STA 302: Regression Analysis

Greeting, Syllabus, Rules

Statistics

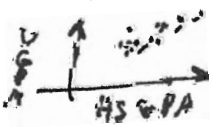
Goal: To obtain knowledge from noisy numerical data.

One approach: Study relationships between variables.

← Case: Sample of n cases (people, rats, cars, hospitals)

← Variable: Piece of information recorded for each case (age, sex, mpg, # nurses etc)

Data file: Rectangular array (Spreadsheet) { Rows = Cases, Cols = Variables

Scatterplot, least squares line 

Related = Not independent

Independent: Conditional distribution of Y given $X = x$ does not depend on x

$$f(y|x) = f(y) \iff \frac{f(x,y)}{f(x)} = f(y)$$

$$\iff f(x,y) = f(x)f(y)$$

x = Independent (Explanatory) Var

y = Dependent (Response) Var

Which is which

Seek to predict y from x (UGPA ← y, HSGPA ← x)

x may contribute to y (Exercise → Blood pressure)

Regression (term based on a misunderstanding) is one way the distribution of y might depend on x. Ind for $i = 1, \dots, n$

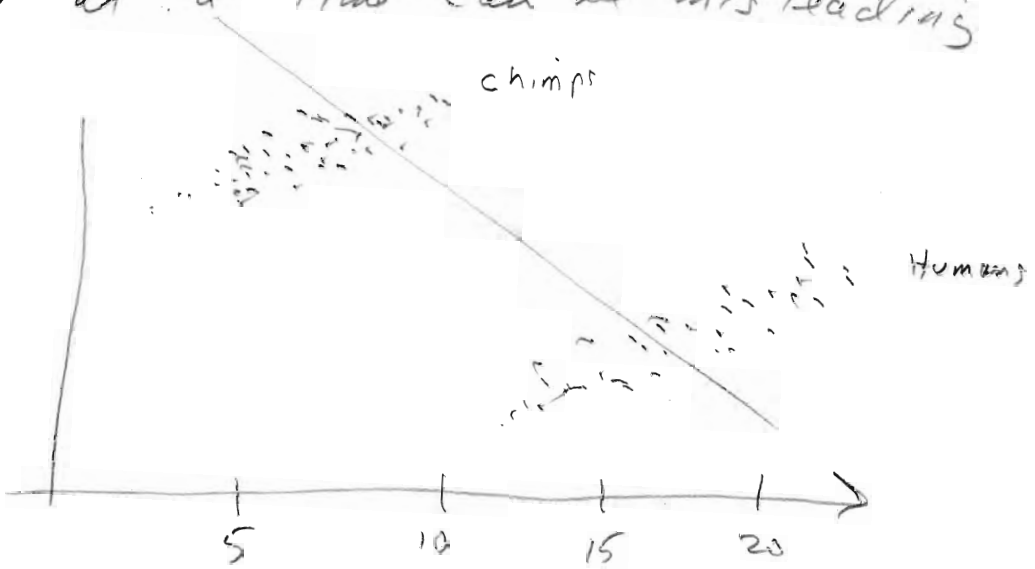
$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad E(\varepsilon_i) = 0,\ V(\varepsilon_i) = \sigma^2$$

or $\varepsilon_i \sim N(0, \sigma^2)$

$$E(Y_i) = \beta_0 + \beta_1 x_i,\ V(Y_i) = \sigma^2$$

A MODEL

But one IV at a time can be misleading



Multiple Regression

In scalar form

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

$\varepsilon_1, \ldots, \varepsilon_n$ independent $E(\varepsilon_i) = 0$, $V(\varepsilon_i) = \sigma^2$

or $\varepsilon_i \sim N(0, \sigma^2)$

In matrix form $\underset{\sim}{Y} = \underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$

$$\underset{\sim}{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \underset{\sim}{X} = \begin{bmatrix} 1 & 14.2 & \cdots & 1 \\ 1 & 11.9 & & 0 \\ 1 & 3.7 & & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 6.75 & & 1 \end{bmatrix} \quad \underset{\sim}{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \underset{\sim}{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\varepsilon \sim MVN(0, \sigma^2 I_n)$$

Tells us we need

- Matrix Algebra
- Random vectors, esp multivariate normal
- Software to compute all this

Tests & CI all depend on the normal distribution

Lead to χ^2, t, F

We'll do the distribution theory

use Change of Vars $E(g(x)) = \int g(x) f_x(x) dx$ a lot.

