GENERALIZED LINEAR MODELS – Introduction (2)

Generalized Pearson statistic

Apart from the deviance another important measure of discrepancy is the generalized Pearson Statistic

\[ X^2 = \sum \frac{(y - \hat{\mu})^2}{V(\hat{\mu})} \]

where \( V(\hat{\mu}) \) is the estimated variance function for the distribution concerned.

Normal distribution \( \rightarrow \) \( X^2 \) is RSS (i.e. residual sum of squares).

Poisson or binomial \( \rightarrow \) original Pearson \( X^2 \) statistic
Deviance and generalized Pearson $X^2$ statistic

- Both the deviance and the generalized Pearson $X^2$ have exact $X^2$ distributions for normal-theory linear models (assuming of course that the model is true) and asymptotic results are available for other distributions.

- The deviance has a general advantage as a measure of discrepancy in that it is additive for NESTED sets of models if MLEs are used, whereas $X^2$ in general is not. However, $X^2$ sometimes may be preferred because of its direct interpretation.

- Note that the quantity $X^2/n-p$, where $n$ the number of observations and $p$ the number of parameters in a model, gives an estimate of the scale or dispersion parameter.
An algorithm for fitting GLM

**Goal:** To show that the MLEs of the parameter $\beta$ in the linear predictor $\eta$ can be obtained by iterative weighted least squares.

In this regression

- The dependent variable is a linearized form of the link function applied to $y$.
- The weights are functions of the fitted values $\hat{\mu}$.

The process is iterative because both the adjusted dependent variable $Z$ and the weight $w$ depend on the fitted values, for which only current estimates are available. The procedure underlying the iteration is as follows:
**ALGORITHM**

(1) Let \( \hat{\eta}_0 \) be the current estimate of the linear predictor with corresponding fitted value \( \hat{\mu}_0 \) derived from the link function \( \eta = g(\mu) \).

(2) Form the **adjusted dependent variate** with typical value

\[
z_0 = \hat{\eta}_0 + \left( y - \hat{\mu}_0 \right) \left( \frac{d\eta}{d\mu} \bigg| \hat{\mu}_0 \right)
\]

(3) The quadratic weight is defined as

\[
W_0^{-1} = \left( \frac{d\eta}{d\mu} \bigg| \hat{\mu}_0 \right)^2 V_0
\]

where \( V_0 \) is the variance function evaluated at \( \hat{\mu}_0 \). Now regress \( z_0 \) onto covariates \( x_1, \ldots, x_p \) with weights \( W_0 \) to give new parameter estimates \( \hat{\beta}_1 \). From these form a new estimate \( \hat{\eta}_1 \) of the linear predictor and repeat until changes in the estimates are sufficiently small.
Note that $z$ is just a linearized form of the link function applied to the data, because, up to first order

$$g(y) \approx g(\mu) + g'(\mu)(y - \mu)$$

hence the right hand side is

$$\eta + (y - \mu) \frac{d\eta}{d\mu}$$

(because $\eta = g(\mu)$)

Moreover, $Var(Z) = W^{-1}$ assuming that $\eta$ & $\mu$ are fixed and known.

Convenient feature: it suggests a simple starting point to get the iteration under way. This consists of using the data themselves as the first estimate of $\hat{\mu}_0$ and from this deriving $\hat{\eta}_0$, $\left( \frac{d\eta}{d\mu} \bigg| \hat{\mu}_0 \right)$ and $V_0$.

Note that adjustments may be required to the data to prevent, for example, our trying to evaluate $\log(0)$ if the log link is used.
Single-factor analysis of variance

Analyses of variance and covariance can be expressed in linear regression terms. For example, consider the one-way analysis of variance model

\[ y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad i = 1, \ldots, p, \quad j = 1, \ldots, n \]

where \( \alpha_i \) is the treatment effect and \( \epsilon_{ij} \) is the error associated with the \( i^{th} \) treatment and \( j^{th} \) observation can be recast as a simple linear regression model by defining \( X_i = \begin{cases} 1, & \text{if group } i \text{ } i = 1, \ldots, p-1. \end{cases} \)

Thus expressed the one-way ANOVA model becomes

\[ y_{ij} = \beta_0 + \beta_1 X_1 + \ldots + \beta_{p-1} X_{p-1} + \epsilon_{ij} \]

that is,

Group 1: \( Y_{1j} = \beta_0 + \beta_1 + \epsilon_{ij} \)

Group 2: \( Y_{2j} = \beta_0 + \beta_2 + \epsilon_{ij} \)

\[ M \]

Group \( p-1: Y_{(p-1)j} = \beta_0 + \beta_{(p-1)} + \epsilon_{ij} \)

Group \( p: Y_{pj} = \beta_0 + \epsilon_{ij} \)
**Regression models for one-way ANOVA**

This regression model is equivalent to the ANOVA model. To see this consider that

\[
\mu_i = \begin{cases} 
\beta_o + \beta_i, & \text{if } i = 1, \ldots, p-1 \\
\beta_o, & \text{if } i = p
\end{cases}
\]

The usual null hypothesis in regression \(H_0: \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0\), means that

\[
\beta_1 = \mu_1 - \mu = 0 \Rightarrow \mu_1 = \mu = \mu_p
\]

\[
\vdots
\]

\[
\beta_{p-1} = \mu_{p-1} - \mu = 0 \Rightarrow \mu_{p-1} = \mu = \mu_p
\]

is thus equivalent to the null hypothesis of the analysis of variance \(H_0: \mu_1 = \mu_2 = \cdots = \mu_p\).

The previous coding scheme is called *reference-coding scheme* since one level of the fixed (categorical) factor is the *reference* level, while the rest are defined as deviations from it. In the model above, we chose level \(p\) as the reference level but we could have easily chosen level 1 (or 2 or 3). The critical point is that coding a factor with \(p\) levels requires \(p-1\) coding variables.\(^1\)

---

\(^1\) This is in the case of a regression model with an intercept. If no intercept exists, then \(p\) coding variables are necessary and no reference category is required.
Example: Effect of gender on plasma retinol levels

Consider the effect of gender on levels of retinol in plasma. The one-way ANOVA is given by the following output:

```
. anova retplasm sex

        Number of obs =     314     R-squared     =  0.0392
        Root MSE      = 204.801     Adj R-squared =  0.0361

Source | Partial SS    df       MS           F     Prob > F
----------+--------------------------------------------------
Model    |  533837.408     1  533837.408      12.73     0.0004
  sex    |  533837.408     1  533837.408      12.73     0.0004
Residual|  13086344.5   312  41943.4117
----------+--------------------------------------------------
  Total  |  13620181.9   313   43514.958
```

