BIOS 662 Fall 2018

Linear Regression, Part I

David Couper, Ph.D.

david_couper@unc.edu

or

couper@bios.unc.edu

https://sakai.unc.edu/portal

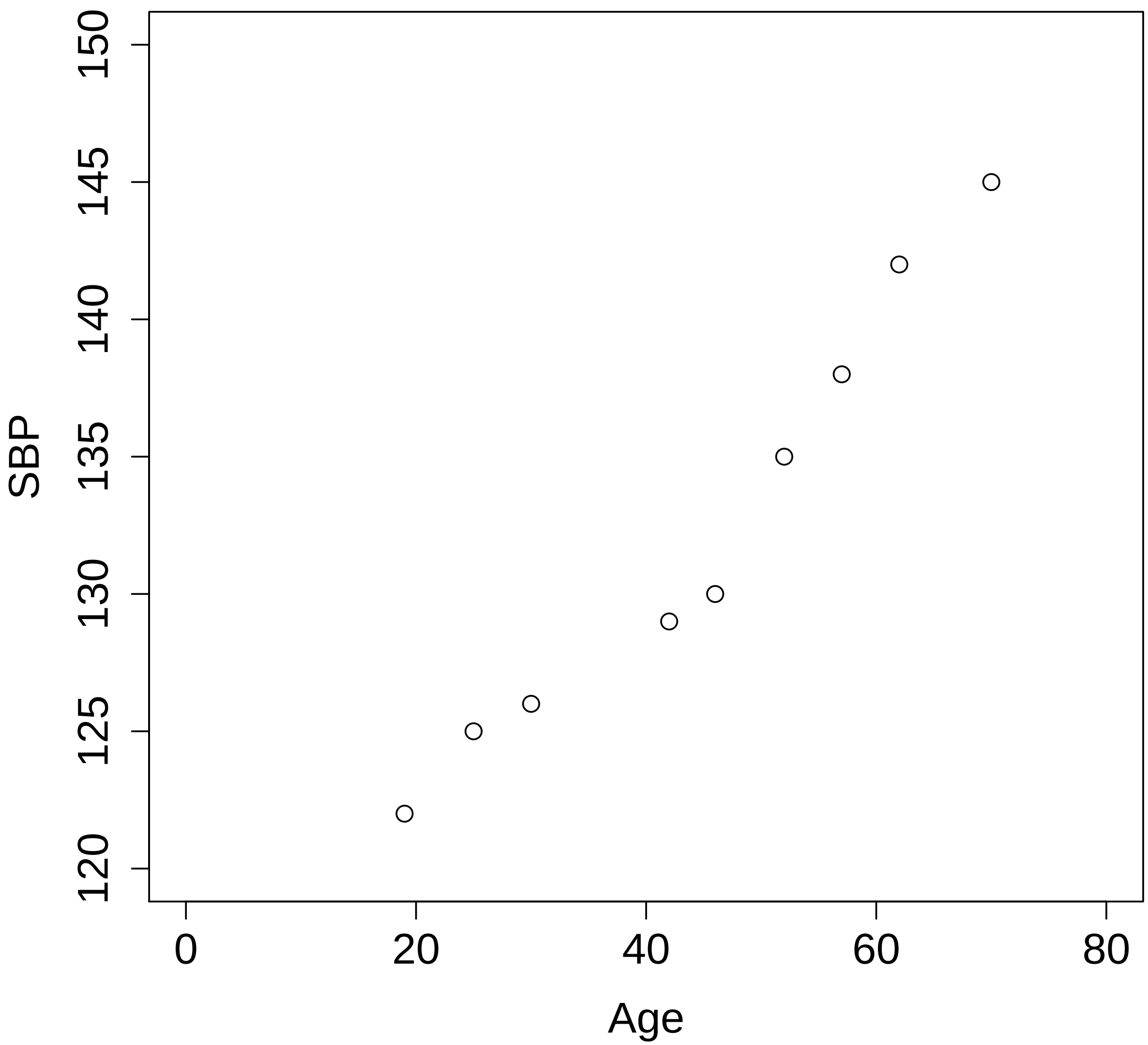
Outline

- Introduction: Assumptions, least-squares estimation
- Confidence intervals and hypothesis testing for regression coefficients
- Confidence interval for mean
- Prediction intervals
- r^2

Example: Systolic Blood Pressure 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 cont.



