PubH 7405:
REGRESSION ANALYSIS

APPLICATIONS A:
COMPARING & COMBINING STATISTICS

We can use what we learned to tackle these 2 problems:

(1) We can compare two or several slopes and can combine to form a common value if these two or several statistics are not different.

(2) We can compare two or several coefficients of correlation and can combine to form a common value if these two or several statistics are not different.

(After learning Multiple Regression, we would re-visit this issue with alternative solution which, in many cases, might be a bit more elegant.)

Example #1:

OCCUPATIONAL HEALTH

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.

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**,
(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.

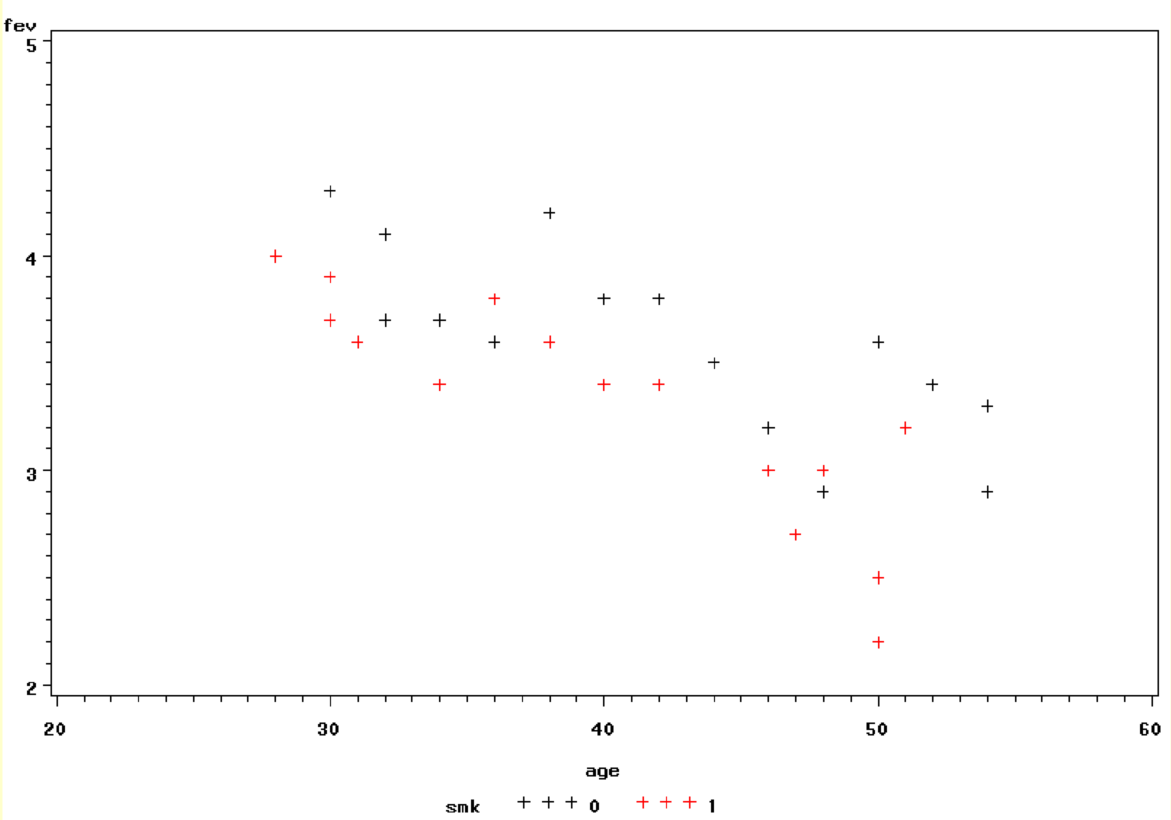
In practice, the fitted regression lines would rarely have precisely the same slope or position – as seen from the scatter diagram. The question is to what extent the differences can be attributed to random variation. There are simple ways to “compare” and, if applicable, to “combine” data forming the common slope.

Example #2:

SMOKING & LUNG HEALTH

Response Variable: Forced Expired Volume (FEV), a measure of lung health.

