HANDOUT 1

Simple linear regression

- Summarize relationship between predictor variable y and explanatory variable x :

$$\underbrace{y = f(x)}_{\text{'explainable' part of } y} + \underbrace{\varepsilon}_{\substack{\text{random error, scatter} \\ \text{'unexplainable' part}}}$$

Trade-off : simple vs. flexible models

linear model $\longleftrightarrow$ smoother, local average

and models in between (e.g. nonlinear)

simple	flexible models
linear model	smoother, local average
bias ↑	bias ↓
variance ↓	variance ↑

Why? Simple models use all of the data to generate the regression line / curve, whereas the local methods use only a subset of the data for each local decision : more data $\Rightarrow$ lower variance.

Typically, we fit a model to the data using <u>least squares</u> LS.

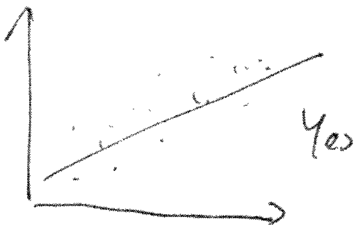
$$\min_{\beta_0,\beta_1} Q \;, \quad Q = \sum_{i=1}^{n} \big(\underbrace{y_i - (\beta_0+\beta_1 x_i)}_{\text{'error'}}\big)^2$$

all observations contribute equally

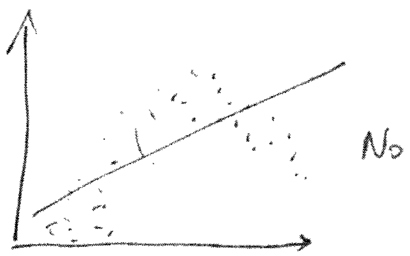
errors are squared, so positive & negative errors (above & below the regression line $\beta_0+\beta_1 x$) are equally important.

BASIC ASSUMPTIONS BEHIND THE LS FIT

(1) Linear model is sufficient to describe the data.



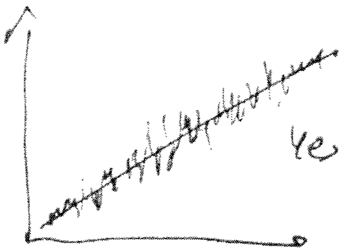
$y_i = \beta_0+\beta_1 x_i + \varepsilon_i$



O/w — regression line is an inadequate summary of the $y \sim x$ relationship.

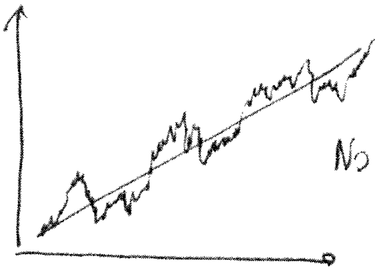
(2) Symmetric errors (since we use $(\;)^2$, letting positive & negative errors contribute equally)

(3) Uncorrelated errors (o/w shouldn't be using a simple sum $\sum_{i=1}^{n}$ in Q)



Yes

$Cov(\varepsilon_i, \varepsilon_j) = 0$



No $\Rightarrow$ Timeseries

(4) Constant variance (o/w shouldn't let all observations contribute equally; large variance - less information about line)

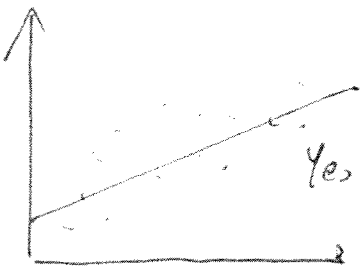
Yes

$V(\varepsilon_i) = \sigma^2$

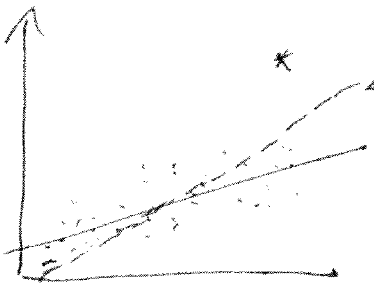
No

$\Rightarrow$ use <u>weighted</u> least squares (WLS)

(5) No outliers (o/w shouldn't let all observations contribute equally)

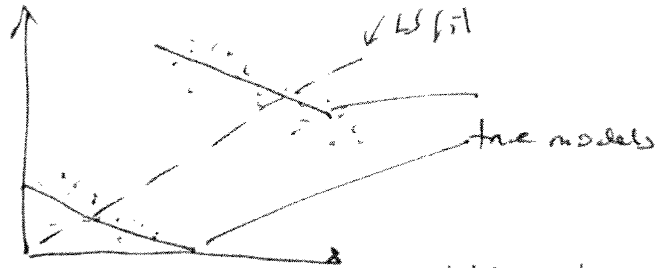


Yes



Other things to look out for

– groups in data



LM not sufficient.

– uneven spread in X

(like the animal data)

Sometimes [① – ⑤] doesn't hold for the data, but a simple transformation (like $\log(\ )$, $\sqrt{(\ )}$, $\frac{1}{(\ )}$) can sort this out.

Graph! Graph! Graph! Try different transformations & models to get as close as possible to ① – ⑤ before proceeding with your analysis.

Note: in regression we model the conditional distribution $y|x$ (not (x,y) jointly). x is assumed known & fixed, only ε is random $\Rightarrow$ y is random through ε.

Extra assumption on ε: $\varepsilon \sim N(0, \sigma^2)$

$\rightarrow$ makes some inferences easier.

(more later)

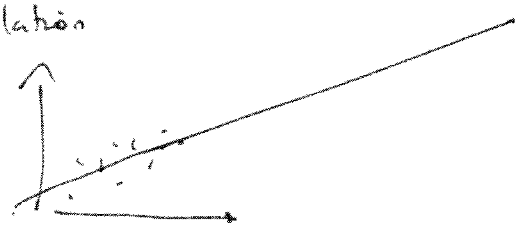
WARNING

- Watch out for omitted variables.

$$\text{true: } y = f(z) + \varepsilon \quad \Rightarrow \quad \text{model } \underbrace{y = h(x) + \varepsilon'}_{\text{artificial relationship}}$$
$$x = g(z) + \eta$$

- missing values – why/how missing?

- extrapolation



interpretation out here is invalid.

- causal interpretation of regression line

- regression effect, regression toward the mean

(if an observation is the most extreme in x, it has nowhere to go but down in its rank in y).

LS estimates

$Q = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$ = cost function to ~~minize~~ minimize

$$\begin{cases} \dfrac{\partial Q}{\partial \beta_0} = \sum (y_i - (\beta_0 + \beta_1 x_i)) \cdot -2 = 0 \\ \dfrac{\partial Q}{\partial \beta_1} = \sum (y_i - (\beta_0 + \beta_1 x_i)) \cdot -2\, x_i = 0 \end{cases}$$

$$\Rightarrow \begin{cases} \sum y_i = n\beta_0 + (\sum x_i)\beta_1 \quad (1) \\ \sum x_i y_i = (\sum x_i)\beta_0 + (\sum x_i^2)\beta_1 \quad (2) \end{cases}$$ "The Normal Equations"

A) Solve for β_0 (1) $\Rightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ (where $\bar{y} = \frac{1}{n}\sum y_i$, $\bar{x} = \frac{1}{n}\sum x_i$)

B) Plug $\hat{\beta}_0$ into (2) $\Rightarrow \sum x_i y_i = (\sum x_i)(\bar{y} - \hat{\beta}_1 \bar{x}) + (\sum x_i^2)\hat{\beta}_1$

$$\Rightarrow \hat{\beta}_1 = \frac{\sum x_i y_i - \sum x_i \sum y_i}{\sum x_i^2 - (\sum x_i)^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

Note; $\hat{\beta}_1 \cong \dfrac{\underbrace{Cov(x,y)}}{\underbrace{Cov(x,x)}_{\text{covariance}}} = \dfrac{Cov(x,y)}{\underbrace{V(x)}_{\text{variance}}} = \underbrace{\text{Correlation}(x,y)}_{=\ \rho \in [-1,1]} \sqrt{\dfrac{V(y)}{V(x)}}$

(only $= 1$ or -1 if $\varepsilon = 0$ exactly) ~~[illegible]~~

Consider x, y normalized such that $V(y) = V(x) = 1$. Then $\boxed{\hat{\beta}_1 = \rho}$

and $|\hat{\beta}_1| \le 1$, slope always less than 1

= regression effect.

Looking ahead — matrix notation

$$\frac{\partial Q}{\partial \underset{\sim}{\beta}} = \begin{pmatrix} \frac{\partial Q}{\partial \beta_0} \\ \frac{\partial Q}{\partial \beta_1} \end{pmatrix} \Rightarrow \underbrace{\begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix}}_{B} = \underbrace{\begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}}_{A} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$ Normal Equations

If A is invertible, solve for

$$\begin{pmatrix} \hat\beta_0 \\ \hat\beta_1 \end{pmatrix} = A^{-1}B$$

$$= \frac{\begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix}}{n\sum x_i^2 - (\sum x_i)^2} \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix}$$

$$= \frac{\begin{pmatrix} \sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i \\ -\sum x_i \sum y_i + n \sum x_i y_i \end{pmatrix}}{\underbrace{n\sum x_i^2 - (\sum x_i)^2}_{= \sum (x_i - \bar x)^2 \cdot n}}$$