Simple Linear Model

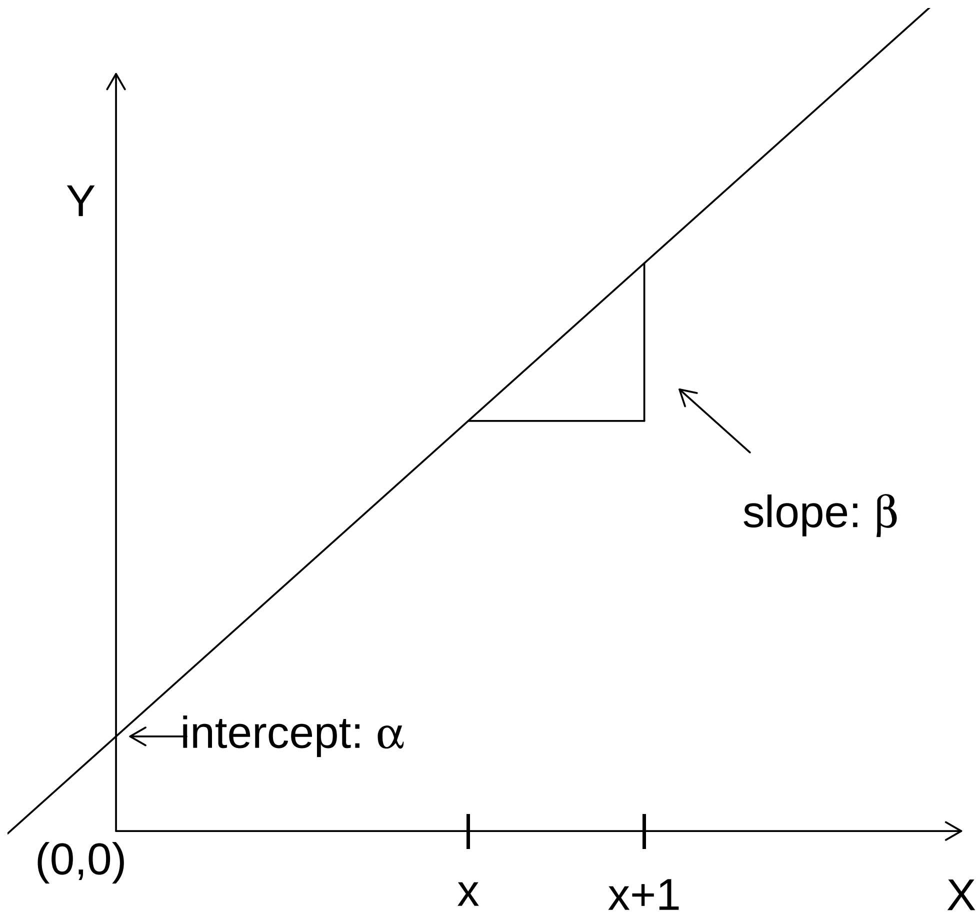
- Line

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

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

- Y dependent variable; response variable
- X independent variable; predictor; covariate

Simple Linear Model



Simple Linear Model with Error

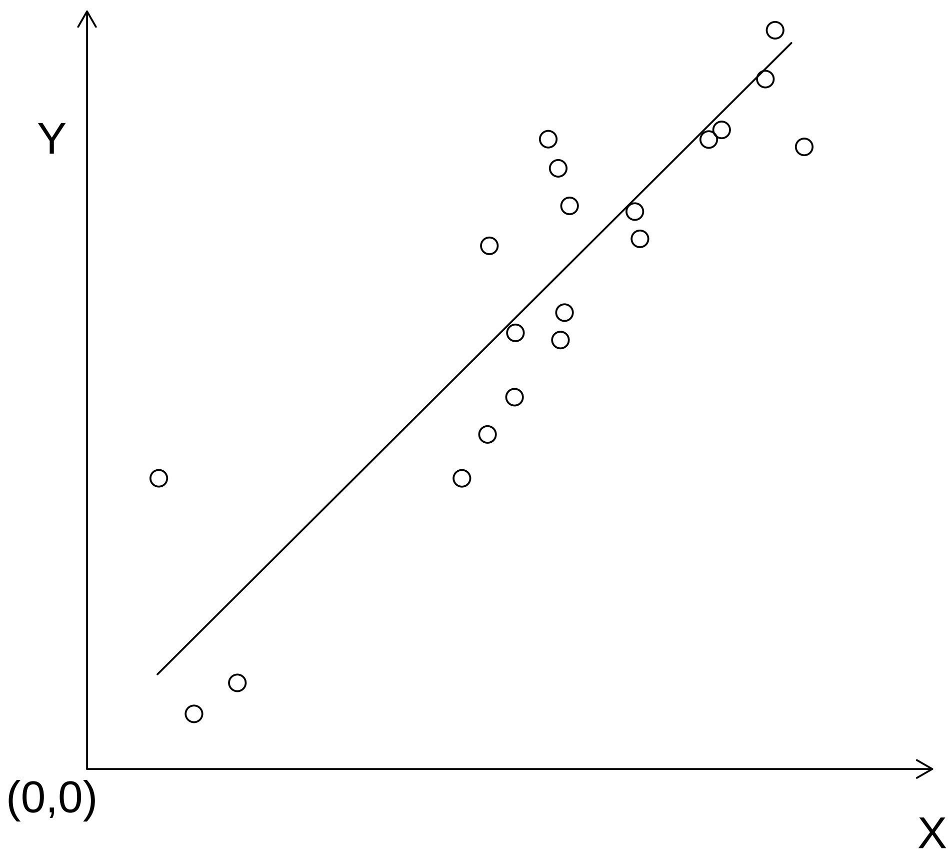
- Linear regression

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

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

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

Simple Linear Model with Error



Model Assumptions

- Data are (Y_i, X_i); $i = 1, 2, \ldots, N$
- Assumptions:

  1. Linearity: $Y_i = \alpha + \beta X_i + \epsilon_i$
  2. Xs 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

- For α:

$$\frac{\partial \sum_i \epsilon_i^2}{\partial \alpha} = -2\sum_i (Y_i - \alpha - \beta X_i)$$

$$= -2N\bar{Y} + 2N\alpha + 2N\beta\bar{X}$$

- For β:

$$\frac{\partial \sum_i \epsilon_i^2}{\partial \beta} = -2\sum_i (Y_i - \alpha - \beta X_i) X_i$$

$$= -2\sum_i X_i Y_i + 2\alpha \sum_i X_i + 2\beta \sum_i X_i^2$$

Least Squares Estimation

- Two equations with two unknowns

$$-2N\bar{Y} + 2N\alpha + 2N\beta\bar{X} = 0 \tag{1}$$

$$-2\sum_i X_i Y_i + 2\alpha \sum_i X_i + 2\beta \sum_i X_i^2 = 0 \tag{2}$$

- From (1)

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

Least Squares Estimation

- Substituting into (2)

$$-\sum_i X_i Y_i + (\bar{Y} - \hat{\beta}\bar{X}) \sum_i X_i + \hat{\beta} \sum_i X_i^2 = 0,$$

implying

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

- Therefore

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

Least Squares Estimation

- Equivalent form:

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

  where

$$[XY] = \sum_i (X_i - \bar{X})(Y_i - \bar{Y})$$

$$[X^2] = \sum_i (X_i - \bar{X})^2$$

- Note that if $X_i = Y_i$ for all i, then $\hat{\beta} = 1$ as one would expect

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

Least Squares Estimation

- Predicted response (also known as *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)

$$\begin{aligned} \hat{\sigma}^2 = s^2_{y \cdot x} &= \frac{1}{N-2} \sum_i (Y_i - \hat{Y}_i)^2 \\ &= \frac{1}{N-2} \sum_i r_i^2 \end{aligned}$$

