MTH 4230 Lab 5

Due Wed., Mar. 4

1 Part A: Extra Sums of Squares, Partial F Tests

1.1 Patient Satisfaction Data Set (Cont'd from Lab 4)

A hospital administrator wished to study the relation between patient **satisfaction** ($\boldsymbol{Y}$) and patient's **age** ($\boldsymbol{X_1}$, in years), **severity** of illness ($\boldsymbol{X_2}$, an index), and **anxiety** ($\boldsymbol{X_3}$, an index). The administrator randomly selected 46 patients and collected the data presented in the file **satisfaction.txt**. This is the **Patient Satisfaction** data set from **Problem 6.15** of the textbook.

1. Read the data into R using `read.table()`.
2. Use `lm()` to fit the **multiple regression model** to the data, with **all three** predictors included in the model.
3. Use `anova()` to obtain the ANOVA table that decomposes the regression sum of squares into the ***extra sums squares*** $\mathbf{SSR}(\boldsymbol{X_2})$, $\mathbf{SSR}(\boldsymbol{X_1}|\boldsymbol{X_2})$, and $\mathbf{SSR}(\boldsymbol{X_3}|\boldsymbol{X_2}, \boldsymbol{X_1})$, for example by typing:

   ```r
   anova(my.reg)
   ```

   (where `my.reg` is the `"lm"` object from Step 2).
4. Test whether X_3 can be dropped from the regression model, given that X_1 and X_2 are retained. Use the ***partial F test*** (which, recall, is equivalent to the *general linear F test*).
5. Show that the F test just performed and the ***t test*** for $\boldsymbol{\beta_3}$ are equivalent (i.e. $t^2 = F$ and the p-values are the same).
6. Recall that the ***coefficient of partial determination*** is

   $$R^2_{X_k|X_1,\ldots,X_{k-1},X_{k+1},\ldots,X_{p-1}} = \frac{\mathrm{SSR}(X_k|X_1,\ldots,X_{k-1},X_{k+1},\ldots,X_{p-1})}{\mathrm{SSE}(X_1,\ldots,X_{k-1},X_{k+1},\ldots,X_{p-1})}.$$

   It measures the **reduction** in **unexplained** $\boldsymbol{Y}$ **variation** (as a proportion) that results from **adding** $\boldsymbol{X_k}$ to a model that **already includes all** the **other predictors**.

   Compute $\boldsymbol{R^2_{X_3|X_1,X_2}} = \mathbf{SSR}(\boldsymbol{X_3}|\boldsymbol{X_1}, \boldsymbol{X_2})/\mathbf{SSE}(\boldsymbol{X_1}, \boldsymbol{X_2})$.

2 Part B: Polynomial Regression

2.1 Nigeria Household Refuse Data Set

In a study of the environmental impact of the increase in solid waste resulting from rapid urban population growth in the Port Harcort area of Nigeria, the **size** (number of residents) and annual **refuse** generation (in metric tons) was determined for each household in a sample of $\boldsymbol{n = 46}$ households in the area. The data are in the file **NigeriaRefuse.txt**.

1. Read the data into *data frame* using `read.table()`.

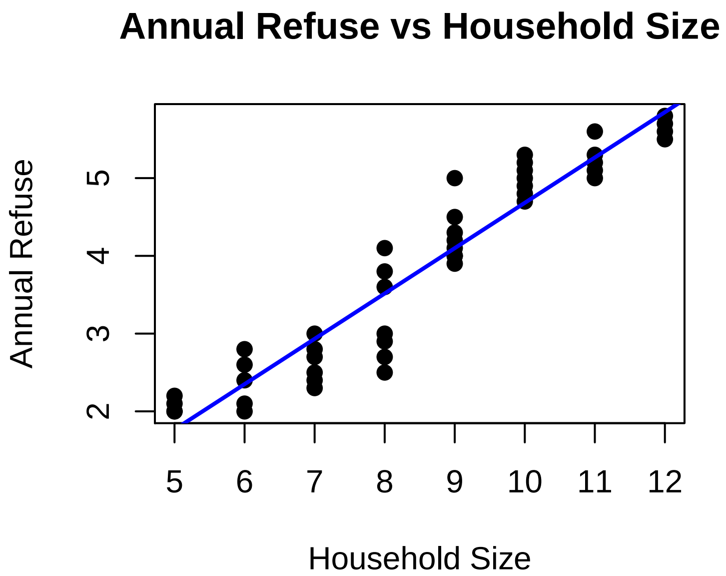
2. Use `plot()` to make a scatterplot of the data, with household **size** on the x axis and annual **refuse** on the y axis, for example by typing:

```
plot(my.data$size, my.data$refuse, pch = 19, xlab = "Household Size",
    ylab = "Annual Refuse", main = "Annual Refuse vs Household Size")
```

