**Sociology 593**
**Exam 1 Answer Key**
**February 15, 2002**

NOTE: Many of these answers are brief and designed to point you in the general direction. Referring back to the lecture notes will give you more detail.

*I.* *True-False.* (20 points) Indicate whether the following statements are true or false. If false, briefly explain why.

1. The Dubin Watson statistic is used to determine what variable, if any, should be entered next in a forward stepwise regression.

False. Durbin Watson is a test for serial correlation.

2. In a bivariate regression, the F value is 81. The T statistic for the beta coefficient is therefore 9.

False. The T value can be either 9 or -9.

3. High multicollinearity results in biased parameter estimates and can make it more difficult to detect significant relationships.

False (or, if you prefer, only half-true). Multicollinearity does not result in biased parameter estimates.

4. A key problem with listwise deletion is that the pieces put together for the regression analysis refer to systematically different subsets of the population, e.g. the cases used in computing $r_{12}$ may be very different than the cases used in computing $r_{34}$.

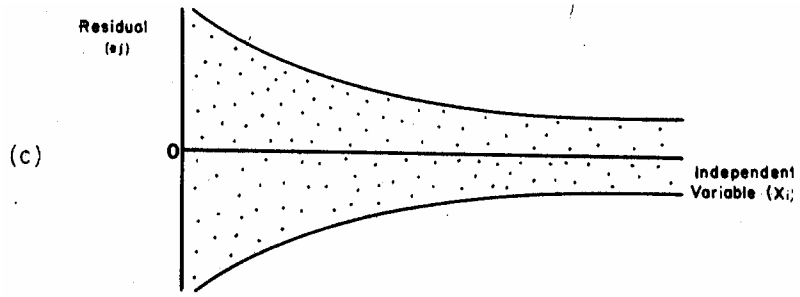
False. This is a problem with pairwise, not listwise deletion of missing data.

5. A researcher encounters an extreme outlier in her data. She should immediately delete the case and then rerun her analyses.

False. You may want to do this eventually, but it shouldn't be your first step. Check for coding errors or improper handling of missing data. Then consider if there are omitted variables that might help to explain the outlier's value.

*II.* *Short answer.* Answer both of the following. (20 points each 40 pts total.)

1. Consider the following plot:




a. Explain how a mis-specified model could produce a pattern such as that shown above. How would you try to solve the problem in this case?

As the notes on heteroskedasticity explain, subgroup differences or omitted variables could produce such a pattern. In such cases, run separate models for each group or add the omitted variables to the model.

b. Explain how a model could be correctly specified and yet you still get a pattern such as the above. Under such conditions, will OLS estimates of the betas be biased? Are there other problems that OLS will have in this situation? Should OLS be used in this case, or is there a superior alternative?

This might occur when a low value on one variable is a necessary but not sufficient condition for a high value on another variable (see hetero notes). OLS estimates will not be biased, but standard errors will be bigger and estimates will be less precise. GLS or WLS is superior (although in practice it may not make that much of a difference unless the heterscedasticity is really extreme).

2. A researcher wants to test whether the effect of attitudes on behavior is different for blacks than it is for whites; that is, she wants to test

$$H_0 : \beta^{White} = \beta^{Black}$$
$$H_A : \beta^{White} \neq \beta^{Black}$$

In reality, the null hypothesis is true: the effect of attitudes on behavior is the same for both blacks and whites. Explain how and why the following conditions might lead the researcher to reach an erroneous conclusion.

a) There is a great deal of random measurement error in the white responses, and very little error in the black responses

In a bivariate regression, random measurement error in the IV will produce a downward bias in the slope coefficient. Since there is more random error for whites than there is for blacks, the downward bias will be greater for them. Hence, it could appear that attitudes have less of an effect on whites than blacks, but this would be just an artifact produced by the measurement problem.

b) Low-income blacks are disproportionately likely to NOT answer questions on attitudes.

Our techniques assume that we have a random and representative sample of the larger population. For blacks, data are missing on a non-random basis. This could cause parameter estimates for blacks to be biased (although we don't know enough to determine what direction that bias might be in).