Example: SBP and Age

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

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

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

$$\hat{\alpha} = 132.4 - 0.45(44.8) = 112.3$$

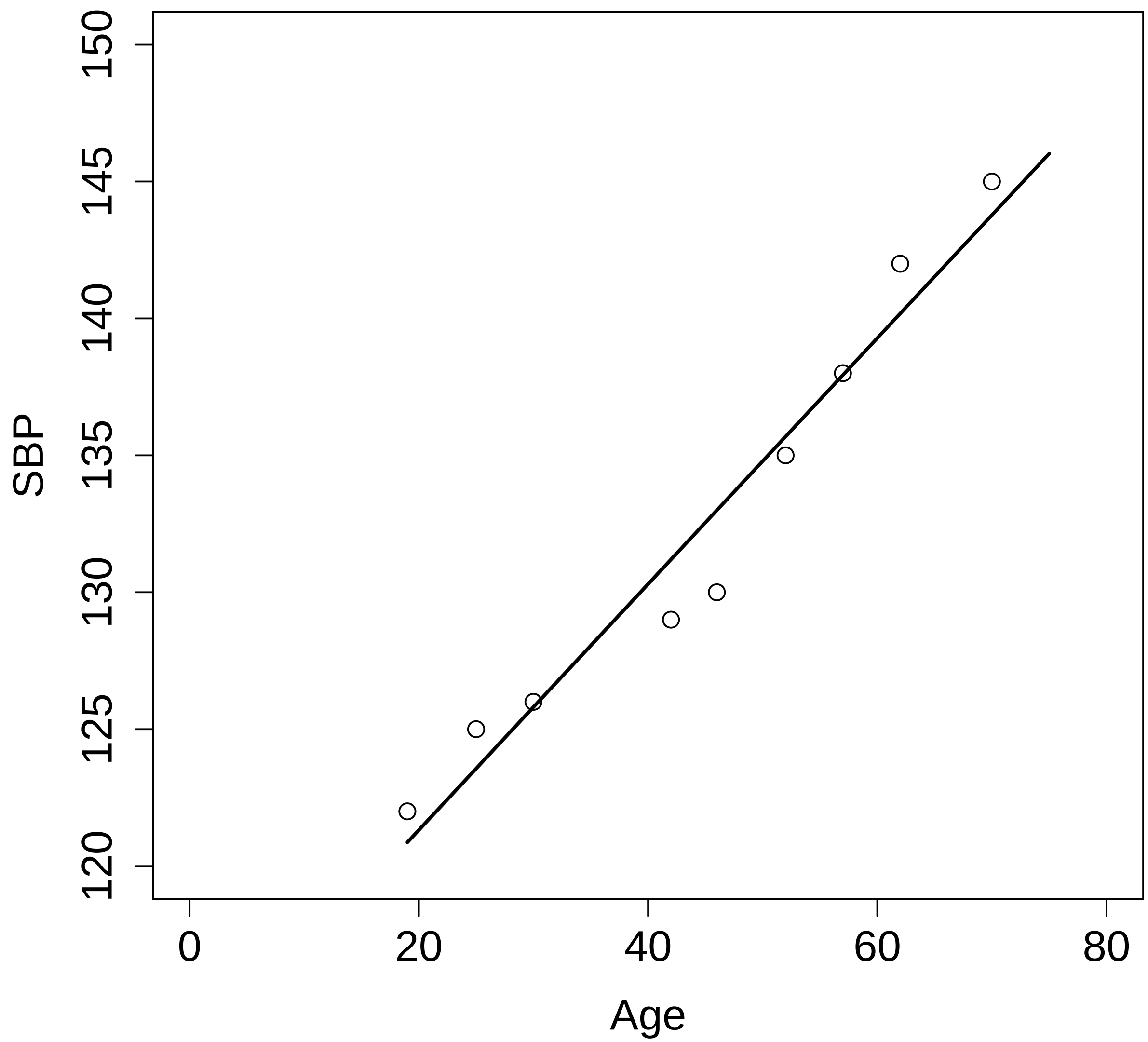
$$\hat{Y}_i = 112.3 + 0.45X_i$$

$$s^2_{y \cdot x} = 3.21$$

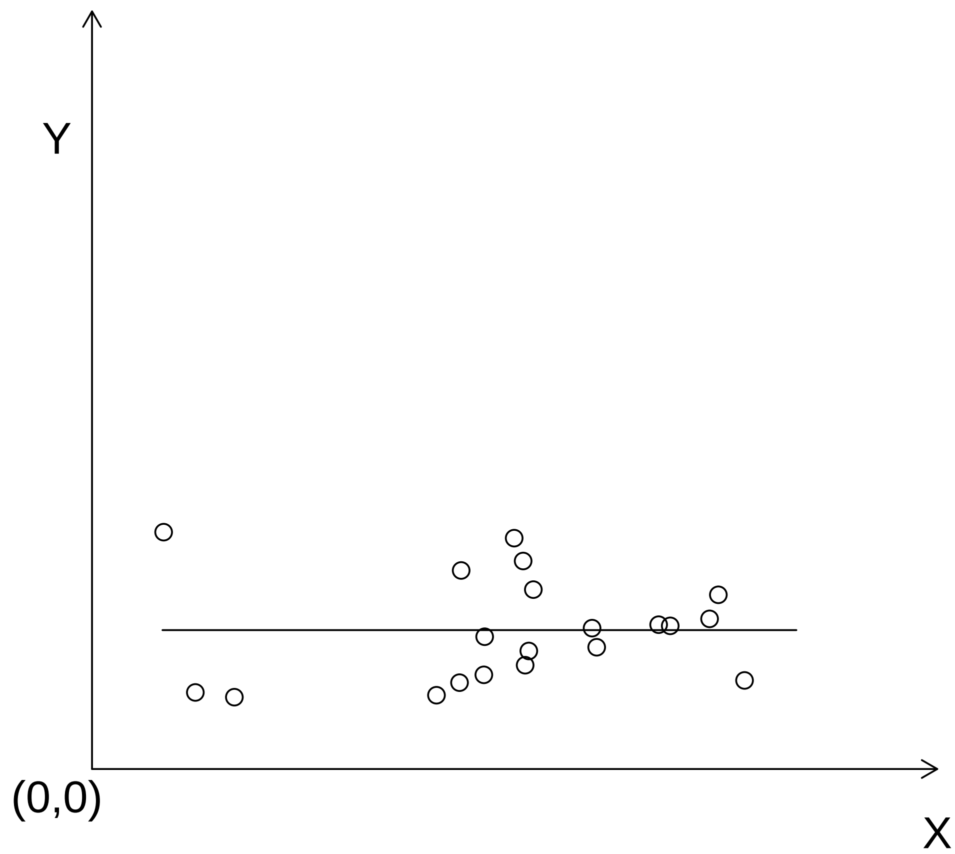
Example: Interpretation

- $\hat{\beta} = 0.45 \Rightarrow$ expected SBP increases 0.45 (mmHg) for each one year increase in age
- $\hat{\alpha} = 112.3 \Rightarrow$? Beware extrapolation (see section 9.4.3 of the text)

