Statistical Linear Models

1: Functional vs. Stochastic Relations

Model: A mathematical approximation of the relationship among real quantities (equation & assumptions about terms)

Functional relationships are perfect. Realizations (X_i, Y_i) solve the relation $Y = f(X)$

Deterministic e.g. $Y = \cos(2.1x) + 4.7$

Although often an approximation to reality (e.g. the solution to a differential equation under simplifying assumptions), the relation itself is "perfect".

Statistical relationships are not perfect. There is a trend plus error. (Signal plus noise).

Stochastic - introduces some error in approximating Y (typically a functional relationship plus noise).

2: Simple Linear Regression Model

For a sample of n pairs $\{(X_i, Y_i)\}_{i=1}^{n}$, let

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad , i = 1, \dots, n$$

where

i) $Y_1, \dots, Y_n$ are realizations of the response variable.

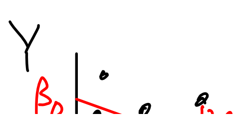
ii) $X_1, \dots, X_n$ are the associated predictor variables.

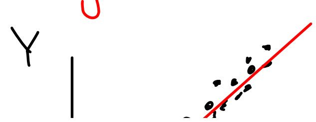
iii) β_0 is the intercept of the regression line.

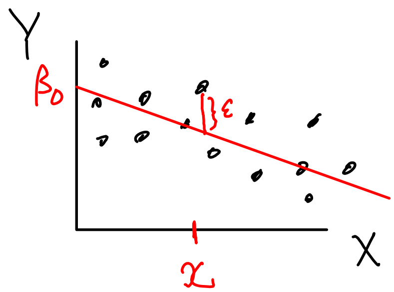
iv) β_1 is the slope of the regression line

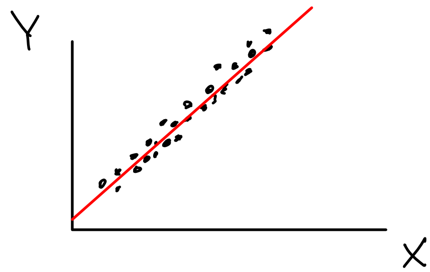
v) $\varepsilon_1, \dots, \varepsilon_n$ are unobserved, uncorrelated random errors.

This model assumes that X & Y are linearly related. (i.e. the mean of Y changes linearly with X).









3: Assumptions about the random errors

we assume that:

$$\boxed{\underbrace{E(\varepsilon_i) = 0}_{\text{mean } 0} \;;\; \underbrace{Var(\varepsilon_i) = \sigma^2}_{\text{constant variance}}, \;\; \underbrace{corr(\varepsilon_i, \varepsilon_j) = 0}_{\text{uncorrelated}} \; i \neq j}$$

$\beta_0 + \beta_1 X_i \longleftarrow$ Deterministic part of the model (fixed but unknown)

$\varepsilon_i \longleftarrow$ Random part of the model.

The goal of statistics is often to seperate signal from noise.

Note: $E[Y_i] = E[\beta_0 + \beta_1 X_i + \varepsilon_i] = \beta_0 + \beta_1 X_i + E[\varepsilon_i)$

$$= \beta_0 + \beta_1 X_i$$

$$Var[Y_i] = var[\beta_0 + \beta_1 X_i + \varepsilon_i] = var(\varepsilon_i) = \sigma^2$$

$$corr[Y_i, Y_j] = 0 \quad \text{for } i \neq j.$$

4: Simple Linear Regression using matrices

The SLR model can be written in matrix terms as

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

or equivalently

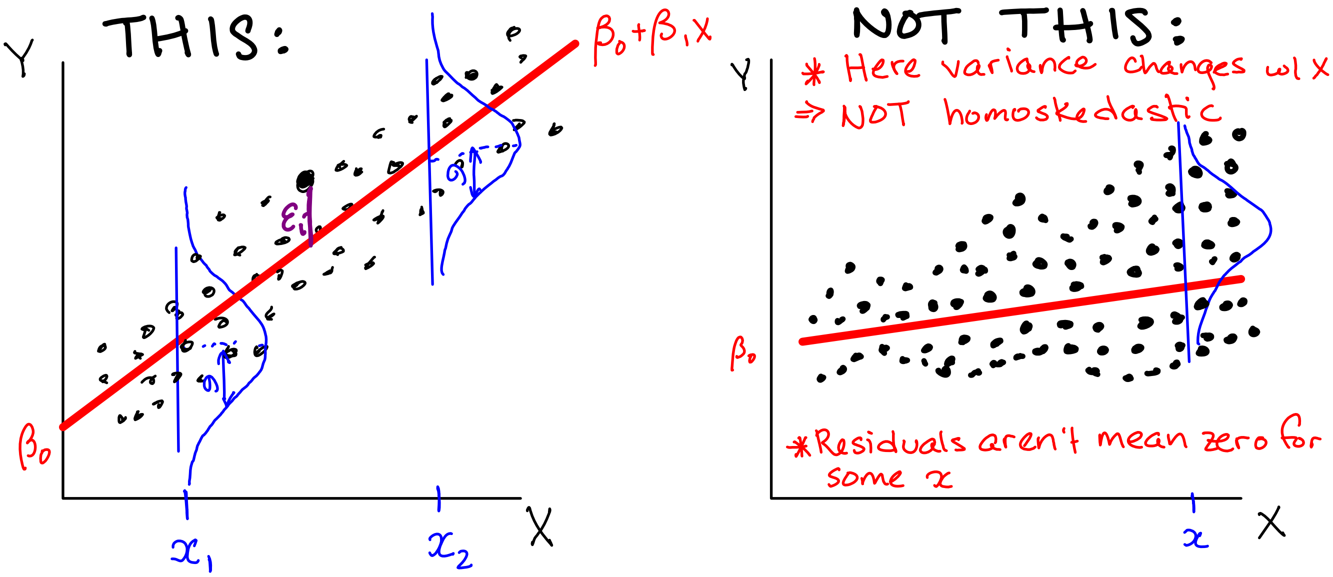
or equivalently

$$\underline{Y} = \underline{X}\,\beta + \varepsilon$$

5: Estimating β_0 and β_1

The assumptions of Simple Linear (as mentioned above)

1. $\varepsilon_i \sim N(0, \sigma^2)$ [on average, the errors cancel out]
2. ε_i are independent [obs. are indep.; e.g. no time dependence]
3. ε_i all have the same variance [homoskedasticity]



β_0 & β_1 are intrinsic parameters that describe the true linear relationship between Y & X. We don't know what they are!

Population:

$Y = \beta_0 + \beta_1 X + \varepsilon$

$\beta_0 = ?$ $\beta_1 = ?$

↓

Sample:

$(X_1, Y_1), \ldots, (X_n, Y_n)$

Used to calculate $\hat{\beta}_0, \hat{\beta}_1$

draw inferences from sample about true property of population.

$(x_1, Y_1), \ldots, (x_n, Y_n)$

Used to calculate $\hat{\beta}_0, \hat{\beta}_1$

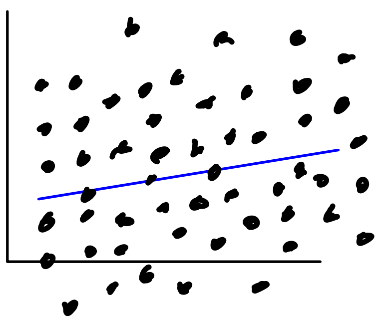
Goal: ① Find best possible estimates, $\hat{\beta}_0$ & $\hat{\beta}_1$ for β_0 and β_1

② Assess "signal-to-noise"

$H_0: \beta_1 = 0$ [line is flat → no linear relationship]

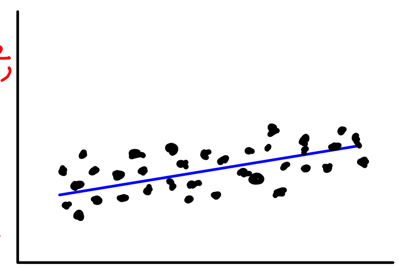
$H_a: \beta_1 \neq 0$ [line isn't flat ⇒ linear relationship exists]

Low signal-to-noise



The two lines have the same estimated slope, but in the data to the left, it is harder for us to conclude that the slope is non-zero

High signal-to-noise



We employ the Method of Least Squares to estimate β_0 & β_1.

