

BIOSTATISTICS METHODS

FOR TRANSLATIONAL & CLINICAL RESEARCH



DAY #4: REGRESSION APPLICATIONS, PART C

MULTILE REGRESSION APPLICATIONS

This last part is devoted to Multiple Regression applications; it covers two topics:

- (1) For interactions, we **re-visit and expands the topic of bioassays**, and
- (2) As an example of non-linear models, I'll show you how to study “**seasonal diseases**”- a case similar to that of quadratic regression - with **two predictor terms representing the same “predictor source”** where we search for an optimal condition for the outcome.

DEFINITION

“Biological assays” or “**bioassays**” are a set of methods for estimating the potency or strength of an “**agent**” by utilizing the “**response**” caused by its application to biological material or experimental living “**subjects**”.

COMPONENTS OF A BIOASSAY

- The **subject** is usually an animal, a human tissue, or a bacteria culture,
- The **agent** is usually a drug,
- The response is usually a change in a particular characteristic or even the death of a subject; the response could be binary or on **continuous** scale.

DIRECT ASSAYS

- In **direct assays**, the doses of the standard and test preparations are “directly measured” for an “event of interest”.
- When an (pre-determined) event of interest occurs, e.g.. the death of the subject, and **the variable of interest is the **dose**** required to produce that response/event for each subject.
- In other words: (Binary) **Response is fixed, Dose is a Random Variable.**

INDIRECT ASSAYS

- In **indirect assays**, the doses of the standard and test preparations are applied and we observe the response that each dose produces; for example, we measure the tension in a tissue or the hormone level or the blood sugar content. For each subject, the dose is fixed in advance, the variable of interest is not the dose but the **response** it produces in each subject.
- **Doses are fixed, Response is a Random Variable**; statistically, indirect assays are more interesting and also more difficult.

For Indirect Assays, depending on the “**measurement scale**” for the response – our Random Variable, we have:

- (1) **Quantal assays**, where **the response is binary**: whether or not an event (like the death of the subject) occur,
- (2) **Quantitative assays**, where measurements for the response are on a **continuous** scale. **This is our main focus; the dependent variable Y.**

The common indirect assay is usually one in which the ratio of equipotent doses is estimated from “curves” or “lines” relating quantitative responses and doses for the two preparations.

The shape of these curves or lines further divides Quantitative Indirect Assays into:

(1) **Parallel-line assays** are those in which the response is linearly related to the log dose,

(2) **Slope ratio assays** are those in which the response is linearly related to the dose itself.

PARALLEL-LINE ASSAYS

- Parallel-line assays are those in which the response is linearly related to the log dose.
- From the definition of “relative potency” ρ , the two equipotent doses are related by $D_S = \rho D_T$.
- The model: $E[Y_S | X_S = \log(D_S)] = \alpha + \beta X_S$, for Standard and, for same dose of Test we have $E[Y_T | X_S = \log(D_S = \rho D_T)] = (\alpha + \beta \log \rho) + \beta X_T$
- We have 2 parallel lines with a **common slope β** and **different intercepts**.

The primary aim of a statistical analysis is to estimate the “relative potency” of an agent or stimulus ; a point estimate as well as confidence limits (i.e. 95% confidence interval). These Estimate are obtained as a result of the “statistical model” – two lines with a common slope and different intercepts.

POINT ESTIMATE

- The model: $E[Y_S|X_S=\log(D_S)] = \alpha + \beta X_S$, for Standard and, for the same dose we have $E[Y_T|X_S=\log(D_S=\rho D_T)] = (\alpha + \beta \log \rho) + \beta X_T$
- Result: We can estimate of $\log \rho$, called M , by subtracting the intercepts and divided by the common slope; let m be a point estimate of M .
- Then we can obtain a point estimate r of ρ by exponentiating “ m ”.

