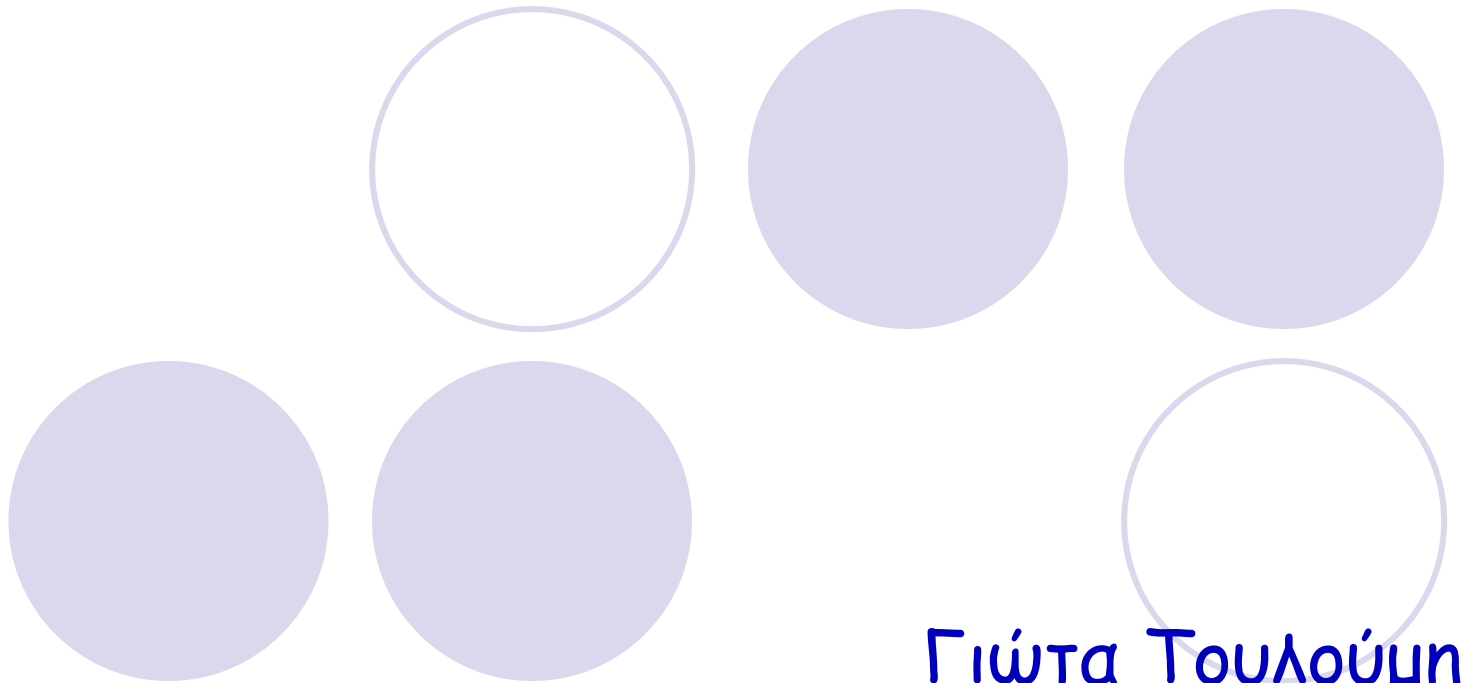


# GENERALIZED LINEAR MODELS



Γιώτα Τουλούμη

Καθηγήτρια Βιοστατιστικής και Επιδημιολογίας  
Εργ. Υγιεινής, Επιδημιολογίας και Ιατρικής Στατιστικής  
Ιατρική Σχολή Πανεπιστημίου Αθήνας

[gtouloum@med.uoa.gr](mailto:gtouloum@med.uoa.gr)

# Notes

$$\text{Fisher Information Matrix } I(\theta) = -\frac{\partial^2}{\partial \theta_i \theta_j} l(\theta) \quad 1 \leq i, j \leq p$$

