MEDICAL COLLEGE OF WISCONSIN

The Straight Dope on Simple Linear Regression

Mei-Jie Zhang, PhD

Sponsored by the Clinical and Translational Science Institute (CTSI) and the Department of Population Health Division of Biostatistics

Speaker Disclosure

In accordance with the ACCME policy on speaker disclosure, the speaker and planners who are in a position to control the educational activity of this program were asked to disclose all relevant financial relationships with any commercial interest to the audience. The speaker and program planners have no relationships to disclose.

Outline

- What is simple linear regression?
- Review of equation of a line
- Fitting a line to data
- Interpretation and prediction
- Confidence intervals and hypothesis tests
- Measuring the strength of association
- Assumptions of linear regression and model checking
- Binary predictors
- Caveats

What is linear regression?

- Method for describing association between a continuous outcome or dependent variable and one or more predictor variables (continuous, binary, categorical, etc.) in a single equation
  - Estimate magnitude, strength, and statistical significance of that relationship
  - Develop an equation to predict outcome for specific values of the predictor variables

Simple linear regression

- "**Simple**" refers to a single predictor variable
- We will focus initially on a continuous predictor, but talk at the end about a binary predictor
- Relationship described by the equation of a line:

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

Example

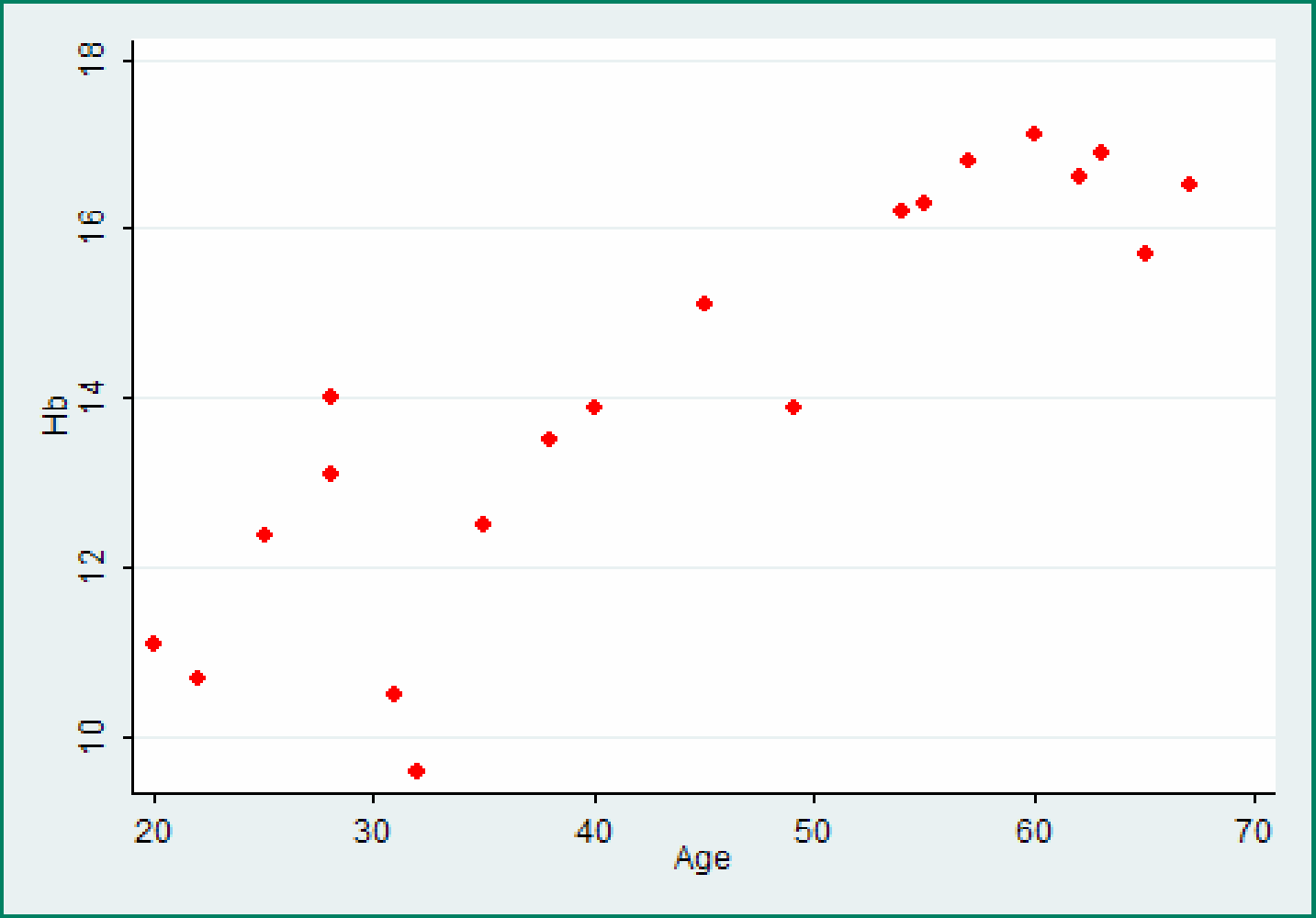
- Survey of anemia in women
- Blood sample drawn, and hemoglobin (*Hb*) and packed cell volume (*PCV*) measured
- Also asked their *Age* and whether they had experienced menopause
- Sample of 20 women
- Table 9.1 in Campbell et al.

Example

- **Researchers want to know: What is relationship between hemoglobin (*Hb*) and *Age*?**
  - **Does *Hb* increase with *Age* (and by how much per year), or decrease, or is it unrelated?**
  - **How strong is that relationship? Can *Hb* be very accurately predicted just from knowing their *Age*?**
  - **Is the relationship statistically significant?**
- **For example: What would be the predicted hemoglobin level for a 45 year old woman?**

Subject	Hb	PCV	Age	Menopause
1	11.1	35	20	0
2	10.7	45	22	0
3	12.4	47	25	0
4	14	50	28	0
5	13.1	31	28	0
6	10.5	30	31	0
7	9.6	25	32	0
8	12.5	33	35	0
9	13.5	35	38	0
10	13.9	40	40	1
11	15.1	45	45	0
12	13.9	47	49	1
13	16.2	49	54	1
14	16.3	42	55	1
15	16.8	40	57	1
16	17.1	50	60	1
17	16.6	46	62	1
18	16.9	55	63	1
19	15.7	42	65	1
20	16.5	46	67	1