$$\text{Intercept}_S = \alpha$$

$$\text{Intercept}_T = \alpha + \beta M$$

$$M = \frac{\text{Intercept}_T - \text{Intercept}_S}{\text{Common Slope}};$$

$M = \log \rho$ is estimated by m

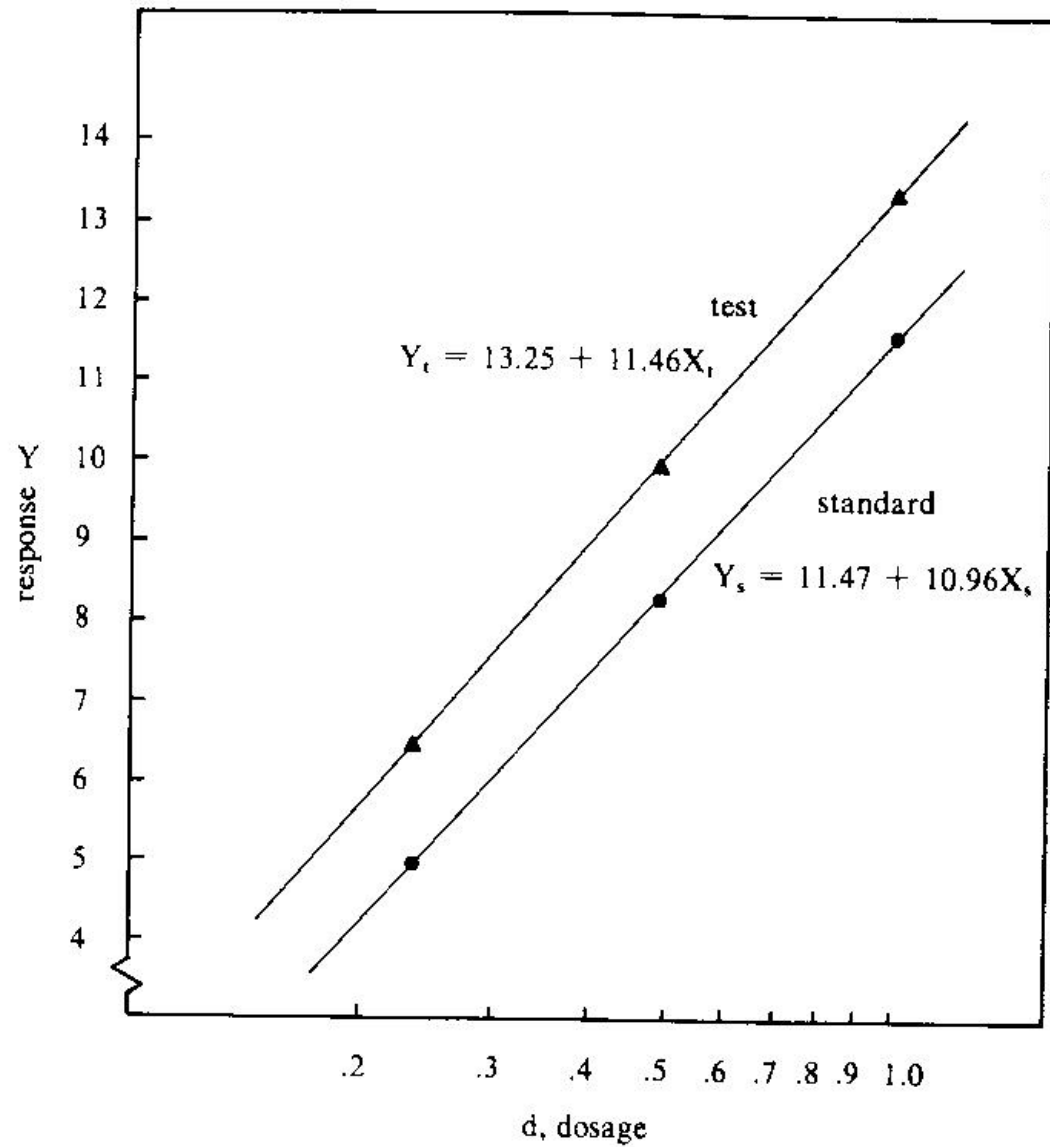
EXAMPLE

- In this example, the test and standard preparations of the agent are tested at the same three dose levels (.25, .50, and 1.0 mg/cc); and there are 8 replications at each dose of each preparation.
- It is designed with 8 dishes/plates, each contains 6 identical bacterial cultures - one in a “well” (randomized complete block design), also called “6-point assay”; the response was the amount of decrease in growth.

	Preparation					
	Standard Preparation			Test Preparation		
Dose (D; mmgcc)	0.25	0.50	1.00	0.25	0.50	1.00
$X = \log_{10}(\text{Dose})$	-0.602	-0.301	0.000	-0.602	-0.301	0.000
Response (Y; mm)	4.9	8.2	11.0	6.0	9.4	12.8
	4.8	8.1	11.5	6.8	8.8	13.6
	4.9	8.1	11.4	6.2	9.4	13.4
	4.8	8.2	11.8	6.6	9.6	13.8
	5.3	7.6	11.8	6.4	9.8	12.8
	5.1	8.3	11.4	6.0	9.2	14.0
	4.9	8.2	11.7	6.9	10.8	13.2
	4.7	8.1	11.4	6.3	10.6	12.8

Graph of the “response” versus “log dose” follows

The Indirect Quantitative Assays



Quantitative parallel line assay responses against log doses.