The $F$ test is significant, implying that gender differences have a statistically significant impact on plasma retinol levels.
The output of the regression for the same model is as follows:

```
. reg

Source | SS       df       MS
---------+-------------------
Model    | 533837.408     1  533837.408               Prob > F      =  0.0004
Residual | 13086344.5   312  41943.4117               R-squared     =  0.0392
---------+-------------------
Total    | 13620181.9   313  43514.958               Root MSE      =  204.80
---------+-------------------

retplasm        Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
----------------+---------------------------------
   _cons        587.7216   12.39511     47.416   0.000        563.333
                  612.1102
   sex
   1     122.3759   34.30232      3.568   0.000       54.88283    189.8691
   2    (dropped)                  
```

The factor `sex==2` (female) has been defined by default as the reference category. Thus, the best estimate for plasma retinol levels for women will be equal to $\hat{\beta}_o = 587.7216$, while the same estimate for males will be $\hat{\beta}_o + \hat{\beta}_1 = 587.7216 + 122.3759 = 710.0976$ as described previously.
We can execute these calculations in a single step using the `reg` command with reference coding

```
. xi: reg retplasm i.sex
i.sex                Isex_1-2    (naturally coded; Isex_1 omitted)

Source |       SS      df      MS
---------+-----------------+
Model |  533837.408     1  533837.408
Residual | 13086344.5   312  41943.4117
---------+-----------------+
Total | 13620181.9  313  43514.958

Number of obs =     314
F(  1,   312) =   12.73
Prob > F =  0.0004
R-squared =  0.0392
Adj R-squared =  0.0361
Root MSE =  204.80

|          | Coef.   Std. Err. | t     | P>|t|       | [95% Conf. Interval] |
|----------|------------------|-------|---------|----------------------|
| Isex_2   | -122.3759        | 34.30232 | -3.568 | 0.000               | -189.8691 -54.88283  |
| _cons    |  710.0976        | 31.98453 | 22.201 | 0.000               | 647.1649 773.0302   |
```

The default reference category is \texttt{sex==1} (male). The means for males are \( \hat{\beta}_0 = 710.0976 \) and for females \( \hat{\beta}_0 + \hat{\beta}_1 = 710.0976 - 122.3759 = 587.7216 \) consistent with the previous results.
If we wished to use the female category as reference, we modify the above code as follows:

```
. char sex[omit] 2

. xi: reg retplasm i.sex

  i.sex                 Isex_1-2     (naturally coded; Isex_2 omitted)

Source |       SS       df       MS
---------+------------------+
Model    |  533837.408     1  533837.408       F(  1,   312) =  12.73
Residual | 13086344.5   312  41943.4117          Prob > F      =  0.0004
---------+------------------+
Total    | 13620181.9   313  43514.958          Adj R-squared =  0.0361
---------+------------------+
        | Number of obs =     314
        | R-squared     =  0.0392
        | Root MSE      =  204.80

retplasm |      Coef.   Std. Err.     t    P>|t|     [95% Conf. Interval]
---------+------------------+
  Isex_1  |   122.3759   34.30232     3.568   0.000      54.88283    189.8691
  _cons   |   587.7216   12.39511     47.416   0.000      563.333    612.1102
---------+------------------+
```

where the command `char sex[omit] 2` specifies explicitly the omitted category 2 (females).
Using the `xi` and `glm` commands and specifying females as reference we have:

```
. char sex[omit] 2

. xi: glm retplasm i.sex
i.sex                 Isex_1-2     (naturally coded; Isex_2 omitted)

Iteration 1 : deviance = 13086344.4522

Residual df  =       312                                No. of obs =       314
Pearson X2    =  1.31e+07                                Deviance   =  1.31e+07
Dispersion    =  41943.41                                Dispersion =  41943.41

Gaussian (normal) distribution, identity link

|    | Coef. | Std. Err. | t    | P>|t| | [95% Conf. Interval] |
|----|-------|-----------|------|-----|----------------------|
| Isex_1 | 122.3759 | 34.30232   | 3.568 | 0.000 | 54.88283    189.8691   |
| _cons | 587.7216  | 12.39511   | 47.416 | 0.000 | 563.333      612.1102   |

(Model is ordinary regression, use `regress` instead)
```

Consistently with the regression-analysis results.
Comments:

1. The command \texttt{xi} defines the level with the \textit{lowest} numerical value as the default reference level. We can manipulate which level is the reference level by defining the \texttt{omit} variable with the command \texttt{char varname[omit] \# where “#” is the numerical value corresponding to the desired reference level. An alternative case is to define as the reference level the most frequent (prevalent) level with \texttt{char _dta[omit] "prevalent"}. In case of string variables the command becomes \texttt{char _dta[omit] "string_literal" where string_literal is the string level that we want to define as reference.}

2. The \texttt{xi} command defines \textit{p-1} variables \texttt{Ivarname\_i}, \((i=1,...,p-1)\), such that \texttt{Ivarname\_i=(varname==i)}.

3. The regression can then be carried out by these variables. To invoke them we use the umbrella term \texttt{i.varname}. 
Regression models for general two-way ANOVA

In the two-way ANOVA the reference coding scheme is implemented as follows:

\[ Y = \mu + \sum_{i=1}^{p-1} \alpha_i X_i + \sum_{j=1}^{q-1} \beta_j Z_j + \sum_{i=1}^{p-1} \sum_{j=1}^{q-1} \gamma_{ij} X_i Z_j + \epsilon_{ij} \]

where \( X_i = \begin{cases} 1, & \text{if treatment } i \\ 0, & \text{otherwise} \end{cases} \quad i = 1, \ldots, p - 1 \) and \( Z_j = \begin{cases} 1, & \text{if block } j \\ 0, & \text{otherwise} \end{cases} \quad j = 1, \ldots, q - 1 \), with \( p \) and \( q \) the number of treatments and blocks respectively.
Implications of coding

The means can be expressed in terms of the coefficients of the regression (this is helpful so we can interpret the output from statistical packages):

\[ \mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}, \quad i = 1, \ldots, p - 1; \quad j = 1, 2, \ldots, q - 1 \]
\[ \mu_{iq} = \mu + \alpha_i, \quad i = 1, \ldots, p - 1 \]
\[ \mu_{pj} = \mu + \beta_j, \quad j = 1, \ldots, q - 1 \]
\[ \mu_{pq} = \mu \]
Example: Effect of sex and vitamin use on plasma retinol levels