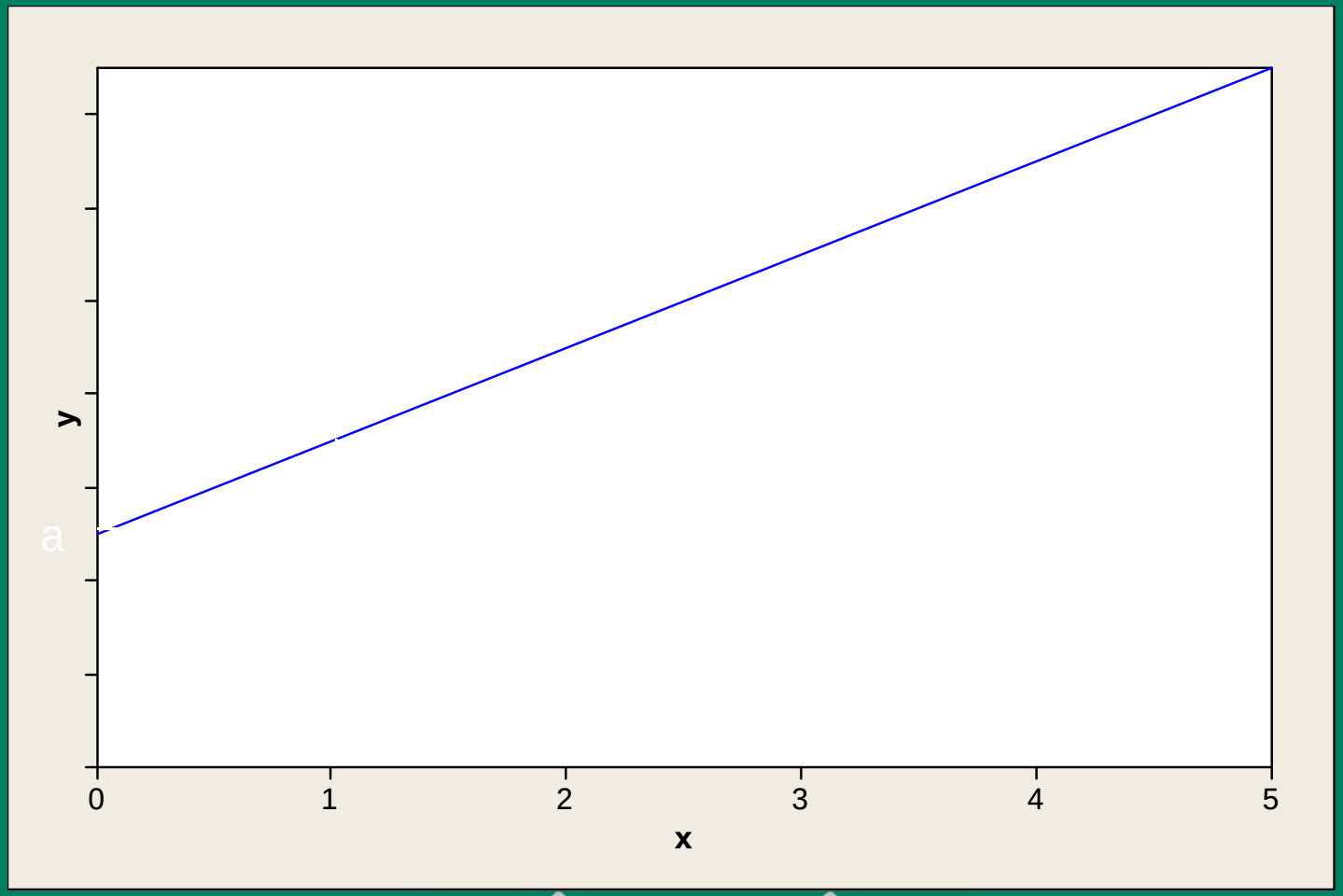
Scatter Plot



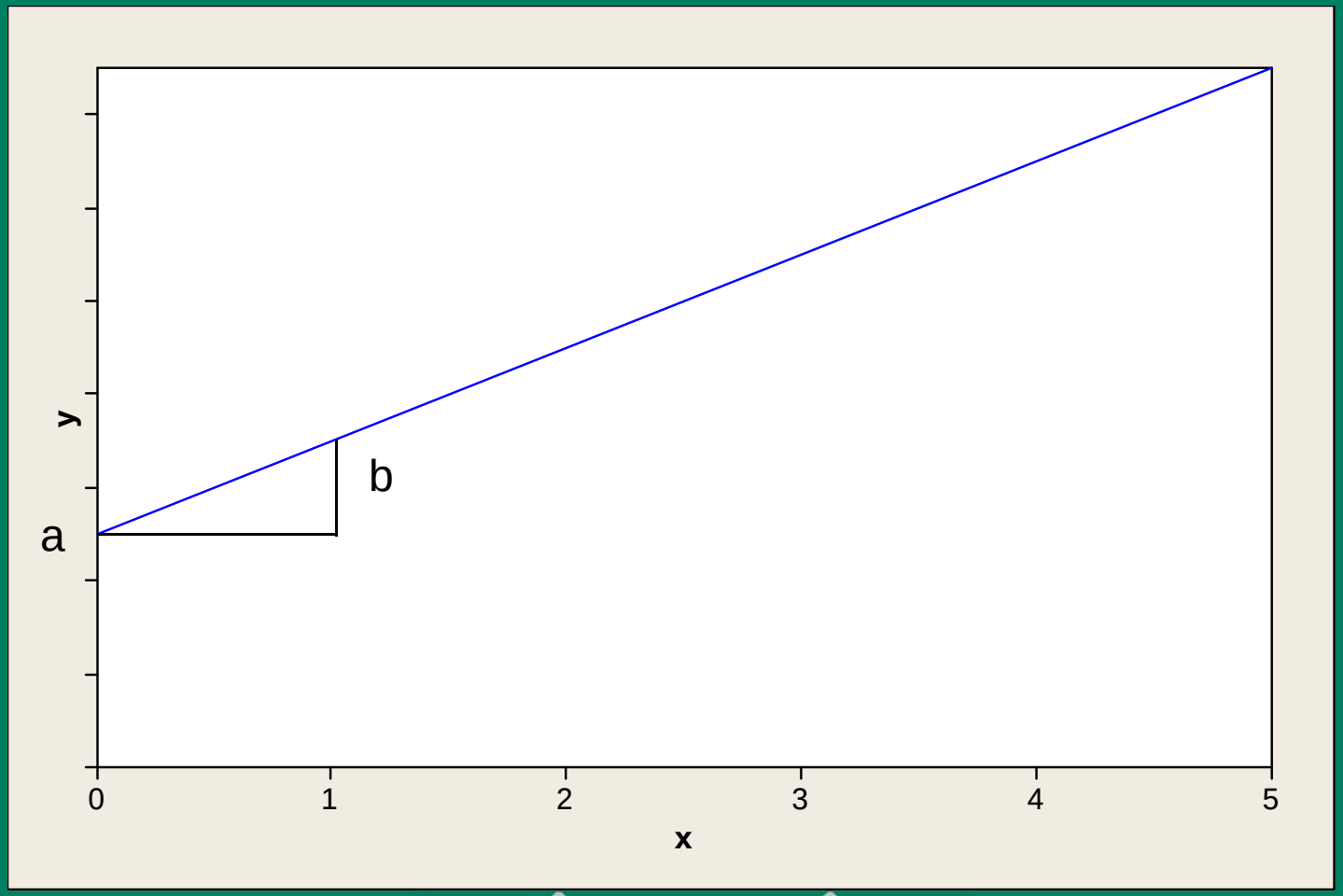
Review of Equation of a Line

- $Y=a+(bX)$
- a=*intercept*: Value of Y when $X=0$
- b=*slope*: Change in Y for one unit increase in X