Doing correctly, we should fit the **two straight lines with a common slope**. Here, each line was fitted separately – not right but can use to see if data fit the model.

When we learn Simple Linear Regression, **we solved the problem by calculating the weighted average of the two estimated slopes**. Another approach, which turns out more simple, is Multiple Linear Regression.

MULTIPLE REGRESSION

- An alternative approach is pooling data from both preparations and using “Multiple Regression”;
- Dependent Variable: **Y = Response**;
Two Independent Variables or predictors are:
X = log(Dose) &
P = Preparation (a “dummy variable” coded as $P = 1$ for “Test” and $P = 0$ for “Standard”)

From the Model :

$$E(Y) = \beta_0 + \beta_1 X + \beta_2 P$$

Standard Preparation :

$$E(Y_s) = \beta_0 + \beta_1 X$$

Test Preparation :

$$\begin{aligned} E(Y_T) &= \beta_0 + \beta_1 X + \beta_2 \\ &= (\beta_0 + \beta_2) + \beta_1 X \end{aligned}$$

Difference of Intercepts = β_2

Common Slope = β_1

Multiple Regression Model :

$$E(Y) = \beta_0 + \beta_1 X + \beta_2 P$$

β_1 is the common slope and

β_2 is the "difference of intercepts " ;

$$M = \log \rho = \frac{\beta_2}{\beta_1}$$

$X = \log(\text{Dose})$ & $P = \text{Preparation}$

DELTA METHOD

If Y is a function of two random variables X_1 and X_2 then we have approximately, with partial derivatives evaluated at the mean values:

$$Var(Y) \cong \left[\frac{\partial y}{\partial x_1} \right]^2 Var(X_1) + 2 \left[\frac{\partial y}{\partial x_1} \right] \left[\frac{\partial y}{\partial x_2} \right] Cov(X_1, X_2) + \left[\frac{\partial y}{\partial x_2} \right]^2 Var(X_2)$$

In the current application: **The point estimate of $\log(\text{Relative Potency})$ is b_2/b_1**

Multiple Regression Model :

$$E(Y) = \beta_0 + \beta_1 X + \beta_2 P$$

β_1 is the common slope and

β_2 is the "difference of intercepts " ;

$$M = \log \rho = \frac{\beta_2}{\beta_1} ; m = \frac{b_2}{b_1}$$

$$Var(m) \cong \frac{b_2^2}{b_1^4} Var(b_1) + 2\left(-\frac{b_2}{b_1^2}\right)\left(\frac{1}{b_1}\right)Cov(b_1, b_2) + \frac{1}{b_1^2} Var(b_2)$$

SLOPE RATIO ASSAYS

- Slope-ratio assays are those in which the response is linearly related to the dose itself.
- From the definition of “relative potency” ρ , the two equipotent doses are related by $D_S = \rho D_T$.
- **The model**: $E[Y_S | X_S = D_S] = \alpha + \beta X_S$, for its equipotent dose $E[Y_T | X_S = D_S] = \alpha + \beta \rho X_T$; the lines have the same intercept - the mean response at zero dose.
- **Result**: We have two straight lines with a **common intercept and different slopes**.

The **primary** aim of a statistical analysis is to estimate the “**relative potency**” ρ of an agent or stimulus ; a point estimate as well as confidence limits (i.e. 95% confidence interval). These Estimate are obtained as a result of the “**statistical model**” – now a different model, **two straight lines with a common intercept but different slopes.**

POINT ESTIMATE

- The model: $E[Y_S|X_S=D_S)] = \alpha + \beta X_S$, for same dose $E[Y_T| X_S=D_S= \alpha + \beta \rho X_T]$; the lines have the same intercept α - the mean response at zero dose. Slopes are different by a factor ρ .
- Result: We can obtain ρ , the relative potency, as the **ratio of two slopes**; then we can obtain a point estimate r as: $r = b_T / b_S$.

