Andrew J. Spieker, PhD
BIOS 6312 - Modern Regression Analysis
Spring 2020
Exam #1

Instructions: Please adhere to the following guidelines:

- This exam is an individual effort (closed book/notes/phone/everything), although you are permitted the use of a scientific calculator.
- Please read the questions carefully and answer only what you are being asked to answer.
- Please be concise. You are asked many times throughout the exam to state something in no more than a sentence or two, meaning you should be able to provide the answer in the amount of space provided.
- See me if you need scratch paper.
- Do not spend too much time on one problem. Mind the point distribution (below).
- Please write legibly.
- Please round any final calculations to **three** significant digits!
- There are ten numbered pages (seven required problems plus two optional problems). There are five pages of appendix material. Check to make sure you have everything before beginning.

#	Score	Points
1		15
2		10
3		10
4		10
5		30
6		10
7		15
Total:		100
Opt. 1		
Opt. 2		

1. A group of co-investigators is using data from a cross-sectional study ($N = 46$) to obtain insights into the association between age and the prostate-specific antigen (PSA, in ng/ml) among an adult male population. Two of the investigators on the team, Michael and Rebecca, take it upon themselves to each perform an analysis of these data using simple linear regression in Stata (Appendix I). When presenting results at the team meeting, they find that their approaches differ in that Rebecca log-transformed PSA and altered the units of age from "years" to "decades" by dividing the `age` variable by ten.

   (a) Circle the letter corresponding to the statement with which you most closely agree:

      A. We should use Rebecca's analysis if we have evidence that PSA is right-skewed.
      B. Since linear regression requires normal errors, we should look at a quantile-quantile plot of the studentized residuals from Rebecca's and Michael's analyses before deciding which to use.
      C. Robust standard errors serve no purpose in Rebecca's analysis because log-transformation always eliminates heteroscedasticity.
      D. A good reason to use the results of Michael's analysis would be if the investigators pre-specified that they would examine the association between age and *mean* PSA.
      E. Since Rebecca's analysis does *not* achieve statistical significance (see Appendix I), we should instead publish Michael's analysis and forget about Rebecca's altogether.

   (b) Use the Stata output of one of the analyses (provided in Appendix I) to briefly summarize this study's evidence regarding an association between age and *geometric* mean PSA. In your response, you need only include (with proper interpretation) a point estimate and its 95% confidence interval, as well as a summary of your conclusion with appropriate inferential measures.

   (c) Devon uses the regression output from Michael's analysis to form a naive 95% prediction interval for PSA among 30 year-old men. After examining the (studentized) residual-versus-predictor plot provided in Appendix II, state the *two* most obvious reasons why this is an extraordinarily bad idea on Devon's part. Do *not* attempt to form the prediction interval.

2. A cross-sectional observational study was conducted to determine the association between number of hours of sleep the night before an exam (X, in hours) and the subsequent exam score (Y, %). $N = 500$ students were randomly sampled from Vanderbilt University and were surveyed regarding their sleep history and most recent exam score. Once the data were collected, the following simple linear regression model was fit using ordinary least squares:

$$\mathbf{E}[Y|X = x] = \beta_0 + \beta_1 x.$$

Your response to sub-questions (a) through (d) below should each be **at most two sentences**.

(a) Based on the scatterplot provided in Appendix III, are you concerned about any violations to the assumption of linearity for the purposes of estimating β_1?

(b) Briefly describe how the scatterplot provided in Appendix III demonstrates graphical evidence of non-normality of the errors.

(c) You find out that one-hundred students in your sample were roommates with another student in your sample. Thinking in terms of whether the regression assumptions are met in this setting, state how this limits your ability to conduct inference on β_1 when using ordinary least squares.

(d) Suppose you had instead fit a no-intercept model, $\mathbf{E}[Y|X = x] = \beta x$. State the additional assumption that would be made by this restriction and briefly comment on whether it is reasonable.

3. We have considered the REACH study for numerous examples in our notes and on homework problems. The REACH study was a randomized controlled trial that sought to evaluate the effect of a text-message intervention on hemoglobin A1c. A subset of the variables from this study include:

$$X = \begin{cases} 0 & \text{if assigned to control group} \\ 1 & \text{if assigned to REACH group} \end{cases},$$

Y = A1c at twelve months (%), and

Z = A1c at baseline (%).