Equation of a Line



Equation of a Line



Linear Regression

- **In most realistic situations, the data do not fit exactly on a line**
  - **Our outcome data *Y* are random variables**
- **Does *Y* increase *on average* as *X* gets larger (or does it decrease on average)?**
- **Linear regression relates the *mean* of *Y* to the predictor *X* using a line**

Linear Regression

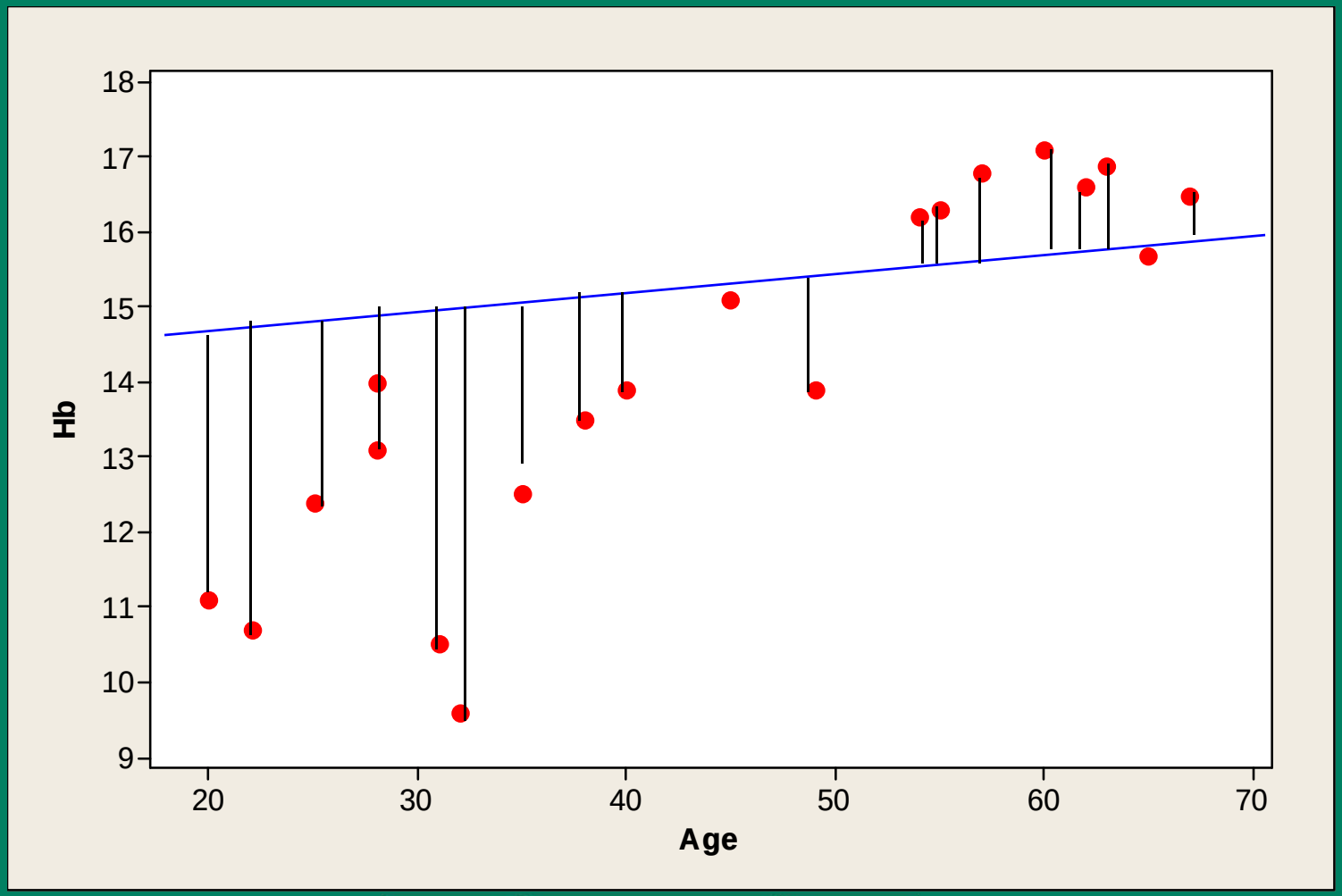
- **Population (true) regression line**
  - $E(Y)=\alpha+\beta*X$
  - $E(Y)$**: mean or expected value of** Y
  - α **and** β **are unknown**
- **Estimated regression line**
  - $E(Y)=a+b*X$
  - a **and** b **are estimates of the population regression line and are subject to sampling variability [estimated by most statistical packages]**

Estimating regression line

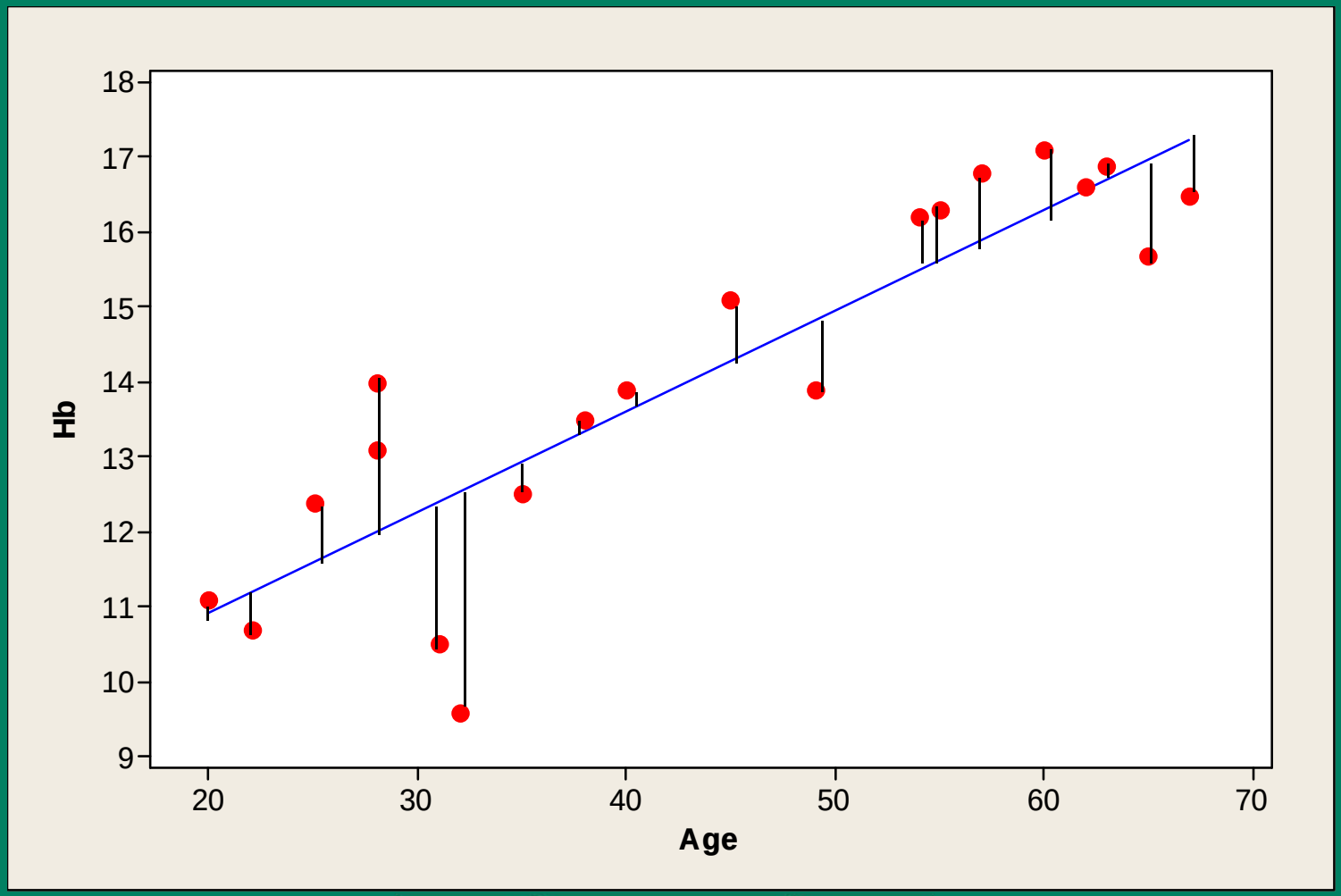
- **Estimated line is the one that "fits the data best"**
- **Each observation, Y, is subject to error**
  - **Observed value:** y_i
  - **Predicted value:** $\hat{y}_i = a + bX_i$
  - **Error or residual:** $e_i = y_i - \hat{y}_i$
- **Best-fitting line minimizes the total errors: actually, sum of squared errors**

