Notes for laboratory session 2

Preliminaries

Consider the ordinary least-squares (OLS) regression of alcohol (`alcohol`) and plasma retinol (`retplasm`). We do this with STATA as follows:

```
. reg  retplasm alcohol

      Source |       SS       df       MS              Number of obs =     314
-------------+------------------------------           F(  1,   312) =   16.19
       Model |   671843.17     1   671843.17           Prob > F      =  0.0001
    Residual |  12948338.7   312  41501.0855           R-squared     =  0.0493
-------------+------------------------------           Adj R-squared =  0.0463
       Total |  13620181.9   313   43514.958           Root MSE      =  203.72

------------------------------------------------------------------------------
    retplasm |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     alcohol |   9.365251   2.327637     4.02   0.000     4.785401     13.9451
       _cons |   578.8857   13.04634    44.37   0.000     553.2158    604.5556
```

Try to locate the following:

a. What is the overall significance of the model and how is it being assessed?
b. What is the effect of alcohol on plasma retinol?
c. For each unit of alcohol consumption increase what is the unit-change in plasma retinol? What is the 95% confidence interval?

Now do the same using the `glm` command of STATA.

```
. glm retplasm alcohol

Iteration 0:   log likelihood = -2113.9991

Generalized linear models                         No. of obs      =       314
Optimization     : ML: Newton-Raphson             Residual df     =       312
                                                  Scale parameter =  41501.09
Deviance         =  12948338.69                   (1/df) Deviance =  41501.09
Pearson          =  12948338.69                   (1/df) Pearson  =  41501.09

Variance function: V(u) = 1                       [Gaussian]
Link function    : g(u) = u                       [Identity]
Standard errors  : OIM

Log likelihood   = -2113.999055                   AIC             =   13.4777
BIC              =  12946544.88

------------------------------------------------------------------------------
    retplasm |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     alcohol |   9.365251   2.327637     4.02   0.000     4.803167    13.92734
       _cons |   578.8857   13.04634    44.37   0.000     553.3153    604.4561
------------------------------------------------------------------------------
```

Try to notice the similarities between the two approaches. Specifically, notice the following:

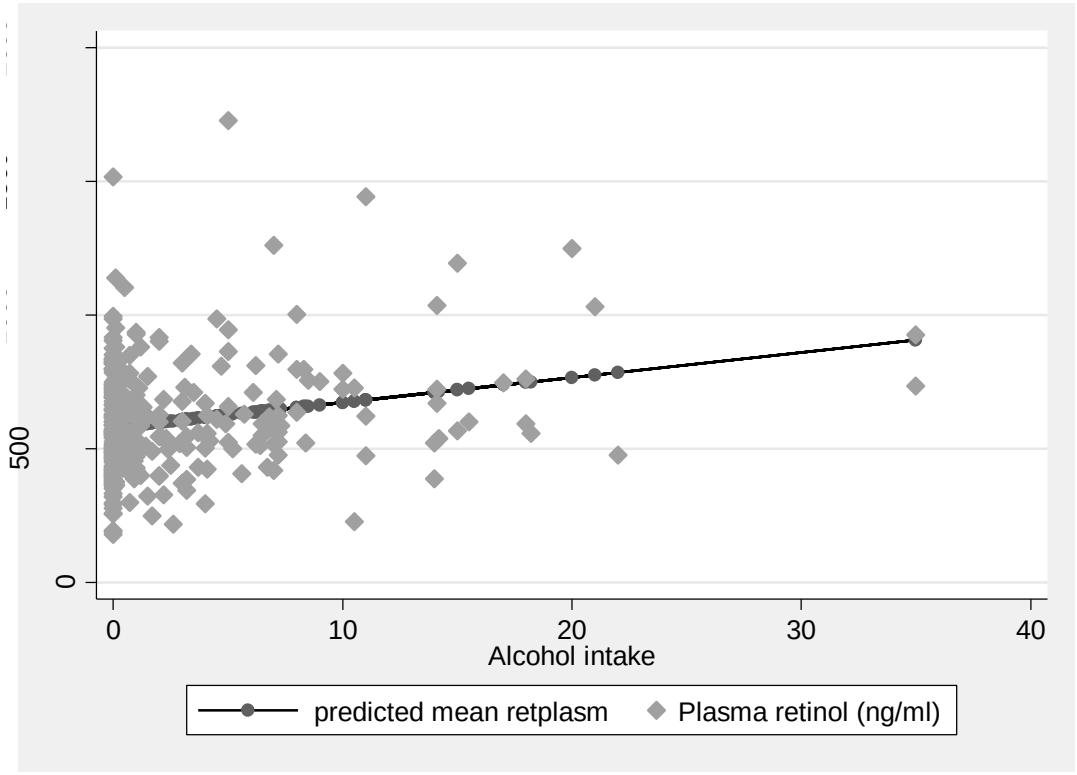
d. Note the type of link and variance function. Why do you think these are the links and variance function used?

