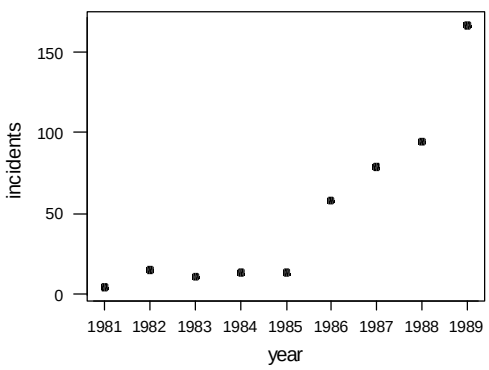
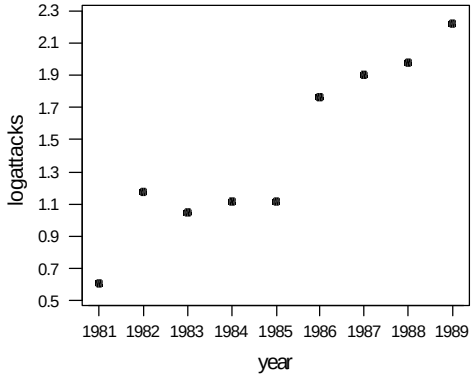
Solution Set: Supplementary Homework 10

Hate Crime in LA

1. *Plot the data, putting year on the X-axis. Is the relationship between year and the number of reported incidents linear? Now "transform" the number of incidents: create a new variable equal to the natural logarithm of the number of incidents. Plot this new variable against year. Is this linear?*





The number of incidents seems to be accelerating with time. Taking the log of the number of incidents makes the pattern seem much more linear (with a few outliers, such as 1982 and 1985).

2. *Regress number (not log) of incidents (the dependent variable) on year (the independent variable). Interpret the results, paying particular attention to the intercept, slope, standard errors of the coefficients, R2, and estimated disturbance variance. Is this a good predictive model?*

```
The regression equation is incidents = - 35448 + 17.9 year

Predictor        Coef       StDev          T        P
Constant      -35448        6868      -5.16    0.001
year          17.883       3.460       5.17    0.001

S = 26.80       R-Sq = 79.2%     R-Sq(adj) = 76.3%
```

In Roman times, there were on average -35,448 hate crimes in Los Angeles. For each year since, hate crimes are expected to increase at a rate of 17.9 per year. This estimate is very unlikely to be the product of chance (a two-tailed test of the null hypothesis that b=0 is rejected at p=.001). The independent variable (years) predicts 79% of the year-to-year variability in hate crime rates. The estimated standard deviation of the disturbances is 26.8 incidents, which leaves a great deal of uncertainty in the prediction of any given year's rate.

3. *Using the slope and intercept estimates, forecast the number of incidents for 1990 and for*

1991. Calculate a 90% confidence interval for each of your forecasts. Which year's confidence interval is larger?

```
Predicted Values: 1990

     Fit  StDev Fit         90.0% CI             90.0% PI
  139.97      19.47   (  103.09,  176.86)  (   77.21,  202.73)

Predicted Values: 1991

     Fit  StDev Fit         90.0% CI             90.0% PI
  157.86      22.60   (  115.04,  200.67)  (   91.44,  224.27)
```

The intervals become wider as X gets farther away from its mean. Thus, the uncertainty is greater for year=1991 than for year=1990.

Reported hate crimes rose to 275 in 1990; 351 in 1991. The forecasts do a poor job.

4. *Replicate #2 using "time" instead of year? Does this change your results? What do you infer from this exercise?*