$$\sum_i (y_i - \hat{y}_i)^2 = \sum_i \left[y_i - (a + bX_i) \right]^2$$

Poorly fitted regression line



Least Squares regression line



Least Squares Regression line

- **Estimates *a* and *b* can be estimated using statistical software, such as**

  SAS, SPSS, STATA, MINITAB, …

- *a=8.24*
- *b=0.134*
- *Hb = 8.24 + 0.134 * Age*

Interpreting regression parameters

- a: Expected or mean value of Y for a person with an X value of *0*
  - $a=8.24$: Mean value of *Hb* for a 0 year old woman is 8.24
  - Not necessarily meaningful
    - Range of *Age* is between 20 and 67
    - Shouldn't extrapolate beyond data range
  - Intercept term is mathematically necessary to fit the line even if it isn't meaningful

Interpreting regression parameters

- b: Change in mean value of Y for every one unit increase in X
  - $b=0.134$: A one year increase in ***Age*** is associated with an increase in mean ***Hb*** of 0.134
  - Individuals who differ in age by one year would have on average a mean difference in ***Hb*** of 0.134.
  - This mean difference is consistent across ***X*** values or ages
    - 21 year olds vs. 20 year olds
    - 56 year olds vs. 55 year olds

Prediction

- **We can use the regression model to predict outcomes for an individual with a specific value of X**
- **Expected *Hb* level for a *30 year old woman* would be:**

$$Hb = 8.24 + 0.134 * 30 = 12.26$$

Confidence Intervals and Hypothesis Tests

- Regression parameters a and b are estimated based on a random sample of 20 women
- a and b are random variables which are subject to sampling variability
  - A different sample of 20 women will have a different estimated regression line
  - Construct confidence intervals for the true regression parameters: in particular, slope β
  - Test whether the association is "real" or not (whether the slope is significantly different from 0)

Confidence Intervals

- Estimate b has a **standard error** ($SE(b)$) associated with it (available from computer package)
- **Sampling distribution** of b is
  - T for a **small** sample size (**<30**), under certain conditions (more later)
  - Z for **large** sample sizes (**≥30**)
  - Usual formulas apply

$$b \pm C_{\alpha/2} \times SE(b)$$

Confidence Intervals (Small sample size)

- $b \pm t \times SE(b)$
- t is critical value from a t-distribution with $(n\text{-}p)$ degrees of freedom
  - n= sample size
  - p= number of parameters in regression model
  - Here we have two parameters: slope and intercept

Confidence Intervals (Large sample size)

- $b \pm Z \times SE(b)$
- Z **is critical value from a standard normal distribution**

  For a 95% Confidence interval, $Z=1.96$

Confidence Intervals: Example

- *b=0.134*; *SE(b)=0.017*
- **d.f. (Degrees of freedom)=20-2=18**
- *t=2.101* **for 95% confidence**
- **Confidence Interval:**

$$0.134 \pm 2.101 \times 0.017 = [0.098, 0.170]$$

- **Note that this interval does not include 0 → Relationship between *Age* and *Hb* is significant**

Hypothesis Tests

- **Is apparent relationship (slope) real or consistent with sampling variability?**
- **Test of null hypothesis: $H_0: \beta=0$**
- **Test Statistic:**

$$\frac{b - 0}{SE(b)}$$

- **This has either a *t*-distribution with *(n-p)* d.f. or a *Z*-distribution under the null hypothesis, depending on sample size**

Hypothesis Tests: Example

- The observed slope of $b=0.134$ has a p-value of <0.001, indicating that there is a statistically significant association between *Age* and *Hb* levels.
- The 95% confidence interval for the slope is from 0.10 to 0.17, indicating that each additional year of *Age* is associated with an increase in average *Hb* of 0.1 to 0.17

Measuring the strength of association

- The coefficient b tells you how much the mean of Y changes with X
- It does not tell you how *strong* the association is
- **Strength of association** refers to how closely the points follow the regression line
- How much does ***Age*** explain differences in ***Hb***?

