Module 2: Introduction to Statistics

Niko Kaciroti, Ph.D.
BIOINF 525 Module 2: W17
University of Michigan

Topic

- Dependence/Association/Relationship
  - Visual Display
    - Scatterplot
  - Covariance and Correlation
    - Pearson and Spearman Correlation

- Regression Model
  - Simple Linear Regression
  - Multiple Regression
    - Nonlinear (Quadratic) Relationship
    - Testing for Interactions

Dependence, Association, Relationship Between X and Y

- Let (x_1,y_1), (x_2,y_2),…., (x_n,y_n) be a sample of pairs of data of variables X (i.e. weight) and Y (i.e. height)

- Hypothesis: Is there a relationship between X and Y?
  - Can one variable predict variation in the second variable?
  - Do changes in X relate to changes in Y?

Dependence, Association, Relationship Between X and Y

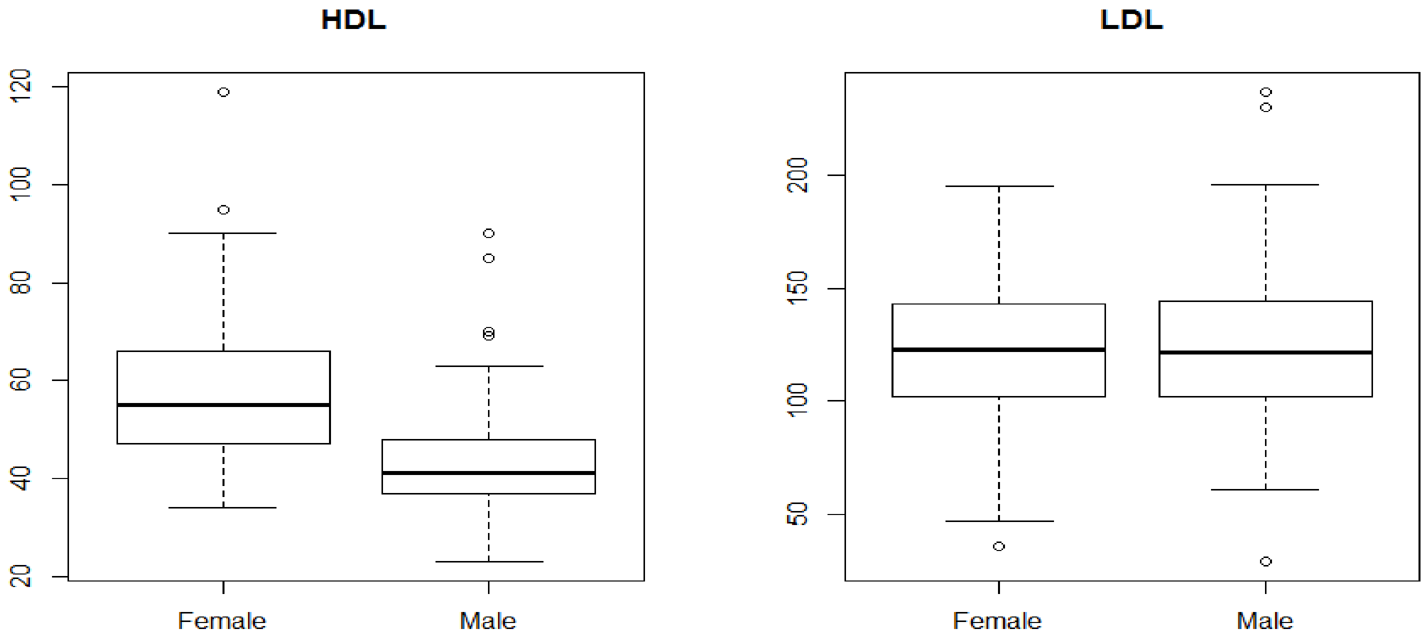
- Dependence between two variables X and Y roughly means that knowing the value of X provides some information about the value of Y

- Other terms used interchangeably for dependence are: Association between X and Y; Relationship between X and Y; X predicts Y

- Different measures of association are used depending if X or Y are discrete or continuous

Dependence, Association, Relationship
(X is Binary, Y is Continuous)

- X is a group variable (Male/Female), Y is Continuous (HDL or LDL).
  - Group differences are a form of dependence



Does HDL depend on the gender of a subject? How about LDL?

Dependence, Association, Relationship
(X is Binary, Y is Binary)

- X is a group variable (Male/Female), Y is binary (Yes/No).
  - OR is a measure of dependence for binary data:

$$\text{OR} = \frac{ODD_{Male}}{ODD_{Female}}$$

  - E.g. Does having HDL <= 40 depend on the gender of the patient? Or, equivalently, are the Odds different between males and females?

$$ODD_{Male}(\text{HDL} \leq 40) = 0.82$$

$$ODD_{Female}(\text{HDL} \leq 40) = 0.11$$

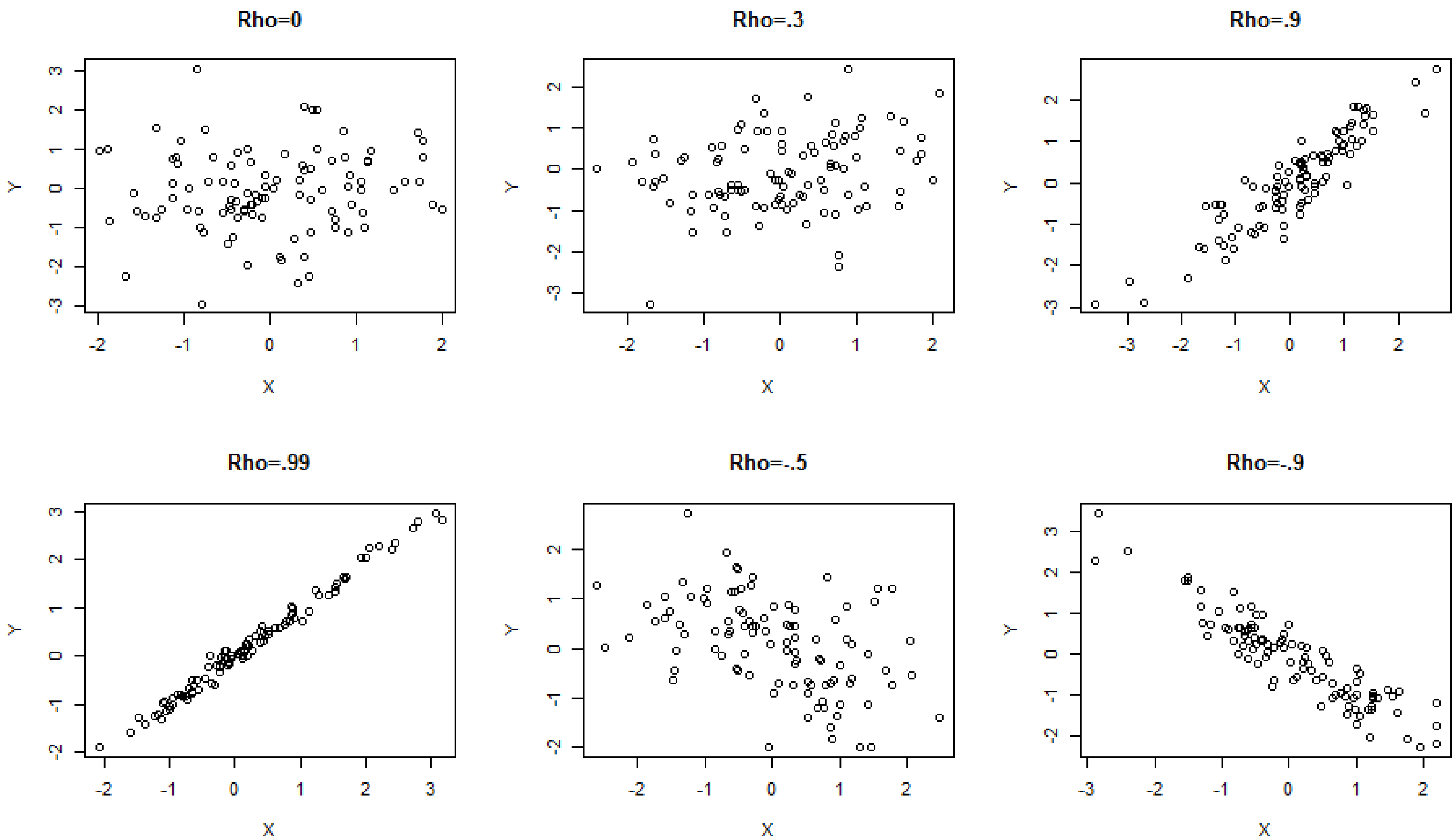
$$\text{OR} = \frac{.82}{.11} = 7.5$$

Dependence, Association, Relationship
(X and Y are Continuous)

- Association between two continuous variables X and Y implies that changes in X are related with changes in Y

