
# Part 2
# Lecture 1 The Statistical Model

1

# Who we are…


Pascal Tyrrell, PhD *Associate Professor*

Department of Medical Imaging , Faculty of Medicine
Department of Statistical Sciences , Faculty of Arts and Science

Paul Corey, PhD *Professor Emeritus*

Biostatistics Program, Dalla Lana Faculty of Public Health
Institute of Medical Science, Faculty of Medicine
Department of Statistical Sciences, Faculty of Arts and Science

# CREATING A STATISTICAL MODEL

A researcher is studying the relationship between the decrease in a person's blood pressure and personal and environmental variables, to which they are exposed.

Knowledge from observational studies and laboratory findings suggest that the decrease in a person's blood pressure is related to their sex, age, and a drug believed to reduce blood pressure.

# MODEL FOR DECREASE IN BLOOD PRESSURE (DBP)

$$MODEL \quad DBP_j = K + ERROR_j \qquad j = 1 \; to \; n$$

$$ERROR_j = DBP_j - K$$

$$Sum \; of \; Squares \; SS = \sum_{j=1}^{j=n} (DBP_j - K)^2$$

The Sum of Squared Deviations about K is a minimum if K is equal to the sample mean $\overline{\text{DBP}}$.

Therefore the sample mean is called the least squares estimate of K.

## MODEL FOR DECREASE IN BLOOD PRESSURE (DBP)

$$MODEL \quad DBP_j = \mu + ERROR_j \quad j = 1 \text{ to } n$$

$$ERROR_j = DBP_j - \mu \qquad Residual = DBP_j - \overline{DBP}$$

$$\text{Sample Variance } s^2 = \frac{\sum_{j=1}^{j=n} (DBP_j - \overline{DBP})^2}{n-1}$$

*The Sum of Squared Deviations about the sample mean divided by $(n-1)$ is an unbiased estimator of the variance $\sigma^2$ that is in the formula of the Gaussian probability density function.*

# STRAIGHT LINE MODEL

$$DBP_j = \beta_0 + \beta_1 \times AGE_j + ERROR_j$$

$$ERROR_j = DBP_j - \beta_0 - \beta_1 \times AGE_j$$

$$j = 1 \; to \; n$$

# FITTING A STRAIGHT LINE

*Least square estimates of the intercept and slope of a straight line model*

$$\widehat{\beta_1} = \frac{\sum_{j=1}^{j=n} (AGE_j - \overline{AGE}) \times DBP_j}{\sum_{j=1}^{j=n} (AGE_j - \overline{AGE})^2} \qquad \widehat{\beta_0} = \overline{DBP} - \widehat{\beta_1} \times \overline{AGE}$$

$$RESIDUAL_j = OBSERVED - PREDICTED = DBP_j - \widehat{\beta_0} - \widehat{\beta_1} \times AGE_j$$

$$Sample\ Variance \qquad s^2 = \frac{\sum_{j=1}^{j=n} (RESIDUAL_j)^2}{n-2}$$

*The Sum of Squared Deviations about the sample mean divided by* $(n-2)$ *is an unbiased estimator of the variance* $\sigma^2$ *that is in the formula of the Gaussian probability density function.*

# MULTIPLE LINEAR REGRESSION

$$DBP_j = \beta_0 + \beta_1 \times DRUG_j + \beta_2 \times SEX_j + \beta_3 \times AGE_j + ERROR_j$$

$j = 1\ to\ n$
$DRUG\ variable\ can\ assume\ values\ Aspirin\ and\ Tynlenol$
$SEX\ variable\ can\ assume\ values\ Female\ and\ Male$

$$Predicted\ \ DBP_j = \widehat{DBP_j} = \widehat{\beta_0} + \widehat{\beta_1} \times DRUG_j + \widehat{\beta_2} \times SEX_j + \widehat{\beta_3} \times AGE_j$$

$$RESIDUAL_j = DBP_j - \widehat{DBP_j} \qquad Sample\ Variance\ \ s^2 = \frac{\sum_{j=1}^{j=n}(RESIDUAL_j)^2}{n-4}$$

$Sum\ of\ Squared\ Residuals\ divided\ by\ (n-4)\ is\ an\ unbiased\ estimator\ of\ the\ variance\ \sigma^2\ that\ is\ in\ the\ formula\ of\ the\ Gaussian\ probability\ density\ function.$

# DRUG SEX INTERACTION IN MULTIPLE LINEAR REGRESSION

$$DBP_j = \beta_0 + \beta_1 \times DRUG_j + \beta_2 \times SEX_j + \beta_3 \times DRUG \times SEX + \beta_4 \times AGE_j + ERROR_j$$