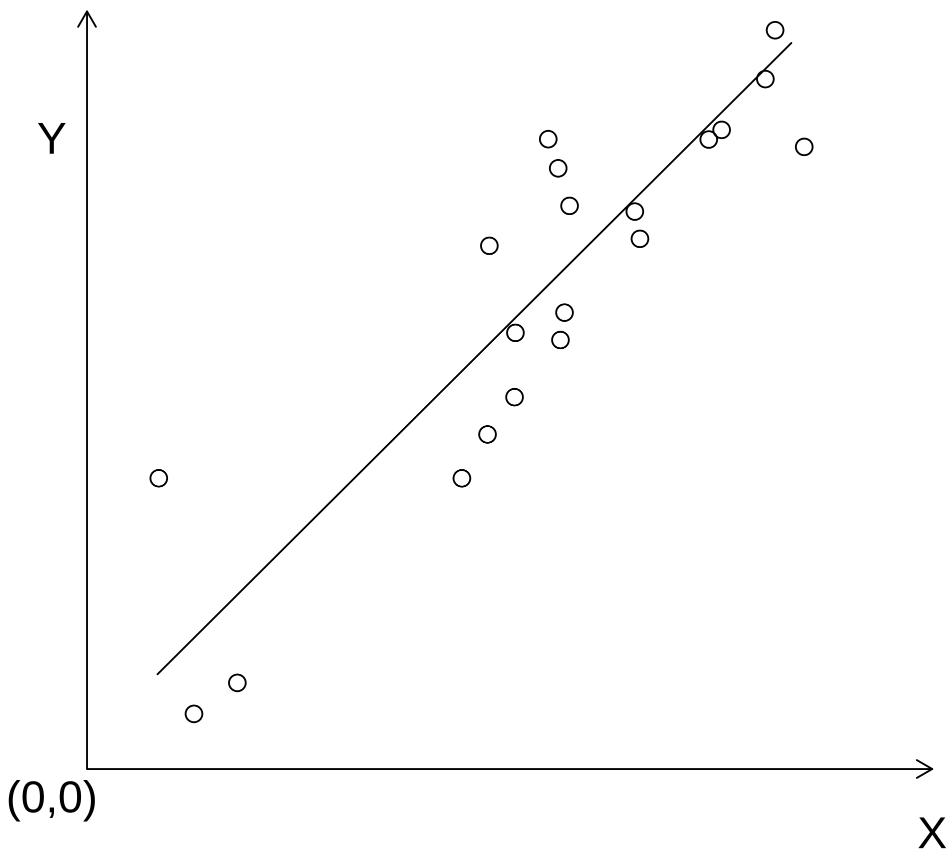
Example: SBP and Age cont.



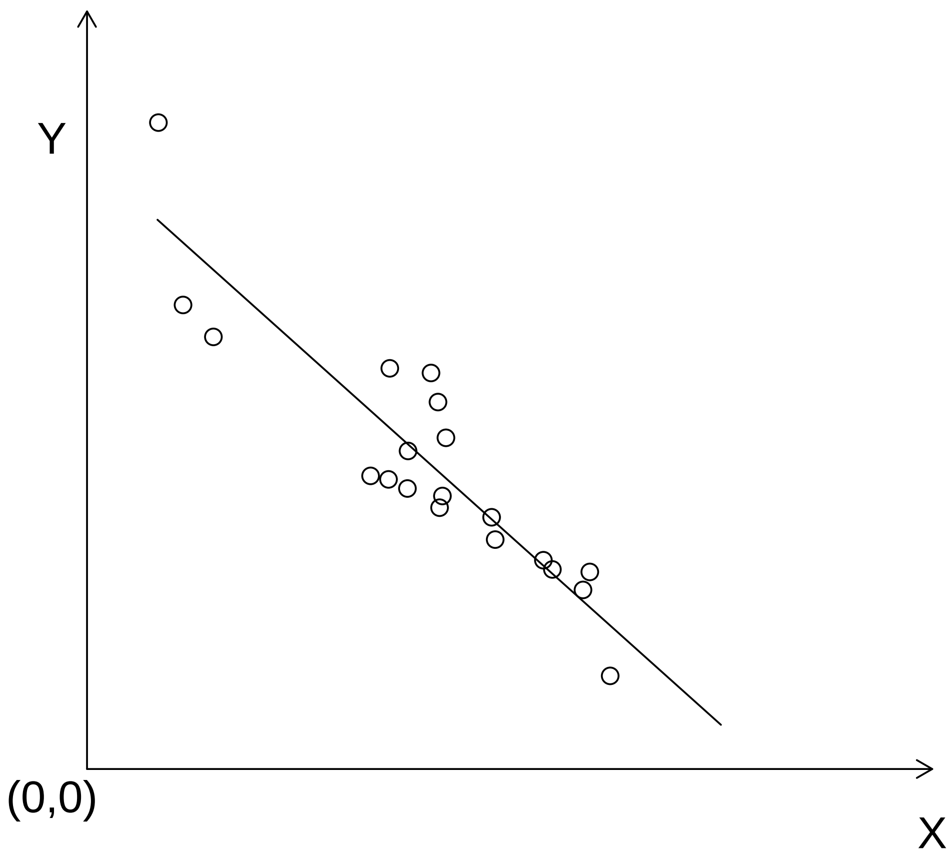
$$\hat{\boldsymbol{\beta}} = \mathbf{0}$$



