AMS 572 Data Analysis I
Simple Linear Regression

Pei-Fen Kuan

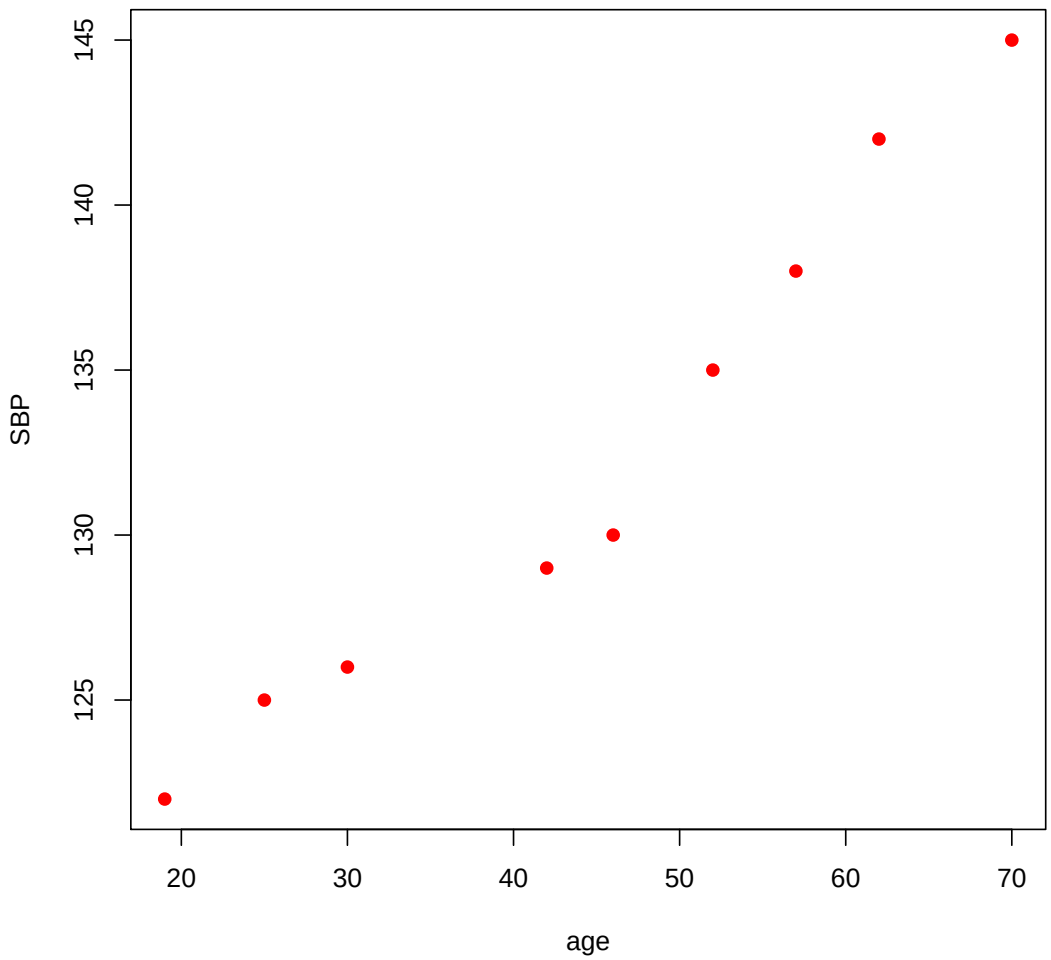
Applied Math and Stats, Stony Brook University

Example: SBP and Age

↑ systolic. blood pressure.

Obs	Age	SBP
1	19	122
2	25	125
3	30	126
4	42	129
5	46	130
6	52	135
7	57	138
8	62	142
9	70	145

Example: SBP and Age



Simple Linear Model

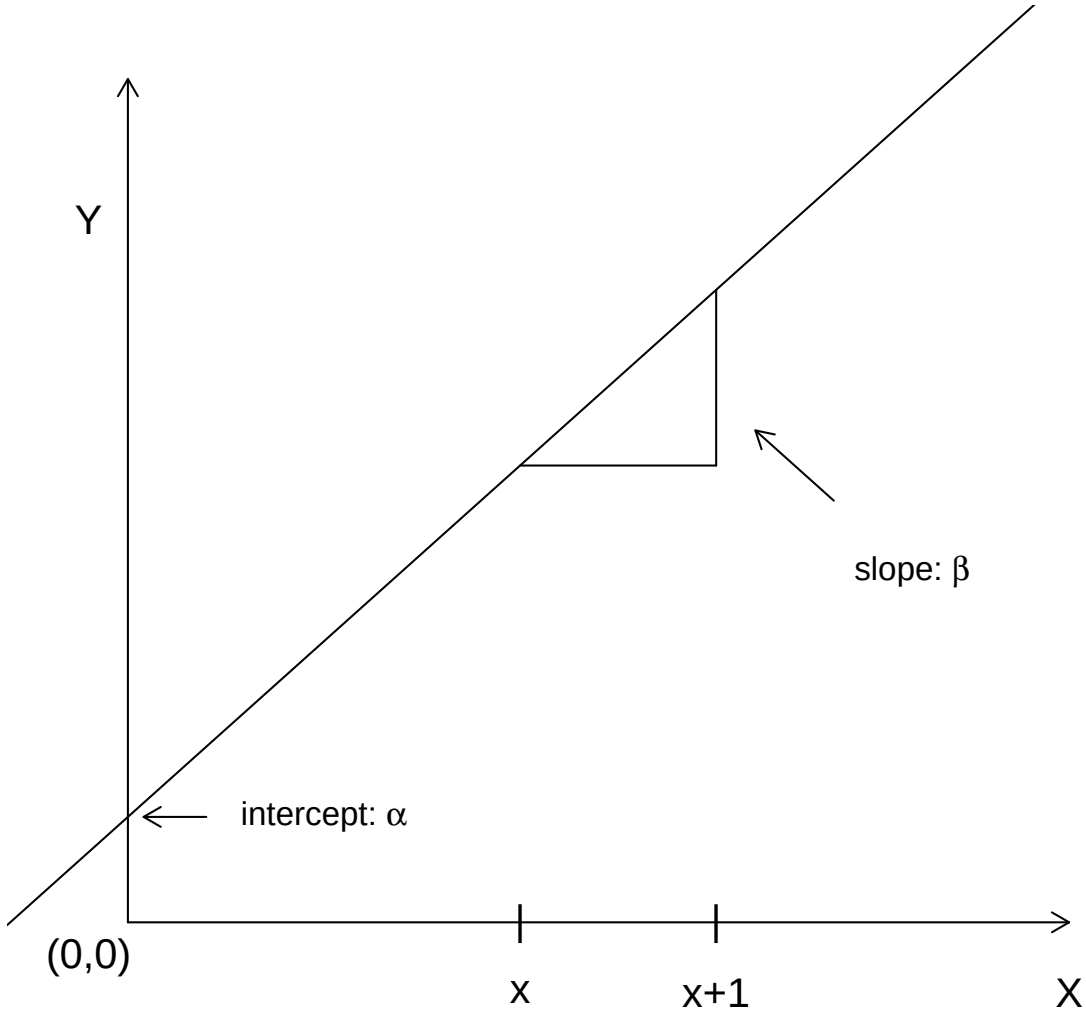
- Line

$$Y = \alpha + \beta X$$

- α = intercept; value of Y when $X = 0$.
- β = slope; change in Y when X change by 1 unit.