---

*III. Computation and interpretation.* (40 points total, 10 pts extra credit)

A researcher is interested in what determines how housework gets divided between couples. She has extracted the following variables from the 1996 General Social Survey (GSS).

| Variable | Question |
|---|---|
| SPWork | How much (of the housework) does your spouse or partner do? The values are 1 "Little or None" 2 "Some" 3 "About half" 4 "Most" 5 "All". |
| Educ | Years of Education<br>(ranges from 0, no formal education, to 20 years) |
| White | 1 = white, 0 = black or other |
| Male | 1 = Male, 0 = Female |
| Party | Coded on a 7 point scale where<br>1 = Strong Republican, 4 = Independent, 7 = Strong Democrat |

She begins by running frequencies on the entire sample. She then runs a regression with SPWork as the dependent variable.

```
FREQUENCIES
  VARIABLES=spwork male educ white party
  /Format = Notable / Statistics = Default .
```

## Frequencies

**Statistics**

| | | How much of the housework does your spouse or partner do? | MALE | Highest year of school completed | WHITE | PARTY |
|---|---|---|---|---|---|---|
| N | Valid | 763 | 2904 | 2895 | 2904 | 2855 |
| | Missing | 2141 | 0 | 9 | 0 | 49 |
| Mean | | 2.6920 | .4425 | 13.36 | .8089 | 4.1860 |
| Std. Deviation | | .95339 | .49677 | 2.929 | .39325 | 1.97977 |
| Minimum | | 1.00 | .00 | 0 | .00 | 1.00 |
| Maximum | | 5.00 | 1.00 | 20 | 1.00 | 7.00 |

```
REGRESSION
  /MISSING LISTWISE/ Descriptives DEF
  /STATISTICS COEFF OUTS R ANOVA ZPP TOL
  /CRITERIA=PIN(.05) POUT(.10)
```

```
/NOORIGIN
/DEPENDENT SpWork
/METHOD=ENTER male white educ party.
```

**Descriptive Statistics**

| | Mean | Std. Deviation | N |
|---|---|---|---|
| How much of the housework does your spouse or partner do? | 2.6909 | .94882 | 744 |
| MALE | .4866 | .50016 | 744 |
| WHITE | .8629 | .34418 | 744 |
| Highest year of school completed | 13.53 | 2.820 | 744 |
| PARTY | 3.9758 | 2.01928 | 744 |

**Correlations**

| | | How much of the housework does your spouse or partner do? | MALE | WHITE | Highest year of school completed | PARTY |
|---|---|---|---|---|---|---|
| Pearson Correlation | How much of the housework does your spouse or partner do? | 1.000 | .479 | .047 | .019 | -.057 |
| | MALE | .479 | 1.000 | -.003 | .000 | -.079 |
| | WHITE | .047 | -.003 | 1.000 | .037 | -.255 |
| | Highest year of school completed | .019 | .000 | .037 | 1.000 | -.076 |
| | PARTY | -.057 | -.079 | -.255 | -.076 | 1.000 |

**Model Summary**

| Model | | R Square | Std. Error of the Estimate |
|---|---|---|---|
| 1 | a | [1] | .83366 |

a. Predictors: (Constant), PARTY, Highest year of school completed, MALE, WHITE

**ANOVA[b]**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 155.298 | 4 | 38.825 | 55.863 | .000[a] |
| | Residual | 513.600 | 739 | .695 | | |
| | Total | 668.898 | 743 | | | |

a. Predictors: (Constant), PARTY, Highest year of school completed, MALE, WHITE

b. Dependent Variable: How much of the housework does your spouse or partner do?

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | Correlations | | | Collinearity Statistics | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | | Zero-order | Partial | Part | Tolerance | VIF |
| 1 | (Constant) | 2.0758 | .194 | | 10.713 | .000 | | | | | |
| | MALE | .9081 | .061 | .479 | 14.799 | .000 | [2] | .478 | .477 | .993 | 1.007 |
| | WHITE | .1279 | [3] | .046 | 1.392 | .164 | .047 | .051 | .045 | .934 | 1.070 |
| | Highest year of school completed | .0055 | .011 | .016 | .509 | .611 | .019 | .019 | .016 | .994 | [4] |
| | PARTY | -.0030 | .016 | [5] | -.192 | .848 | -.057 | -.007 | -.006 | .924 | 1.082 |

a. Dependent Variable: How much of the housework does your spouse or partner do?

