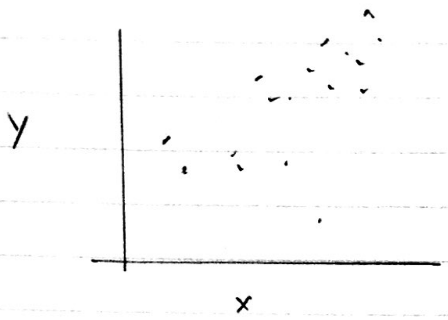
STAT 423

WIN 2016

Regression Models



Things to Learn

prob. model
regression model
linear regression model
prediction error
cross validation

Questions:

(Description) for each given x-value, what is dist of y-vals?
(Predictive) for a given x, what would we predict for a new y-value?
(Inferential) How does the dist of y change with x?

Note: In these questions, x and y have separate roles.
* Interest is in distribution of y for specific x's. *

Common Terminology:

x	y
input variable	output var
control variable	response
predictor	outcome
covariate	"variable of interest"
~~independent var~~	~~dependent var~~

Questions about the dist of y as a funct. of x can be addressed with regression models

Probability Model: $P(y) : y \in \mathbb{R} \quad \int_{\mathbb{R}} p(y)\,dy = 1, \quad \sum_{y \in \mathcal{Y}} p(y) = 1$

Population quantities

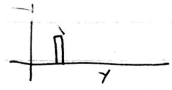
$$E[y] = \int y\,p(y)\,dy$$

$$V[y] = \int (y - E(y))^2\,p(y)\,dy$$

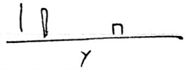
$$SD(y) = V(y)^{1/2}$$

Interpretation (Repeated Experiments)

Run 1 of an experiment $\Longrightarrow$



Run 2 "

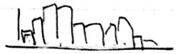


⋮

Run 27 $\Longrightarrow$

⋮

Run 1000 $\Longrightarrow$

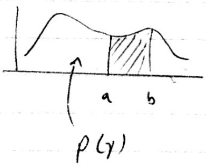


"Run ∞"

population $p(y)$

Objective of stat. inference: Learn about the process $(p(y))$ a finite # of experiments. $\{y_1 \ldots y_n\}$

Interpretation (Repeated sampling)



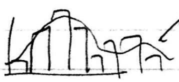
Given a large (or "infinite") population of y-values let $p(y)$ be the corresponding pdf.

Let Y be a randomly selected element of the population:

$$Pr(Y \in (a,b)) = \int_a^b p(y)\,dy$$

Sampling Model: In a repeated experiment or survey, we obtain a sample $\{y_1 \ldots y_n\}$ of iid observations from $p(y)$.

$$Y_1 \ldots Y_n \overset{\text{iid}}{\sim} p(y).$$



sample histogram

Objective of Stat. inference

Learn about the population $(p(y))$ from $\{y_1 \ldots y_n\}$.

Regression Model:

A regression model of Y given X is a probability model for y, for each possible x.

$\{ p(y|x) : x \in \mathcal{X} \}$

↖ set of possible x-values.

Example: x = mothers height (inches)   Questions: descript ..
y = daughters height   predict ..
inferential ..

$$\left.\begin{array}{l} p(y|x=50) \\ p(y|x=51) \\ \vdots \\ p(y|x=75) \end{array}\right\} \leftarrow$$ represents the population of heights of daughters whose mothers were 51 in tall.

$= \{ p(y|x) : x \in \mathcal{X} \}$

Mean + Variance Models

$p(y\|x)$	$E(Y\|x)$	$V(Y\|x)$
$p(y\| 50)$	$E[Y\|X=50] = \int y\, p(y\|50)\, dy$	$\mathrm{var}(Y\|X=50)$
$p(y\| 51)$	$\vdots$	$\vdots$
$\vdots$		
$p(y\| 75)$	$E[Y\|X=75] = \int y\, p(y\|75)\, dy$	$\mathrm{var}(Y\|X=75)$

Linear Regression Model: A regression model is linear (in expectation) if

$$E[Y|X=x] = \beta_0 + \beta_1 x$$ for some (unknown) values of β_0, β_1.

contrast to a "nonparametric" mean model

$E[Y|X=x] = \mu_x$, μ_x unrelated to $\mu_{x'}$

Why a linear regression model

Linear regression: $E[Y|X=x] = \beta_0 + \beta_1 x$

$E[Y|X=x+1] = \beta_0 + \beta_1 (x+1)$

$E[Y|X=x+1] - E[Y|X=x] = \beta_1$, for all x

Seems like a very strong assumption!

Nonparametric regression: $E[Y|X=x] = \mu_x$

$E[Y|X=x+1] = \mu_{x+1}$

$\mu_{x+1} - \mu_x$ could be anything,

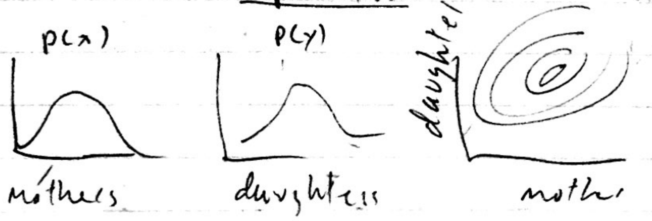
* is less restrictive
* makes fewer assumptions

Why might {being more restrictive, making more assumptions} be useful?

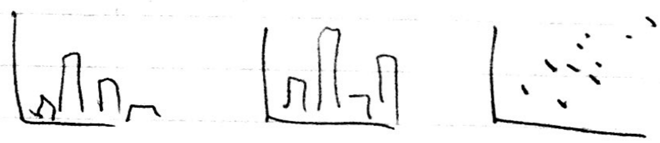
Example: X = height of mother $x \in \{50, 51, \ldots, 75\}$

Y = height of daughter $y \in \{50, 51, \ldots, 75\}$

Population



Sample Data



Questions:

(Descriptive): What is the sample/empirical dist of y for each x?

(Inferential) What is $\left\{\begin{array}{ll} \text{population dist} & p(y|x) \\ \text{expectation} & E(y|x) \\ \text{SD} & SD(y|x) \end{array}\right\}$ of y for each x?

(Prediction) If a mother with height x is randomly selected from the population, what would you predict her daughters height x is?