UCD School of Mathematics and Statistics

STAT40840: Data programming with SAS
Laura Kirwan

Lecture 9

www.ucd.ie/mathstat

Lecture 9: Analysing Data

9.1 General Linear Models

9.2 Generalised Linear Models

9.3 Linear Mixed Models

9.4 Generalised Linear Mixed Models

Objectives – 9.1

- Use a SAS procedure to fit a Linear Model
- Use LSMEANS to compare factor means

GLM procedure

- The GLM procedure uses the method of least squares to fit general linear models.

- Among the statistical methods available in PROC GLM are regression, analysis of variance, analysis of covariance, multivariate analysis of variance, and partial correlation.

GLM procedure

- PROC GLM analyzes data within the framework of general linear models.
- PROC GLM handles models relating one or several continuous dependent variables to one or several independent variables.
- The independent variables can be either *classification* variables, which divide the observations into discrete groups, or *continuous* variables.

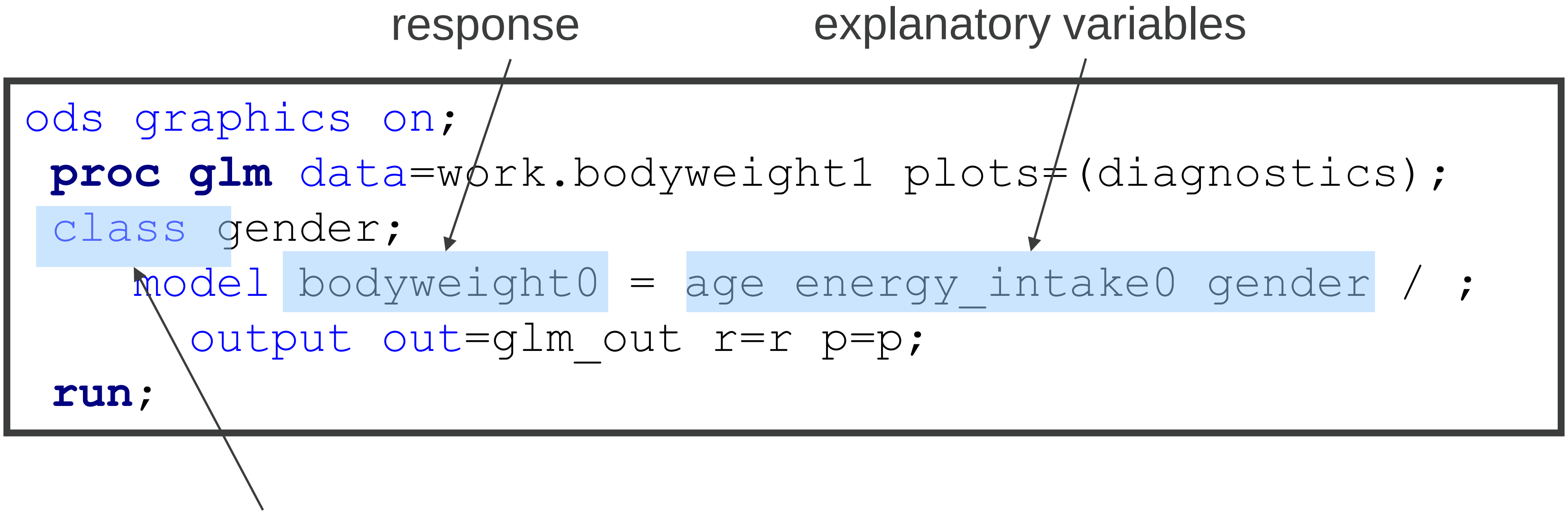
Scenario

We will continue to use the bodyweight dataset (from assignment1).

We will use the GLM procedure to fit a general linear model including both continuous and categorical explanatory variables

GLM procedure

General linear model



response

explanatory variables

```
ods graphics on;
 proc glm data=work.bodyweight1 plots=(diagnostics);
  class gender;
     model bodyweight0 = age energy_intake0 gender / ;
       output out=glm_out r=r p=p;
 run;
```

Specifying “gender” as a categorical variable

L9_D1.sas

GLM procedure

General linear model

Diagnostic plots

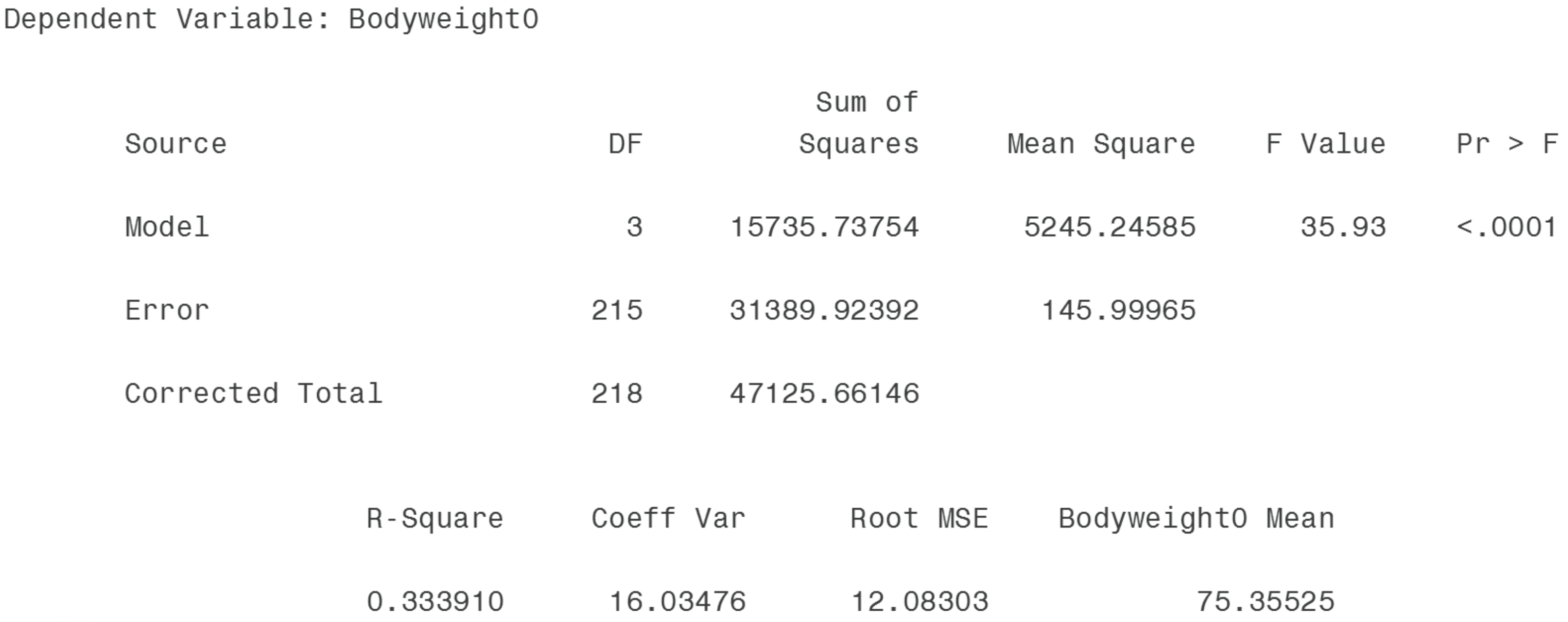
```
ods graphics on;
 proc glm data=work.bodyweight1 plots=(diagnostics);
 class gender;
    model bodyweight0 = age energy_intake0 gender / ;
      output out=glm_out r=r p=p;
 run;
```