Main Predictor: Age



(1) Generally, FEV is linearly related to Age,

(2) Non-smokers have better lung health as compared to smokers; the Regression Line for smokers looks "steeper". To see if smoking could modify the natural effect of Age on Lung Health; we might want to compare the slopes. If they are not different, the common parameter would be more precisely estimated using all data.

Example #3:

WHITE CELLS & LEUKEMIA

Leukemia is a cancer characterized by an over-proliferation of white blood cells; the higher the white blood count (WBC), the more severe the disease; WBC is an important predictor of Survival Time (the Response). Another important factor is the presence (“AG positive”) or absence (“AG negative”) of Auer rods and/or significant granulature of the leukemic cells in the bone marrow at diagnosis.

AG-Positive, n = 17		AG-Negative, n = 16	
White Blood count	Survival Time	White Blood Count	Survival Time
(WBC)	(weeks)	(WBC)	(weeks)
2,300	65	4,400	56
750	156	3,000	65
4,300	100	4,000	17
2,600	134	1,500	7
6,000	16	9,000	16
10,500	108	5,300	22
10,000	121	10,000	3
17,000	4	19,000	4
5,400	39	27,000	2
7,000	143	28,000	3
9,400	56	31,000	8
32,000	26	26,000	4
35,000	22	21,000	3
100,000	1	79,000	30
100,000	1	100,000	4
52,000	5	100,000	43
100,000	65		

Look at the effects of “AG factor”, that morphologic characteristic of white cells! We might want to compare the two coefficients of correlation (between WBC and survival time); there is a strong effect modification here.

AG-Positive, n = 17		AG-Negative, n = 16	
White Blood count (WBC)	Survival Time (weeks)	White Blood Count (WBC)	Survival Time (weeks)
2,300	65	4,400	56
750	156	3,000	65
4,300	100	4,000	17
2,600	134	1,500	7
6,000	16	9,000	16
10,500	108	5,300	22
10,000	121	10,000	3
17,000	4	19,000	4
5,400	39	27,000	2
7,000	143	28,000	3
9,400	56	31,000	8
32,000	26	26,000	4
35,000	22	21,000	3
100,000	1	79,000	30
100,000	1	100,000	4
52,000	5	100,000	43
100,000	65		

It can be easily seen that, among AG-positive patients, WBC and Survival Time are negatively correlated – as noted that “the higher the white blood count (WBC), the more severe the disease”. But that is not necessarily true for AG-negative patients: AG modifies the effect of WBC.

Example #4:

KIDNEY ACTIVITIES

One way to learn about the function of the kidney is to study the rates at which it produces and consumes different substances. One important aspect is the rate at which oxygen is consumed since this is considered to be a measure of how hard the kidney is working. Another aspect of interest is the rate at which kidney reabsorbs ionic sodium from the urine. This later activity, known as sodium pumping, requires energy. Thus there should be a direct relationship between the intensity of sodium pumping and the rate of oxygen consumption.

An interesting and important Regression Model:

In the regression of Oxygen Consumption Y (Response) against Sodium Re-absorption X (Predictor Explanatory Variable), the Intercept was interpreted as the base rate of oxygen consumption (which is less important) and the reciprocal of the Slope was used as the estimate of the “pumping efficiency”, an important measure of renal health. Investigators could use pumping efficiency as an outcome to compare different experiment conditions or to evaluate (new) medication/supplement.

Tasks:

Studies cloud be carried in human or animals (dogs or rats); in animals, it could be live or kidneys could be removed from animals and maintained with fluids. Experiments could be done or observations could be made to obtain several data points from each kidney. The we could compare the slopes across subjects in the same experiment condition and combined to form a common pumping efficiency. How do we do that? How do we compare several slopes?

THE DEMAND CURVE

A fundamental concept of consumer demand, in Behavioral Economics, is the Demand Curve relating the consumption of a commodity (Q, the dependent variable) to its (unit) price (P, the independent variable). According to the theory, the consumption of most goods will decrease with increases in price (Watson and Holman, 1977). At the backbone of the Demand Curve is the concept of Elasticity.