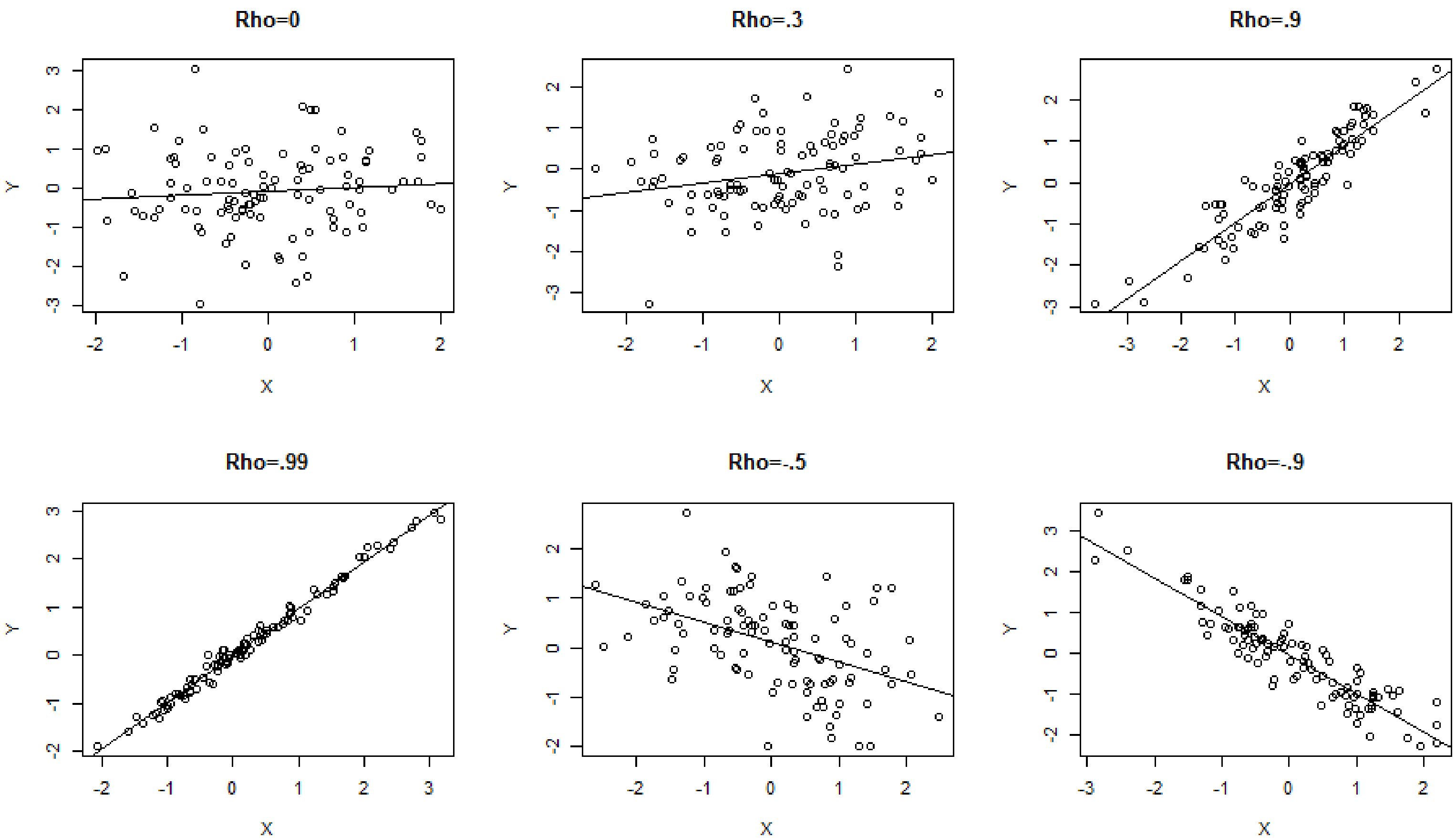
- Scatterplot can be initially used to visually explore for possible associations
  - A scatterplot is a graphical display of the data by plotting pairs of x and y
  - The presence of any pattern indicates dependence

Scatterplot Examples



Which scatterplot indicate strongest dependence?

Scatterplot Examples



How to Measure the Association For Continuous X and Y

- The scatterplot can help in identifying patterns and the direction of an association. However, it does not provide a numerical estimate of the association

- **Covariance** is used to capture the linear association and the direction of the association (positive or negative) between two variables X and Y

Covariance Between Two Variables

- Let (x_1,y_1), (x_2,y_2),…., (x_n,y_n) be a sample of pairs of data of variables X and Y. The covariance is defined as:

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)}{N}$$

$$\widehat{Cov}(X,Y) = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{N-1}$$

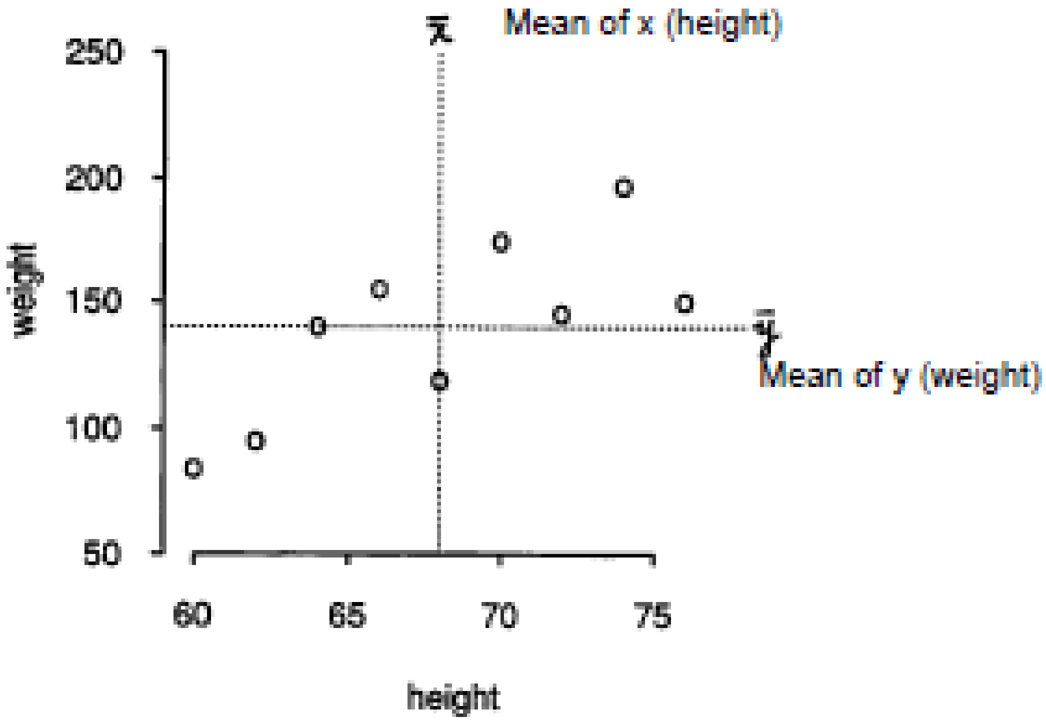
$$Var(X) = Cov(X,X) = \frac{\sum_{i=1}^{n}(x_i-\mu_x)^2}{N}$$

Covariance Between Two Variables

- Example data on height and weight for 9 people. Are they related?

Height	Weight
60	84
62	95
64	140
66	155
68	119
70	175
72	145
74	197
76	150

Scatterplot: Plot of Height vs. Weight



Intuitive Interpretation of Covariance

$$\widehat{Cov}(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{N-1}$$

- The covariance can be viewed intuitively as a sum of "matches" (or "mismatches") in terms of a subject being on the same side of the mean for each variable X or Y

- A "match" is when $x_i - \overline{x}$ and $y_i - \overline{y}$ have the same sign.
  - For example, if x_i is greater than the mean ($x_i - \overline{x} > 0$) then y_i is also greater than the mean ($y_i - \overline{y} > 0$)

- A "mismatch" is when $x_i - \overline{x}$ and $y_i - \overline{y}$ have the opposite sign.
  - If x_i is above the mean ($x_i - \overline{x} > 0$) and y_i is below the mean ($y_i - \overline{y} < 0$), or vice versa

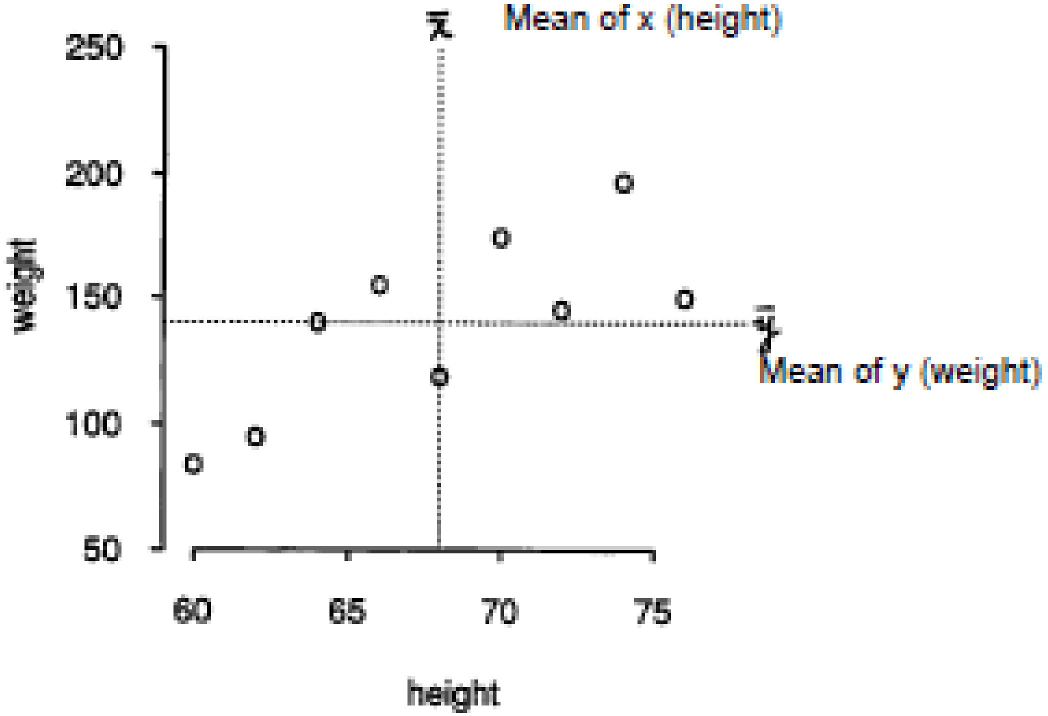
Intuitive Interpretation of Covariance

$$\widehat{Cov}(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{N-1} \qquad (1)$$

For a particular subject *i*, a "match" leads to a positive product in Equation (1), whereas a "mismatch" leads to a negative product

- If Eq. (1) is dominated by "matches", then Cov(X,Y) > 0 and the association between X and Y is said to be positive
- If Eq. (1) is dominated by "mismatches", the Cov(X,Y) < 0 and the association is negative
- If there are more or less the same "matches" and "mismatches", then there is no relationship between X and Y

Scatterplot: Plot of Height vs. Weight



How many “mismatched” points are in the plot?

Covariance Between Two Variables

- Example data on height and weight for 9 people. Are they related?

