

Statistical Modeling Using SAS

Xiangming Fang

Department of Biostatistics
East Carolina University

SAS Code Workshop Series 2011



Outline

- 1 Linear Regression
- 2 Logistic Regression
- 3 General Linear Regression
- 4 Other Regression Models



Linear Regression Models

Simple linear regression model:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Multiple linear regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_m X_m + \varepsilon$$

Assumptions:

- Relationships between the response and predictors are linear
- Observations are independent
- Errors are normally distributed
- Errors have a common variance

SAS Procedures: PROC REG, PROC GLM, PROC GENMOD



Example: Question and Data

Question: How does the weight depend on the lifestyle and physiological measurements? Variables:

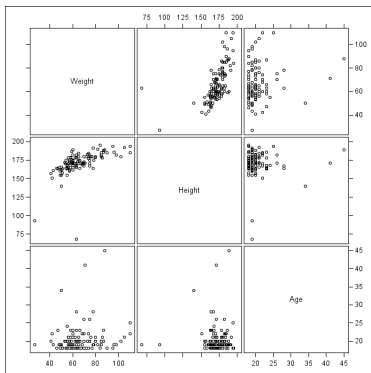
- Weight: Weight in kg
- Height: Height in cm
- Age: Age in years
- Gender
- Smoker: Regular smoker? Yes/No
- Alcohol: Regular drinker? Yes/No
- Exercise: Frequency of exercise - High, Moderate, Low



Example: Bivariate Analysis

```
PROC SGSCATTER DATA=WEIGHT;  
MATRIX WEIGHT HEIGHT AGE;  
RUN;
```

Figure: Scatter Plot



Example: Bivariate Analysis

```
PROC CORR DATA=WEIGHT;
VAR WEIGHT HEIGHT AGE;
RUN;
```

Pearson Correlation Coefficients, N = 110
 Prob > |r| under H0: Rho=0

	Weight	Height	Age
Weight	1.00000	0.57968	0.14906
Weight		<.0001	0.1201
Height	0.57968	1.00000	0.02545
Height	<.0001		0.7919
Age	0.14906	0.02545	1.00000
Age	0.1201	0.7919	



Example: Bivariate Analysis

```
PROC TTEST DATA=WEIGHT;
CLASS GENDER;
VAR WEIGHT;
RUN;
```

Method	Variances	DF	t Value	Pr > t
Pooled	Equal	108	-7.79	<.0001
Satterthwaite	Unequal	101.46	-8.02	<.0001

```
PROC TTEST DATA=WEIGHT;
CLASS SMOKER;
VAR WEIGHT;
RUN;
```

Method	Variances	DF	t Value	Pr > t
Pooled	Equal	108	-0.53	0.5973
Satterthwaite	Unequal	13.294	-0.60	0.5556

```
PROC TTEST DATA=WEIGHT;
CLASS ALCOHOL;
VAR WEIGHT;
RUN;
```

Method	Variances	DF	t Value	Pr > t
Pooled	Equal	108	-2.09	0.0390
Satterthwaite	Unequal	74.989	-2.00	0.0492



Example: Bivariate Analysis

```

PROC ANOVA DATA=WEIGHT;
CLASS EXERCISE;
MODEL WEIGHT = EXERCISE ;
MEANS EXERCISE;
MEANS EXERCISE /LSD;
RUN;

```

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Exercise	2	1155.346345	577.673173	2.59	0.0799

Level of Exercise	N	-----Weight-----	
		Mean	Std Dev
High	14	71.5714286	11.6666143
Low	37	62.1216216	16.0798379
Moderate	59	67.7288136	14.8600721

Exercise Comparison		Difference Between Means	95% Confidence Limits		
High	- Moderate	3.843	-4.962	12.648	
High	- Low	9.450	0.156	18.743	***
Moderate	- High	-3.843	-12.648	4.962	
Moderate	- Low	5.607	-0.604	11.818	
Low	- High	-9.450	-18.743	-0.156	***
Low	- Moderate	-5.607	-11.818	0.604	



Example: Model Selection

Select the 'best' model based on

- Results from bivariate analysis
- Scientific knowledge
- Significance of each predictor in the regression model
- SELECTION option in PROC REG
 - Provides 8 methods to select the final model
 - Mostly used: BACKWARD, FORWARD, STEPWISE



