GEOG 400/500 Project 2 **Name:**

First step: Simple linear regressions. Again.

Put an "X" next to your hypothesized direction of association between each Xi and Y (direct or inverse):

Y and X1	Direct	Inverse	%cen cty 99 → Gunmurd97%
Y and X2	Direct	Inverse	%gbn Bush→ Gunmurd97%
Y and X3	Direct	Inverse	2K per cap → Gunmurd97%
Y and X4	Direct	Inverse	%>24:BA → Gunmurd97%
Y and X5	Direct	Inverse	NRA carry99 → Gunmurd97%
Y and X6	Direct	Inverse	% yngmen → Gunmurd97%

Alpha you plan to use to test the null hypotheses, "there is no association between Xi & Y"

Alpha =

Justification for the *alpha* you picked:

Which associations turned out significant at your chosen *alpha* level? (mark with an "X")
And which of the **significant** associations had the same direction you predicted (direct or inverse)? Mark with an "X."

	Calculated P-Value 4 decimal places	Is P-Value < *alpha?* Y or N	Do significant associations match your prediction? Y or N?	Your predic-tion I or D?
Y and X1	Prob-value	Significant?	Prediction correct?	

Y and X2	Prob-value	Significant?	Prediction correct?
Y and X3	Prob-value	Significant?	Prediction correct?
Y and X4	Prob-value	Significant?	Prediction correct?
Y and X5	Prob-value	Significant?	Prediction correct?
Y and X6	Prob-value	Significant?	Prediction correct?

Interpretation/speculations:

Kitchen sink multiple regression model → Y variable is

Name the two variables in your **best** simple linear regression model above: → Best X_i

Below, compare the multiple regression model with the best simple linear model (3 decimal places works here):

Multiple regression R	*vs.*	Best bivariate model's R
Multiple regression R^2_{adj}	*vs.*	Best bivariate model's R^2

By comparing movement in R^2/R^2_{adj}, did you produce a **much** better explanation of Y with 6 X_is? Y or N

At 4 decimal places of accuracy, do you see a noticeable improvement in the significance of the model?

Whether it was worth your bother or not, please write down your model, showing a and b coëfficients at **4** decimal places:
Don't forget the signs of the b coëfficients and placing the X_1, X_2, ...X_6 after the coëfficients and before the sign of the next term.

Y =

Looking at the t-scores and prob-values for each variable in the kitchen sink model and comparing them with their corresponding values in the simple linear regressions you did earlier, which (if any) have significant p-values once they're allowed to interact?

X_1	X_2	X_3	X_4	X_5	X_6

Put an "X" below the variable(s).

Why is it that several of the variables are significant considered alone (bivariate simple linear regressions) but drop out of significance when they are all put together in a common multiple regression model?

Prune the model of all X_i variables that have p-values larger than your pre-selected *alpha* standard. Names of rejected variables:

Refining the multiple regression model through backwards elimination:

Rerun the multiple regression, but ONLY with the X_i variables that still have p-values smaller than *alpha* in the multiple regression

While R^2 can be expected to decline, the key diagnostics are the changes in significance and in the F statistic that defines it.

Did F:

Increase | Decrease | Stay roughly the same-ish

Write down your new model (4 decimal places, with proper signs and the original X_i numbers: e.g., $Y= 0.0571 + 0.3452X_2 - 0.3671X_6$

Y =

Now, re-examine the new t-scores and p-values. The new regression will have altered them from either the simple linear regression or the kitchen sink everything-in-it regression. Did the interactions among variables in the new model cause any of your X_i variables to generate new p-values **higher** than your *alpha* standard (meaning you throw them out in any further round)?

Yes | No | Which?

Dump any X_i variable with effects that are no longer significant in comparison with your *alpha* standard. Now, rerun the regression.

Write down your newest model:

Y =

What happened to the new(est) R^2_{adj}?

Increased	Decreased	Stayed roughly the same

Dealing with an outlier

After doing your six scatterplots of each X_i on Y, identify the outlier record:

Redo the kitchen-sink multiple regression (all X_i variables at once) but with the outlier removed.

Write down this newest kitchen-sink masterpiece, leaving X_i in their original order, coëfficients at 4 decimal places, proper signs:

Y =

As before, refine through backwards elimination. Throw out all X_i variables with p-values above your *alpha* standard.

Write down the newest model.

Y =

Which of the two outlier-free models (kitchen-sink or the backwards pruned one) has the best F score/significance value?

Interpretation of the performance of your models, both with and without the outlier. Use sheet below, if necessary.

Overflow if you need more space: