Answers to the questions

Preliminaries

a. What is the overall significance of the model and how is it being assessed?

The significance of the model is assessed by the *F* test which is equal to 16.19 with an associated p value equal to 0.0001.

b. What is the effect of alcohol on plasma retinol?

In this case, this can be assessed by the *F* test. For general situations, we look at the *t* test associated with the slope of the line for alcohol consumption. The *t* test here has value 4.02, which compared to a *t* statistic with 312 degrees of freedom. The associated p value is less than 0.001. Recall that, in this case, the square of the *t* statistic is equal to the value of the *F* statistic.

c. For each unit of alcohol consumption increase how what is the unit-change in plasma retinol? What is the 95% confidence interval?

For each unit of alcohol consumption (i.e., for one additional drink per day) the change in plasma retinol levels will be 9.37 ng/ml. The 95% confidence interval is [4.79, 13.95] of plasma retinol. This is produced by adding or subtracting to the average estimate $t_{312;0.975}$ standard error of $\hat{\beta}$.

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

The identity link is used because in the linear model, the function $g(\mu) = \mu = \sum_{i=1}^{n} \beta_i X_i$. The variance function is

$V(\mu) = b''(\theta) = 1 (recall\, b(\theta) = \frac{\theta^2}{2}\, in\, normal\, regression)$.

Model building

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

The criterion produced by the `anova` procedure with the `sequential` option is the *conditional F* test statistic, which in this case is *F*(`fat|alcohol`)=4.09. This compared to an *F* distribution with 1 and 311 degrees of freedom (the denominator degrees of freedom are associated with the residual of the larger model) produce a p value of 0.0439.

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

If the alpha level used is 5%, this *F* statistic would imply that the incremental effect of fat intake, over and above alcohol consumption is significant.

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

The one degree of freedom is associated with the difference in the dimensionality (number of explanatory effects) in the two models.

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

To replicate the results provided by the `test` command we proceed as follows:

$$\frac{D(\texttt{alcohol}) - D(\texttt{alcohol, fat, fiber})}{D(\texttt{alcohol, fat, fiber})/310} = \frac{12948338.7\text{-}12778518.55}{41221.03} = 4.12$$

Then we can check the p value associated with this statistic as follows:

```
. display chi2tail(2,4.12)
.12745397
```

Notice that the chi-square statistic is double the *F* statistic that would form the usual *F* criterion that we used in previous experience. This is because in deriving the chi-square we did not divide by the difference in the dimensionality of the two models (2). Recall that the usual *F* criterion is derived as follows:

$$\frac{[SSE(\texttt{alcohol}) - SSE(\texttt{alcohol, fat, fiber})]/2}{SSE(\texttt{alcohol, fat, fiber})/310} = \frac{(12948338.7\text{-}12778518.55)/2}{41221.03} = 2.06$$

To see what the p value associated with this test is we do the following:

```
. di Ftail(2,310,2.06)
.12919517
```

This is close to the previous p values associated with the chi-square statistic (The chi-square test is an asymptotic approximation of the F-test).