We choose estimates $\hat{\beta}_0$ & $\hat{\beta}_1$ so that our line is as close as possible to

$$Y_i \approx \hat{\beta}_0 + \hat{\beta}_1 X_i$$

$$\begin{bmatrix} \hat{\beta}_0 + \hat{\beta}_1 X_1 \\ \hat{\beta}_0 + \hat{\beta}_1 X_2 \\ \vdots \\ \hat{\beta}_0 + \hat{\beta}_1 X_n \end{bmatrix} \approx \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

Two vectors are close if their Euclidean distance is small:

$$\|\vec{V}_1 - \vec{V}_2\| = \sqrt{\sum_{i=1}^{n} (V_1(i) - V_2(i))^2}$$

Choose $\hat{\beta}_0$ & $\hat{\beta}_1$ to minimize $\sqrt{\sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2}$

The square root function is monotonic, so we can ignore it and just minimize $\sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$

and just minimize $\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2$

Solve me!

$$\begin{cases} \dfrac{\partial}{\partial\hat{\beta}_0}\sum_{i=1}^{n}(Y_i-\hat{\beta}_0-\hat{\beta}_1X_i)^2=0 \\ \dfrac{\partial}{\partial\hat{\beta}_1}\sum_{i=1}^{n}(Y_i-\hat{\beta}_0-\hat{\beta}_1X_i)^2=0 \end{cases} \Rightarrow \begin{cases} \sum_{i=1}^{n}-2(Y_i-\hat{\beta}_0-\hat{\beta}_1X_i)=0 \\ \sum_{i=1}^{n}-2X_i(Y_i-\hat{\beta}_0-\hat{\beta}_1X_i)=0 \end{cases}$$

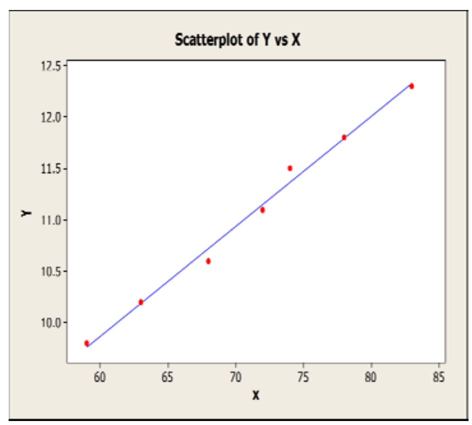
Solving yields "Least-Squares Estimators"

$$\boxed{\begin{aligned} \hat{\beta}_0 &= \bar{Y}-\hat{\beta}_1\bar{X} \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2} = \frac{S_{XY}}{S_{XX}} \end{aligned}}$$

Example: Data were collected measuring the age of a person (X) and their hip bone loss (Y)

Find the equation of the regression line.

X	Y	$X_i-\bar{X}$	$Y_i-\bar{Y}$	$(X_i-\bar{X})^2$	$(X_i-\bar{X})(Y_i-\bar{Y})$
59	9.8	-12	-1.2	144	14.9
63	10.2	-8	-0.8	64	6.7
68	10.6	-3	-0.4	9	1.3
72	11.1	1	0.1	1	0.1
74	11.5	3	0.5	9	1.4
78	11.8	7	0.8	49	5.3
83	12.3	12	1.3	144	15.1
$\bar{X}=71$	$\bar{Y}=11.0$			$S_{xx}=420$	$S_{xy}=44.8$



$\hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}} = \dfrac{44.8}{420} = 0.10\bar{6}$

$\hat{\beta}_0 = \bar{Y}-\hat{\beta}_1\bar{X} = 11.0-0.10\bar{6}(71) = 3.46$

$$\boxed{Y = 3.46 + .10\bar{6}X + \varepsilon}$$

This is our proposed model.

Now that we've fit a line to our sample data, we need

Now that we've fit a line to our sample data, we need to draw inferences about the population:

1. Were our initial assumptions reasonable?
- normally dist'd, zero-mean, independent, homoskedastic errors

If answer to (1) is "Yes" proceed.
Otherwise, transform data or fit a new model.