Produce the predicted regression line for alcohol consumption, along with a scatter plot of the observed values.

```
. quietly glm retplasm alcohol

. predict yhat
(option mu assumed; predicted mean retplasm)

.sc yhat retplasm alcohol, c(l .) ms(i o)scheme(s2mono)
```



Model building

Assess the effect of adding variable `fat` after `alcohol` has been added to the model. Recall from your previous experience that this can be done with the general linear model procedure as follows:

```
. anova retplasm c.alcohol c.fat,  seq
```

```
                           Number of obs =     314     R-squared     =  0.0617
                           Root MSE      = 202.716     Adj R-squared =  0.0556

                  Source |    Seq. SS     df       MS           F     Prob > F
              -----------+----------------------------------------------------
                   Model |  839986.495     2  419993.248      10.22     0.0001
                         |
                 alcohol |   671843.17     1   671843.17      16.35     0.0001
                     fat |  168143.325     1  168143.325       4.09     0.0439
                         |
                Residual |  12780195.4   311  41093.8758
              -----------+----------------------------------------------------
                   Total |  13620181.9   313   43514.958
```

e. What is the criterion of whether fat intake has a significant effect on plasma retinol levels *after* adjusting for alcohol intake?

f. What is the decision about whether we should include fat intake in a model of plasma retinol levels?

Now using the `glm` command in STATA:

```
. glm  retplasm alcohol fat

Iteration 0:   log likelihood = -2111.9469

Generalized linear models                         No. of obs      =       314
Optimization     : ML: Newton-Raphson             Residual df     =       311
                                                  Scale parameter =  41093.88
Deviance         =  12780195.36                   (1/df) Deviance =  41093.88
Pearson          =  12780195.36                   (1/df) Pearson  =  41093.88

Variance function: V(u) = 1                       [Gaussian]
Link function    : g(u) = u                       [Identity]
Standard errors  : OIM

Log likelihood   = -2111.946946                   AIC             =    13.471
BIC              =   12778407.3

------------------------------------------------------------------------------
    retplasm |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     alcohol |   9.990265   2.336708     4.28   0.000     5.410401    14.57013
         fat |  -.6975524   .3448463    -2.02   0.043    -1.373439   -.0216661
       _cons |   630.7705    28.7483    21.94   0.000     574.4249    687.1162
------------------------------------------------------------------------------
```

We carry out the calculations that lead to the decision about adding or not of fat in the model. Consider using the deviance of the joint model (with fat and alcohol included) versus the model with only alcohol included.

We do this as follows: The deviance of the former model is $D(X_1) = 12948338.6\,9$, while the one for the latter model is $D(X_1, X_2) = 12780195.3\,6$, where X_1 is `alcohol` and X_2 is `fat`. The criterion is $\frac{D(X_1) - D(X_1, X_2)}{D(X_1, X_2)/n-3} = 4.09$.

g. We can compare this to a chi-square distribution with one degree of freedom. Why?

This is done as follows:

```
. display chi2tail(1,4.09)
.04313765
```

The p-value is 0.043<0.05 which suggests that fat should be included into the model. Alternatively, you can use the `test` command as follows:

```
. test fat

 ( 1)  [retplasm]fat = 0

           chi2(  1) =    4.09
         Prob > chi2 =    0.0431
```

If you wanted to assess whether two variables added, after alcohol consumption has been entered in the model, are significant, you can use the same method. Consider the following:

```
.  glm retplasm alcohol fat fiber

Iteration 0:   log likelihood = -2111.9263

Generalized linear models                         No. of obs      =       314
Optimization     : ML: Newton-Raphson             Residual df     =       310
                                                  Scale parameter =  41221.03
Deviance         =  12778518.55                   (1/df) Deviance =  41221.03
Pearson          =  12778518.55                   (1/df) Pearson  =  41221.03

Variance function: V(u) = 1                       [Gaussian]
Link function    : g(u) = u                       [Identity]
Standard errors  : OIM

Log likelihood   = -2111.926345                   AIC             =  13.47724
BIC              =  12776736.24

------------------------------------------------------------------------------
    retplasm |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     alcohol |   9.964244   2.343874     4.25   0.000     5.370336    14.55815
         fat |  -.6767318   .3604768    -1.88   0.060    -1.383253    .0297898
       fiber |  -.4526052   2.244069    -0.20   0.840    -4.850899    3.945688
       _cons |   635.0317   35.71259    17.78   0.000     565.0363    705.0271
------------------------------------------------------------------------------
```

Now you can test the addition of fat and fiber intake in the model as follows:

```
. test fat fiber

 ( 1)  [retplasm]fat = 0
 ( 2)  [retplasm]fiber = 0

           chi2(  2) =    4.12
         Prob > chi2 =    0.1275
```

The results imply that the *joint* effect of fat and fiber intake is not significant when considered in addition to alcohol intake.

h. Can you replicate these results by hand, by considering this model and compare it to the one with only alcohol consumption included?