Creating output dataset

L9_D1.sas

Viewing the Output

```
ods graphics on;
 proc glm data=work.bodyweight1 plots=(diagnostics);
 class gender;
    model bodyweight0 = age energy_intake0 gender / ;
      output out=glm_out r=r p=p;
 run;
```

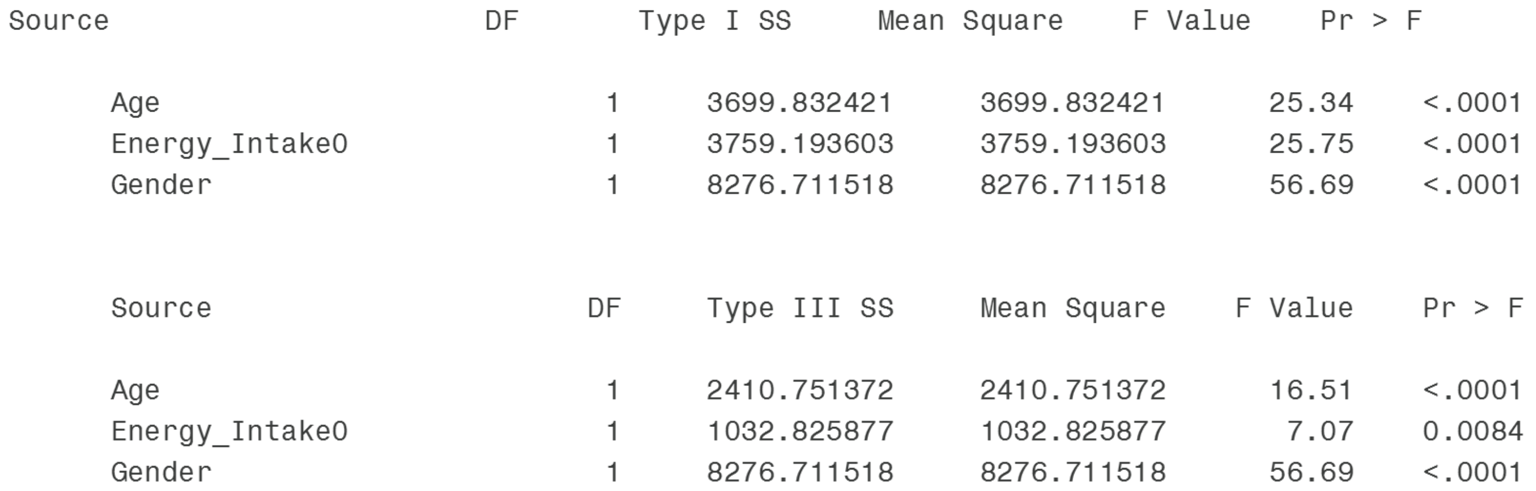


```
Dependent Variable: Bodyweight0

                                          Sum of
       Source                 DF         Squares     Mean Square    F Value    Pr > F

       Model                   3     15735.73754      5245.24585      35.93    <.0001

       Error                 215     31389.92392       145.99965

       Corrected Total       218     47125.66146


                    R-Square     Coeff Var      Root MSE    Bodyweight0 Mean

                    0.333910      16.03476      12.08303            75.35525
```

L9_D1.sas

Viewing the Output

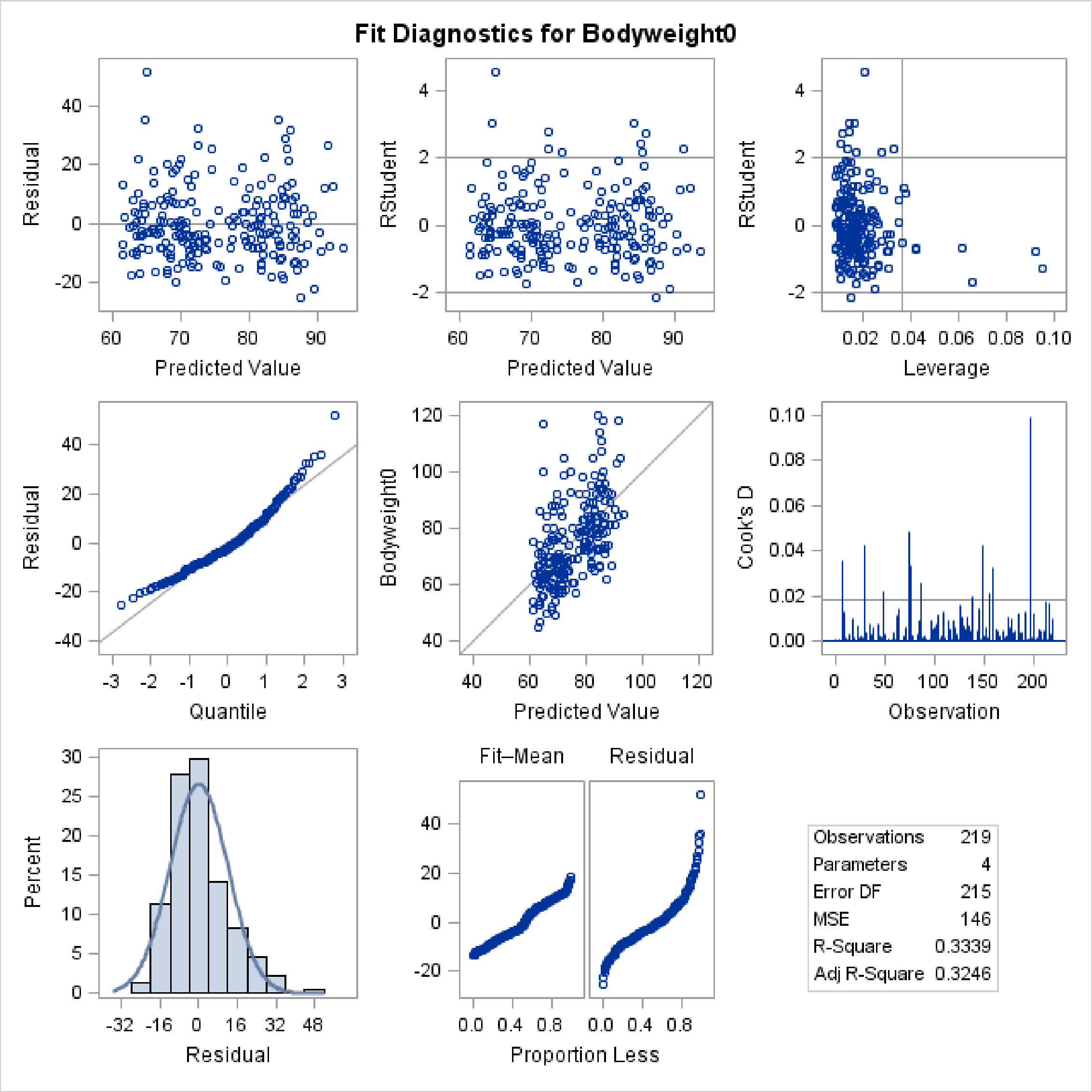
```
ods graphics on;
 proc glm data=work.bodyweight1 plots=(diagnostics);
 class gender;
    model bodyweight0 = age energy_intake0 gender / ;
      output out=glm_out r=r p=p;
 run;
```



```
Source                  DF       Type I SS      Mean Square    F Value    Pr > F

     Age                          1     3699.832421     3699.832421      25.34    <.0001
     Energy_Intake0               1     3759.193603     3759.193603      25.75    <.0001
     Gender                       1     8276.711518     8276.711518      56.69    <.0001


     Source                      DF     Type III SS     Mean Square    F Value    Pr > F

     Age                          1     2410.751372     2410.751372      16.51    <.0001
     Energy_Intake0               1     1032.825877     1032.825877       7.07    0.0084
     Gender                       1     8276.711518     8276.711518      56.69    <.0001
```