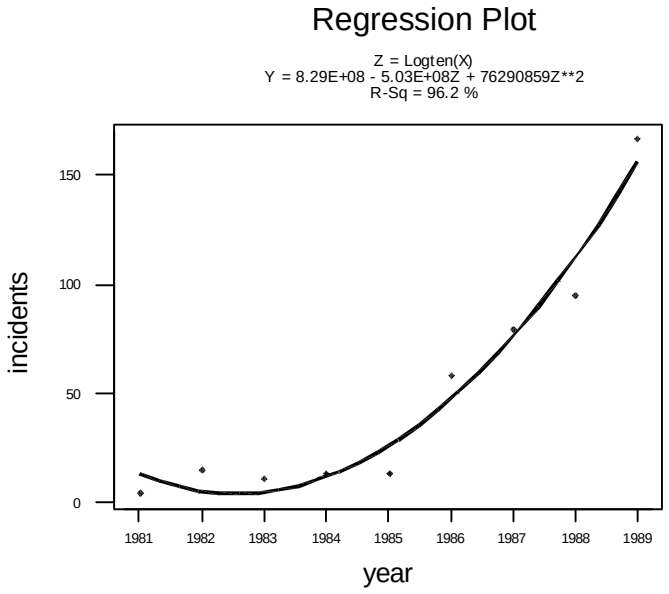
```
The regression equation is incidents = - 21.0 + 17.9 time

Predictor        Coef       StDev          T        P
Constant       -20.98       16.47      -1.27    0.243
time           17.883       3.460       5.17    0.001

S = 26.80       R-Sq = 79.2%     R-Sq(adj) = 76.3%
```

Same results, except for the intercept, which now reflects the expected value of Y in 1981. Adding a constant to the independent variable only affects the estimate of the intercept.

5. *Now, revise the model by including an additional independent variable, the square of year. (Again, the dependent variable is the number of incidents.) Does this improve the "fit" of your regression model to the data from the 1980s? What statistics led you to your conclusion?*

```
The regression equation is
incidents =14339526 - 14466 year + 3.65 year2

Predictor        Coef       StDev          T        P
Constant     14339526     2792133       5.14    0.002
year           -14466        2813      -5.14    0.002
year2          3.6483      0.7086       5.15    0.002

S = 12.44       R-Sq = 96.2%     R-Sq(adj) = 94.9%
```

The fit improves. Note the increase in the R-squared and decrease in s. Substantively, the fact that the squared term is positive means that the effect of year is accelerating over time, as the shown in the following graph:



Note the problem with quadratic models: the implication is that hate crimes in 1981 are expected to be higher than 1982. If we work backwards in time, the expected hate crime rate goes upwards (the regression line is a parabola).

6. *Repeat #3 using your revised regression model. Does the revised model lead to an improved "out of sample forecast"?*

```
Predicted Values: 1990
     Fit  StDev Fit         90.0% CI            90.0% PI
  206.86      15.82   (  176.11,  237.61)  (  167.75,  245.97)

Predicted Values: 1991

     Fit  StDev Fit         90.0% CI            90.0% PI
  264.87      23.28   (  219.63,  310.11)  (  213.58,  316.16)
```

These numbers are still too low, but they are closer.

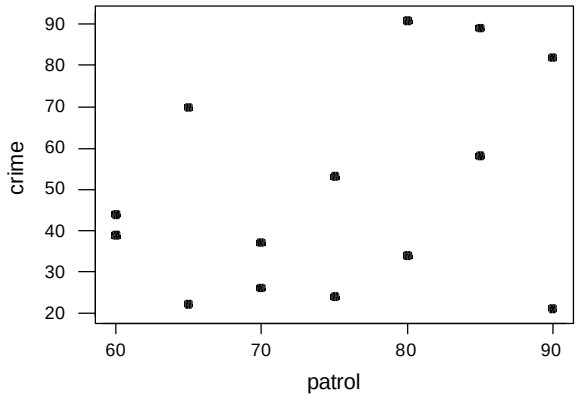
Deterrence

1. *Write a regression formula for the dependent variable. Review the assumptions underlying OLS regression analysis. What assumptions assure an unbiased estimate of the slope and intercept? What assumptions assure unbiased standard errors?*

Crime = a + b (Patrol) + u

2. Which of the OLS assumptions are likely to be violated in this instance?

It may be that more patrols go out on days that experience more crime, in which case X and u would be correlated, leading to bias. Biased standard errors may result from the fact that above-expected crime one day may produce above-expected crime the next day.



3. *Plot the data. Comment on your ocular regression.*

There is little apparent pattern to these data. One notices, if anything, a slightly positive association, which runs counter to the notion that patrols deter crime.

