Biost 517 / Biost 514 Applied Biostatistics I / Biostatistics I

Timothy A. Thornton, Ph.D.
Associate Professor of Biostatistics
University of Washington

Lecture 14:

Relationship between Two Variables; Scatterplots; Correlation; Introduction to Simple Linear Regression

Response and Predictor Variables

- Scientific questions are often concerned with the relationship between two (or more) variables of interest
- Insight is often provided by using statistical methods to assess the association between variables
  - –Response variable (outcome, dependent variable)
  - –Predictor of Interest (grouping variable, “independent” variable)
  - –The scientific question is addressed by comparing the distribution of the response variable across groups that are defined by the predictor/grouping variable

Two Sample Setting: Binary Predictor

- When the predictor (or grouping) variable is binary or categorical, we have covered a variety of methods for detecting associations based on comparing some summary measure of the response variable across groups

- For binary response and binary predictor, inference for an association is obtained by comparing either the proportions or odds of the response variable
  - hypothesis testing and confidence intervals for risk differences, risk ratios, or odds ratios

- For continuous response variable and binary predictor variable, can use the mean as a summary measure for inference on an association and compare differences in mean between two groups
  - hypothesis testing and confidence intervals for differences in the mean; e.g. two-sample t-tests

Infinite Sample Setting: Continuous Predictor

- For a continuous predictor, conceptually there are an infinite number of groups
  - –E.g., if the predictor is age, then with enough precision no two people in your sample have exactly the same age
  - –Even at a typical level of precision, there may be 1 person in many groups, and some groups that are not represented in your sample.
-

Continuous Predictor

• For a continuous predictor, thresholds can be used to make discrete groups for comparisons

- There are some significant drawbacks:
  - Results can be sensitive to the choice of threshold, i.e., the groupings
  - Two subjects with similar values of the predictor but happen to land on the opposite side of a threshold are placed in different groups: not appealing
  - If there are multiple groups, not ideal that the analysis does not appropriately takes into account the natural ordering of the groups
- Usually ideal to perform statistical analysis of using continuous groupings. Don't do thresholding of quantitative traits if it can be avoided!

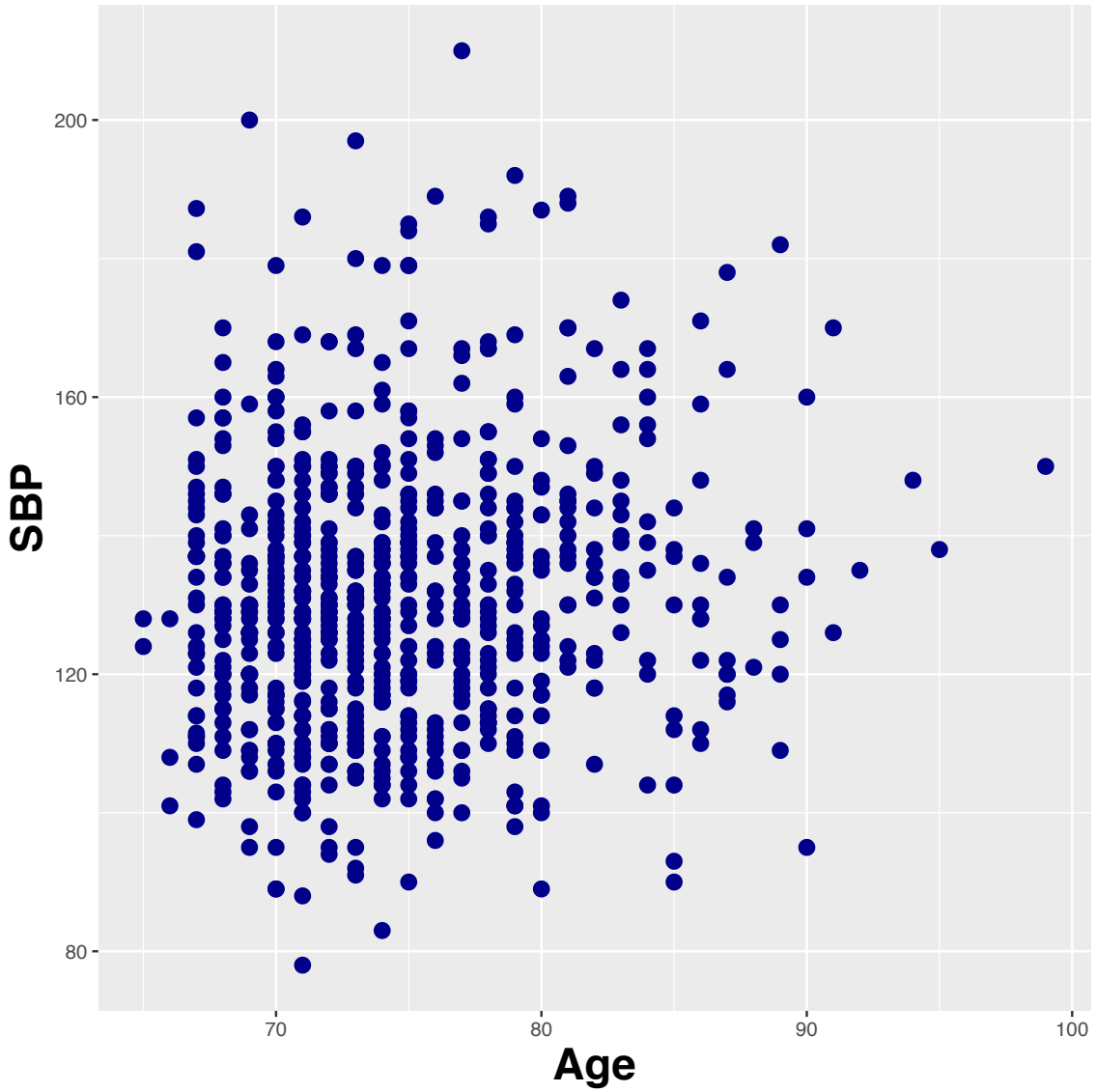
Examining Relationship Between Two Variables

- To investigate the relationship between two variables, we must measure both variables on the same individual.
- Two variables are **associated** with each other if some values of one variable tend to occur more often with some values of the second variable than with other values of the variable
- We are often interested in the following when examining the relationship between two variables:
  - Exploring the relationship between the variables (exploratory analysis)
  - Determining if one of the variables can explain the variation in the other for a scientific question of interest

Scatterplot: Graphical Display of Relationship for Quantitative Variables

- A scatter plot shows the relationship between two quantitative variables measured on the same individuals. It is most useful for the graphical display of two continuous variables.
- The values of one variable are on the x-axis, and the values of the other are on the y-axis. Each individual is represented by a point in the graph.
- If there is a clear predictor/explanatory variable, it is convention for the explanatory variable to go on the x-axis and the response variable to go on the y-axis.