Moment-generating functions

X is a RV

$$M_X(t) = E(e^{Xt}) = \begin{cases} \sum_x e^{xt} p_X(x) \\ \int_{-\infty}^{\infty} e^{xt} f_X(x)\, dx \end{cases}$$

Correspond uniquely to probability distributions. We will use (without proof) MGF of

$X \sim N(\mu, \sigma^2)$ $\quad M_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$

$X \sim \chi^2(\nu)$ $\quad M_X(t) = (1-2t)^{-\nu/2}$

No derivation required here

$e^{\mu t + \frac{1}{2}\sigma^2 t^2}$ $\qquad (1-2t)^{-r/2}$

MGFs correspond uniquely to probability distributions (no proof : Advanced)

Ex $\quad X \sim N(\mu, \sigma^2), \quad Z = \dfrac{X-\mu}{\sigma}$

$$M_Z(t) = E\, e^{\frac{X-\mu}{\sigma} t} = E\left(e^{X(t/\sigma)} e^{-\frac{\mu t}{\sigma}}\right)$$

$$= e^{-\frac{\mu t}{\sigma}} M_X(t/\sigma) = e^{-\mu t/\sigma} e^{\mu(t/\sigma) + \frac{1}{2}\sigma^2 (t/\sigma)^2}$$

$$= e^{-\mu t/\sigma} e^{\mu t/\sigma} e^{\frac{1}{2}t^2} = e^{t^2/2} = e^{0t + \frac{1}{2}1t^2} \qquad N(0,1)$$

Properties (well known)

$$M_{aX}(t) = M_X(at)$$

$$M_{(X+c)}(t) = e^{ct} M_X(t)$$

$$E(e^{(X+c)t}) = E(e^{Xt + ct})$$
$$= E(e^{Xt} e^{ct})$$
$$= e^{ct} E(e^{Xt})$$

If $X_1 \perp X_2$ are ind,

$$M_{X_1 + X_2}(t) = M_{X_1}(t) M_{X_2}(t) \qquad \text{same } t$$

Proof

Next page

$$M_{X_1+X_2}(t) = \iint e^{(x_1+x_2)t} f_{X_1,X_2}(x_1,x_2)\,dx_1\,dx_2$$

$$= \iint e^{x_1 t} e^{x_2 t} f_{X_1,X_2}(x_1,x_2)\,dx_1\,dx_2$$

$$\overset{\text{ind}}{=} \iint e^{x_1 t} e^{x_2 t} f_{X_1}(x_1) f_{X_2}(x_2)\,dx_1\,dx_2$$

$$= \int e^{x_2 t} f_{X_2}(x_2) \underbrace{\left[\int e^{x_1 t} f_{X_1}(x_1)\,dx_1\right]}_{M_{X_1}(t)} dx_2$$

$$= M_{X_1}(t)\, M_{X_2}(t)$$

And by the uniqueness of MGFs, this is an if and only if. That is

X_1 & X_2 are independent if and only if $M_{X_1+X_2}(t) = M_{X_1}(t)\, M_{X_2}(t)$

Extends to $X_1 \cdots X_n$ ind, then

$$M_{\sum X_i}(t) = \prod_{i=1}^{n} M_{X_i}(t)$$

So, let $X_1 \cdots X_n \overset{iid}{\sim} N(\mu, \sigma^2)$,

$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Find dist of $\bar{X}_n$.

