is the model value.

a. Compute the least-squares line for predicting the true potential from the model value.

b. Compute 95% confidence intervals for β_0 and β_1.

c. Two molecules differ in their model value by 0.5. By how much do you estimate that their true potentials will differ?

d. A molecule has a model value of 4.3. Find a 95% confidence interval for its true potential.

e. Can you conclude that the mean potential of molecules whose model value is 4.5 is greater than 0.93? Explain.

a. Compute the least-squares line for predicting the true potential from the model value.

```
>> HW72lm = fitlm(HW72)

Linear regression model:
    TRUE ~ 1 + Model

Estimated Coefficients:
                   Estimate       SE        tStat       pValue
                   ________    ________    _______    __________

    (Intercept)    -0.37909     0.11329    -3.3463     0.0029224
    Model           0.30201    0.026479     11.406    1.0478e-10

Number of observations: 24, Error degrees of freedom: 22
Root Mean Squared Error: 0.0514
R-squared: 0.855,  Adjusted R-Squared 0.849
F-statistic vs. constant model: 130, p-value = 1.05e-10
```

y-hat = -0.37909 + 0.30201x

b. Compute 95% confidence intervals for β_0 and β_1.

$t_{22,0.025} = 2.0739$

-0.3791 ± (2.074)(0.1133) or -0.3791 ± 0.2350 or (−0.614, −0.144)

 0.3020 ± (2.074)(0.0265) or 0.3020 ± 0.0549 or (0.247, 0.357)

```
Estimated Coefficients:
                   Estimate    SE          tStat      pValue
    (Intercept)    -0.37909     0.11329    -3.3463     0.0029224
    Model           0.30201    0.026479     11.406    1.0478e-10
```

c. Two molecules differ in their model value by 0.5. By how much do you estimate that their true potentials will differ?

```
(0.5)(0.30201) = 0.151
```

d. A molecule has a model value of 4.3. Find a 95% confidence interval for its true potential.

```
>> [pred ci] = predict(HW72lm,4.3)
pred =
    0.9196
ci =
    0.8977    0.9415
```

```
The CI is (0.8977. 0.9415), a CI for the mean true value when the model value is 4.3.
or

The prediction interval was not requested, but it can be found as follows:

>> [pred ci] = predict(HW72lm,4.3,'Prediction','observation')
ci =

    0.8107    1.0284

The prediction interval for an observation with model value 4.3 is (0.8107, 1.0284)
```

e. Can you conclude that the mean potential of molecules whose model value is 4.5 is greater than 0.93? Explain.

```
>> [pred ci] = predict(HW72lm,4.5)

pred =

    0.9800


ci =

    0.9545    1.0054
```

Since the CI does not include 0.93,we reject the null hypothesis.

Alternatively, the CI is the prediction ± $t_{22,0.025}$se(prediction) so 0.02545 = $t_{22,0.025}$se(prediction) and since the t-value is 2.0739, we have se(prediction) = 0.02545/2.0739 = 0.01227
The test statistic is (0.9800 – 0.9300)/0.01227 = 4.074 and p = 0.00025. This is of course a two-sided test.

e. Can you conclude that the mean potential of molecules whose model value is 4.5 is greater than 0.93? Explain.

Computing by hand, we need

$n = 24$

$s^2 = 0.0514^2$

SSX = var(HW72.Model)*23

= 0.1640*23 = 3.7709

$(x - xbar)^2$ = (4.5 - mean(HW82.Model))^2

= (4.5 – 4.2599)^2 = 0.0577

SE(prediction)

= 0.0514*sqrt(1/24 + 0.0577/3.7709)

= 0.01227

3. Two radon detectors were placed in different locations in the basement of a home. Each provided an hourly measurement of the radon concentration, in units of pCi/L. The data are presented in the file `HW7-3.csv`.

a. Compute the least-squares line for predicting the radon concentration at location 2 from the concentration at location 1.

b. Plot the residuals versus the fitted values. Does the linear model seem appropriate?

c. Divide the data into two groups: points where $R_1 < 4$ in one group, points where $R_1 \geq 4$ in the other. Compute the least-squares line and the residual plot for each group. Does the line describe either group well? Which one?

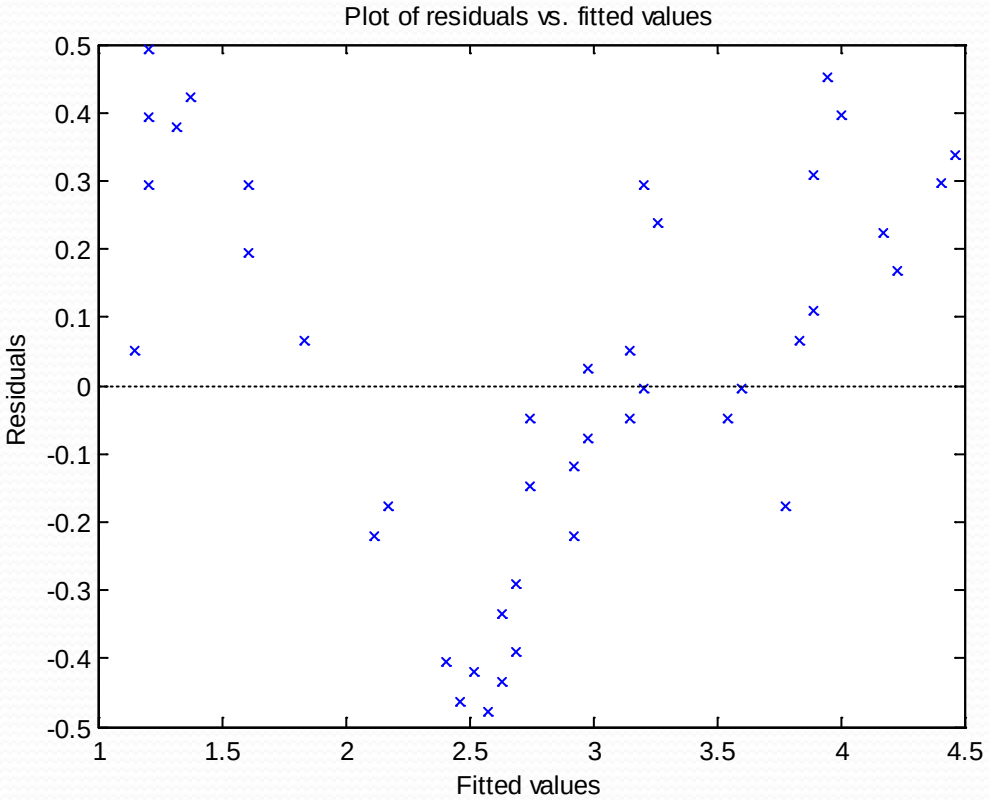
d. Explain why it might be a good idea to fit a linear model to part of these data, and a nonlinear model to the other.

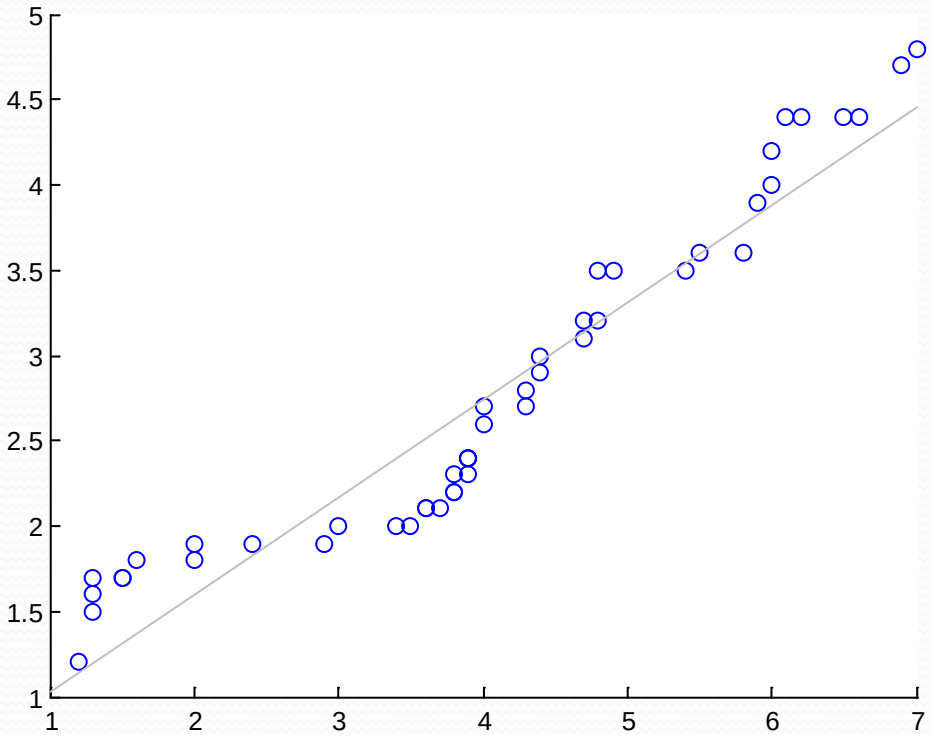
a. Compute the least-squares line for predicting the radon concentration at location 2 from the concentration at location 1.

```
>> HW73lm = fitlm(HW73)

Linear regression model:
    R2 ~ 1 + R1

Estimated Coefficients:
                   Estimate    SE          tStat     pValue
    (Intercept)    0.46358      0.11724    3.9541      0.000269
    R1             0.57112     0.026908    21.224    4.4867e-25

>> plotResiduals(HW73lm,'fitted')
>> scatter(HW73.R1,HW73.R2)
>> lsline
```

b. Plot the residuals versus the fitted values. Does the linear model seem appropriate?

```
No. The relationship appears curved.
```