$j = 1$ to n

$$Predicted\ DBP_j = \widehat{DBP_j} = \widehat{\beta_0} + \widehat{\beta_1} \times DRUG_j + \widehat{\beta_2} \times SEX_j + \widehat{\beta_3} \times DRUG \times SEX + \widehat{\beta_4} \times AGE_j$$

$j = 1$ to n

$$RESIDUAL_j = DBP_j - \widehat{DBP_j} \qquad Sample\ Variance\ s^2 = \frac{\sum_{j=1}^{j=n}(RESIDUAL_j)^2}{n-5}$$

*If the estimate $\widehat{\beta_3}$ is large it may mean that there is DRUG * SEX interaction which means that the size and possibly also the sign of the drug effect is different among males and females. Sample variance $s^2$ is unbiased estimator of $\sigma^2$.*

What values should these betas have to minimize the error? Good question! All statistical programs produce values for these unknown betas so that the **variation** among the error terms is minimized. How is the variation measured? A statistic called the variance measures the variation among the error terms. Values given to the beta constants will be such that the sum of the error terms is zero and their variance is minimized. Is this variance an estimate of the variance that is in the Gaussian probability model. **YES !!!!!!**

```
TITLE1  "  COMPARING SAME MEANS USING GLM PROCEDURE  "  ;
DATA  STUDY ; INPUT  COLOUR $  NAME $   ID   RTIME ;
DATALINES ;
GREEN      ABEL     1       232.6
RED        ABEL     1       232.0
GREEN      ADAM     2       257.5
RED        ADAM     2       250.5
GREEN      AMOS     3       253.1
RED        AMOS     3       237.1
GREEN      ANDY     4       205.4
RED        ANDY     4       201.5
GREEN      BART     5       226.0
RED        BART     5       211.1
RUN;    ** NOTE: MOST DATASETS HAVE A LINE OF DATA FOR EACH SUBJECT;
```

# USING GLM PROCEDURE TO COMPARE TWO OR MORE SAMPLE MEANS

```
TITLE1  "  ASSUMING A COMPLETELY RANDOMIZED DESIGN  "   ;
PROC  GLM  DATA = STUDY ;  CLASS COLOUR ;
MODEL  RTIME = COLOUR ;
LSMEANS  COLOUR / TDIFF  PDIFF  STDERR  CL ;  RUN  ;

TITLE1  "  ASSUMING A RANDOMIZED BLOCK DESIGN  "   ;
PROC  GLM  DATA = STUDY ;  CLASS COLOUR  ID ;
MODEL  RTIME  =  COLOUR ID ;  **Note  ID in MODEL statement ;
LSMEANS  COLOUR / TDIFF  PDIFF  STDERR  CL  ;  RUN  ;
```

NOTE: ID Variable can be replaced by the NAME variable in the CLASS and MODEL statements.

# ASSUMING A COMPLETELY RANDOMIZED DESIGN

The GLM Procedure

Dependent Variable: RTIME

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 1 | 179.776000 | 179.776000 | 0.43 | 0.5323 |
| Error | 8 | 3377.500000 | 422.187500 | | |
| Corrected Total | 9 | 3557.276000 | | | |

| R-Square | Coeff Var | Root MSE | RTIME Mean |
|---|---|---|---|
| 0.050538 | 8.907232 | 20.54720 | 230.6800 |

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| COLOUR | 1 | 179.7760000 | 179.7760000 | 0.43 | 0.5323 |

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| COLOUR | 1 | 179.7760000 | 179.7760000 | 0.43 | 0.5323 |

MY NOTE: Square Root 0.426 = 0.65 and before we had t = 0.65

< SAME AS BEFORE !

## The GLM Procedure
## Least Squares Means

| COLOUR | RTIME LSMEAN | Standard Error | H0:LSMEAN=0 | H0:LSMean1=LSMean2 | |
|---|---|---|---|---|---|
| | | | Pr > \|t\| | t Value | Pr > \|t\| |
| GREEN | 234.920000 | 9.188988 | <.0001 | 0.65 | 0.5323 |
| RED | 226.440000 | 9.188988 | <.0001 | | |

| COLOUR | RTIME LSMEAN | 95% Confidence Limits | |
|---|---|---|---|
| GREEN | 234.920000 | 213.730156 | 256.109844 |
| RED | 226.440000 | 205.250156 | 247.629844 |

**Least Squares Means for Effect COLOUR**

| i | j | Difference Between Means | 95% Confidence Limits for LSMean(i)-LSMean(j) | |
|---|---|---|---|---|
| 1 | 2 | 8.480000 | -21.486965 | 38.446965 |