Height	Weight
60	84
62	95
64	140
66	155
68	119
70	175
72	145
74	197
76	150

Covariance Between Two Variables

- Example data on height and weight for 9 people. Are they related?

	Height	Weight	Height – 68	Weight - 140
	60	84	-8	-56
	62	95	-6	-45
	64	140	-4	0
	66	155	-2	15
	68	119	0	-21
	70	175	2	35
	72	145	4	5
	74	197	6	57
	76	150	8	10
Mean	68	140		

Covariance Between Two Variables

- Example data on height and weight for 9 people. Are they related?

	Height	Weight	Height – 68	Weight - 140	Product
	60	84	-8	-56	448
	62	95	-6	-45	270
	64	140	-4	0	0
	66	155	-2	15	-30
	68	119	0	-21	0
	70	175	2	35	70
	72	145	4	5	20
	74	197	6	57	342
	76	150	8	10	80
Mean	68	140			Cov(H,W)=1200/8=150

Properties of Covariance

- *Cov(X+a,Y)=Cov(X,Y)*

$$Cov(X+a,Y) = \frac{\sum_{i=1}^{n}(x_i + a - (\mu_x + a))(y_i - \mu_y)}{N}$$

$$= \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{N} = Cov(X,Y)$$

- If there is a systematic error when measuring X or Y the covariance (association) is not effected
  - Examples of systematic error are when the measurement instruments are not calibrated; Different labs may have different calibrations

- “Good” property: It allows replication of the results from different labs etc.

Properties of Covariance

- *Cov(aX,bY)=a*b*Cov(X,Y)*

$$Cov(aX,bY) = \frac{\sum_{i=1}^{n}(ax_i - a\mu_x)(by_i - b\mu_y)}{N}$$

$$= \frac{a * b * \sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{N} = a * b * Cov(X,Y)$$

- The covariance will change if X or Y are multiplied by a scalar
- “Bad” property: The covariance will change if the units change (e.g. from inches to feet). However the associations should not change regardless of the unit of measure

Correlation of Two Variables
(Pearson Correlation)

- Correlation is derived by standardizing the covariance, so its value <u>does not</u> depend on the unit of measurement

$$\rho = corr(x,y) = \frac{Cov(X,Y)}{SD(X) * SD(Y)} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

$$corr(ax, by) = \frac{\sum_i^n (ax_i - \overline{ax})(by_i - \overline{by})}{\sqrt{\sum_i^n (ax_i - \overline{ax})^2 \sum_i^n (by_i - \overline{by})^2}} = \frac{ab \sum_i^n (x_i - \overline{x})(y_i - \overline{y})}{ab\sqrt{\sum_i^n (x_i - \overline{x})^2 \sum_i^n (y_i - \overline{y})^2}} = corr(x,y)$$

- The correlation between x and y is the same regardless of what unit is used for x and y

Correlation of Two Variables
(Pearson Correlation)

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\mathrm{SD}(X)\mathrm{SD}(Y)}$$

- The correlation coefficient ρ, is referred to as the Pearson correlation. It is a measure of the <u>linear</u> relationship between X and Y

- Correlation can be positive or negative: $-1 \leq \rho \leq 1$
  - $\rho > 0$: Increases on X are related with increases on Y
  - $\rho < 0$: Increases on X are related with decreases on Y
  - $\rho = 0$: No association between X and Y
  - $|\rho| = 1$: Perfect correlation, Y is a linear transformation of X, Y=a+bX

If $\rho = -1$, is $b > 0$ or $b < 0$?

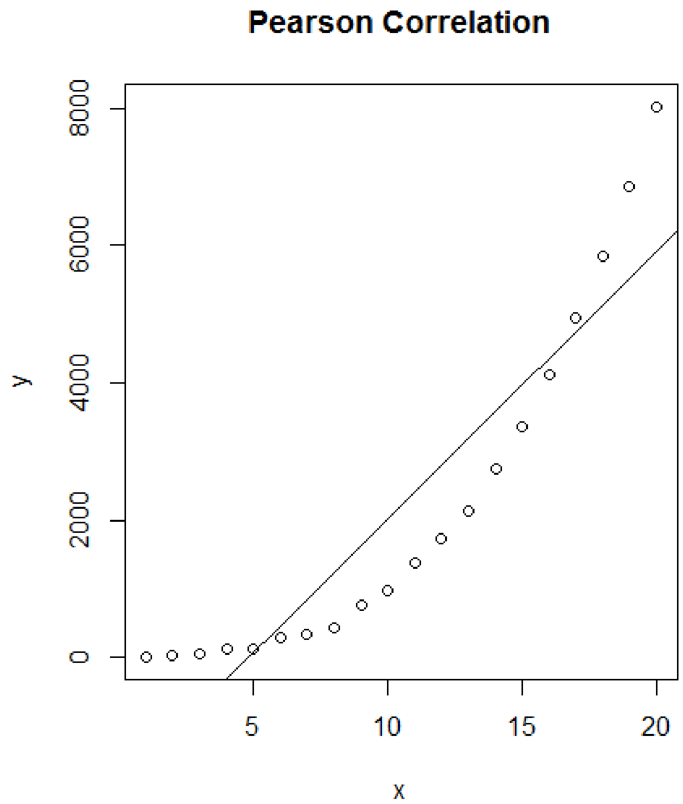
Correlation of Two Variables
(Spearman Correlation)

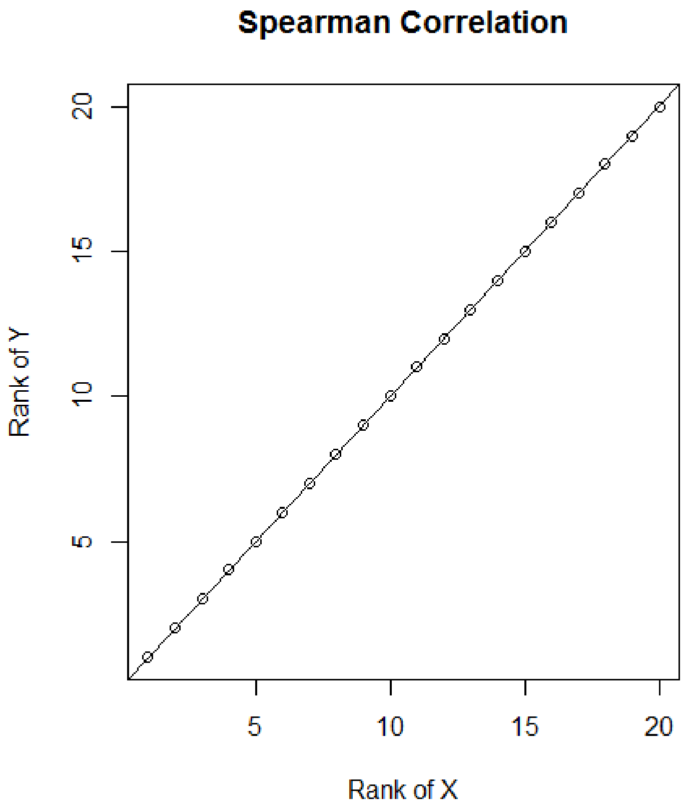
- Spearman correlation is a <u>nonparametric</u> correlation that does not depend on the linearity between X and Y. It is also not affected by outliers

- For each pair, x and y, calculate their corresponding ranks, rank(x) and rank(y). The Spearman correlation is the same as the Pearson correlation, but applied on the ranks of X and Y:

$$corr_S(X,Y) = corr_P\ (rank(X), rank(Y))$$

Pearson vs. Spearman Correlation

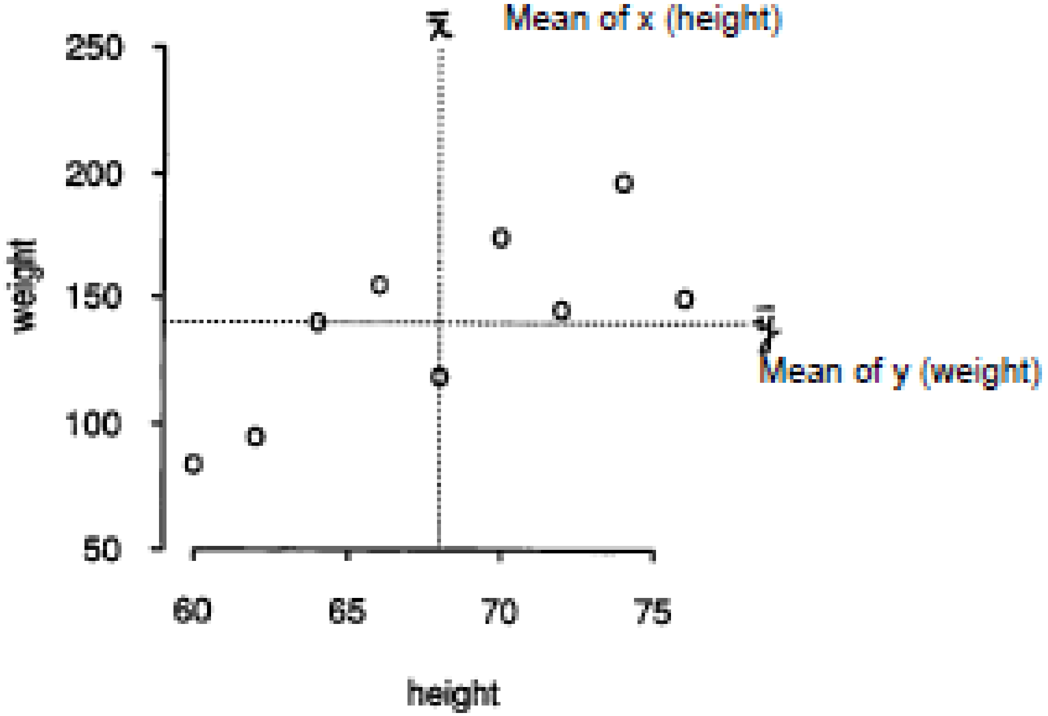




Pearson r=.92

Spearman r=1

Test for Correlation



The estimate for correlation ρ is $r = .76$, with some margin of error.
How do we test if ρ is different from 0?