L9_D1.sas

Viewing the Output



Fit Diagnostics for Bodyweight0

Observations	219
Parameters	4
Error DF	215
MSE	146
R-Square	0.3339
Adj R-Square	0.3246

L9_D1.sas

GLM procedure

General linear model

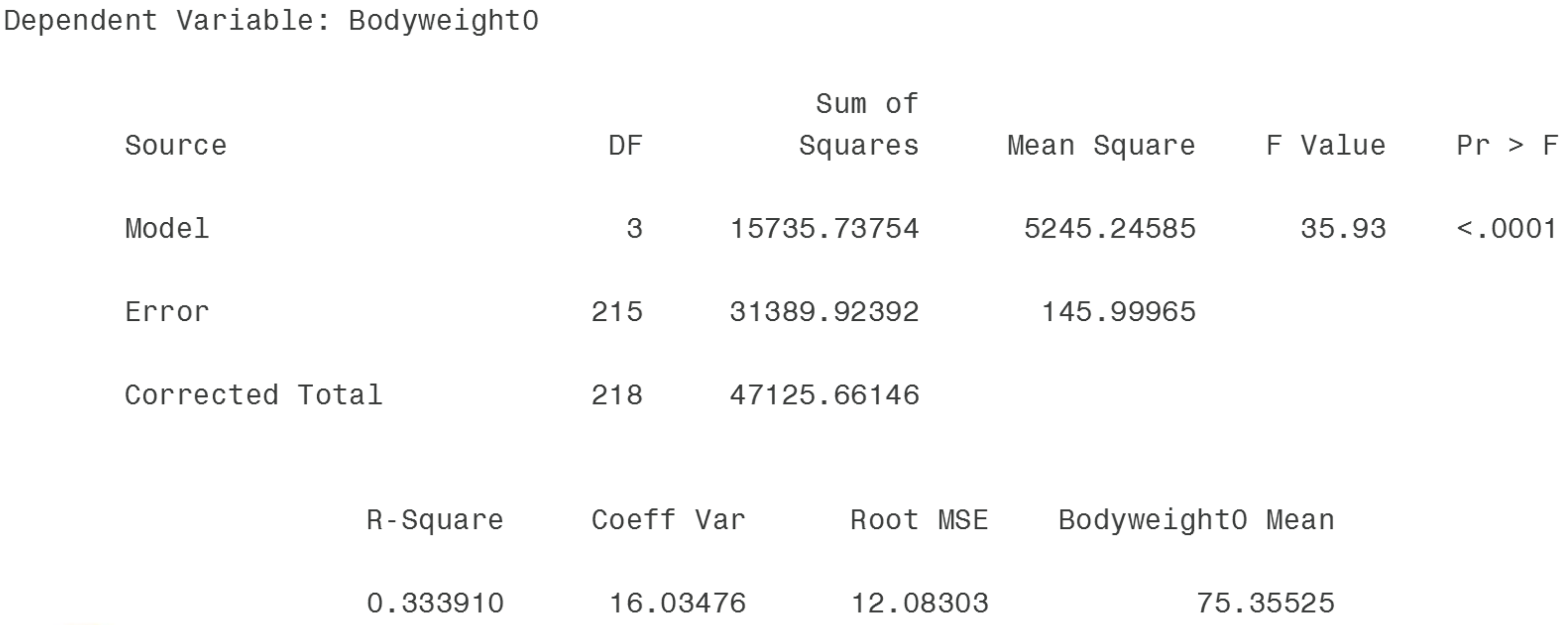
```
*GLM - Comparing means;
 ods graphics on;
 proc glm data=work.bodyweight1 plots=(diagnostics);
 class gender;
    model bodyweight0 = age energy_intake0 gender / ;
      lsmeans gender / pdiff;
 run;
```

Comparing Means

L9_D2.sas

Viewing the Output

```
ods graphics on;
 proc glm data=work.bodyweight1 plots=(diagnostics);
 class gender;
    model bodyweight0 = age energy_intake0 gender / ;
      output out=glm_out r=r p=p;
 run;
```



```
Dependent Variable: Bodyweight0

                                         Sum of
    Source                    DF        Squares     Mean Square    F Value    Pr > F

    Model                      3    15735.73754      5245.24585      35.93    <.0001

    Error                    215    31389.92392       145.99965

    Corrected Total          218    47125.66146


                  R-Square     Coeff Var      Root MSE    Bodyweight0 Mean

                  0.333910      16.03476      12.08303            75.35525
```

L9_D2.sas

Viewing the Output

```
*GLM - Comparing means;
 ods graphics on;
 proc glm data=work.bodyweight1 plots=(diagnostics);
 class gender;
    model bodyweight0 = age energy_intake0 gender / ;
      lsmeans gender / pdiff;
 run;
```

```
Least Squares Means


                                               H0:LSMean1=
                          Bodyweight0            LSMean2
                 Gender        LSMEAN           Pr > |t|

                 0         82.4322901             <.0001
                 1         69.4081596
```

L9_D2.sas

Objectives - 9.2

- Use a SAS procedure to fit a Generalised Linear Model
- Explore a variety of link functions

GENMOD procedure

- The GENMOD procedure fits generalized linear models.
- The class of generalized linear models is an extension of traditional linear models that allows the mean of a population to depend on a linear predictor through a nonlinear link function and allows the response probability distribution to be any member of an exponential family of distributions.

GENMOD procedure

- The GENMOD procedure fits a generalized linear model to the data by maximum likelihood estimation of the parameter vector . There is, in general, no closed form solution for the maximum likelihood estimates of the parameters.
- The GENMOD procedure estimates the parameters of the model numerically through an iterative fitting process.
- The dispersion parameter is also estimated by maximum likelihood or, optionally, by the residual deviance or by Pearson’s chi-square divided by the degrees of freedom.
- Covariances, standard errors, and p-values are computed for the estimated parameters based on the asymptotic normality of maximum likelihood estimators.