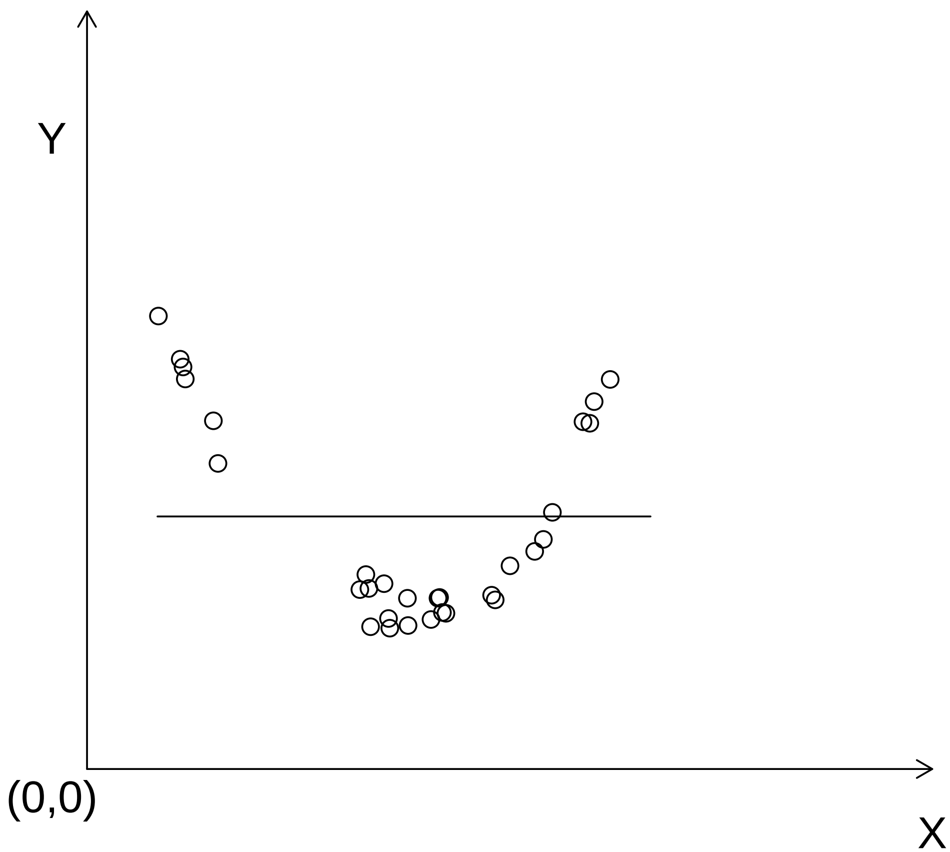
$$\hat{\boldsymbol{\beta}} > \mathbf{0}$$



$\hat{\boldsymbol{\beta}} < \mathbf{0}$



$$\hat{\boldsymbol{\beta}} = \mathbf{0}$$



Confidence Intervals 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 the model,

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

- Thus

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

Confidence Intervals and Hypotheses Tests

- Equivalently

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

- $100(1-\alpha)\%$ CI for β

$$\hat{\beta} \pm z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{\sum_i (X_i - \bar{X})^2}}$$

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

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

Confidence Intervals and Hypotheses Tests

- If σ^2 is unknown, use $s^2_{y \cdot x}$ and t_{N-2}
- $100(1-\alpha)\%$ CI for β

$$\hat{\beta} \pm t_{N-2,1-\alpha/2} \sqrt{s^2_{y \cdot x} / \sum_i (X_i - \bar{X})^2}$$

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

$$t = \frac{\hat{\beta} - \beta_0}{\sqrt{s^2_{y \cdot x} / \sum_i (X_i - \bar{X})^2}}$$

Confidence Intervals and Hypotheses Tests: SBP

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

$$C_{0.05} = \{t : |t| > t_{7,0.975} = 2.365\}$$

- Observed test statistic implies reject H_0

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

- 95% CI

$$0.449 \pm 2.365\sqrt{3.21/2417.56} = (0.363, 0.535)$$

Confidence Intervals and Hypotheses Tests

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

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

- $H_0 : \alpha = \alpha_0$

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

SBP Example in R

```
> 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
```

SBP Example in SAS

```
proc reg;
   model sbp=age;
```

```
The REG Procedure
Model: MODEL1
Dependent Variable: sbp

                              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
```

Confidence Interval 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:

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

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

Confidence Interval for $E(Y|X = x)$

- Recall that $\bar{Y}$ and $\hat{\beta}$ are independent normally distributed random vaiables
- Thus $\hat{\mu}_x$ is normally distributed and

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

Confidence Interval for $\boldsymbol{E(Y|X = x)}$

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

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

- Note that $\text{Var}(\hat{\mu}_x)$ is a function of $x - \bar{X}$
- So, the further x is from $\bar{X}$, the wider the CI will be
- Design considerations: Note 9.3 in the text