Plot of residuals vs. fitted values
Residuals
Fitted values
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
-0.5
1
1.5
2
2.5
3
3.5
4
4.5


5
4.5
4
3.5
3
2.5
2
1.5
1
1
2
3
4
5
6
7

c. Divide the data into two groups: points where $R_1 < 4$ in one group, points where $R_1 \geq 4$ in the other. Compute the least-squares line and the residual plot for each group. Does the line describe either group well? Which one?

```
>> fitlm(HW73(1:24,:))
Estimated Coefficients:
                   Estimate    SE          tStat     pValue
    (Intercept)      1.233     0.071943    17.139    3.2661e-14
    R1             0.26358     0.024188    10.897    2.4736e-10
```

$R_2 = 1.233 + 0.26358R_1$

```
>> fitlm(HW73(25:47,:))
Estimated Coefficients:
                   Estimate    SE          tStat      pValue
    (Intercept)    -0.19001     0.17932    -1.0596       0.30136
    R1              0.70999    0.032981     21.527    8.5715e-16
```

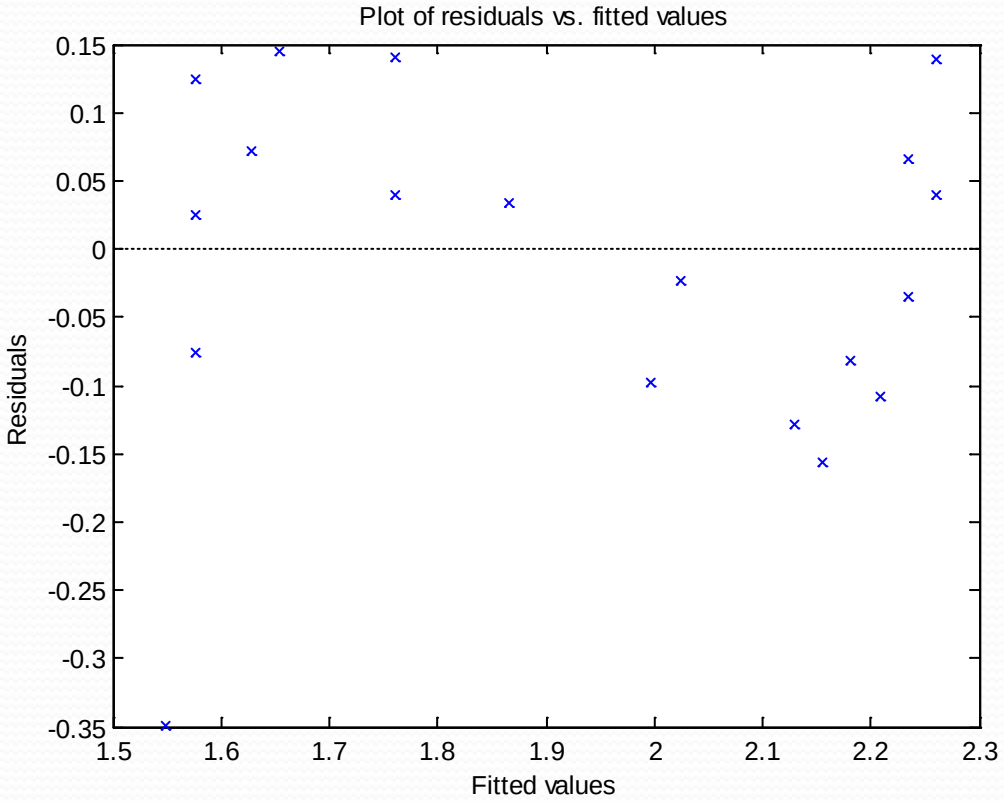
$R_2 = -0.19001 + 0.70999R_1$

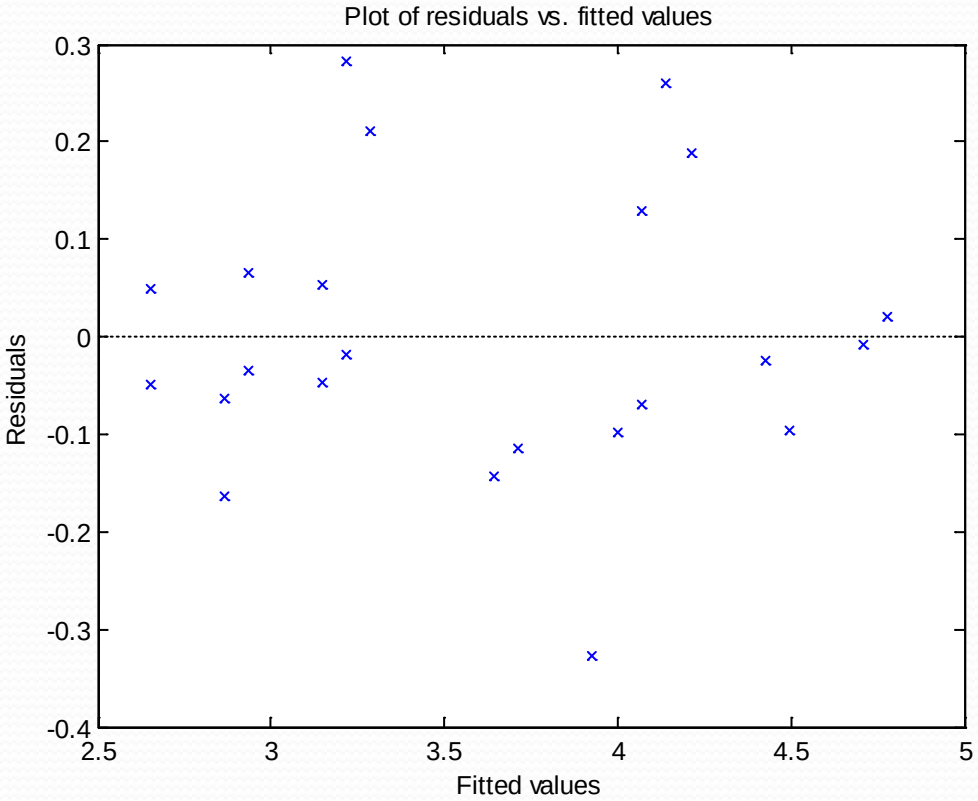
```
>> plotResiduals(fitlm(HW73(1:24,:)),'fitted')
>> plotResiduals(fitlm(HW73(25:47,:)),'fitted')
```

The first residual plot shows an S-shaped pattern, the second one, no pattern.

d. Explain why it might be a good idea to fit a linear model to part of these data, and a nonlinear model to the other.

The linear model seems to work well when R > 4 but not for smaller values. We might use a quadratic or cubic curve for that part (or a cubic spline overall, though that is beyond the scope of the course).