ELASTICITY

At the discrete level, a section of the demand curve is characterized by a parameter called Elasticity (E) which is defined as the ratio of two rates or proportions:

$$E = \frac{\dfrac{(Q_2 - Q_1)}{(1/2)(Q_1 + Q_2)}}{\dfrac{(P_2 - P_1)}{(1/2)(P_1 + P_2)}}$$

“Elasticity” could be used to compare liability between products; that with the same price increase, consumption of one product would reduce faster than that of the other. For example, the difference could represent different levels of dependency or addiction.

ELASTICITY on Continuous Scale

For a point on the demand curve, i.e. continuous scale, the elasticity E becomes:

$$E = \frac{\dfrac{(Q_2 - Q_1)}{(1/2)(Q_1 + Q_2)}}{\dfrac{(P_2 - P_1)}{(1/2)(P_1 + P_2)}} \longrightarrow E = \left[\frac{P}{Q}\right]\left[\frac{dQ}{dP}\right] = \frac{d[\ln Q]}{d[\ln P]}$$

which represents the slope on the demand curve when both price (P) & consumption (Q) are expressed on the log scale (we do not have to graph with both on log scale).

DEMAND CURVE
FOR TOBACCO RESEARCH

The demand curve established for food consumption has been adopted for use in tobacco research in areas of product liability and relative reinforcing efficacy (RRE), a concept in psychopharmacological research (Bickel and Madden 1999).

There are studies both in humans (surveys of smokers) & animals (experiments with rats)

ANIMAL EXPERIMENTS

- Human research suggests that there are sex differences in the addiction-related behavioral effects of nicotine; a study was conducted to examine this issue in rats.
- Male and female rats were trained to self-administer nicotine (0.06 mg/kg) under a FR 3 schedule during daily 23-hour sessions.
- Rats were then exposed to saline extinction and reacquisition of NSA, followed by weekly reductions in the unit dose (0.03 to 0.00025 mg/kg) until extinction levels of responding were achieved.
- Fifteen rats (8 males, 7 females) were tested at 8 doses: 0.03, 0.02, 0.01, 0.007, 0.004, 0.002, 0.001, 0.0005, mg/kg/infusion.

SURVEYS OF SMOKERS

- **Data are collected by the cigarette purchase task (CPT) survey, also called TPT, in which participants were asked to respond to the following set of questions**
- **How many cigarettes would you smoke if they were______ each?: 0¢ (free), 1¢, 5¢, 13¢, 25¢, 50¢, $1, $2, $3, $4, $5, $6, $11, $35, $70, $140, $280, $560, $1,120.**
- **This set of questions are asked during an online survey in the preceding order until the respondents gives “0” as an answer, then no more further questions will be asked.**

STANDARDIZATION

A Standardized Demand Curve could be formed as follows; let:

$$t = \ln(P/P_B)$$

$$S(t) = \frac{Q}{Q_B}$$

"Survival Fraction" $S(t) = Q/Q_B$ going down from 1.0 as "time" t increases. In this setup, individual curves have the same shape as the "global curve".

HOW TO EXPRESS ELASTICITY?

$$S(t) = \frac{Q}{Q_0} \;\&\; t = \ln(Q/Q_0) = \ln Q - \ln Q_0$$

$$\ln[S(t)] = \ln(Q) - \ln(Q_0)$$

$$h(t) = -\frac{d}{dt}\ln[S(t)] = -\frac{d(\ln Q)}{d(\ln P)}$$

$$= -\text{Elasticity}$$

If we view the Standardized Demand Curve as a survival curve, Elasticity Function is simply the negative of the Hazard Function.

A Possibility: WEIBULL MODEL

St Demand Curve $S(t) = \exp[-(\alpha t)^{\beta}]$

Elasticity $E(t) = -\alpha\beta(\alpha t)^{\beta-1}$

Could it be more simple? Yes, if β=1, the standardized demand curve has linear elasticity.

DATA ANALYSIS STRATEGIES

- There could be two choices:

(1) Starting with individual curves, then combining results to form the population or global curve.

(2) Going right to population curve, and treat individual data as repeated observations.

- The first strategy is more simple – a straight application of Simple Linear Regression - and it would show individual differences.

DATA TRANSFORMATION

$$t = \ln(P/P_0)$$

$$S(t) = \frac{Q}{Q_B}$$

$$S(t) = \exp[-(\alpha t)^{\beta}]$$

$$E(t) = -\alpha\beta(\alpha t)^{\beta-1}$$

→

$$\ln[-\ln S(t)] = \beta\ln\alpha + \beta\ln t$$

$$\ln[-\ln\frac{Q}{Q_B}] = \beta\ln\alpha + \beta\ln[\ln(P/P_B)]$$

We have a simple linear regression after two double log transformations; goodness-of-fit could be judged visually (R^2 is a good measure).

We combine individual results by calculating weighted averages of slopes and intercepts using inverse of variance as the weight; and use these weighted averages to form Standardized Demand & Elasticity functions.

METHOD

FOR COMPARING & COMBING STATISTICS

Review: ANALYSIS OF VARIANCE

- SST measures the “total variation” in the combined sample with (n-1) degrees of freedom, $n=\Sigma n_i$ is the total size. It is decomposed into: SST=SSW+SSB
- SSW measures the variation within samples with $\Sigma(n_i-1)=(n-k)$ degrees of freedom, and
- SSB measures the variation between sample means with (k-1) degrees of freedom; k=# of groups

SSB measures the variation, or difference, between sample means:

$$SSB = \sum_{i,j} (\bar{x}_i - \bar{x})^2 = \sum_i n_i (\bar{x}_i - \bar{x})^2$$

(which is a concept similar to the “variance”: variation among sample means); decision is based on the F-statistic: F = MSB/MSW.

Grand mean is a “weighted Average”; the weight is the inverse of the variance:

$$\bar{x} = \frac{\sum_{i,j} x_{ij}}{N}$$

$$= \frac{\sum_i n_i \bar{x}_i}{\sum_i n_i}$$

$$= \frac{\sum_i \frac{n_i}{\sigma^2} \bar{x}_i}{\sum_i \frac{n_i}{\sigma^2}}$$

Between Sum of Squares is a weighted sum of deviations (from weighted mean); the weight is the inverse of the variance: :

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

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

$$= \text{Weighted Sum of Deviations}$$

Regression: 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:**

$$\text{Estimated slope } = b_{1i}$$

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

GENERAL METHODOLOGY

Let $\mathbf{w_i}$ **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**:

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

Weighted average →

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

Similar to ANOVA →

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

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**.

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

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

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

Recall:

$$\mathrm{b}_1 = \frac{\sum (x-\bar{x})y}{\sum (x-\bar{x})^2}$$

$$Var(b_1) = \frac{\sum (x-\bar{x})^2 Var(y)}{\{\sum (x-\bar{x})^2\}^2}$$

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

Therefore, we can use as the weight, the Sum of Squares of X:

$$\mathbf{SSX = \sum (x_i - \bar{x})^2}$$

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:

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

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

The result becomes much more simple when we only need to compare two slopes (or any two statistics).

$$\tilde{b} = \frac{w_1 b_1 + w_2 b_2}{w_1 + w_2}$$

$$b_1 - \tilde{b} = b_1 - \frac{w_1 b_1 + w_2 b_2}{w_1 + w_2}$$

$$= \frac{w_2 (b_1 - b_2)}{w_1 + w_2}$$

$$b_2 - \tilde{b} = \frac{w_1 (b_1 - b_2)}{w_1 + w_2}$$

$$\mathbf{b}_1 - \tilde{\mathbf{b}} = \frac{\mathbf{w}_2(\mathbf{b}_1 - \mathbf{b}_2)}{\mathbf{w}_1 + \mathbf{w}_2}$$

$$\mathbf{b}_2 - \tilde{\mathbf{b}} = \frac{\mathbf{w}_1(\mathbf{b}_1 - \mathbf{b}_2)}{\mathbf{w}_1 + \mathbf{w}_2}$$

$$\begin{aligned}
\mathbf{G} &= \mathbf{w}_1(\mathbf{b}_1 - \tilde{\mathbf{b}})^2 + \mathbf{w}_2(\mathbf{b}_2 - \tilde{\mathbf{b}})^2 \\
&= [\frac{\mathbf{w}_1\mathbf{w}_2^2 + \mathbf{w}_2\mathbf{w}_1^2}{(\mathbf{w}_1 + \mathbf{w}_2)^2}]^2(\mathbf{b}_1 - \mathbf{b}_2)^2 \\
&= \frac{\mathbf{w}_1\mathbf{w}_2}{(\mathbf{w}_1 + \mathbf{w}_2)}(\mathbf{b}_1 - \mathbf{b}_2)^2
\end{aligned}$$

$$G = \frac{w_1 w_2}{(w_1 + w_2)}(b_1 - b_2)^2$$

$$= \frac{(b_1 - b_2)^2}{\frac{1}{w_1} + \frac{1}{w_2}}$$

$$= \frac{(b_1 - b_2)^2}{\mathrm{Var}(b_1) + \mathrm{Var}(b_2)}$$

This is equivalent to referring the following statistic to percentiles of the Standard Normal distribution – a very common practice:

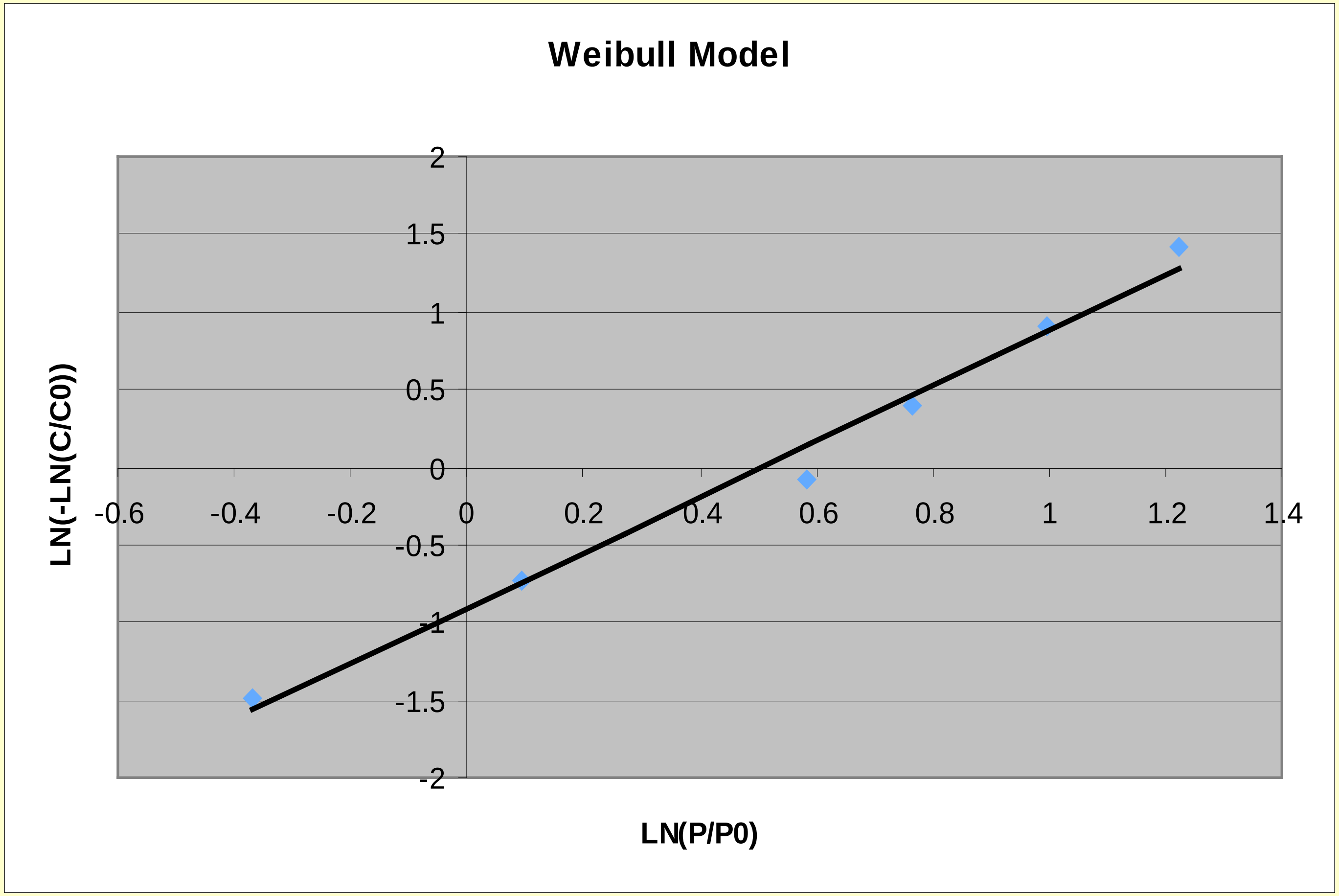
$$\mathbf{z} = \frac{\mathbf{b_1} - \mathbf{b_2}}{\sqrt{\mathbf{Var(b_1)} + \mathbf{Var(b_2)}}}$$

(Variance of difference is equal sum of variances)

NUMERICAL EXAMPLE: RATS DATA

Price	Males							
50	2.0220	1.4820	1.4400	2.2200	1.3800	1.6800	2.0820	1.6560
100	1.6110	1.0800	0.9810	1.4610	1.0110	1.3200	1.7490	0.7710
150	1.2460	0.8200	0.8140	1.0400	0.9000	1.0460	1.3940	0.7200
300	0.8000	0.6130	0.4600	0.6430	0.5870	0.6000	0.6170	0.4600
429	0.4571	0.1589	0.1449	0.4179	0.4809	0.4760	0.1281	0.2751
750	0.1692	0.0588	0.0520	0.2200	0.2200	0.2492	0.0172	0.0960
1500	0.0320	0.0106	0.0140	0.0346	0.0880	0.0834		0.0180
3000				0.0117	0.0290			
6000					0.0057			
12000					0.0022			

Price	Females						
50	3.0000	2.2200	1.6200	3.3600	1.9980	1.8420	2.3220
100	1.4400	0.9690	0.8310	1.8990	1.1700	1.2210	1.3500
150	1.0260	0.9000	0.7340	1.1260	1.0000	1.0940	0.9260
300	0.6830	0.2630	0.2870	0.6830	0.5500	0.7200	0.6700
429	0.1680	0.1470	0.0959	0.5670	0.4571	0.5159	0.4501
750	0.0652	0.0320	0.0428	0.3348	0.2188	0.2732	0.2320
1500			0.0094	0.0586	0.0606	0.1314	0.0400
3000					0.0193	0.0227	0.0157
6000					0.0108	0.0089	
12000							



Weibull Model fits well

RESULTS

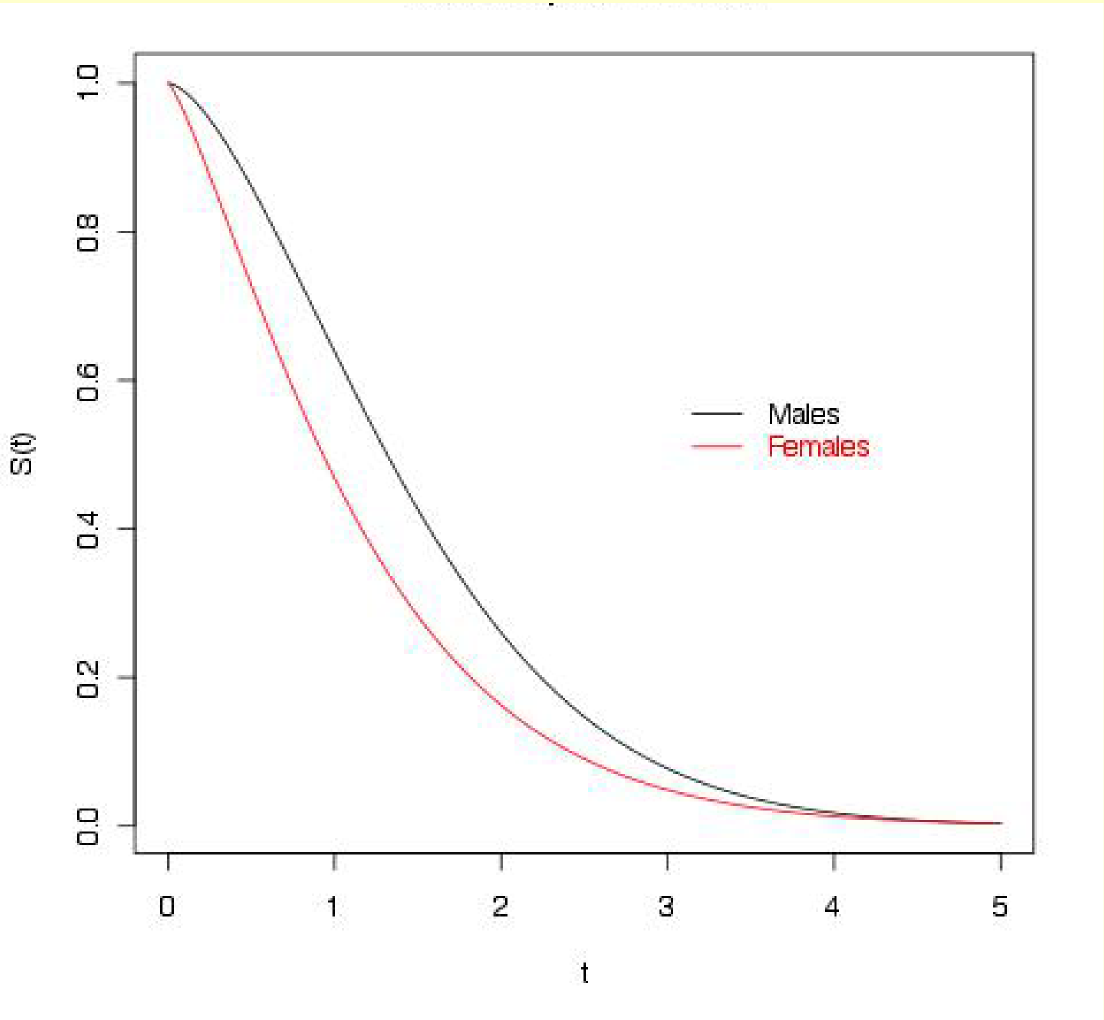
Males

Intercept	-0.911	-0.667	-0.544	-0.452	-0.891	-0.883	-0.995	-0.123
SE(Intercept)	0.08	0.158	0.122	0.081	0.102	0.034	0.125	0.151
Slope	1.793	1.752	1.651	1.416	1.56	1.558	2.463	1.115
SE(Slope)	0.104	0.205	0.158	0.091	0.093	0.044	0.194	0.197
R^2	0.987	0.948	0.965	0.98	0.976	0.997	0.982	0.889
Global Parameter:								
Alpha = 0.603								
Beta = 1.585+/-0.032								
R^2 = 0.971								