Scatterplot of SBP versus Age in Elderly Adults

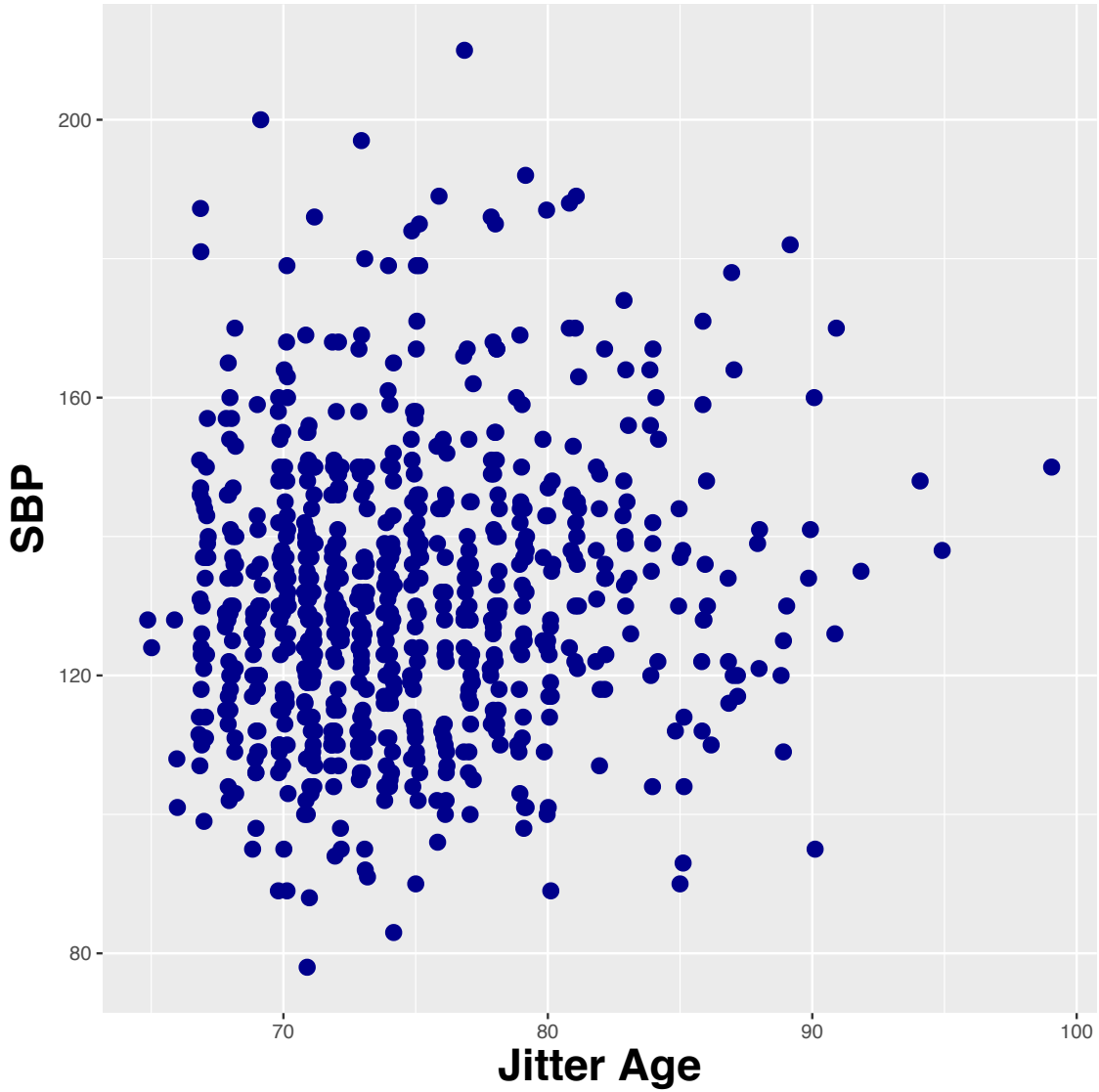


Jittered Scatterplots

- If variables are discretely measured, jittering can be helpful to allow all for most points to be seen in a scatterplot
- "jittering": adding a little noise to the data to break ties
- Can use the **jitter()** function in R to add random noise to a variable

```
> mri$JitterAge<-jitter(mri$age)
> head(mri$age)
[1] 72 81 90 72 70 72
> head(mri$JitterAge)
[1] 72.19175 81.04411 89.95812 71.85067 69.96484 72.14121
```

Scatterplot of SBP versus Age Jittered



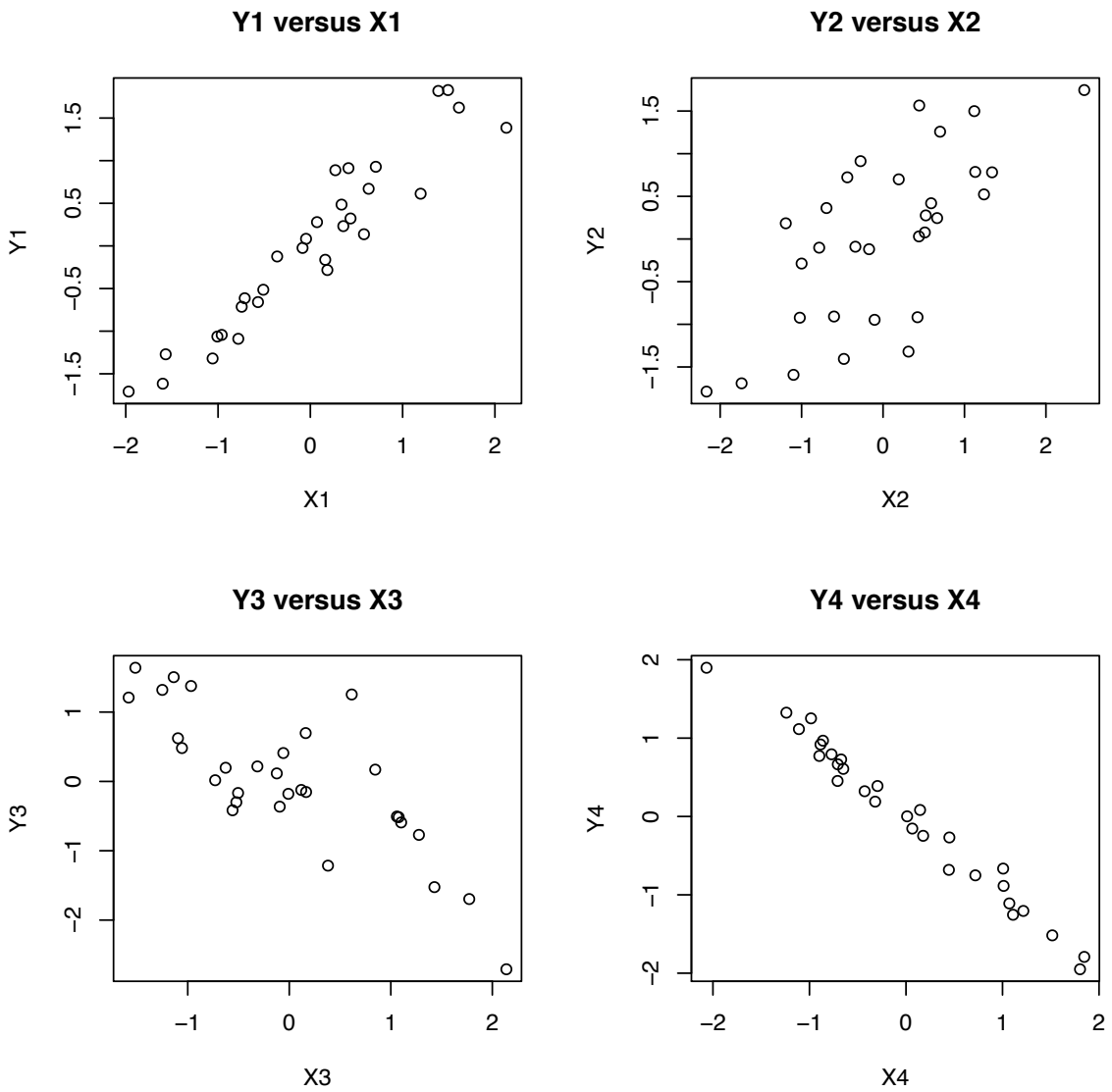
Assessing a Scatterplot

- First look for the overall pattern
  - What is the *form* of the relationship?
  - What is the *shape* of the relationship?
  - What is the *direction* of the relationship?
- Trends in location across groups
  - First order trends (linear)
  - Second order trends (curves, U-shape, S-shape)
- Trends in the within-group spread of data
- Are there any deviations from the overall pattern? An individual that falls outside the overall pattern is an **outlier**.

Assessing a Scatterplot: Association

- **Positive association**: Above-average values of both variables tend to occur together in the scatterplot, and the same for below-average values.
- **Negative association**: Above-average values of one variable tend to occur with below-average values of the other in the scatterplot, and vice-versa.
- The strength of the association is determined by how closely the points follow an overall pattern.