Plot of residuals vs. fitted values
Residuals
Fitted values
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
-0.2
-0.25
-0.3
-0.35
1.5
1.6
1.7
1.8
1.9
2
2.1
2.2
2.3


Plot of residuals vs. fitted values
Residuals
Fitted values
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
2.5
3
3.5
4
4.5
5

4. In an experiment to determine the factors affecting tensile strength in steel plates, the tensile strength (in kg/mm²), the manganese content (in parts per thousand), and the thickness (in mm) were measured for a sample of 20 plates. The following MINITAB output presents the results of fitting the model
Tensile strength = $\beta_0 + \beta_1$ Manganese $+\beta_2$ Thickness.

```
Strength  =  26.641  +  3.3201  Manganese  –  0.4249  Thickness

Predictor    Coef      StDev      T          P
Constant   26.641    2.72340   9.78      0.000
Manganese  3.3201    0.33198  10.00      0.000
Thickness –0.4249    0.12606  –3.37      0.004

S  =  0.8228        R-Sq  =  86.2%     R-Sq(adj)  =  84.6%

Analysis  of  Variance
Source              DF             SS        MS         F     P
Regression           2            72.01    36.005    53.19     0.000
Residual  Error      17           11.508    0.6769
Total                19           83.517
```

4. In an experiment to determine the factors affecting tensile strength in steel plates, the tensile strength (in kg/mm^2), the manganese content (in parts per thousand), and the thickness (in mm) were measured for a sample of 20 plates. The following MINITAB output presents the results of fitting the model Tensile strength = $\beta_0 + \beta_1$ Manganese$+\beta_2$ Thickness.

a. Predict the strength for a specimen that is 10 mm thick and contains 8.2 ppt manganese.

b. If two specimens have the same thickness, and one contains 10 ppt more manganese, by how much would you predict their strengths to differ?

c. If two specimens have the same proportion of manganese, and one is 5 mm thicker than the other, by how much would you predict their strengths to differ?

d. Find a 95% confidence interval for the coefficient of manganese.

e. Find a 99% confidence interval for the coefficient of thickness.

f. Can you conclude that $\beta_1 \neq 3$.

g. Can you conclude that $\beta_2 \neq -0.1$?

4. In an experiment to determine the factors affecting tensile strength in steel plates, the tensile strength (in kg/mm^2), the manganese content (in parts per thousand), and the thickness (in mm) were measured for a sample of 20 plates. The following MINITAB output presents the results of fitting the model
Tensile strength = $\beta_0 + \beta_1$ Manganese$+\beta_2$ Thickness.

Predictor	Coef	StDev	T	P
Constant	26.641	2.72340	9.78	0.000
Manganese	3.3201	0.33198	10.00	0.000
Thickness	−0.4249	0.12606	−3.37	0.004

a. Predict the strength for a specimen that is 10 mm thick and contains 8.2 ppt manganese.

$26.641 + 3.3201(8.2) - 0.4249(10) = 49.62$

b. If two specimens have the same thickness, and one contains 10 ppt more manganese, by how much would you predict their strengths to differ?

$(3.3201)(10) = 33.201$ kg/mm^2

c. If two specimens have the same proportion of manganese, and one is 5 mm thicker than the other, by how much would you predict their strengths to differ?

$(0.4249)(5) = 2.1245$ kg/mm^2

4. In an experiment to determine the factors affecting tensile strength in steel plates, the tensile strength (in kg/mm^2), the manganese content (in parts per thousand), and the thickness (in mm) were measured for a sample of 20 plates. The following MINITAB output presents the results of fitting the model
Tensile strength = β_0+ β_1 Manganese+β_2 Thickness.

Predictor	Coef	StDev	T	P
Constant	26.641	2.72340	9.78	0.000
Manganese	3.3201	0.33198	10.00	0.000
Thickness	−0.4249	0.12606	−3.37	0.004

d. Find a 95% confidence interval for the coefficient of manganese.

```
>> tinv(0.025,17) = 2.1098
3.3201 ± (2.1098)(0.33198)
3.3201 ± 0.7004
(2.6197, 4.0205)
```

e. Find a 99% confidence interval for the coefficient of thickness.

```
>> tinv(0.005,17) = 2.8982
-0.4249 ± (2.8982)(0.12606)
-0.4249 ± 0.3654
(-0.7903, -0.0595)
```

Predictor	Coef	StDev	T	P
Constant	26.641	2.72340	9.78	0.000
Manganese	3.3201	0.33198	10.00	0.000
Thickness	−0.4249	0.12606	−3.37	0.004

f. Can you conclude that $\beta_1 \neq 3$? Perform the appropriate hypothesis test.

First, 3 is in the CI (2.6197, 4.0205), so the hypothesis is not rejected. Also, (3.3201 −3)/0.33198 = 0.9642, so p = 2(0.1742) = 0.3484. The data are not inconsistent with $\beta_1 = 3$

g. Can you conclude that $\beta_2 \neq -0.1$? Perform the appropriate hypothesis test.

We can't use the CI from the previous slide if we want to use a p = 0.05 criterion. Directly, (−0.4249 + 0.1)/0.12606 = −2.577, p = 2(0.0098) = 0.0196. The data are not consistent with $\beta_2 = -0.1$.

5. The article "Multiple Linear Regression for Lake Ice and Lake Temperature Characteristics" (S. Gao and H. Stefan, Journal of Cold Regions Engineering, 1999:59–77) presents data on maximum ice thickness in mm (y), average number of days per year of ice cover (x_1), average number of days the bottom temperature is lower than 8°C (x_2), and the average snow depth in mm (x_3) for 13 lakes in Minnesota.

a. Fit the model $y = \beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\varepsilon$. For each coefficient, find the P-value for testing the null hypothesis that the coefficient is equal to 0.

b. If two lakes differ by 2 in the average number of days per year of ice cover, with other variables being equal, by how much would you expect their maximum ice thicknesses to differ?

c. Do lakes with greater average snow depth tend to have greater or lesser maximum ice thickness? Explain.

a. Fit the model $y = \beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\varepsilon$. For each coefficient, find the P-value for testing the null hypothesis that the coefficient is equal to 0.

```
>> fitlm(HW75,'y~x1+x2+x3')

Estimated Coefficients:
                    Estimate    SE          tStat      pValue
    (Intercept)    -372.98       242.37    -1.5389      0.15821
    x1              3.5368       1.1948     2.9602     0.015955
    x2              3.7345       1.1069     3.3738    0.0082067
    x3             -2.1661     0.85177      -2.543     0.031554
```

The p-values are given in the last column. All the coefficients are clearly non-zero except for the intercept.

b. If two lakes differ by 2 in the average number of days per year of ice cover, with other variables being equal, by how much would you expect their maximum ice thicknesses to differ?

If x1 differs by 2, the max ice thickness is predicted to differ by (2)(3.5368) = 7.07mm

c. Do lakes with greater average snow depth tend to have greater or lesser maximum ice thickness? Explain.

Greater average snow depth → less ice thickness (the coefficient is negative). (Snow is an insulator.)

6. The article "Vehicle-Arrival Characteristics at Urban Uncontrolled Intersections" (V. Rengaraju and V. Rao, Journal of Transportation Engineering, 1995:317–323) presents data on traffic characteristics at 10 intersections in Madras, India. The data set HW7-6.csv provides data on road width in m (x_1), traffic volume in vehicles per lane per hour (x_2), and median speed in km/h (y).

a. Fit the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$. Find the P-values for testing that the coefficients are equal to 0. Use the `anova` command to obtain the analysis of variance table.

b. Fit the model $y = \beta_0 + \beta_1 x_1 + \varepsilon$. Find the P-values for testing that the coefficients are equal to 0. Use the `anova` command to obtain the analysis of variance table.

c. Fit the model $y = \beta_0 + \beta_1 x_2 + \varepsilon$. Find the P-values for testing that the coefficients are equal to 0. Use the `anova` command to obtain the analysis of variance table.

d. Which of the models (a) through (c) do you think is best? Why?

a. Fit the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$. Find the P-values for testing that the coefficients are equal to o.

```
>> moda = fitlm(HW76,'y~x1+x2')

Linear regression model:
    y ~ 1 + x1 + x2

Estimated Coefficients:
                   Estimate        SE         tStat      pValue
                   ________    _________    _______    _________

    (Intercept)      25.613       10.424     2.4572     0.043647
    x1              0.18387      0.12353     1.4885      0.18024
    x2             -0.015878    0.0040542    -3.9164    0.0057757


Number of observations: 10, Error degrees of freedom: 7
Root Mean Squared Error: 3.07
R-squared: 0.712,  Adjusted R-Squared 0.63
F-statistic vs. constant model: 8.65, p-value = 0.0128
>> anova(moda)

              SumSq     DF    MeanSq      F         pValue
              ______    __    ______    ______    _________

    x1        20.938    1     20.938    2.2155      0.18024
    x2        144.95    1     144.95    15.338    0.0057757
    Error     66.154    7     9.4506
```

Variable x1 seems unimportant and variable x2 seems important.

