AMS 572 Data Analysis I
Simple Linear Regression

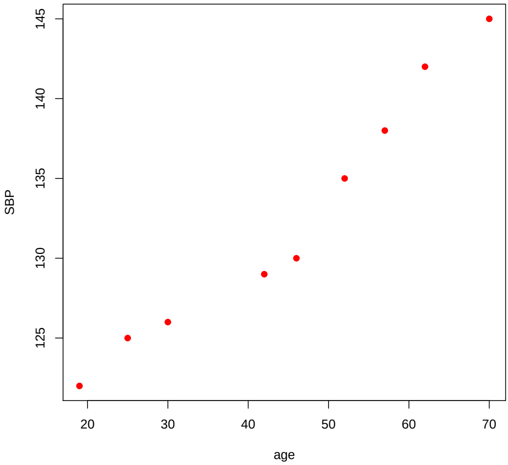
Pei-Fen Kuan

Applied Math and Stats, Stony Brook University

Example: SBP and Age

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

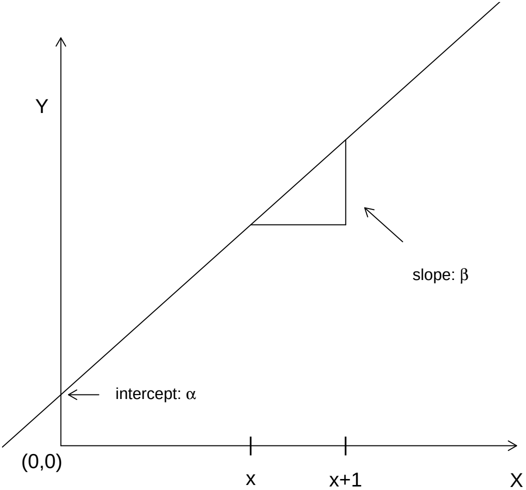
Example: SBP and Age



Simple Linear Model

- Line
$$Y = \alpha + \beta X$$
- α = intercept; value of Y when $X = 0$
- β = slope; change in Y when X changes 1 unit
- Y : dependent variable, response variable, outcome variable
- X : covariate, independent variable, predictor

Simple Linear Model



Simple Linear Model with Error

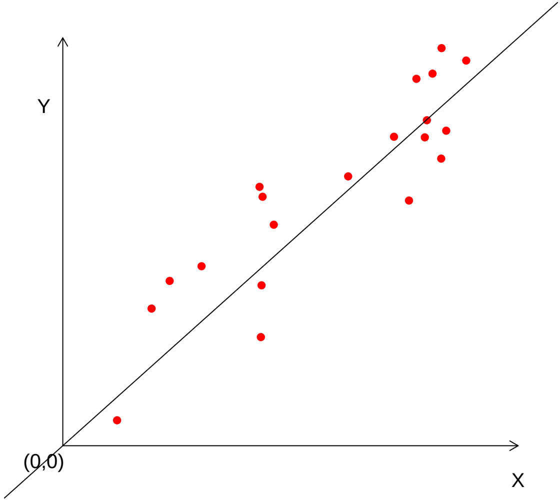
- Linear regression

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

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

- 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 $N(0, \sigma^2)$

Least Squares Estimation

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

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

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

Least Squares Estimation

- Solving the partial derivatives, we get

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

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

Least Squares Estimation

- 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)

$$\begin{aligned} \hat{\sigma}^2 = \text{MSE} &= \frac{1}{N-2} \sum_i (Y_i - \hat{Y}_i)^2 \\ &= \frac{1}{N-2} \sum_i r_i^2 \end{aligned}$$

Example: SBP and Age

$$\bar{Y} = 132.4; \bar{X} = 44.8$$

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

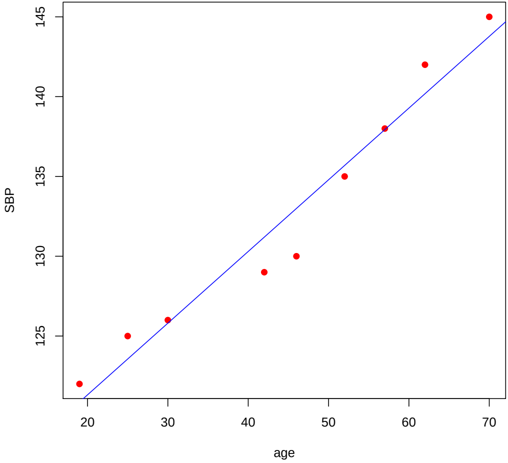
$$\hat{\beta} = \frac{54461 - 9(132.4)(44.8)}{20463 - 9(44.8)^2} = 0.45$$

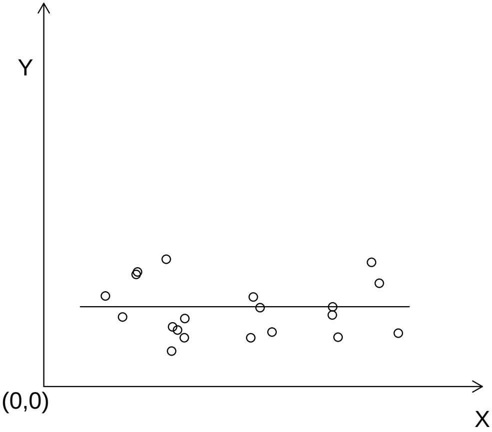
$$\hat{\alpha} = 132.4 - .45(44.8) = 112.3$$