Assessing a Scatterplot: Association



Correlation

- The **linear relationship** is an important form of relationship between two variable. It is a first order trend between to variables.
- We are not good at judging linear relationships by eye, and can be fooled by scaling and stretching of a scatterplot.
- **Correlation** is a numerical measure of the *direction* and *strength* of the linear relationship between two quantitative variables.
- The **sample correlation coefficient** of two variables is often represented by ***r*** in statistics

Correlation

- Suppose we have n pairs of observations for two random variables X and Y: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. The correlation for the two random variables is

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) = \frac{S_{xy}}{\sqrt{S_x}\sqrt{S_y}}$$

where

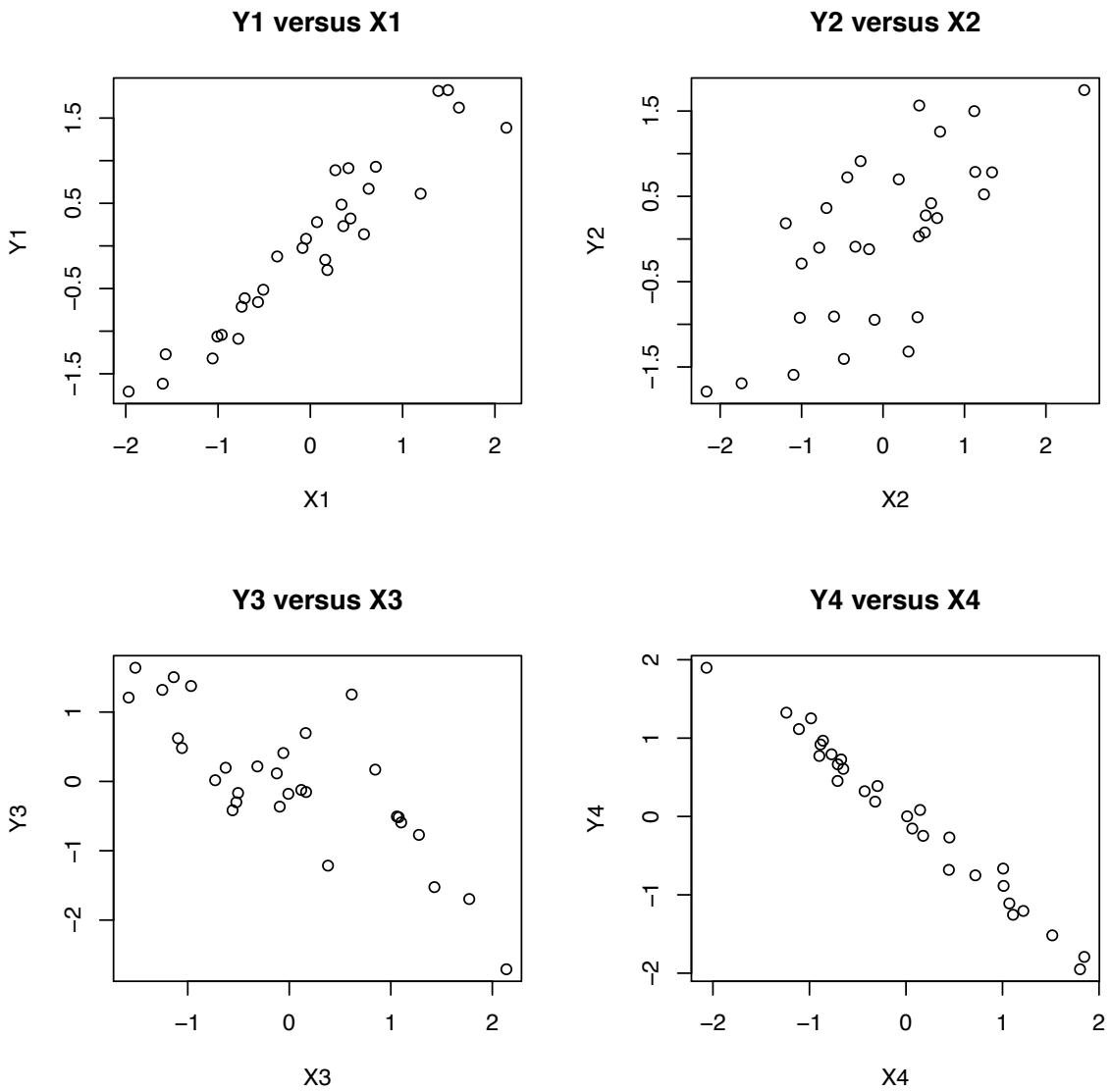
$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \qquad S_x = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Facts about Correlation

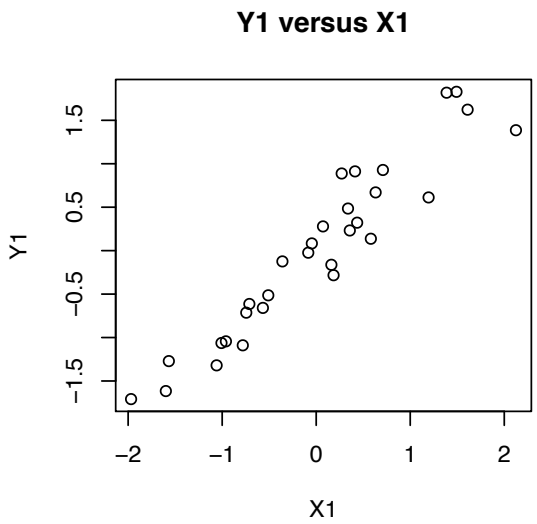
- Doesn't distinguish between the explanatory and response variable. (You can call either one x and the other y.)
- Can be computed only for two *quantitative* variables.
- Does not depend on the units of measurement.
- $-1 \leq r \leq 1$.
  - $r \approx 0$: very weak linear relationship;
  - $r > 0$: positive association; $r < 0$: negative association;
  - $r = -1$ or $r = 1$: *only* when all the data points on the scatterplot lie exactly along a straight line;
- Measures *only linear relationships*, not curved relationships.
- Not resistant, so outliers can greatly change its value.
- Adding a constant value to either the x or y coordinate for every individual will not change the correlation.

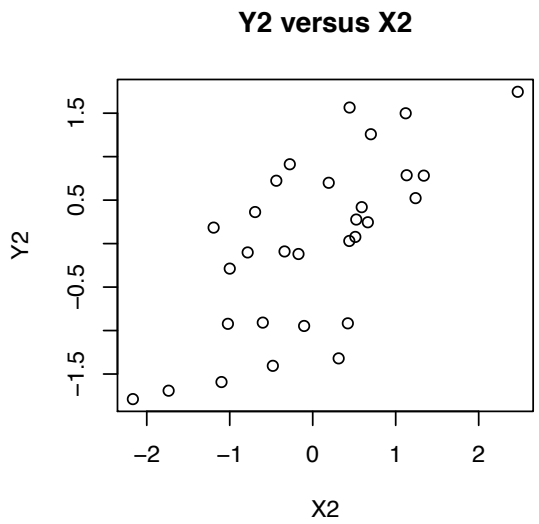
Can you guess the correlation?



Can you guess the correlation?