Coefficient of determination

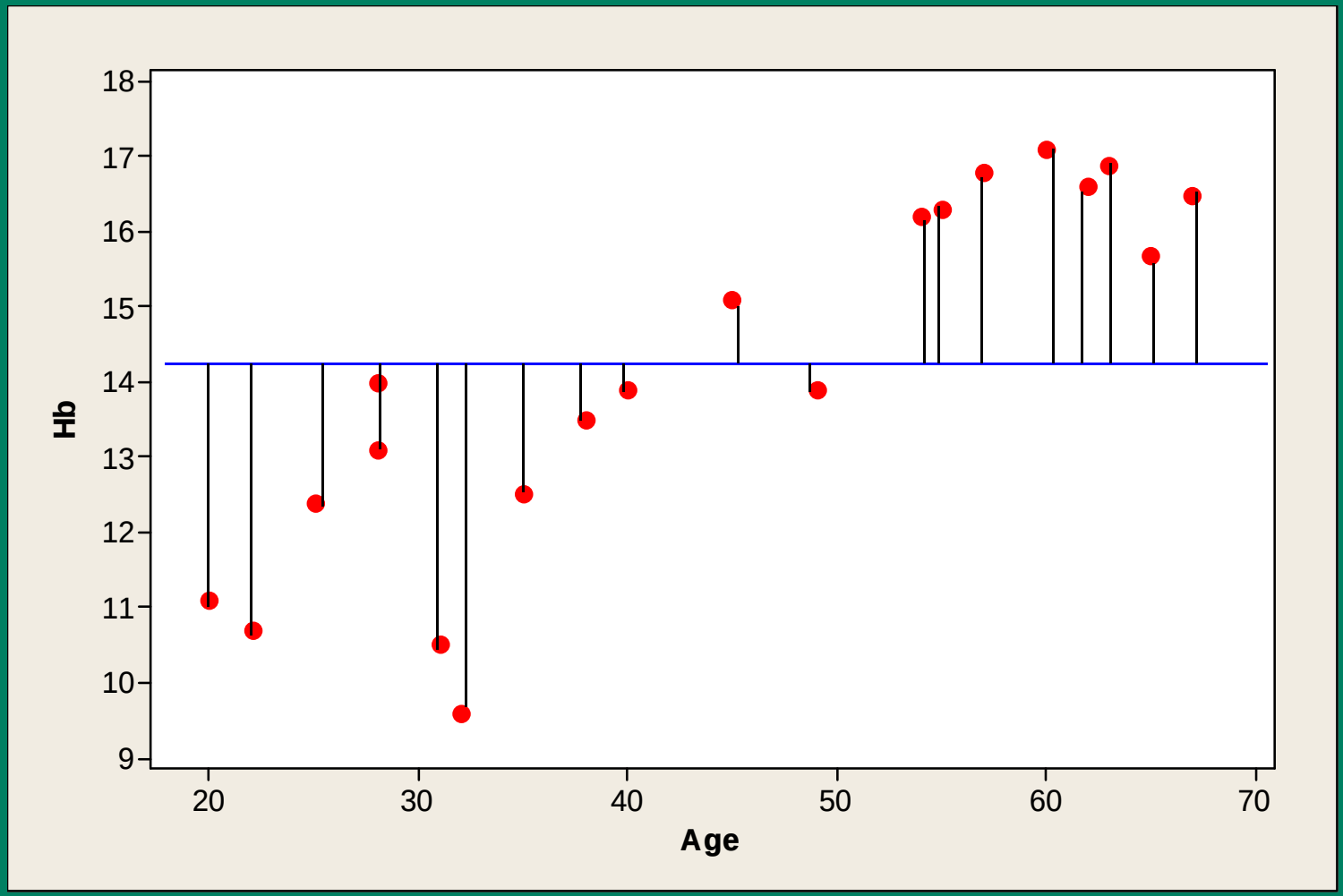
- R^2: Measure of the percent of the variability in the outcome Y which is accounted for by the regression on X
- R^2 near **1** (or 100%): X is accounting for all the variation in Y, so the points fall on the regression line
- R^2 near **0** (or 0%): X does not explain any of variation in Y, so the **scatter-plot** looks completely random (no discernible linear trend)

Coefficient of Determination: Example

- In our study, $R^2 = 77.4\%$, so that ***Age*** explained 77.4% of the variability in ***Hb*** values
- There is still 22.6% unexplained
  - May be other factors not considered yet which also affect ***Hb***

Variability around mean

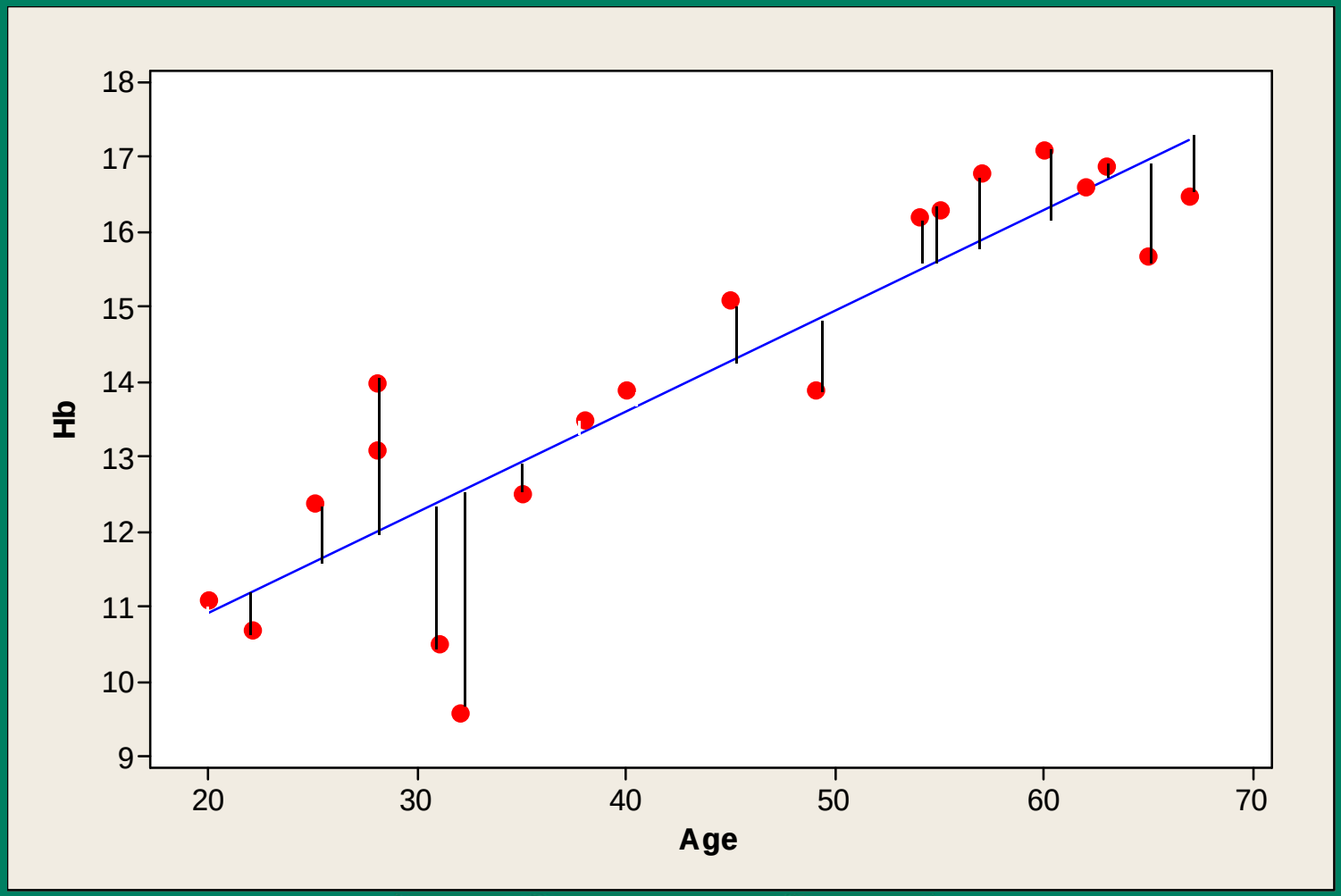
(Sum of Squared Residuals Around Mean = 109.62)



Mean Hb = 14.1

Variability around regression line

(Sum of Squared Residuals = 24.81)



Coefficient of Determination

- **Variability unexplained by regression**
  - **Sum of Squared Residuals around regression divided by Sum of Squared Residuals around mean**
  - **24.81/109.62=22.6%**
- **Variability explained by regression**
  - **Complement of above**
  - $R^2=1-0.226=74.4\%$

Correlation

- **(Pearson's) Correlation, r: Related measure of strength *and direction* of association**
- **Square root of R^2, except that it has the same sign as the slope, b**
- **R^2 near 100%: r near 1 or -1**
- **R^2 near 0%: r near 0**
- **$r>0$: positive association**
- **$r<0$: negative association**
- **$r=0$: no association**
- **In our example, slope is positive so :**

