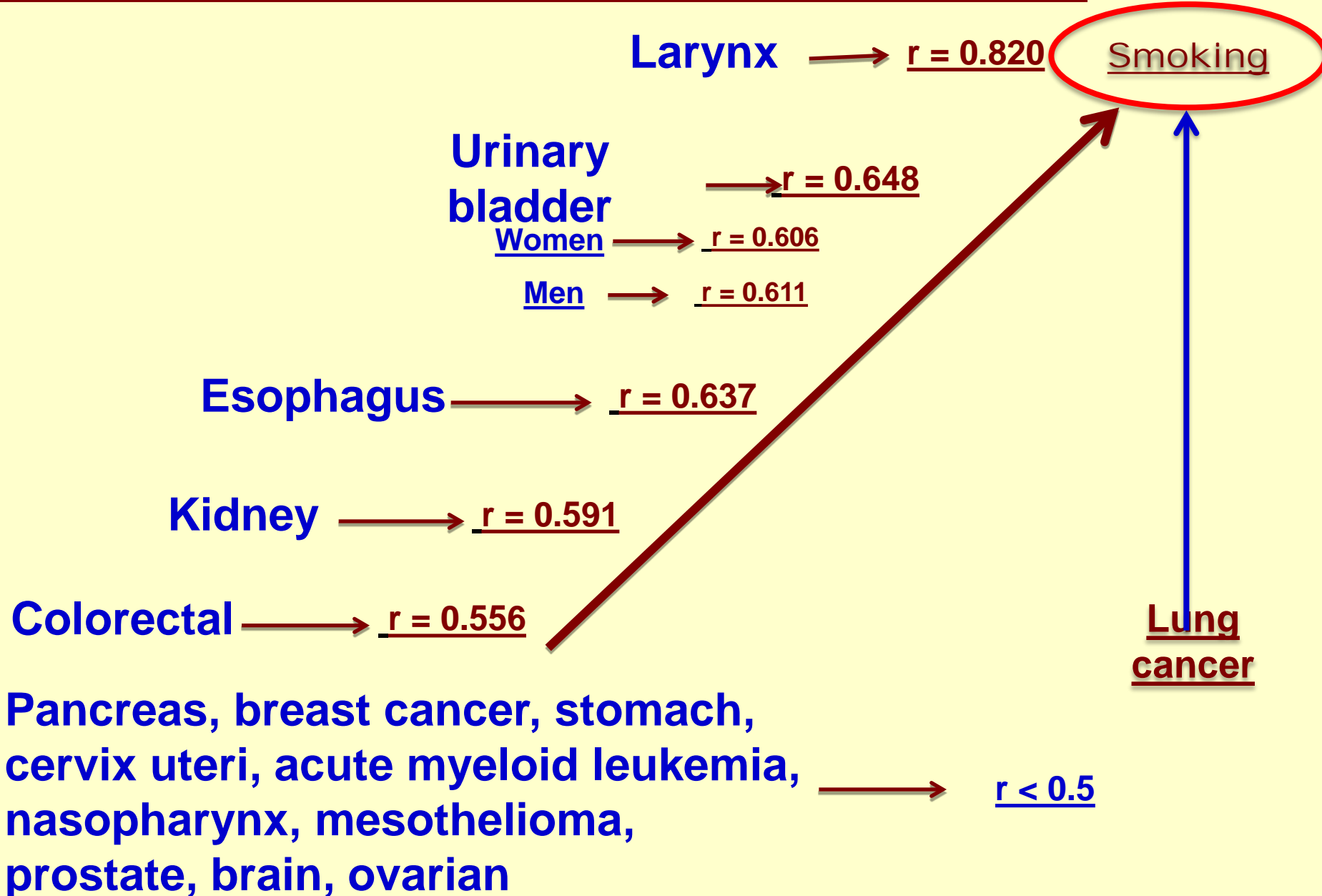# PubH 7405: REGRESSION ANALYSIS

#1. SMOKING & CANCERS

There is strong association between lung cancer and smoking; this has been thoroughly investigated. A study was conducted (Chest, 2010) to explore whether Cigarette Smoking was the cause on cancers other lung cancer. The reasoning was that if lung cancer is associated with other cancers, those cancers are also associated with smoking (guilty by association?).
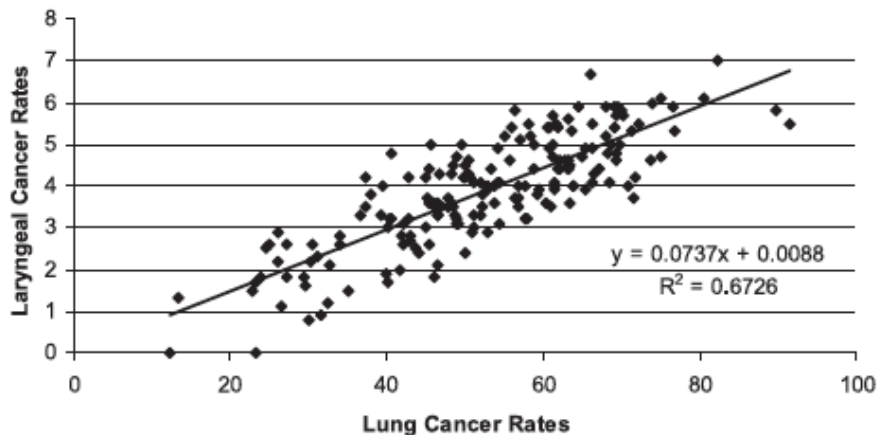
Methods

- **Data source: Surveillance, Epidemiology and End Results (SEER) database; used extracted data and computed incidence rates for 197 counties (n=197, each county is an observation unit)**