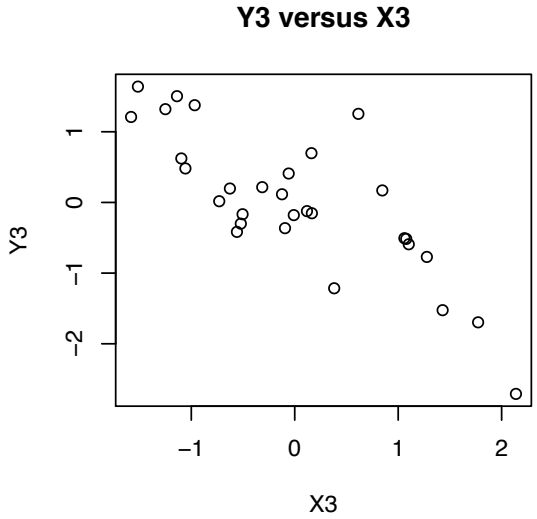
r=0.95

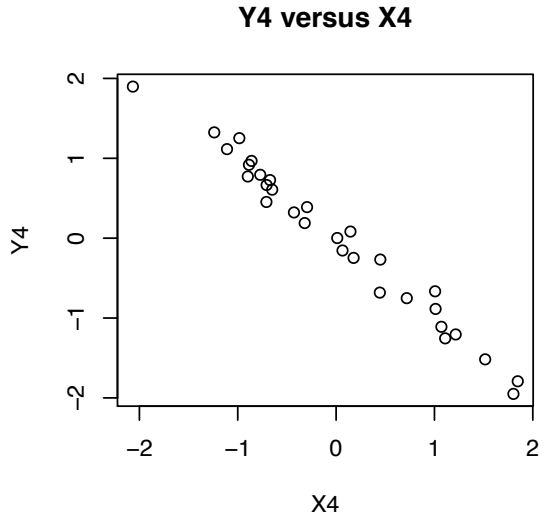




r=0.70

r=-0.8

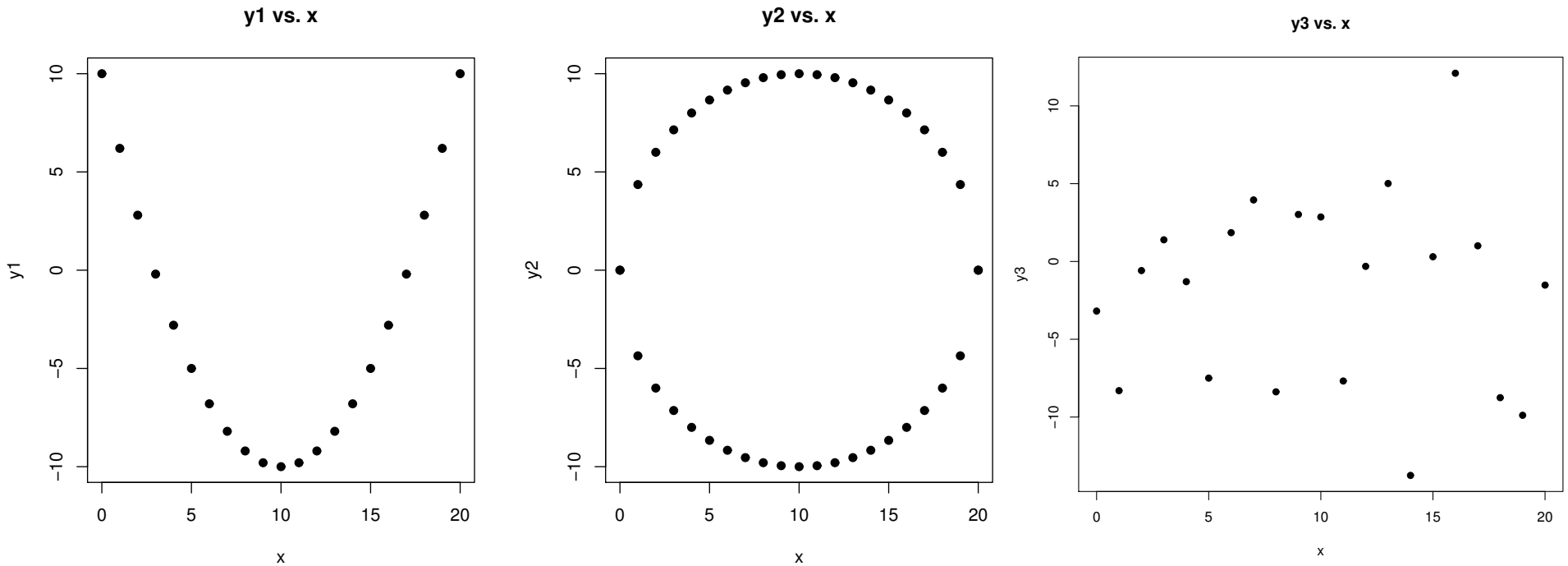




r=-0.99

Caution: Correlations of Zero

- A correlation of zero does not (necessarily) mean there is no relationship between two variables
- All of the following scatterplots of variables that have a zero correlation:



Caution: Correlations of Zero

- Correlation is not a complete description of bivariate (two-variable) data – even if the relationship truly is linear.
- You should also report additional summary statistics for each variable:
  - Mean
  - standard deviation
  - Median
  - 1st quartile
  - 2nd quartile (Median)
  - 3rd quartile
  - Minimum
  - Maximum
-

Blood Pressure Example: Scientific and Statistical Questions

- Let's go back to the blood pressure example in elderly adults
- Interested in association between blood pressure and age
- Scientific question:
  - Does aging affect blood pressure?
- Statistical question: Does the distribution of blood pressure differ across age groups?
  - Acknowledges variability of response
  - Acknowledges uncertainty of cause and effect
    - Differences could be related to calendar time of birth instead of age (e.g., a birth-cohort effect). Oldest individuals could have been born during the great depression which could have led to earlier disease progression that affects blood pressure.
  - Would need a longitudinal study to make inference about how blood pressure changes as a function of age.

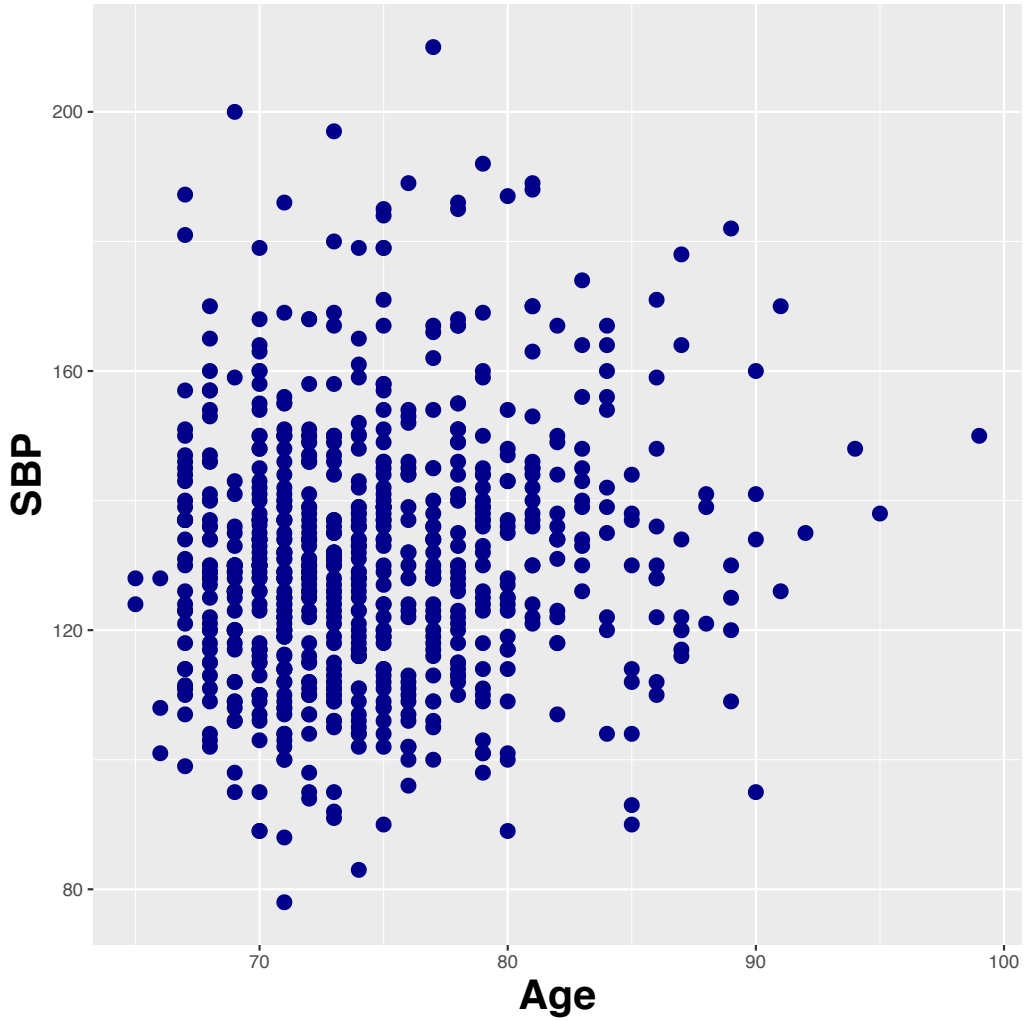
Cohort Study with SBP Measurements

- Cohort study of elderly adults aged 65 years and older
- Observation Study for the incidence of cardiovascular disease
- Data available on systolic blood pressure (SBP) for the study subjects
- Response: SBP
  - continuous