Now $\Rightarrow$ Line 1: $$\frac{\bar y \sum x_i^2 - \bar x \sum x_i y_i}{\sum (x_i - \bar x)^2} = \bar y \frac{\sum x_i^2}{\sum (x_i - \bar x)^2} - \hat\beta_1 \bar x - \frac{\bar x \bar y \sum x_i}{\sum (x_i - \bar x)^2} = \bar y \underbrace{\left(\frac{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2}{\sum (x_i - \bar x)^2} \right)}_{=1} - \hat\beta_1 \bar x$$

$$= \bar y - \hat\beta_1 \bar x \ \checkmark$$

Line 2: $$\frac{\sum x_i y_i - \sum x_i y_i}{n \sum (x_i - \bar x)^2} = \frac{\sum (x_i - \bar x)(y_i - \bar y)}{\sum (x_i - \bar x)^2} = \hat\beta_1 \ \checkmark$$

Matrix notation

Very important!

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i=1,\dots,n \iff \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}}_{\underset{\sim}{X}} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$\iff \underset{n\times 1}{\underset{\sim}{y}} = \underset{n\times 2}{\underset{\sim}{X}} \underset{2\times 1}{\underset{\sim}{\beta}} + \underset{n\times 1}{\underset{\sim}{\varepsilon}}$$

vectors/matrices

multivariate case

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_{p-1} x_{i\,p-1} + \varepsilon_i \iff \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & x_{11} & x_{21} & \cdots & x_{p-1\,1} \\ 1 & x_{21} & x_{22} & & \vdots \\ 1 & x_{31} & x_{32} & & \\ \vdots & \vdots & \vdots & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{p-1,n} \end{pmatrix}}_{\underset{\sim}{X}} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \dots$$

$\underset{\sim}{X}$ is called the <u>design matrix</u>

$$\iff \underset{n\times 1}{\underset{\sim}{y}} = \underset{n\times p}{\underset{\sim}{X}} \underset{p\times 1}{\underset{\sim}{\beta}} + \underset{n\times 1}{\underset{\sim}{\varepsilon}}$$

$$Q = \sum_{i=1}^{n} \Big(\underbrace{y_i - (\beta_0 + \beta_1 x_{i1} + \dots + \beta_{p-1} x_{i\,p-1})}_{\substack{\text{error} \\ e_i}}\Big)^2 = \sum_{i=1}^{n} e_i^2 = (e_1\ e_2 \ \dots\ e_n)\begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

$$= \underset{\sim}{e}'\underset{\sim}{e}, \text{ where } \underset{\sim}{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

and e' = transpose of e

Now, can also write errors $e_i = (y_i - (\beta_0 + \beta_1 x_{i1} + \dots \beta_{p-1} x_{i,p-1}))$
in vector form $\underset{\sim}{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} = \underset{\sim}{y} - \underset{\sim}{X}\underset{\sim}{\beta}$

$\Rightarrow Q = \underset{\sim}{e}'\underset{\sim}{e} = (\underset{\sim}{y} - \underset{\sim}{X}\underset{\sim}{\beta})'(\underset{\sim}{y} - \underset{\sim}{X}\underset{\sim}{\beta})$

We want to minimize Q with respect to $\underset{\sim}{\beta}$.

$\Rightarrow$ Matrix derivative

$$\frac{\partial Q}{\partial \underset{\sim}{\beta}} = \underbrace{-2\underline{X}'}_{\text{inner derivative}}(y - \underline{X}\beta) = 0$$

$\Rightarrow \boxed{(\underline{X}'\underline{X})\underset{\sim}{\beta} = \underline{X}'y}$ Normal equations on matrix form.

"Interpretation" $(\underline{X}'\underline{X}) \sim Cov(\underline{X})$, where diagonal of $\underline{X}'\underline{X}$ matrix is the $Var(X_1), Var(X_2) \dots$ and off-diagonal elements are the covariances $Cov(X_1, X_2) \dots$

So if some X's are linearly dependent, or close to
$\rightarrow (\underline{X}'\underline{X})^{-1}$ may not exist, or be numerically unstable

$(\underline{X}'y) \sim Cov(\underline{X}, y)$ — measure of linear dependence between the

4) $(X'X)^{-1}$ exists $\Rightarrow$ $\underline{\hat{\beta} = (X'X)^{-1}X'y}$

Problems: if $(X'X)^{-1}$ numerically unstable, meaning X's are related $\Rightarrow$ difficult to interpret the meaning of the slopes $\hat{\beta}_1 \ldots \hat{\beta}_{p-1}$ - they are not just measuring the contribution of one x, if x's are related.

Back to simple linear regression $\Rightarrow$

Properties

- unbiased estimates

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$= \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2} = \sum k_i y_i$$

where $k_i = \sum \frac{(x_i - \bar{x})}{\sum_j (x_j - \bar{x})^2}$

That is, $\hat{\beta}_1$ is a weighted average of y-values!

Note, extreme x_i's, far from $\bar{x}$, contribute the most to the slope estimate

large k_i here