- **Method: Conducted Linear regression for incidences of lung cancer (proxy for smoking; independent variable) and incidence of other cancers (dependent variable).**

Some results: Correlation coefficients

Larynx → **r = 0.820**   **Smoking**

Urinary bladder → **r = 0.648**

Women → **r = 0.606**

Men → **r = 0.611**

Esophagus → **r = 0.637**

Kidney → **r = 0.591**

Colorectal → **r = 0.556**

Lung cancer

Pancreas, breast cancer, stomach, cervix uteri, acute myeloid leukemia, nasopharynx, mesothelioma, prostate, brain, ovarian → **r < 0.5**
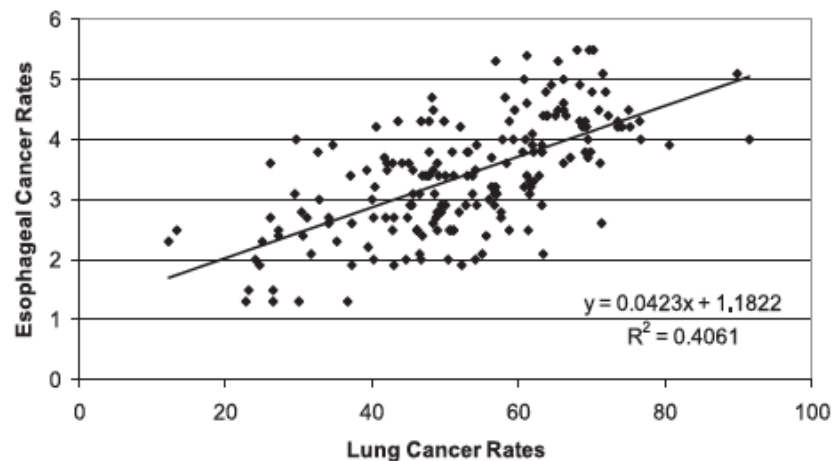
FIGURE 1. Linear regression of laryngeal cancer vs lung cancer. Linear regression formula and R^2 value are noted.

FIGURE 2. Linear regression of esophageal cancer vs lung cancer. Linear regression formula and R^2 value are noted.