- Predictor of interest (grouping): Age
  - continuous
    - an infinite number of ages are possible
    - we probably will not sample every one of them

Descriptive Statistics: Correlation and Scatterplot of SBP versus Age

- Correlation between SBP and Age is $r = 0.12$
- Scatterplot: Response on y-axis, predictor on x-axis



SBP Example: Descriptive Statistics in R

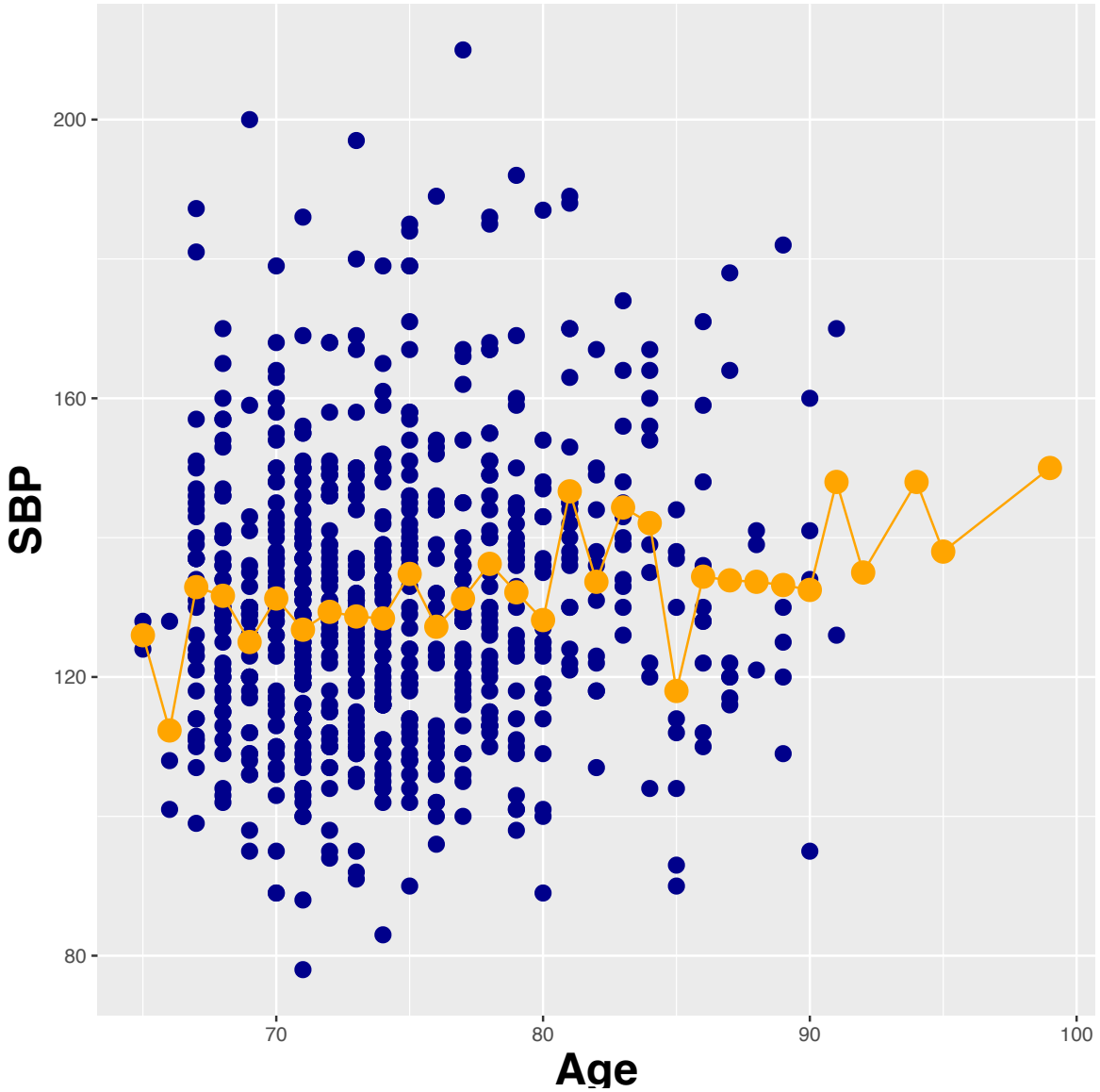
- Tabular: Stratified descriptive statistics
  - Strata by scientifically relevant categories (not quintiles)
  - Can use the R package uwIntroStats for this:

```
library(uwIntroStats)
mri$age5<-cut(mri$age,breaks=seq(60,100,5))
descrip(mri$sbp,strata=mri$age5)
```

		N	Mean	Std Dev	Min	25%	Mdn	75%	Max
sbp:	All	735	131.11	19.66	78	118	130	142	210
sbp:	Str (60,65]	2	126	2.83	124	125	126	127	128
sbp:	Str (65,70]	175	129.99	18.7	89	117	129	140	200
sbp:	Str (70,75]	298	129.37	19.02	78	116	129	139	197
sbp:	Str (75,80]	157	131.44	20.94	89	117	129	144	210
sbp:	Str (80,85]	67	138.73	20.02	90	128	138	148.5	189
sbp:	Str (85,90]	30	133.73	21.32	95	120	129	141	182
sbp:	Str (90,95]	5	143.4	16.82	126	135	138	148	170
sbp:	Str (95,100]	1	150	NA	150	150	150	150	150

Scatterplot of SBP versus AGE

- Scatterplot with mean SBP for each age plotted



Limitations with Two-way comparisons

- Conceptually, we could
  - Estimate the mean SBP for each AGE
  - For any two ages, test the hypothesis that the mean SBP is equal
  - If any such test is significant, conclude that we have evidence of an association between SBP and AGE (and, more specifically, that mean SBP varies with AGE)
- Multiple comparisons problem
- Not leveraging information from people with similar ages
  - When comparing 80-year-olds to 70-year-olds, not using any information about 69 year olds or 81-year-olds

Regression

- Regression can be viewed as an extension of the two sample statistical analysis setting to an “infinite sample” setting
- Allows for infinitely many groups to be compared. Grouping variables can be continuous, binary, or categorical
- In broad terms, regression refers to inference on differences in the response between different values of the predictor of interest

General Regression Setting

- The general regression model with a predictor involves:
  - Random variable Y that is the "response" variable.
  - Random variable X that is the "predictor of interest" (POI)
  - A parameter θ that is a summary measure of the distribution of Y
- Common choices for θ:
  - Mean
  - Geometric mean (providing Y is always positive)
  - Median (or, less commonly, some other quantile such as 25th or 75th percentile)
  - Probability that Y > c for some specified, scientifically relevant value of c
  - Odds that Y > c for some specified, scientifically relevant value of c
  - Hazard function (instantaneous rate of failure at a specified time, conditional on being at risk of failure at that time)

General Regression Setting

- Goal is to gain inference on the parameter θ for different values of X, i.e., $\theta|X = x$.
- Consider our SBP example in the elderly cohort. Let
  - Y be SBP (response)
  - θ is the mean of Y (parameter), i.e. $\mathrm{E[Y]}$
  - X is age (POI)
- We would interpret $\theta|X = 72$ to be $\mathrm{E[Y}|X = 72]$, i.e., the mean of SBP for elderly individuals in the populations who are aged 72

General Regression Model

- General notation for regression model

$$g(\theta \mid X = x) = \beta_0 + \beta_1 x$$

$g(\ )$ "link" function used for modeling

β_0 "Intercept"

β_1 "Slope for X (POI)"

- The link function "links" the POI to the population parameter
- Interpretation of the regression parameters depends on the link functions (will discuss this in more detail later).
- Typical link functions used with regression are either
  - "identity link" ($g(\theta) = \theta$) to describe an additive model
  - "log link" ($g(\theta) = \log(\theta)$) to describe a multiplicative model