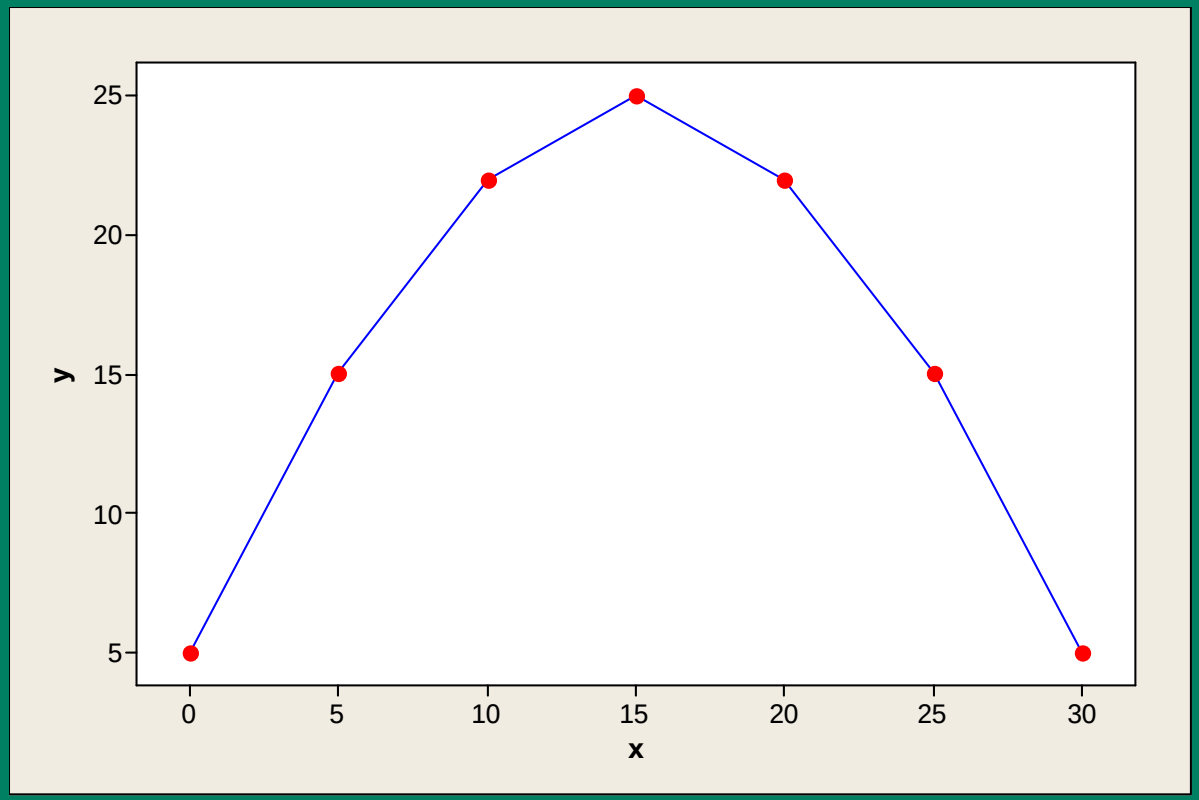
$$r = \sqrt{0.774} = 0.88$$

Correlation

- Correlation often used as a measure of association
- r is between -1 and 1
- Unit-less measure of association
- Slope is dependent on scale of X variable, so it cannot tell us strength of association
- r and R^2 measure strength of linear association

Nonlinear relationship

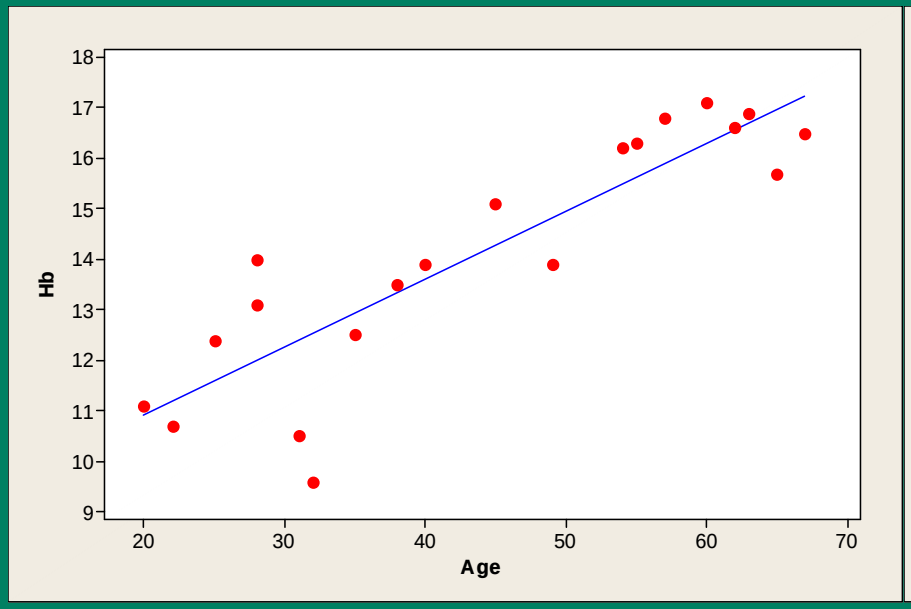
- **Strong nonlinear (quadratic) association, but no linear association: $b=0$, $r=0$, $R^2=0\%$**



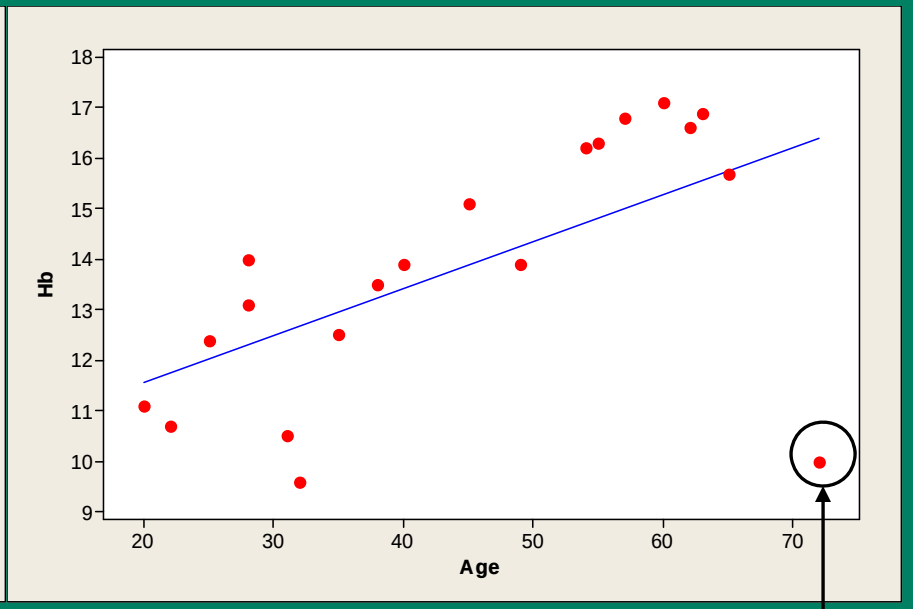
Outliers

- **Correlation and regression lines can be *sensitive to outliers*:**
  - **Outlier is a point which does not seem to fit the pattern of the rest of the points**
  - **Pull the regression line in direction of outlier**