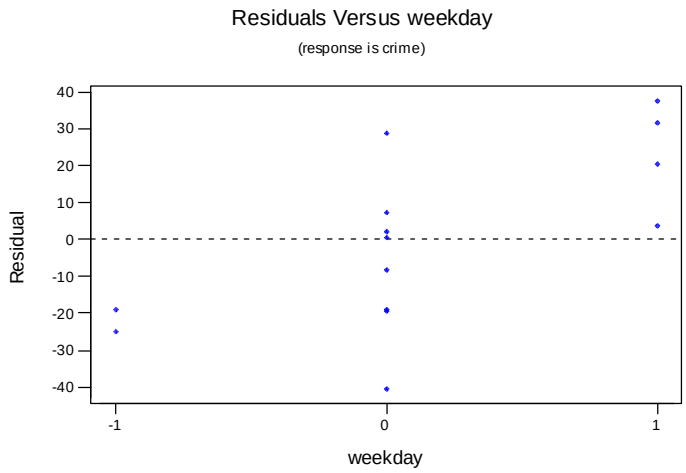
4. *Perform a regression analysis to estimate the slope, intercept, disturbance variance (also known as the error variance), standard error of the slope, and the R2. Provide a substantive interpretation of the findings.*

```
The regression equation is crime = - 12.9 + 0.829 patrol

Predictor        Coef       StDev          T        P
Constant       -12.86       49.50      -0.26    0.799
patrol         0.8286      0.6542       1.27    0.229

S = 24.48       R-Sq = 11.8%     R-Sq(adj) = 4.4%
```

The constant implies that when there are no patrols, the crime rate is negative. The slope implies that for each one-unit increase in patrols, crime is expected to increase by .83. The small t-value of 1.27, however, does not allow us to reject the null hypothesis that b=0, particularly if the test were a one-tailed test involving the alternative that b < 0! The disturbance variance is the square of s, or 599. The weak R-square indicates that this model has little predictive value.

These results do not support the hypothesis that patrols deter crime.

5. *Create a new variable, scored "1" if the day of the week is Friday or Saturday, "0" if the day of the week is Monday-Thursday, and "-1" if the day of the week is Sunday. Plot the residuals against this new variable.*



6. *Drop all observations that fall on Friday, Saturday, or Sunday. Repeat Problem #4.*

```
The regression equation is crime = 71.1 - 0.414 patrol

Predictor        Coef       StDev          T        P
Constant        71.13       40.85       1.74    0.132
patrol        -0.4139      0.5576      -0.74    0.486

S = 16.73       R-Sq = 8.4%      R-Sq(adj) = 0.0%
```

Now the slope is negative, which conforms to theoretical expectations. With so few observations, however (only 6 degrees of freedom), the slope estimate is smaller than its standard error. Thus, we cannot reject the null hypothesis that the data were generated by a true slope of zero. The model fails as a predictive device, explaining very little variance in crime rates.

7. *Based on the results in #6, predict the number of crimes that occur when the number of patrol cars is 50, 75, and 100. Find prediction intervals for each of these three cases.*

```
Predicted Values: 50, 75, 100

     Fit  StDev Fit          95.0% CI              95.0% PI
   50.44      13.87   (   16.49,   84.38)  (   -2.75,  103.62)
   40.09       6.08   (   25.22,   54.96)  (   -3.47,   83.65)
   29.74      16.43   (  -10.48,   69.97)  (  -27.65,   87.14)
```

More patrols apparently reduce crime, although point predictions in these three cases (PI) are difficult to differentiate.

8. *Create a new variable, called FRISAT, scored "1" if the day of the week is Friday or Saturday, "0" if the day of the week is Monday-Thursday or Sunday. Then create another variable, SUNDAY, scored "1" if the day of the week is Sunday; otherwise, this variable has a value of "0". Using multiple regression, regress the number of crimes on FRISAT, SUNDAY, and number of patrol cars. Compare your answers to your results for #4 and #6.*

```
The regression equation is
crime = 51.2 - 0.140 patrol + 39.0 frisat - 18.5 sunday

Predictor        Coef       StDev          T        P
Constant        51.24       37.09       1.38    0.197
patrol        -0.1395      0.5051      -0.28    0.788
frisat          39.02       11.33       3.44    0.006
sunday         -18.47       13.15      -1.40    0.190

S = 16.56       R-Sq = 66.4%     R-Sq(adj) = 56.3%
```

Patrols have a weak and statistically insignificant deterrent effect. Fridays and Saturdays have expected crime rates that are 39 units higher than rates associated with weekdays. Sundays have rates that are -18.5 units lower than rates associated with weekdays. However, we cannot reject the null that the Sunday slope t-ratio of -1.4 could have been generated by random chance had the true slope in fact been zero.