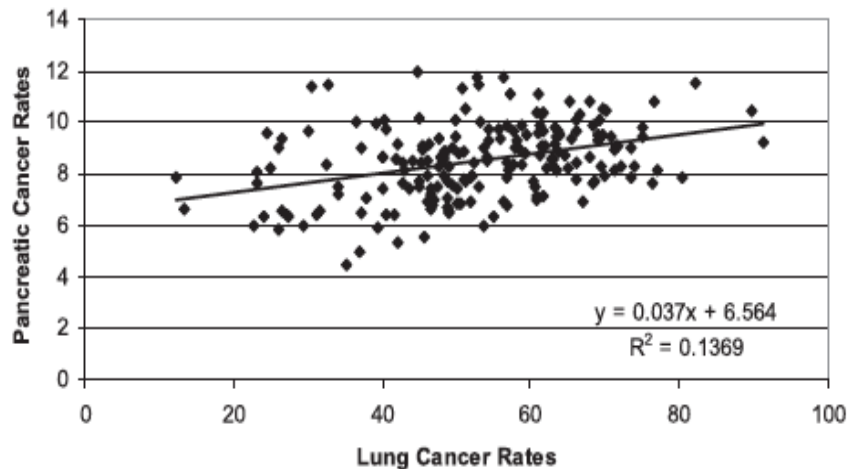**FIGURE 3.** Linear regression of pancreatic cancer vs lung cancer. Linear regression formula and R^2 value are noted.
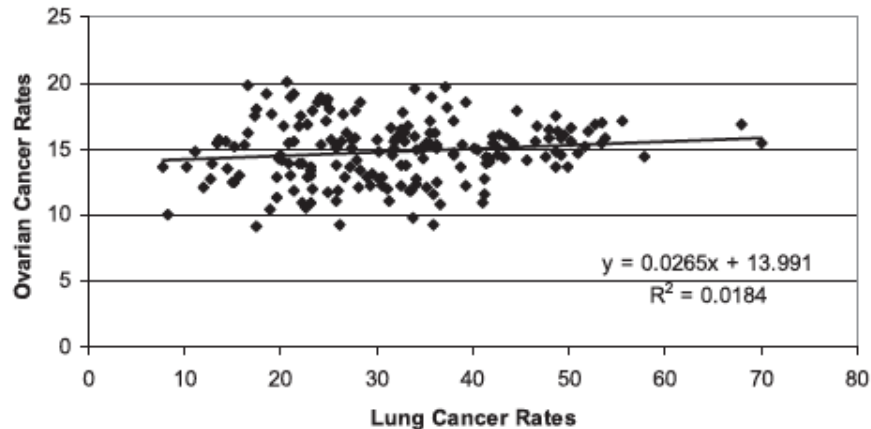

FIGURE 6. Linear regression of ovarian cancer vs lung cancer. Linear regression formula and R^2 value are noted.

ETIOLOGIC FRACTION

$$\% \text{ of Overall incidence due to smoking} = \left(1 - \frac{Incidence \text{ in absence of lung cancer}}{Overall \text{ incidence rates}}\right) * 100$$

Site	Overall Incidence Rates[a]	Incidence in the Absence of Lung Cancer	% of Overall Incidence Due to Smoking
Urinary bladder	20.8	10.6	49.0
Urinary bladder, women	9.4	4.6	51.1
Urinary bladder, men	36.9	18.6	49.6
Larynx	4.6	0.009	99.8
Esophagus	4.5	1.2	73.3
Colorectal	57.5	32.6	43.3
Kidney	10.6	6.0	43.4

[a]Per 100,000 population at risk.

Incidence of cancer in the absence of smoking

$$\% \text{ of Overall incidence due to smoking} = \left(1 - \frac{\text{Incidence in absence of lung cancer}}{\text{Overall incidence rates}}\right) * 100$$

Site	Overall Incidence Rates[a]	Incidence in the Absence of Lung Cancer	% of Overall Incidence Due to Smoking
Urinary bladder	20.8	10.6	49.0
Urinary bladder, women	9.4	4.6	51.1
Urinary bladder, men	36.9	18.6	49.6
Larynx	4.6	0.009	99.8
Esophagus	4.5	1.2	73.3
Colorectal	57.5	32.6	43.3
Kidney	10.6	6.0	43.4

[a] Per 100,000 population at risk.

≈100%

73.3%

#2. LUNG TUMORIGENESIS

- A group of mice were injected with NNK (a toxin from tobacco products) dissolved in saline when mice are 6 weeks old.

- About 16-20 weeks after treated by NNK, **most mice have lung tumors**; there will be an average of 10 surface tumors per lung and an average total tumor volume per lung = 400 mm^3 +/- 100 (SD)

A DOSE-RANGING EXPERIMENT

- Among a group of NNK-treated mice (with tumors after 16 weeks), say n=50, 10 mice are selected and sacrificed to measure tumor volumes – serve as baseline (or data for controls)

- The other 40 mice are randomized into 10 groups of 4 mice each treated by 10 different doses of a cancer agent/drug; the doses are spread over a very wide range from very low to very high

- Aim is to **calculate the dose for 50% reduction of tumor volume (ED50)** , the "**median effective dose**" which characterizes the **agent's potency**.

DATA SUMMARIES

- Let "d" be one of the doses; x = log (d)
- v_0 = average tumor volume of control group
- v_x = average tumor volume of group treated with dose "d"; and
- $p_x = (v_0 - v_x)/v_0$ the **per cent of tumor reduction** to treatment with dose d.

A REGRESSION MODEL

$$\ln \frac{p_x}{1 - p_x} = \beta_0 + \beta_1 x$$

After estimating intercept and slope, β_0 by "b_0" and β_1 "b_1", we can calculate the median effective dose by setting $p_x = .5$:

$$ED_{50} = \exp(-b_0/b_1)$$

Interesting Question:
How do you find the Standard Error of ED_{50}?

Where Does This Model Come From?

$$\ln \frac{p_x}{1 - p_x} = \beta_0 + \beta_1 x$$

MEDIAN EFFECT PRINCIPLE

When a dose **D** of an agent is applied to a pharmacological system, the fractions f_a and f_u of the system affected and unaffected satisfy the so-called "median effect principle" :

$$\frac{f_a}{f_u} = \left\{ \frac{d}{ED_{50}} \right\}^m$$

where ED_{50} is the "median effective dose" and "m" is a Hill-type coefficient; m = 1 for first-degree or Michaelis-Menten system. The median effect principle has been investigated much very thoroughly in pharmacology.

$$\frac{f_a}{f_u} = \left\{ \frac{d}{ED_{50}} \right\}^m$$

$$= \frac{f_a}{1 - f_a}$$

$$\ln \frac{f_a}{1 - f_a} = \ln \frac{p}{1 - p}$$

$$= m\ln(ED_{50}) + m\ln(d)$$

$$= \beta_0 + \beta_1 x$$

#3. PROSTATE CANCER

- The prostate is part of a man's reproductive system. It is a gland surrounding the neck of the bladder & it contributes a secretion to the semen.

- A healthy prostate is about the size of a walnut and is shaped like a donut. The urethra (the tube through which urine flows) passes through the hole in the middle of that "donut".

- If the **prostate grows** too large, it squeezes the urethra causing a variety of **urinary problems**.

- Cancer begins in **cells**, building blocks of tissues
- When normal process goes wrong, **new cells form unnecessarily <u>and</u> old cells do not die** when they should. **Extra mass of cells called a tumor; and malignant tumors are cancer**.
- No one knows the exact causes of prostate cancer … yet, but age is a significant factor. Most men with prostate cancer are over 65; if they live long enough a large proportion of men would eventually have prostate cancer.

PROSTATE CANCER SCREENING

- There are risk factors (age, family history) and symptoms (inability to urinate, frequent urination at night, etc…)

- Common screening is a blood test to measure **Prostate-Specific Antigen (PSA).**

- However, a high level could be caused by benign prostatic hyperplasia (BPH – growth of benign cells); so the test is not very specific.

- **Increasing PSA over time does not necessarily mean prostate cancer**

A PROSTATE CANCER MODEL

- **Serum PSA in patients diagnosed with prostate cancer follows an exponential growth curve.**

- **A retrospective study of banked serum samples (Carter et al., *Cancer Research* 52, 1992) showed that the exponential growth begins 7-9 years before the tumor is detected clinically.**

EXPONENTIAL GROWTH MODEL

$$PSA_t = PSA_0 \exp(\beta_1 t)$$

$$Y_t = \ln PSA_t$$

$$= \beta_0 + \beta_1 t$$

PSA-DT

- **PSA-DT**, the prostate specific antigen **doubling time**, has been used to predict clinical outcomes such as time to progression and **prostate cancer specific mortality**.
- Calculated by different ways; sometimes not calculated correctly, e.g. using 2 data points.
- Simple but not easy; still difficulties.

$$\mathbf{PSA}_t = \mathbf{PSA}_0 \exp(\beta_1 t)$$

$$\mathbf{Y}_t = \mathbf{lnPSA}_t$$

$$= \beta_0 + \beta_1 t$$

$$\mathbf{Y}_2 - \mathbf{Y}_1 = \beta_1 (t_2 - t_1)$$

$$\mathbf{DT} = \frac{\mathbf{ln2}}{\beta_1}$$

Interesting Question:

How do you find the Standard Error of DT?

$$PSA - DT = \frac{\ln 2}{\beta_1}$$

$$Var(PSA - DT) \cong (\ln 2)^2 (\frac{-1}{b_1^2})^2 Var(b_1)$$

$$SE(PSA - DT) = \frac{\ln 2}{b_1^2} SE(b_1)$$

It has nothing to do with "b_0", the estimated intercept.

#4. A VACCINE MODEL

Maintenance of long-term antibody responses is critical for protective immunity against many pathogens. After a person is vaccinated, his/her antibody is usually reached maximum level A_0 at about $t_0 = 2$ weeks which can be considered as "time zero" or time origin (relabeled as t=0); after that antibody level is decreased following an "Exponential Decay Model".

EXPONENTIAL DECAY MODEL

$$A_t = A_0 \exp(-\lambda_1 t)$$

$$Y_t = \ln A_t$$

$$= \lambda_0 - \lambda_1 t$$

The "Exponential Decay Model" for antibodies is similar to the "Exponential Growth Model" for prostate-specific antigens; the only difference is the "<u>negative slope</u>". The counterpart of the "Doubling Time" is the "Half Life"

HALF LIFE

The "half life" M is defined: $A(M) = (1/2)A_0$; people are advised to get re-vaccinated at about t = M. For example, you should get re-vaccinated for "tetanus" after 10 years because its half life is about 10 years.

$$A_t = A_0 \exp(-\lambda_1 t)$$

$$Y_t = \ln A_t$$

$$= \lambda_0 - \lambda_1 t$$

$$Y_2 - Y_1 = \lambda_1(t_2 - t_1)$$

$$M = \frac{\ln 2}{\lambda_1}$$

Same Interesting Question:

How do you find the Standard Error of M?

#5. PHARMACOLOGY BASICS

There are two different types of drugs:
Agonists- they stimulate and activate the receptors
Antagonists - they stop the agonists from stimulating the receptors
Once the receptors are activated, they either trigger a particular response directly on the body, or they trigger the release of hormones and/or other endogenous drugs in the body to stimulate a particular response.
The **action of drugs** on the human body is called **pharmacodynamics** and what the **body does with the drug** is called **pharmacokinetics**

DOSE-RESPONSE RELATIONSHIP

For many agonists, the dose-effect relationships are approximately hyperbolic of the form:

$$y = y_{max} \frac{D}{D + K}$$

Where y is the effect or response, D is the dose, and K is some constant. We can turn this into a linear relationship between the reciprocal effect 1/y against the reciprocal dose:

$$\frac{1}{y} = \frac{1}{y_{max}} + \left(\frac{K}{y_{max}}\right)\frac{1}{D}$$

ENZYME KINETICS

The Michaelis-Menten equation describes the relationship between the velocity v (a response) and the substrate concentration (a predictor) as:

$$v = \frac{v_{\max}[S]}{[S] + K_M}$$

Where K_M is known as the Michaelis constant. Thant relationship can be transformed into linear form by reciprocating both sides:

$$\frac{1}{v} = \left(\frac{K_M}{v_{\max}}\right)\frac{1}{[S]} + \frac{1}{v_{\max}}$$