Observed Fisher Information Matrix:  $I(\hat{\theta}_{ML})$

$$\text{Hessian } H(\theta) = \frac{\partial^2}{\partial \theta_i \theta_j} l(\theta) \quad 1 \leq i, j \leq p$$

$$\text{Var}(\hat{\theta}_{ML}) = [I(\hat{\theta}_{ML})]^{-1}$$

$$\text{SE}(\hat{\theta}_{ML}) = \frac{1}{\sqrt{I(\hat{\theta}_{ML})}}$$

$$H(\beta) = \begin{pmatrix} \frac{\partial^2 l}{\partial \beta_0^2} & \frac{\partial^2 l}{\partial \beta_0 \beta_1} & \frac{\partial^2 l}{\partial \beta_0 \beta_2} \\ \frac{\partial^2 l}{\partial \beta_1 \beta_0} & \frac{\partial^2 l}{\partial \beta_1^2} & \frac{\partial^2 l}{\partial \beta_1 \beta_2} \\ \frac{\partial^2 l}{\partial \beta_2 \beta_0} & \frac{\partial^2 l}{\partial \beta_2 \beta_1} & \frac{\partial^2 l}{\partial \beta_2^2} \end{pmatrix}$$

# 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 asymptomatic 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

- 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 variable**:

$$z_0 = \hat{\eta}_0 + (y - \hat{\mu}_0) \left( \frac{d\eta}{d\mu} \Big|_{\hat{\mu}_0} \right)$$

3. The quadratic weight is defined as:

$$W_0^{-1} = \left( \frac{d\eta}{d\mu} \Big|_{\hat{\mu}_0} \right)^2 V_0 \quad \text{where } V_0 \text{ is the variance function evaluated at } \hat{\mu}_0.$$

Now, regress  $z_0$  onto covariates  $x_1, \dots, x_p$  with weights  $W_0$  to give new parameter estimate  $\hat{\beta}_1$ . Form new  $\hat{\eta}_1$ . Repeat until changes in estimates are sufficiently small.

# ALGORITHM (2)



Taylor expansion of  $f(x)$  about a point  $(x - a)$ :

$$f(x) = f(a) + f'(a)(x - a) + \frac{f''(a)}{2!}(x - a)^2 + \dots + \frac{f^{(n)}(a)}{n!}(x - a)^n$$

---

Note that  $z$  is just a linearized form of the link function applied to the data, because, up to first order

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

hence the right hand side is

$$\eta + (y - \mu) \frac{d\eta}{d\mu} \quad (\text{because } \eta = g(\mu))$$

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

# Starting values

---

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

$$\hat{\eta}_0, \left( \frac{d\eta}{d\mu} \Big|_{\hat{\mu}_0} \right) \text{ and } V_0$$

Note that adjustments may be required to the data to prevent, for example, 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

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad i = 1, \dots, p, \quad j = 1, \dots, n$$

where  $\alpha_i$  is the treatment effect and  $\varepsilon_{ij}$

is the error associated with the  $i^{\text{th}}$  treatment and  $j^{\text{th}}$  observation can be recast as a simple linear regression model by defining

$$X_i = \begin{cases} 1, & \text{if group } i \\ 0, & \text{otherwise} \end{cases} \quad i = 1, \dots, p-1$$

# Single-factor analysis of variance

Thus expressed the one-way ANOVA model becomes

---

$$y_{ij} = \beta_0 + \beta_1 X_{1j} + \cdots + \beta_{p-1} X_{(p-1)j} + \varepsilon_{ij}$$

that is,

$$\text{Group 1: } Y_{1j} = \beta_0 + \beta_1 + \varepsilon_{ij}$$

$$\text{Group 2: } Y_{2j} = \beta_0 + \beta_2 + \varepsilon_{ij}$$

⋮

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

$$\text{Group } p: Y_{pj} = \beta_0 + \varepsilon_{ij}$$

# Regression models for one-way ANOVA

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

$$\mu_i = \begin{cases} \beta_0 + \beta_i, & \text{if } i = 1, \dots, p-1 \\ \beta_0, & \text{if } i = p \end{cases}$$

The usual null hypothesis in regression

$H_0: \beta_1 = \beta_2 = \dots = \beta_{p-1} = 0$ , means that:

$$\beta_1 = \mu_1 - \mu = 0 \Rightarrow \mu_1 = \mu = \mu_p$$

⋮

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

is thus equivalent to the null hypothesis of the ANOVA  $H_0: \mu_1 = \mu_2 = \dots = \mu_p$

# Regression models for one-way ANOVA

---

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 described previously, 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 (when a regression model with an intercept is fitted).