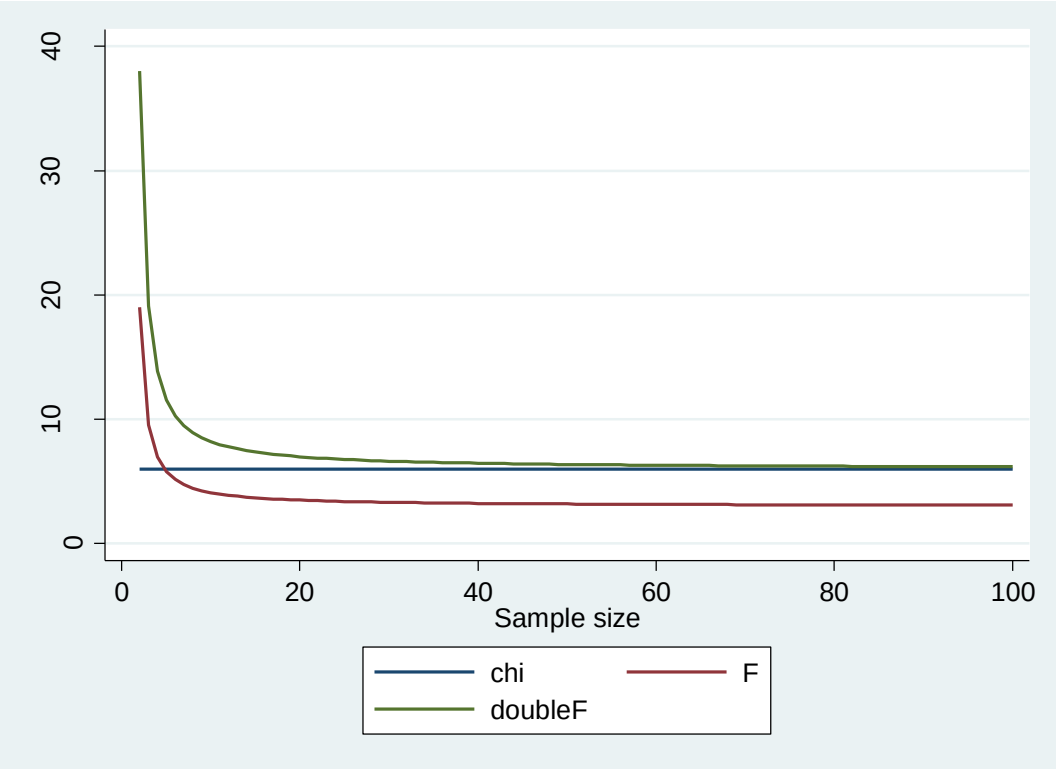
Check the following output

```
. reg retplasm alcohol fat fiber

. test fat fiber

 ( 1)  fat = 0.0
 ( 2)  fiber = 0.0

       F(  2,   310) =    2.06
            Prob > F =    0.1292
```

```
. qui glm retplasm alcohol fat fiber

. test fat fiber

 ( 1)  [retplasm]fat = 0.0
 ( 2)  [retplasm]fiber = 0.0

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

Check the following code and graph showing the relationship between χ^2 and F distributions (2 d.f. in the nominator)

```
drop _all
set obs 100
gen n=_n
gen chi=invchi2tail(2,0.05)
gen F=invFtail(2,n,0.05)
gen doubleF=2*F
sc chi F doubleF n if F<100,c(l l l) ms(i i i ) ///
lw(medthick medthick medthick) xti(Sample size) yti(Critical value)
```



Alternatively, using a Wald type test approach:

```
. qui glm retplasm alcohol fat fiber
. mat li e(b)

e(b)[1,4]
      retplasm:   retplasm:   retplasm:   retplasm:
       alcohol         fat       fiber       _cons
y1   9.9642436  -.67673183  -.45260517   635.03172

. mat li e(V)

symmetric e(V)[4,4]
                    retplasm:   retplasm:   retplasm:   retplasm:
                     alcohol         fat       fiber       _cons
retplasm:alcohol    5.493744
    retplasm:fat  -.12020051   .12994354
  retplasm:fiber   .28951807  -.23165708   5.0358435
  retplasm:_cons   -9.034527  -6.6917157  -47.411313   1275.3889

. mat b=e(b)

. mat b1=b[1..1,2..3]

. mat li b1

b1[1,2]
      retplasm:   retplasm:
           fat       fiber
y1  -.67673183  -.45260517

. mat V=e(V)

. mat V1=V[2..3,2..3]

. mat li V1

symmetric V1[2,2]
                  retplasm:   retplasm:
                       fat       fiber
  retplasm:fat   .12994354
retplasm:fiber  -.23165708   5.0358435

. mat x2=b1*syminv(V1)*b1'

. mat li x2

symmetric x2[1,1]
           y1
y1  4.1197455
```

Recall that $\hat{\theta} \sim N(\theta, \hat{I}(\theta)^{-1})$

and

$$W = \left(\hat{\theta} - \theta_{\mathrm{o}}\right)' \hat{I}(\theta) \left(\hat{\theta} - \theta_{\mathrm{o}}\right) \sim \chi^2_n$$