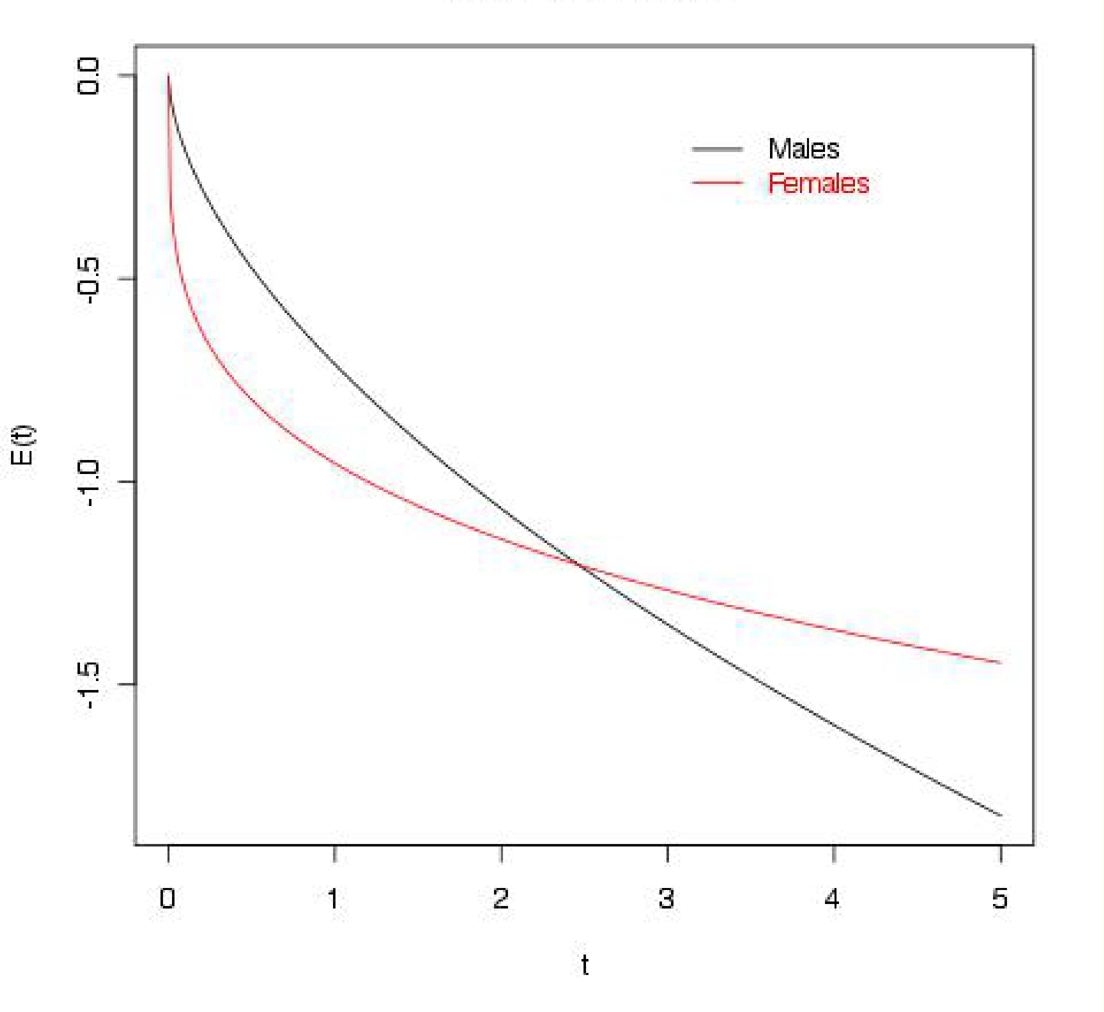
Females

Intercept						
SE(Intercept)						
Slope						
SE(Slope)						
R^2						
Global Parameters:						
Alpha = 0.803						
Beta = 1.258+/-0.047						
R^2 = 0.956						

STANDARDIZED DEMAND CURVES



ELASTICITY CURVES



(1) For both males and females, we have “decreasing elasticity”; consumption reduction accelerates as prices increases;

(2) What’s interesting is the two curves are crossing at a very high price; for lower prices the consumption for females drops faster first but it becomes slower at higher prices.

(3) One possible explanation is that female rats are weaker (larger α, early reduction) but more addicted (smaller β, more resistant to reduction, difference narrows down).

Due As Homework

#7.1 We have a data set on 86 smokers (File: Cigarettes); three outcome or response variables are Carbon monoxide, Cotinine (a derivative of Nicotine), and NNAL (a derivative of NNN, a toxin only comes from tobacco products). Data for 3 other explanatory variables are also included: Age, Gender (1=female), and Cigarettes per Day (CPD). Let Y= Cotinine & X=CPD; compare the slopes for men and women.

#7.2 Refer to the “Rat Demand Data” file and focus on the group of female rats (Only (a) is required):

(a) Compare the slopes and, if the difference is not significant, calculate the weighted average and its standard error.

(b) Compare the intercepts and, if the difference is not significant, calculate the weighted average and its standard error.

(c) Show how to obtain the global parameters α and β.

#7.3 For the Kidney activities, we have a “control” data set for which investigator lowered pressure in order to perturb the kidney equilibrium point so as to provide a range of values for regression; there were 5 observations for each of 10 kidneys, data are given on next page). Compare the slopes and, if the difference is not significant, calculate the weighted average and its standard error.

(This exercise is optional)

Kidney i	Slope, b_i	SSX_i
1	0.00967	1384
2	0.04784	360
3	0.03134	753
4	0.01928	3153
5	0.01928	3050
6	0.01747	4575
7	0.04817	1570
8	0.01893	4175
9	0.04233	719
10	0.02706	885