1. We test our assumptions via residual plots

Residual at data point i : $e_i = \underbrace{Y_i}_{\text{observation}} - \underbrace{\hat{Y}_i}_{\text{what the line predicts } = \hat{\beta}_0 + \hat{\beta}_1 X_i}$

a. Normality: normal probability plot should be straight

b. Independence: time ordered plot should look random and without patterns

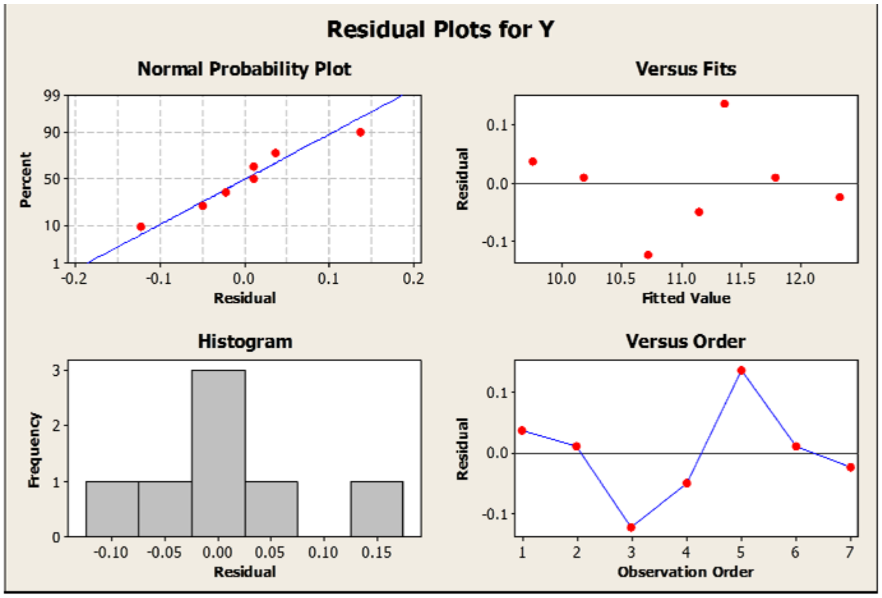
c. Homoskedasticity: residuals vs. X plot should have even width

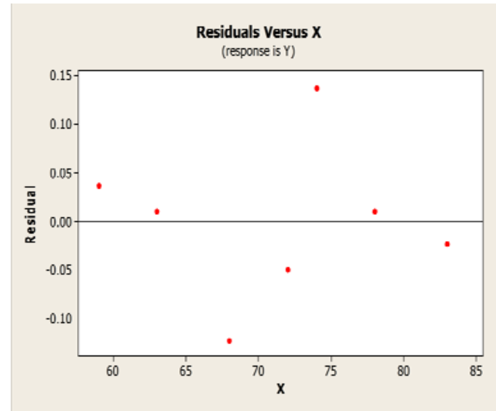
d. Mean of Zero: residuals vs. X should be centered about 0

e. Other: residuals vs. fitted values $\hat{Y}$

residuals vs. excluded variables: any patterns suggest variables to add to model.

Example:

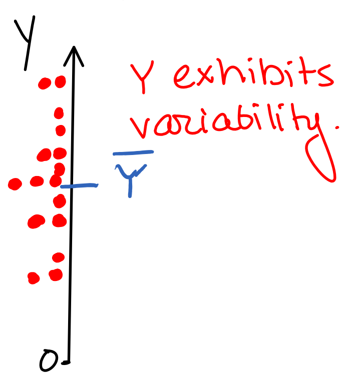




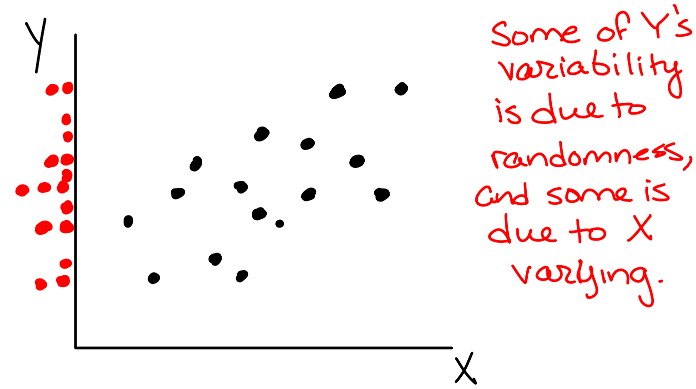
2. How much of Y's variability can be explained by a linear model with X? R^2 !!

Caution: Many novices treat R^2 as the be-all-and-end-all of a linear regression. We'll see why we need to be wary of R^2 and why it's only one component of a regression analysis!

