1. Consider an example where $z$ and $y$ denote pretest and posttest mental capacity scores for subjects randomized to one of three treatments.

<table>
<thead>
<tr>
<th>Trt 1</th>
<th>Trt 2</th>
<th>Trt 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$z$</td>
<td>$y$</td>
<td>$z$</td>
</tr>
<tr>
<td>24</td>
<td>45</td>
<td>23</td>
</tr>
<tr>
<td>28</td>
<td>50</td>
<td>33</td>
</tr>
<tr>
<td>38</td>
<td>59</td>
<td>31</td>
</tr>
<tr>
<td>42</td>
<td>60</td>
<td>34</td>
</tr>
<tr>
<td>24</td>
<td>47</td>
<td>18</td>
</tr>
<tr>
<td>39</td>
<td>66</td>
<td>24</td>
</tr>
<tr>
<td>45</td>
<td>76</td>
<td>41</td>
</tr>
<tr>
<td>19</td>
<td>50</td>
<td>34</td>
</tr>
<tr>
<td>22</td>
<td>36</td>
<td>39</td>
</tr>
</tbody>
</table>

$z_1 = 30 \quad \hat{y}_1 = ? \quad \hat{z}_2 = 30.7 \quad \hat{y}_2 = ? \quad \hat{z}_3 = 31.9 \quad \hat{y}_3 = ?$

(a) To read the data in, we'll look at a few of features of the DATA step:

- INFILE statement with FIRSTOBS option
- the single trailing @ sign
- do loops

```
DATA raoeg;
  INFILE "raoeg12.1.dat" FIRSTOBS=4;
  DO trt=1 to 3;
    INPUT z y @;
    OUTPUT;
    *PUT trt z y;
  END;
RUN;
PROC PRINT DATA=raoeg;RUN;
```
Because of the @ statement, the INPUT statement does NOT look to the next line of the datafile for the next values of z and y. Instead, it holds the record until SAS either

- executes an INPUT statement with no trailing @ or
- reaches the bottom of the datastep (after trt = 3 loop finishes in this case)

Now that INPUT z y @; has been executed, OUTPUT; writes the variables to the dataset raoeg. Then the END; statement iterates to trt = 2 and the step inside the DO loop is repeated twice. After the trt = 3 step, SAS exits the DO loop and exits the datastep.

You can always use PUT statements to see what is going on during the datastep iterations and PROC PRINT; to see the finished datasets.

(b) Compute mean pretest scores (z) and post test scores (y) using PROC MEANS;

\[ \text{PROC MEANS MEAN STD;} \]
\[ \quad \text{CLASS trt;} \]
\[ \quad \text{VAR y z;} \]
\[ \text{RUN;} \]

Carry out a one-way ANOVA to test whether or not the post-test population means \( \mu_1, \mu_2, \mu_3 \) are plausibly equal in light of these data. That is, are \( \bar{y_i} + 1, \bar{y_i} + 2, \bar{y_i} + 3 \) significantly different? Use the code below.

\[ \text{PROC GLM;} \]
\[ \quad \text{CLASS trt;} \]
\[ \quad \text{MODEL y=trt;} \]
\[ \text{RUN;} \]

To see the treatment means, add a MEANS trt; statement after the MODEL statement.
The $F$-ratio for the test of equality of means should be $F = 6.7$ with a $p$-value of 0.0043. Find these in the output.

(c) Add the SOLUTION option to the MODEL statement:

```
MODEL y=trt/SOLUTION;
```

This will adopt the default parameterization for treatment effects discussed in class (the antibiotic example.) Note that $\hat{\beta}_0$ is $\bar{y}_3$.

(d) Note the $MS[E]$ term. What is the estimated standard deviation of posttest scores receiving a given treatment?

(e) Write the residuals from the one-way ANOVA to a variable called “e” in a new dataset called “raoeg2” with an OUTPUT statement:

```
OUTPUT OUT=raoeg2 R=e;
```

(f) Investigate any linear (or other) association between these residuals and the pretest scores $z$ using PROC REG:

```
PROC REG DATA=raoeg2 SIMPLE;
  MODEL e=z;
  PLOT p.*z;
RUN;
```

Note the strong association (see the $F$-test for the $\beta = 0$ hypothesis.) Note the proportion of variation in the residuals that could be explained using the pretest score $z$, $r^2$. Look at the plot and make note of the estimated slope, $\hat{\beta}$. Also, make note of the overall average pretest score $\bar{z}$. It will subsequently serve as a point of reference for comparing treatment effects.

(g) Using our general linear model framework with two indicator variables $x_1$ and $x_2$ for the three treatment groups, fit a model using the following regression function for the mean posttest score:

$$\mu(x_1, x_2, z) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta z$$

You can use PROC GLM; to do this:

```
PROC GLM DATA=raoeg2;
  CLASS trt;
  MODEL y=trt z;
RUN;
```

Estimate the standard deviation for posttest scores for a given treatment and fixed pretest score $z = z_0$. Note the
reduction in $MS[E]$ from the model that did not include $z$ (in part(d)). That is, note the reduction in unexplained error, or the increase in explained “Model” variance.

(h) Add the **SOLUTION** option to the **MODEL** statement:

```plaintext
MODEL y=trt z/SOLUTION;
```

Observe the changes in the estimated coefficients, $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$. These now account for the pretest score. Note $\hat{\beta}$, the estimated slope for the linear dependence of $y$ on $z$ is the same as it was when the residuals $e$ from the one-way ANOVA were regressed on $z$. (The standard error changes because the degrees of freedom in the regression aren’t quite correct. The standard errors from the full model are correct.)

(i) To obtain a nice plot of the full fitted model, add an **OUTPUT** statement and a **PROC GPLOT**:

```plaintext
PROC GLM DATA=raoeg;
   CLASS trt;
   MODEL y=trt z;
   OUTPUT OUT=raoeg3 P=fitted;
RUN;

PROC GPLOT;
   PLOT fitted*z=trt;
RUN;
```

(j) To compute adjusted means, an **LSMEANS** statement can be used after the **MODEL** statement. Before running this, reason out the proper direction for the inequalities below. The direction in which a mean $\bar{y}_{i+1}$ is adjusted depends on where its covariate mean $\bar{z}_i$ is relative to the reference point $\bar{z}$ and the sign of the estimated partial slope for $z$, $\hat{\beta}$.

\[
\begin{align*}
\bar{y}_{1,adj} &< \bar{y}_{i+1} \\
\bar{y}_{2,adj} &< \bar{y}_{i+2} \\
\bar{y}_{3,adj} &< \bar{y}_{i+3}
\end{align*}
\]

Use the **STDERR** and **CL** options to report standard errors and confidence intervals for adjusted means:

```plaintext
LSMEANS trt/STDERR CL;
```

Check to see that your assessments of the inequalities were right.