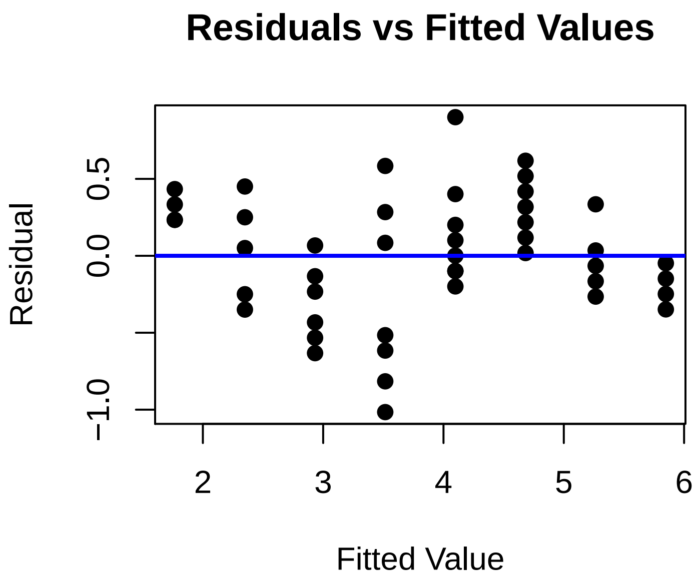
3. Use `lm()` to fit a ***simple linear regression model*** to the data, with household **size** as the predictor and annual **refuse** as the response.

4. Use `abline()` to add the regression line to the plot of Step 2 by typing something like:

```
abline(my.reg, col = "blue", lwd = 2)
```



5. Use `plot()` to make a plot of the residuals (y axis) versus the fitted values (x axis). Add a horizontal line at $y = 0$ by typing `abline(h = 0)`. Your plot should look like this:



6. Notice there's a nonlinear pattern in the scatterplot of Step 2, which leads to the pattern in residual plot of Step 5.

   We want to know if a ***polynomial regression model*** will fit the data substantially better.

   Add three columns to your *data frame*, one containing the **squares**, another the **cubes**, and another the **quartics** (4th powers) of the household **sizes** by typing something like:

```
my.data$size2 <- my.data$size^2
my.data$size3 <- my.data$size^3
my.data$size4 <- my.data$size^4
```

   Now check:

```
head(my.data)
```

7. Fit a ***4th order polynomial regression model***

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \epsilon$$

   to the data, where Y = annual **refuse** and X = household **size**. Fit the model by typing something like:

```
my.reg <- lm(refuse ~ size + size2 + size3 + size4, data = my.data)
```

8. Use `summary()` to look at the results.
9. We want to know if any of the higher order terms (e.g. X^4, X^3, or X^2) can be dropped from the model. Use `anova()` to obtain the ANOVA table that decomposes the variation in annual refuse into ***extra sums squares***:
   - SSR(X)
   - SSR($X^2|X$)
   - SSR($X^3|X, X^2$)
   - SSR($X^4|X, X^2, X^3$)

   and carries out the associated ***partial F tests***.
10. Now use `lm()` to fit the ***3rd order polynomial regression model***

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$$

   to the data.
11. Look at the results using `summary()`.
12. Using the result of Step 11, add the fitted ***3rd order polynomial***

$$\hat{Y} = b_0 + b_1 X + b_2 X^2 + b_3 X^3$$

   to the scatterplot of Step 5 by typing:

```
curve(expr = 12.95114 - 4.76726*x + 0.64641*x^2 - 0.02502*x^3,
      from = 5, to = 12, col = "blue", add = TRUE)
```

   You should end up with something like this: