

PubH 7405: REGRESSION ANALYSIS



SLR: BIOMEDICAL APPLICATIONS

In applications, sometimes data are classified into groups, and **within each group a separate regression model of Y on X may be postulated.** For example, the regression of “forced expiratory volume” (Y) on age (X) may be considered separately for men in different occupational groups because **different occupations may have different effects on the lung health of workers.** **Differences** between the regression lines, especially the **slopes**, are our primary interest.

AN EXAMPLE

Suppose we study “vital capacity” among men working in the cadmium industry; the main purpose of the study was to see **whether exposure to fumes was associated with a change in respiratory function**. However, we must take into account the effect of “age” because **respiratory performance declines with age**. The men in the sample were divided into three groups:

- (1) those who were **exposed for at least 10 years**,
- (2) those who were **exposed for less than 10 years**, and
- (3) Control group consisted of men **not exposed to fumes**.

We then **consider three regression lines**:

$Y =$ Vital Capacities (liters) versus $X =$ Age

It is well-known that respiratory test performance declines with age. But the question is **whether being exposed to fume in the cadmium industry would accelerate the declining process.** That is to focus on the **difference of slopes.**

We could consider to merge groups 1 and 2 then compare to group 3; however, the result would be **masked** by a phenomenon called “**healthy worker effects**” (healthier people are more likely to choose more dangerous occupations).

The main focus could be placed on the **comparison of group 1 versus group 2** – by showing an **attenuation of health worker effects**: the **decline is steeper** in group 1 (**longer exposure**) than in group 2 (**shorter exposure**).

When possible differences between the regression lines, for example the slopes, are of interest, there are two possibilities:

(1) If the slopes clearly differ, from one group to another, then we have no choice but to draw separate group-specific inferences.

(2) If the slopes do not differ, the lines are parallel with a common slope; that common slope can and should be estimated using combined data from all groups.

GROUP-SPECIFIC RESULTS

Suppose there are k groups with n_i pairs of observations (i.e. “correlation data”) in **the i th group**; a regression line is fitted with the following results for the slope:

Estimated slope = b_{1i}

$$s^2(b_{1i}) = \frac{MSE_i}{\sum_i (x - \bar{x})^2}$$

GENERAL METHODOLOGY

Let w_i be the inverse of the variance of the i th slope; under the null hypothesis that the true slopes of the k groups are all equal, the following statistic G follows approximately a **Chi-square distribution with $(k-1)$ degrees of freedom**:

Weighted average \longrightarrow

$$w_i = \frac{1}{s^2(b_{1i})}$$
$$\tilde{b}_1 = \frac{\sum w_i b_{1i}}{\sum w_i}$$

Similar to ANOVA \longrightarrow

$$G = \sum w_i (b_{1i} - \tilde{b}_1)^2$$

The use of this G statistic is similar to the F statistic in one-way ANOVA:

$$F = \frac{MSB}{MSW}$$

$$SSB = \sum n_i (\bar{x}_i - \bar{x})^2$$

$$= \sum \frac{\sigma^2}{Var(\bar{x}_i)} (\bar{x}_i - \bar{x})^2$$

= Weighted Average of Deviations

If the statistic G is not statistically significant, the null hypothesis that the true slopes of the k groups are all equal is “tentatively accepted”, the common slope is best estimated by the weighted average (of the k individual slopes). **The sampling distribution of this weighted average is approximately normal.**

$$w_i = \frac{1}{s^2(b_{1i})}$$

$$\tilde{b}_1 = \frac{\sum w_i b_{1i}}{\sum w_i}$$

$$\sigma^2(\tilde{b}_1) = \frac{1}{\sum w_i}$$

We could use the same method to compare and combine **intercepts** and **coefficients of correlation**; for the later one we take advantage of the Fisher's transformation:

$$z = \frac{1}{2} \ln \left\{ \frac{1+r}{1-r} \right\}$$

$$\text{Var}(z) = \frac{1}{n-3}$$

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 = $400 \text{ mm}^3 \pm 100$ (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 $p_x = .5$:

$$\mathbf{ED}_{50} = \mathbf{exp}(-\mathbf{b}_0/\mathbf{b}_1)$$

Interesting Question:

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

& Where Does This Model Come From?

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.

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

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

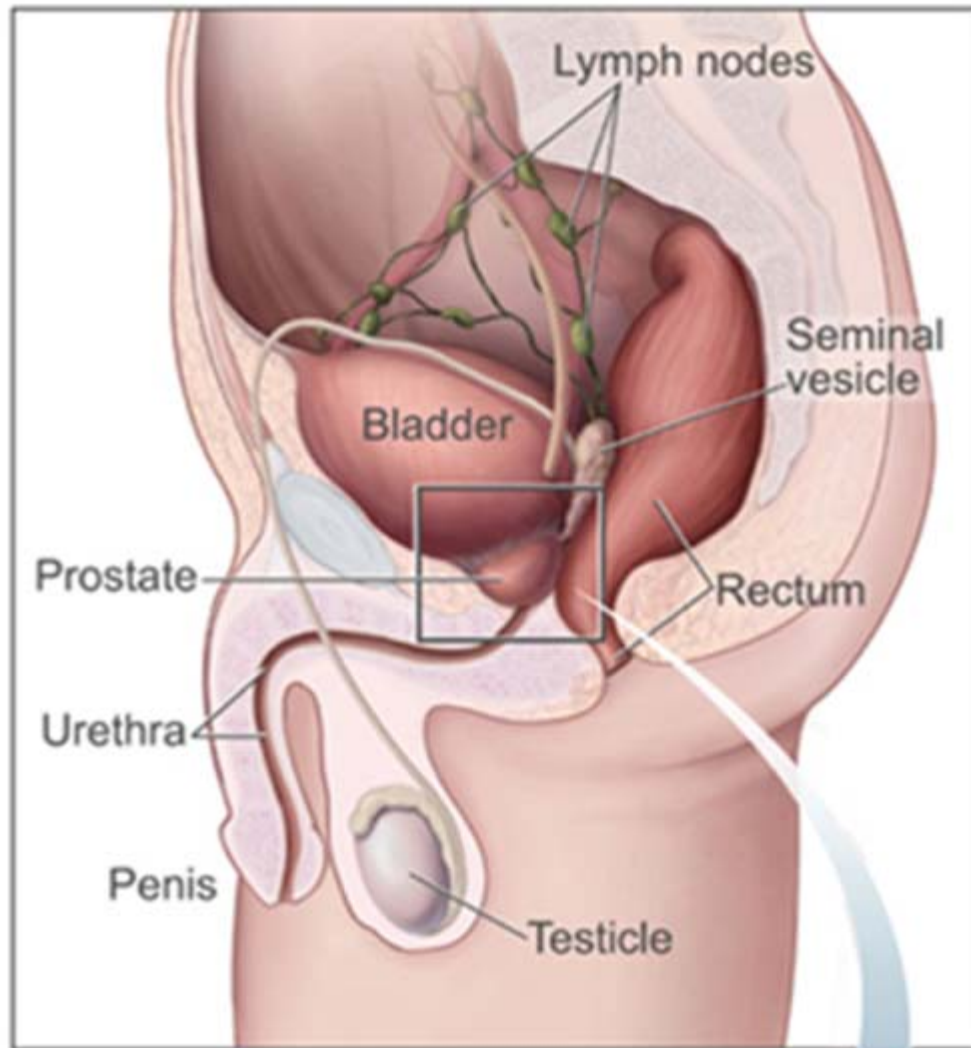
$$\ln \frac{f_a}{1 - f_a} = \ln \frac{p}{1 - p}$$

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

$$= \beta_0 + \beta_1 x$$

PROSTATE

- 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**.



This shows the prostate and nearby organs.

Prostate

PROSTATE CANCER

- Cancer begins in **cells**, building blocks of tissues
- When normal process goes wrong, **new cells form unnecessarily and 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

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

$$Y_t = \ln \text{PSA}_t$$

$$= \beta_0 + \beta_1 t$$

The slope β is a parameter representing disease severity because when the slope is larger, the level of PSA increases faster.

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.

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

$$Y_t = \ln \text{PSA}_t$$

$$= \beta_0 + \beta_1 t$$

$$Y_2 - Y_1 = \beta_1 (t_2 - t_1)$$

$$\text{DT} = \frac{\ln 2}{\beta_1}$$

Interesting Question:

How do you find the Standard Error of DT?

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

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

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

**It has nothing to do with "b₀",
the estimated intercept.**

POSSIBLE PROBLEMS

- Exponential growth model may apply to about two thirds of all cancer patients – why?
- **Starting point is hard to determine** because exponential growth may start years before tumor detection. However, it may need verification for “biochemical progression” because – without it – non-exponential phase may be captured in the data and slope would be under-estimated (and PSA-DT be over-estimated – severity under-estimated)

If $t_0 < t_1$, we still have the same exponential model:

$$PSA(t) = PSA(t_1) \exp[\beta(t - t_1)]; \quad t > t_1$$

(We can use any time t_1 in the exponential growth stage as “origin” instead of the unknown time of disease inception t_0).

$$PSA(t) = PSA(t_0) \exp[\beta(t - t_0)]$$

$$PSA(t_1) = PSA(t_0) \exp[\beta(t_1 - t_0)]$$

$$\begin{aligned} \frac{PSA(t)}{PSA(t_1)} &= \frac{\exp[\beta(t - t_0)]}{\exp[\beta(t_1 - t_0)]} \\ &= \exp[\beta(t - t_0) - \beta(t_1 - t_0)] \\ &= \exp[\beta(t - t_1)] \end{aligned}$$

$$PSA(t) = PSA(t_1) \exp[\beta(t - t_1)]$$

A possible “hidden” trap is the possible presence of some “nadir value” (non-zero floor value); without subtracting it, the slope would be under estimated too. It is possible that the theory of “exponential” only applies to the PSA level above the nadir value. That might be why, when used without subtracting the nadir value, the exponential growth model only fits about 70% of patients.

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 (reabeled 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 “negative slope”. 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.

$$\mathbf{A}_t = \mathbf{A}_0 \exp(-\lambda_1 t)$$

$$\begin{aligned} \mathbf{Y}_t &= \ln \mathbf{A}_t \\ &= \lambda_0 - \lambda_1 t \end{aligned}$$

$$\mathbf{Y}_2 - \mathbf{Y}_1 = \lambda_1 (t_2 - t_1)$$

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

Same Interesting Question:

How do you find the Standard Error of M?

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. That 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:

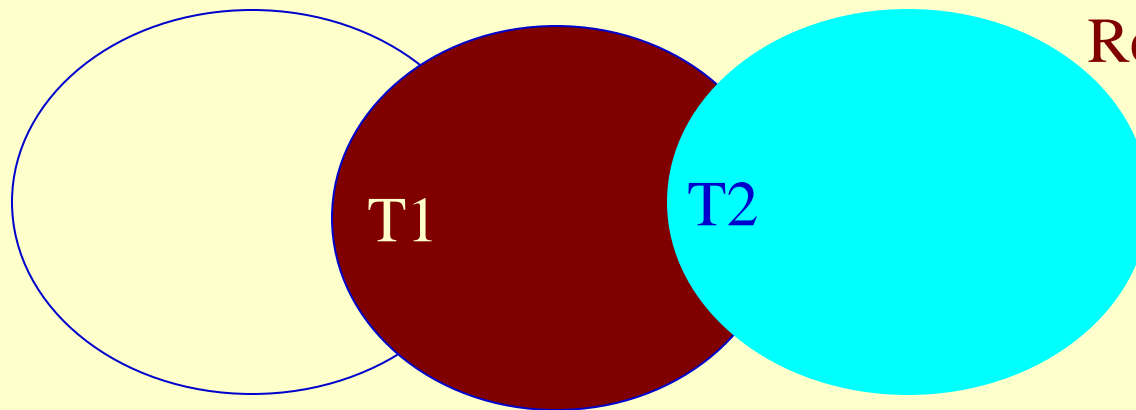
$$\frac{dy}{dt} = k(a - y)$$

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



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.

DEFINITION

- “**Biological assays**” or **bioassays** are methods for estimating the potency or strength of an agent or stimulus by utilizing the response or effect or reaction caused by its application to biological material or experimental living subjects.
- Simple examples:
 - (1) Six **aspirin** tablets can be fatal to a child;
 - (2) Certain dose 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 “test preparation” of the stimulus - having an **unknown** “potency” - is “assayed” to find the response.
- (2) we find the **dose** of the “standard preparation” which produces the same 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 two 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 measured” for (or until) an “event of interest”. **Response is fixed (binary), dose is random.**
- When an 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. 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**.

Note: Direct Assays are only applicable when “intra subject 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) Quantal assays, where the response is binary: whether or not an event (like the death of the subject) occurs,
- (2) Quantitative assays, where measurements for the response are on a continuous scale.

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 $\mathbf{D}_S = \rho\mathbf{D}_T$.
- The model: $E[Y_S | \mathbf{X}_S = \log(\mathbf{D}_S)] = \alpha + \beta\mathbf{X}_S$, for Standard and, for same dose of Test we have $E[Y_T | \mathbf{X}_S = \log(\mathbf{D}_S = \rho\mathbf{D}_T)] = (\alpha + \beta\log\rho) + \beta\mathbf{X}_T$
- 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:

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

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

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 weighted average of the two estimated slopes. Another approach, which turns out more simple, is Multiple Linear Regression.

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**.

Due As Homework

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

7.1 Refer the first data set on the left, a 4-point assay of Corticotropin and find a point estimate of the Relative Potency.

7.2 Refer to the second data set, next page:

- Compare the two slopes;
- if the result in (a) is not statistically significant, calculate the weighted average the slope and its standard error;
- Find a point estimate of the Relative Potency

Only #7.2.3 are required

	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

7.3 Refer to dataset “Cigarettes”, let $Y = \log(\text{NNAL})$ and $X = \text{CPD}$, and consider two regression lines, a line for females (Gender = 1) and a line for males (Gender = 2)

a) Compare the two slopes;

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

7.4 Answer the 2 questions of Exercise 7.3 using dataset “Infants” with $X = \text{Gestational Weeks}$ and $Y = \text{Birth Weight}$; 2 regression lines for two categories of **toxemia** (toxemia = 1 for pregnancy condition with metabolic disorder