Consider the following two linear regression models:

$$\begin{aligned} \mathbf{E}[Y|X = x] &= \beta_0^* + \beta_1^* x, \text{ and} \\ \mathbf{E}[Y|X = X, Z = z] &= \beta_0 + \beta_1 x + \beta_2 z \end{aligned}$$

For this problem, you may assume that the trial is well conducted so that there is *no* systematic confounding, that each model is correctly specified, and that Z and Y are linearly correlated.

(a) Which of the following four statements is true?

I. $\beta_1^* = \beta_1$
II. $\beta_1^* > \beta_1$
III. $\beta_1^* < \beta_1$
IV. It is impossible to tell without more information.

Answer: _____

(b) You employ ordinary least squares (OLS) linear regression to estimate each model. Which of the following four statements is true?

I. $\widehat{\beta}_1^* = \widehat{\beta}_1$
II. $\widehat{\beta}_1^* > \widehat{\beta}_1$
III. $\widehat{\beta}_1^* < \widehat{\beta}_1$
IV. It is impossible to tell without more information.

Answer: _____

(c) Again based on ordinary least squares estimation, which of the following four statements is true?

I. $\text{Var}(\widehat{\beta}_1^*) = \text{Var}(\widehat{\beta}_1)$
II. $\text{Var}(\widehat{\beta}_1^*) > \text{Var}(\widehat{\beta}_1)$
III. $\text{Var}(\widehat{\beta}_1^*) < \text{Var}(\widehat{\beta}_1)$
IV. It is impossible to tell without more information.

Answer: _____

(d) In at most two sentences, state the purpose of adjusting for baseline A1c (i.e., the second model).

4. We have considered a study of forced expiratory volume (FEV) in children for numerous examples in our notes and on homework problems. The FEV study was a cross-sectional observational study that sought to evaluate whether smoking was associated with FEV, a measure of lung capacity in which higher values signify better lung function. A subset of the variables from this study include:

$$X = \begin{cases} 0 & \text{if a non-smoker} \\ 1 & \text{if a smoker} \end{cases},$$

$$Y = \text{FEV (L/sec), and}$$

$$Z = \text{Age (years).}$$

Consider the following two linear regression models:

$$\begin{aligned} \mathbf{E}[Y|X = x] &= \beta_0^* + \beta_1^* x, \text{ and} \\ \mathbf{E}[Y|X = X, Z = z] &= \beta_0 + \beta_1 x + \beta_2 z. \end{aligned}$$

You may assume for this problem that both models are correctly specified.

(a) Which of the following is true regarding ordinary least squares estimation of β_1^* (i.e., from the unadjusted model), assuming homoscedasticity?

I. It is exactly the same as a t-test assuming equal variances between smoking groups.
II. It is exactly the same as a t-test allowing unequal variances between smoking groups.
III. It is approximately the same as a t-test assuming equal variances between smoking groups.
IV. It is approximately the same as a t-test allowing unequal variances between smoking groups.

Answer: _____

(b) Which of the following is true regarding ordinary least squares estimation of β_1^* (i.e., from the unadjusted model), allowing heteroscedasticity?

I. It is exactly the same as a t-test assuming equal variances between smoking groups.
II. It is exactly the same as a t-test allowing unequal variances between smoking groups.
III. It is approximately the same as a t-test assuming equal variances between smoking groups.
IV. It is approximately the same as a t-test allowing unequal variances between smoking groups.

Answer: _____

(c) Again based on ordinary least squares estimation, which of the following four statements is true?

I. $\text{Var}(\widehat{\beta}_1^*) = \text{Var}(\widehat{\beta}_1)$
II. $\text{Var}(\widehat{\beta}_1^*) > \text{Var}(\widehat{\beta}_1)$
III. $\text{Var}(\widehat{\beta}_1^*) < \text{Var}(\widehat{\beta}_1)$
IV. It is impossible to tell without more information.

Answer: _____

(d) In at most two sentences, state the purpose of adjusting for age (i.e., the second model).

