An Example of Fuller's (1984) Method for Testing the Bias of Unweighted Estimates of Regression Parameters in a Linear Regression Model (using SAS and R; any software can be used for the SAS portion of this example).

July 1, 2012

The data file BMI.DAT contains data from the 2006 BRFS. The variables are as follows:

AGE: age of subject in years, centered around 50

BMI: BMI of subject

WGT: final sampling weight

First, read in the data:

```
data brfs;
   infile "J:\Regression Analysis\bmi.dat";
   input age bmi wgt;
run;
```

Fit a model predicting BMI as a linear function of age. Using the method of Fuller (1984) described in K&G 4.6-2, estimate the bias in the unweighted estimator of the slope relating age to BMI, and test whether this estimate differs from 0 at the $\alpha = 0.05$ level.

First, we compute the unweighted slope and intercept:

```
proc reg data = brfs;
   model bmi = age;
run;
quit;
```

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	27.38628	0.01449	1890.11	<.0001
age	**1**	**-0.00099942**	**0.00084223**	**-1.19**	**0.2354**

The unweighted estimate of the intercept is 27.386, while the unweighted estimate of the slope is -0.00099942. We are primarily focused on the estimate of the slope.

Next, compute the weighted slope and intercept:

```
proc surveyreg data = brfs;
   model bmi = age;
   weight wgt;
run;
```

Estimated Regression Coefficients

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	27.3580690	0.02650062	1032.36	<.0001
age	**0.0109521**	**0.00148287**	**7.39**	**<.0001**

NOTE: The denominator degrees of freedom for the t tests is 168146.

The weighted estimate of the intercept is 27.358, while the weighted estimate of the slope is 0.0109521. There appears to be a fairly substantial bias when failing to use the sampling weights to estimate the slope, as the sign of the slope actually changes (from negative to positive) compared to the unweighted estimate, and the slope would now be considered significantly different from 0 when using the sampling weights. **The estimate of the bias in the unweighted estimator of the slope is 0.0109521 - -0.00099942 = 0.01195152.**

The results from K&G 4.6-2 suggest that we can fit a simple linear regression model to a data set including two additional predictor variables, one being the sample weights (or the column of 1s for the intercept multiplied by the sampling weights), and one being a weighted version of the predictor variable AGE (the original predictor multiplied by the sampling weights), and test the bias in the unweighted regression model by testing whether the vector of two additional regression parameters for these predictors is equal to **0**. However, in this example we are focused on only a subset of the parameters, namely the slope only. From the result shown previously in K&G 4.6-2, the estimates of the two additional parameters can be written as

$$\hat{\tau}_u = A(\hat{\beta} - \hat{\beta}_u),$$

where the matrix *A* is defined on page 191 of K&G. To test whether a *subset* of *q* of the unweighted regression coefficients in the original model are biased, one uses the test statistic

$$W^* = (\hat{\beta}^* - \hat{\beta}_u^*)'[\hat{\text{cov}}(\hat{\beta}^* - \hat{\beta}_u^*)]^{-1}(\hat{\beta}^* - \hat{\beta}_u^*)$$

where

$$\hat{\beta}^{*} - \hat{\beta}_{u}^{*}$$

is the *q*-dimensional vector of the *differences* in the estimates of the unweighted and weighted regression coefficients (in this case, just the difference in the unweighted and weighted slope, 0.01195), and

$$\hat{\text{cov}}(\hat{\beta}^{*} - \hat{\beta}_{u}^{*})$$

is the appropriate submatrix of

$$\hat{\text{cov}}(\hat{\beta} - \hat{\beta}_{u})$$

where

$$\hat{\text{cov}}(\hat{\beta} - \hat{\beta}_{u}) = A^{-1}\,\text{cov}(\hat{\tau}_{u})A'^{-1}$$

In this part of the example, given that there are two parameters being estimated in both models, the variance-covariance matrix of the vector of differences in the unweighted and weighted parameter estimates should be a 2 x 2 matrix, and we are primarily interested in the variance of the difference between the weighted and unweighted estimates of the slope for computing the test statistic W^{*}.