Common Intercept = α

$$\text{Slope}_S = \beta$$

$$\text{Slope}_T = \beta\rho$$

$$\rho = \frac{\text{Slope}_T}{\text{Slope}_S};$$

$$\rho \text{ is estimated by } r = \frac{b_T}{b_S}$$

Unlike the case of Parallel Line Bioassays, where we **could** (weighted) average the two slopes. The of Slope Ratio Bioassays are different; **even if we want, we cannot use Simple Linear Regression** to fit two lines with a common intercept. We have to turn to get help from Multiple Linear Regression; the good news is we can do it in more than one way.

MULTIPLE REGRESSION

- An alternative approach is pooling data from both preparations and using “Multiple Regression”;
- Dependent Variable: **Y = Response**;
Two Independent Variables are:
X = Dose &
P = Preparation (a “dummy variable” coded as $P = 1$ for “Test” and $P = 0$ for “Standard”)

Multiple Regression Model #1:

$$E(Y) = \beta_0 + \beta_1 X + \beta_2 PX$$

β_0 is the common intercept and slopes are :

$$\beta_T = \beta_1 + \beta_2$$

$$\beta_S = \beta_1$$

$$\rho = \frac{\beta_1 + \beta_2}{\beta_1}$$

$X = \text{Dose} \ \& \ P = \text{Preparation}$

MULTIPLE REGRESSION #2

Let Y be the response, X_S and X_T the doses.
Consider the following model in which for any observation on S, set $X_T=0$, for any observation on T, set $X_S=0$; the model may include control observations for which we set $X_S = X_T = 0$:

$$E(Y) = \beta_0 + \beta_S X_S + \beta_T X_T;$$

β_0 = Common Intercept

$$\rho = \beta_T / \beta_S$$

In both Multiple Regression models for Slope-Ratio assays, point estimate of Relative potency is obtained from computer output; and **we can use Delta Method to calculate Variance/Standard Error.** That is possible with the option “COVB” in PROC REG.