Outliers (Example)



r=0.88



r=0.602

Outlier

Spearman's Correlation

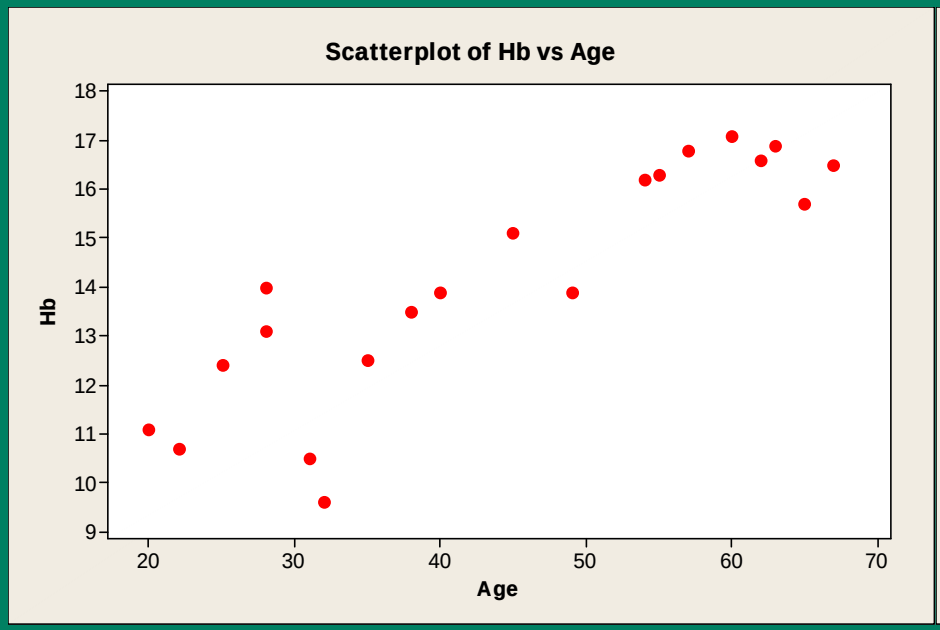
- **Spearman's correlation is an alternative measurement to describe the association between two numerical variables:**
  - **Less sensitive to outliers**
  - **Does not require assumption of normality for both variables, like Pearson's correlation does**
  - **Obtained by computing Pearson's correlation on the ranks of the two variables**

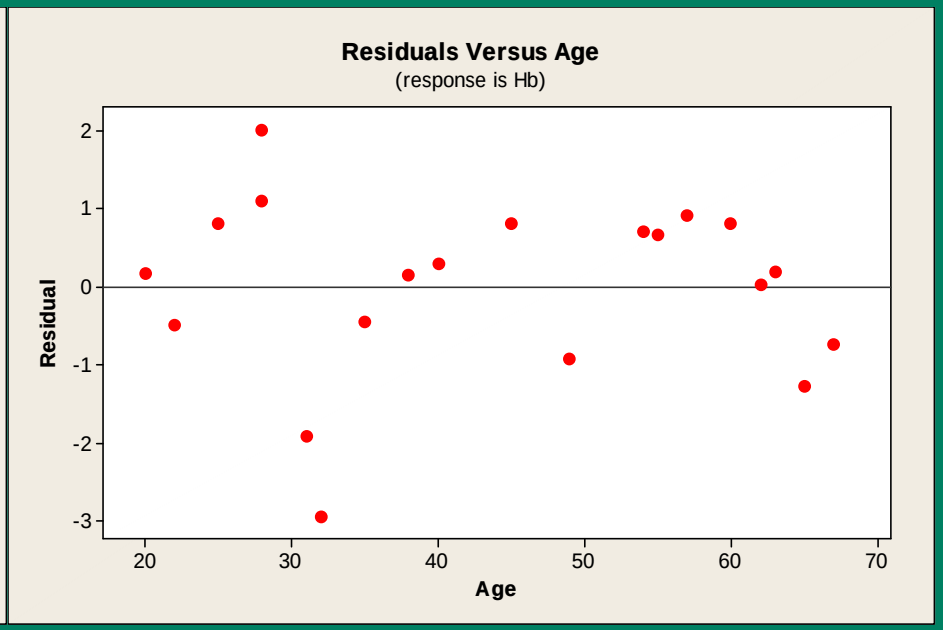
Assumptions of Linear Regression

- Relationship is approximately linear
- Prediction error is unrelated to predicted value
- Residuals about fitted line are normally distributed
- Residuals are independent

Checking linearity

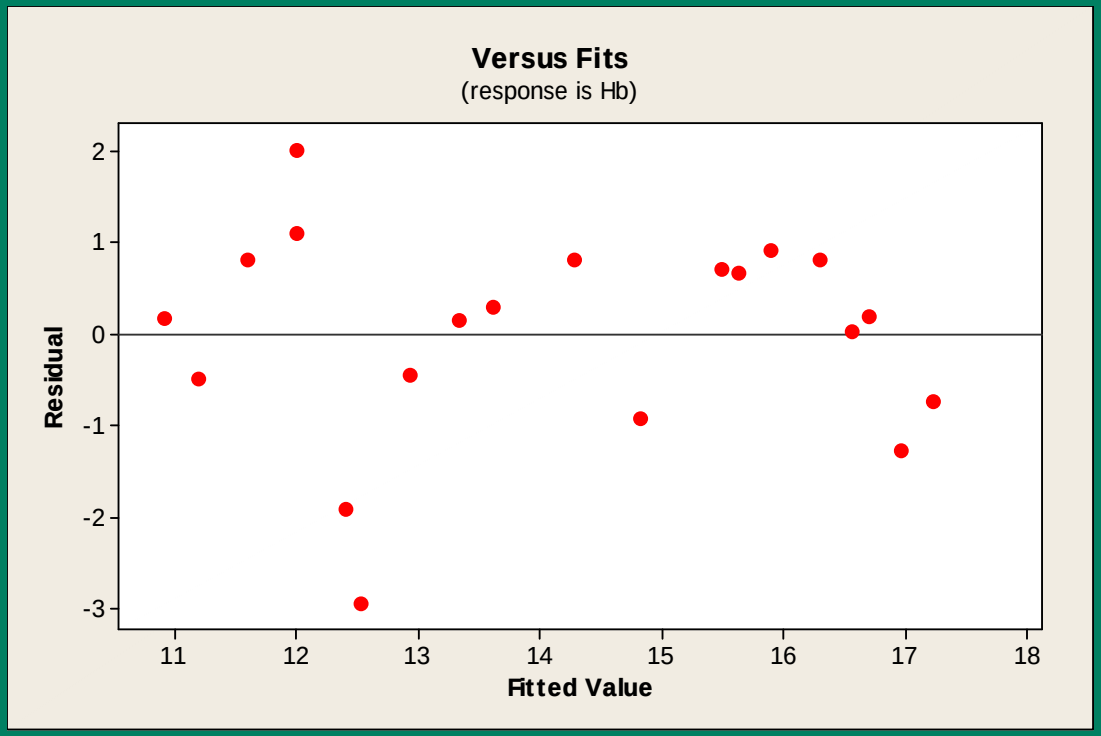
- **Scatter plot**: Does a straight line seem to fit
- Plot of residuals against predictor variable