GENMOD procedure

A number of popular link functions and probability distributions are available in the GENMOD procedure. The built-in link functions are as follows:

identity:	$g(\mu) = \mu$
logit:	$g(\mu) = \log(\mu / (1-\mu))$
probit:	$g(\mu) = \Phi^{-1}(\mu)$
	(where Φ is the standard normal cumulative distribution function)
power:	$g(\mu) = \begin{cases} \mu^{\lambda} & \text{if } \lambda \neq 1 \\ \log(\mu) & \text{if } \lambda = 1 \end{cases}$
log:	$g(\mu) = \log(\mu)$
complementary log-log:	$g(\mu) = \log(-\log(1- \mu))$

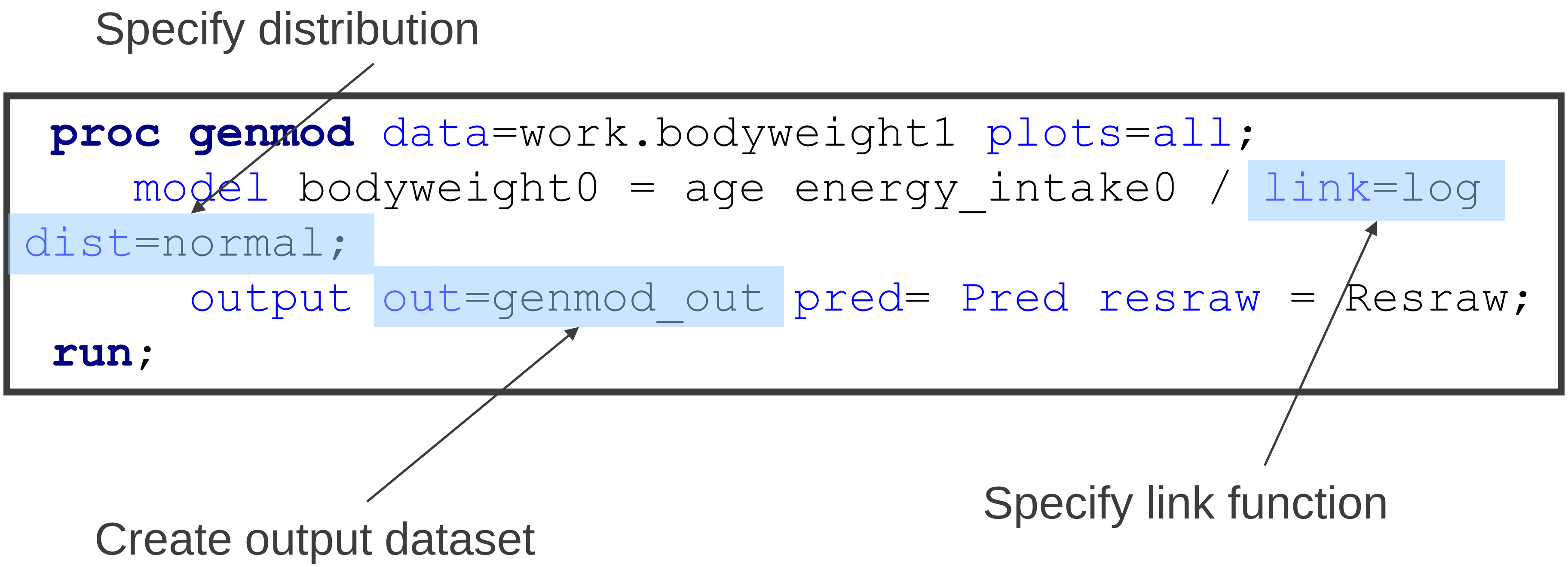
GENMOD procedure

The available distributions are as follows:

- normal
- binomial (proportion)
- Poisson
- gamma
- inverse Gaussian
- negative binomial
- geometric
- multinomial
- zero-inflated Poisson
- zero-inflated negative binomial

GENMOD procedure

Generalised linear model



Specify distribution

```
proc genmod data=work.bodyweight1 plots=all;
    model bodyweight0 = age energy_intake0 / link=log
dist=normal;
        output out=genmod_out pred= Pred resraw = Resraw;
run;
```