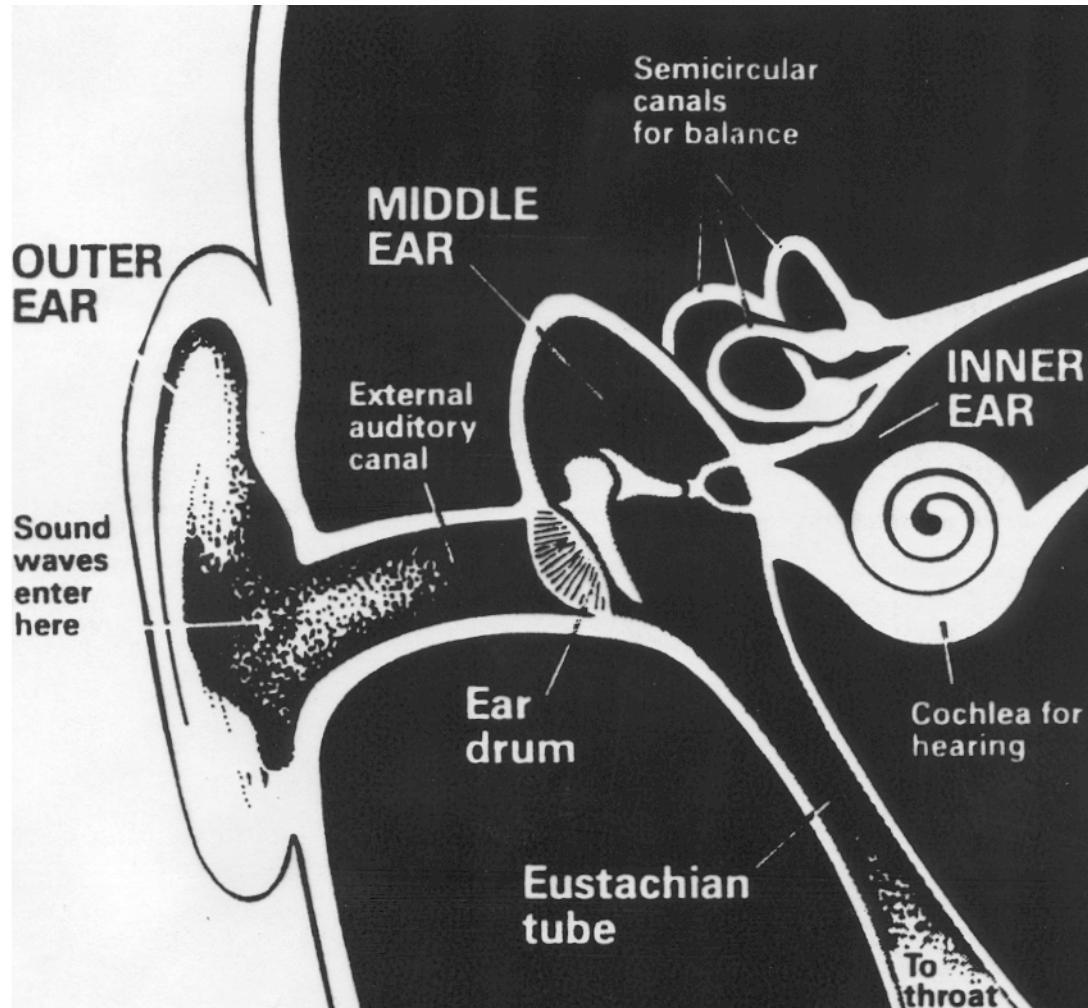
POLYNOMIAL REGRESSION

- The second-order or quadratic model could be used when the true relationship may be unknown but the second degree polynomial provides a better fit than a linear one.
- If a quadratic model fits, perhaps an useful application would be **to maximize/minimize the “Mean Index”**, $\beta_1 x + \beta_2 x^2$, in order to determine the value of X at which the **Mean of Y attains its maximum or minimum value** (depending on the sign of β_2).
- The following application is based on this idea of optimization involving **two related predictor variables**.

A MYTH OR A SCIENCE?

- Just like “Palm Reading”, Horoscope is very popular in all continents!
- Believers say that your DOB partially determines your life: How happy you are, How successful you are, How healthy you are; Even revealing something about your spouse and your kids !
- So, let first focus on the aspect of health.

OTITIS MEDIA: INFLAMMATION OF THE MIDDLE EAR



OTITIS MEDIA

- Is Inflammation of the middle-ear space, often referred to as “Children Ear Infection”.
- Is the **2nd most prevalent disease on earth**, affects 90% of children by age 2 (in the US).
- **Costs 3.8 billions in direct costs** (physician visits, tube placements, antibiotics, etc...) and 1.2 billions in indirect costs (lost works by mothers, etc...); in 1995 dollars.
- Causes hearing loss, **learning disabilities**, and other middle-ear sequelae.

AS A CHILDREN DISEASE

- It is the most common diagnosis at physician visits ahead of well-child, URI, injury, and sore throat.
- It is responsible for 24.5 million physician visits in 1990, increased 150% from 1975; probably due to increased awareness and more aggressive diagnoses.

Perhaps it is the most typical “public health problem”; and its most interesting characteristic is that it is a “**seasonal disease**” – just like the most prevalent disease on earth, the **common cold**.

DOB AS A DISEASE INDICATOR

- Other examples
- Nilsson et. al. Season of births as predictor of **atopic manifestations**. *Archives of Disease in Childhood*, 1997. (food allergies, asthma, etc...)
- Torrey et. al. Seasonal birth patterns of **neurological disorders**. *Neuro-epidemiology*, 2000.

LITERATURE REVIEWS

- Most or all investigators, with emphasis on “season”, often **divide the year into four seasons** and compare them using either a Chi-square test or a One-way Analysis of Variance F-test depending on the endpoint being discrete or continuous!
- What are the **problems** with this vastly **popular** approach to investigating seasonality?

POINTS TO CONSIDER

- There is no universal agreement on the definition of seasons, plus regional differences (for example, the Midwest's winter is more than 3 months; we have 2 seasons a year, **Winter & Road Repair**).
- There are **no reasons to believe** that the risk associated with time of the year is similar within seasons and different between seasons. If there is any risk associated with time of the year, that **level of risk must change gradually** , not abruptly, as time progresses.