FIRST-ORDER DRUG DECAY

Consider a drug A which decomposes according to the reaction: A $\rightarrow$ B + C. The original concentration of A is "a". After time "t" the number of moles decomposed is "y"; the amount of A is "a-y", and y moles of B or C have been formed. The differential equation for the chemical reaction:

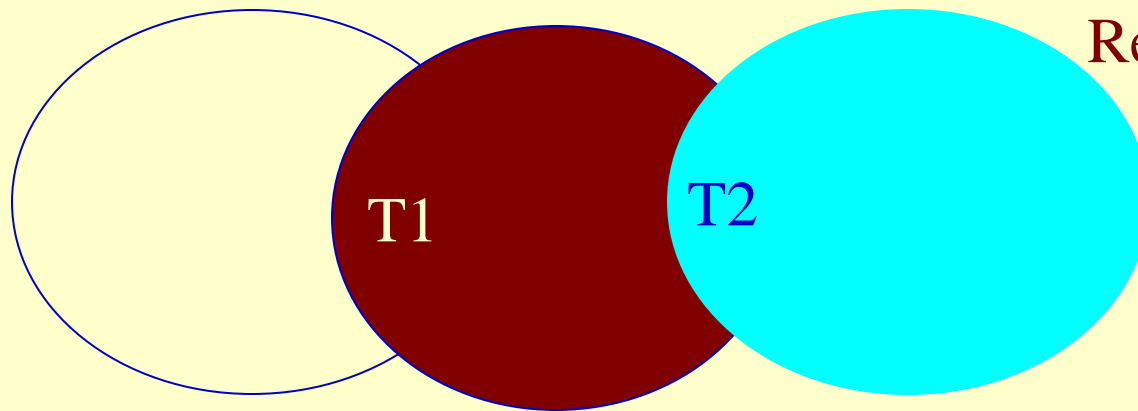$$\frac{dy}{dt} = k(a - y)$$

The solution is a straight line through the origin with slope "k", a special case of linear regression without an intercept.

$$\ln\left(\frac{y}{a - y}\right) = kt$$

Clinical Research

Population Research

T1

T2

Laboratory Research

Translational Research is the component of basic science that interacts with clinical research (T1) or with population research (T2).

People often emphasize more on the first area of translational research, T1; they are research efforts and activities needed to bring discoveries in the laboratories to the bed sides; (T2 starts gaining popularity in last few years).

And it is hard to pinpoint precisely the starting point of "T1"; many believe that translational research starts with "biological assays" – or bioassays, but some could point to In Vitro or In Vivo which are pre-clinical.

#6. BIOASSAYS

- "Biological assays" or bioassays are methods for <u>estimating the potency</u> or strength of an agent or stimulus by utilizing the response or effect or reaction caused by <u>its application to biological material or experimental living subjects</u>.

- Simple examples:
  (1) <u>Six</u> aspirin tablets can be fatal to a child;
  (2) <u>Certain dose</u> of a drug can kill a cat.

UNDERLYING RATIONALE

If the relationship between stimulus level and response exists (by means of an algebraic expression − or a "regression model") then it can be used to study the potency of a dose from the response it produces.

BASIC PROCESS

(1) a "<u>test</u> preparation" of the stimulus - having an unknown "potency" - is "assayed" to find the response.

(2) we find the dose of the "<u>standard</u> preparation" which produces the <u>same</u> response (as that by test preparation). (Then we infer the dose of the test versus the dose of the standard: the ratio of the two doses causing the same response)

There are <u>two </u>types of bioassays
(they are both stochastic, of course):
(1) direct assays and
(2) indirect assays.

DIRECT ASSAYS

- In direct assays, the doses of the standard and test preparations are "directly <u>measured</u>" for (or until) an "event of interest". **<u>Response is fixed (binary), dose is random.</u>**

- When an event of interest occurs, e.g.. the death of the subject, and the <u>variable of interest</u> is the dose required to produce that response/event for each subject. The value is called **" individual effect dose" (IED).**

- For example, we can increase the dose until the heart beat (of an animal) ceases to get IED.