## Example: Effect of gender on plasma retinol levels

For assessing the effect of gender on plasma retinol levels, the one-way ANOVA is given by the 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 effect on plasma retinol levels

# The output of the regression for the same model is:

```
. reg
```

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

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  $\hat{\beta}_0 = 587.7216$ , while the same estimate for males will be  $\hat{\beta}_0 + \hat{\beta}_1 = 587.7216 + 122.3759 = 710.0976$  as described previously.

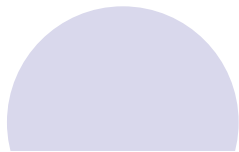
We can execute these calculations in a single step using the `reg` command with reference coding

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

-----+-----
Source |           SS          df           MS              Number of obs =       314
-----+-----
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        R-squared     =     0.0392
                                           Adj R-squared =     0.0361
                                           Root MSE     =    204.80

-----+-----
retplasm |           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 `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	Number of obs =	314	
Residual	13086344.5	312	41943.4117	F( 1, 312) =	12.73	
				Prob > F =	0.0004	
				R-squared =	0.0392	
				Adj R-squared =	0.0361	
				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
-----
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
-----+-----


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

Consistently with the regression-analysis results.



## Comments:

1. The command `xi` defines the level with the *lowest* numerical value as the default reference level. We can manipulate which level is the reference level by defining the `omit` variable with the command `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 `char _dta[omit] "prevalent"`. In case of string variables the command becomes `char _dta[omit] "string_literal"` where `string_literal` is the string level that we want to define as reference.
2. The `xi` command defines  $p-1$  variables `Ivarname_i`, ( $i=1, \dots, p-1$ ), such that `Ivarname_i = (varname == i)`.
3. The regression can then be carried out by these variables. To invoke them we use the umbrella term `i.varname`.




## Regression models for general two-way ANOVA

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

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

where  $X_i = \begin{cases} 1, & \text{if treatment } i \\ 0, & \text{otherwise} \end{cases}$   $i = 1, \dots, p-1$  and  $Z_j = \begin{cases} 1, & \text{if block } j \\ 0, & \text{otherwise} \end{cases}$   $j = 1, \dots, q-1$ , 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):

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}, \quad i = 1, \dots, p-1; j = 1, 2, \dots, q-1$$

$$\mu_{iq} = \mu + \alpha_i, \quad i = 1, \dots, p-1$$

$$\mu_{pj} = \mu + \beta_j, \quad j = 1, \dots, q-1$$

$$\mu_{pq} = \mu$$

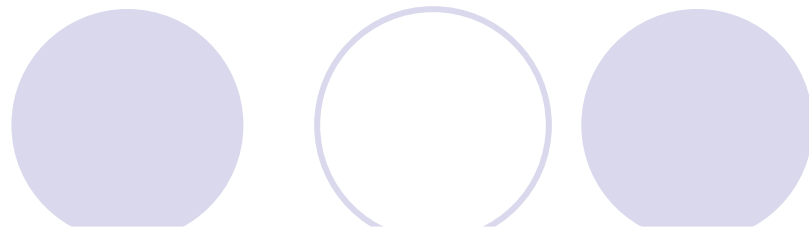
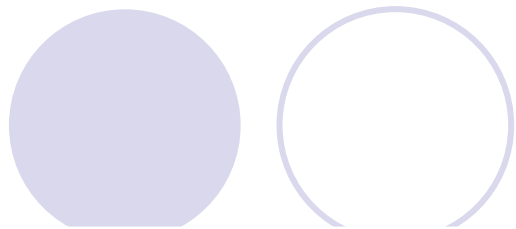


### Example: Effect of sex and vitamin use on plasma retinol levels

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:

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

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)

retplasm	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
Isex_1	166.3468	47.76693	3.482	0.001	72.35604	260.3376
Ivituse_1	33.46968	29.28935	1.143	0.254	-24.16284	91.10221
Ivituse_2	39.49589	31.87656	1.239	0.216	-23.22748	102.2193
IsXv_1_1	-11.72721	76.51943	-0.153	0.878	-162.2942	138.8398
IsXv_1_2	-255.6611	105.4603	-2.424	0.016	-463.175	-48.14725
_cons	563.2184	21.84213	25.786	0.000	520.2397	606.1971

(Model is ordinary regression, use regress instead)

where  $\text{\_Isex\_1}$  is  $X_1$  (males vs females | no vitamin use),  $\text{\_Ivituse\_1}$  is  $Z_1$  (“fairly often” – frequent vitamin users vs non users | gender=female) and  $\text{\_Ivituse\_2}$  is  $Z_2$  (“not often” – infrequent vitamin users vs non users | gender=female). The interactions are  $\text{\_IsXv\_1\_1}$ ,  $X_1Z_1$  [Gender effect\* | frequent vitamin use vs Gender effect | no vitamin use] and  $\text{\_IsXv\_1\_2}$ ,  $X_1Z_2$  [Gender effect | infrequent vitamin use vs Gender effect | no vitamin use]