5. This problem pertains to interpretation of regression parameters, particularly when dealing with common variable transformations. For each sub-question, there is a true-or-false question included pertaining to the model posed in that sub-question. Please circle the letter corresponding to your answer (no follow-up explanation is required). Then, you are asked to use plain language to interpret one of the parameters in the model, which means that you are *not* to include mathematical/statistical notation (e.g., $\mathbf{E}[Y|X = x]$); however, you are permitted to use notation to refer to variables (e.g., "X"), and you are permitted to use numbers/percentages (e.g., "10%").

   (a) Consider the following regression model:

   $$\mathbf{E}[Y|X = x] = \beta_0 + \beta_1 x.$$

   T F When X increases by one unit, Y increases by β_1 units.

   Provide a plain-language interpretation for $3\beta_1$.

   (b) Consider the following regression model:

   $$\mathbf{E}[\log(Y)|X = x] = \beta_0 + \beta_1 \log(x).$$

   T F If Y denotes a measure of concentration, it absolutely *must* be log-transformed.

   Provide a plain-language interpretation for $e^{\log(1.2)\beta_1}$.

   (c) Consider the following regression model (assume X is categorical with levels 0, 1, and 2):

   $$\mathbf{E}[Y|X = x] = \beta_0 + \beta_1 1(x = 1) + \beta_2 1(x = 2)$$

   T F To determine if X and (mean) Y are associated, we could test $H_0 : \beta_1 = \beta_2 = 0$.

   Provide a plain-language interpretation for β_1.

(d) Consider the following regression model:

$$\mathbf{E}[Y|X = x] = \beta_0 + \beta_1(x - 10) + \beta_2(x - 10)^2$$

T F To determine if X and Y are associated, it is sufficient to test $H_0 : \beta_1 = 0$.

Provide a plain-language interpretation for β_1.

(e) Consider the following regression model:

$$\mathbf{E}[Y|X = x, Z = z] = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz.$$

T F β_0 denotes the mean of Y among the subgroup with $X = Z = 0$.

Provide a plain-language interpretation for β_3.

(f) Consider the following regression model:

$$\mathbf{E}[Y|X = x, Z = z, W = w] = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 w + \beta_4 xz + \beta_5 xw + \beta_6 zw + \beta_7 xzw$$

T F If the three-way interaction between X, Y, and Z is of most clinical interest, we need only estimate β from the reduced model $\mathbf{E}[Y|X = x, Z = z, W = w] = \beta xzw$.

Provide a plain-language interpretation for β_4.

6. A study was conducted to evaluate potential predictors of systolic blood pressure, (SBP, in mmHg, denoted by Y). As a simple example, consider alcohol consumption, X_1 (oz/week) and race category, X_2, as predictors. The race variable in this example is coded as follows:

$$X_2 = \begin{cases} 1 & \text{White} \\ 2 & \text{Black} \\ 3 & \text{Asian} \\ 4 & \text{Other} \end{cases}$$

The following regression model is fit using Stata:

$$\mathbf{E}[Y|X_1 = x_1, X_2 = x_2] = \beta_0 + \beta_1 x_1 + \beta_2 1(x_2 = 2) + \beta_3 1(x_2 = 3) + \beta_4 1(x_2 = 4).$$

For this problem, refer to the Stata output in Appendix IV, which includes the regression output and an F-test of the hypothesis $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$.

(a) Report the mean squared error, with an interpretation that holds regardless of heteroscedasticity.

(b) In a sentence, either state your conclusions on whether there is evidence that mean SBP differs between black and Asian individuals (adjusted for alcohol consumption) ***or*** state that this cannot be done on the basis of the output and suggest a modification to the model that would permit this.

(c) In a sentence, summarize your conclusions on whether there is evidence of an overall association between alcohol consumption and mean SBP, adjusted for race.

(d) In a sentence, summarize your conclusions on whether there is evidence of an overall association between race and mean SBP, adjusted for alcohol consumption.

7. This is a continuation of Problem 6. The investigators decided that they would like to allow for an interaction between alcohol consumption (X_1) and race category (X_2). For this problem, refer to the Stata output in Appendix V. It may help you to know that $Z_{0.975} = 1.96$.

   (a) Form a naive 95% prediction interval for SBP among white individuals with no history of alcohol consumption and briefly state the assumptions required for the prediction interval to be approximately valid — you are *not* expected to evaluate the extent to which they hold.

   (b) Examine the F-test results from the `testparm` command. In a sentence, and in plain language, what is the hypothesis being tested by this command?

   (c) Note from the output of the `testparm` command the that the critical value is $F(4, 727)$. Explain briefly where the numbers 4 and 727 come from.

   (d) Glen states that because each of the the p-values corresponding to the three coefficients reported under `race#c.alcoh` is larger than 0.05, there is not sufficient evidence of an interaction between race and alcohol consumption. Is this argument valid? If not, explain in no more than two sentences what test you would prefer to conduct to evaluate evidence of such an interaction.

8. ***Optional problem 1***: This is an optional problem — do not attempt it until you have completed the rest of the exam. A small bonus can be earned from a correct response.

   Consider a randomized trial with two primary treatment groups, as denoted by X:

$$X = \begin{cases} 0 \text{ control} \\ 1 \text{ experimental treatment} \end{cases}$$

   Further suppose that a subset of individuals receiving the experimental treatment (i.e., those with $X = 1$) also receive an *additional* treatment, denoted by Z:

$$Z = \begin{cases} 0 \text{ no additional treatment} \\ 1 \text{ additional treatment} \end{cases}$$

   The investigators are interested in evaluating whether the supplemental treatment, Z, has an additional effect on mean Y beyond that of X alone. Without thinking it all the way through, a statistician on the study fits the following model:

$$\mathbf{E}[Y|X = x, Z = z] = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz.$$

   In a couple of sentences, explain what will go wrong here and why. Then, briefly propose one or two alternative ways of answering the clinical question at hand. If you have time to propose two alternatives, briefly characterize the relative advantages and disadvantages of each of those two approaches.

9. ***Optional problem 2***: This is an optional problem — do not attempt it until you have completed the rest of the exam. A small bonus can be earned from a correct response.

   Let $\mathbf{X}$ denote an $N \times (p+1)$ design matrix of full rank. Recall that the hat matrix, $\mathbf{H}$, is defined as:

   $$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T.$$

   Prove that $\mathbf{H}$ is an orthogonal projection matrix by first proving that it is symmetric (i.e., $\mathbf{H}^T = \mathbf{H}$) and then proving it is idempotent (i.e., that $\mathbf{H}^2 = \mathbf{H}$).

   Since the ordinary least squares estimate is given by $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, this in essence proves that $\widehat{\boldsymbol{\beta}}$ is estimated in a way such that $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{H}\mathbf{y}$ is the orthogonal projection of $\mathbf{y}$ onto the linear subspace spanned by the columns of the design matrix $\mathbf{X}$.

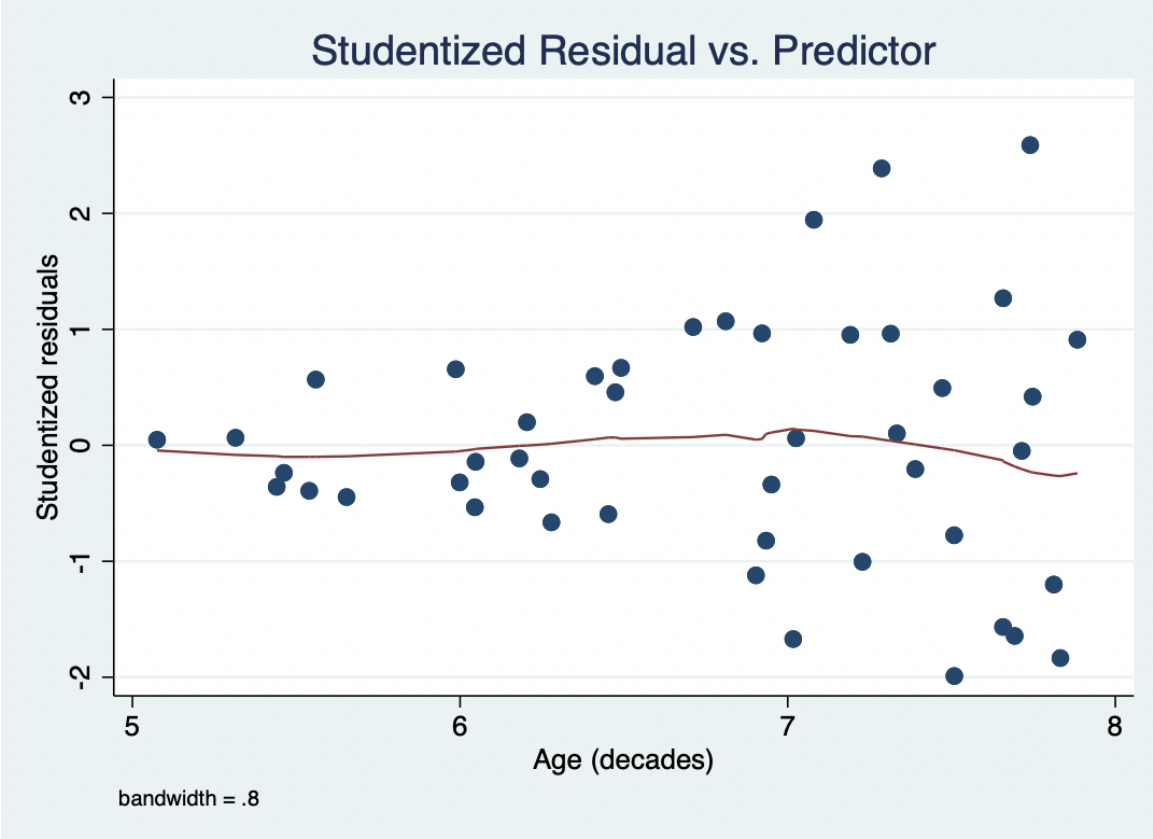
Andrew J. Spieker, PhD
BIOS 6312 - Modern Regression Analysis
Spring 2020
Exam #1