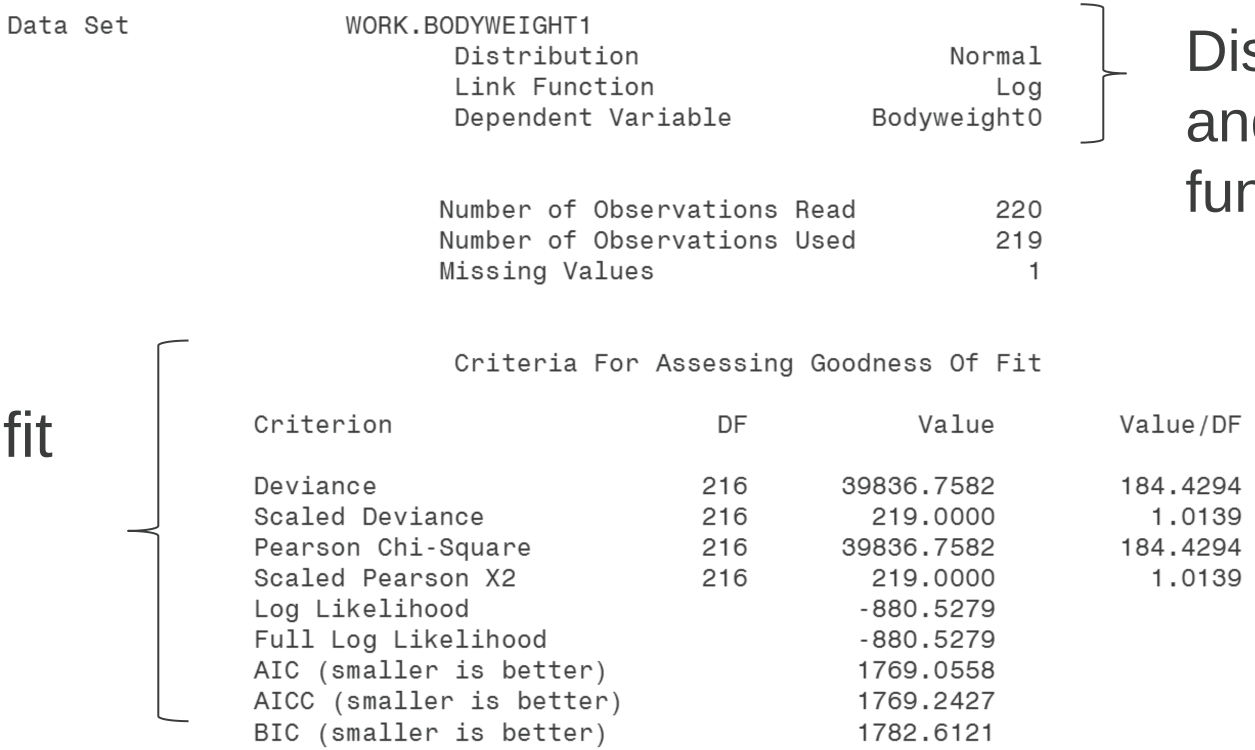
Specify link function

Create output dataset

L9_D3.sas

Viewing the Output

```
 proc genmod data=work.bodyweight1 plots=all;
      model bodyweight0 = age energy_intake0 / link=log
dist=normal;
          output out=genmod_out pred= Pred resraw = Resraw;
 run;
```



```
          Data Set              WORK.BODYWEIGHT1
                                       Distribution                    Normal
                                       Link Function                      Log
                                       Dependent Variable         Bodyweight0


                                      Number of Observations Read         220
                                      Number of Observations Used         219
                                      Missing Values                        1


                                       Criteria For Assessing Goodness Of Fit

                             Criterion                 DF           Value        Value/DF

                             Deviance                 216      39836.7582        184.4294
                             Scaled Deviance          216        219.0000          1.0139
                             Pearson Chi-Square       216      39836.7582        184.4294
                             Scaled Pearson X2        216        219.0000          1.0139
                             Log Likelihood                     -880.5279
                             Full Log Likelihood                -880.5279
                             AIC (smaller is better)            1769.0558
                             AICC (smaller is better)           1769.2427
                             BIC (smaller is better)            1782.6121
```

Distribution and link function

Model fit

L9_D3.sas

Viewing the Output

```
 proc genmod data=work.bodyweight1 plots=all;
    model bodyweight0 = age energy_intake0 / link=log
dist=normal;
      output out=genmod_out pred= Pred resraw = Resraw;
 run;
```

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	3.9595	0.0593	3.8432	4.0757	4455.50	<.0001
Age	1	0.0042	0.0009	0.0025	0.0060	23.14	<.0001
Energy_Intake0	1	0.0001	0.0000	0.0000	0.0001	21.12	<.0001
Scale	1	13.4871	0.6444	12.2814	14.8113		

L9_D3.sas

Objectives - 9.3

- Use a SAS procedure to fit a Linear Mixed Model
- Explore a variety of structures for variance-covariance matrix

MIXED procedure

- The MIXED procedure fits a variety of mixed linear models to data and enables you to use these fitted models to make statistical inferences about the data.

- A mixed linear model is a generalisation of the standard linear model used in the GLM procedure, the generalisation being that the data are permitted to exhibit correlation and nonconstant variability.

MIXED procedure

- The mixed model generalizes the standard linear model for response y as follows: $y = X\beta + Z\gamma + \varepsilon$

- Here, β is an unknown vector of fixed-effects parameters with known design matrix X, γ is an unknown vector of random-effects parameters with known design matrix Z, and ε is an unknown random error vector whose elements are no longer required to be independent and homogeneous.

- To further develop this notion of variance modeling, assume that γ and ε are Gaussian random variables that are uncorrelated and have expectations 0 and variances **G** and **R**, respectively.

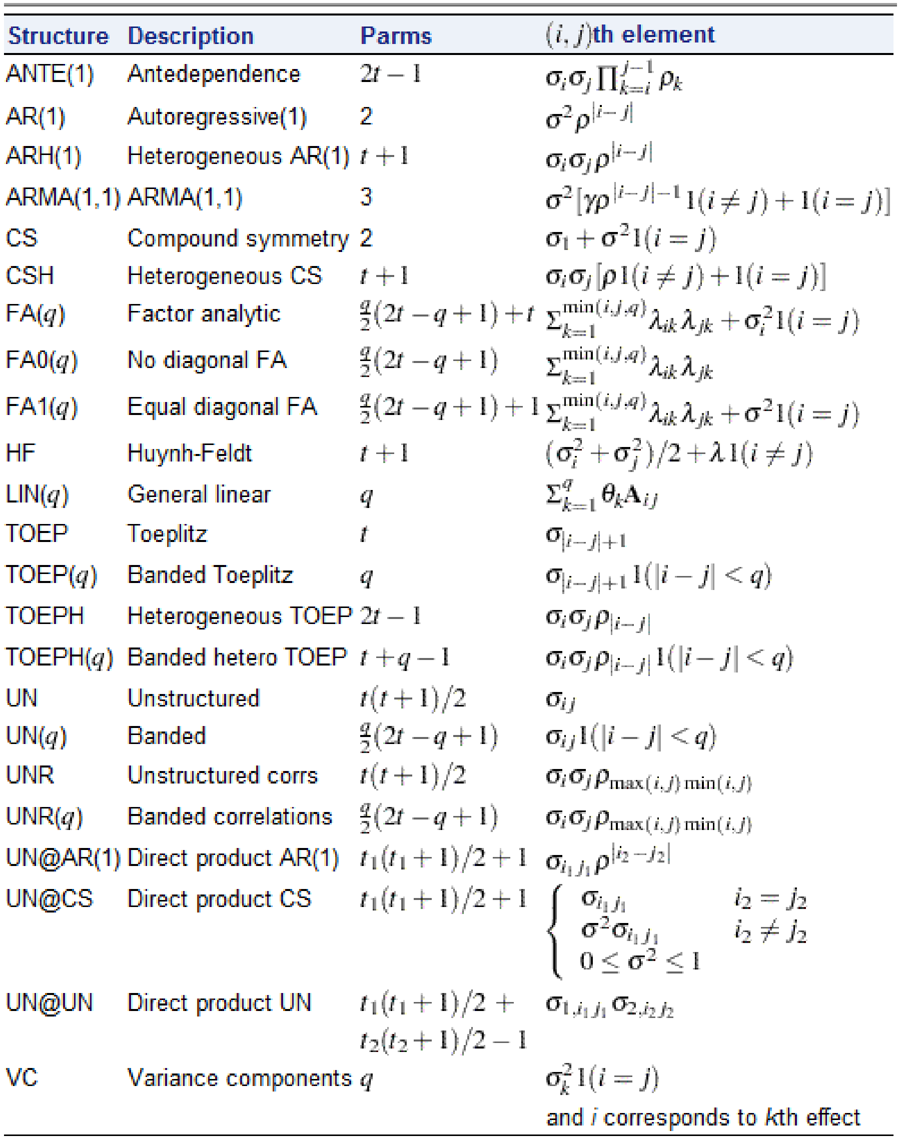
MIXED procedure

- The REPEATED statement is used to specify the **R** matrix in the mixed model.

- The RANDOM statement defines the random effects constituting the γ vector in the mixed model, and specifies the **G** matrix. It can be used to specify traditional variance component models and to specify random coefficients. The random effects can be classification or continuous, and multiple RANDOM statements are possible.