Example: Model Selection

```
PROC REG DATA=WEIGHT;  
BACKWARD: MODEL WEIGHT = HEIGHT AGE GENDER10 DRINKER EXERCISE_LOW  
                /SELECTION=BACKWARD SLSTAY=0.3;  
FORWARD: MODEL WEIGHT = HEIGHT AGE GENDER10 DRINKER EXERCISE_LOW  
                /SELECTION=FORWARD SLENTRY=0.3;  
INCLUDE: MODEL WEIGHT = HEIGHT AGE GENDER10 DRINKER EXERCISE_LOW  
                /SELECTION=BACKWARD SLSTAY=0.3 INCLUDE=1;  
RUN;
```

- SLSTAY: maximum p-value for a predictor to stay in the model
- SLENTRY: maximum p-value for a predictor to be added to the model
- INCLUDE: the first n predictors are forced to be included in the model



Example: Final Model

```
PROC REG DATA=WEIGHT;
MODEL WEIGHT = HEIGHT AGE GENDER10 EXERCISE_LOW ;
RUN;
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	12794	3198.43478	27.42	<.0001
Error	105	12247	116.63714		
Corrected Total	109	25041			

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-10.18400	13.10769	-0.78	0.4389
Height	1	0.36873	0.07105	5.19	<.0001
Age	1	0.37147	0.26717	1.39	0.1674
GENDER10	1	12.56047	2.28958	5.49	<.0001
EXERCISE_LOW	1	-3.35304	2.21628	-1.51	0.1333



Example: Model Diagnosis

```
PROC REG DATA=WEIGHT;  
MODEL WEIGHT = HEIGHT AGE GENDER10 EXERCISE_LOW /SPEC WHITE R VIF ;  
OUTPUT OUT=RESID RESIDUAL=RESID;  
PLOT RESIDUAL.*PREDICTED. ;  
PLOT RESIDUAL.*NQQ.;  
RUN;  
PROC UNIVARIATE DATA=RESID PLOT NORMAL;  
VAR RESID;  
RUN;
```

- SPEC: testing for constant variance
- WHITE: adjusting for non-constant variance
- R: residual analysis - identifying potential outliers
- VIF: variance inflation factor - collinearity problem when $VIF > 10$
- OUTPUT: saving the residuals to a separate dataset
- 1st PLOT: plotting residuals vs. predicted values - checking for constant variance
- 2nd PLOT: normal Q-Q plot of residuals
- PLOT and NORMAL in PROC UNIVARIATE: plot and test for normality



What if Model Assumptions are violated

- Non-normality and/or non-constant variance
 - Transformation on the response variable
 - Use more robust methods
- Collinearity exists
 - Remove one of the predictors with collinearity
 - Increase sample size if possible
 - Leave the model as is. Collinearity does not reduce the predictive power or reliability of the model as a whole, at least within the sample data themselves; it only affects calculations regarding individual predictors.
- Outliers
 - Verify the data
 - Remove the outliers if you are sure these values are unreasonable, especially when the sample is large
 - Keep the outliers, but treat them differently.



Example: Model Diagnosis

Normality does not hold.

Test	--Statistic--	-----p Value-----
Shapiro-Wilk	W 0.955559	Pr < W 0.0010
Kolmogorov-Smirnov	D 0.091738	Pr > D 0.0224
Cramer-von Mises	W-Sq 0.197098	Pr > W-Sq 0.0057
Anderson-Darling	A-Sq 1.254681	Pr > A-Sq <0.0050

Constant variance does not hold.

Test of First and Second Moment Specification		
DF	Chi-Square	Pr > ChiSq
12	22.90	0.0286

Scatter plot of residuals vs. predicted values indicates a log transformation or a power transformation with power less than 1 might help with the constant variance violation. In fact, it helps with the normality assumption too.



Example: After Log Transformation

Normality assumption holds.

Test	--Statistic---	-----p Value-----
Shapiro-Wilk	W 0.978217	Pr < W 0.0685
Kolmogorov-Smirnov	D 0.075347	Pr > D 0.1266
Cramer-von Mises	W-Sq 0.076499	Pr > W-Sq 0.2333
Anderson-Darling	A-Sq 0.495452	Pr > A-Sq 0.2182

Constant variance assumption holds.

Test of First and Second Moment Specification		
DF	Chi-Square	Pr > ChiSq
12	14.60	0.2639



Example: Final Model and Results

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	3.14185	0.78546	33.41	<.0001
Error	105	2.46888	0.02351		
Corrected Total	109	5.61073			

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	2.92409	0.18611	15.71	<.0001
Height	1	0.00614	0.00101	6.09	<.0001
Age	1	0.00555	0.00379	1.46	0.1464
GENDER10	1	0.18292	0.03251	5.63	<.0001
EXERCISE_LOW	1	-0.06121	0.03147	-1.95	0.0544



Introduction

Linear regression model applies when the outcome variable is continuous. What if the outcome variable is binary (0 or 1)? For example,

- whether a subject is a case of certain disease
- whether an individual successfully completed some task
- Yes/No answer to a survey question

Answer: Logistic Regression

SAS Procedures: PROC LOGISTIC, PROC GENMOD



Logistic Regression Model

$$\ln \left[\frac{P(Y = 1)}{1 - P(Y = 1)} \right] = \beta_0 + \beta_1 X_1 + \dots + \beta_m X_m$$

Assumptions:

- The log odds of $Y=1$ has linear relationships with predictors
- Observations are independent.

Interpretation of regression coefficient β_j :

The odds ratio (OR) for one unit increase in X_j is e^{β_j} .

Odds ratio: the ratio of odds in 2 different groups

Interpretation of OR:

- If $OR = 1$, then $P(Y = 1)$ is the same in both groups
- If $OR > 1$, then $P(Y = 1)$ is larger in numerator group than in denominator group.
- If $OR < 1$, then $P(Y = 1)$ is less in numerator group than in denominator group.



Example: Question and Data

A study is aimed to investigate the analgesic effects of treatments on elderly patients with neuralgia. Two test treatments and a placebo are compared. The response variable is whether the patient reported pain or not.

Variables:

- Pain: the response variable - Yes/No
- Treatment: a categorical variable with three levels - A, B, and P (placebo)
- Sex: gender of patients
- Age: age of patients, in years
- Duration: the duration of complaint, in months, before the treatment began



Example: A Tentative Model