b. Fit the model $y = \beta_0 + \beta_1 x_1 + \varepsilon$. Find the P-values for testing that the coefficients are equal to 0.

```
>> modb = fitlm(HW76,'y~x1')

Linear regression model:
    y ~ 1 + x1

Estimated Coefficients:
                   Estimate      SE        tStat     pValue
                   ________    _______    _______    _______

    (Intercept)     14.444      16.754    0.86215     0.4137
    x1             0.17334     0.20637    0.83993    0.42534


Number of observations: 10, Error degrees of freedom: 8
Root Mean Squared Error: 5.14
R-squared: 0.081,  Adjusted R-Squared -0.0338
F-statistic vs. constant model: 0.705, p-value = 0.425
>> anova(modb)

              SumSq     DF    MeanSq       F       pValue
              ______    __    ______    _______    _______

    x1        18.616    1     18.616    0.70547    0.42534
    Error     211.11    8     26.389
```

Variable x1 does not seem to be useful.

c. Fit the model $y = \beta_0 + \beta_1 x_2 + \varepsilon$. Find the P-values for testing that the coefficients are equal to 0.

```
>> modc = fitlm(HW76,'y~x2')

Linear regression model:
    y ~ 1 + x2

Estimated Coefficients:
                   Estimate        SE         tStat      pValue
                   _________    _________    _______    _________

    (Intercept)        40.37       3.4545     11.686    2.623e-06
    x2              -0.015747    0.0043503    -3.6197    0.0067859


Number of observations: 10, Error degrees of freedom: 8
Root Mean Squared Error: 3.3
R-squared: 0.621,  Adjusted R-Squared 0.573
F-statistic vs. constant model: 13.1, p-value = 0.00679
>> anova(modc)

              SumSq     DF    MeanSq      F         pValue
              ______    __    ______    ______    _________

    x2        142.63    1     142.63    13.102    0.0067859
    Error     87.091    8     10.886
```

Variable x2 seems to be a useful predictor.

d. Which of the models (a) through (c) do you think is best? Why?

x1 is not significant in the combined model, or in the model with just x1, so model c) looks best. x2 is significant there. The RMSE is smaller in a), but not by enough to be statistically significant. Some criteria such as the AIC would keep variable x1.

7. The article "Modeling Resilient Modulus and Temperature Correction for Saudi Roads" (H. Wahhab, I. Asi, and R. Ramadhan, Journal of Materials in Civil Engineering, 2001:298–305) describes a study designed to predict the resilient modulus of pavement from physical properties. The file `HW7-7.csv` presents data for the resilient modulus at 40°C (104°F) in 10^6 kPa (y), the surface area of the aggregate in m^2/kg (x_1), and the softening point of the asphalt in °C (x_2).

The full quadratic model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 + \varepsilon.$$

Which submodel of this full model do you believe is most appropriate? Justify your answer by fitting two or more models and comparing the results.

```
>> model = fitlm(HW77,'y~x1*x2+x1^2+x2^2')

Linear regression model:
    y ~ 1 + x1*x2 + x1^2 + x2^2

Estimated Coefficients:
                    Estimate        SE         tStat      pValue
                   __________    _________    ________    _______

    (Intercept)       -8.9597       50.447     -0.1776    0.86297
    x1                -1.0985       3.8399    -0.28608    0.78129
    x2                0.40155       1.1682     0.34374    0.73894
    x1:x2            0.081573      0.10349     0.78826    0.45081
    x1^2             -0.27101      0.26102     -1.0383    0.32622
    x2^2            -0.0073347    0.0055679     -1.3173    0.22028


Number of observations: 15, Error degrees of freedom: 9
Root Mean Squared Error: 0.66
R-squared: 0.511,  Adjusted R-Squared 0.239
F-statistic vs. constant model: 1.88, p-value = 0.193
>> anova(model)

              SumSq     DF    MeanSq        F       pValue
             _______    __    _______    _______    _______

    x1       0.10484    1     0.10484    0.24046    0.63561
    x2        2.3467    1      2.3467     5.3821    0.04549
    x1:x2    0.27091    1     0.27091    0.62135    0.45081
    x1^2     0.47005    1     0.47005     1.0781    0.32622
    x2^2     0.75662    1     0.75662     1.7353    0.22028
    Error     3.9241    9     0.43601
```

Neither quadratic term is significant, nor is the product of x1 and x2. We remove the product since it has the worst p-value of the terms that can be removed. Note that variable x2 is significant in the anova ('h') because it is tested against a model not including x1:x2 and x2^2.

```
>> model = fitlm(HW77,'y~x1+x2+x1^2+x2^2')

Linear regression model:
    y ~ 1 + x1 + x2 + x1^2 + x2^2

Estimated Coefficients:
                    Estimate        SE         tStat      pValue
                   __________    _________    ________    _______

    (Intercept)       -42.932       25.718     -1.6693    0.12601
    x1                 1.4936       1.9449     0.76795    0.46027
    x2                 1.1357      0.69164      1.6421    0.13161
    x1^2            -0.091797      0.12576    -0.72995    0.48217
    x2^2            -0.0081813    0.0053589     -1.5267    0.15783


Number of observations: 15, Error degrees of freedom: 10
Root Mean Squared Error: 0.648
R-squared: 0.477,  Adjusted R-Squared 0.268
F-statistic vs. constant model: 2.28, p-value = 0.132
>> anova(model)

              SumSq     DF    MeanSq        F         pValue
             _______    __    _______    _______    ________

    x1       0.10484     1    0.10484    0.24992     0.62795
    x2        2.3467     1     2.3467     5.5939    0.039604
    x1^2     0.22352     1    0.22352    0.53282     0.48217
    x2^2     0.97776     1    0.97776     2.3308     0.15783
    Error      4.195    10     0.4195
```

Neither quadratic term is significant so we remove x1^2 since it has the worse p-value of the two terms that can be removed.

```
>> model = fitlm(HW77,'y~x1+x2+x2^2')

model =


Linear regression model:
    y ~ 1 + x1 + x2 + x2^2

Estimated Coefficients:
                    Estimate        SE         tStat     pValue
                   __________    _________    _______    _______

    (Intercept)       -32.114       20.568    -1.5614    0.14673
    x1               0.078541      0.15373    0.51089    0.61953
    x2                0.96533      0.63709     1.5152    0.15791
    x2^2           -0.0068648    0.0049379    -1.3902    0.19195


Number of observations: 15, Error degrees of freedom: 11
Root Mean Squared Error: 0.634
R-squared: 0.449,  Adjusted R-Squared 0.299
F-statistic vs. constant model: 2.99, p-value = 0.0773
>> anova(model)

ans =

               SumSq     DF    MeanSq        F         pValue
              _______    __    _______    _______    ________

    x1       0.10484      1    0.10484    0.26101     0.61953
    x2        2.3356      1     2.3356     5.8144    0.034534
    x2^2     0.77634      1    0.77634     1.9327     0.19195
    Error     4.4185     11    0.40169
```

We could remove x1 or x2^2. Even though x1 has a worse p-value, it may make sense to remove the quadratic term first, so we do that.

```
>> model = fitlm(HW77,'y~x1+x2')

Linear regression model:
    y ~ 1 + x1 + x2

Estimated Coefficients:
                   Estimate       SE        tStat      pValue
                   ________    ________    _______    ________

    (Intercept)     -3.6833      2.2758    -1.6185     0.13153
    x1             0.062295     0.15913    0.39146     0.70232
    x2             0.080868    0.034816     2.3227    0.038575


Number of observations: 15, Error degrees of freedom: 12
Root Mean Squared Error: 0.658
R-squared: 0.353,  Adjusted R-Squared 0.245
F-statistic vs. constant model: 3.27, p-value = 0.0736
>> anova(model)

               SumSq      DF     MeanSq        F         pValue
              ________    __    ________    _______    ________

    x1       0.066339     1    0.066339    0.15324     0.70232
    x2         2.3356     1      2.3356     5.3951    0.038575
    Error      5.1949    12     0.43291
```

Now we remove x1 leaving only x2 as a predictor.

```
>> model = fitlm(HW77,'y~x2')

Linear regression model:
    y ~ 1 + x2

Estimated Coefficients:
                   Estimate       SE        tStat      pValue
                   ________    ________    _______    ________

    (Intercept)     -3.4482      2.1224    -1.6247     0.12822
    x2             0.084601    0.032375     2.6131    0.021464


Number of observations: 15, Error degrees of freedom: 13
Root Mean Squared Error: 0.636
R-squared: 0.344,  Adjusted R-Squared 0.294
F-statistic vs. constant model: 6.83, p-value = 0.0215
>> anova(model)

             SumSq     DF    MeanSq       F         pValue
             ______    __    _______    ______    ________

    x2       2.7636     1     2.7636    6.8285    0.021464
    Error    5.2612    13    0.40471
```