Appendix Material for Exam 1

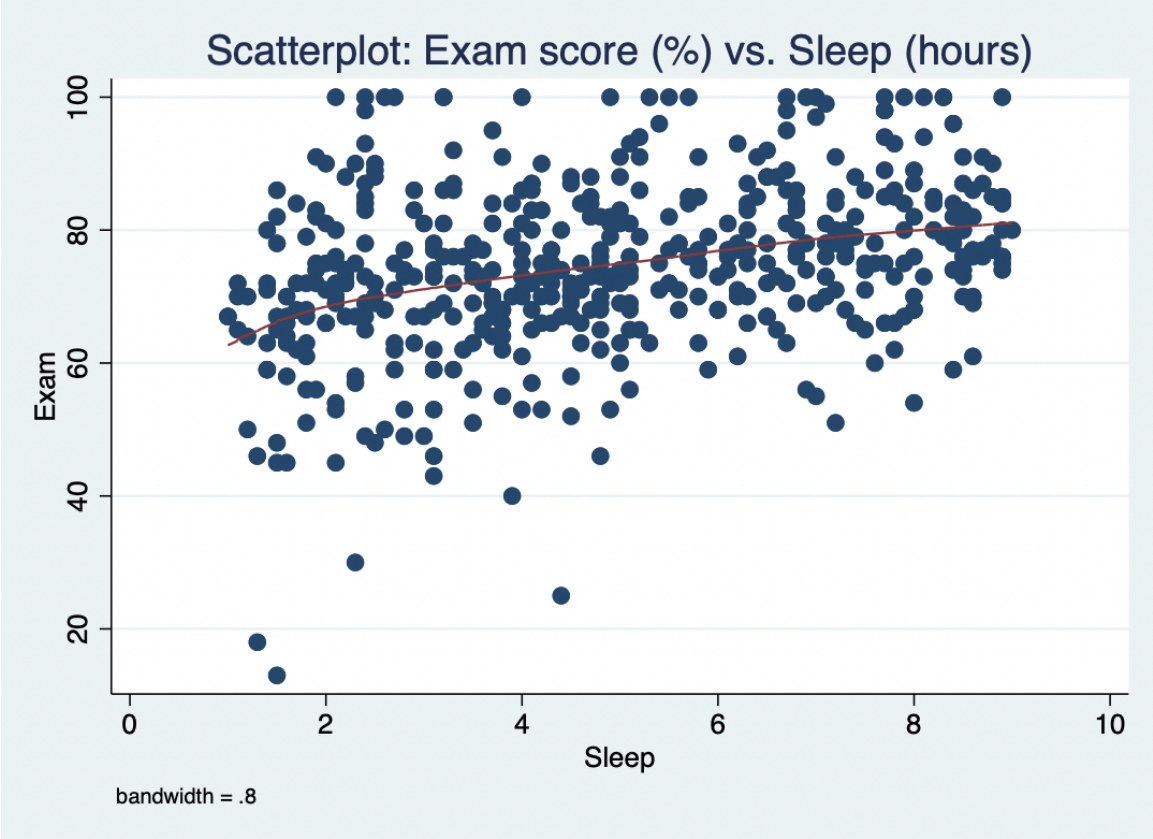
APPENDIX I: Stata output for Problem 1

```
* MICHAEL'S ANALYSIS

. regress psa age, robust

Linear regression                               Number of obs     =         46
                                                F(1, 44)          =       7.93
                                                Prob > F          =     0.0072
                                                R-squared         =     0.1332
                                                Root MSE          =     1.6947

------------------------------------------------------------------------------
             |               Robust
         psa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0821603    .029172     2.82   0.007     .0233679    .1409526
       _cons |  -2.272699   1.795634    -1.27   0.212    -5.891561    1.346162
------------------------------------------------------------------------------



* REBECCA'S ANALYSIS

. gen logpsa = log(psa)
. gen age_decades = age/10

. regress logpsa age_decades, robust

Linear regression                               Number of obs     =         46
                                                F(1, 44)          =       3.84
                                                Prob > F          =     0.0565
                                                R-squared         =     0.0662
                                                Root MSE          =     .56421

------------------------------------------------------------------------------
             |               Robust
      logpsa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 age_decades |    .185762   .0948398     1.96   0.057     -.005375     .376899
       _cons |  -.2202967   .5902497    -0.37   0.711    -1.409867    .9692734
------------------------------------------------------------------------------
```