Example: Linear Regression Model

- Linear regression uses an identity link.
- Model can be used to answer scientific questions that can be assessed from linear trends
- SBP Example: Want to estimate "best" fitting line for average SBP within age groups

$$E\left(SBP \middle| Age \right) = \beta_0 + \beta_1 \times Age$$

- An association will exist if the slope (β_1) is nonzero
  - In that case, the average SBP will be different across different age groups

"Rule of Thumb"

- The regression model thus produces something similar to "a rule of thumb"

  – E.g., "Normal SBP is 100 plus half your age"

$$E\left(SBP \middle| Age \right) = 100 + 0.5 \times Age$$

- Linear regression estimates parameters using "least squares"
  - Most efficient average-based estimates for homoscedastic data
    - Asymptotically normal distribution for estimates
  - Most efficient estimation when data is normal within groups
    - Normal distribution for estimates even in small sample sizes

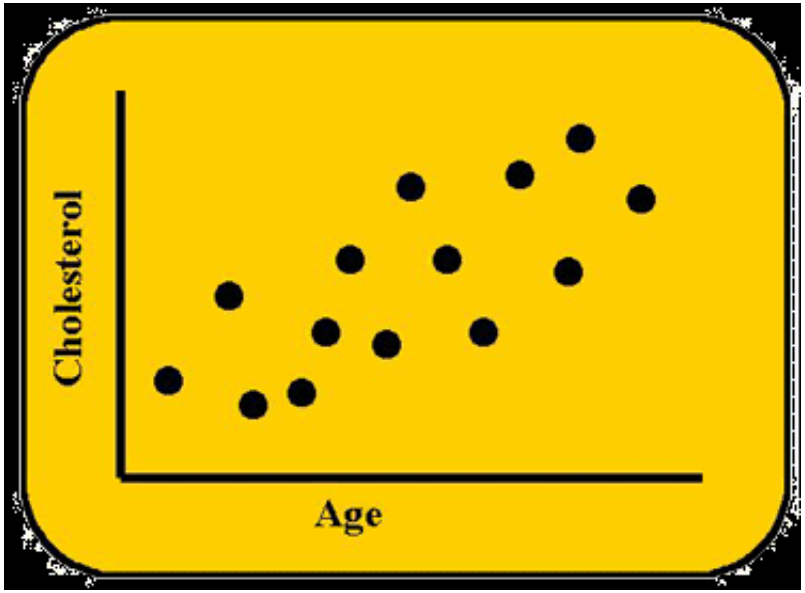
Least Squares Line

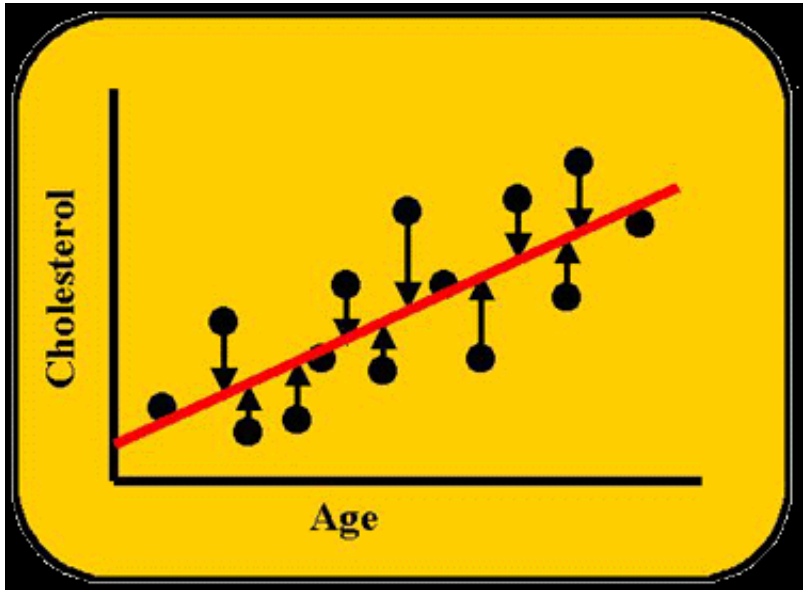
- Find the straight line that minimizes total squared vertical distance from data to line
  - Conceptually: Trial and error search
    - Guess a formula for a line
    - Compute total squared distance from data to line
    - Iterate until smallest number found
  - Calculus:
    - Find a formula based on derivatives
  - Real life:
    - Computers find such estimates easily

Least Squares Line

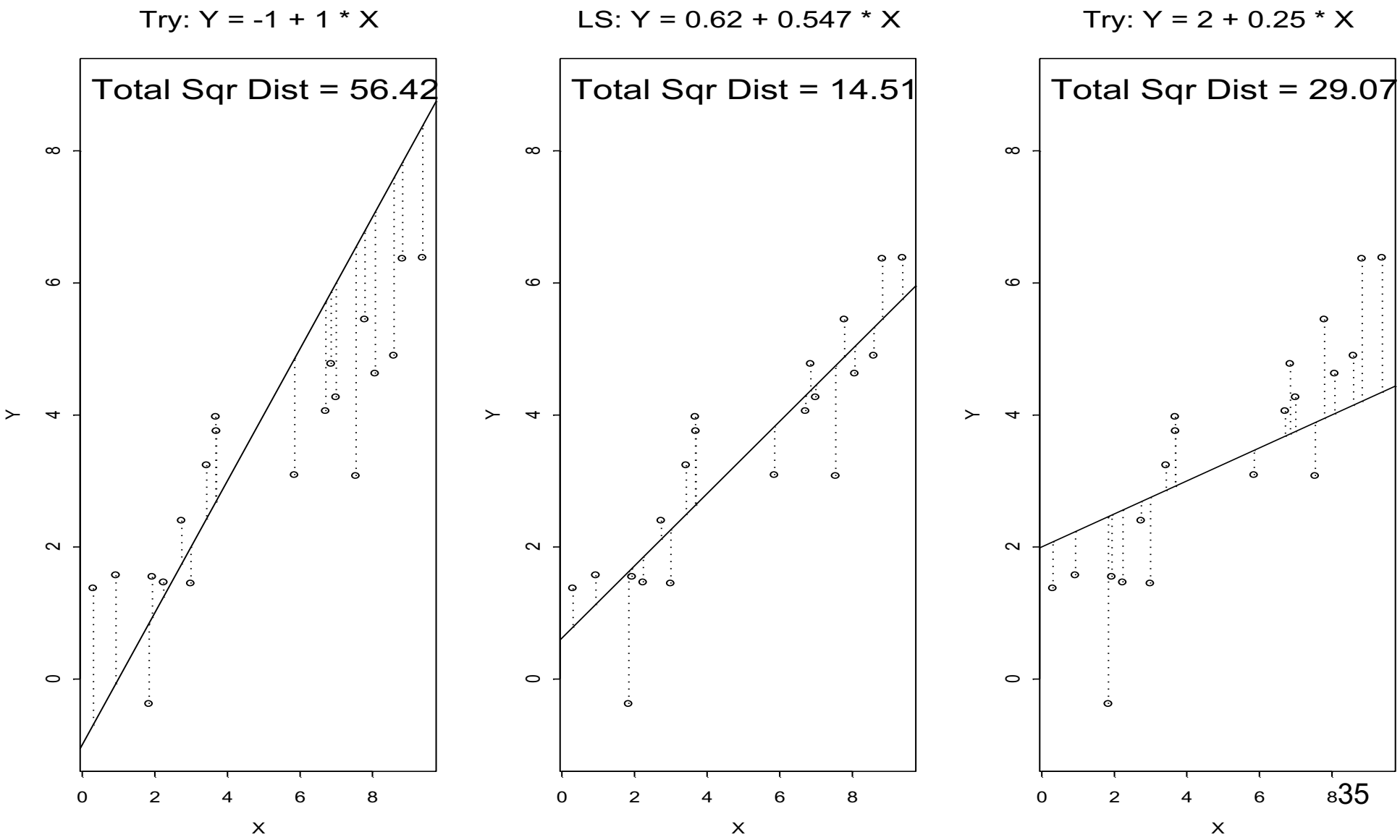
- The regression line is sometimes called the “least squares line.”