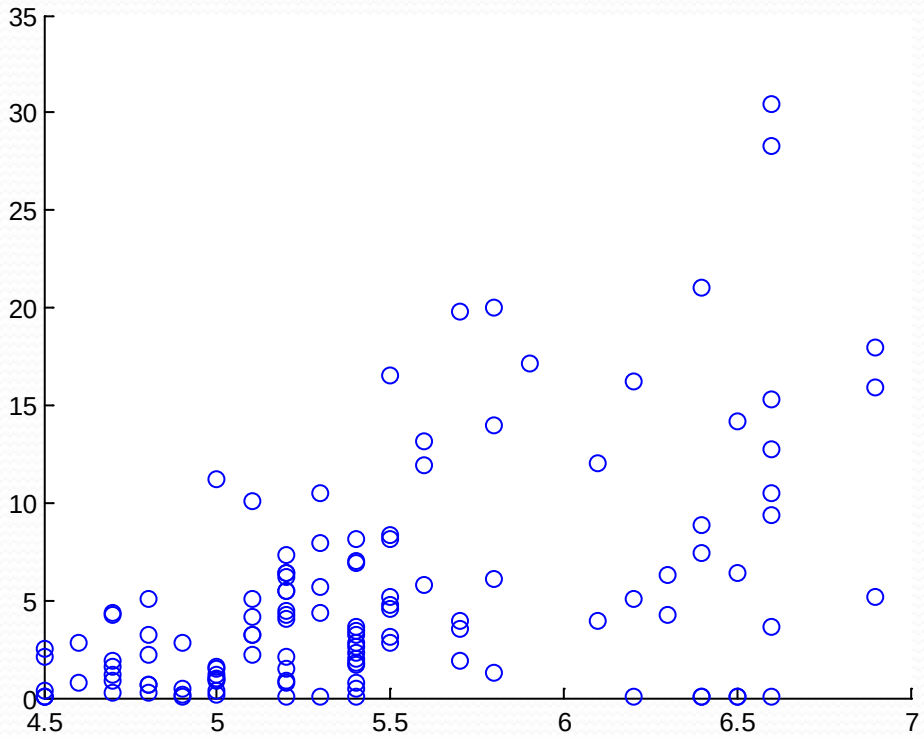
We keep x2 because it appears to be a useful predictor.

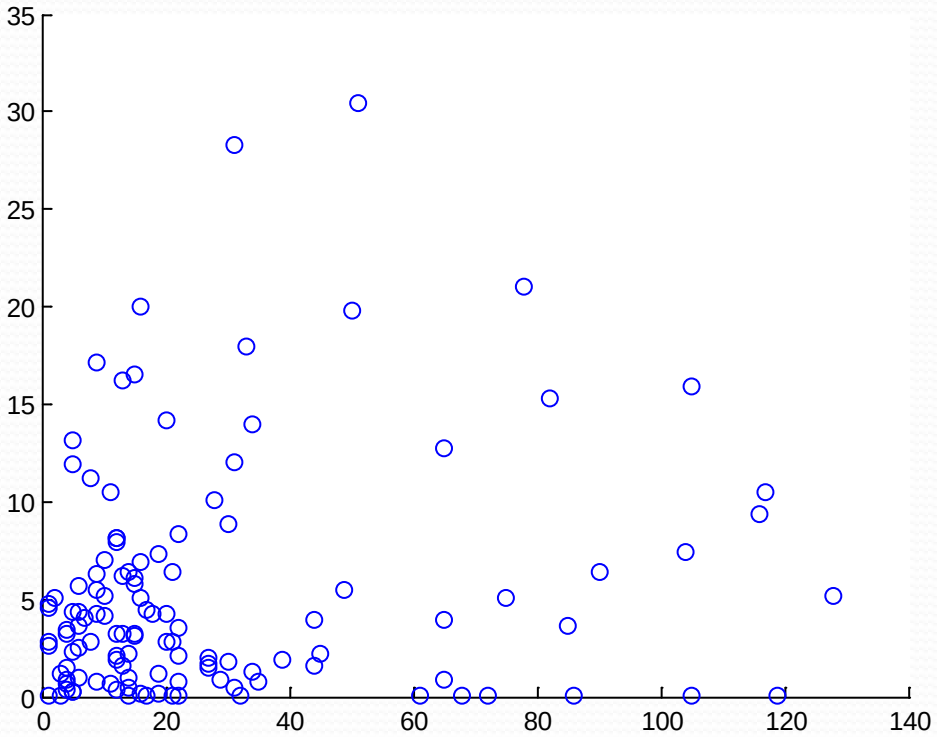
```
>> stepwiselm(HW77,'y~x1*x2+x1^2+x2^2')
1. Removing x1:x2, FStat = 0.62135, pValue = 0.45081
2. Removing x1^2, FStat = 0.53282, pValue = 0.48217
3. Removing x1, FStat = 0.26101, pValue = 0.61953
4. Removing x2^2, FStat = 1.9574, pValue = 0.18711

Linear regression model:
    y ~ 1 + x2

Estimated Coefficients:
                   Estimate    SE          tStat      pValue
    (Intercept)     -3.4482      2.1224    -1.6247     0.12822
    x2              0.084601    0.032375     2.6131    0.021464

Number of observations: 15, Error degrees of freedom: 13
Root Mean Squared Error: 0.636
R-squared: 0.344,  Adjusted R-Squared 0.294
F-statistic vs. constant model: 6.83, p-value = 0.0215
```

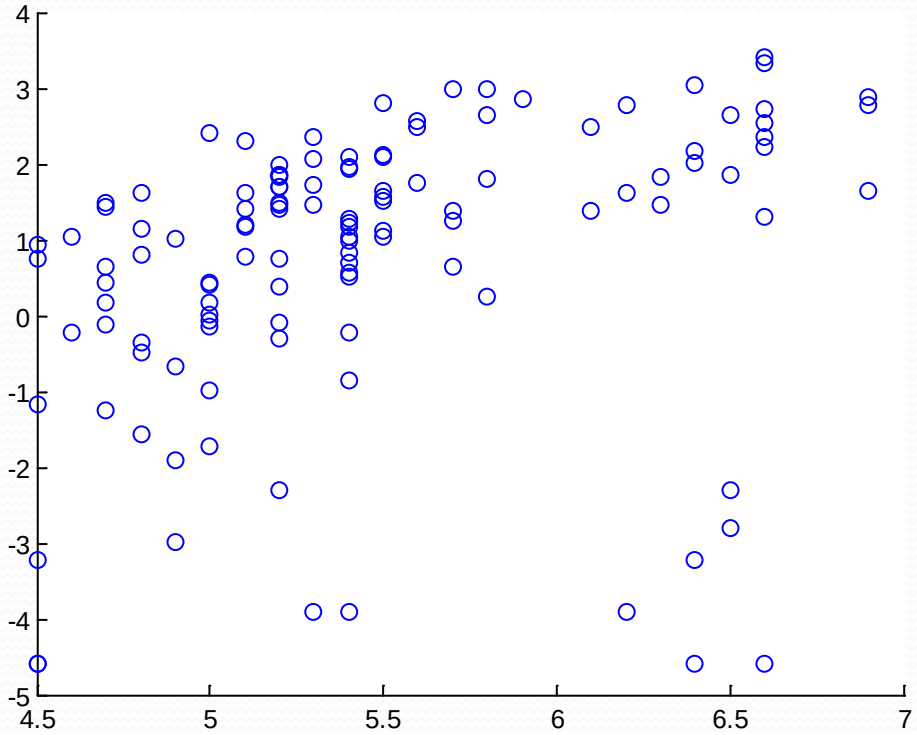
8. The article "Seismic Hazard in Greece Based on Different Strong Ground Motion Parameters" (S. Koutrakis, G. Karakaisis, et al., Journal of Earthquake Engineering, 2002:75–109) presents a study of seismic events in Greece during the period 1978–1997. Of interest is the duration of "strong ground motion," which is the length of time that the acceleration of the ground exceeds a specified value. For each event, measurements of the duration of strong ground motion were made at one or more locations. The file `HW7-8.csv` presents, for each of 121 such measurements, the data for the duration of time y (in seconds) that the ground acceleration exceeded twice the acceleration due to gravity, the magnitude m of the earthquake, the distance d (in km) of the measurement from the epicenter, and two indicators of the soil type s_1 and s_2, defined as follows: s_1 = 1 if the soil consists of soft alluvial deposits, s_1 = 0 otherwise, and s_2 = 1 if the soil consists of tertiary or older rock, s_2 = 0 otherwise. Cases where both s_1 = 0 and s_2 = 0 correspond to intermediate soil conditions. The article presents repeated measurements at some locations, which we have not included here.