```
PROC LOGISTIC DATA=NEURALGIA;
CLASS Treatment Sex;
MODEL Pain (event='No')= Treatment Sex Treatment*Sex Age Duration ;
RUN;
```

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	19.2236	7.1315	7.2661	0.0070
Treatment	A	1	0.8483	0.5502	2.3773	0.1231
Treatment	B	1	1.4949	0.6622	5.0956	0.0240
Sex	F	1	0.9173	0.3981	5.3104	0.0212
Treatment*Sex	A F	1	-0.2010	0.5568	0.1304	0.7180
Treatment*Sex	B F	1	0.0487	0.5563	0.0077	0.9302
Age		1	-0.2688	0.0996	7.2744	0.0070
Duration		1	0.00523	0.0333	0.0247	0.8752

event='No' indicates the probability of no pain is modeled. If Pain is coded with 0 (with pain) and 1 (no pain), then option DESCENDING should be specified in PROC LOGISTIC.



Example: Model Selection

```
PROC LOGISTIC DATA=NEURALGIA;  
CLASS Treatment Sex;  
MODEL Pain (event='No')= Treatment Sex Treatment*Sex Age Duration  
  /SELECTION=BACKWARD INCLUDE=1;  
RUN;
```

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	32.7358	4	<.0001
Score	25.6611	4	<.0001
Wald	14.5666	4	0.0057



Example: Model Selection

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	19.0804	6.7882	7.9007	0.0049
Treatment A	1	0.8772	0.5274	2.7662	0.0963
Treatment B	1	1.4246	0.6036	5.5711	0.0183
Sex F	1	0.9118	0.3960	5.3013	0.0213
Age	1	-0.2650	0.0959	7.6314	0.0057

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits
Treatment A vs P	24.022	3.295 175.121
Treatment B vs P	41.528	4.500 383.262
Sex F vs M	6.194	1.312 29.248
Age	0.767	0.636 0.926



Example: Model Diagnosis