- It is the line that minimizes the sum of squared (vertical) distances from each response to the line.





Conceptual Example



Linear Regression in R

- In R, linear regression can be preformed in R with the ***lm()*** function
  - lm is “linear model”
- Commands:

```
mylm=lm(respvar~predvar,data=mydata)
```

where “respvar” is the response variable and “predvar” is the predictor variable in the linear regression model

- Using the ***summary()*** function with a linear model object created using the ***lm()*** function, e.g., “*summary(mlym)”,* will provide regression parameter estimates and inference
  - Point Estimates for Intercept and slope
  - Standard Errors
  - Test statistics and 2-sided *p*-values for testing the null hypothesis that each of the regression parameters are equal to 0.

Example: Estimates, Inference in R

```
model=lm(sbp~age,data=mri)
summary(model)

Call:
lm(formula = sbp ~ age, data = mri)

Residuals:
    Min      1Q  Median      3Q     Max
-51.568 -13.843  -0.568  10.432  77.845

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  98.9494     9.8894   10.01  < 2e-16 ***
age           0.4312     0.1323    3.26  0.00116 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.54 on 733 degrees of freedom
Multiple R-squared:  0.01429,   Adjusted R-squared:  0.01295
F-statistic: 10.63 on 1 and 733 DF,  p-value: 0.001165
```

$$E\left(SBP \middle| Age \right) = 98.9 + 0.431 \times Age$$

Linear Regression Inference

- The regression output for R provides
  - Estimates
    - Intercept: estimated mean when age = 0
      - Estimated intercept: 98.9
    - Slope: estimated difference in average SBP for two groups differing by one year in age
      - Slope is labeled by variable name: “age”
      - Estimated slope: .431
  - Standard errors
  - P values testing for
    - Intercept of zero (who cares?)
    - Slope of zero (test for linear trend in means)

Linear Regression Inference

- Confidence intervals can easily get be obtained in R using the either of the following commands: *confint (model)* or *confint.default(model),* where the later command assumes asymptotic normality

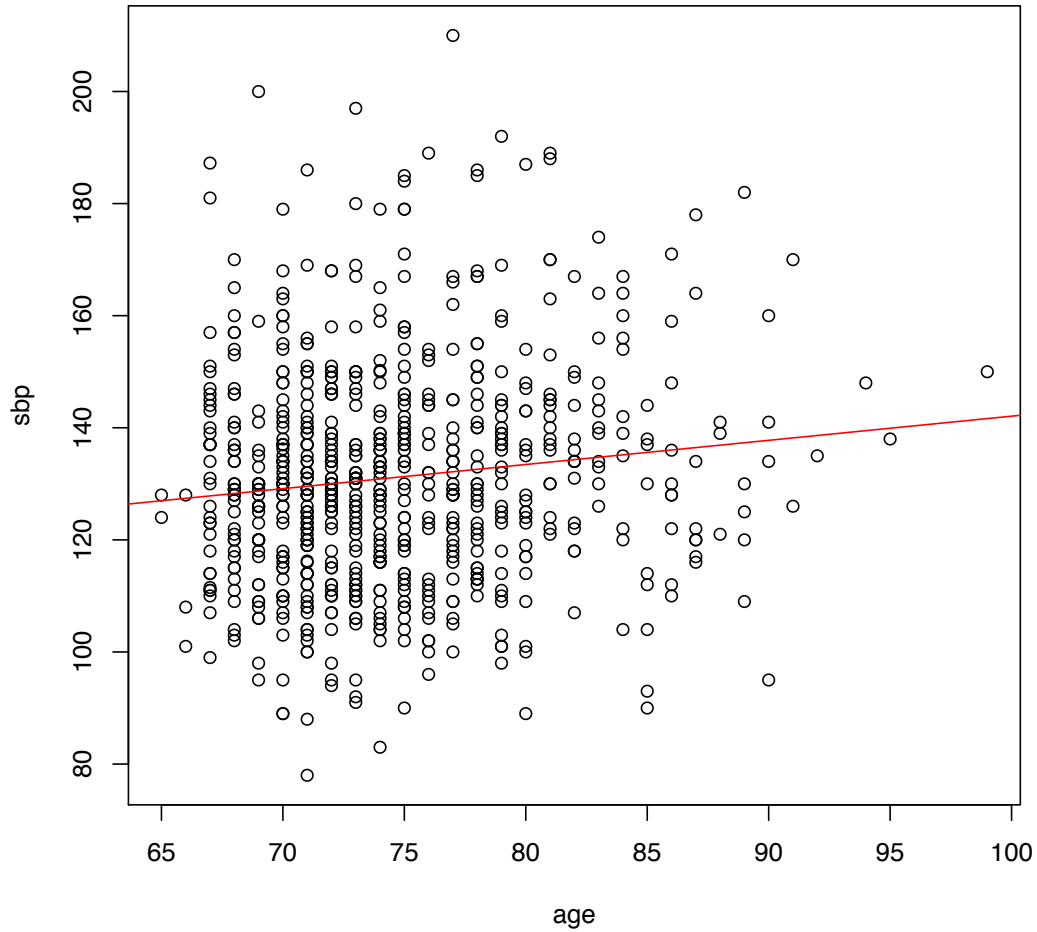
```
>confint.default(model)

                2.5 %      97.5 %
(Intercept) 79.5666299 118.3322457
age          0.1719901   0.6904915
```

Linear Regression of SBP on AGE

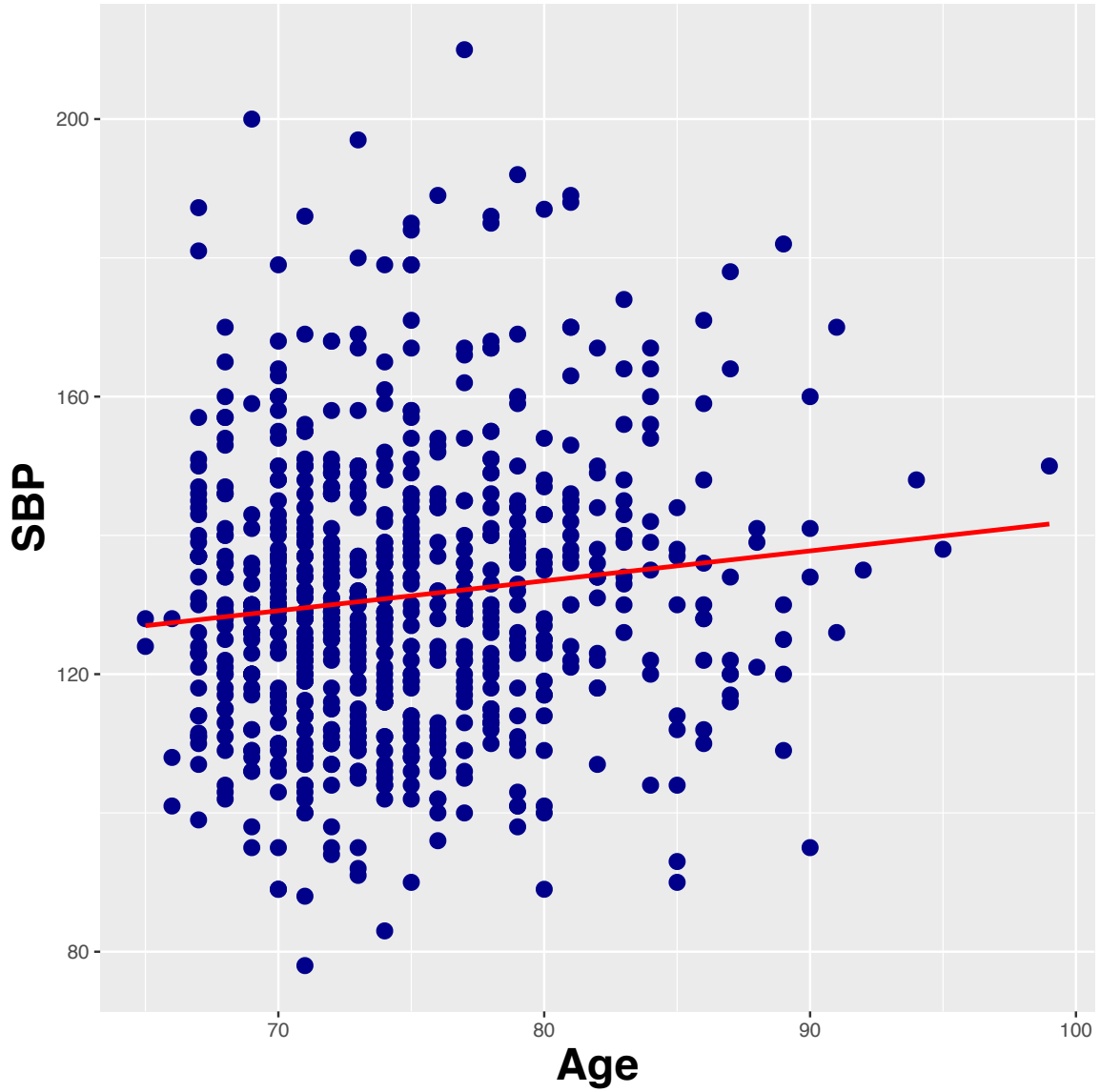
```
>plot(sbp~age,data=mri)

>abline(model,col="red")
```