**MY NOTE: Confidence Interval includes zero !!**

## ASSUMING A RANDOMIZED BLOCK DESIGN

### The GLM Procedure

### Dependent Variable: RTIME

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 5 | 3465.762000 | 693.152400 | 30.30 | 0.0028 |
| Error | 4 | 91.514000 | 22.878500 | | |
| Corrected Total | 9 | 3557.276000 | | | |

| R-Square | Coeff Var | Root MSE | RTIME Mean |
|---|---|---|---|
| 0.974274 | 2.073499 | 4.783147 | 230.6800 |

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| COLOUR | 1 | 179.776000 | 179.776000 | 7.86 | 0.0487 |
| ID | 4 | 3285.986000 | 821.496500 | 35.91 | 0.0022 |

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| COLOUR | 1 | 179.776000 | 179.776000 | 7.86 | 0.0487 |
| ID | 4 | 3285.986000 | 821.496500 | 35.91 | 0.0022 |

Matched Pairs Design

<< MY NOTE: 3465.76 / 3557.27 = 0.974

<< Same as before

The GLM Procedure
Least Squares Means

| COLOUR | RTIME LSMEAN | Standard Error | H0:LSMEAN=0 | H0:LSMean1=LSMean2 | |
|---|---|---|---|---|---|
| | | | Pr > \|t\| | t Value | Pr > \|t\| |
| GREEN | 234.920000 | 2.139089 | <.0001 | 2.80 | 0.0487 |
| RED | 226.440000 | 2.139089 | <.0001 | | |

| COLOUR | RTIME LSMEAN | 95% Confidence Limits | |
|---|---|---|---|
| GREEN | 234.920000 | 228.980938 | 240.859062 |
| RED | 226.440000 | 220.500938 | 232.379062 |

Least Squares Means for Effect COLOUR

| i | j | Difference Between Means | 95% Confidence Limits for LSMean(i)-LSMean(j) | |
|---|---|---|---|---|
| 1 | 2 | 8.480000 | 0.080898 | 16.879102 |

**MY NOTE: 95% Confidence Interval does not contain zero.**

# WHAT IS THE ADVANTAGE OF GENERAL LINEAR MODEL (GLM) OVER THE TTEST PROCEDURE ??

```
TITLE1 " COMPARING TWO DRUGS ";
PROC GLM DATA=STUDY; CLASS DRUG SEX ;
MODEL RTIME =
DRUG   SEX   DRUG * SEX AGE WEIGHT / SS3;

LSMEANS DRUG/TDIFF PDIFF STDERR CL;
LSMEANS SEX /TDIFF PDIFF STDERR CL; RUN ;
```

Reaction time variable RTIME  is OUTCOME variable.

PRIMARY QUESTION:  does the variable DRUG predict OUTCOME ?  In other words are the DRUG and OUTCOME variables associated ?

If SEX, AGE and WEIGHT predict OUTCOME including them reduces the residual variance, increases $R^2$ and reduces their p values. If the effect of drug is greater for males than females that is called interaction ( DRUG*SEX ).

If the predictor variables SEX, AGE and WEIGHT are also associated with the DRUG variable then excluding them produces biased estimates of the DRUG effect. They are then called CONFOUNDERS.

Product variable DRUG * SEX is called INTERACTION. If it is large it means that the size of the DRUG effect is different for males and females.


IF I DROP BOTH OF THESE, WHICH DO YOU THINK WILL LAND FIRST?
©1983 CREATORS SYNDICATE, INC.
3-1
A FIN SAYS THE ROCK HITS FIRST.
BLAP
BINK
I FORGOT. HE LIVES IN A VACUUM.

In the cartoon the loser assumed his friend would let go of the rock and feather at the same time. That was not part of the bet.

In many studies researchers may mistakenly think that the comparison was fair and valid. In a study the proportion of males in the exposed and unexposed groups may be quite different AND if males are at higher risk of disease the comparison would be biased.

# PROCEDURES SIMILAR TO GLM USED FOR ALL 4 OUTCOMES

| THREE DESIGNS | COMPLETELY RANDOMIZED | RANDOMIZED BLOCK SPLIT PLOT | |
|---|---|---|---|
| TREATMENT LAYOUT | ONE WAY | | 2 BY 2 FACTORIAL |
| OUTCOME VARIABLE | CONTINUOUS | Blood Pressure | Serum Cholesterol |
| | BINARY | Death Yes/No | Cure Yes/No |
| | COUNT | Number of deaths | Number of falls |
| | SURVIVAL TIME | Time to Death | Time to Cure |


# End of Lecture 1

*Next up in Part 2 Lecture 2: Study Design*