Example: SBP and Age

- Suppose we want a 95% CI for the mean SBP when age = 40

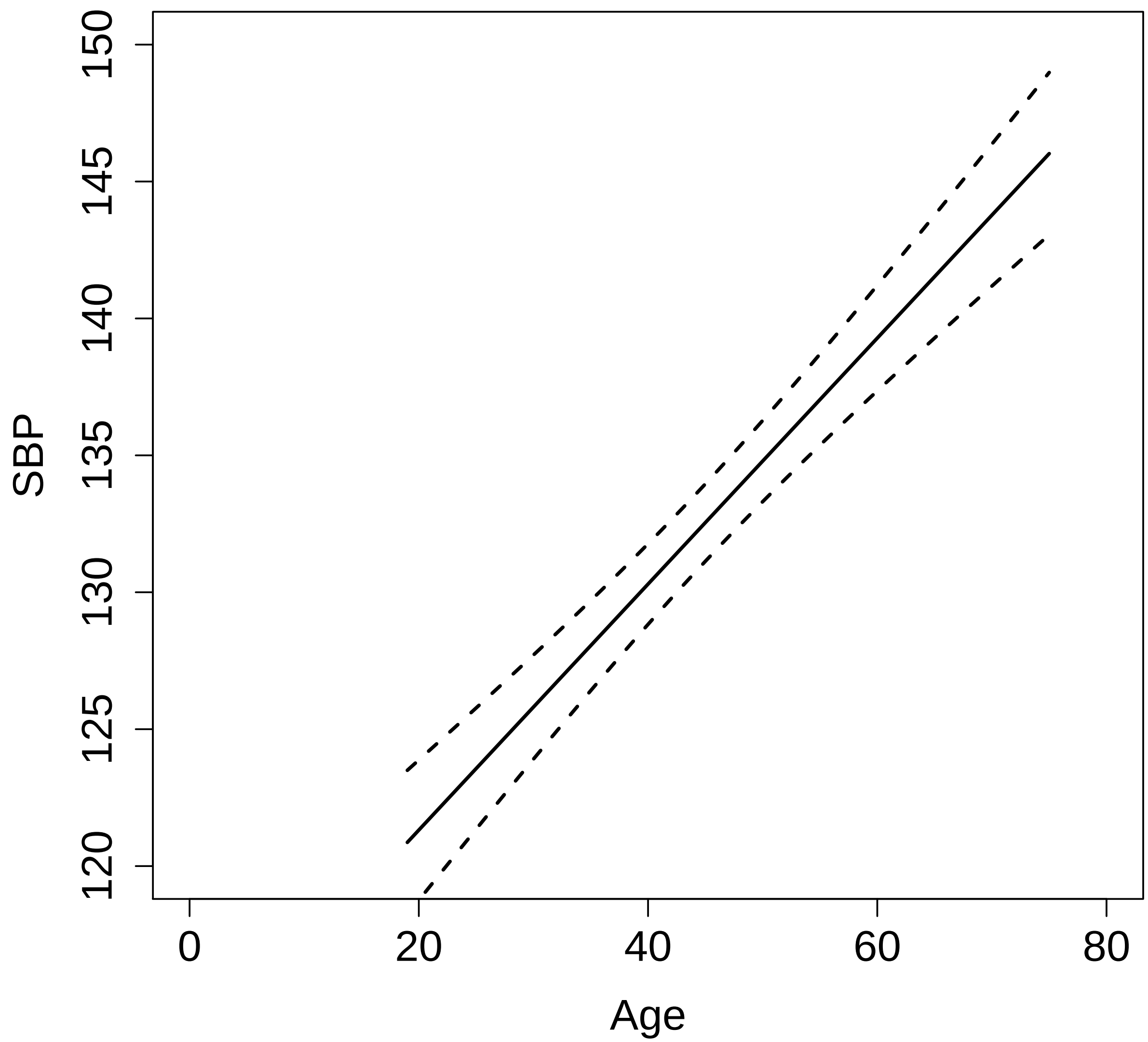
$$\hat{\mu}_{40} = 112.3 + 0.45(40) = 130.3$$

- Confidence interval:

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

$$(128.8, 131.8)$$

Example: SBP and Age cont.



Confidence Interval for $E(Y|X = x)$

- These "bands" should be interpreted in a pointwise fashion only
- The text's usage of the term "bands" is non-standard (p. 303-4)
- Usual interpretation of *confidence band*: covers the entire regression line with $100(1-\alpha)\%$ confidence
- Cf. Section 2.6 of *Applied Linear Statistical Models*, Neter et al., 4th edition, 1996

Prediction

- Suppose we want a prediction interval (PI) for a new or future observation, given $X = x$

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

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

$$E(Y_x - \hat{Y}_x) = \alpha + \beta x - (\alpha + \beta x) = 0$$

$$\mathrm{Var}(Y_x - \hat{Y}_x) = \mathrm{Var}(Y_x) + \mathrm{Var}(\hat{Y}_x) - 2\mathrm{Cov}(Y_x, \hat{Y}_x)$$

Prediction

- Because Y_x is not part of the sample, Y_x and $\hat{Y}_x$ are independent
- Therefore

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

Prediction

- Because ϵ is normally distributed, it follows that

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

- If σ^2 is not known,

$$\frac{Y_x - \hat{Y}_x}{s_{y\cdot x}\sqrt{1 + \frac{1}{N} + \frac{(x-\bar{X})^2}{\sum(X_i-\bar{X})^2}}} \sim t_{N-2}$$

Prediction

- $100(1-\alpha)\%$ prediction interval for a new or future observation at $X = x$

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

- Cf. Section 5-10 of *Applied Regression Analysis and Multivariable Methods*, Kleinbaum et al., 3$^{\text{rd}}$ edition, 1998