Stated more formally:

Null hypothesis: b3 = 0
Alternative: b3 < 0 (Sunday rates are lower than weekday rates)

df=14 - 4 = 10
Critical t for a 5% test = -1.81
Test statistic = -1.40; therefore accept null

The model has much more predictive accuracy than the initial regression, suggesting that a good deal of the day-to-day variability in crime rates is associated with the weekly crime cycle.

New Haven Colony

1. *What is the correlation between size of household and estate? Between size of household and First Division land? What is the mean acreage per household member appropriated in the first division? Calculate a 95% confidence interval for this mean.*

```
Correlation of household and estate = 0.468, P-Value = 0.125
Correlation of household and firstdiv = 0.701, P-Value = 0.011

Variable            N       Mean     Median     TrMean      StDev    SE Mean
divperhh           12       7.20       5.83       6.62       4.45       1.28
```

with 12-1=11 degrees of freedom we take 7.2 +/- (1.28)*(2.201)=4.38 to 10.2

2. *Score sex as a dummy variable. Score church elder-ness as a dummy variable. Using multiple regression, regress First Division acreage on household size, sex, and "elderness." Interpret the results.*

```
The regression equation is
firstdiv = 0.4 + 5.38 household + 6.8 sexnum + 10.2 eldernum

Predictor        Coef       StDev          T        P
Constant         0.36       10.95       0.03    0.975
househol        5.383       1.879       2.86    0.021
sexnum           6.76       11.01       0.61    0.556
eldernum        10.23       13.95       0.73    0.484

S = 17.04       R-Sq = 53.2%     R-Sq(adj) = 35.7%
```

sexnum = 1 if female; eldernum=1 if yes

Bigger households got more land. There is some indication that women and elders got more, too, but these coefficients fall well short of statistical significance.

3. *Add the size of the estate as a predictor (independent variable) in the previous regression. How does this change the results? What do you infer about the process by which land was divided?*

```
The regression equation is
firstdiv =0.000000 + 2.50 household -0.000000 sexnum +0.000000 eldernum
           + 0.0500 estate

Predictor        Coef       StDev          T        P
Constant   0.00000000  0.00000000          *        *
househol      2.50000     0.00000          *        *
sexnum    -0.00000000  0.00000000          *        *
eldernum   0.00000000  0.00000000          *        *
estate      0.0500000   0.0000000          *        *

S = 0           R-Sq = 100.0%    R-Sq(adj) = 100.0%
```

Land was distributed according to a rigid formula in which only household size and estate contribution mattered.

4. *Repeat the regression in #3 using instead Meadow acreage as the dependent variable. How do the results differ? Which division of land was more egalitarian and why?*

```
The regression equation is Meadow =0.000000 + 0.500 household +0.000000 sexnum
+0.000000 eldernum + 0.0500 estate

Predictor        Coef       StDev          T        P
Constant   0.00000000  0.00000000          *        *
househol     0.500000    0.000000          *        *
sexnum     0.00000000  0.00000000          *        *
eldernum   0.00000000  0.00000000          *        *
estate      0.0500000   0.0000000          *        *
S = 0           R-Sq = 100.0%    R-Sq(adj) = 100.0
```

Household size had less of an effect on Meadow land distribution; estate had the same effect. Whether the Meadow division was more or less egalitarian depends on one's definition of the term. If it is more egalitarian to distribute land by according to people rather than according to initial investment, then the First Division is more egalitarian. On the other hand, one could argue that the First Division gave undue quantities of land to households with more members and that a more equitable distribution would be the same size distribution to each household.