In this case we have two blocks, gender with $p=2$ categories (male, female) and vitamin use with $q=3$ categories (“fairly often”, “not often”, “no use”). With females and the no-vitamin-use categories used as reference categories, each observation is given by the following equation:

$$y_{ijk} = \mu + \alpha_1 X_{1k} + \sum_{j=1}^{2} \beta_j Z_{jk} + \sum_{j=1}^{2} \gamma_{ij} X_{1k} Z_{jk} + \epsilon_{ijk}$$

where $X_1 = \begin{cases} 1, \text{male} \\ 0, \text{otherwise} \end{cases}$, $Z_1 = \begin{cases} 1, \text{"fairly often"} \\ 0, \text{otherwise} \end{cases}$ and $Z_2 = \begin{cases} 1, \text{"not often"} \\ 0, \text{otherwise} \end{cases}$
The STATA output (using the *xi* command) is as follows:

```
. char sex[omit] 2
. char vituse[omit] 3

. xi: glm retplasm i.sex*i.vituse
i.sex                  Isex_1-2         (naturally coded; Isex_2 omitted)
i.vituse               Ivitus_1-3       (naturally coded; Ivitus_3 omitted)
i.sex*i.vituse         IsXv_#-#        (coded as above)

Iteration 1 : deviance = 12783793.5808

Residual df = 308                 No. of obs = 314
Pearson X2 = 1.28e+07              Deviance = 1.28e+07
Dispersion = 41505.82              Dispersion = 41505.82

Gaussian (normal) distribution, identity link
```
(STATA output continued)

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>retplasm</td>
<td>Coef.</td>
<td>Std. Err.</td>
<td>t</td>
<td>P&gt;</td>
<td>t</td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>--------</td>
<td>-----------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td>Isex_1</td>
<td>166.3468</td>
<td>47.76693</td>
<td>3.482</td>
<td>0.001</td>
<td>72.35604</td>
<td>260.3376</td>
</tr>
<tr>
<td>Ivitus_1</td>
<td>33.46968</td>
<td>29.28935</td>
<td>1.143</td>
<td>0.254</td>
<td>-24.16284</td>
<td>91.10221</td>
</tr>
<tr>
<td>Ivitus_2</td>
<td>39.49589</td>
<td>31.87656</td>
<td>1.239</td>
<td>0.216</td>
<td>-23.22748</td>
<td>102.2193</td>
</tr>
<tr>
<td>IsXv_1_1</td>
<td>-11.72721</td>
<td>76.51943</td>
<td>-0.153</td>
<td>0.878</td>
<td>-162.2942</td>
<td>138.8398</td>
</tr>
<tr>
<td>IsXv_1_2</td>
<td>-255.6611</td>
<td>105.4603</td>
<td>-2.424</td>
<td>0.016</td>
<td>-463.175</td>
<td>-48.14725</td>
</tr>
<tr>
<td>_cons</td>
<td>563.2184</td>
<td>21.84213</td>
<td>25.786</td>
<td>0.000</td>
<td>520.2397</td>
<td>606.1971</td>
</tr>
</tbody>
</table>

(Model is ordinary regression, use regress instead)

where _Isex_1 is $X_1$ (males vs females | no vitamin use), _Ivituse_1 is $Z_1$ (“fairly often” – frequent vitamin users vs non users | gender=female) and _Ivituse_2 is $Z_2$ (“not often” – infrequent vitamin users vs non users | gender=female). The interactions are _IsXv_1_1, $X_1Z_1$ [Gender effect* | frequent vitamin use vs Gender effect | no vitamin use] and _IsXv_1_2, $X_1Z_2$ [Gender effect | infrequent vitamin use vs Gender effect | no vitamin use]

*Mean difference in retinol plasma levels (male-female)
The model

\[ E[Y] = b_0 + b_1 \text{Male} + b_2 \text{Freq\_Use} + b_3 \text{Infreq\_Use} + b_4 \text{Male} \times \text{Freq\_Use} + b_5 \text{Male} \times \text{Infreq\_Use} \]

<table>
<thead>
<tr>
<th>Gender</th>
<th>Vitamin Use</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>Frequent</td>
<td>Infrequent</td>
<td>No Use</td>
</tr>
<tr>
<td>b_0 + b_1 + b_2 + b_4</td>
<td>b_0 + b_1 + b_3 + b_5</td>
<td>b_0 + b_1</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>b_0 + b_2</td>
<td>b_0 + b_3</td>
<td>b_0</td>
</tr>
<tr>
<td>Difference (Male-Female)</td>
<td>b_1 + b_4</td>
<td>b_1 + b_5</td>
<td>b_1</td>
</tr>
</tbody>
</table>
Estimates of model parameters

Constant :  \( \_\text{cons}=\hat{\beta}_0=563.2184 \)

Main effects :  \( \_\text{Isex}_1=\hat{\beta}_1=166.3468, \_\text{Ivituse}_1=\hat{\beta}_2=33.46968, \_\text{Ivituse}_2=\hat{\beta}_3=39.49589 \)

Interactions :  \( \_\text{IsXv}_1\_1 =\hat{\beta}_4 =-11.72721, \_\text{IsXv}_1\_2 =\hat{\beta}_5 =-255.6611 \)

The estimates of the various parameters are given as follows:

1. Females
   a. Frequent users (“fairly often”): 563.2184+33.46968=\textbf{596.68808}
   b. Infrequent users (“not often”): 563.2184+39.49589=\textbf{602.71429}
   c. Non-users: \textbf{563.2184}

2. Males
   a. Frequent users: 563.2184+166.3468+33.46968+(-11.72721)=\textbf{751.30767}
   b. Infrequent users: 563.2184+166.3468+39.49589+(-255.6611)=\textbf{513.39999}
   c. Non-users: 563.2184+166.3468 =\textbf{729.5652}
The descriptive statistics of the plasma retinol levels by gender and vitamin use are given in the STATA output below:

```
.tabulate sex vituse, summarize(retplasm)
```