MIXED procedure

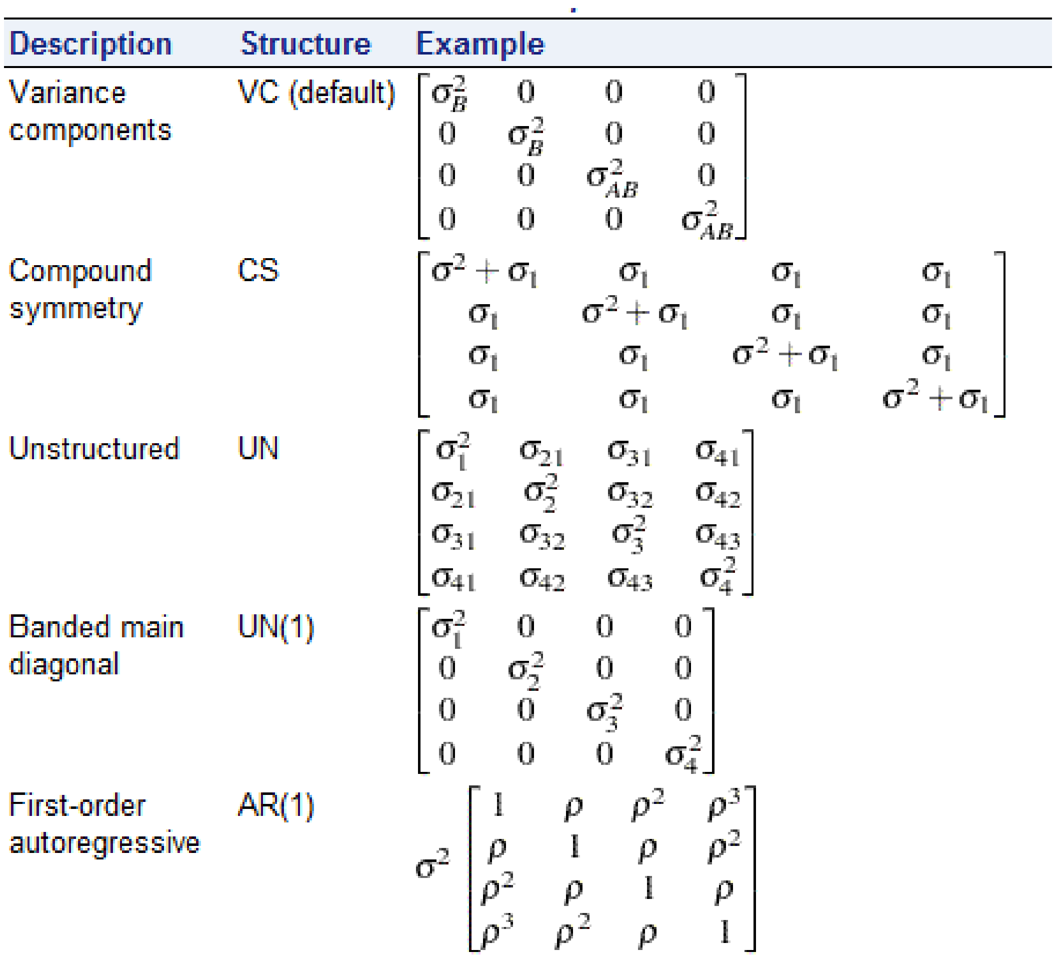
Covariance structures



Structure	Description	Parms	(i,j)th element
ANTE(1)	Antedependence	$2t-1$	$\sigma_i\sigma_j\prod_{k=i}^{j-1}\rho_k$
AR(1)	Autoregressive(1)	2	$\sigma^2\rho^{\lvert i-j\rvert}$
ARH(1)	Heterogeneous AR(1)	$t+1$	$\sigma_i\sigma_j\rho^{\lvert i-j\rvert}$
ARMA(1,1)	ARMA(1,1)	3	$\sigma^2[\gamma\rho^{\lvert i-j\rvert-1}1(i\neq j)+1(i=j)]$
CS	Compound symmetry	2	$\sigma_1+\sigma^2 1(i=j)$
CSH	Heterogeneous CS	$t+1$	$\sigma_i\sigma_j[\rho 1(i\neq j)+1(i=j)]$
FA(q)	Factor analytic	$\frac{q}{2}(2t-q+1)+t$	$\Sigma_{k=1}^{\min(i,j,q)}\lambda_{ik}\lambda_{jk}+\sigma_i^2 1(i=j)$
FA0(q)	No diagonal FA	$\frac{q}{2}(2t-q+1)$	$\Sigma_{k=1}^{\min(i,j,q)}\lambda_{ik}\lambda_{jk}$
FA1(q)	Equal diagonal FA	$\frac{q}{2}(2t-q+1)+1$	$\Sigma_{k=1}^{\min(i,j,q)}\lambda_{ik}\lambda_{jk}+\sigma^2 1(i=j)$
HF	Huynh-Feldt	$t+1$	$(\sigma_i^2+\sigma_j^2)/2+\lambda 1(i\neq j)$
LIN(q)	General linear	q	$\Sigma_{k=1}^{q}\theta_k\mathbf{A}_{ij}$
TOEP	Toeplitz	t	$\sigma_{\lvert i-j\rvert+1}$
TOEP(q)	Banded Toeplitz	q	$\sigma_{\lvert i-j\rvert+1}1(\lvert i-j\rvert<q)$
TOEPH	Heterogeneous TOEP	$2t-1$	$\sigma_i\sigma_j\rho_{\lvert i-j\rvert}$
TOEPH(q)	Banded hetero TOEP	$t+q-1$	$\sigma_i\sigma_j\rho_{\lvert i-j\rvert}1(\lvert i-j\rvert<q)$
UN	Unstructured	$t(t+1)/2$	σ_{ij}
UN(q)	Banded	$\frac{q}{2}(2t-q+1)$	$\sigma_{ij}1(\lvert i-j\rvert<q)$
UNR	Unstructured corrs	$t(t+1)/2$	$\sigma_i\sigma_j\rho_{\max(i,j)\min(i,j)}$
UNR(q)	Banded correlations	$\frac{q}{2}(2t-q+1)$	$\sigma_i\sigma_j\rho_{\max(i,j)\min(i,j)}$
UN@AR(1)	Direct product AR(1)	$t_1(t_1+1)/2+1$	$\sigma_{i_1j_1}\rho^{\lvert i_2-j_2\rvert}$
UN@CS	Direct product CS	$t_1(t_1+1)/2+1$	$\begin{cases}\sigma_{i_1j_1} & i_2=j_2\\ \sigma^2\sigma_{i_1j_1} & i_2\neq j_2\\ 0\leq\sigma^2\leq 1\end{cases}$
UN@UN	Direct product UN	$t_1(t_1+1)/2+$ $t_2(t_2+1)/2-1$	$\sigma_{1,i_1j_1}\sigma_{2,i_2j_2}$
VC	Variance components	q	$\sigma_k^2 1(i=j)$ and i corresponds to kth effect