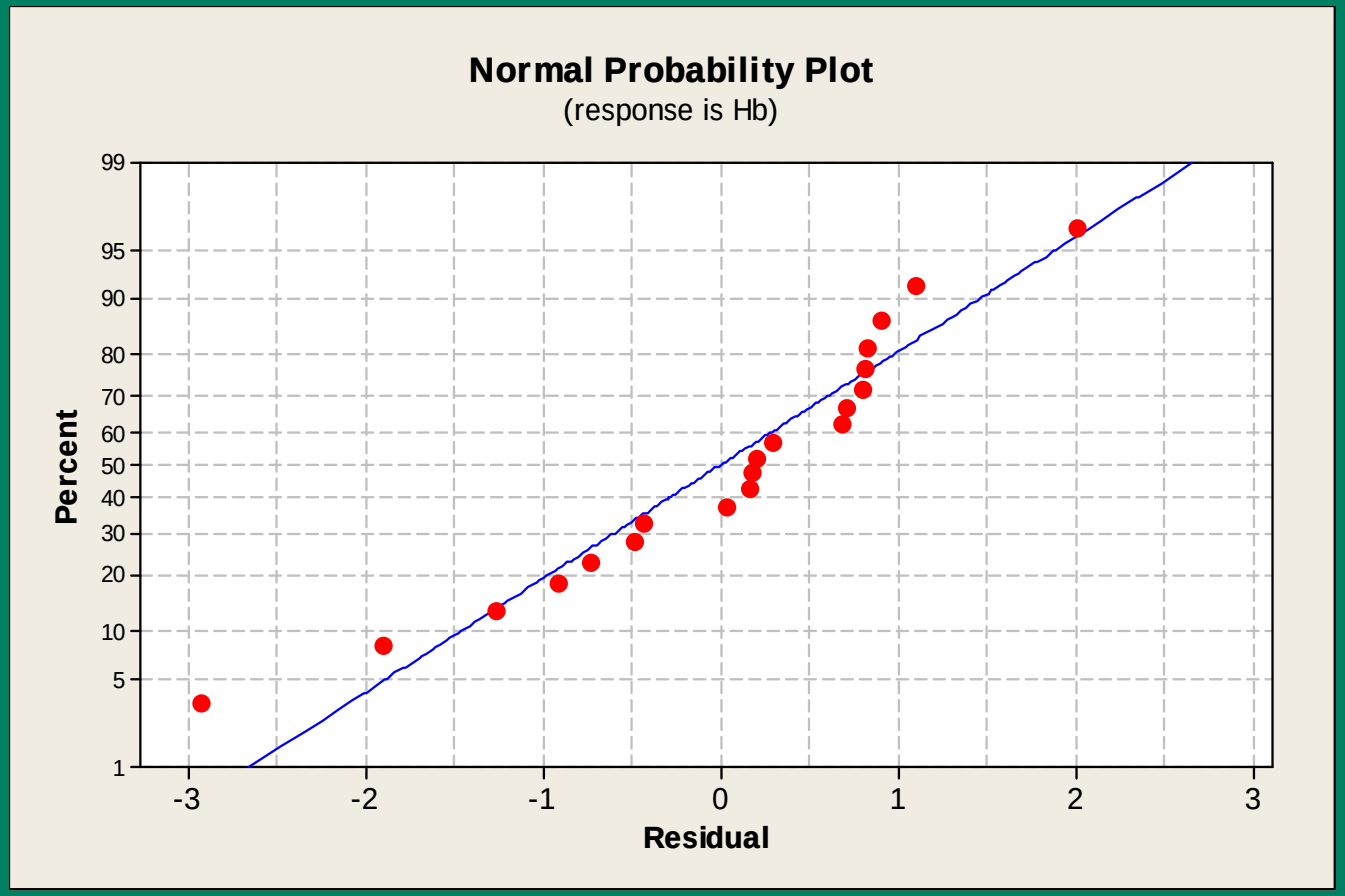
Checking constant variance

- Plot residuals against fitted values
- Does spread of data increase or decrease with larger fitted values (cone-shaped)?



Checking normality of residuals

- **Normal Plot**



Checking independence

- If observations are on separate individuals, this is not a concern
- Potential problem if:
  - observations are ordered over time
  - Multiple observations made on some individuals
- Plot residuals against time ordering
- (Not applicable in our example)

Sensitivity to assumptions

- **Linear relationship assumption is crucial, and results are sensitive to this**
  - **Transformation of *Y***
  - **Quadratic terms in *X***
  - **Categorization of *X***
- **Non-normality or non-constant variance unlikely to impact equation estimates much, but will have a modest impact on SE's and p-values**

Binary predictors

- ***Linear regression*** **is a method for relating a continuous outcome with one or more predictor variables**
  - **Predictors can be continuous or categorical**
- **Two groups with continuous outcome**
  - **Could simply compare outcome using t test**
  - **Example: Compare hemoglobin levels between women who are post-menopausal and those who are pre-menopausal**

Binary Predictors

- Regression formulation: set up **binary indicator variable**
  - Group 1: $X=0$ (e.g. pre-menopausal)
  - Group 2: $X=1$ (e.g. post-menopausal)
- Regression model: $E(Y)=a+b*X$
- Group 1: $E(Y)=a+b(0)=a$
- Group 2: $E(Y)=a+b(1)=a+b$
- b represents difference in means between group 2 and group 1
- In our example, b is the difference in mean *Hb* between post-menopausal and pre-menopausal women
- *t-test* is a special case of linear regression

Regression caveats

- Large b does not mean strong linear relationship
  - Strength of association is how closely the points fit the line, which is measured by r or R^2
- Very small p-value also does not mean strong linear relationship
  - Only indicates that slope is not 0
- Correlation does not imply causality
  - Only indicates association between variables, not a cause/effect relationship
  - May be confounding factors which account for the apparent relationship

References

- **Medical Statistics: A Textbook for the Health Sciences**, 4th Ed, by Campbell, et al., Wiley, 2007
- **Biostatistics**, 8th Ed, by Daniel, Wiley, 2005
- **Practical Statistics For Medical Research**, 2nd Ed, by Altman, Chapman & Hall, 2008.

Resources

- The Clinical and Translation Science Institute (CTSI) supports education, collaboration, and research in clinical and translational science:
  www.ctsi.mcw.edu
- The Biostatistics Consulting Service provides comprehensive statistical support

www.mcw.edu/biostatsconsult.htm

Free drop-in consulting

- **MCW/Froedtert/CHW:**
  - Monday, Wednesday, Friday 1 – 3 PM @ CTSI Administrative offices (LL772A)
  - Tuesday, Thursday 1 – 3 PM @ Health Research Center, H2400
- **VA: 1st and 3rd Monday, 8:30-11:30 am**
  - VA Medical Center, Building 70, Room D-21
- **Marquette: 2nd and 4th Monday, 8:30-11:30 am**
  - Olin Engineering Building, Room 338D