- Y : dependent variable, response variable
- X : covariate, independent variable, predictor

Simple Linear Model



Simple Linear Model with Error

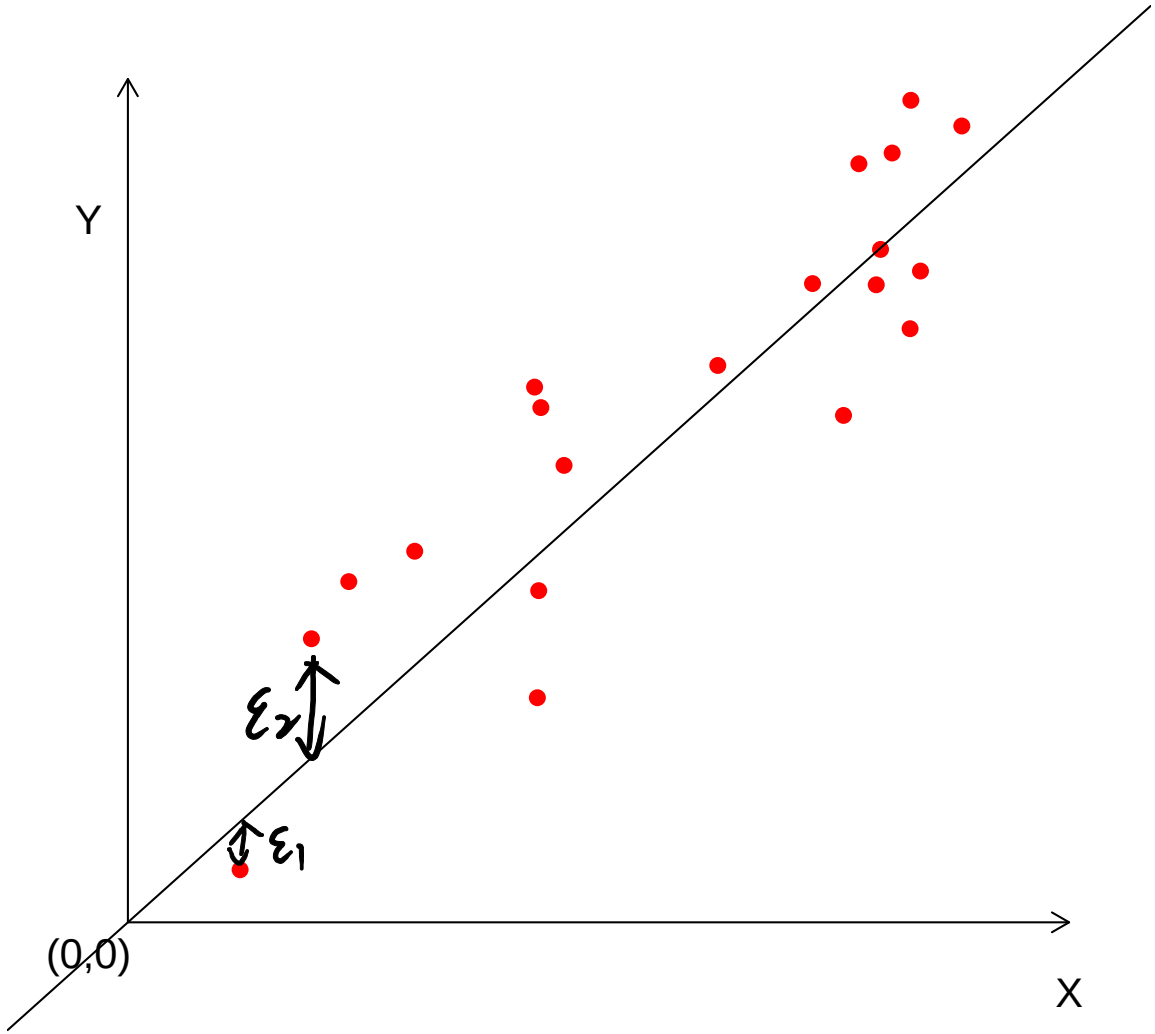
- Linear regression

$$Y = \alpha + \beta X + \epsilon$$

$$\epsilon = Y - \alpha - \beta X$$

- ϵ equals vertical distance from Y to line defined by $\alpha + \beta X$

Simple Linear Model with Error



Model Assumptions

- Data are (Y_i, X_i); $i = 1, 2, \ldots, N$
- Assume:
  1. Linearity: $Y_i = \alpha + \beta X_i + \epsilon_i$
  2. X's are fixed constants
  3. ϵ_i iid $N(0, \sigma^2)$

$$Y_i \underset{\text{indep}}{\sim} N(\alpha + \beta X_i, \sigma^2)$$

Least Squares Estimation

$f = \sum (Y_i - \alpha - \beta X_i)^2$

$\frac{\partial f}{\partial \alpha} = -2 \sum (Y_i - \alpha - \beta X_i) = 0.$

$\frac{\partial f}{\partial \beta} = -2 \sum (Y_i - \alpha - \beta X_i) X_i = 0.$

Solve for α, β.

$\hat{\alpha} = \bar{Y} - \hat{\beta} \bar{X}. \quad \hat{\beta} = \frac{\sum X_i Y_i - N \bar{X} \bar{Y}}{\sum X_i^2 - N \bar{X}^2}$

- Least squares estimators are values of α and β that minimize

$$\sum_{i=1}^{N} \epsilon_i^2 = \sum_{i=1}^{N} (Y_i - \alpha - \beta X_i)^2$$

$\bar{X} = \frac{\sum X_i}{N}$

$\bar{Y} = \frac{\sum Y_i}{N}.$

- Set partial derivatives equal to 0, solve for α and β
- Can also derive these estimators via maximum likelihood

Least Squares Estimation

$$\hat{\alpha} = \bar{Y} - \bar{X}\hat{\beta}$$

- Equivalent form:

$$\hat{\beta} = \frac{\sum_i X_i Y_i - N\bar{X}\bar{Y}}{\sum_i X_i^2 - N\bar{X}^2}$$

- Note if $X_i = Y_i$ for all i, then $\hat{\beta} =$ 1
- Also if $Y_i = \bar{Y}$ for all i, then $\hat{\beta} =$ 0

Least Squares Estimation

$Y_i = \alpha + \beta X_i + \varepsilon_i$

$\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

$Y_i \sim N(\alpha + \beta X_i, \sigma^2)$

- Predicted response (aka *fitted values*)