Test for Correlation

- Testing the null hypothesis that X is not associated with Y: H_0: $\rho = 0$ vs. H_A: $\rho \neq 0$

- The following test is used for testing H_0:

$$t_{n-2} = \frac{r}{se(r)} = \frac{r}{\sqrt{\frac{1-r}{n-2}}}$$

- If data are normally distributed, then t_{n-2} follows a t-distribution with n-2 degrees of freedom. The usual p-value < 0.05 criteria is then used to reject H_0

Test for Correlation in R

- cor.test(height,weight)

Pearson's product-moment correlation

t = 3.0805, df = 7, p-value = 0.0178

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval: 0.1904203, 0.9460844

sample estimates: cor

0.7586069

Test for Difference on Correlation Coefficients By Group

- Another question of interest is for testing whether the relationship between X and Y is different by groups.
  - E.g. Correlation between weight and height is different for males (ρ_m) vs. females (ρ_f):

$$H_0: \rho_m = \rho_f \quad \text{vs. } H_A: \rho_m \neq \rho_f$$

- We will test this hypothesis later using the regression model approach with interaction terms

Topic

- Dependence/Association/Relationship
  - Visual Display
    - Scatterplot
  - Covariance and Correlation
    - Pearson and Spearman Correlation

- **Regression Model**
  - **Simple Linear Regression**
  - Multiple Regression
    - Nonlinear (Quadratic) Relationship
    - Testing for Interactions

Correlation vs. Regression Model

- Correlation is a measure of association
  - It shows: if X and Y are related; the magnitude of the relationship; and its direction
  - However, correlation does not show how to predict Y from X (or X from Y)

- Regression is a modeling technique
  - It builds models for the variable Y as a function of one (or more) variable X
  - It measures the association between X and Y, <u>and</u> also can be used to predict Y from X

Simple Linear Regression

- Simple linear regression model describes the value of variable Y as a linear function of another variable X plus some error terms

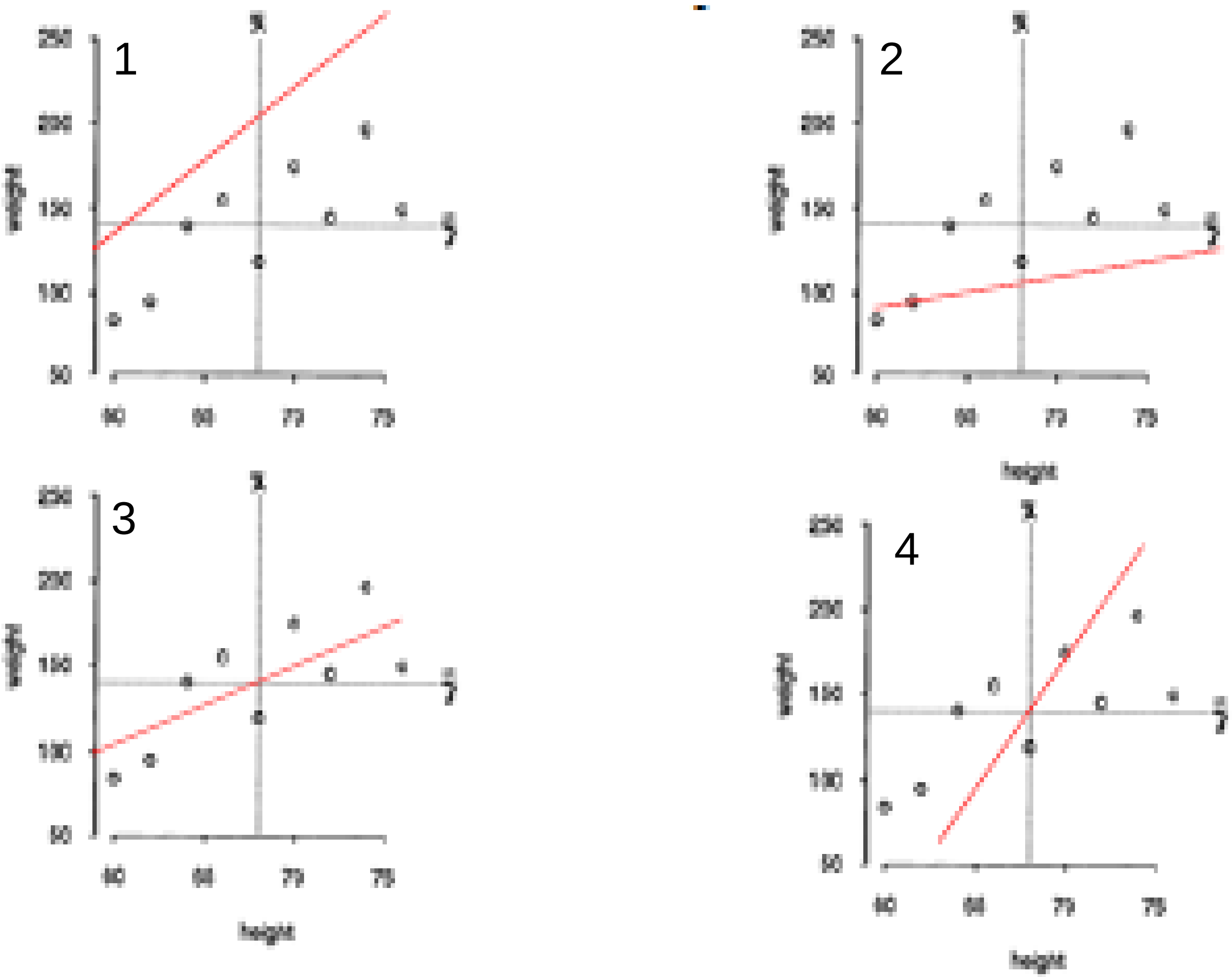
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- When *X* may explain changes in *Y*, then *X is* called an **explanatory** variable (or **predictor** variable, or **independent** variable, or **covariate**)

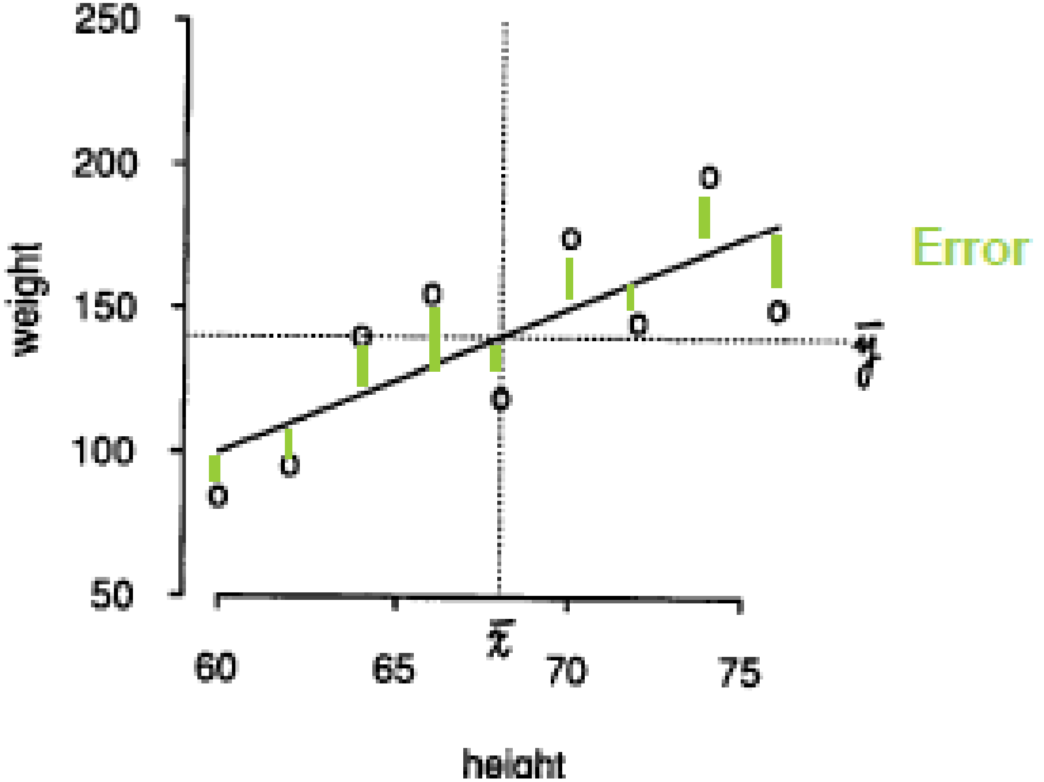
- The variable *Y* is called the **response** variable (or the **outcome** variable, or the **dependent** variable)

- $\varepsilon_i \sim N(0, \sigma^2)$ is the **error** term (or **residual**)

What Line Best Describes the Relationship of Weight and Height?



How Far is the Observed Weight from the Predicted Weight?



Estimating the Line That Best Describes the Relationship of Weight and Height?