A NEWER APPROACH

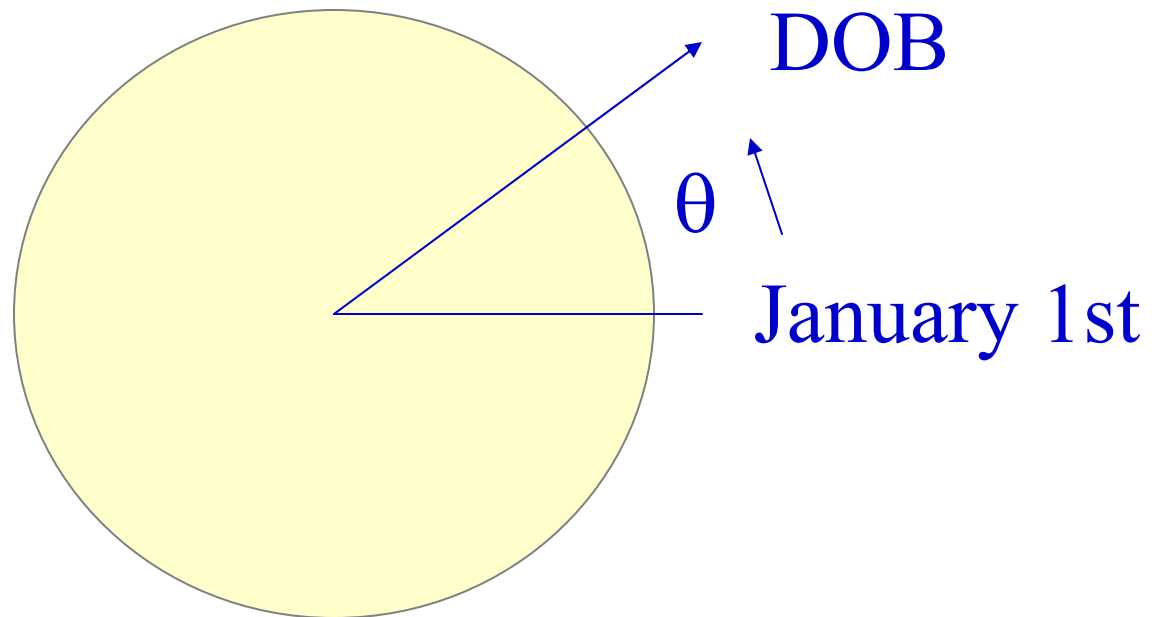
- DOB will be treated as a continuous variable with a circular distribution.
- Many endpoints may be considered but let focus on a continuous endpoint (representing a “possible cause”) as a show case of MLR
- I like to see this as a study of an important public health problem (a disease with high prevalence) focussing on a widespread exposure (**indoor pollution**).

DATA FOR ILLUSTRATION

- Study population: 592 infants within Health Partners system. Pregnant women were enrolled, and followed. **Cord blood** was collected at time of birth and **blood samples** were analyzed by ELISA for **pneumococcal IgG antibody** for 7 serotypes.
- Specific Aims: we focus on **DOB** and see how it's related to the disease by seeing how it's related to antibodies.

DOB REPRESENTATION

The DOB for each infant is represented by an angle, called θ , between 0^0 and 360^0 :



CONVERSION FORMULA

- The angle θ representing each DOB is calculated using the following conversion:

$$\theta = (\text{DOB} - \text{January } 1^{\text{st}}) / (365 \text{ or } 366) * 360^0$$

- Each θ is an angle ranging from 0^0 to 360^0
- For example, if an infant was born on October 2nd, 1991, then:

$$\theta = (10/02-01/01) / (365) * 360^0; \text{ or } \mathbf{270.25^0}$$

PARAMETERIZATION: RESULT

- The DOB, represented by θ , becomes a continuous variable with a **circular** distribution without a true zero origin point (that is, **any other date can be used as time origin in the place of January 1st; we would have the same results regardless of choices**).
- The DOB, represented by θ , is characterized by two (2) component: sine (**$\sin\theta$**) & cosine (**$\cos\theta$**).

LOW ANTIBODY: A CAUSE?

- Otitis Media is often referred to as “**Ear Infection**”; bacteria have a definite role here.
- The disease occurs early in life before infants have their own completed **immune system**; their ability to fight infection consist only of what they are inherited from their mother.
- Streptococcus pneumoniae (Spn) causes about 50 percent of recurrent Accute Otitis Media episodes.
- Therefore, the role of cord blood Pneumococcal **IgG antibody** is important.

HYPOTHESIS

- Low maternal concentration of IgG antibody results in low neonatal antibody and early onset of OM, which leads to recurrent and chronic diseases.
- I'll only use type 19F (Ant19F), on log scale - because skewed distribution), one of the seven serotype, for illustration.

In other words:

The dependent variable is:

$Y = \text{Antibody, type 19F} - \text{on log scale}$

Target predictors? **$\text{Sine}(\theta)$ and $\text{Cos}(\theta)$** ,
together they represent the DOB.

