1. Consider the following very simple regression model:

$$Y_i = \beta_0 + \epsilon_i$$

where $\epsilon_1, \ldots, \epsilon_n \sim N(0, \sigma^2)$. For this problem, assume that σ^2 is known.

(a) Let $\widehat{\beta}_0$ be the least squares estimator. Show that $\widehat{\beta}_0 = \overline{Y}$.

(b) Show that the leverage scores are $h_{ii} = 1/n$.

(c) Find the vector of fitted values $\widehat{\mathbf{Y}} = (\widehat{Y}_1, \ldots, \widehat{Y}_n)$.

(d) Recall that the jackknife residual is

$$t_i = \frac{Y_i - \widehat{Y}_{i(-i)}}{s_i}$$

where $s_i^2 = \text{Var}(Y_i - \widehat{Y}_{i(-i)})$. (Remember that we are assuming that σ^2 is known. We do not need to estimate σ^2.) Find an explicit expression for t_i in terms of n, σ, $\overline{Y}$ and Y_i.

Hint: Note that Y_i and $\widehat{Y}_{i(-i)}$ are independent.

(e) Show that $t_i \to \infty$ as $Y_i \to \infty$.

2. Recall that in a linear regression model, h_{ii} denotes the i^{th} leverage score. Prove that $0 \leq h_{ii} \leq 1$.

Hint: Recall that $\mathbf{H}$ is idempotent and hence the diagonal elements of $\mathbf{H}$ and $\mathbf{H}^2$ are the same.

3. Get the Anscombe datasets:

```
attach(anscombe)
names(anscombe)
```

For each of the four datasets make four plots: (i) the data and the fitted line (ii) the usual residuals versus x, (iii) the jacknnife (studentized) residuals versus x and (iv) Cook's distance versus x.

For each dataset, comment on what you see in the plots. For data set four, the jackknife residual and the Cook's distance are undefined for one of the observations. (You will get an NaN from R). Explain why.

4. Download this data set:

```
http://stat.cmu.edu/~larry/health.txt
```

There are four variables. `y` is a measure of health risk. A high number means a high probability of getting a serious illness in the next year. This variable was computed using some complicated, expensive tests. We want to try to predict `y` from some other, simple, inexpensive variables. These other variables are `age`, `tri` (triglycerides) and `chol` (cholesterol). The goal is to predict `y` from `age`, `tri` and `chol`.

(i) Do a `pairs` plot. Comment on any interesting patterns.

(ii) Fit a linear model

```
lm(y ~ age + tri + chol)
```

(iii) Plot the jackknife residuals and Cook's distances against each covariate. You should notice some potential problems. Describe the plots. In particular, comment on any potential non-linear relationships and on any potential outliers and influential points.

(iv) To get a better model, let's try adding some quadratic terms. (Don't bother adding interactions.) So fit a model of the form:

$$Y = \beta_0 + \beta_1 \text{age} + \beta_2 \text{tri} + \beta_3 \text{chol} + \beta_4 \text{age}^2 + \beta_5 \text{tri}^2 + \beta_6 \text{chol}^2 + \epsilon.$$

Repeat the diagnostics and comment on what you see. At this point you will notice that you have fixed one problem (nonlinearity) but uncovered another (a potentially large, influential outlier). What data point has the largest residual and influence? Can you suggest an explanation for this?

(v) Remove the problematic datapoint. Refit the quadratic model. Confirm that the residuals and influence diagnostics are improved. Summarize the fitted model. In particular, give 95 percent confidence intervals for all the parameters (except the intercept). Make sure you do a Bonferroni correction. (In other words, replace $1 - \alpha$ with $1 - \alpha/m$ where m is the number of confidence intervals you are computing.)

5. Download these secret data:

```
d = read.table("http://stat.cmu.edu/~larry/secretdata.txt")
y =  d[,1]
x1 = d[,2]
x2 = d[,3]
x3 = d[,4]
x4 = d[,5]
```

(a) Do a `pairs` plot.

(b) Fit a linear mosel as usual:

```
out = lm(y ~ x1 + x2 + x3 + x4)
```

Summarize the fitted model.

(c) Plot the residuals versus each covariate.

(d) Plot the fitted values versus the residuals:

```
plot(fitted(out),resid(out))
```

See anything interesting?