$$M_{\bar{X}}(t) = M_{\frac{1}{n}\Sigma X_i}(t) = M_{\Sigma X_i}(t/n)$$

$$= \prod_{i=1}^{n} M_{X_i}(t/n) = \prod_{i=1}^{n} e^{\mu t/n + \frac{1}{2}\sigma^2 (t/n)^2}$$

$$= e^{\sum_{i=1}^{n}(\mu t/n) + \sum_{i=1}^{n} \frac{1}{2}\sigma^2 t^2/n^2}$$

$$= e^{n(\mu t/n) + n\left(\frac{1}{2}\sigma^2 t^2/n^2\right)}$$

$$= e^{\mu t + \frac{1}{2}\frac{\sigma^2}{n} t^2}$$

$$= e^{\mu t + \frac{1}{2}\frac{\sigma^2}{n} t^2}$$

MGF of $N\left(\mu, \frac{\sigma^2}{n}\right)$

So $\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

$X_1 \cdots X_n \overset{ind}{\sim} \chi^2(r_i)$ $\quad Y = \Sigma X_i$

$$M_Y(t) = \prod_{i=1}^{n} M_{X_i}(t) = \prod_{i=1}^{n} (1-2t)^{-r_i/2}$$

$$= (1-2t)^{-\frac{\Sigma r_i}{2}}$$

$\chi^2(\Sigma r_i)$

~~Do the sad problems~~

A KEY TOOL for deriving the F and t distributions

Let $X_1 \perp X_2$ be independent, $Y = X_1 + X_2$

$X_1 \sim \chi^2(r_1)$, $Y \sim \chi^2(r_1 + r_2)$ $\quad r_1 > 0$, $r_2 > 0$

Then $X_2 \sim \chi^2(r_2)$

Proof By independence, $M_Y(t) = M_{X_1}(t) M_{X_2}(t)$

$$\Rightarrow (1-2t)^{-\frac{r_1+r_2}{2}} = (1-2t)^{-\frac{r_1}{2}} M_{X_2}(t)$$

Multiply both sides by $(1-2t)^{r_1/2}$

$$\Rightarrow (1-2t)^{-r_2/2} = M_{X_2}(t)$$

MGF of $\chi^2(r_2)$, done

One more ...

$Z \sim N(0,1)$, $Y = Z^2$. Show $Y \sim \chi^2(1)$

$$M_Y(t) = M_{Z^2}(t) = \int_{-\infty}^{\infty} e^{z^2 t} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(z^2 - 2z^2 t)} dz$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(1-2t)}{2} z^2} dz \quad \leftarrow \text{like } \tfrac{1}{\sigma^2}$$

$$= \frac{1}{(1-2t)^{1/2}} \int_{-\infty}^{\infty} \frac{(1-2t)^{1/2}}{\sqrt{2\pi}} e^{-\frac{(1-2t)}{2} z^2} dz$$

$$= 1$$

$$= (1-2t)^{-1/2}$$

MGF of $\chi^2(1)$

There are some homework problems

THIS will save technical headaches later.

Because MGFs correspond uniquely to probability distributions (densities do not), distributions can be <u>defined</u> in terms of their MGFs.

So we will define $X \sim N(\mu, \sigma^2)$ as meaning $M_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$, where $\sigma^2 \geq 0$.

Now observe. Let Y be a <u>degenerate</u> RV with $P(Y = \mu) = 1$,

So $M_Y(t) = E(e^{Yt}) = \sum_{y : P_Y(y) > 0} e^{yt} P(y)$

$= e^{\mu t} \cdot 1 = e^{\mu t}$

MGF of $N(\mu, 0)$

So in this sense, degenerate RVs are normal

Think of $\lim_{\sigma^2 \to 0} f_X(x)$, $X \sim N(\mu, \sigma^2)$