MIXED procedure

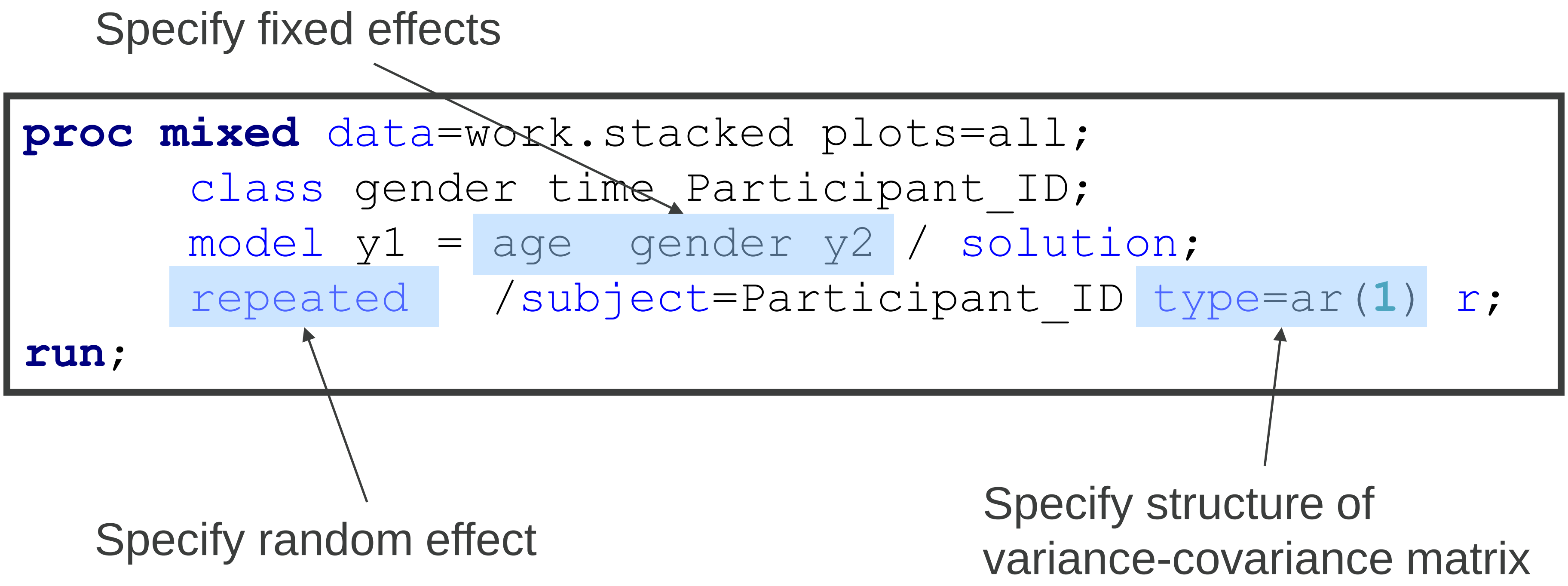
Common covariance structures



Description	Structure	Example
Variance components	VC (default)	$\begin{bmatrix} \sigma_B^2 & 0 & 0 & 0 \\ 0 & \sigma_B^2 & 0 & 0 \\ 0 & 0 & \sigma_{AB}^2 & 0 \\ 0 & 0 & 0 & \sigma_{AB}^2 \end{bmatrix}$
Compound symmetry	CS	$\begin{bmatrix} \sigma^2+\sigma_1 & \sigma_1 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma^2+\sigma_1 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma^2+\sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma_1 & \sigma^2+\sigma_1 \end{bmatrix}$
Unstructured	UN	$\begin{bmatrix} \sigma_1^2 & \sigma_{21} & \sigma_{31} & \sigma_{41} \\ \sigma_{21} & \sigma_2^2 & \sigma_{32} & \sigma_{42} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \sigma_{43} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 \end{bmatrix}$
Banded main diagonal	UN(1)	$\begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 \\ 0 & 0 & \sigma_3^2 & 0 \\ 0 & 0 & 0 & \sigma_4^2 \end{bmatrix}$
First-order autoregressive	AR(1)	$\sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$

MIXED procedure

Linear Mixed model



Specify fixed effects

```
proc mixed data=work.stacked plots=all;
        class gender time Participant_ID;
        model y1 = age  gender y2 / solution;
        repeated   /subject=Participant_ID type=ar(1) r;
run;
```

Specify random effect

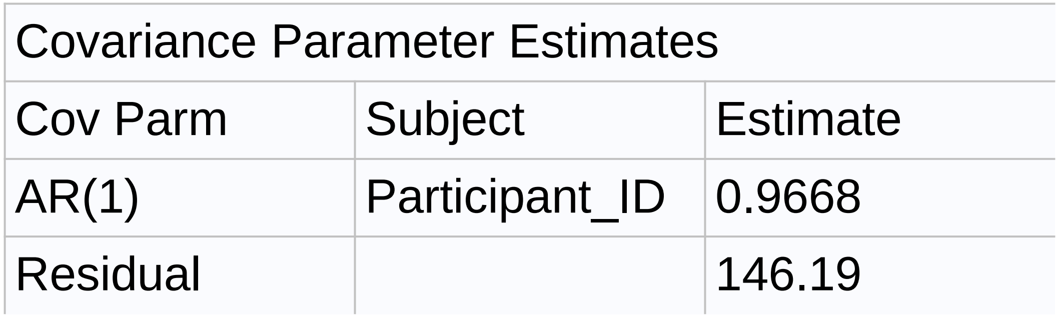
Specify structure of
variance-covariance matrix

L9_D4.sas

Viewing the Output

```
proc mixed data=work.stacked plots=all;
      class gender time Participant_ID;
      model y1 = age  gender y2 / solution;
      repeated   /subject=Participant_ID type=ar(1) r;
run;
```

Random Effects



Covariance Parameter Estimates		
Cov Parm	Subject	Estimate
AR(1)	Participant_ID	0.9668
Residual		146.19

L9_D4.sas

Viewing the Output

```
proc mixed data=work.stacked plots=all;
      class gender time Participant_ID;
      model y1 = age  gender y2 / solution;
      repeated   /subject=Participant_ID type=ar(1) r;
run;
```

Fixed Effects

Solution for Fixed Effects

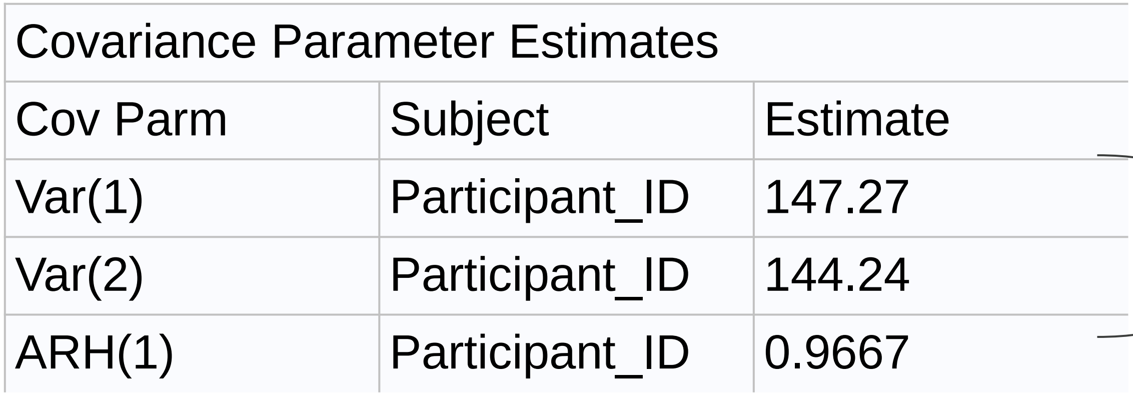
Effect	Gender	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		56.5605	2.7823	217	20.33	<.0001
Age		0.2271	0.05941	217	3.82	0.0002
Gender	0	13.9088	1.6442	217	8.46	<.0001
Gender	1	0	.	.	.	.
y2		0.000903	0.000263	174	3.44	0.0007