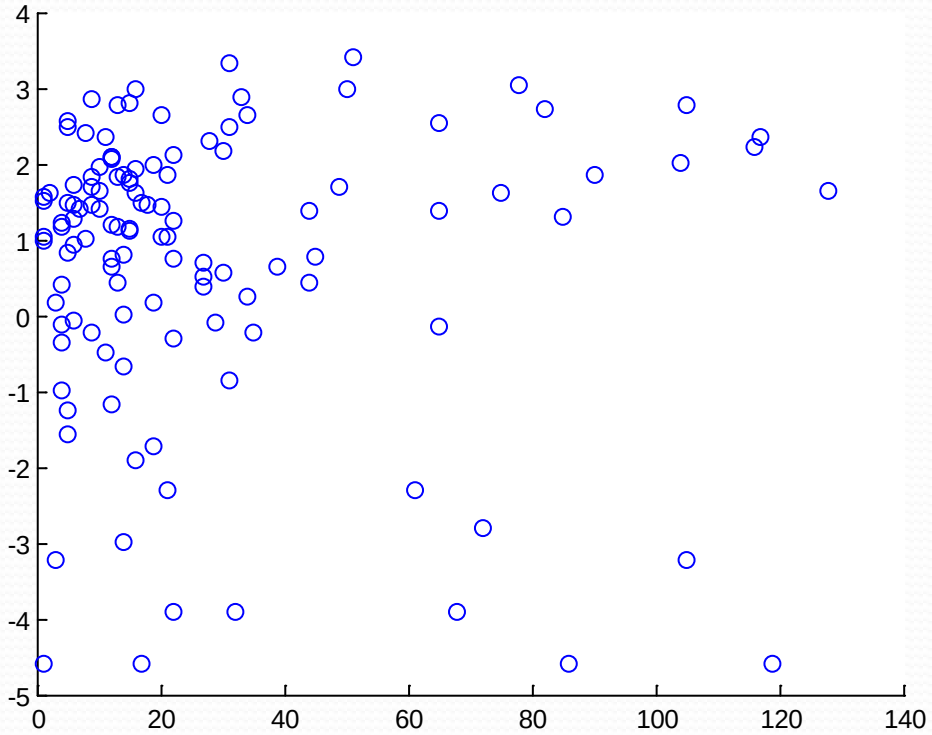
Use the data to construct a linear model to predict duration y from some or all of the variables m, d, s_1, and s_2. Be sure to consider transformations of the variables, as well as powers of and interactions between the independent variables. Describe the steps taken to construct your model. Plot the residuals versus the fitted values to verify that your model satisfies the necessary assumptions. In addition, note that the data is presented in chronological order, reading down the columns. Make a plot to determine whether time should be included as an independent variable.

```
>> fitlm(HW78,'y~m+d+s1+s2')


Linear regression model:
    y ~ 1 + m + d + s1 + s2

Estimated Coefficients:
                   Estimate     SE          tStat      pValue
    (Intercept)      -33.881      4.8669    -6.9615    2.1494e-10
    m                  7.592     0.95277     7.9684    1.2345e-12
    d              -0.076981    0.020769    -3.7065    0.00032359
    s1               0.68398     0.97954    0.69826       0.48641
    s2               -1.9942      1.2088    -1.6497       0.10171


Number of observations: 121, Error degrees of freedom: 116
Root Mean Squared Error: 4.71
R-squared: 0.389,  Adjusted R-Squared 0.368
F-statistic vs. constant model: 18.4, p-value = 9.51e-12

>> scatter(HW78.m, HW78.y)
>> scatter(HW78.d, HW78.y)
>> scatter(HW78.m,log(HW78.y))
>> scatter(HW78.d,log(HW78.y))
```