Typical Experiment (for direct assays):

A group of subjects (e.g. animals like mice) are randomly divided into two subgroups and then IED of a standard preparation is measured in each subject of group 1; the IED of the test or unknown preparation is measured in each subject of group 2. The aim is to estimate the "**relative potency**", that is the "**ratio of concentrations**" of the test relative to standard to produce the same biological effect/event. In this example, it is the **ratio of sample means**.

<u>Note:</u> Direct Assays are <u>only</u> applicable when "<u>intra subject</u> dose escalation" is possible. By **technical/ethical reasons**, this is not done in Human Trials; we would rely on Indirect Assays.

Keep in mind that the "concentration" and the "dose" are inversely proportional - when concentration is high, we need a smaller dose to reach the same response. In other words , we define the "relative potency" or the "ratio of concentrations" of the test to standard as the **"ratio of doses" of the standard to test**:

$$\rho = \frac{Dose_S}{Dose_T}$$

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**; The response could be binary or continuous.

- Statistically, indirect assays are more interesting (and, of course, also more difficult).

MEASUREMENT SCALE

Depending on the "measurement scale" for the response (of indirect assays), we have:

(1) <u>Quantal assays</u>, where the response is binary: whether or not an event (like the death of the subject) occurs,

(2) <u>Quantitative assays</u>, where measurements for the response are on a continuous scale.

UNDERLYING RATIONALE

If the relationship between stimulus level and response exists (by means of an algebraic expression) then it can be used to study the potency of a dose from the response it produces.

The common indirect assay is usually one in which the ratio of equipotent doses is estimated from curves relating quantitative responses and doses for the two preparations. The shape of these "curves" 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 $\mathbf{D_S = \rho D_T}$.

- The model: $E[Y_S|\mathbf{X_S}=\log(D_S)] = \alpha + \beta X_S$, for Standard and, for same dose of Test we have $E[Y_T|\mathbf{X_S}=\log(D_S= \rho D_T)= (\alpha + \beta\log\rho) + \beta X_T$

- We have 2 parallel lines with a **common slope** β and **different intercepts**.

We have 2 parallel lines with a common slope and different intercept:

$$\beta_{1S} = \beta_{1T} = \beta$$

$$\beta_{0S} = \alpha$$

$$\beta_{0T} = \alpha + \beta \log \rho$$

$$\log \rho = \frac{\beta_{0T} - \beta_{0S}}{\beta}$$

In general, the point estimate of the relative potency involves 2 intercepts and two slopes:

$$\log r = \frac{b_{0T} - b_{0S}}{\dfrac{w_T b_{1T} + w_S b_{1S}}{w_T + w_S}}$$

Optimal choice for each weight is the inverse of variance of the slope; e.g. $\mathbf{W_T = 1/Var(b_{1S})}$

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 <u>replications</u> at each dose of each preparation.

- It is designed with 8 <u>dishes/plates</u>, 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
$X = \log_{10}(Dose)$	-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