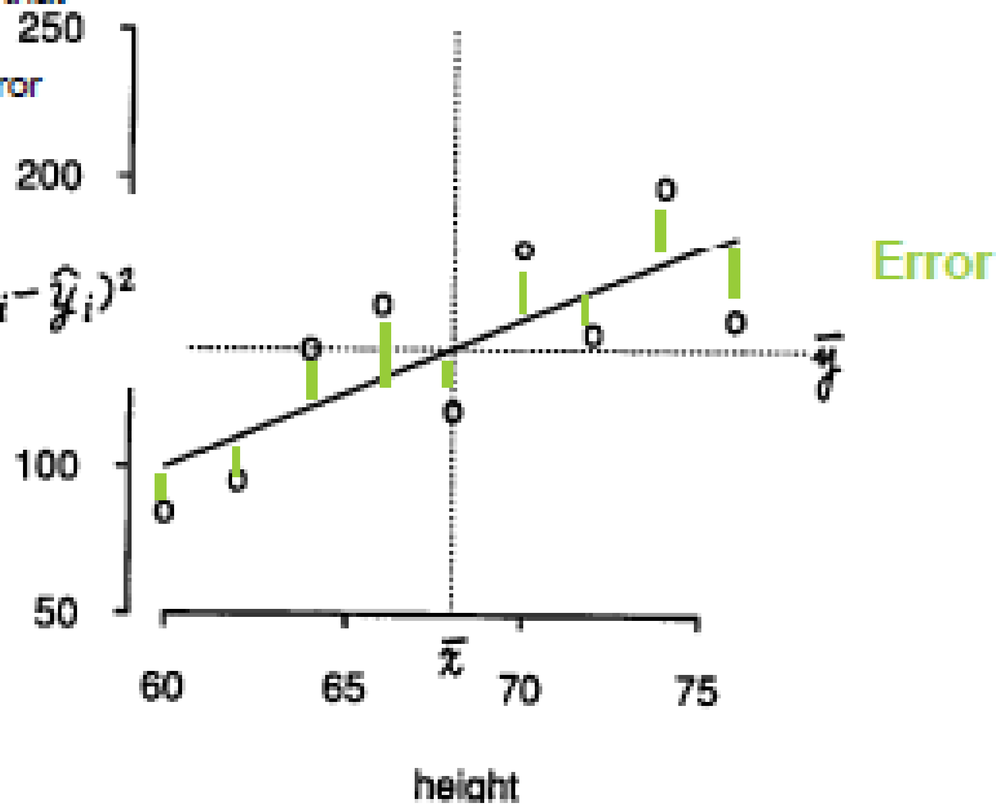
- Find the line for which the predicted values ($\hat{Y}_i$) are closest to the actual values (Y_i)

- First, for each subject *i* define the error between the predicted value and the actual value, ($\hat{Y}_i - Y_i$), then minimize the sum of errors across all subjects

Estimating the Line That Best Describes the Relationship of Weight and Height?

Goal: make a line that minimizes
Sum of squares error

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$



Least Squares Estimate for Regression Parameters β_0 and β_1

- Least squares is a technique used to estimate parameters in a regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- Least squares minimizes the sum of squares for the residuals:

$$\text{SSR} = \sum_{i=1}^{n} {\varepsilon_i}^2 = (Y_1 - \beta_0 - \beta_1 X_1)^2 + (Y_2 - \beta_0 - \beta_1 X_2)^2 + \ldots + (Y_n - \beta_0 - \beta_1 X_n)^2$$

Least Squares Estimate for Regression Parameters β_0 and β_1

- The "least squares estimate" are given by the values of b_0 and b_1 as follows:

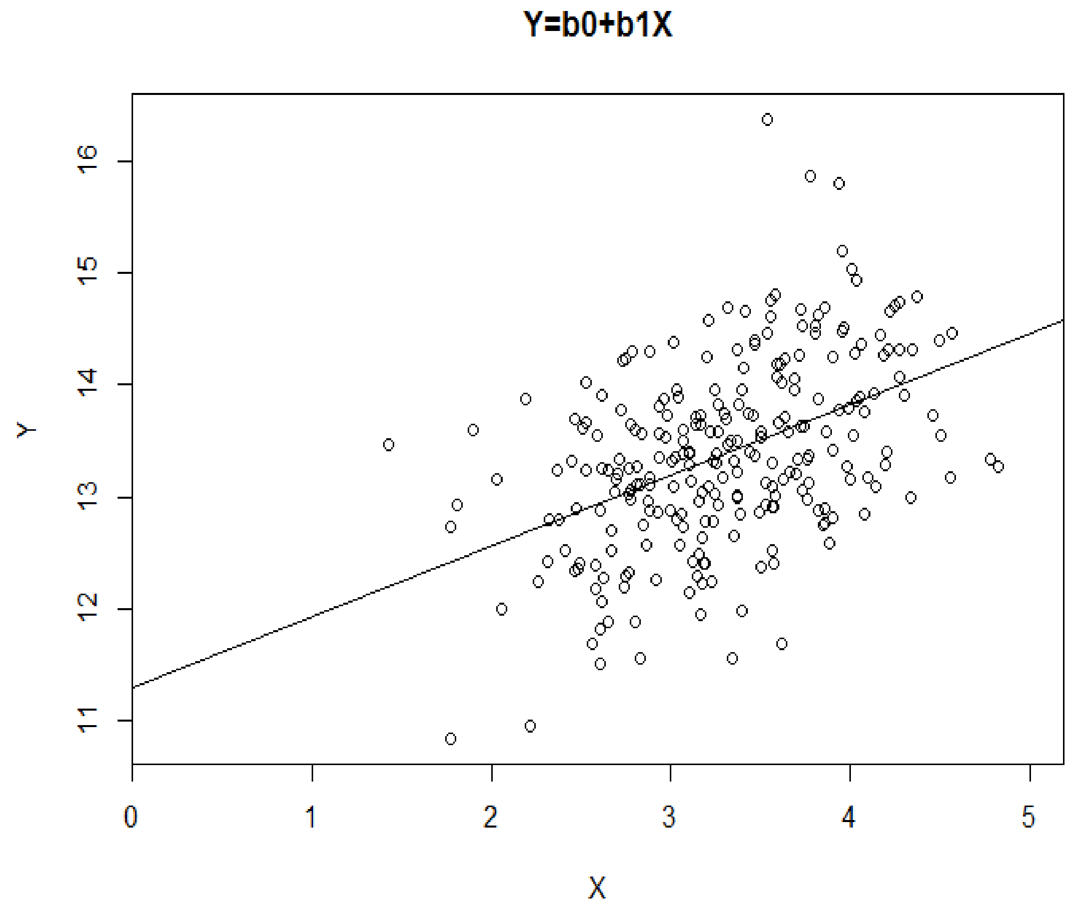
$$b_1 = \frac{\sum_i Y_i(X_i - \overline{X})}{\sum_i (X_i - \overline{X})^2} = \widehat{corr}(Y, X) * \frac{\widehat{SD}(Y)}{\widehat{SD}(X)}$$

$$b_0 = \overline{Y} - b_1\overline{X}$$

- After we have calculated the estimates, b_0 and b_1, the "fitted values" (or predicted values) for Y are given by:

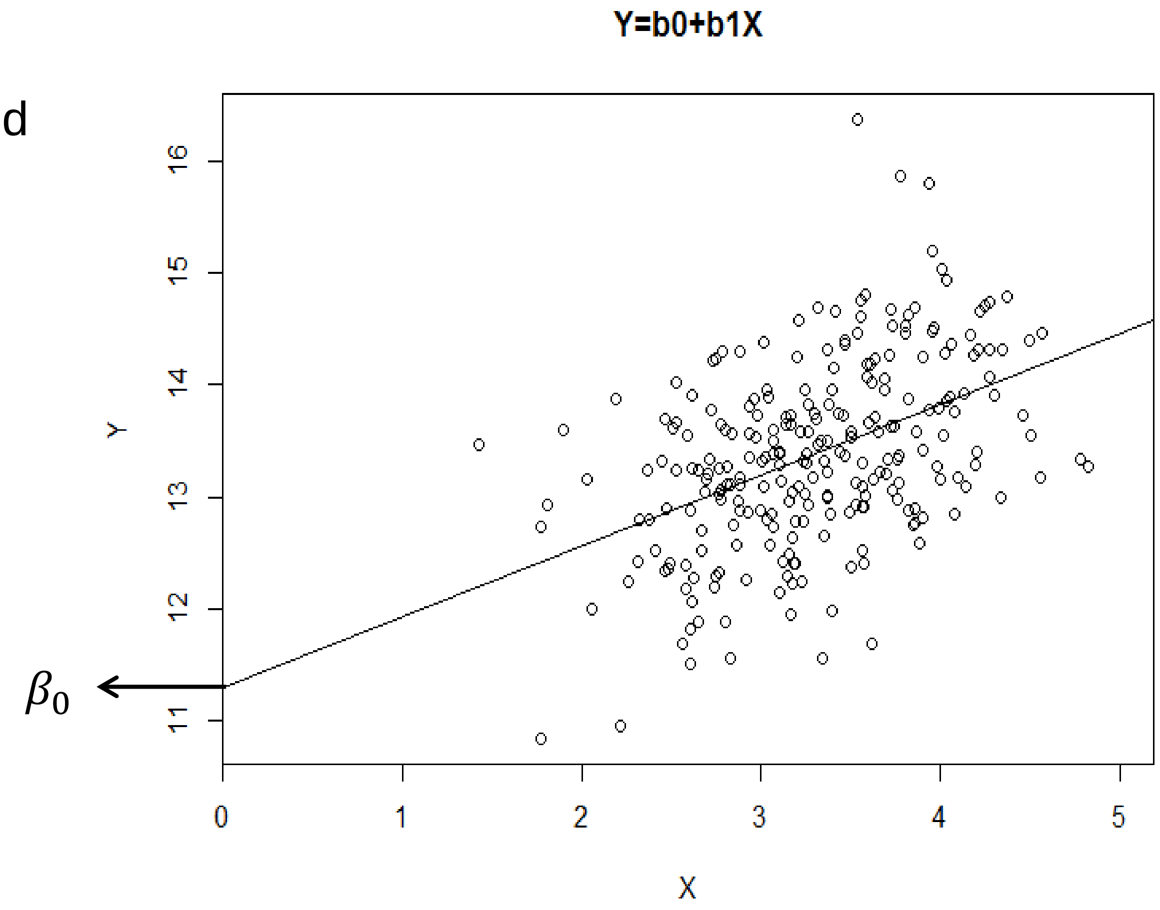
$$\hat{Y}_i = b_0 + b_1 X_i$$

Geometric Interpretation of the Regression Parameters Intercept (β_0) and Slope (β_1)



Geometric Interpretation of the Regression Parameters Intercept (β_0) and Slope (β_1)