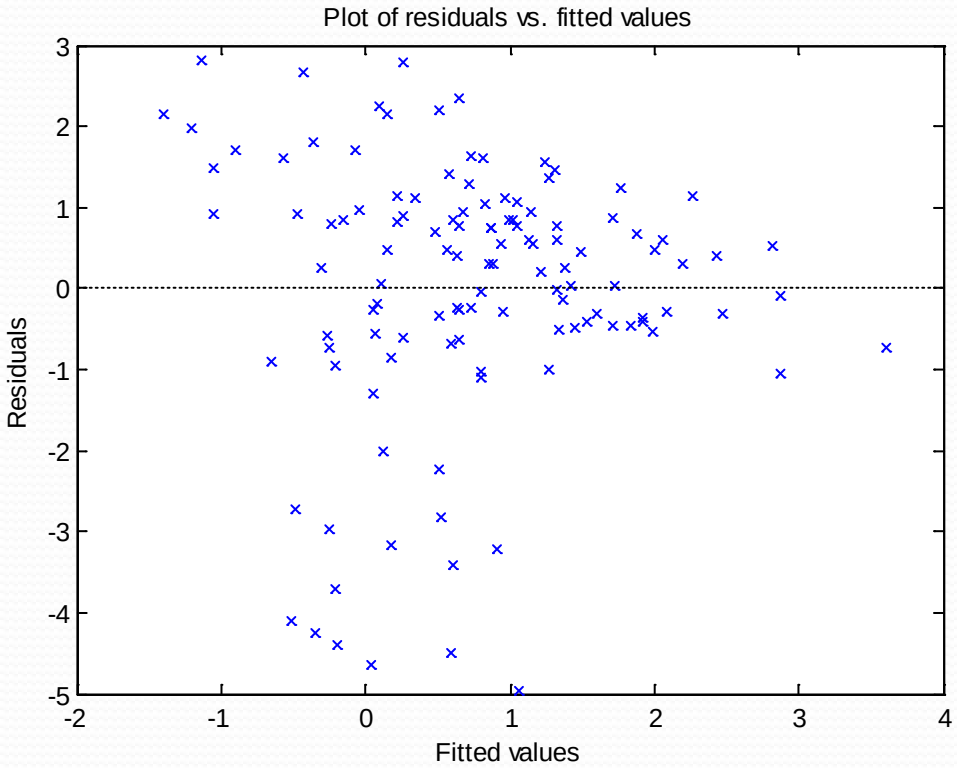

35
30
25
20
15
10
5
0
4.5
5
5.5
6
6.5
7

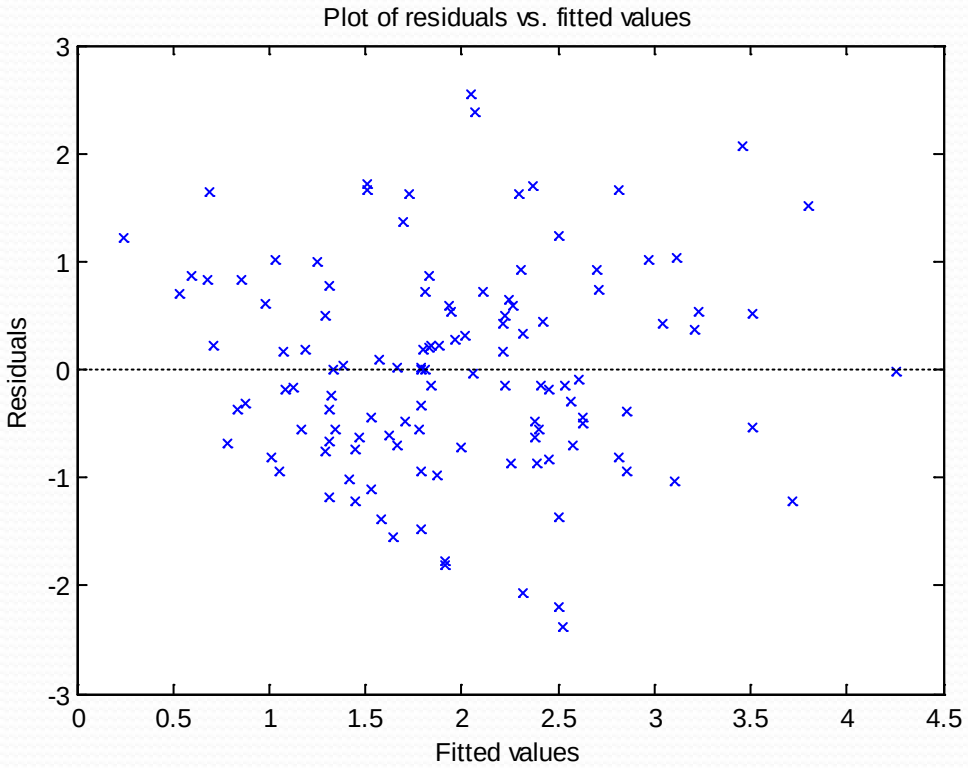

35
30
25
20
15
10
5
0
0
20
40
60
80
100
120
140


4
3
2
1
0
-1
-2
-3
-4
-5
4.5
5
5.5
6
6.5
7


4
3
2
1
0
-1
-2
-3
-4
-5
0
20
40
60
80
100
120
140

```
>> HW78a = [HW78 table(log(HW78.y),'VariableNames',{'logy'})]
>> HW78afit1 = fitlm(HW88a,'logy~m+d+s1+s2')

Linear regression model:
    logy ~ 1 + m + d + s1 + s2

Estimated Coefficients:
                   Estimate        SE         tStat        pValue
                   ________    _________    ________    __________

    (Intercept)        -8.13        1.719     -4.7294    6.3937e-06
    m                 1.8332      0.33653      5.4474    2.9055e-07
    d              -0.027724    0.0073358     -3.7793    0.00024996
    s1              -0.29199      0.34598    -0.84394       0.40044
    s2               -1.1798      0.42696     -2.7632     0.0066576

Number of observations: 121, Error degrees of freedom: 116
Root Mean Squared Error: 1.66
R-squared: 0.246,  Adjusted R-Squared 0.22
F-statistic vs. constant model: 9.48, p-value = 1.15e-06

>> plotResiduals(HW78afit1,'fitted')
>> HW78b = [HW78a table(sqrt(HW78.y),'VariableNames',{'sqrty'})]
>> plotResiduals(fitlm(HW78b,'sqrty~m+d+s1+s2'),'fitted')
```


Plot of residuals vs. fitted values
Residuals
Fitted values
3
2
1
0
-1
-2
-3
-4
-5
-2
-1
0
1
2
3
4


Plot of residuals vs. fitted values
Residuals
Fitted values
3
2
1
0
-1
-2
-3
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5

```
>> crosstab(HW78.s1,HW78.s2)
    43    26
    52     0

>> HW78c = [HW78b table(HW78.s1+2*HW78.s2,'VariableNames',{'s12'})]
>> tabulate(HW78c.s12)
  Value    Count   Percent
      0       43     35.54%
      1       52     42.98%
      2       26     21.49%
>> HW78fit2 = fitlm(HW78c,'sqrty~m+d+s12','CategoricalVars',8)

Linear regression model:
    sqrty ~ 1 + m + d + s12

Estimated Coefficients:
                   Estimate        SE          tStat        pValue
                   ________    _________    _________    __________

    (Intercept)      -5.9497       1.0112      -5.8839    3.9605e-08
    m                  1.563      0.19795       7.8955    1.8055e-12
    d              -0.017569    0.0043151      -4.0716    8.5648e-05
    s12_1           -0.01674      0.20352    -0.082255       0.93459
    s12_2           -0.63125      0.25115      -2.5135      0.013327
```

```
>> HW78fit3 = fitlm(HW78c,'sqrty~m*d*s12+m^2+d^2','CategoricalVars',8)
>> anova(HW78fit3)
```

	SumSq	DF	MeanSq	F	pValue
m	58.207	1	58.207	62.253	2.7683e-12
d	12.381	1	12.381	13.242	0.00042325
s12	7.6971	2	3.8485	4.116	0.018961
m:d	0.86925	1	0.86925	0.92967	0.33712
m:s12	4.2806	2	2.1403	2.2891	0.1063
d:s12	0.84531	2	0.42266	0.45204	0.63754
m^2	0.37634	1	0.37634	0.4025	0.52716
d^2	0.0026134	1	0.0026134	0.0027951	0.95794
m:d:s12	4.896	2	2.448	2.6182	0.077605
Error	100.05	107	0.93501		

```
>> HW78d = HW78c
>> HW78d(:,[1 4 5 6]) = []
>> HW78fit4 = stepwiselm(HW78d,'sqrty~m*d*s12+m^2+d^2','CategoricalVars',4)
1. Removing d^2, FStat = 0.0027951, pValue = 0.95794
2. Removing m^2, FStat = 0.52405, pValue = 0.47069

Linear regression model:
    sqrty ~ 1 + m*d + m*s12 + d*s12 + m:d:s12
Estimated Coefficients:
                    Estimate        SE        tStat       pValue
                    ________    ________    ________    ________

    (Intercept)      -3.0645      2.3504     -1.3038     0.19505
    m                 1.1048     0.42472      2.6013    0.010577
    d                -0.2403     0.10611     -2.2647     0.02551
    s12_1            -5.5512      2.9857     -1.8593    0.065686
    s12_2           -0.80749       3.648    -0.22135     0.82523
    m:d             0.035229    0.016679      2.1122     0.03695
    m:s12_1          0.94322     0.55089      1.7122    0.089707
    m:s12_2        -0.075238     0.66643     -0.1129     0.91032
    d:s12_1          0.26996     0.11508      2.3459    0.020791
    d:s12_2          0.28283     0.13398       2.111    0.037059
    m:d:s12_1      -0.043186     0.01813      -2.382     0.01895
    m:d:s12_2      -0.043778    0.021088     -2.0759    0.040252


Number of observations: 121, Error degrees of freedom: 109
Root Mean Squared Error: 0.96
R-squared: 0.441,  Adjusted R-Squared 0.385
F-statistic vs. constant model: 7.83, p-value = 7.25e-10
```

```
>> anova(HW78fit4)
```

	SumSq	DF	MeanSq	F	pValue
m	58.764	1	58.764	63.713	1.5773e-12
d	13.293	1	13.293	14.412	0.00024184
s12	7.5885	2	3.7943	4.1138	0.018951
m:d	0.212	1	0.212	0.22986	0.63259
m:s12	3.831	2	1.9155	2.0768	0.13026
d:s12	0.63682	2	0.31841	0.34522	0.70883
m:d:s12	5.4636	2	2.7318	2.9619	0.055901
Error	100.53	109	0.92233		

```
>> anova(fitlm(HW78d,'sqrty~m+d+s12+m:d+m:s12+d:s12','CategoricalVars',4))
```

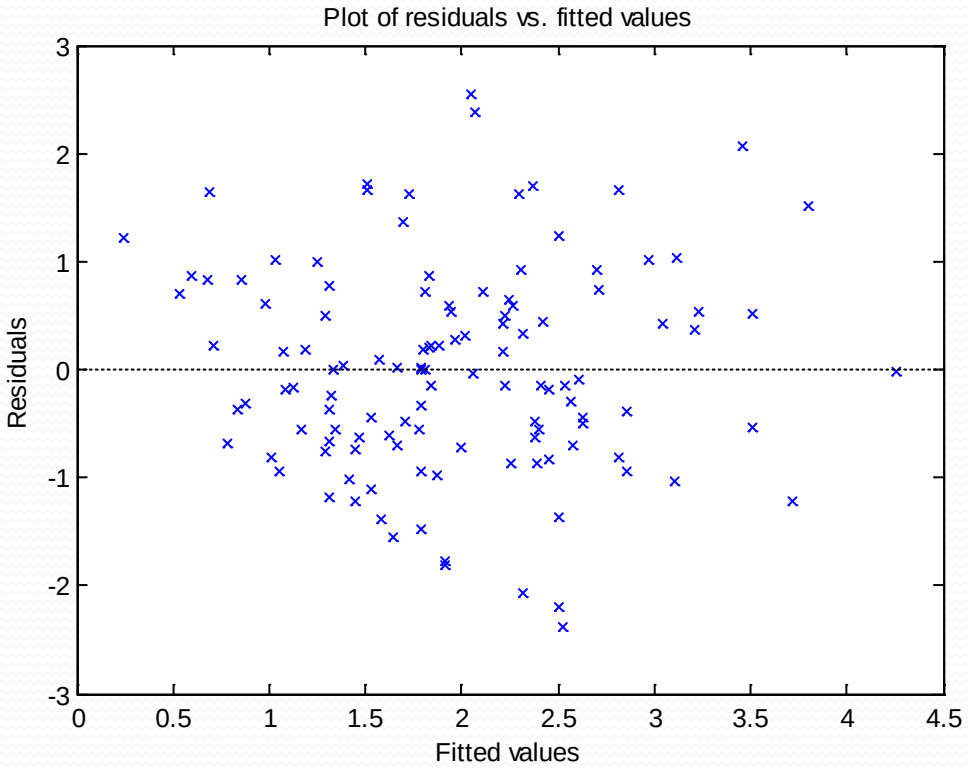
	SumSq	DF	MeanSq	F	pValue
m	58.764	1	58.764	61.537	2.9102e-12
d	13.293	1	13.293	13.92	0.00030239
s12	7.5885	2	3.7943	3.9733	0.021546
m:d	0.212	1	0.212	0.22201	0.63844
m:s12	3.831	2	1.9155	2.0059	0.13938
d:s12	0.63682	2	0.31841	0.33344	0.71717
Error	106	111	0.95493		