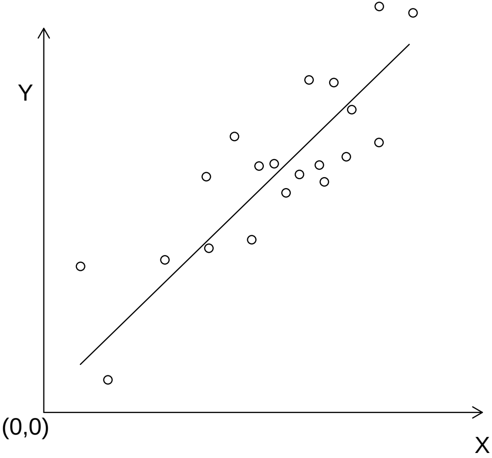
Example: Interpretation

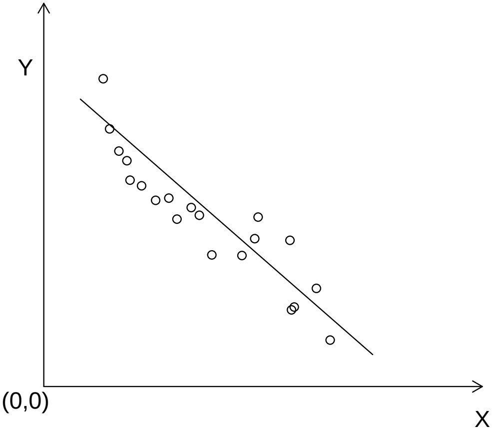
- $\hat{\beta} = 0.45 \rightarrow$ expected SBP increases 0.45 (mmHg) for each one year increase in age
- How about $\hat{\alpha} = 112.3$?

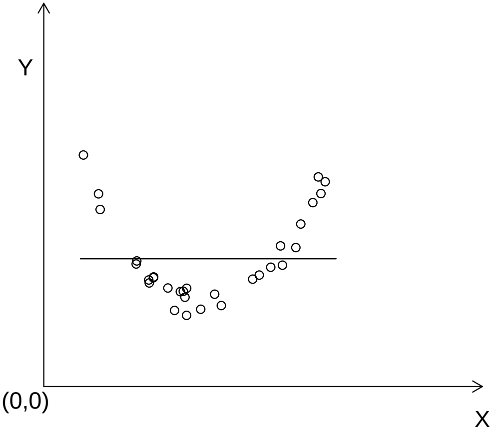
Example: SBP and Age




Y
(0,0)
X


Y
(0,0)
X


Y
(0,0)
X


Y
(0,0)
X

CI and Hypotheses Tests

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

CI and Hypotheses Tests

- Equivalently

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

- $(1-\alpha) * 100\%$ CI for β

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

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

$$z = \frac{\hat{\beta} - \beta_0}{\sqrt{\sigma^2 / \sum_i (X_i - \bar{X})^2}}$$

CI and Hypotheses Tests

- If σ^2 is unknown, use MSE and t_{N-2}
- $(1-\alpha)*100\%$ CI for β

$$\hat{\beta} \pm t_{N-2,\alpha/2}\sqrt{\text{MSE}/\sum_i (X_i - \bar{X})^2}$$

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

$$t = \frac{\hat{\beta} - \beta_0}{\sqrt{\text{MSE}/\sum_i (X_i - \bar{X})^2}}$$

CI and Hypotheses Tests: SBP

- For SBP example, $H_0 : \beta = 0$ versus $H_a : \beta \neq 0$

$$C_{.05} = \{t : |t| > t_{7,0.025} = 2.365\}$$

- Observed test statistic implies reject H_0

$$t = \frac{0.449 - 0}{\sqrt{3.21/2417.56}} = 12.32$$

- 95% CI

CI and Hypotheses Tests

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

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

- $H_0 : \alpha = \alpha_0$

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

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

                                 Sum of          Mean
Source                 DF       Squares        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

                       Parameter       Standard
     Variable    DF     Estimate          Error    t Value    Pr > |t|

     Intercept    1    112.33169        1.73773      64.64      <.0001
     age          1      0.44917        0.03644      12.33      <.0001
```

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:

$$\begin{aligned} \hat{\mu}_x &= \hat{\alpha} + \hat{\beta} x \\ &= \bar{Y} + \hat{\beta}(x - \bar{X}) \end{aligned}$$

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

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

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

$$\begin{aligned} Var(\hat{\mu}_x) &= Var(\bar{Y}) + (x - \bar{X})^2 Var(\hat{\beta}) \\ &= \sigma^2 \left[\frac{1}{N} + \frac{(x-\bar{X})^2}{\sum_i (X_i - \bar{X})^2} \right] \end{aligned}$$

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

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

$$\hat{\mu}_x \pm t_{N-2,\alpha/2}\sqrt{\text{MSE}\left\{\frac{1}{N} + \frac{(x-\bar{X})^2}{\sum_i (X_i - \bar{X})^2}\right\}}$$

Example: SBP and Age

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

- CI

Example: SBP and Age