MULTIPLE REGRESSION

- We are interested in whether there is a relationship between infants' DOBs and their type 19F IgG antibody concentrations, so we fit a linear multiple regression model.
- $Y = \text{Ant19F}$ is the response variable and $X_1 = \sin(\theta)$ and $X_2 = \cos(\theta)$ are two covariates:

Model:

$$\text{Mean}(\text{Ant19F}) = \beta_0 + \beta_1 * \sin(\theta) + \beta_2 * \cos(\theta)$$

This model is **similar to a polynomial regression**; just like (X and X^2), the two predictors $\text{Sin}(\theta)$ and $\text{Cos}(\theta)$ are representing the **same “predictor source”**— so that coefficients β_1 and β_2 do not follow the usual interpretation. Therefore, **we will investigate their roles together.**

MULTIPLE REGRESSION RESULTS

<u>Factor</u>	<u>Coefficient Estimate</u>	<u>St Error</u>	<u>p-value</u>
Intercept	$b_0=.57$.08	<.001
$\sin(\theta)$	$b_1=.38$.11	<.001
$\cos(\theta)$	$b_2=.01$.11	.095

The p-value for $\sin(\theta)$ is very small, indicating that **DOB, part of which is represented by $\sin(\theta)$** , is a significant factor in predicting infants' antibody 19F concentration.

ANTIBODY: RESULTS

- Since DOB is represented by two variables (sin(θ) and cos(θ)), it is difficult to interpret the two regression coefficients individually.
- By taking the derivative of:

$$\text{Mean}(\text{Ant19F}) = b_0 + b_1 * \sin(\theta) + b_2 * \cos(\theta)$$

relative to θ and set it to zero, we get two angles in (0,360): 88.5⁰ (max.; Ant19F=2.59) and 268.5⁰ (min.; Ant19F=1.21), which correspond to April 1st & September 30th.

An Illustration of The Horoscope

- I divide the year into 4 seasons : (i) **February 16 - May 15** (centered at April 1, date with maximum antibody), (ii) **May 16 - August 15**, (iii) **August 16 - November 15** (centered at September 30, date with minimum antibody), and (iv) **November 16 - February 15**. The following Table shows a rather symmetric distribution of antibody level, type 19F

<u>Season</u>	<u>GM of Antibody</u>	<u>95% Confidence Interval</u>
2/16-5/15	2.264	(1.714,2.992)
5/16-8/15	1.942	(1.432,2.633)
8/16-11/15	1.375	(0.930,2.032)
11/16-2/15	1.735	(1.290,2.333)

RESULTS/IMPLICATION

- The results imply that infants born during the Spring season (as April 1st suggests) tend to have a higher antibody concentration than infants born during the Fall season (as September 30th suggests).
- Therefore infants born in Spring should have a lower risk of disease than infants born in Fall. **We proved Horoscope!**

Why Antibody has anything to do with
DOB? Does horoscope have some
scientific supports?

ANTIBODY: INTERPRETATION

- The finding that infants born in the Fall have low antibody levels compared to those born in the Spring probably due to different levels of maternal exposure preceding the infant's birth.
- Pregnant women have the greatest exposure potential to pneumococcal bacteria during winter, peak antibody levels would follow that exposure, resulting in greater amounts of antibody transferred to infants born in the Spring.
- Result: amount of time a pregnant women staying indoor would be predictive of cord blood antibody.

SOURCES OF EXPOSURE

- Some difference in the findings (Low-antibody kids in **late September**, representing **the lack of exposure by the mothers**). It's not sole determinant.
- The other source is the **exposure by the newborns** (houses are closely sealed in the Winter, starting in **November to March**, likely leading to more severe indoor pollution when newborns suffer).
- Disease occurrence due to both sources, lack of exposure by mother (Fall) and exposure by the child (Winter): total combined effect is peaked in **late October to late December** when disease occurs.