```
PROC LOGISTIC DATA=NEURALGIA;  
CLASS Treatment Sex;  
MODEL Pain (event='No')= Treatment Sex Age /INFLUENCE LACKFIT ;  
RUN;
```

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
8.2913	8	0.4055

- **INFLUENCE**: produces regression diagnosis including Pearson and Deviance residuals which are expected to fall between -2 and 2 with 95% chance; can be used to identify potential outliers and evaluate model fit.
- **LACKFIT**: generates the Hosmer and Lemeshow Goodness-of-Fit Test; the null hypothesis is the model provides an adequate fit.



Example: Model Predictive Ability

Association of Predicted Probabilities and Observed Responses

Percent Concordant	90.3	Somers' D	0.811
Percent Discordant	9.1	Gamma	0.816
Percent Tied	0.6	Tau-a	0.401
Pairs	875	c	0.906

- Somers'D, Gamma, Tau-a, and c measure the correlation between the predicted probabilities and the observed dichotomous response variable.
- A value of -1 or +1 indicates perfect agreement; zero indicates no agreement.



More on Model Predictive Ability

Four aspects of predictive ability (borrow terminology from disease screening tests):

- Sensitivity: the probability that screening test is positive given that the person has the disease.
- Specificity: the probability that screening test is negative given that the person does not have the disease.
- Positive predictive value: the probability that a person has the disease given a positive test result.
- Negative predictive value: the probability that a person does not have the disease given a negative test.

Receiver Operating Characteristic (ROC) curve:

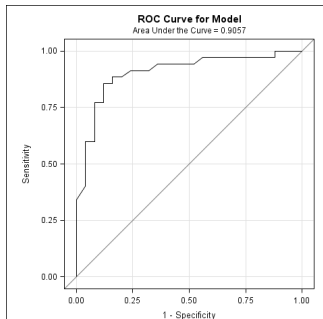
- a graphical plot of the sensitivity, or true positive rate, vs. false positive rate ($1 - \text{specificity}$)
- if a model fits well, we expect a ROC curve with area under curve greater than 0.5



Example: Model Predictive Ability

```
ODS GRAPHICS ON;  
PROC LOGISTIC DATA=NEURALGIA;  
CLASS Treatment Sex;  
MODEL Pain (event='No')= Treatment Sex Age / OUTROC=ROC ;  
RUN;  
ODS GRAPHICS OFF;
```

Figure: ROC Curve



General Linear Models

Linear regression models assume independent data, but it is often in practice that the outcome variables are repeated (or multiple) measures on the same individuals. If the outcome variables are continuous, then this is where general linear models apply.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_m X_m + \varepsilon$$

Assumptions:

- Relationships between the response and predictors are linear
- Errors are normally distributed
- Errors have a common variance
- However, the errors are now allowed to be correlated

SAS Procedures: PROC GLM and PROC MIXED



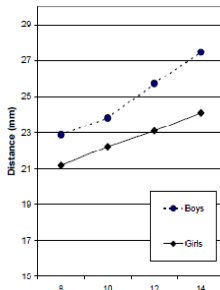
Classical Methods for Repeated Measures Data

- 1 **Multivariate Analysis of Variance (MANOVA) and Repeated Measures ANOVA**
 - Tests for treatment effect;
 - Comparisons between time points;
 - Can only incorporate time-independent covariates;
 - MANOVA assumes unstructured correlation matrix; repeated measures ANOVA assumes compound symmetric (or exchangeable) structure;
 - Implemented by PROC GLM in SAS.
- 2 **More Advanced Methods: Linear Mixed Models**
 - Tests for treatment effect;
 - Modelling the relationship between the response and time;
 - Can incorporate both time-dependent and time-independent covariates;
 - Implemented by PROC MIXED in SAS.



Example: Potthoff and Roy Growth Data (1964)

A study was conducted involving 27 children, 16 boys and 11 girls. On each child, the distance (mm) from the center of the pituitary to the pterygomaxillary fissure was made at ages 8, 10, 12, and 14 years of age. The objectives of the study were to (1) Determine whether distances over time are larger for boys than for girls; (2) Determine whether the rate of change of distance over time is similar for boys and girls.



Example: MANOVA Test for Gender Effect