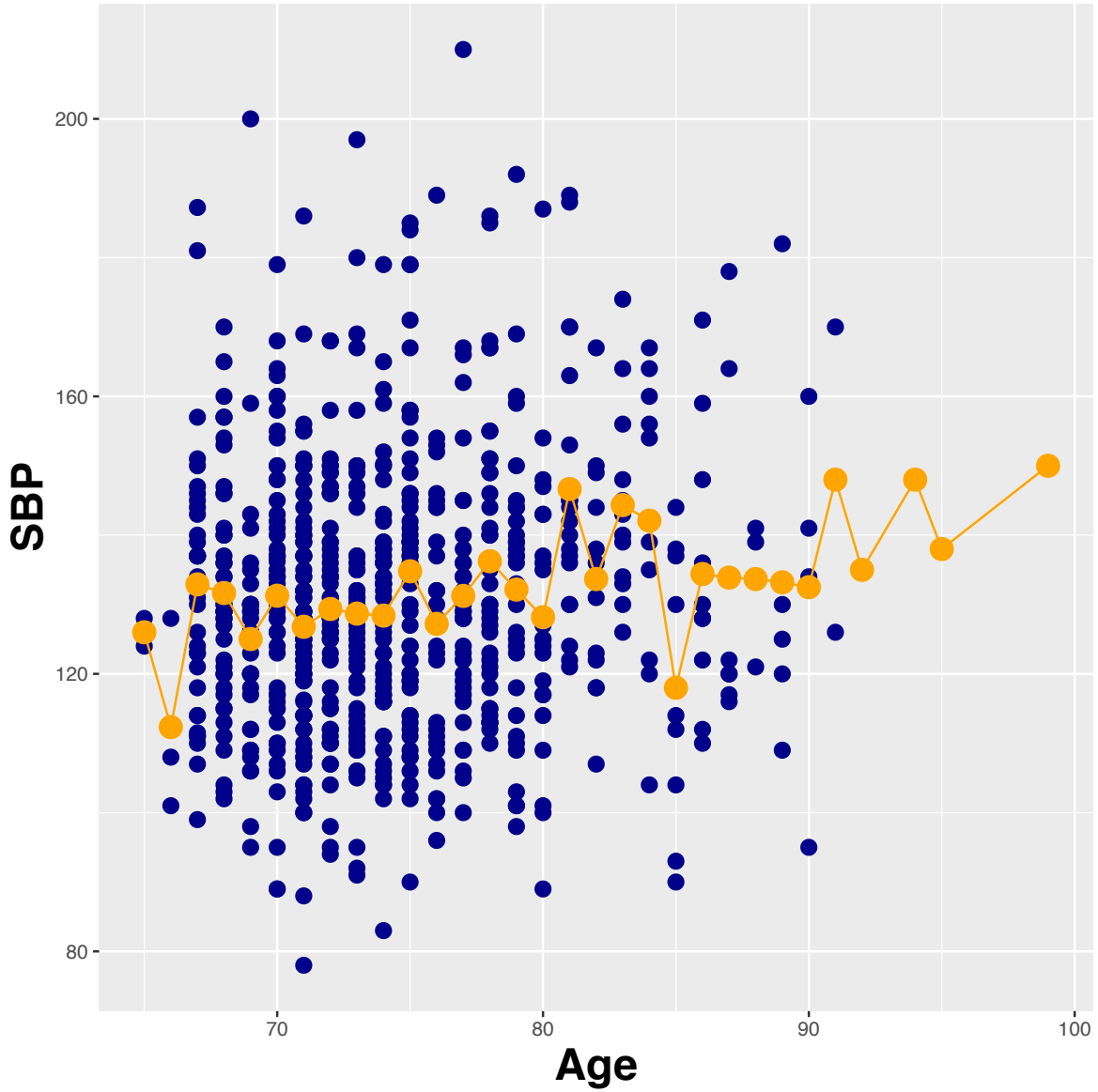
Linear Regression of SBP on AGE

- Graph with the ggplot2 package in R



Scatterplot of SBP versus AGE

- Scatterplot with mean SBP for each age plotted



Use of Regression

- The regression “model” serves to
  - Make estimates in groups with sparse data by “borrowing information” from other groups
  - Define a comparison across groups to use when answering scientific question

Borrowing Information

- Use other groups to make estimates in groups with sparse data

- Intuitively: 67 and 69 year olds would provide some relevant information about 68 year olds

- Assuming straight line relationship between mean SBP and age tells us how to adjust data from other (even more distant) age groups
  - If we do not know about the exact functional relationship, we might want to borrow information only close to each group
    - (splines in BIOST 515/518)

Defining “Contrasts”

- Define a comparison across groups to use when answering scientific question

- If straight line relationship in means, slope is difference in mean SBP between groups differing by 1 year in age
  - Regression in some sense considers all possible pairwise contrasts, and then averages them in a special way

- If nonlinear relationship in means, slope is average difference in mean SBP between groups differing by 1 year in age
  - Statistical jargon: a “contrast” across the means

Example: Interpretation

"From the linear regression analysis, we estimate that for each year difference in age between two populations, the difference in mean SBP is 0.43 mmHg. A 95% CI suggests that this observation is not unusual if the true difference in mean SBP per year difference in age were between 0.17 and 0.69 mmHg. Because the two sided P value is P < .005, we reject the null hypothesis that there is no linear trend in the average SBP across age groups."

Example: Interpretation

- Note specification of point estimate, CI, and p value
  - Response: SBP (measured in mmHg)
  - Summary measure: mean
  - Contrast of summary measure across groups: difference
  - Predictor of interest: age
  - Difference in POI across groups being compared: 1 year

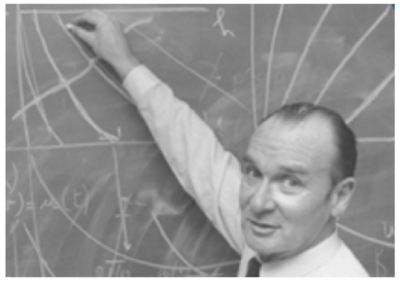
"From linear regression analysis, we estimate that for **each year difference in age** between two populations, the **difference** in **mean SBP** is 0.43 mmHg. A 95% CI suggests that this observation is not unusual if the true difference in mean SBP per year difference in age were between 0.17 and 0.69 mmHg. Because the two sided P value is P < .0005, we reject the null hypothesis that there is no linear trend in the average SBP across age groups."

Disclaimer/Warning for Regression Models

In statistics, as in fashion, a model is an idealization of reality.

Peter McCullagh
JRSSD (1999) 48:1

Models basically play the same role in economics as in fashion: they provide an articulated frame on which to show off your material to advantage ...; a useful role, but fraught with the dangers that the designer may get carried away by his personal inclination for the model, while the customers may forget that the model is more streamlined than reality.



Jacques Drèze
Economic Journal (1985) 95:380