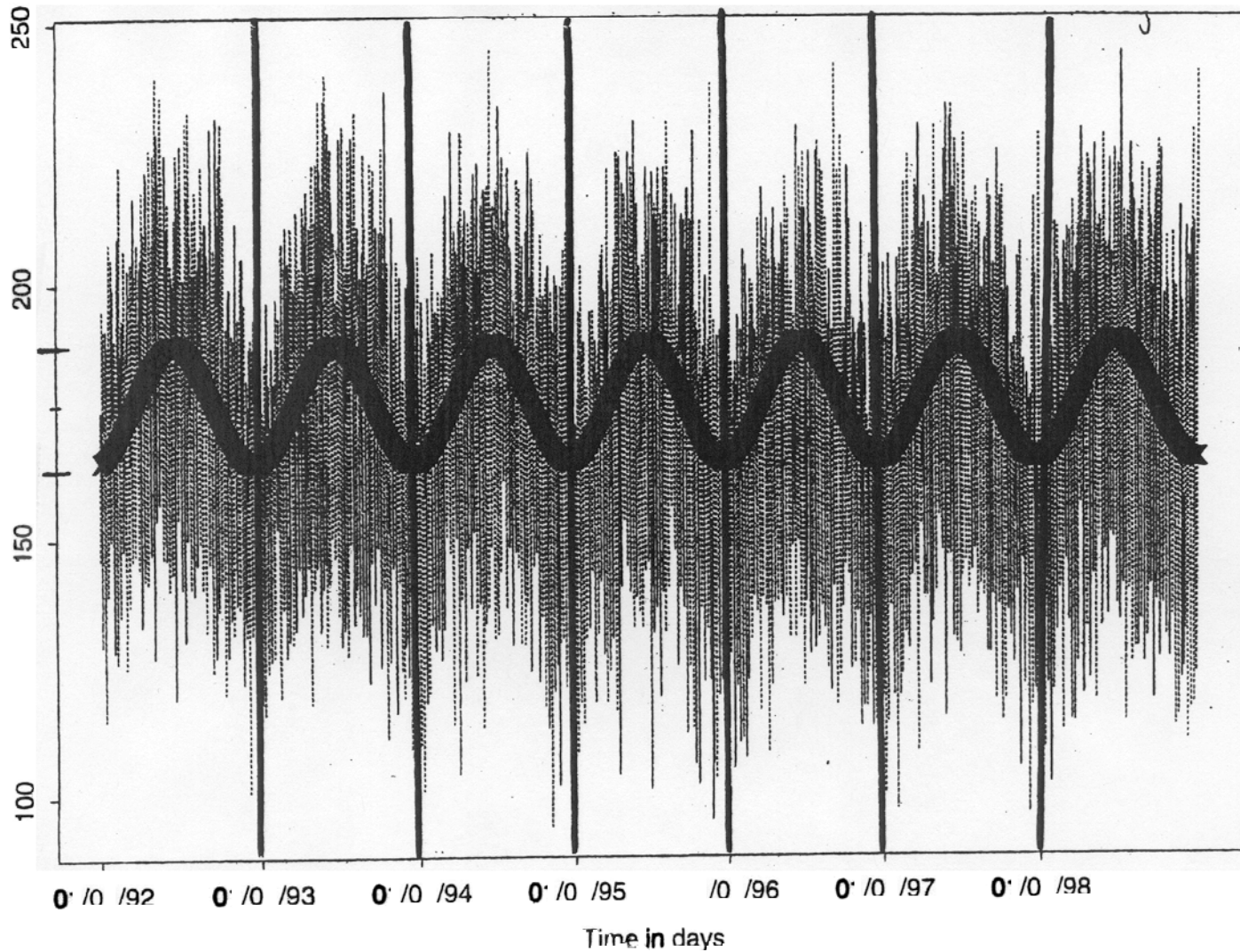
MORE GOOD NEWS:

The other piece of good news is that more kids in Minnesota are born in Spring-Summer than Fall-Winter:

- Maximum: 188 births/day in June-July
 - Minimum: 165 births/day in Dec.-January
- (Family planning ? By parents, and by doctors?)

SEASONAL BIRTHS PATTERN

188
165



There are other smaller cycles:

(1) Hours of the Day: Blood Pressure
or Time to take medication?

(2) Days of the Week: “Weekend
Effects” in scheduling surgeries, say
C-sections?

Suggested Exercise

#5 Use the data set in the following Table,

- Use data from the Standard Preparation & Scatter diagram to verify the linear regression of Y versus $X = \log(\text{dose})$ fits better than Y versus $X = \text{Dose}$;
- Formally test (at $\alpha = .05$) that the lines are parallel;
- Find a point estimate of the $\log(\text{Relative Potency})$;
- Calculate the Standard Error of the estimate in (c).

	Vitamin D3 (Standard)			Cod-liver Oil (Test)			
Dose	5.76	9.6	16	32.4	54	90	150
Response	33.5	36.2	41.6	32	32.6	35.7	44
	37.3	35.6	37.9	33.9	37.7	42.8	43.3
	33	36.7	40.5	30.2	36	38.9	38.4
		37	42			40.3	44.2
		39.5					43.7