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



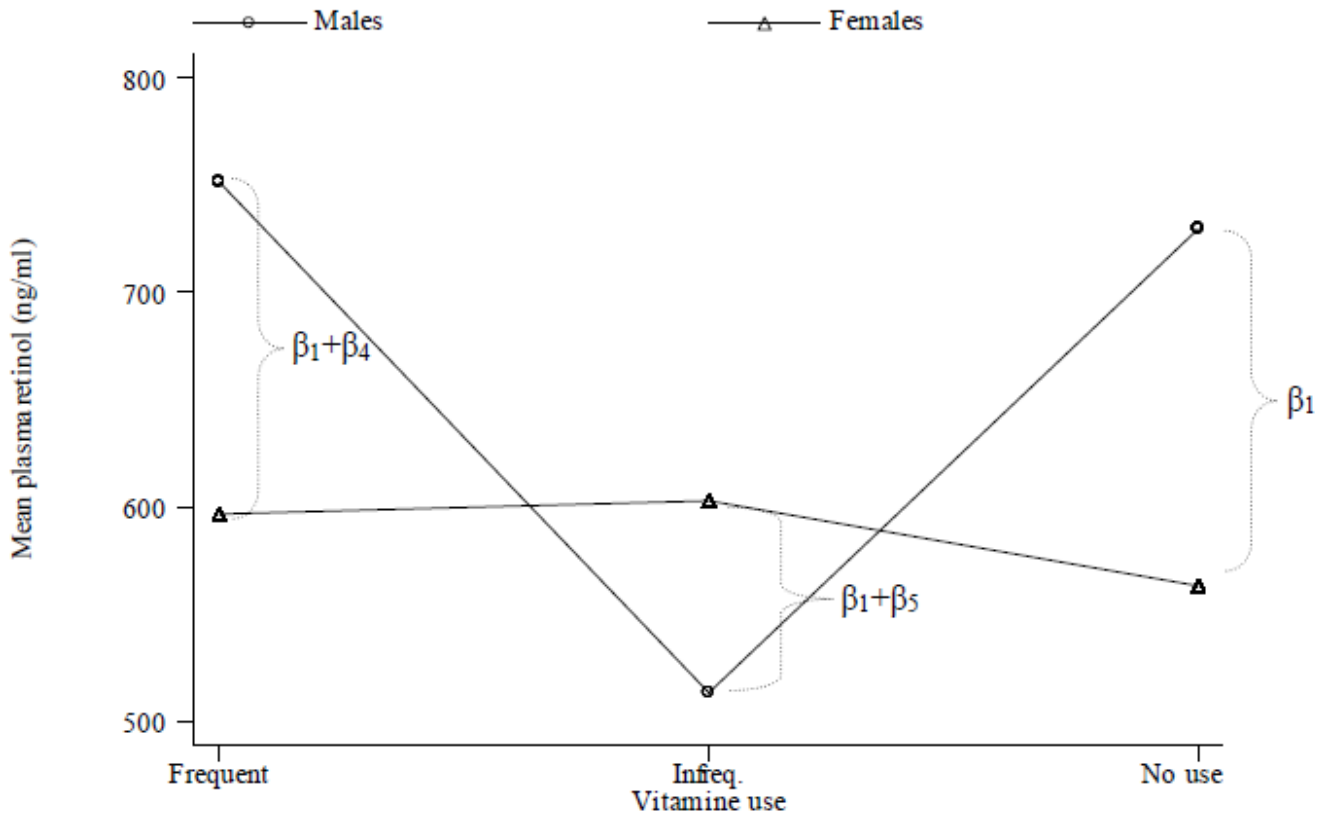
### The model

$$E[Y]=b_0+b_1*Male+b_2*Freq\_Use+b_3*Infreq\_Use+b_4*MaleX Freq\_Use +b_5* MaleX Infreq\_Use$$

	<b>Vitamin Use</b>		
<b>Gender</b>	<i>Frequent</i>	<i>Infrequent</i>	<i>No Use</i>
<i>Male</i>	$b_0+b_1+b_2+b_4$	$b_0+b_1+b_3+b_5$	$b_0+b_1$
<i>Female</i>	$b_0+b_2$	$b_0+b_3$	$b_0$
<b><i>Difference (Male-Female)</i></b>	<b><math>b_1+b_4</math></b>	<b><math>b_1+b_5</math></b>	<b><math>b_1</math></b>



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.

$$E[Y]=b_0+b_1*Male+b_2*Freq\_Use+b_3*Infreq\_Use+b_4*MaleX\ Freq\_Use +b_5* MaleX\ Infreq\_Use$$

## Estimates of model parameters

Constant :  $\_cons = \hat{\beta}_0 = 563.2184$

Main effects :  $\_Isex\_1 = \hat{\beta}_1 = 166.3468,$

$\_Ivituse\_1 = \hat{\beta}_2 = 33.46968, \_Ivituse\_2 = \hat{\beta}_3 = 39.49589$

Interactions :  $\_IsXv\_1\_1 = \hat{\beta}_4 = -11.72721, \_IsXv\_1\_2 = \hat{\beta}_5 = -255.6611$

The estimates of the various parameters are given as follows:

### 1. Females

a. Frequent users (“fairly often”):  $563.2184 + 33.46968 = \mathbf{596.68808}$

b. Infrequent users (“not often”):  $563.2184 + 39.49589 = \mathbf{602.71429}$

c. Non-users:  $\mathbf{563.2184}$

### 2. Males

a. Frequent users:  $563.2184 + 166.3468 + 33.46968 + (-11.72721) = \mathbf{751.30767}$

b. Infrequent users:  $563.2184 + 166.3468 + 39.49589 + (-255.6611) = \mathbf{513.39999}$

c. Non-users:  $563.2184 + 166.3468 = \mathbf{729.5652}$




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)

Means, Standard Deviations and Frequencies of Plasma retinol (ng/ml)

      Sex |      Vitamine use
          | Frequent  Infrequent No Use  | Total
-----+-----+-----+-----+-----+
Males    | 751.30769    513.4  729.56522 | 710.09756
          | 329.43269  298.59303  290.0285 | 305.52208
          |      13      5      23 |      41
-----+-----+-----+-----+
Females  | 596.68807  602.71429  563.21839 | 587.72161
          | 203.71816  184.6959  159.92785 | 185.43069
          |      109      77      87 |      273
-----+-----+-----+-----+
Total    | 613.16393  597.26829    598 | 603.70064
          | 223.83038  192.02109  204.39088 | 208.60239
          |      122      82    110 |      314
```



## 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:  $Comb1 = b_{men} + b_{freq} + b_{men*freq} + \_cons$ .

If the model has been defined as: `reg retplasm men freq infreq menfreq meninfr`

We can get the estimate, the SE and the 95% CI of the parameter `Comb1` in Stata by the command

```
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:  $\mathbf{C}=(1,1,0,1,0,1)$ . This constraint asks for the sum of those parameters indicated by 1s in  $\mathbf{C}$ . Stata command: `matrix C=(1,1,0,1,0,1)`

2) Estimate the linear combination as:  $\mathbf{C}*\mathbf{b}'$  where  $\mathbf{b}$  the  $1 \times p$  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 obtained as: `mat A=e(b)`.