**Casewise Diagnostics[a]**

| Case Number | Std. Residual | How much of the housework does your spouse or partner do? | Predicted Value | Residual | DFBETA | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | (Constant) | PARTY | Highest year of school completed | MALE | WHITE |
| 62 | 3.417 | 5.00 | 2.1518 | 2.8482 | .0031 | .0008 | .0019 | -.0074 | -.0277 |
| 176 | 3.268 | 5.00 | 2.2752 | 2.7248 | .0121 | -.0019 | .0001 | -.0078 | .0014 |
| 1800 | 3.447 | 5.00 | 2.1266 | 2.8734 | .0242 | .0016 | -.0001 | -.0071 | -.0261 |
| 2518 | 3.289 | 5.00 | 2.2580 | 2.7420 | .0132 | .0000 | -.0007 | -.0072 | .0045 |

a. Dependent Variable: How much of the housework does your spouse or partner do?

a) (10 pts) The researcher is very puzzled by her sample size. As her frequencies show, there are almost 3,000 cases in the 1996 GSS. But, there are only 744 cases in her regression analysis. Explain to the researcher why so many cases are missing. Cite evidence from the printout to support your answer. Other than the reduced sample size, would you yourself be very worried about the missing data, e.g. do you think your results will be seriously biased because of the MD? Are there any additional simple computer runs you would recommend for checking whether MD may be a problem?

She is using listwise deletion, which causes a case to be dropped if it has MD on any variable in the analysis. As the frequencies show, almost all of the missing data is in the SPWork variable. My guess is that most or all of this data is "missing by design." Only people with a spouse or partner were asked the question; it wouldn't make sense to ask this question of others. Hence, it is not surprising that there is great deal of missing data, since the question will be "not applicable" for many respondents. Nevertheless, we should do a little additional checking before we conclude there is no reason for concern.

The descriptive statistics for the full sample are not too different from the descriptive statistics for the cases included in the regression (the regression has slightly more males, 48.66% compared to 44.25%; mean years of education are 13.53 and 13.33 respectively; both the regression sample and the full sample are very close to middle of the road, with means of 3.9758 and 4.1860). These small differences could easily be due to the fact that people with partners are a little different than people without partners.

Just to be safe, however, I would recommend running a complete frequency on SPWork, to make sure that most of the people who were asked the question answered it. I did this, and only 17 people failed to answer the question; all the other MD cases were coded as "Not Applicable." If the refusal rate had been much higher than this, e.g. hundreds of people asked the question had failed to answer, I might have done additional runs, comparing the amount of missing data across groups, such as race and gender.

In short, it is not surprising this question has a lot of missing data – indeed, we would probably expect it to – but we should still check to make sure that most people who could have answered the question did indeed do so. Overall, it appears the researcher has little reason to be concerned about missing data in this case.

b) (10 pts) Fill in the missing quantities [1] – [5].

Here is the uncensored printout:

**Model Summary[b]**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .482[a] | .232 | .228 | .83366 |

a. Predictors: (Constant), PARTY, Highest year of school completed, MALE, WHITE

b. Dependent Variable: How much of the housework does your spouse or partner do?

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | Correlations | | | Collinearity Statistics | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | | Zero-order | Partial | Part | Tolerance | VIF |
| 1 | (Constant) | 2.0758 | .194 | | 10.713 | .000 | | | | | |
| | MALE | .9081 | .061 | .479 | 14.799 | .000 | .479 | .478 | .477 | .993 | 1.007 |
| | WHITE | .1279 | .092 | .046 | 1.392 | .164 | .047 | .051 | .045 | .934 | 1.070 |
| | Highest year of school completed | .0055 | .011 | .016 | .509 | .611 | .019 | .019 | .016 | .994 | 1.006 |
| | PARTY | -.0030 | .016 | -.006 | -.192 | .848 | -.057 | -.007 | -.006 | .924 | 1.082 |

a. Dependent Variable: How much of the housework does your spouse or partner do?