If we just look at Y:



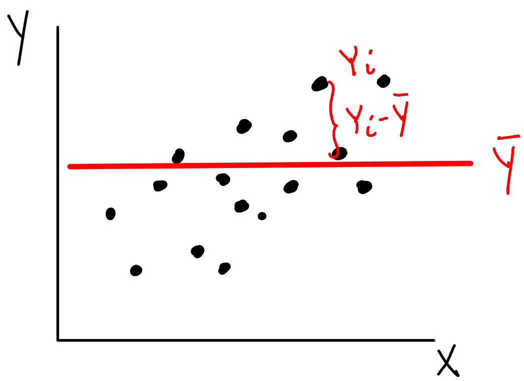
If we consider (X, Y) pairs:

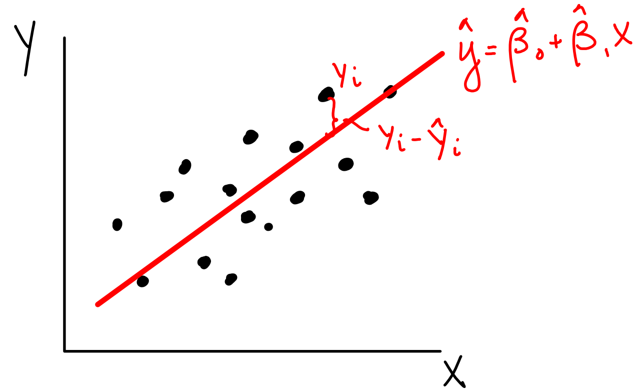


If Y is independent of X, the best model is a flat line at $\bar{Y}$:

If Y relates linearly with X, the best model is $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$:

at Y:





SST = Sum of Squares Total

$$= \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

SSE = Sum of Squared Errors

$$= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

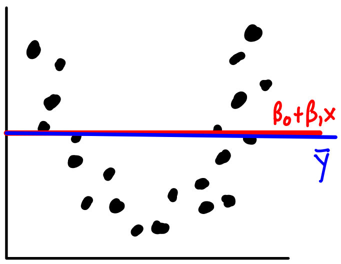
If there's a linear relationship between X & Y then SSE should be small relative to SST.

$$\boxed{R^2 = 1 - \frac{SSE}{SST}}$$

"Coefficient of determination" = percentage of Y's variability explained by X.

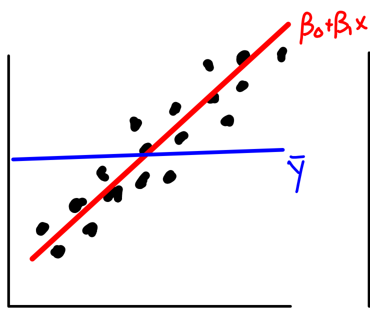
The closer to 1 R^2 is, the more our linear model outperforms a flat line model. This doesn't necessarily mean the linear model is good!

For each of the following pictures guess whether R^2 is high or low. For which pictures is a linear model a good choice?



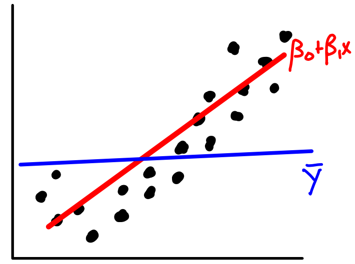
$R^2 \approx 0$

Linear Model



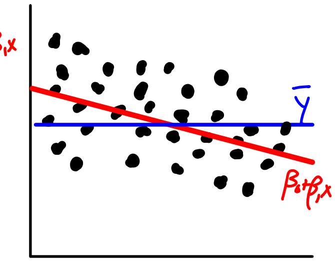
R^2 high

Linear Model IS



R^2 high

Linear model.



R^2 low

Linear Model Is

$R^2 \approx 0$ Linear Model NOT appropiate

R^2 high Linear Model IS appropriate

R^2 high Linear model NOT appropriate

R^2 low Linear Model Is appropriate

* ALWAYS PLOT Y VS. X BEFORE DECIDING WHETHER OR NOT TO USE LINEAR REGRESSION ! *

R^2 also ignores issues like: homoskedasticity, independence, normality, magnitude of slope. Need to assess fit in several ways.

3. Does a linear relationship exist?
   - We already examined linearity graphically
   - Is the true slope non-zero?

If satisfied with model after answering (2) & (3), proceed. Otherwise, fit a different model.

4. Can we make predictions from our model?