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$$

- Residual

$$r_i = Y_i - \hat{Y}_i$$

- Estimate variance by mean square error (MSE)

$$\hat{\sigma}^2 = s^2_{y.x} = \frac{1}{N-2} \sum (Y_i - \hat{Y}_i)^2$$

$$= \frac{1}{N-2} \sum r_i^2$$

Example: SBP and Age

$$\bar{Y} = 132.4; \bar{X} = 44.8 \qquad n = 9$$

$$\sum_i X_i Y_i = 54461; \sum_i X_i^2 = 20463$$

$$\hat{\beta} = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sum X_i^2 - n\bar{X}^2} = \frac{54461 - 9(44.8)(132.4)}{20463 - 9(44.8)^2}$$

$$= 0.45$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = 132.4 - 0.45(44.8)$$

$$= 112.3$$

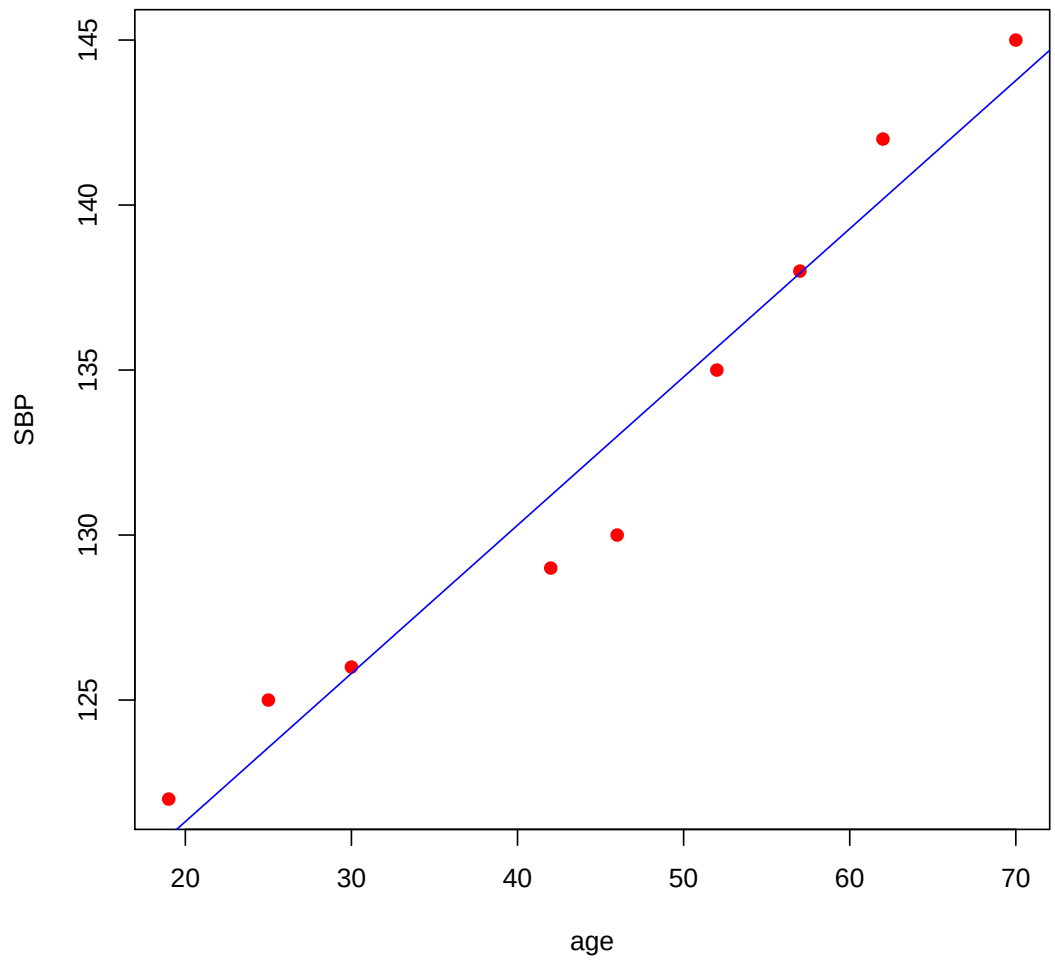
$$\hat{Y}_i = 112.3 + 0.45 X_i$$

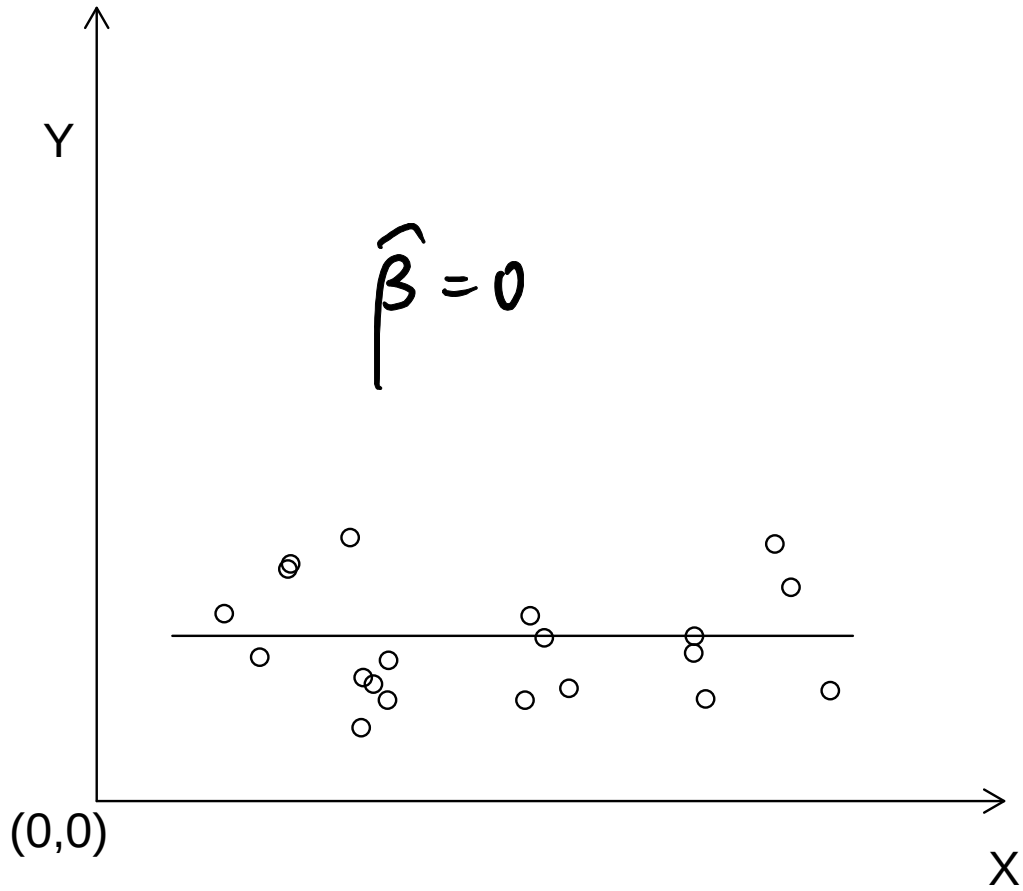
Example: Interpretation

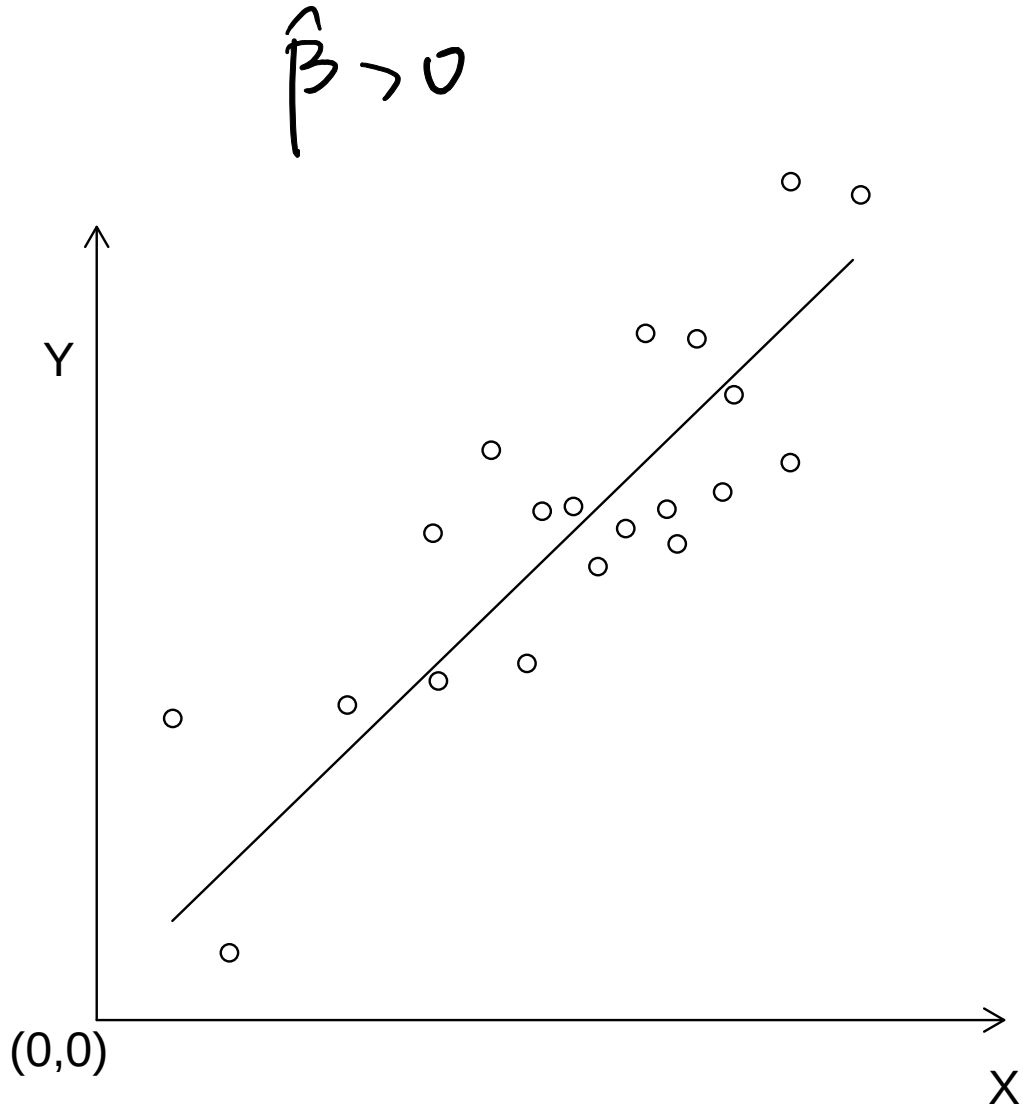
- $\hat{\beta} = 0.45 \rightarrow$ expected SBP increases 0.45 (mmHg) for each one year increase in age
- $\hat{\alpha} = 112.3 \rightarrow$ Beware of extrapolation!!

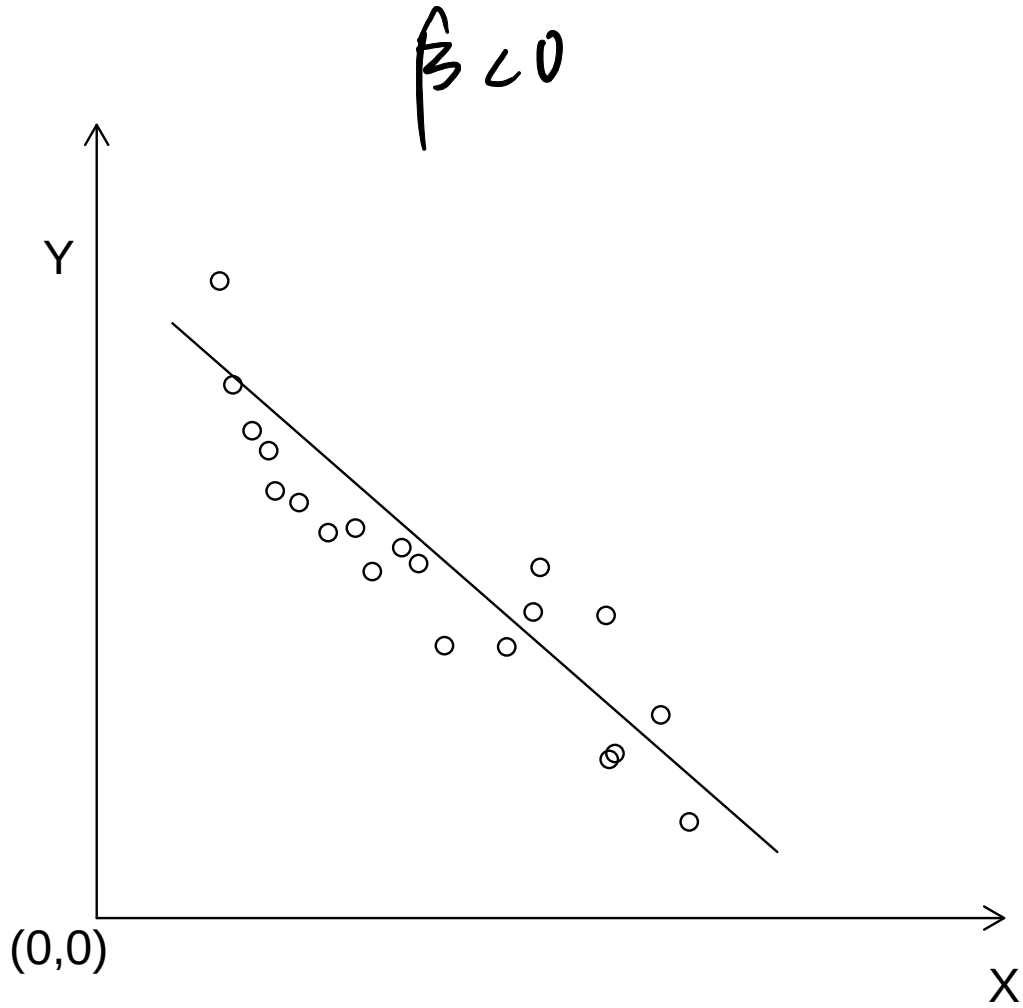
Since no x(age) values is observed for age < 19, the intercept does not make much sense.