```

PROC GLM DATA=prdentals;
CLASS Gender;
MODEL Dist8 Dist10 Dist12 Dist14 = Gender ;
MANOVA H=Gender;
RUN;

```

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.60230061	3.63	4	22	0.0203
Pillai's Trace	0.39769939	3.63	4	22	0.0203
Hotelling-Lawley Trace	0.66030051	3.63	4	22	0.0203
Roy's Greatest Root	0.66030051	3.63	4	22	0.0203



Assumption for Repeated Measures ANOVA

Suppose there are 3 time points. Let $\mathbf{Y}_{hi} = (Y_{hi1}, Y_{hi2}, Y_{hi3})'$, then

$$\text{var}(\mathbf{Y}_{hi}) = \begin{bmatrix} \sigma_b^2 + \sigma_e^2 & \sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 + \sigma_e^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 + \sigma_e^2 \end{bmatrix}$$

This is the so-called compound symmetric or exchangeable covariance structure.

In SAS, Mauchly's test for sphericity (compound symmetry) can be requested by specifying the PRINTE option in the REPEATED statement



What If the Sphericity Assumption is Unreasonable?

- Use the unstructured MANOVA approach.
- Adjust the degrees of freedom of the ANOVA F to make the test perform better.
 - Greenhouse-Geisser (G-G) adjustment
 - Hunyh and Feldt (H-F) adjustment



Example: Repeat Measures ANOVA

```
PROC GLM DATA=prdentals;
CLASS Gender ;
MODEL Dist8 -- Dist14 = Gender / NOUNI;
REPEATED Age 4 (8 10 12 14) / PRINTE ;
RUN;
```

Sphericity Tests

Variables	DF	Mauchly's Criterion	Chi-Square	Pr > ChiSq
Transformed Variates	5	0.4998695	16.449181	0.0057
Orthogonal Components	5	0.7353334	7.2929515	0.1997



Example: Repeated Measures ANOVA

Repeated Measures Analysis of Variance Tests of Hypotheses for Between Subjects Effects

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Gender	1	140.4648569	140.4648569	9.29	0.0054
Error	25	377.9147727	15.1165909		

Repeated Measures Analysis of Variance Univariate Tests of Hypotheses for Within Subject Effects

Source	DF	Type III SS	Mean Square	F Value	Pr > F	Adj Pr > F	G - G	H - F
Age	3	209.4369739	69.8123246	35.35	<.0001	<.0001	<.0001	<.0001
Age*Gender	3	13.9925295	4.6641765	2.36	0.0781	0.0878	0.0781	0.0781
Error (Age)	75	148.1278409	1.9750379					

Greenhouse-Geisser Epsilon 0.8672
Huynh-Feldt Epsilon 1.0156



Example: Repeated Measures ANOVA

Repeated measures ANOVA also generates MANOVA testing for the Age effect and the interaction effect of Age and Gender.

MANOVA Test for the Hypothesis of no Age Effect

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.19479424	31.69	3	23	<.0001
Pillai's Trace	0.80520576	31.69	3	23	<.0001
Hotelling-Lawley Trace	4.13362211	31.69	3	23	<.0001
Roy's Greatest Root	4.13362211	31.69	3	23	<.0001

MANOVA Test for the Hypothesis of no Age*Gender Effect

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.73988739	2.70	3	23	0.0696
Pillai's Trace	0.26011261	2.70	3	23	0.0696
Hotelling-Lawley Trace	0.35155702	2.70	3	23	0.0696
Roy's Greatest Root	0.35155702	2.70	3	23	0.0696



More Models and SAS Procedures

- 1 Generalized Linear Models: Non-Normal response, eg, binary outcome, counts, etc.
 - PROC GENMOD: can handle both independent and dependent (say, repeated measures) data
- 2 Linear Mixed Models: Normal response, with both fixed and random effects
 - PROC MIXED
- 3 Generalized Linear Mixed Models: Non-Normal response, with both fixed and random effects
 - PROC GLIMMIX
- 4 Non-Linear Models: Non-linear relationship
 - PROC NLIN
- 5 Non-Linear Mixed Models: Non-linear relationship, with both random and fixed effects
 - PROC NLMIXED
- 6 Generalized Additive Models: Normal or non-normal response, unknown form of relationship
 - PROC GAM