<table>
<thead>
<tr>
<th>Sex</th>
<th>Frequent</th>
<th>Infrequent</th>
<th>No Use</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Males</td>
<td>751.30769</td>
<td>513.4</td>
<td>729.56522</td>
<td>710.09756</td>
</tr>
<tr>
<td></td>
<td>329.43269</td>
<td>298.59303</td>
<td>290.0285</td>
<td>305.52208</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>5</td>
<td>23</td>
<td>41</td>
</tr>
<tr>
<td>Females</td>
<td>596.68807</td>
<td>602.71429</td>
<td>563.21839</td>
<td>587.72161</td>
</tr>
<tr>
<td></td>
<td>203.71816</td>
<td>184.6959</td>
<td>159.92785</td>
<td>185.43069</td>
</tr>
<tr>
<td></td>
<td>109</td>
<td>77</td>
<td>87</td>
<td>273</td>
</tr>
<tr>
<td>Total</td>
<td>613.16393</td>
<td>597.26829</td>
<td>598</td>
<td>603.70064</td>
</tr>
<tr>
<td></td>
<td>223.83038</td>
<td>192.02109</td>
<td>204.39088</td>
<td>208.60239</td>
</tr>
<tr>
<td></td>
<td>122</td>
<td>82</td>
<td>110</td>
<td>314</td>
</tr>
</tbody>
</table>
```
The graphical representation of the data above is given as follows:

We see that there is a significant interaction caused by an unexpected low plasma level of retinol among men that used vitamins infrequently.
SEs and 95% CI for linear combination of the estimates in STATA

In Stata we can estimate the mean, the SE and the 95% CI of a linear combination of the parameters by using the command lincom after the model fit.

For example, to estimate mean plasma of retinol in men with frequent vitamin use we need to estimate the combination: \( \text{Comb1} = b_{\text{men}} + b_{\text{freq}} + b_{\text{men*freq}} + \_\text{cons} \).

If the model has been defined as: \text{reg retplasm men freq infeq menfreq meninfr} \\
We can get the estimate, the SE and the 95% CI of the parameter Comb1 in Stata by the command \text{Lincom men+freq+menfreq+_cons} \\
After fitting the model.
General method to estimate the mean and the variance of a linear combination of the estimates

In general, we can estimate the mean and the variance of the linear combination such as Comb1 following the steps:

1) Define constraints: \( C=(1,1,0,1,0,1) \). This constraint asks for the sum of those parameters indicated by 1s in \( C \). Stata command: matrix \( C=(1,1,0,1,0,1) \)

2) Estimate the linear combination as: \( C\boldsymbol{b}' \) where \( \boldsymbol{b} \) the 1xp vector of bs. Applying that in our example we will get the mean level of plasma retinol for men with frequent vitamin use. Stata command: mat estcomb1=C*A. \( A \) is the vector of the estimates obtaines as: mat A=e(b).

3) Estimate the variance of the linear combination as: \( C\boldsymbol{V(b)}\boldsymbol{C}' \) where \( \boldsymbol{V(b)} \) the variance-covariance matrix of the estimates. This will produce the variance of the combination. Stata command: varcomb1=C*B*C’. \( B \) is the variance-covariance matrix of the parameters: mat B=e(V).
The estimates and the 95% CI of the parameters in the retinol example

<table>
<thead>
<tr>
<th></th>
<th>Frequent vitamin user</th>
<th>Infrequent vitamin use</th>
<th>Not use</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate (95% CI)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>751.31 (640 – 862)</td>
<td>513.40 (334 – 693)</td>
<td>729.57 (646 – 813)</td>
</tr>
<tr>
<td>Female</td>
<td>596.69 (558 – 635)</td>
<td>602.71 (557 – 648)</td>
<td>563.22 (520 – 606)</td>
</tr>
</tbody>
</table>

Note that for females all three estimates are consistent while for males those with infrequent use have significantly lower mean level of plasma retinol. Apart from men with infrequent vitamin use, women have on average lower levels of plasma retinol.
Regression models for the analysis of covariance

The analysis of covariance can also be expressed in terms of a linear regression by reparametrizing the fixed effect in the usual way. The complete ANACOVA model (including interaction) is as follows:

\[ y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ + \epsilon \]

where \( X \) and \( Z \) may be vector-valued.

For example, consider the effect of gender and age on plasma retinol levels. We code the gender variable as before, i.e., \( X_1 = \begin{cases} 1, \text{male} \\ 0, \text{otherwise} \end{cases} \) and \( Z = \text{age} \) and \( XZ \) is the age/gender interaction.

With this parametrization, the model for the males and females are:

Males: \[ y_M = (\beta_0 + \beta_1) + (\beta_2 + \beta_3)Z + \epsilon \]

\[ \beta_{0M} \]
\[ \beta_{1M} \]

Females: \[ y_F = \beta_0 + \beta_2 Z + \epsilon \]
The test of parallelism

From the parametrization of the ANACOVA model we see that the effect of gender impacts the intercept of the line, while the interaction term affects the slope.

If there is no interaction (i.e., if $\beta_3 = 0$), the two lines are parallel (or they coincide if $\beta_2 = 0$). Thus, testing the null hypothesis $H_0 : \beta_3 = 0$ is equivalent to testing whether the lines formed by the regression of plasma retinol levels by age in the two genders are parallel.
Consider the following output:

```
.xi: glm retplasm i.sex*age
i.sex       Isex_1-2   (naturally coded; Isex_2 omitted)
i.sex*age   IsXage_#  (coded as above)
```

Iteration 1 : deviance = 12658374.0533

Residual df = 310
Pearson X2 = 1.27e+07
Dispersion = 40833.46

Gaussian (normal) distribution, identity link

| Coef. | Std. Err. | t     | P>|t|   | [95% Conf. Interval] |
|--------|-----------|-------|-------|-------------------------------|
| age    | 2.810887  | .8693928 | 3.233 | 0.001 | 1.10023, 4.521545 |
| Isex_1 | 235.3007  | 151.7706 | 1.550 | 0.122 | -63.33017, 533.9316 |
| IsXage_1| -2.421536 | 2.502083 | -0.968 | 0.334 | -7.344749, 2.501676 |
| _cons  | 451.2649  | 43.94161 | 10.270 | 0.000 | 364.8033, 537.7264 |

(Model is ordinary regression, use regress instead)
Test of parallelism (continued)

From the STATA output above we have that there is no significant interaction between gender and age. This is obtained from the p-value of the z test for $H_0 : \beta_3 = 0$, which is 0.334. Thus, the data do not contradict the assumption of parallelism. This is shown graphically in the following figure:
A more parsimonious model is as follows:

```stata
. char sex[omit] 2
.
xi: glm retplasm i.sex age
i.sex                 Isex_1-2     (naturally coded; Isex_2 omitted)

Iteration 1 : deviance = 12696620.8404

Residual df = 311          No. of obs = 314
Pearson X2 = 1.27e+07       Deviance = 1.27e+07
Dispersion = 40825.15       Dispersion = 40825.15

Gaussian (normal) distribution, identity link
------------------------------------------------------------------------------
retplasm |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+---------------------------------------------------------------------
  Isex_1  |   92.42252   35.20318      2.625   0.009       23.15599     161.689
  age     |   2.518526   .8151396      3.090   0.002       .9146404    4.122412
  _cons   |   465.4578   41.41804     11.238   0.000       383.9628    546.9528
------------------------------------------------------------------------------
(Model is ordinary regression, use regress instead)

Which leads to a significant gender effect (p value=0.009) at the 5% level.
```
Model selection

To motivate model selection in the generalized linear model, I present the mechanics of model selection in the linear model.

Consider the process of starting with a “full” model in the sense that it is a model containing all variables that we are willing to consider. Then the criterion of removing a variable is based on an $F$ test as follows (here we consider $p$ variables plus the intercept in all models):

$$\frac{SSE(X_{p_1}) - SSE(X_{p_2})}{SSE(X_{p_2})/(n - p_2 - 1)} \sim F_{1,n-p_2-1}$$

where $SSE(X_{p_1})$ and $SSE(X_{p_2})$ are the residual sum of squares of the full model and the sub-model respectively.
**Example:** Plasma retinol levels (continued). The output from the full model is as follows:

```
.xi: reg retplasm age i.sex i.smokstat quetelet i.vituse calories fat fiber alcohol chol
i.sex                          _Isex_1-2 (naturally coded; _Isex_2 omitted)
i.smokstat                     _Ismokstat_1-3 (naturally coded; _Ismokstat_1 omitted)
i.vituse                       _Ivituse_1-3 (naturally coded; _Ivituse_3 omitted)

Source | SS     df       MS
-------------------------------+------------------
Model | 1896984.44 12  158082.037  F(12, 301) = 4.06
Residual | 11723197.4 301  38947.4997 Prob > F = 0.0000
R-squared = 0.1393
Adj R-squared = 0.1050
Total | 13620181.9 313  43514.958  Root MSE = 197.35

| retplasm | Coef.   Std. Err. | t    | P>|t| | [95% Conf. Interval] |
|----------|-------------------|------|------|----------------------------------|
| age      | 2.653472          .8756372 | 3.03 | 0.003 | .9303267 4.376618 |
| _Isex_1  | 76.8363           37.37679 | 2.06 | 0.041 | 3.283403 150.3892 |
| _Ismokstat_1 | 44.90691 25.13723 | 1.79 | 0.075 | -4.560058 94.37388 |
| _Ismokstat_2 | -65.74155 36.25566 | -0.22 | 0.866 | -72.00408 70.68925 |
| _Ismokstat_3 | 1.581298 1.917623 | 0.82 | 0.410 | -2.192347 5.354944 |
| quetelet | 35.40501          27.26527 | 1.30 | 0.195 | -18.24968 89.05969 |
| _Ivituse_1 | 27.8062 29.71094 | 0.94 | 0.350 | -30.66125 86.27365 |
| _Ivituse_2 | 0.758574 .598645 | 1.27 | 0.206 | -.0419486 .1936634 |
| calories  | 1.512089          .9335381 | 1.62 | 0.106 | -3.349177 .3249986 |
| fiber     | -4.207861         3.100573 | -1.36 | 0.176 | -10.30941 1.893684 |
| alcohol   | 7.371856          2.602759 | 2.83 | 0.005 | 2.249949 12.49376 |
| chol      | -0.775529         104.8078 | -0.74 | 0.460 | -2.838016 .1286959 |
| _cons     | 416.1679          83.85834 | 4.96 | 0.000 | 251.145 581.1907 |
```
While the output from the model excluding cholesterol levels is,

```
.xi: reg retplasm age i.sex i.smokstat quetelet i.vituse calories fat fiber alcohol
i.sex             _Isex_1-2           (naturally coded; _Isex_2 omitted)
i.smokstat        _Ismokstat_1-3     (naturally coded; _Ismokstat_1 omitted)
i.vituse          _Ivituse_1-3       (naturally coded; _Ivituse_3 omitted)

Source | SS     df    MS                    Number of obs =  314
-------------+---------------------------------------------
Model | 1875659.49 11  170514.499             Prob > F     =  0.0000
Residual | 11744522.4 302  38889.1469           R-squared     =  0.1377
-------------+---------------------------------------------
Total | 13620181.9 313   43514.958           Adj R-squared =  0.1063

                      | Coef.     Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+---------------------------------------
retplasm |                          |     |           |               |
age | 2.678755   .8743146     3.06   0.002     .9582353    4.399275
_Isex_1 | 72.77019   36.94293     1.97   0.050     .0720315    145.4683
_Ismokstat_2 | 46.0355   25.07212    1.84   0.067    -3.302663    95.37367
_Ismokstat_3 | .1518775   36.212   0.00   0.997    -71.10792    71.41168
quetelet | 1.536417   1.915227    0.80   0.423    -2.232463    5.305297
_Ivituse_1 | 36.63589   27.19409    1.35   0.179   -16.87799    90.14978
_Ivituse_2 | 28.56312   29.67107    0.96   0.336   -29.82509    86.95134
calories | .0674277   .0587265    1.15   0.252    -0.0481373   .1829927
fat | -1.592929   .9264287   -1.72   0.087    -3.416001   .2301444
fiber | -3.812586   3.051921   -1.25   0.213    -9.818308   2.193137
alcohol | 7.534269   2.591544     2.91   0.004     2.434499    12.63404
    _cons | 412.199    83.62392     4.93   0.000   247.6397    576.7583
```
The criterion for removing cholesterol level from consideration is

\[
\frac{SSE(X_{p_2}) - SSE(X_{p_1})}{SSE(X_{p_2})/(n - p_2 - 1)} = \frac{11744522.4 - 11723197.4}{11723197.4/301} = 0.5475
\]

This can also be given by using the `test` command after regression on the full model as follows:

```
. quietly xi: reg  retplasm age i.sex i.smokstat quetelet i.vituse calories fat fiber alco > hol chol
. test chol
   ( 1)  chol = 0.0

   F(  1,   301) =  0.55
   Prob > F =  0.4599
```

and is equivalent to the \(t\) test listed in the output of the full model above (recall that an \(F\) test with 1 degree of freedom in the numerator is equal to the square of the \(t\) test with equal degrees of freedom as in the denominator of the \(F\) test).
Model selection in the GLM

A similar concept as the residual sums of squares in the GLM is the deviance. In addition, the log-likelihood can be used in the derivation of likelihood-ratio tests. We consider these two concepts here.

The likelihood ratio $\lambda$ is the fraction of the maximized likelihood of the sub-model and the full model respectively. For large samples, $-2\log\lambda \sim \chi^2_\nu$ where $\nu$ is the difference in the dimension of the two models. For two models that are different by a single variable, $\nu=1$ of course.

In general, the likelihood-ratio criterion is $-2\log\lambda = \frac{D(X_{p_1}) - D(X_{p_2})}{\phi}$, where $\phi$ is a scale parameter, and $p_1 < p_2$. In particular, in the linear model this is:

$$-2\log\lambda = \frac{SSE(X_{p_1}) - SSE(X_{p_2})}{SSE(X_{p_2})/(n-p_2-1)}$$
Example: Plasma retinol levels (continued)

In our example, we can derive the likelihood-ratio test as follows:

```stata
xi: glm retplasm i.sex age i.smokstat i.vituse quetelet calories fat fiber alcohol cho
> l
i.sex             _Isex_1-2           (naturally coded; _Isex_2 omitted)
i.smokstat        _Ismokstat_1-3      (naturally coded; _Ismokstat_1 omitted)
i.vituse          _Ivituse_1-3         (naturally coded; _Ivituse_3 omitted)
Iteration 0:   log likelihood = -2098.3936
Generalized linear models                          No. of obs =       314
Optimization     : ML: Newton-Raphson
Residual df     =       301
Scale param     =   38947.5
Deviance         =  11723197.42                    (1/df) Deviance =  38947.5
Pearson          =  11723197.42                    (1/df) Pearson =  38947.5
Variance function: V(u) = 1                        [Gaussian]
Link function    : g(u) = u                        [Identity]
Standard errors  : OIM
Log likelihood   = -2098.39358                    AIC             =  13.44837
BIC             =  11723122.68

+---------------------------------------------+----------+-----------------+-----------------+
|            |   Coef.   |     Std. Err.   |      z       |     P>|z|     |  [95% Conf. Interval] |
|---------------------------------------------+----------+-----------------+-----------------+
|retplasm| -76.8363  |   37.37679      |  2.06         |  0.040     | 3.579146 150.0935   |
|age    |  2.653472 |   0.8756372     |  3.03         |  0.002     |  0.937255 4.36969   |
|I_smokstat_1 |  44.90691 |   25.13723      |  1.79         |  0.074     | -4.36116 94.17499   |
|I_vituse_1 | -6574155  |   36.25566      | -18.03395    | 88.84396   |
|I_vituse_2 |  35.40501 |   27.26527      |  1.30         |  0.194     | -30.42617 86.03856  |
|quetelet |  1.581298 |   1.917623      |  0.82         |  0.410     | -2.177174 5.33977   |
|calories |  0.0758574|   0.0598645     |  1.27         |  0.205     | -0.041475 0.193897  |
|fat    | -1.512089 |   0.935381      | -1.62         |  0.105     | -3.34179 0.317612   |
|fiber  | -4.207861 |   3.100573      | -1.36         |  0.175     | -10.28487 1.869151  |
|alcohol|  7.371856  |   2.602759      |  2.83         |  0.005     |  2.270543 12.47317  |
|cholesterol | -0.0775529|   0.1048078     | -0.74         |  0.459     | -0.282972 0.1278666 |
|cons   |  416.1679  |   83.85834      |  4.96         |  0.000     |  251.8086 580.5272  |
|---------------------------------------------+----------+-----------------+-----------------|
```
Example: Plasma retinol levels (continued)

In our example, we can derive the likelihood-ratio test as follows:

```
xi: glm retplasm i.sex age i.smokstat i.vituse quetelet calories fat fiber alcohol cho
>i
```

```
i.sex             _Isex_1-2           (naturally coded; _Isex_2 omitted)
i.smokstat        _Ismokstat_1-3      (naturally coded; _Ismokstat_1 omitted)
i.vituse          _Ivituse_1-3        (naturally coded; _Ivituse_3 omitted)
```

Iteration 0: log likelihood = -2098.3936

```
Generalized linear models                     No. of obs = 314
Optimization : ML: Newton-Raphson             Residual df = 301
Scale param = 38947.5                        (1/df) Deviance = 38947.5
Deviance = 11723197.42                       (1/df) Pearson = 38947.5
Pearson = 11723197.42
```

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

```
Log likelihood = -2098.39358                  AIC = 13.44837
BIC = 11723122.68
```

```
|     | Coef.   | Std. Err. | z     | P>|z|     | [95% Conf. Interval] |
|-----|---------|-----------|-------|--------|---------------------|
| _Isex_1 | 76.8363 | 37.37679 | 2.06  | 0.040  | 3.579146            | 150.0935 |
| age     | 2.653472 | .8756372 | 3.03  | 0.002  | .9372552            | 4.36969 |
| _Ismokstat_2 | 44.90691 | 25.13723 | 1.79  | 0.074  | -4.36116            | 94.17499 |
| _Ismokstat_3 | -6574155 | 36.25566 | -1.82 | 0.074  | -71.7121           | 70.40238 |
| _Ivituse_1 | 35.40501 | 27.26527 | 1.30  | 0.194  | -18.03395           | 88.84396 |
| _Ivituse_2 | 27.8062  | 29.71094 | 0.94  | 0.349  | -30.42617           | 86.03856 |
| quetelet | 1.581298 | 1.917623 | 0.82  | 0.410  | -2.177174           | 5.33977 |
| calories | .0758574 | .0598645 | 1.27  | 0.205  | -.041475            | .1931897 |
| fat      | -1.512089 | .9335381 | -1.62 | 0.105  | -3.34179            | .317612 |
| fiber    | 4.207861  | 3.100573 | 1.36  | 0.175  | -10.28487           | 1.869151 |
| alcohol  | 7.371856  | 2.602759 | 2.83  | 0.005  | 2.270543            | 12.47317 |
| chol     | -.0775529 | .1048078 | -0.74 | 0.459  | -.2829723           | .1278666 |
```
The likelihood-ratio test can be constructed as follows:

\[-2\log \lambda = \frac{SSE(X_{p_1}) - SSE(X_{p_2})}{SSE(X_{p+1})/(n - p - 1)} = \frac{11744522.37 - 11723122.68}{11723122.68/301} = 0.5477\]

Its asymptotic (long-term) distribution is a chi-square with one degree of freedom.

```
di chiprob(1, (11744522.38 - 11723122.68)/((11723122.68)/(301)))
.45926647
```

which is similar to the results of the $F$ test previously. Notice that we get the same results if we subtract the maximized log-likelihoods as follows:

\[-2\log \lambda = -2\left|2098.39358 - 2098.67891\right| = 0.57066\]

with asymptotic distribution that is also chi-square with one degree of freedom.

```
di chiprob(1, -2*(2098.39358 - 2098.67891))
.44999677
```
Wald tests

The easiest way to assess the impact of the factor cholesterol in the model is with the `test` command, which generates the Wald test described previously.

```
. quietly xi: glm retplasm i.sex age i.smokstat i.vituse quetelet calories fat fiber alco
   > hol chol
```

In STATA 7.0, this is given by

```
. test chol
( 1) [retplasm]chol = 0.0
    chi2( 1) = 0.55
    Prob > chi2 = 0.4593
```

In STATA 6.0, we can derive the chi-square (Wald) test as follows:

```
. di chiprob(1,( -.0775529/.1048078)^2)
   0.4593282
```
Finally, we show here the model-selection for the complete problem.