APPENDIX II: Diagnostic plot for Problem 1



APPENDIX III: Scatterplot for Problem 2



APPENDIX IV: Stata output for Problem 6

```
. regress sbp alcoh i.race, robust

Linear regression                               Number of obs     =        735
                                                F(4, 730)         =       2.86
                                                Prob > F          =     0.0228
                                                R-squared         =     0.0162
                                                Root MSE          =     19.556

------------------------------------------------------------------------------
             |               Robust
         sbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       alcoh |    .365078   .1571445     2.32   0.020     .0565689     .673587
             |
        race |
          2  |   4.508028   2.175304     2.07   0.039     .2374294    8.778626
          3  |  -.2145398   2.962762    -0.07   0.942     -6.03109     5.60201
          4  |   6.091013   6.691777     0.91   0.363    -7.046412    19.22844
             |
       _cons |   129.6117   .8431304   153.73   0.000     127.9564    131.2669
------------------------------------------------------------------------------

. testparm i.race

 ( 1)  2.race = 0
 ( 2)  3.race = 0
 ( 3)  4.race = 0

       F(  3,   730) =    1.68
            Prob > F =    0.1693
```

APPENDIX V: Stata output for Problem 7

```
. regress sbp c.alcoh##i.race, robust

Linear regression                               Number of obs     =        735
                                                F(7, 727)         =       1.70
                                                Prob > F          =     0.1049
                                                R-squared         =     0.0175
                                                Root MSE          =     19.584

------------------------------------------------------------------------------
             |               Robust
         sbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       alcoh |   .4325421   .1891784     2.29   0.023      .061141    .8039431
             |
        race |
          2  |   4.977631   2.418321     2.06   0.040     .2299052    9.725356
          3  |   .2108627   3.197077     0.07   0.947    -6.065743    6.487469
          4  |   8.523285   7.436498     1.15   0.252    -6.076288    23.12286
             |
race#c.alcoh |
          2  |  -.2112467    .425641    -0.50   0.620    -1.046879    .6243856
          3  |  -.2397076   .3406035    -0.70   0.482    -.9083915    .4289763
          4  |  -.7565144     .86767    -0.87   0.384    -2.459952    .9469235
             |
       _cons |   129.4707   .8654675   149.60   0.000     127.7716    131.1698
------------------------------------------------------------------------------

. testparm c.alcoh race#c.alcoh

 ( 1)  alcoh = 0
 ( 2)  2.race#c.alcoh = 0
 ( 3)  3.race#c.alcoh = 0
 ( 4)  4.race#c.alcoh = 0

       F(  4,   727) =    1.54
            Prob > F =    0.1877
```