L9_D4.sas

Alternative covariance structure

```
proc mixed data=work.stacked plots=all;
      class gender time Participant_ID;
      model y1 = age  gender y2 / solution;
      repeated   /subject=Participant_ID type=arh(1) r;
run;
```

Random Effects



Covariance Parameter Estimates		
Cov Parm	Subject	Estimate
Var(1)	Participant_ID	147.27
Var(2)	Participant_ID	144.24
ARH(1)	Participant_ID	0.9667

Heterogeneous variance

L9_D5.sas

Objectives - 9.4

- Use a SAS procedure to fit a Generalised Linear Mixed Model
- Explore a variety of structures for variance-covariance matrix

GLIMMIX procedure

- The GLIMMIX procedure fits statistical models to data with correlations or nonconstant variability and where the response is not necessarily normally distributed. These models are known as generalized linear mixed models (GLMM).

- GLMMs, like linear mixed models, assume normal (Gaussian) random effects. Conditional on these random effects, data can have any distribution in the exponential family.

GLIMMIX procedure

- The exponential family comprises many of the elementary discrete and continuous distributions. The binary, binomial, Poisson, and negative binomial distributions, for example, are discrete members of this family. The normal, beta, gamma, and chi-square distributions are representatives of the continuous distributions in this family.

- In the absence of random effects, the GLIMMIX procedure fits generalized linear models (fit by the GENMOD procedure).

GLIMMIX procedure

Generalised Linear Mixed model

```
proc glimmix data=work.stacked plots=all;
       class gender time Participant_ID;
       model y1 = age  gender y2 / solution link=log dist=normal;
       random   _residual_ /subject=Participant_ID type=arh(1)  g;
run;
```

Specify random effect

Specify link function and distribution

L9_D6.sas

GLIMMIX procedure

- The RANDOM statement defines the Z matrix of the mixed model, the random effects in the γ vector, the structure of G, and the structure of R .

- You can specify the _RESIDUAL_ keyword before the option slash (/) to indicate a residual-type (R-side) random component that defines the matrix. Basically, the _RESIDUAL_ keyword takes the place of the *random-effect* if you want to specify R-side variances and covariance structures.

GLIMMIX procedure

Heterogenous variance-covariance matrix

```
proc glimmix data=work.stacked plots=all;
    class gender time Participant_ID;
    model y1 = age  gender y2 / solution link=power(0.5) dist=normal;
    random _residual_ /subject=Participant_ID type=arh(1) group=gender;
    covtest 'Equal Covariance Matrices'  homogeneity;
run;
```

Provides a mechanism to obtain statistical inferences for the covariance parameters.

Identifies groups by which to vary the covariance parameters

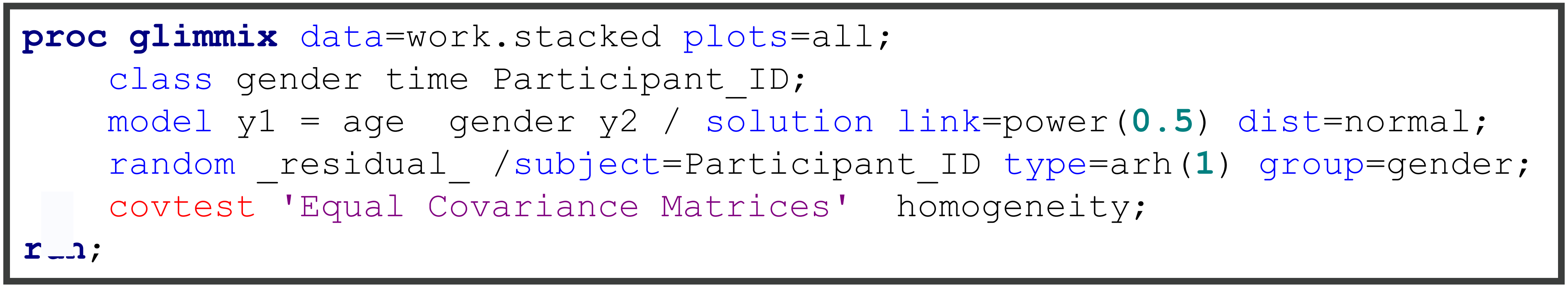
L9_D6.sas

GLIMMIX procedure

- The RANDOM statement defines the Z matrix of the mixed model, the random effects in the γ vector, the structure of G, and the structure of R .

- You can specify the _RESIDUAL_ keyword before the option slash (/) to indicate a residual-type (R-side) random component that defines the matrix. Basically, the _RESIDUAL_ keyword takes the place of the *random-effect* if you want to specify R-side variances and covariance structures.

Heterogenous covariance structure



```
proc glimmix data=work.stacked plots=all;
    class gender time Participant_ID;
    model y1 = age  gender y2 / solution link=power(0.5) dist=normal;
    random _residual_ /subject=Participant_ID type=arh(1) group=gender;
    covtest 'Equal Covariance Matrices'  homogeneity;
run;
```

Covariance Parameter Estimates				
Cov Parm	Subject	Group	Estimate	Standard Error
Var(1)	Participant_ID	Gender 0	143.30	20.3454
Var(2)	Participant_ID	Gender 0	145.94	20.9289
ARH(1)	Participant_ID	Gender 0	0.9597	0.008335
Var(1)	Participant_ID	Gender 1	150.43	19.6600
Var(2)	Participant_ID	Gender 1	141.57	19.0865
ARH(1)	Participant_ID	Gender 1	0.9730	0.005322

L9_D6.sas

Heterogenous covariance structure

```
proc glimmix data=work.stacked plots=all;
    class gender time Participant_ID;
    model y1 = age  gender y2 / solution link=power(0.5) dist=normal;
    random _residual_ /subject=Participant_ID type=arh(1) group=gender;
    covtest 'Equal Covariance Matrices'  homogeneity;
run;
```

Tests of Covariance Parameters
Based on the Residual Pseudo-Likelihood

Label	DF	-2 Res Log P-Like	ChiSq	Pr > ChiSq	Note
Equal Covariance Matrices	3	393.92	4.31	0.2302	DF

L9_D6.sas