```
.xi: sw glm retplasm i.sex (i.smokstat) (i.vituse) age quetelet calories fat > fiber alcohol chol, pr(.1)
i.sex             Isex_1-2    (naturally coded; Isex_2 omitted)
i.smokstat        Ismoks_1-3  (naturally coded; Ismoks_1 omitted)
i.vituse          Ivitus_1-3   (naturally coded; Ivitus_3 omitted)

begin with full model
p = 0.4599 >= 0.1000  removing chol
p = 0.4231 >= 0.1000  removing quetelet
p = 0.4163 >= 0.1000  removing Ivitus_1 Ivitus_2
p = 0.1572 >= 0.1000  removing Ismoks_2 Ismoks_3
p = 0.1806 >= 0.1000  removing fiber
p = 0.5284 >= 0.1000  removing calories

Residual df =     309                                      No. of obs =       314
Pearson X2   = 1.21e+07                                    Deviance   = 1.21e+07
Dispersion   = 39055.6                                    Dispersion = 39055.6

Gaussian (normal) distribution, identity link
---------------------------------------------------------------------
retplasm |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interv
---------|--------------------------------------------|---------------------
Isex_1 |     74.055   36.44476      2.032   0.043       2.343714    145.7663
fat    |   -0.6188433   .3501419    -1.767   0.078      -1.307807    .0701208
alcohol|     8.724091   2.340494      3.727   0.000       4.11877    13.32941
age    |     2.389427   .8229901     2.903   0.004      .7700534    4.008801
_cons   |    498.7073    54.22216     9.197   0.000      392.0159    605.3986
---------------------------------------------------------------------
(Model is ordinary regression, use regress instead)
```
Pearson Residuals

The Pearson residuals are defined as

\[ r_{i,p} = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}} \]

and it is the raw residual scaled by the estimated standard deviation of \( Y \). The name is taken from the fact that for the Poisson distribution the Pearson residual is just the signed square root of the component of the Pearson \( X^2 \) goodness-of-fit statistic, i.e.

\[ \sum_{i=1}^{n} r_{i,p}^2 = X^2 \]

A disadvantage of the Pearson residual is that the distribution of \( r_{i,p} \) for non-normal distributions is markedly skewed, and it may fail to have properties similar to those of a normal-theory residual.
Deviance Residuals

If the deviance is used as a measure of discrepancy of a generalized linear model, then each unit contributes a quantity \( d_i \) to that measure, so

\[
\sum d_i = D
\]

Thus, if we define

\[
r_{i,D} = \text{sign}(y - \mu) \sqrt{d_i}
\]

we have a quantity that increases with \( y_i - \mu_i \) and for which \( \sum r_{i,D}^2 = D \).
Residuals- Linear regression

Recall that variance of the true residuals is assumed to be constant. The variance of the fitted (observed) residuals is NOT constant, since there is variance in estimation of the line and of the expected values. Therefore, for model checking we need to standardize the observed residuals.

Lets explore it in normal regression

\[ E(Y|X = x) = X\beta \]

Design matrix

\[ H = X(X^T X)^{-1} X^T \]

Hat matrix

\[ h_{ii} = h_i = x_i^T (X^T X)^{-1} x_i \]

i\(^{th}\) leverage

\[ e = y - \hat{y} \]

Residuals and

\[ Var(e) = (I - H)\sigma^2 \]
Residuals- Linear regression (continue)

standardized residual
\[ r_i = \frac{e_i}{S \sqrt{1 - h_i}} \]

here \( e_i, S \) are not independent since \( e_i \) enters in the calculation of \( S \).

sample variance

studentized residuals
\[ r_i^* = \frac{e_i}{S(i) \sqrt{1 - h_i}} \]

here numerator & denominator are independent

sample variance with \( i \)th observation omitted

The distribution of \( r_i^* \sim t_{n-p-1} \)

The \( i \)th leverage is large if \( h_i \geq 2 p' / n \) where \( p' = \) total # of covariates in the model including intercept, \( n = \) total # of observations.
Standardized residuals in GLMs

The key quantities for GLM diagnostics are:

- Pearson residuals
- Deviance

The general definition of standardized residuals is:

\[
\begin{align*}
    r'_p &= \frac{y - \hat{\mu}}{\sqrt{\hat{\phi} V(\hat{\mu})(1 - h)}} \\
    r'_D &= \frac{r'_D}{\sqrt{\hat{\phi}(1 - h)}}
\end{align*}
\]

(standardized deviance residual)
Model checking

The predicted values and the residuals from the optimal model (the one including gender, fat and alcohol intake and age) are produced by STATA commands as follows:

```
. quietly reg retplasm sex fat alcohol age
. predict yhat
   (option xb assumed; fitted values)
. predict r, resid
. predict rstan, rstand
. predict rstud, rstud
```
Model checking: residuals

The assumptions of the model that must be checked are independence, normality and homoskedasticity. We usually work with the standardized residuals \( r_{std,i} = \frac{r_i}{\hat{\sigma} \sqrt{1-h_{ii}}} \) (produced with the option \texttt{rstan}) or the studentized residuals \( r_{stud,i} = \frac{r_i}{\hat{\sigma}_i \sqrt{1-h_{ii}}} \) (with option \texttt{rstud}), where \( \hat{\sigma} \) is an estimate of the standard deviation derived with all the observations, while and \( \hat{\sigma}_i \) is the estimate with the \( i^{th} \) observation missing. On the other hand, \( h_{ii} \) is the \( i^{th} \) diagonal element of the hat matrix (recall that in regression \( \hat{y} = X(X'X)^{-1}X'y = Hy \), where \( H \) is the “hat” matrix). The leverage points are a measure of distance (outlier, potential influential point). We use the Cook’s distances as a combined measure of influence and distance since they are \( D_i = \frac{r_{std,i}^2 h_{ii}}{(1-h_{ii})} \).
Homoskedasticity

This refers to the homogeneity of variance. We can see what the stud. residuals look like as follows:

```
    . graph r yhat, yline(0) xlab ylab border
```

We see that there is no obvious problem with lack of homoskedasticity in these data.
Normality

The next assumption of the general linear model is that of normality of the residuals. This can be checked using the `qnorm` command in STATA as follows:
Q-Q plots

These are plots that compare the distribution of a variable to a known distribution. They can be used alternatively to compare the distributions of two variables. In general, if the distributions are approximately equal the points on the graph should lie on a straight line.

In the plot above, we see that there are problems with the distribution of retinol levels at the “tails” which are shorter for small values and “fatter” for larger values.

This Q-Q plot can be produced manually following these steps:

- Sort the residuals from smaller to largest (we are still working with studentized residuals)
- Imagining that the residuals have a normal distribution then their ranked values should be close to standard normal distribution percentiles, that is, $z_{(i)} = \Phi^{-1}\left(\frac{i}{n+1}\right), i=1,\ldots,n$ (this is actually the way STATA produces a q-q plot).
Q-Q plot (continued)

To create the q-q plot above manually we proceed as follows:

. sort rstud
. gen zi=invnorm(_n/(_N+1))
. label var zi "Inverse Normal"
. graph rstud zi zi, xlab ylab c(.l) s(o) rlab yline xline
The Shapiro-Wilks test of normality

To formally test the hypothesis of normality, we can use the Shapiro-Wilks test as follows:

\[ . \text{swilk rstud} \]

| Variable | Obs | W     | V     | z     | Prob>|z |
|----------|-----|-------|-------|-------|------|
| rstud    | 314 | 0.93618 | 14.159 | 6.235 | 0.0000 |

The test p value is 0.000 < 0.05 which means that the normality assumption is not fulfilled.
In order to find which transformation to use, a general method is that of Box and Cox. The general Box-Cox transformation is as follows:

\[
y^* = \begin{cases} 
  \frac{y^{\lambda} - 1}{\lambda}, & y \neq 0 \\
  \log(y), & y = 0
\end{cases}
\]

Several possible choices of \( \lambda \) are tried. The best choice is given through a likelihood criterion.

Some usual transformations are given as follows:

- \( \lambda = -1 \) Inverse transformation
- \( \lambda = 1 \) No transformation is necessary
- \( \lambda = 0.5 \) Square-root transformation
- \( \lambda = 0 \) Logarithmic transformation
Box-cox transformation

To implement the Box-Cox technique in STATA we proceed as follows:

```
. boxcox retplasm, lstart(-1) graph generate (newret)
```

(note: iterations performed using zero = .001)

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Lambda</th>
<th>Zero</th>
<th>Variance</th>
<th>LL</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-1.0000</td>
<td>89.67819</td>
<td>51344.1853</td>
<td>-1702.87019</td>
</tr>
<tr>
<td>1</td>
<td>0.1327</td>
<td>2.56310</td>
<td>37070.2753</td>
<td>-1651.72960</td>
</tr>
<tr>
<td>2</td>
<td>0.1676</td>
<td>0.00243</td>
<td>37062.7391</td>
<td>-1651.69768</td>
</tr>
<tr>
<td>3</td>
<td>0.1676</td>
<td>0.00000</td>
<td>37062.7420</td>
<td>-1651.69770</td>
</tr>
</tbody>
</table>

Transform: \( \frac{(retplasm^{L-1})}{L} \)

<table>
<thead>
<tr>
<th>L</th>
<th>[95% Conf. Interval]</th>
<th>Log Likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1676</td>
<td>(not calculated)</td>
<td>-1651.6977</td>
</tr>
</tbody>
</table>

Test:  
- \( L = -1 \)  \( \chi^2(1) = 104.14 \)  \( Pr>\chi^2 = 0.0000 \)
- \( L = 0 \)   \( \chi^2(1) = 2.19 \)   \( Pr>\chi^2 = 0.1387 \)
- \( L = 1 \)   \( \chi^2(1) = 49.39 \)   \( Pr>\chi^2 = 0.0000 \)
A value of zero for lamda is not unreasonable, suggesting a logarithmic transformation of (retplasm). The new variable newret contains the transformed values of retplasm (with lamda = 0.1676).
Further model checking

Running the model `newret` with as the dependent variable we have:

```stata
. xi: sw reg newret age i.sex (i.smokstat) quetelet (i.vituse) calories fat fi > ber alcohol betadiet retdiet, pr(.1)
```

i.sex                 Isex_1-2 (naturally coded; Isex_2 omitted)
i.smokstat            Ismoks_1-3 (naturally coded; Ismoks_1 omitted)
i.vituse              Ivitus_1-3 (naturally coded; Ivitus_1 omitted)

begin with full model

\[
p = 0.6699 \geq 0.1000 \text{ removing retdiet}
\]
\[
p = 0.6327 \geq 0.1000 \text{ removing quetelet}
\]
\[
p = 0.5945 \geq 0.1000 \text{ removing Ivitus_2 Ivitus_3}
\]
\[
p = 0.5018 \geq 0.1000 \text{ removing betadiet}
\]
\[
p = 0.2852 \geq 0.1000 \text{ removing Isex_1}
\]
\[
p = 0.1146 \geq 0.1000 \text{ removing Ismoks_2 Ismoks_3}
\]

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 314</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>34.2403962</td>
<td>5</td>
<td>6.84807924</td>
<td>F(  5,   308) = 7.93</td>
</tr>
<tr>
<td>Residual</td>
<td>265.811233</td>
<td>308</td>
<td>.863023485</td>
<td>Prob &gt; F   = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>300.05163</td>
<td>313</td>
<td>.958631405</td>
<td>Adj R-squared  = 0.0997</td>
</tr>
</tbody>
</table>

| newret | Coef. | Std. Err. | t    | P>|t| | [95% Conf. Interval] |
|--------|-------|-----------|------|------|----------------------|
| age    | .0161012 | .0038357  | 4.198 | 0.000 | .0085538 .. .0236487 |
| fat    | -.0092023 | .0042933  | -2.143 | 0.033 | -.0176502 .. -.0007544 |
| calories | .0004638 | .0002721  | 1.704 | 0.089 | -.0000716 .. .0009992 |
| fiber  | -.0235039 | .0140948  | -1.668 | 0.096 | -.0512381 .. .0042303 |
| alcohol | .0368435 | .0115156  | 3.199 | 0.002 | .0141843 .. .0595027 |
| _cons | 10.6249 | .2747965  | 38.665 | 0.000 | 10.08418 .. 11.16561 |
Checks for outliers and influential observations

We produce residuals, leverage values and Cook’s distances as follows:

```
predict rstud, rstud
.predict d,cooksd
.predict h, hat
```

A studentized residual greater than 2 in absolute value, a leverage greater than $2p/n=0.0382$, where $p$ is the number of predictors plus the intercept, and a Cook’s distance of 1 or higher are indicative of an outlier or of excessive influence, or both respectively.

```
.list rstud d h if abs(rstud)>2.0 | h>.0382

 rstud          d          h
  1.  -2.880156   .0283224   .0205401
  2.  -3.392216   .0174023   .0092961
   .
  313.  3.294111   .0123143   .0069778
  314.  3.587223   .0196182   .0094104
(37 cases)
.list rstud d h if abs(rstud)>2.0 & h>.0382
( 0 cases)
```
Model checking (continued)

To summarize the Cook’s distances we proceed as follows:

```
. summarize d

Variable | Obs | Mean   | Std. Dev. | Min   | Max
---------|-----|--------|-----------|-------|-----
       d | 314 | .0031528 | .0058947 | 9.35e-11 | .0390723
```

There are no observations with Cook’s distance above 1, although there are several points with large residuals or leverage. However, the number of points that we are testing for large residuals is so large, that the criterion of 2.0 or higher is probably very liberal (as 314 repeated tests are being conducted!). Thus, the fit is probably acceptable.