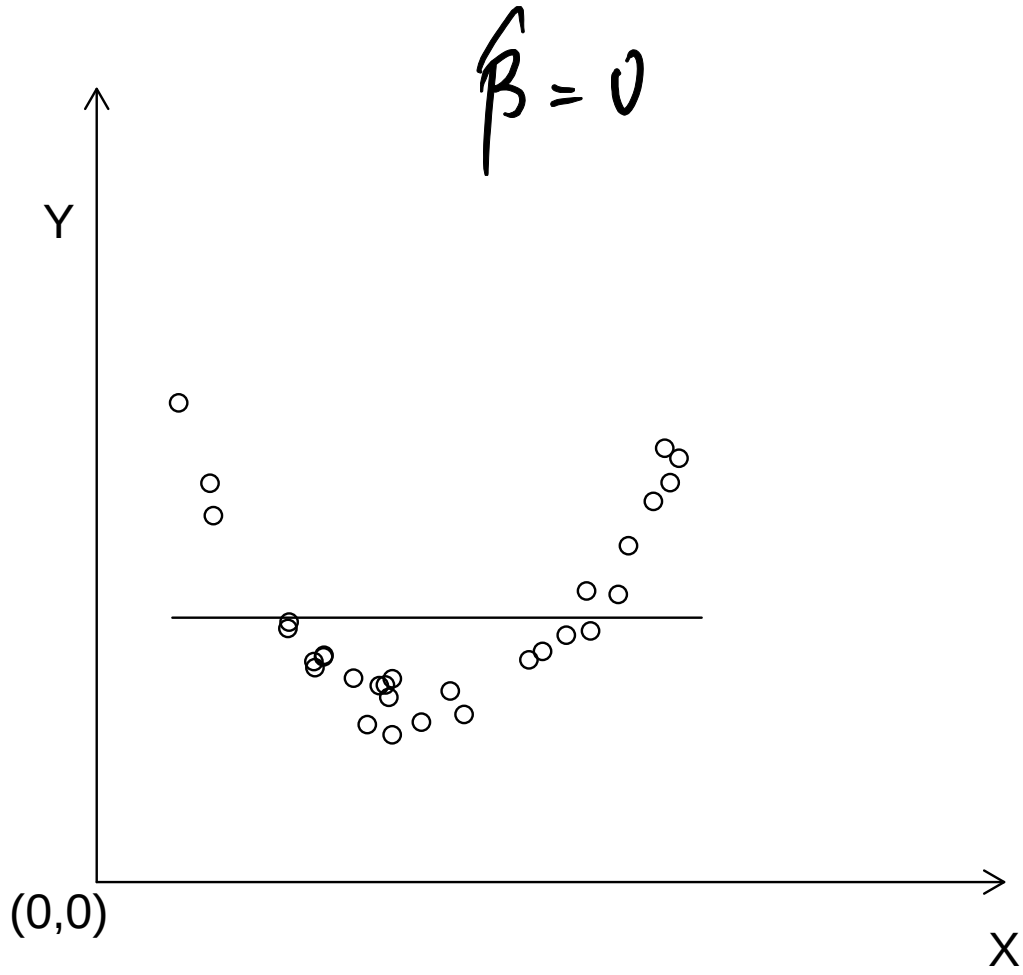
Example: SBP and Age




Y
$\hat{\beta} = 0$
(0,0)
X


$\hat{\beta} > 0$
Y
(0,0)
X


$\hat{\beta} < 0$
Y
(0,0)
X


$\beta = 0$
Y
(0,0)
X

CI and Hypotheses Tests

$$\hat{\beta} = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sum X_i^2 - n\bar{X}^2} = \frac{\sum X_i Y_i - \bar{X}\sum Y_i}{\sum (X_i - \bar{X})^2}$$

- Can write

$$\hat{\beta} = \sum c_i Y_i$$

where

$$c_i = \frac{X_i - \bar{X}}{\sum_j (X_j - \bar{X})^2}$$

- Under model,

$$Y_i \sim N(\alpha + \beta X_i, \sigma^2)$$

- Thus

$$\hat{\beta} \sim N\left(\sum c_i(\alpha + \beta X_i), \sigma^2 \sum c_i^2\right)$$

CI and Hypotheses Tests

- Equivalently

$$PQ = \frac{\hat{\beta} - \beta}{\sqrt{\frac{\sigma^2}{\sum (X_i - \bar{X})^2}}} \sim N(0,1)$$

$$\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum_i (X_i - \bar{X})^2}\right)$$

- $(1-\alpha) * 100\%$ CI for β

$$\hat{\beta} \pm Z_{\alpha/2} \cdot \sqrt{\frac{\sigma^2}{\sum_i (X_i - \bar{X})^2}}$$

- Test for $H_0 : \beta = \beta_0$

$$Z = \frac{\hat{\beta} - \beta_0}{\sqrt{\frac{\sigma^2}{\sum (X_i - \bar{X})^2}}}$$

CI and Hypotheses Tests

- If σ^2 is unknown, use $S^2_{Y|X}$ (i.e MSE) and t_{n-2}
- $(1-\alpha)*100\%$ CI for β

$$\hat{\beta} \pm t_{n-2,\alpha/2} \cdot \sqrt{\frac{S^2_{Y|X}}{\sum (X_i - \bar{X})^2}}$$

- Test for $H_0 : \beta = \beta_0$

$$t = \frac{\hat{\beta} - \beta_0}{\sqrt{\frac{S^2_{Y|X}}{\sum (X_i - \bar{X})^2}}}$$

CI and Hypotheses Tests: SBP

$N = 9$ 2 sided test.

- For SBP example, $H_0 : \beta = 0$ versus $H_A : \beta \neq 0$, $\alpha = 0.05$

$$C_{0.05} = \{t : |t| > t_{7, 0.025} = 2.365\}.$$

- Observed test statistic implies reject H_0

$$t = \frac{0.45 - 0}{\sqrt{3.21/2417.56}} = 12.32.$$

- 95% CI for β

$$0.45 \pm 2.365\sqrt{\frac{3.21}{2417.56}}$$

$$= (0.363, 0.535)$$

$\hat{\beta} = 0.45$

$$S_{YX}^2 = \frac{\sum (Y_i - \hat{Y}_i)^2}{7} = 3.21$$

obs	Age	SP	SBP
1	122	$112.3 + 0.45(1)$	
	$= x_1$	$= \hat{Y}_1$	
10	145	$112.3 + 0.45(10)$	
	$= x_9$	$= \hat{Y}_9$	

CI and Hypotheses Tests

$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$

$E(\hat{\alpha}) = E(\bar{Y} - \hat{\beta}\bar{X})$

$= \alpha + \beta\bar{X} - \beta\bar{X} = \alpha.$

$Var(\hat{\alpha}) = Var(\bar{Y} - \hat{\beta}\bar{X})$

$= Var(\bar{Y}) + Var(\hat{\beta}\bar{X})$

$= \frac{\sigma^2}{N} + \frac{\bar{X}^2\sigma^2}{\sum(X_i - \bar{X})^2}$

- It can be shown that $\bar{Y}$ and $\hat{\beta}$ are independent

show $Cov(\bar{Y}, \hat{\beta}) = 0$

- Therefore

$$\hat{\alpha} \sim N\left(\alpha, \sigma^2\left\{\frac{1}{N} + \frac{\bar{X}^2}{\sum_i (X_i - \bar{X})^2}\right\}\right)$$

- $H_0 : \alpha = \alpha_0$

$$t = \frac{\hat{\alpha} - \alpha_0}{S_{YX}\sqrt{\frac{1}{N} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}}} \sim t_{N-2}$$

SAS Code

```
data sbpdat;
input age sbp @@;
datalines;
19 122 25 125 30 126 42 129 46 130 52 135 57 138 62 142 70 145
;
run;

proc reg data=sbpdat;
model sbp = age;
run;
```

SAS Output

The REG Procedure
Model: MODEL1
Dependent Variable: sbp

Number of Observations Read	9
Number of Observations Used	9

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	487.74667	487.74667	151.91	<.0001
Error	7	22.47555	3.21079		
Corrected Total	8	510.22222			

Root MSE	1.79187	R-Square	0.9559
Dependent Mean	132.44444	Adj R-Sq	0.9497
Coeff Var	1.35292		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	112.33169	1.73773	64.64	<.0001
age	1	0.44917	0.03644	12.33	<.0001

$= \hat{\beta}_2$

$= \hat{\beta}$

$\sqrt{\text{Var}(\hat{\beta})} = \sqrt{\frac{3.21}{247.56}}$

R Code and Output

```
> age <- c(19,25,30,42,46,52,57,62,70)
> sbp <- c(122,125,126,129,130,135,138,142,145)
> fit <- lm(sbp~age)
> summary(fit)

Call:
lm(formula = sbp ~ age)

Residuals:
    Min      1Q  Median      3Q     Max
-2.9934 -0.6884  0.1933  1.2265  1.8199

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 112.33169  1.73773   64.64 5.57e-11 ***
age           0.44917  0.03644   12.32 5.31e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.792 on 7 degrees of freedom
Multiple R-squared:  0.9559, Adjusted R-squared:  0.9497
F-statistic: 151.9 on 1 and 7 DF,  p-value: 5.313e-06
```

CI for $E(Y|X = x)$

- Goal: CI for the mean of Y given $X = x$
- Let $\mu_x = E(Y|X = x)$
- Estimator for μ_x:

$$\hat{\mu}_x = \hat{\alpha} + \hat{\beta} x$$

$$= \bar{Y} - \hat{\beta}\bar{x} + \hat{\beta} x$$

$$= \bar{Y} + \hat{\beta}(x - \bar{x})$$

- $E(\hat{\mu}_x) = \mu_x$

CI for $E(Y|X = x)$

- Recall $\bar{Y}$ and $\hat{\beta}$ are independent Normal RVs
- Thus $\hat{\mu}_x$ is Normal and

$$Var(\hat{\mu}_x) = Var(\bar{Y}) + (x - \bar{x})^2 Var(\hat{\beta})$$

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

$$= \sigma^2 \left[\frac{1}{N} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right].$$

CI for $E(Y|X = x)$

- Therefore, a $(1-\alpha)100\%$ CI for μ_x is (σ^2 unknown)

$$\hat{\mu}_x \pm t_{N-2,\alpha/2}\sqrt{s^2_{Y,X}\left[\frac{1}{N}+\frac{(X-\bar{X})^2}{\sum(X_i-\bar{X})^2}\right]}$$

- Note that $\widehat{Var}(\hat{\mu}_x)$ is a function of $X-\bar{X}$ so the farther X is from $\bar{X}$, the longer the CI will be.

Example: SBP and Age

$N = 9$

- Suppose we want a 95% CI of the mean SBP if age = 40

$$\hat{\mu}_{40} = 112.3 + 0.45(40) = 130.3$$

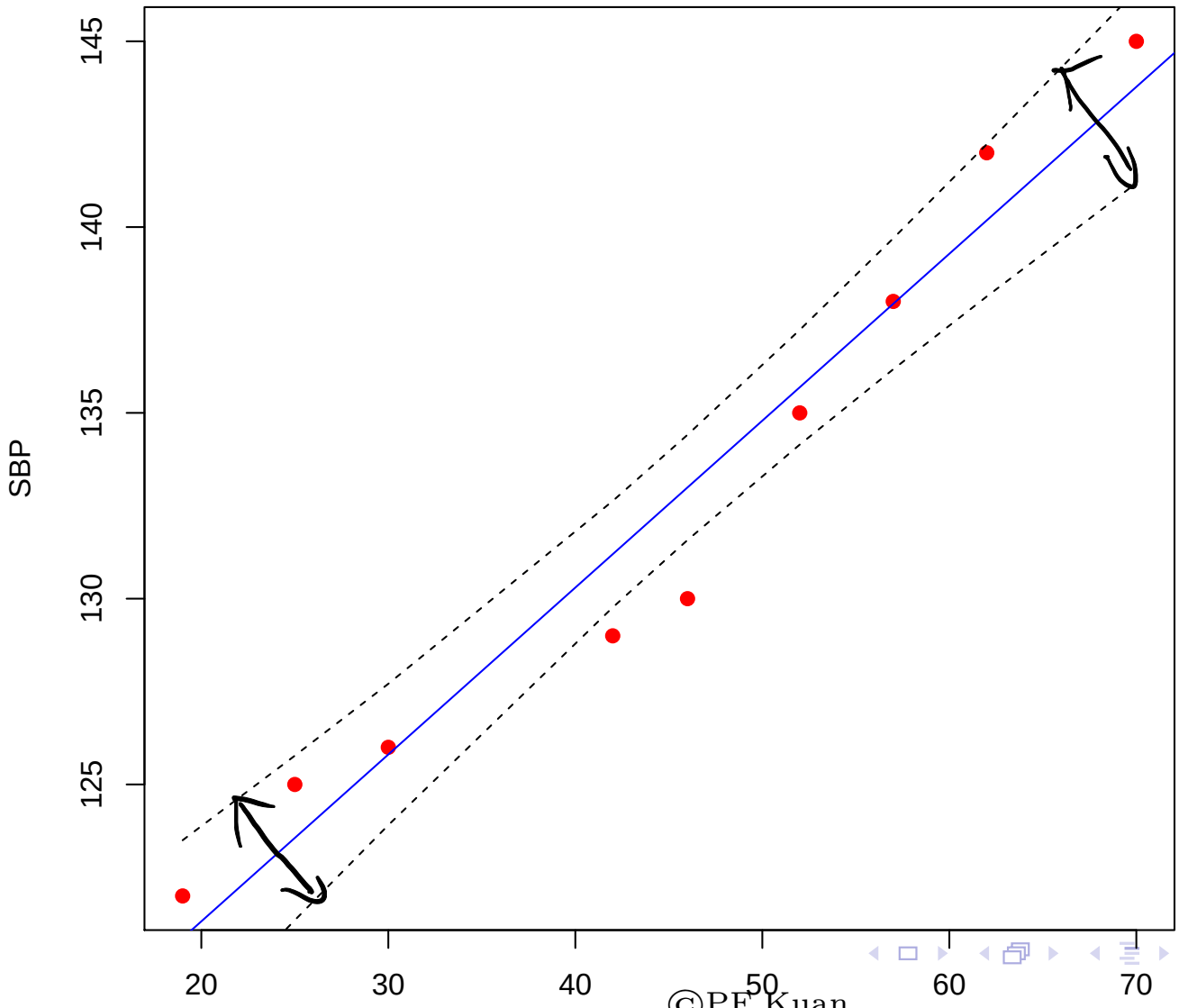
- CI

$t_{7,0.025} = 2.365$, $S_{\hat{Y}} = 1.79$.

$$130.3 \pm 2.365(1.79)\sqrt{\frac{1}{9} + \frac{(40-44.8)^2}{2417.59}}$$

$$(128.8, 131.8)$$

Example: SBP and Age



Prediction

- Want prediction interval (PI) for future observation given $X = x$

$$\hat{Y}_x = \hat{\alpha} + \hat{\beta} x$$

- Note: Y_x is a RV, so we consider the RV $Y_x - \hat{Y}_x$

$$E(Y_x - \hat{Y}_x) = \alpha + \beta x - E(\hat{Y}_x)$$

$$= \alpha + \beta x - \alpha - \beta x = 0.$$

$$V(Y_x - \hat{Y}_x) = \mathrm{Var}(Y_x) + \mathrm{Var}(\hat{Y}_x)$$

$$= \sigma^2 + \sigma^2 \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right]$$

$$= \sigma^2 \left[1 + \frac{1}{n} + \frac{(x - \bar{y})^2}{\sum (x_i - \bar{x})^2} \right]$$

Prediction

- Since the ϵ's are normally distributed, it follows

$$Y_x - \hat{Y}_x \sim N\left(0, \sigma^2\left\{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right\}\right)$$

- If σ^2 is not known,

$$\frac{Y_x - \hat{Y}_x}{S_{y,x}\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}}} \sim t_{n-2}$$

Prediction

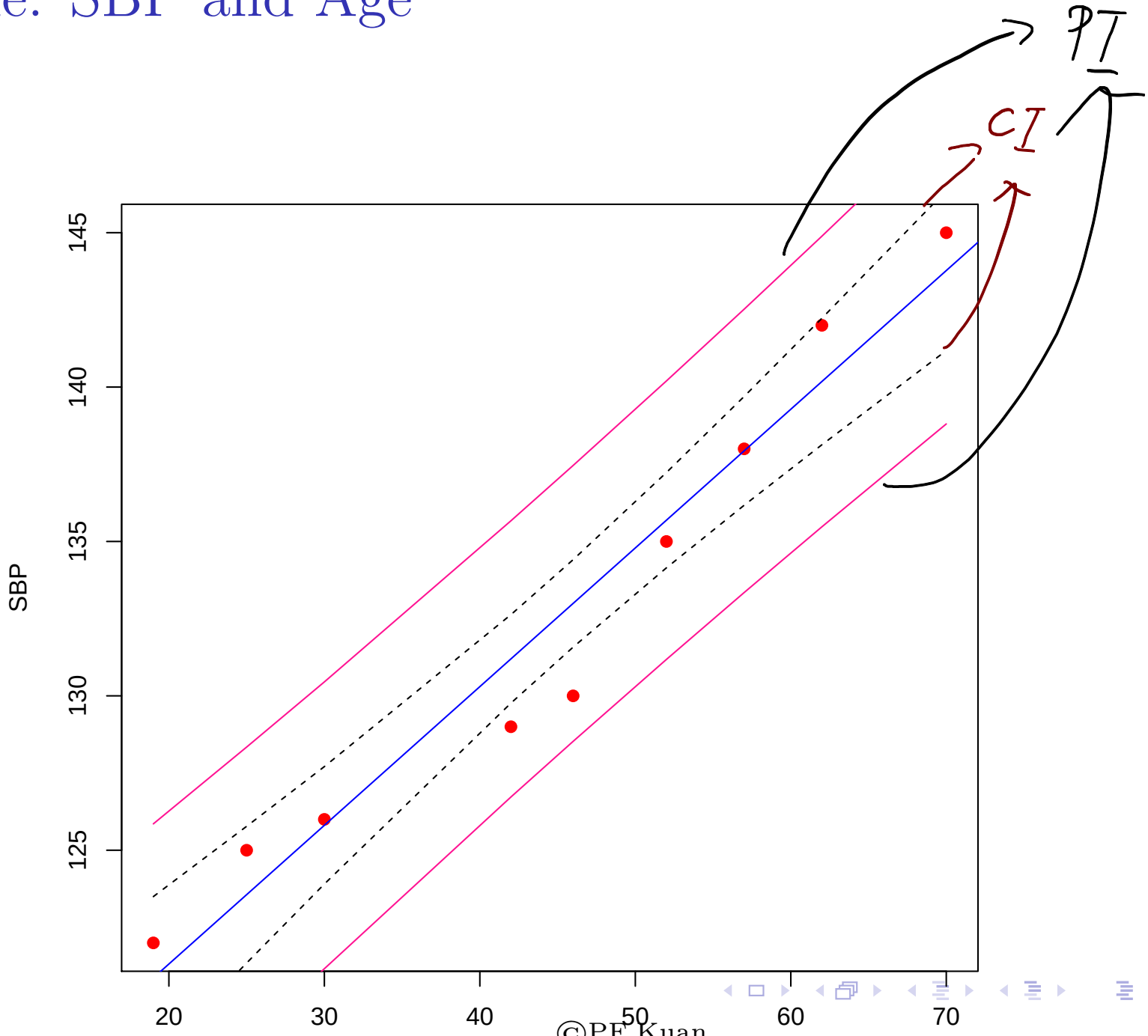
- $(1-\alpha)100\%$ PI for future observation at $X = x$

$$\hat{Y}_x \pm t_{N-2,1-\alpha/2} s_{y.x} \sqrt{1 + \frac{1}{N} + \frac{(x-\bar{X})^2}{\sum (X_i - \bar{X})^2}}$$

Prediction

- Suppose we want a 95% PI for an individual who is 40 years old:
- Point estimate:
- PI:

Example: SBP and Age



SAS Code

```
data sbpdat;
input age sbp @@;
datalines;
19 122 25 125 30 126 42 129 46 130 52 135 57 138 62 142 70 145 40 NA
;
run;

proc reg data=sbpdat;
model sbp = age;
output out=foo lcl=LCL lclm=LCLM p=P uclm=UCLM ucl=UCL;

proc print data=foo;
run;
```

SAS Output

The REG Procedure
Model: MODEL1
Dependent Variable: sbp

Number of Observations Read	10
Number of Observations Used	9
Number of Observations with Missing Values	1

$H_0: \beta = 0$ vs $H_a: \beta \neq 0$

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model (Regression)	1	487.74667	487.74667	151.91	<.0001
Error	7	22.47555	3.21079		
Corrected Total	8	510.22222			

Root MSE	1.79187	R-Square	0.9559
Dependent Mean	132.44444	Adj R-Sq	0.9497
Coeff Var	1.35292		

Parameter Estimates

Variable	DF	Parameter Estimate		Standard Error	t Value	Pr > |t|
Intercept	1	112.33169	$\hat{\alpha}$	1.73773	64.64	<.0001
age	1	0.44917	$\hat{\beta}$	0.03644	12.33	<.0001

SAS Output

$\hat{\mu}_x = \hat{E}(Y|X=x) = \hat{\gamma}_x = \hat{\alpha} + \hat{\beta}x.$

CI

PI.

The SAS System 23:41 Saturday, September 5, 2015 4

Obs	age	sbp	P	LCLM	UCLM	LCL	UCL
1	19	122	120.866	118.234	123.498	115.878	125.854
2	25	125	123.561	121.347	125.774	118.780	128.341
3	30	126	125.807	123.905	127.708	121.162	130.451
4	42	129	131.197	129.764	132.629	126.724	135.669
5	46	130	132.993	131.577	134.410	128.526	137.461
6	52	135	135.688	134.145	137.232	131.179	140.198
7	57	138	137.934	136.172	139.696	133.345	142.523
8	62	142	140.180	138.131	142.229	135.474	144.887
9	70	145	143.773	141.181	146.366	138.806	148.741
10	40	.	130.298	128.827	131.770	125.813	134.784

R Code and Output

> summary(fit).

```
> fit <- lm(sbp~age)
> predict(fit,data.frame(age=40),interval='confidence')
       fit      lwr      upr
1 130.2984 128.8273 131.7696
> predict(fit,data.frame(age=40),interval='prediction')
       fit      lwr      upr
1 130.2984 125.8132 134.7836
```

Sum of Squares Decomposition

- Can decompose total sum of squares

$$\sum_i (Y_i - \bar{Y})^2 = \sum_i (\hat{Y}_i - \bar{Y})^2 + \sum_i (Y_i - \hat{Y}_i)^2$$

$$SST = SSR + SSE$$

- Total sample variance of the Y's:

$$s_y^2 = \frac{SST}{N-1} = \frac{\sum_i (Y_i - \bar{Y})^2}{N-1}.$$

(Unadjusted) r^2

- The unadjusted r^2 is given by

$$r^2 = \frac{SSR}{SST}$$

- *Coefficient of determination*
- Proportion of total variation attributable to regression
- SBP Example

$$r^2 = \frac{487.75}{510.22} = 0.9559$$

Adjusted r^2

$r_a^2 = 1 - \frac{N-2}{N-1}(1-r^2)$

$r_a^2 \le r^2$. $r_a \approx r$ when $N \to \infty$

- Note that the sample variance of the Y's is $s_y^2 = 63.78$ while $s_{y.x}^2 = 3.21$
- Thus X "explains"

In general, $r_a^2 = 1 - \frac{SSE/N-k-1}{SST/N-1}$.

$$\frac{63.78 - 3.21}{63.78} = 0.9497$$

proportion of the variance in Y.

- This quantity is called the *adjusted* r^2

$S_{Y,X}^2 = \frac{\sum(Y_i - \hat{Y}_i)^2}{N-k-1}$ $r_a^2 = \frac{s_y^2 - s_{y.x}^2}{s_y^2} = 1 - \frac{S_{Y,X}^2}{S_Y^2} = 1 - \frac{SSE/N-2}{SST/N-1}$

k: number of predictor.

$= 1 - MSE/MST$.

$Y = \alpha + \beta_1 X_1 + \varepsilon$. $Y = \alpha_2 + \beta_3 X_1 + \beta_3 X_2 + \varepsilon$

Unadjusted r^2

Information criterion for model comparison selection

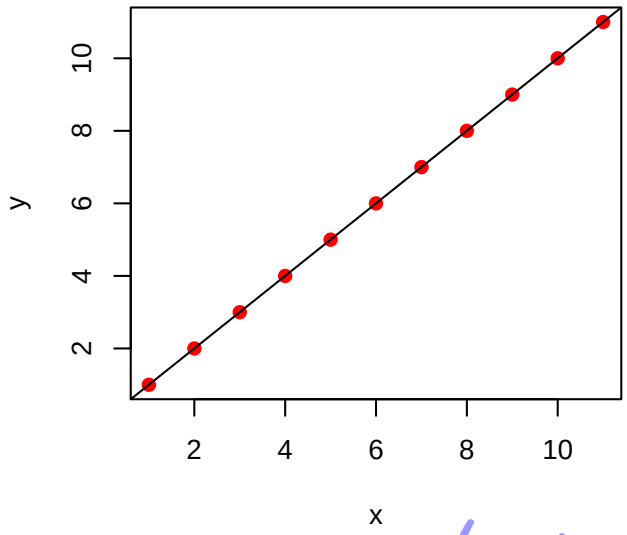
K←AIC. Akaike's IC.

N.K←BIC: Bayesian IC

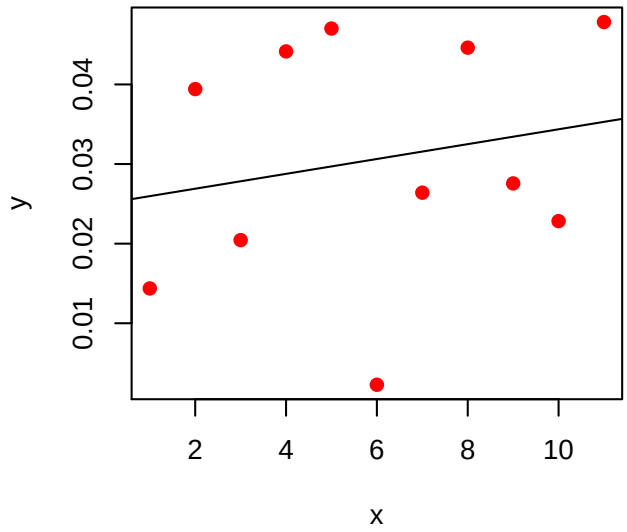
- Proportion of total variation attributable to regression
- Degree of linear association
- Ranges between 0 and 1
- $r^2 = 0 \rightarrow$ no linear association between X & Y.
- $r^2 = 1 \rightarrow$ perfect fit

Examples of r^2

1

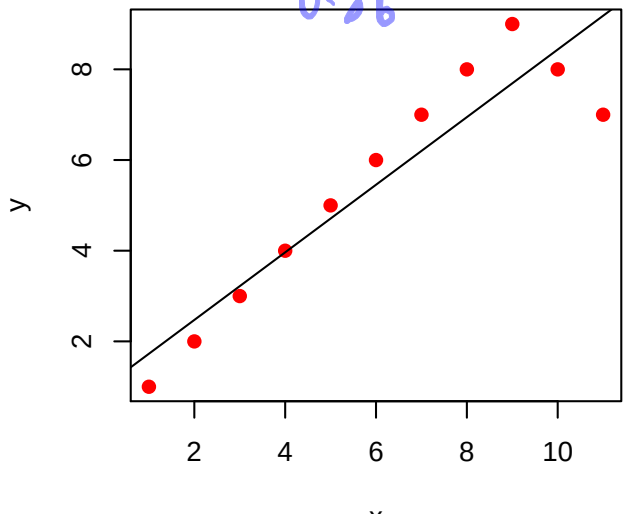


0.64



piecewise regression / splines.

0.86



0