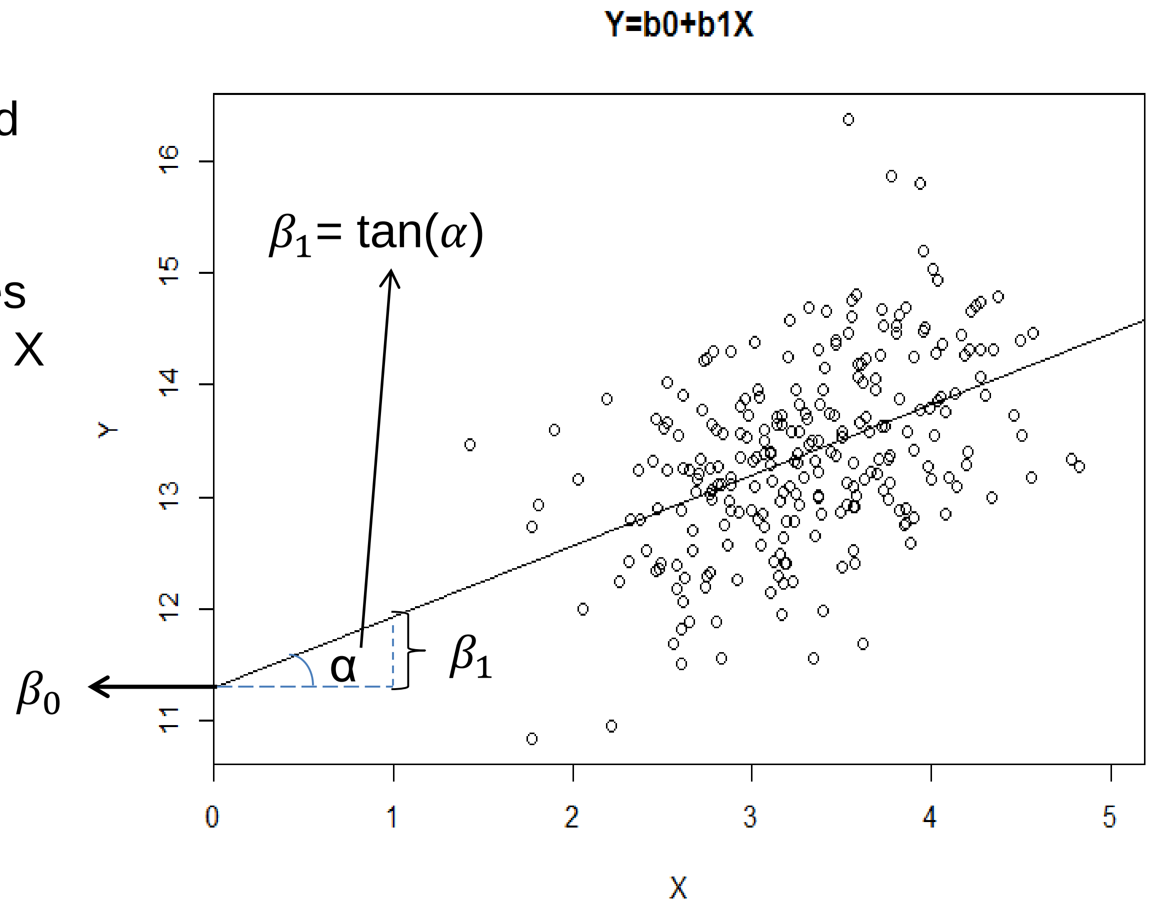
<u>Intercept</u>: $\beta_0$is the expected value of Y when X=0



Geometric Interpretation of the Regression Parameters Intercept (β_0) and Slope (β_1)

Intercept: β_0 is the expected value of Y when X=0

Slope: β_1 measures changes in Y for one unit increase in X



Testing for Relationship Between X and Y Using Regression Model

- Test whether Y is related to X: H_0: $\beta_1 = 0$ vs. H_A: $\beta_1 \neq 0$.

- The following test is used for testing H_0:

$$t_{n-1} = \frac{b_1}{se(b_1)}$$

- When $\varepsilon_i \sim N(0, \sigma^2)$, then t_{n-1} follows a t-distribution with n-1 degrees of freedom. The usual p-value < 0.05 criteria is then used to reject H_0

Simple Linear Regression in R

- summary(lm(weight~height))

- Coefficients:

	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	-200.000	110.690	-1.807	0.1137
height	5.000	1.623	3.080	0.0178 *

R-squared: 0.5755

R-Square: Measure of Goodness of Fit of a Regression Model

- A regression model provides a "good" fit if the predicted values $\hat{Y}$ are closely related to the actual values Y

- R^2 measures the goodness of fit. It is equal to the squared correlation between Y and $\hat{Y}$ (or Y and X):

$$R^2 = {r_{y\hat{y}}}^2 = {r_{yx}}^2$$

Assumptions for Linear Regression Model

- There are several assumptions made in a linear regression model:

$$Y_i=\beta_0+\beta_1 X_i + \varepsilon_i$$

  - The observations are independent
  - The relationship between x and y is linear
    - Scatterplot
  - $\varepsilon_i \sim N(0, \sigma^2)$ are normally distributed with zero mean and constant variance
    - Q-Q Plot, Shapiro-Wilk's test

Topic

- Dependence/Association/Relationship
  - Visual Display
    - Scatterplot
  - Covariance and Correlation
    - Pearson and Spearman Correlation

- Regression Model
  - Simple Linear Regression
  - **Multiple Regression**
    - **Nonlinear (Quadratic) Relationship**
    - **Testing for Interactions**

Multiple Regression

- Multiple regression model is an extension of the simple linear regression. It permits any number of predictor variables. Multiple regression simply means “multiple predictors”

- The model is similar to the case with one predictor; it just has more X’s and β’s.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

β_0: Intercept

β_k: Slope for X_k, for k=1,2,…,p

ε_i: Error term (residual)

Least Square Estimate

- The least square estimates for multiple regression are defined in the same way, by minimizing the “residuals” $\varepsilon_i = Y_i - \beta_0 - \beta_1 X_i - \beta_2 X_{2i} - \cdots - \beta_p X_{pi}$. Thus, the parameter estimates are chosen to minimize the “sum of squared residuals”:

$$\text{SSR} = \sum_{i=1}^{n} (\underbrace{Y_i - \beta_0 - \beta_1 X_i - \beta_2 X_{2i} - \cdots - \beta_p X_{pi}}_{\varepsilon_i{}^2})^2$$

Features of Multiple Regression

- Multiple regression model improves the prediction of Y by using multiple variables

- It is used to estimate partial association of X and Y. That is, how much X contributes in predicting Y that is unique to X and does not overlap with other covariates

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$$

  – β_1, is unadjusted/overall association between X_1 and Y

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \varepsilon_i$$

  – β_1 is the adjusted association between X_1 and Y, adjusted for X_2,…, X_p

- R^2 is used to measure the overall association of X_1,X_2,…, X_p with Y

Testing for Relationship Between X_k and Y Using Multiple Regression

- Test for H_0: $\beta_k = 0$ vs. H_A: $\beta_k \neq 0$.

- The following test is used:

$$t_{n-1} = \frac{b_k}{se(b_k)}$$

- If $\varepsilon_i \sim N(0, \sigma^2)$, then t_{n-1} follows a t-distribution with n-1 degrees of freedom. The p-value < 0.05 criteria is then used to reject H_0

Multiple Regression Example in R
(TROPHY Data)

- We want to test whether LDL, Insulin, Age, and DBP are related to or predict BMI24?
  - Then fit the following multiple regression

$$BMI24_i = \beta_0 + \beta_1 LDL_i + \beta_2 Insulin_i + \beta_3 Age_{1i} + \beta_4 DBP_i + \varepsilon_i$$

Multiple Regression Example in R
(TROPHY Data)

R Output:

Coefficients:	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	22.1	7.24	3.0	0.00285 **
LDL	0.03	0.014	2.33	0.02189 *
Insulin	0.25	0.05	4.56	1.32e-05 ***
Age	-0.05	0.059	-0.86	0.39085
DBP0	0.036	0.078	0.46	0.64564

Multiple R-squared: 0.2101

Adjusted R-squared: 0.1814

F-statistic: 7.314 on 4 and 110 DF, p-value: 2.92e-05.

Interpretation of R-Square

- The total sum of squares for Y, which is a measure of variation, can be decomposed as follows:

$$\underbrace{\sum_i^n (y_i - \overline{y})^2}_{SS_{Tot}} = \underbrace{\sum_i^n (y_i - \hat{y}_i)^2}_{SS_{err}} + \underbrace{\sum_i^n (\hat{y}_i - \overline{y})^2}_{SS_{Reg}}$$

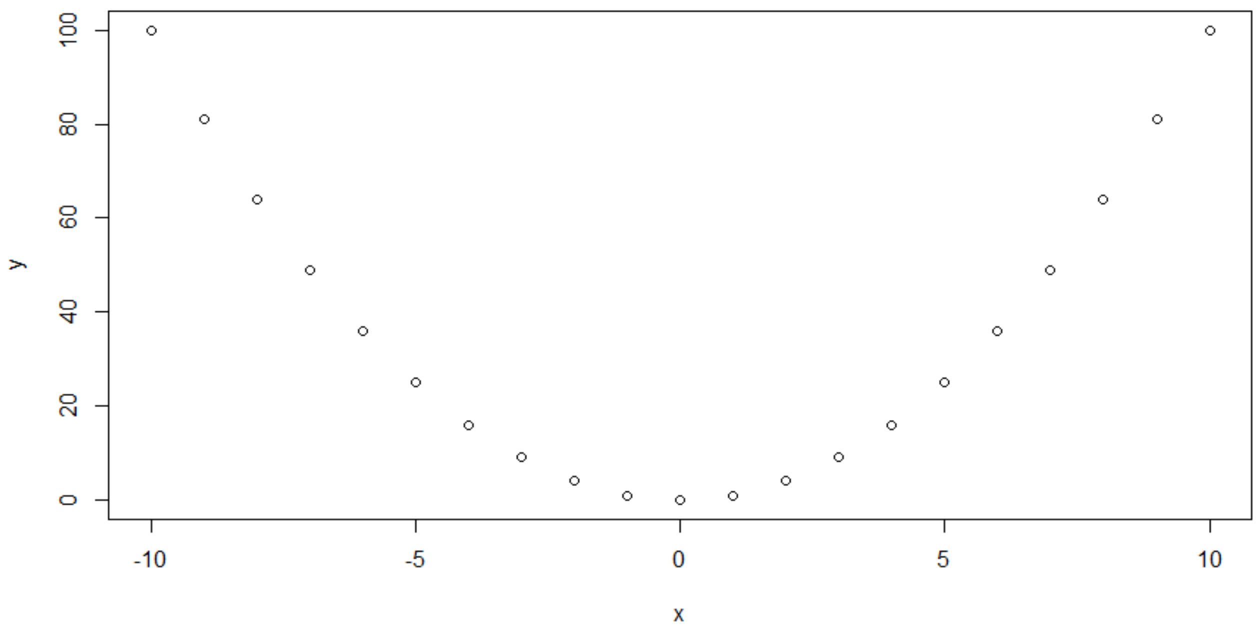
$R^2 = \frac{SS_{Reg}}{SS_{Tot}}$: It is the proportion of the variance on Y explained by the model

1-$R^2 = \frac{SS_{err}}{SS_{Tot}}$: It is the proportion of the unexplained variance

- R^2=.21, means that 21% of the variation on BMI24 is explained by the model or by LDL, Insulin, Age, and DBP

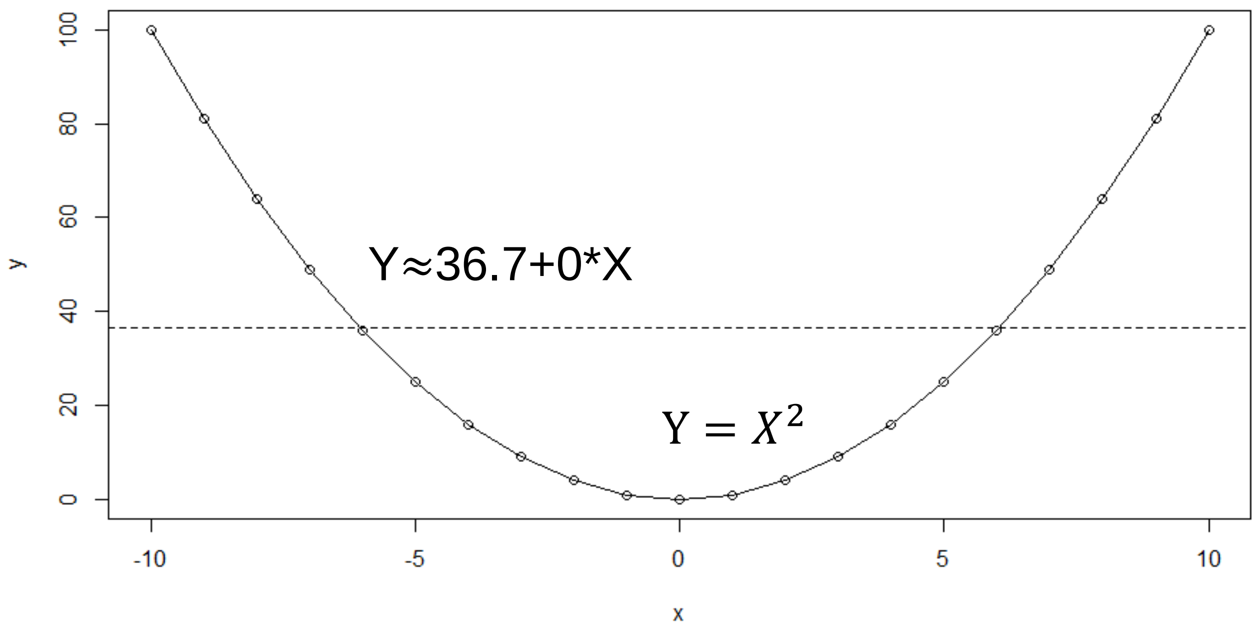
Nonlinear Scatterplot

What do you do if the scatterplot of the raw data suggests that the association between Y and X is not linear, (i.e. Y≈ X^2)?



Nonlinear Scatterplot

What do you do if the scatterplot of the raw data suggests that the association between Y and X is not linear, (i.e. Y≈ X^2)?



Nonlinear (Quadratic) Regression Model

- Linear regression can be extended by including a quadratic term. Then, multiple regression can be used to fit a quadratic regression:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X^2{}_{1i} + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

- Along similar lines, you could include X^3 or log(X), etc., depending on the type of relationship between X and Y. Here β_2 is the curvature coefficient

- H_0: β_2= 0 vs. H_A: $\beta_2 \neq 0$. If H_0 is rejected, the relationship between X and Y is not linear

Testing if the Association Between X and Y Varies by Group

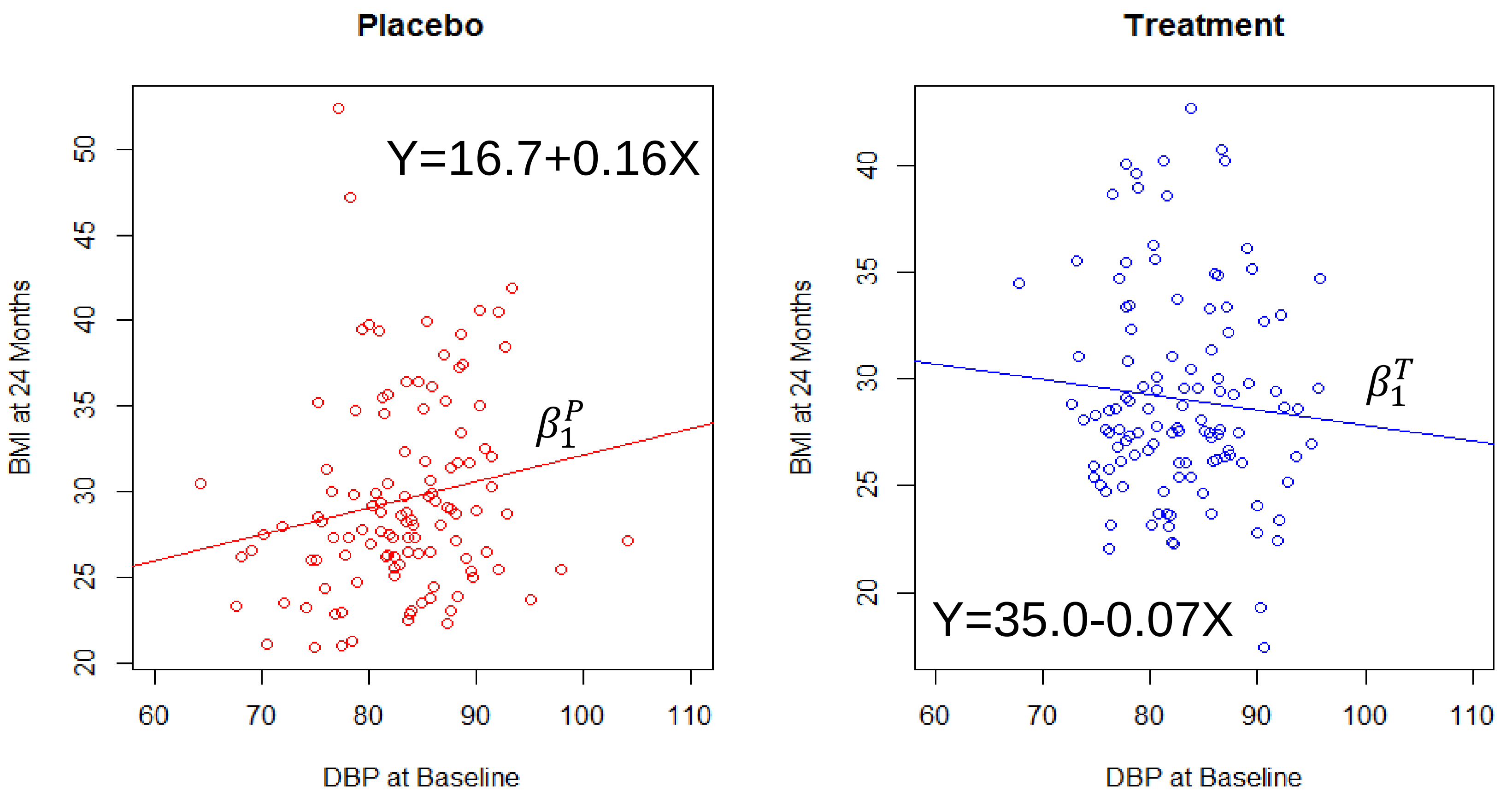
- Q: Is the association between DBP and BMI24 different between subjects in the Treatment group versus subjects in the Placebo group?

- First, fit separate models by group:

  – Treatment Group: $BMI24_i = \beta_0^T + \beta_1^T DBP_i + \varepsilon_i$

  – Placebo Group: $BMI24_i = \beta_0^P + \beta_1^P DBP_i + \varepsilon_i$

Subgroup Analysis: Model the Relationship of X on Y for Each Treatment Group



How to test $H_0\colon \beta_1^T = \beta_1^P$?

Interactions

- **Interaction term** is defined as the product of two predictors (*i.e.* Trt x DBP). We will fit the following multiple regression:

$$BMI24_i = \beta_0 + \beta_1 Trt_i + \beta_2 DBP_i + \beta_3 Trt_i * DBP_i + \varepsilon_i$$

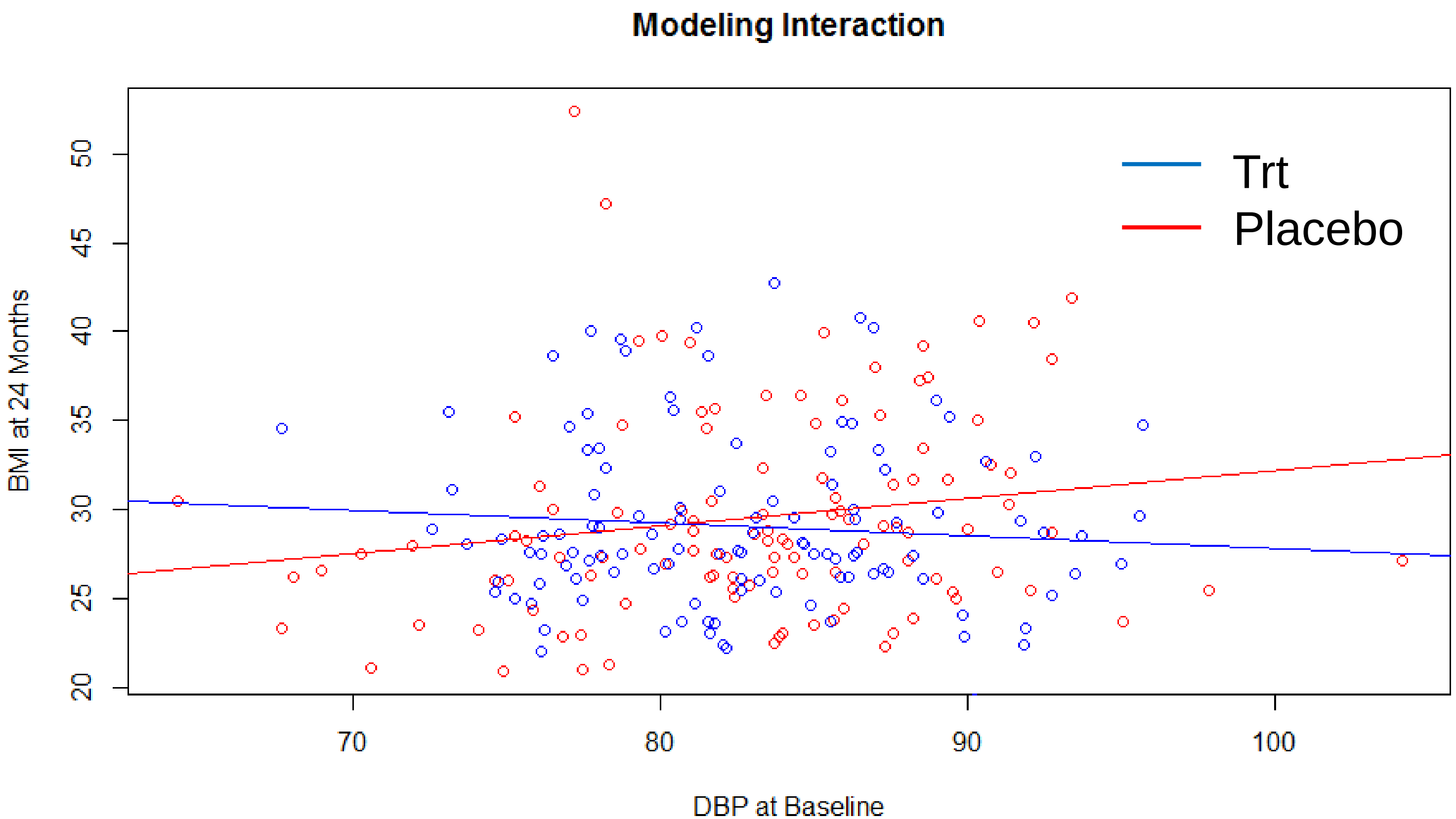
Placebo: $Trt_i = 0$: $BMI24_i = \beta_0 + \beta_2 DBP_i + \varepsilon_i$

Treatment: $Trt_i = 1$: $BMI24_i = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) DBP_i + \varepsilon_i$

- If the relationship between X and Y is the same for each group, then $\beta_2 = \beta_2 + \beta_3$, which implies β_3 must be 0
  - Use multiple regression to test: H_0: β_3=0.

Modeling Interactions

(TROPHY Data)



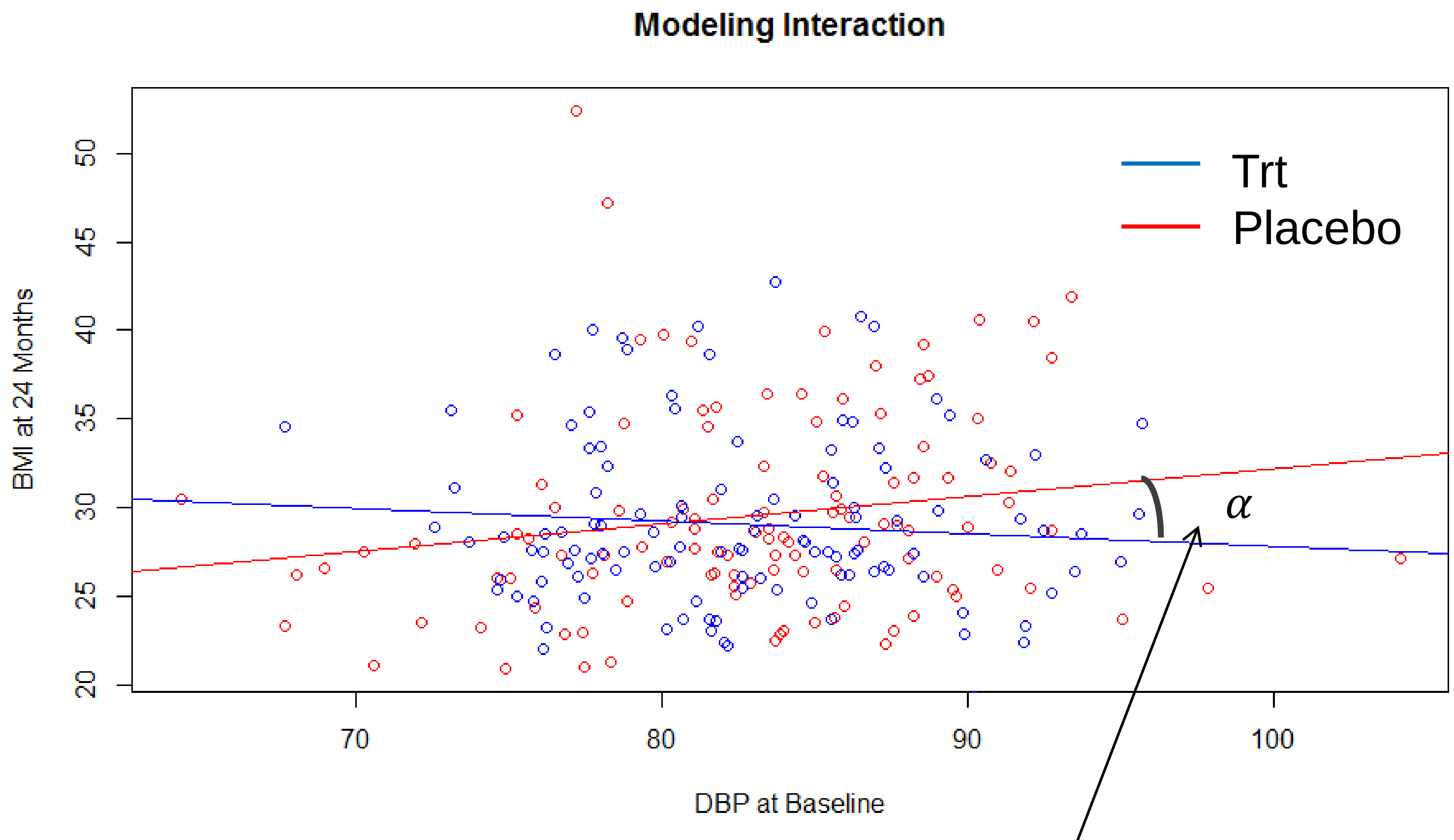
Modeling Interactions
(TROPHY Data)

R Output

Coefficients:	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	16.7	6.28	2.65	0.00853 **
Trt01	18.4	9.46	1.94	0.05319
DBP0	0.16	0.075	2.061	0.04048 *
DBP0:Trt01	-0.23	0.11384	-1.997	0.04698 *

Modeling Interactions

(TROPHY Data)



$$Y \approx 16.7 + 18.4Trt + 0.16X \; - 0.23\, Trt * X$$

Summary Points

- Correlation is a measure of association between continuous X and Y
  - Pearson Correlation (Linear association):

$$\rho = corr(x, y) = \frac{Cov(X, Y)}{SD(X) * SD(Y)} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

    - $|\rho| \leq 1$
    - $|\rho| = 1$: Y is a linear function of X, Y=a+bX
    - $\rho = 0$: No association between X and Y

  - T-test for testing $H_0: \rho = 0$ of no association between of X and Y

$$t_{n-2} = \frac{r}{se(r)} = \frac{r}{\sqrt{\frac{1-r}{n-2}}}$$

  - Spearman Correlation (Nonparametric):

$$corr_S(X, Y) = corr_P\ (rank(X), rank(Y))$$

Summary Points

- Simple Linear Regression (Model Y as a linear function of X)

$$Y_i=\beta_0+\beta_1 X_i + \varepsilon_i \text{ where } \varepsilon_i \sim N(0,\sigma^2)$$

  – β_0 is the intercept: Expected value of Y when X=0.
  – β_1 is the slope: How much Y changes if X changes by 1
  – Least squares estimate of β_0 and β_1:

$$b_1 = \frac{\sum_i Y_i(X_i - \overline{X})}{\sum_i (X_i - \overline{X})^2} = \widehat{corr}(Y,X) * \frac{\widehat{SD}(Y)}{\widehat{SD}(X)}$$

$$b_0 = \overline{Y} - b_1\overline{X}$$

  – T-test for testing $H_0\colon \beta_1 = 0$ of no association between of X and Y

$$t_{n-1} = \frac{b_1}{se(b_1)}$$

Summary Points

- Multiple Regression (Model Y as a linear function of several $X_k's$)

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \varepsilon_i \text{ where } \varepsilon_i \sim N(0, \sigma^2)$$

  - β_0 (Intercept): Expected value of Y when all X_k=0
  - β_k (Slope): How much Y changes if X_k changes by 1 (adjusting for other X's)
  - T-test for testing $H_0: \beta_k = 0$ of no partial association between of X_k and Y

$$t_{n-1} = \frac{b_k}{se(b_k)}$$

  - β_3 (Interaction terms): Does the effect of X on Y varies by group (i.e. Trt)

$$Y_i = \beta_0 + \beta_1 Trt_i + \beta_2 X_{2i} + \beta_3 Trt_i X_i + \varepsilon_i \text{ where } \varepsilon_i \sim N(0, \sigma^2)$$

  - T-test of $H_0: \beta_3 = 0$; The association between X and Y does not vary by group