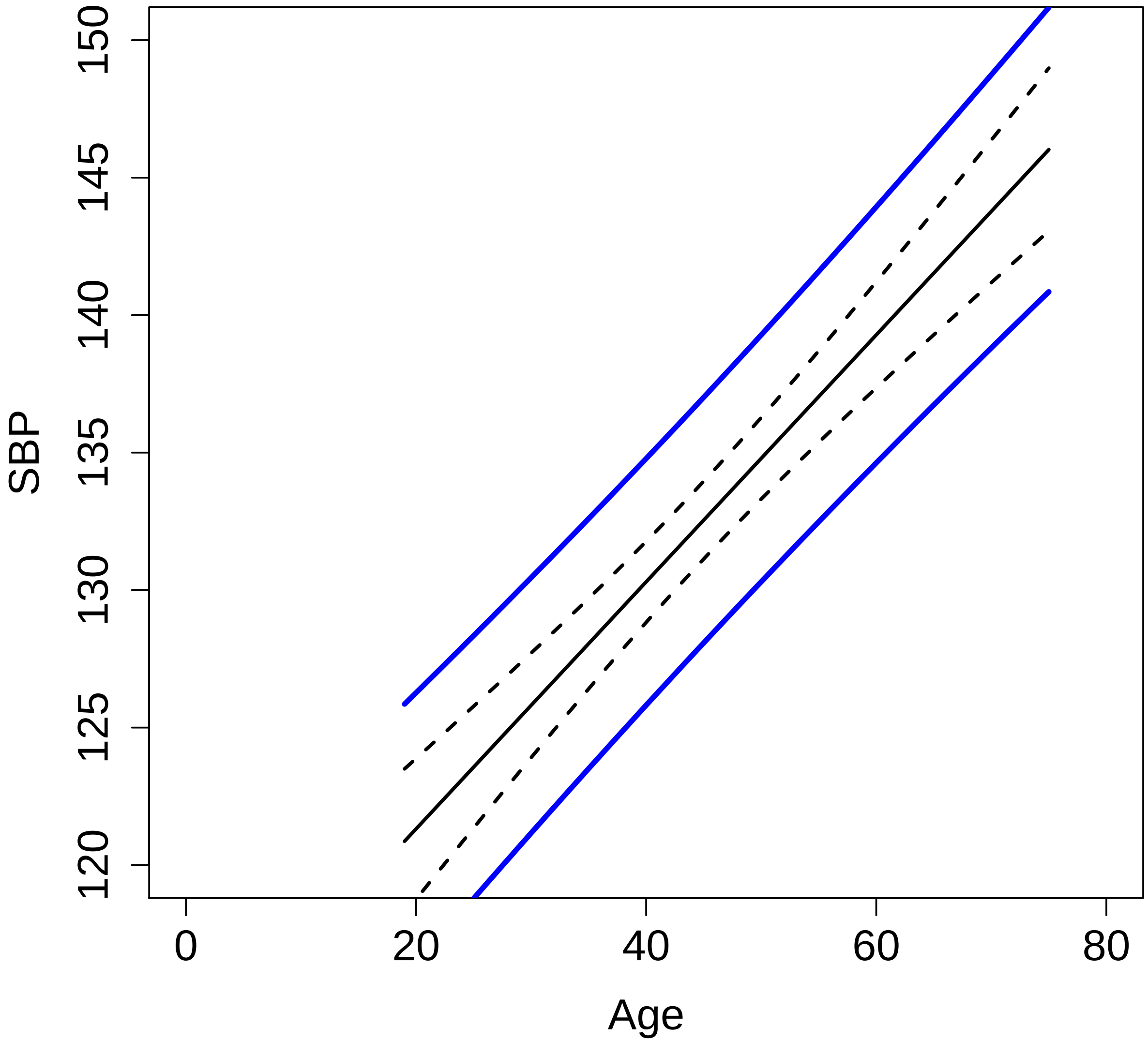
Prediction

- Suppose we want a 95% prediction interval for an individual who is 40 years old
- Point estimate: $\hat{Y}_{40} = 130.3$
- Prediction interval:

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

$$(125.8, 134.8)$$

Example: SBP vs Age



SBP Example in R

```
> 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
```

SBP Example in SAS

- In the input dataset add an observation with age = 40 and missing SBP

```
proc reg;
   model sbp=age;
   output out=ci lcl=LCL lclm=LCLM p=P uclm=UCLM ucl=UCL;

proc print data=ci;
```

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

Sum of Squares Decomposition

- We can decompose the total sum of squares

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

$$\text{SST} = \text{SSR} + \text{SSE}$$

- Total sample variance of the Ys:

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

Unadjusted r^2

- The unadjusted r^2 is given by

$$r^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

- r^2 is called the *coefficient of determination*
- Proportion of total variation attributable to regression
- SBP example:

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

Adjusted r^2

- Note that the sample variance of the Ys is $s_y^2 = 63.78$ while $s_{y\cdot x}^2 = 3.21$

- Thus X “explains” the proportion

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

of the variance of Y

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

$$r_a^2 = \frac{s_y^2 - s_{y\cdot x}^2}{s_y^2} = 1 - \frac{s_{y\cdot x}^2}{s_y^2} = 1 - \frac{\mathrm{SSE}/(N-2)}{\mathrm{SST}/(N-1)}$$

Adjusted and Unadjusted r^2

- Note that

$$r_a^2 = 1 - \frac{\text{SSE}/(N-2)}{\text{SST}/(N-1)}$$

and

$$r^2 = 1 - \frac{\text{SSE}}{\text{SST}}$$

- Implying

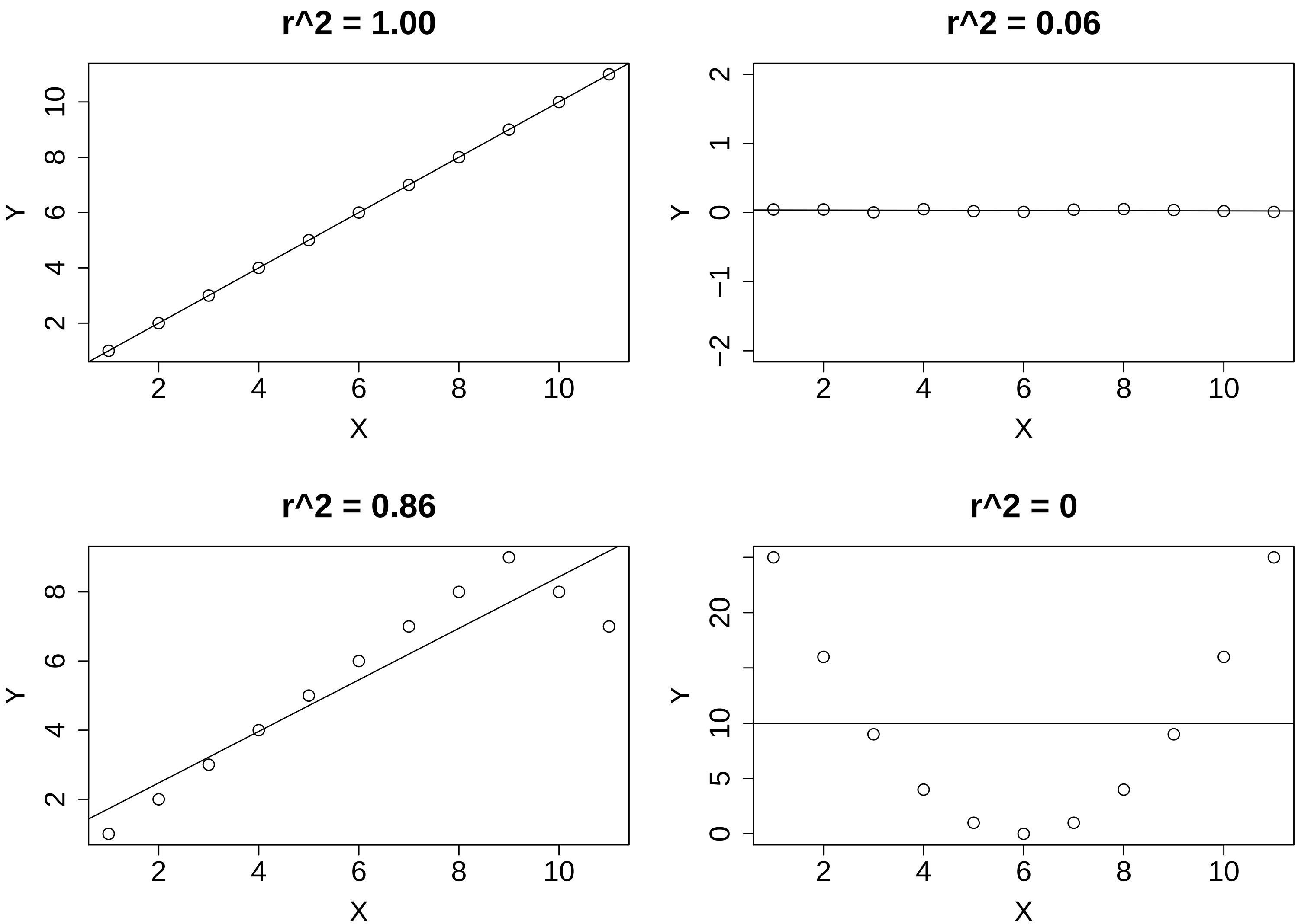
$$r_a^2 = 1 - \frac{N-1}{N-2}(1 - r^2)$$