Can "see" that Relative potency is about 1.5

The Indirect Quantitative Assays

$Y_t = 13.25 + 11.46X_t$

test

standard

$Y_s = 11.47 + 10.96X_s$
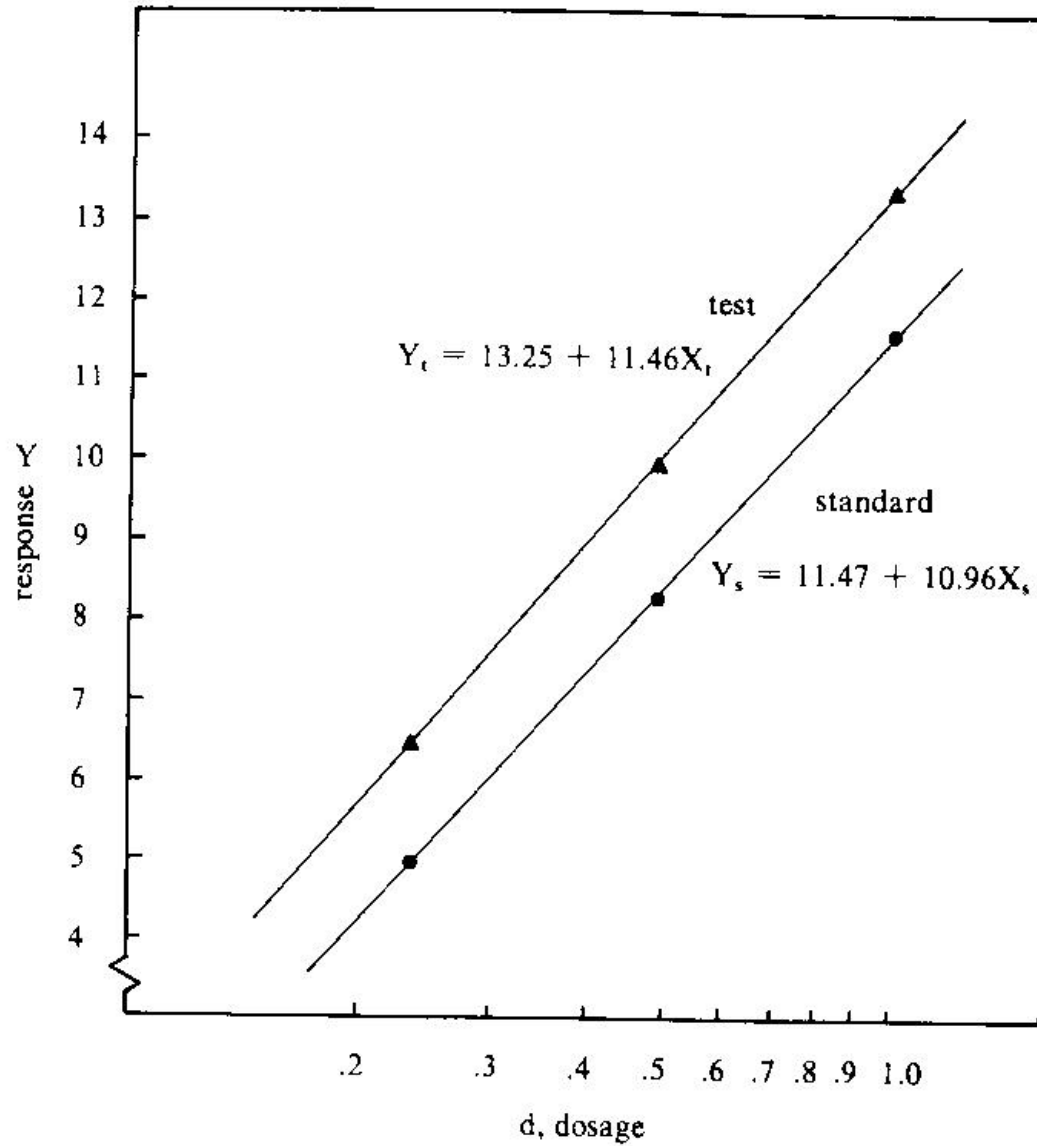
Quantitative parallel line assay responses against log doses.

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 can solve the problem by calculating the <u>weighted average</u> of the two estimated slopes**. Another approach, which turns out more simple, is Multiple Linear Regression.

Due As Homework

#8.1 **Refer the first data set below, a 4-point assay of Corticotropin and find a point estimate of the Relative Potency.**

Dose				
Standard		Test		
0.015	0.045	0.015	0.045	Total
45.07	60.2	49.75	66.35	221.37
44.12	62.93	35.83	45.58	191.46
39.64	48.44	44.94	54.26	187.28
31.48	48.95	34.76	56.39	171.58
160.31	220.52	165.28	225.58	771.69

#8.2 Refer to the following data set:
   a) Compare the two slopes;
   b) if the result in (a) is not statistically significant, calculate the weighted average the slope and its standard error;
   c) Find a point estimate of the Relative Potency

	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

#8.3 Refer to dataset "Infants", let Y= Birthweight and X = Gestational Weeks, and consider two regression lines for two categories of toxemia (toxemia = 1/0 for pregnancy condition with metabolic disorder)

 a) Compare the two slopes;

 b) If the result in (a) not statistically significant, obtain the weighted average and its standard error

(This exercise #8.3 is optional)