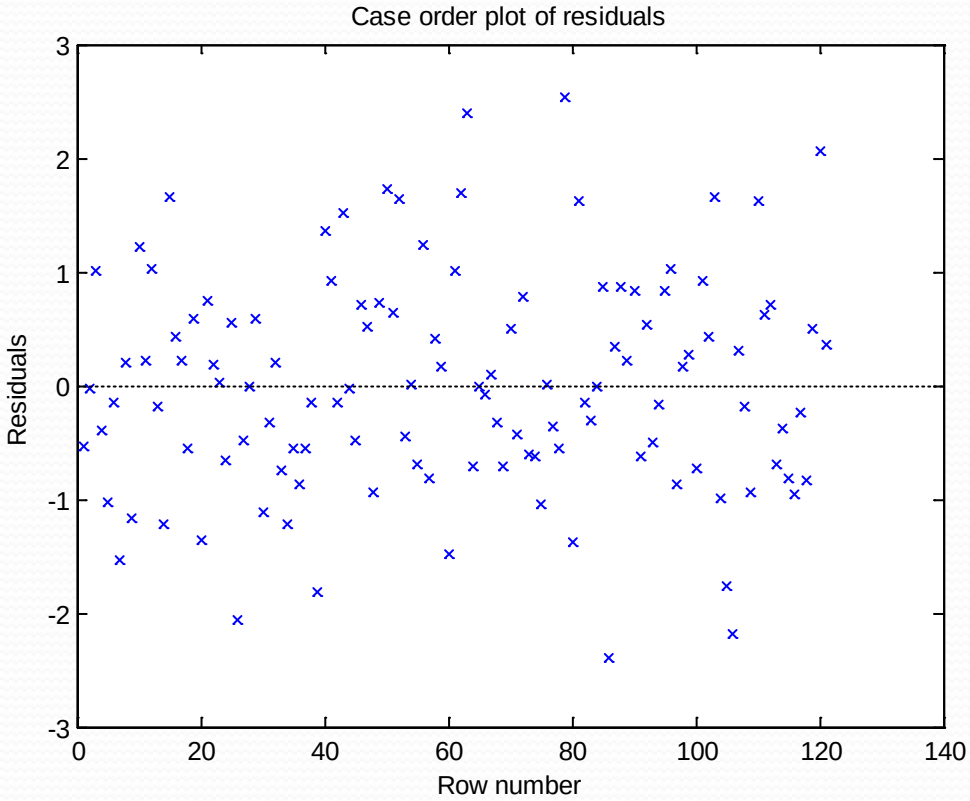
```
>> anova(fitlm(HW78d,'sqrty~m+d+s12+m:d+m:s12','CategoricalVars',4))

              SumSq     DF     MeanSq        F          pValue
              _____    ___    _______    _______    __________

    m          59.688      1     59.688     63.252    1.5236e-12
    d          13.293      1     13.293     14.086    0.00027743
    s12        7.5885      2     3.7943     4.0208      0.020565
    m:d       0.17096      1    0.17096    0.18116        0.67119
    m:s12      4.1634      2     2.0817      2.206        0.11487
    Error      106.63    113    0.94367
>> anova(fitlm(HW78d,'sqrty~m+d+s12+m:s12','CategoricalVars',4))

              SumSq     DF     MeanSq        F          pValue
              _____    ___    _______    _____    __________

    m         59.688      1     59.688    63.709    1.2558e-12
    d         13.293      1     13.293    14.188     0.0002634
    s12       7.5262      2     3.7631    4.0166      0.020623
    m:s12     4.2614      2     2.1307    2.2742        0.10752
    Error     106.81    114    0.93689
```

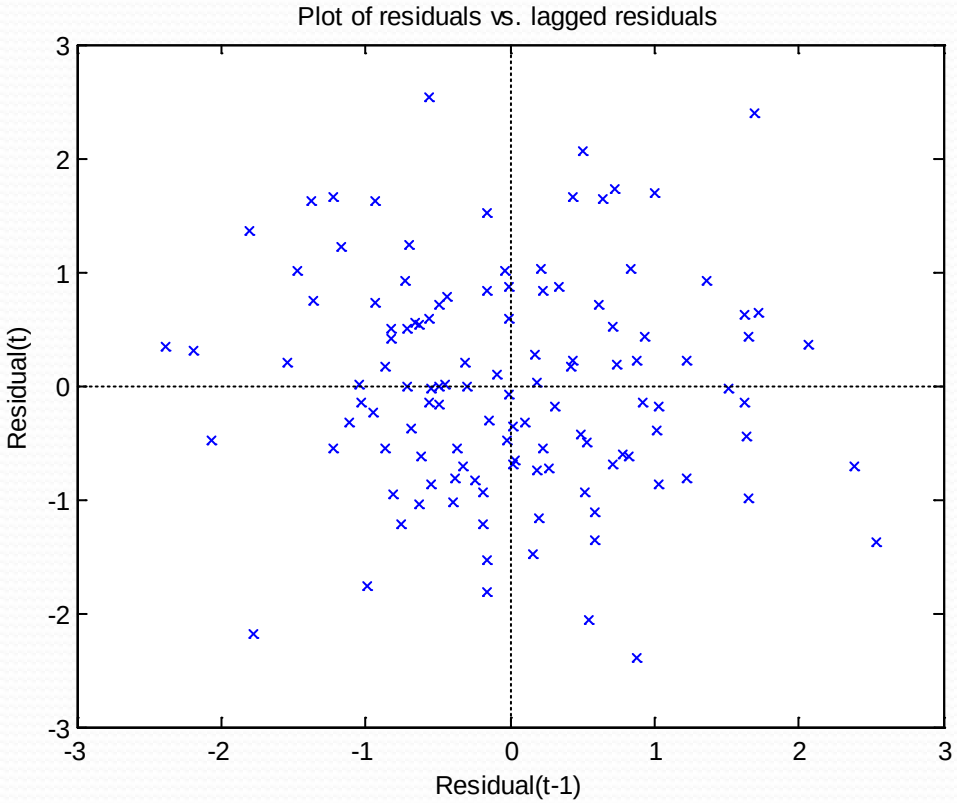
```
>> anova(fitlm(HW78d,'sqrty~m+d+s12','CategoricalVars',4))

ans =
```

	SumSq	DF	MeanSq	F	pValue
m	59.688	1	59.688	62.34	1.8055e-12
d	15.873	1	15.873	16.578	8.5648e-05
s12	7.5262	2	3.7631	3.9302	0.022308
Error	111.07	116	0.95747		

```
>> HW78fit5 = fitlm(HW78d,'sqrty~m+d+s12','CategoricalVars',4)


Linear regression model:
    sqrty ~ 1 + m + d + s12

Estimated Coefficients:
                   Estimate        SE          tStat        pValue
                   ________    _________    _________    __________

    (Intercept)      -5.9497       1.0112      -5.8839    3.9605e-08
    m                  1.563      0.19795       7.8955    1.8055e-12
    d              -0.017569    0.0043151      -4.0716    8.5648e-05
    s12_1           -0.01674      0.20352    -0.082255       0.93459
    s12_2           -0.63125      0.25115      -2.5135      0.013327


Number of observations: 121, Error degrees of freedom: 116
Root Mean Squared Error: 0.979
R-squared: 0.383,  Adjusted R-Squared 0.361
F-statistic vs. constant model: 18, p-value = 1.63e-11
>> plotResiduals(HW78fit5,'fitted')
>> plotResiduals(HW78fit5, 'caseorder')
>> plotResiduals(HW78fit5,'lagged')
```


Plot of residuals vs. fitted values
Residuals
Fitted values
3
2
1
0
-1
-2
-3
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5


Case order plot of residuals
Residuals
Row number
3
2
1
0
-1
-2
-3
0
20
40
60
80
100
120
140


Plot of residuals vs. lagged residuals
Residual(t)
Residual(t-1)
3
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3

```
>> [xc lags] = xcorr(HW78fit5.Residuals.Raw,5,'coef')

xc =

   -0.1364
   -0.0565
   -0.0449
    0.0610
   -0.0626
    1.0000        #lag 0 correlation is always 1.0
   -0.0626        #the autocorrelations are all small (large would be near ±1)
    0.0610
   -0.0449
   -0.0565
   -0.1364

lags =

    -5    -4    -3    -2    -1     0     1     2     3     4     5


So no problem with time correlations
```

Other comments

- We could have used log distance instead of distance.
- Log magnitude would not probably be a good idea because Richter Scale magnitude is already on the log scale.
- Use of the square root of shaking time made the model simple and with no apparent violations of the assumptions.
- Many other approaches would lead to useful predictive models.