3) Estimate the variance of the linear combination as:  $\mathbf{C}*\mathbf{V}(\mathbf{b})*\mathbf{C}'$  where  $\mathbf{V}(\mathbf{b})$  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

	Frequent vitamin user	Infrequent vitamin use	Not use
	Estimate (95% CI)	Estimate (95% CI)	Estimate (95% CI)
Male	751.31 (640 – 862)	513.40 (334 – 693)	729.57 (646 – 813)
Female	596.69 (558 – 635)	602.71 (557 – 648)	563.22 (520 – 606)

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:)

$$y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ + \varepsilon$$


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.,  $X_1 = \begin{cases} 1, & \text{male} \\ 0, & \text{otherwise} \end{cases}$ ,  $Z = \text{age}$  and  $XZ$  is the age/gender interaction.

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

$$\text{Males: } y_M = \underbrace{(\beta_0 + \beta_1)}_{\beta_{0M}} + \underbrace{(\beta_2 + \beta_3)}_{\beta_{1M}} Z + \varepsilon$$

$$\text{Females: } y_F = \beta_0 + \beta_2 Z + \varepsilon.$$



## 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 retpiasm i.sex*age
i.sex          Isex_1-2      (naturally coded; Isex_2 omitted)
i.sex*age      IsXage_#      (coded as above)

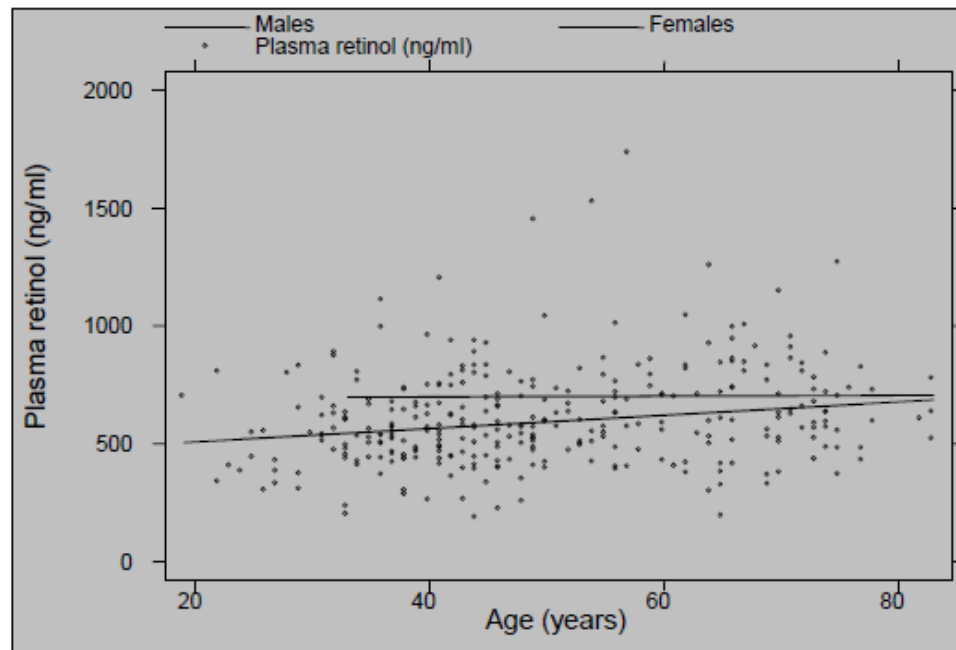
Iteration 1 : deviance = 12658374.0533

Residual df =      310                No. of obs =      314
Pearson X2  = 1.27e+07                Deviance   = 1.27e+07
Dispersion = 40833.46                Dispersion = 40833.46

Gaussian (normal) distribution, identity link
-----
retpiasm |      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:

```
. 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.