First, we compute an estimate of the τ_u vector in SAS per the method in K&G 4.6-2, along with the variances and covariances of the estimated τ_u vector:

```
data brfs2;
   set brfs;
   agewgt = age * wgt;
run;

proc reg data = brfs2;
   model bmi = age wgt agewgt / covb;
run;
quit;
```

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	27.39997	0.01762	1554.99	<.0001
age	1	-0.00672	0.00102	-6.56	<.0001
wgt	**1**	**-0.00000179**	**0.00006874**	**-0.03**	**0.9793**
agewgt	**1**	**0.00004151**	**0.00000425**	**9.76**	**<.0001**

Covariance of Estimates

Variable	Intercept	age	wgt	agewgt
Intercept	0.0003104901	-2.998061E-6	-6.832997E-7	1.249969E-9
age	-2.998061E-6	1.048282E-6	1.2591293E-9	-2.452524E-9
wgt	-6.832997E-7	1.2591293E-9	**4.725735E-9**	**3.134604E-11**
agewgt	1.249969E-9	-2.452524E-9	**3.134604E-11**	**1.807246E-11**

Next, the data and this variance-covariance matrix were imported into R:

```
brfs <- read.table("J:\\Regression Analysis\\bmi.dat",h=T)

cov.tau <-
cbind(c(0.0000000047257,0.000000000031346),c(0.000000000031346,0.0000000000
180725))

cov.tau

            [,1]         [,2]

[1,] 4.7257e-09 3.13460e-11

[2,] 3.1346e-11 1.80725e-11
```

Next, we compute the *A* matrix per page 191 of K&G:

```
w <- Diagonal(x = brfs$wgt)


dim(w)

[1] 168147 168147


unwgt.model <- lm(bmi ~ age, brfs)


unwgt.model # check that the unweighted results match with SAS output


Call:

lm(formula = bmi ~ age, data = brfs)
```

```
Coefficients:

(Intercept)          age

 27.3862770   -0.0009994


x <- model.matrix(unwgt.model) # extract the X matrix from the model


a <- solve((t(x)%*%w%*%w%*%x)-
(t(x)%*%w%*%x%*%solve(t(x)%*%x)%*%t(x)%*%w%*%x)) %*% (t(x)%*%w%*%x)
```

Next, we confirm the result in K&G 4.6-2 given the computed *A* matrix, by refitting the weighted and unweighted models and then re-computing the estimate of the τ_u vector:

```
wgt <- solve(t(x)%*%w%*%x)%*%t(x)%*%w%*%brfs$bmi


wgt

2 x 1 Matrix of class "dgeMatrix"

            [,1]

[1,] 27.35806896

[2,]  0.01095214


unwgt <- solve(t(x)%*%x)%*%t(x)%*%brfs$bmi


unwgt

                    [,1]

(Intercept) 27.386276975

age         -0.000999425


a %*% (wgt - unwgt)

2 x 1 Matrix of class "dgeMatrix"

              [,1]

[1,] -1.786781e-06

[2,]  4.150864e-05
```

The resulting vector of parameter estimates matches with the boldfaced estimates of the τ_u vector in the SAS output above.

Then, we compute the variance-covariance matrix of the differences in the unweighted and weighted estimates of the parameters in the full model:

```
solve(a)%*%cov.tau%*%t(solve(a))
```

```
2 x 2 Matrix of class "dgeMatrix"
              [,1]          [,2]
[1,]  4.373068e-04 -4.166977e-06
[2,] -4.166977e-06  1.481237e-06
```

We are primarily interested in the variance of the difference between the weighted and unweighted estimates of the slope, which is boldfaced above. Given this information and the weighted and unweighted estimates of the slope computed above, we compute the test statistic *W** for testing the bias:

```
(0.0109521 - -0.00099942)*(0.0109521 - -0.00099942) / 0.000001481237
[1] 96.43212
```

Per K&G 4.6-2, this test statistic *W** can be multiplied by ($d - q + 1$), where d is the degrees of freedom and q is the number of parameters being tested (making this term equal to d), divided by (dq), which is also equal to d (resulting in *W**), and then referred to an F distribution with $q = 1$ and $d - q + 1 = 168{,}147$ (= 168,147 – 1 + 1) degrees of freedom:

```
1-pf(96.43212,1,168147)
[1] 0
```

We thus have strong evidence against the null hypothesis that there is no bias in the unweighted estimate of the slope relating age to BMI; the unweighted analysis would result in a biased estimate of the relationship of age with BMI, making it necessary to use the sampling weights to estimate this slope.