To confirm:

[1] $R^2 = SSR/SST = 155.298/668.898 = .232$

[2] $r_{Male,SPWork} = .479$ (see the earlier table of correlations)

[3] Note that $T_{White} = B_{White}/s_{bWhite}$, so $s_{bWhite} = B_{White}/T_{White} = .1279/1.392 = .092$

[4] $VIF_{Educ} = 1/TOL_{Educ} = 1/.994 = 1.006$

[5] $$b'_{Party} = b_{Party} * \frac{s_{Party}}{s_y} = \frac{-.0030 * 2.01928}{.94882} = -.006$$

c) (10 pts) If you were the researcher, would you be worried about multicollinearity? Why or why not? Is there reason to be greatly concerned about outliers in the data? Assuming the outliers are not coding errors, what do you think would be the better strategy – toss the outliers out, or try to add other explanatory variables to the model?

The low T values might at first make you concerned about multicollinearity. However, the correlations among the Xs are pretty low and the tolerances are all extremely high, so multicollinearity does not seem to be a problem.

As the casewise diagnostics show, there are a few outliers with standardized residuals above three. But, especially given the large sample, the number and size of these outliers does not seem extremely unreasonable. Further, all are coded 5, which is a legitimate value, making it less likely that there is a coding problem. Also, DFBETA shows you the regression estimates would change little if any one of the outliers were dropped (they might change more if all 4 were dropped, but it still doesn't look like it would make that much difference).

Hence, I probably would not drop these cases. If possible, I would double-check to make sure these cases were coded correctly. Then, I might consider adding other variables that would help to explain their extreme values. There are no doubt lots of other variables that explain how much housework gets done by a person's spouse. For example, in these cases, maybe the respondent works full-time or travels a lot while the spouse stays at home.

d) (10 pts) Interpret the results. About how even is the (perceived) housework split – do respondents think they are doing most of the work, do they think their spuse/partner is doing most of the work, or do they think the split is about equal? The researcher chose these variables because she thought a person's gender, race, education and party identification would all affect how much housework their spouse did. To what extent is she right, and to what extent is she wrong? Cite evidence from the printout to support your answer.

Respondents think their spouse is doing a little less than half the housework (see the mean for SPWork). Of course, this could be a biased perception on their part. The only variable that significantly affects SPWork is MALE – which means that when the respondent is male, the partner/spouse (i.e. the female) does more work. This is probably not a shocking finding. None of the other variables have a significant effect; but as explained below, the way the model is set up, you probably wouldnot expect them too.

e) (10 pts extra credit) A colleague is very critical of the researcher's model. She agrees with the choice of gender as an IV, but argues that, at least as the model is set up, it makes little theoretical sense to expect education, race, and party id to have much of an effect. What do you think she is basing her argument on? [HINT: For each of the IVs in the model, do you think spouses will tend to have similar values, or different values? Do you think respondents' characteristics unilaterally determine how much housework their spouse/partner does?]

This model probably has some major specification problems. The amount of housework done by one's partner may be partly influenced by your own characteristics, but it is also influenced by your partner's characteristics. Ergo, the corresponding spouse characteristics should probably also be included in the model (except for gender; unless the question is also asked of same-sex couples). This is a problem of omitted variable bias, which we have alluded to and will soon discuss more fully.

A further complication is that, in the U.S. at least, people tend to marry people who are similar to themselves (except for gender of course). Since partners tend to be of the same race, have similar levels of education, and tend to have similar political beliefs, whatever effect these variables may have will tend to offset each other (e.g. better educated people might prefer to do less housework, but so will their better educated spouses, so for many or most couples the effects of education will tend to cancel out.)

Hence, a better model would have the respondent's race, education and party id and the corresponding variables for the partner. If the researcher is correct in thinking that these vars are important, we would likely find, for example, that both Respondent's Education and Partner's Education have significant effects but the signs are the opposite of each other, e.g. the better educated the respondent is, the more work the partner does, but the better educated the partner is, the less work the partner does.

This is an example of suppressor effects, something we have talked about in the past and will soon talk about again.