- Thus $r^2 \approx r_a^2$ for large N

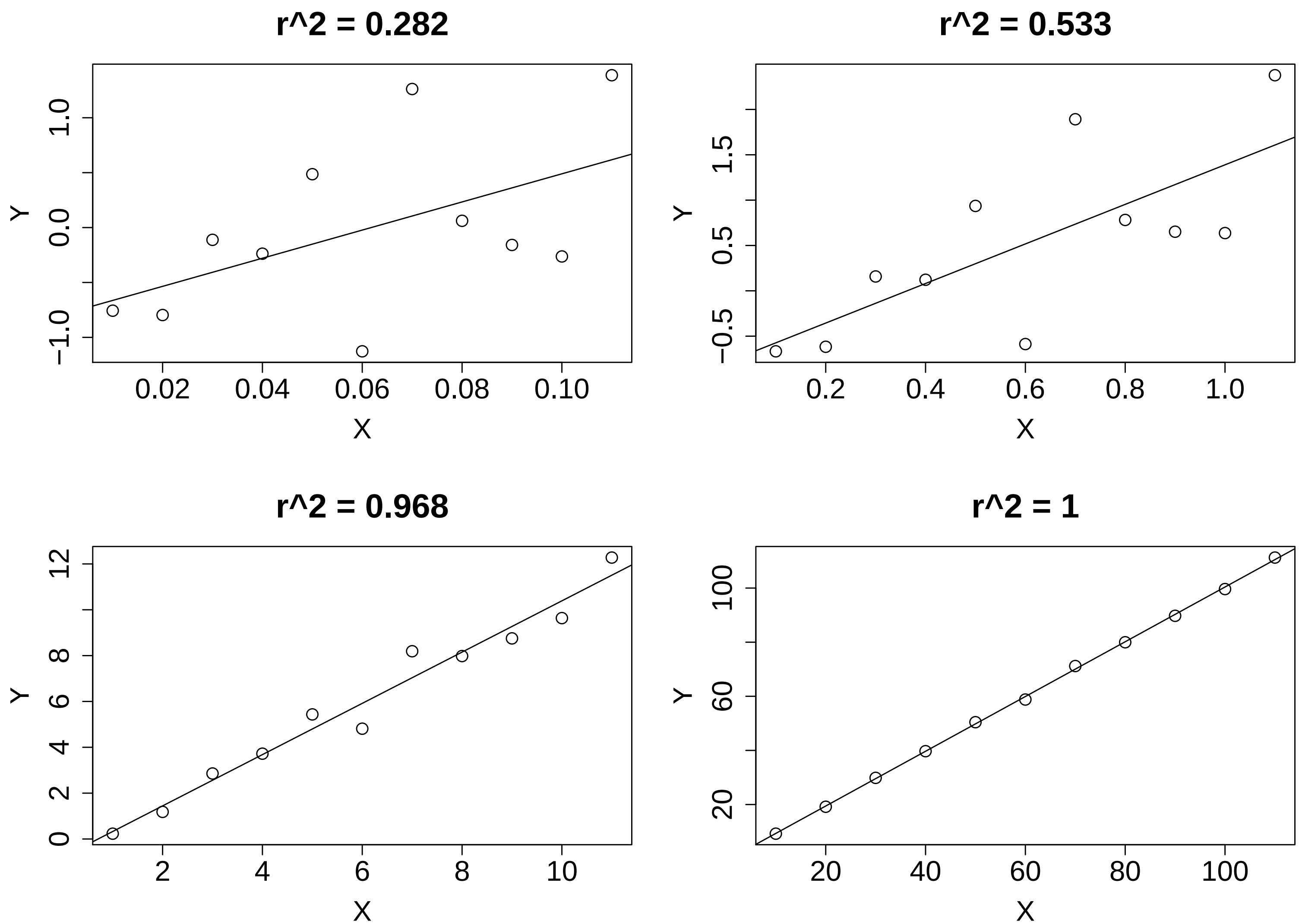
Unadjusted r^2

- 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 and Y; however, a non-linear association may still exist!
- $r^2 = 1$ indicates perfect fit; assessment of fit also by diagnostics
- $r^2 = 1 - \text{SSE}/\text{SST}$ typically increases with range/spacing of X

Examples of r^2



Examples of r^2: $Y = 0 + 1 \cdot X + \epsilon$, $\epsilon \sim N(0,1)$



Examples of r^2: $Y = 0 + 1 \cdot X + \epsilon$, $\epsilon \sim N(0, 4)$