

# Linear Regression

Cohen Chapter 10

EDUC/PSY 6600

Fit the analysis to the data, *not* the data to the analysis.

- **Statistical Maxim**

# Motivating Example

- Dr. Ramsey conducts a *non-experimental* study to evaluate what she refers to as the 'strength-injury hypothesis.' It states that overall body strength in elderly women determines the number and severity of accidents that cause bodily injury. If the results support her hypothesis, she plans to conduct an experimental study to assess whether weight training reduces injuries in elderly women.
- Data from 100 women who range in age from 60 to 70 years old are collected. The women initially undergo a series of measures that assess upper and lower body strength, and these measures are summarized into an overall index of body strength.
- Over the next 5 years, the women record each time they have an accident that results in a bodily injury and describe fully the extent of the injury. On the basis of these data, Dr. Ramsey calculates an overall injury index for each woman.
- A simple regression analysis is conducted with the overall index of body strength as the predictor (independent) variable and the overall injury index as the outcome (dependent) variable.

# Correlation vs. Regression

## Correlation

- Relationship between two variables  
(no outcome or predictor)
- Strength and direction of  
relationship

# Correlation vs. Regression

## Correlation

- Relationship between two variables (no outcome or predictor)
- Strength and direction of relationship

## Regression

- Outcome and predictor (directional)
- Simple and Multiple Linear Regression

# Regression Basics

- Y usually predicted variable
  - A.k.a: Dependent, criterion, outcome, response variable
  - Predicting Y from X = 'Regressing Y on X'
- X usually variable used to predict Y
  - A.k.a: Independent, predictor, explanatory variable
- Different results when X & Y switched

Regression analysis is procedure for obtaining *the* line that best fits data (Assuming relationship is best described as linear)

# Regression Basics

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

$\hat{Y}_i$  = predicted (unobserved) value of Y for a given case i

$b_0$  = y-intercept:

Constant,  $\hat{Y}$  when  $X = 0$ , only interpreted if  $X = 0$  is meaningful

Alternative notation:  $a$  or  $a_{XY}$

$b_1$  = slope of regression line for 1st IV

Constant, Rate of change in Y for every 1-unit change in X

Alternative notation:  $b_{XY}$

$X_i$  = value of predictor for a given case i

# Accuracy of Prediction

## Correlation $\neq$ Causation

- All points do not fall on regression line
  - Prediction works for most, but not all in sample
- W/out knowledge of X, best prediction of Y is mean  $\bar{Y}$ 
  - $s_y$  : best measure of prediction error
- With knowledge of X, best prediction of Y is from the equation  $\hat{Y}$ 
  - Standard error of estimate (SEE or  $s_{Y.X}$ ): best measure of prediction error
  - Estimated SD of residuals in population



# Accuracy of Prediction

Standard Error of Estimate

Residual or Error Variance  
or Mean Square Error

$$s_{Y.X} = \sqrt{\frac{\sum(Y_i - \hat{Y})^2}{N - 2}} = \sqrt{\frac{SS_{residual}}{df}}$$

$$s_{Y.X}^2 = \frac{\sum(Y_i - \hat{Y})^2}{N - 2} = \frac{SS_{residual}}{df}$$

$df = N - 2$  (2 df lost in estimating regression coefficients)

Seeking smallest  $s_{Y.X}$  as it is a measure of variation of observations around regression line

# Line of Best Fit

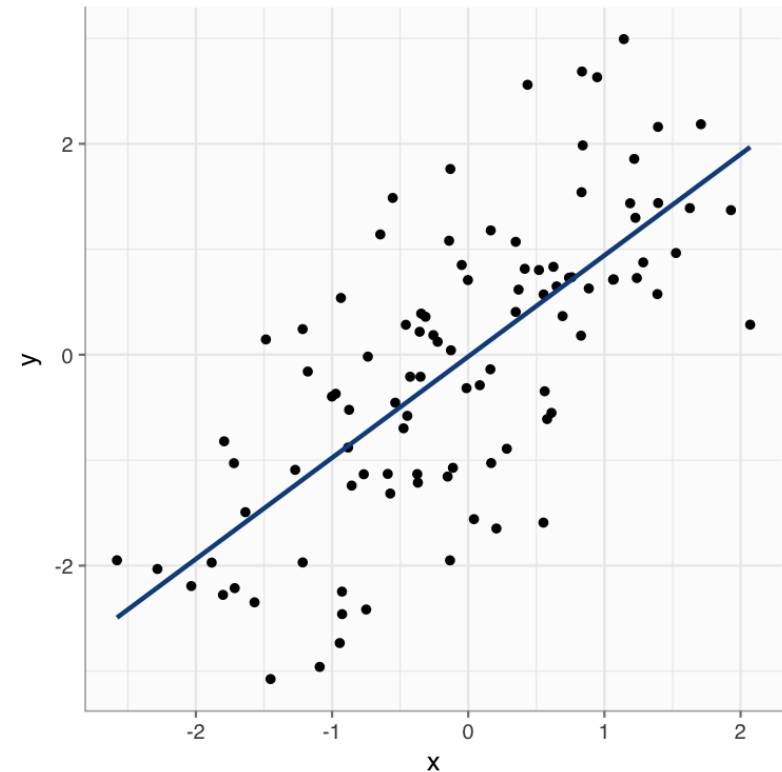
The relationship (prediction) is usually not perfect so regression coefficients ( $b_0, b_1$ ) computed to minimize error as much as possible

**Error of Residuals:** difference between observed  $Y$  and  $\hat{Y}$  -->  $e_i = Y_i - \hat{Y}_i$

**Technique:** Ordinary Least Squares (OLS) regression

Goal: minimize  $SS_{error}$  ( $SS_{residuals}$ )

$$SS_{residuals} = \sum_{i=1}^n (Y_i - \hat{Y}_i)$$



# Line of Best Fit

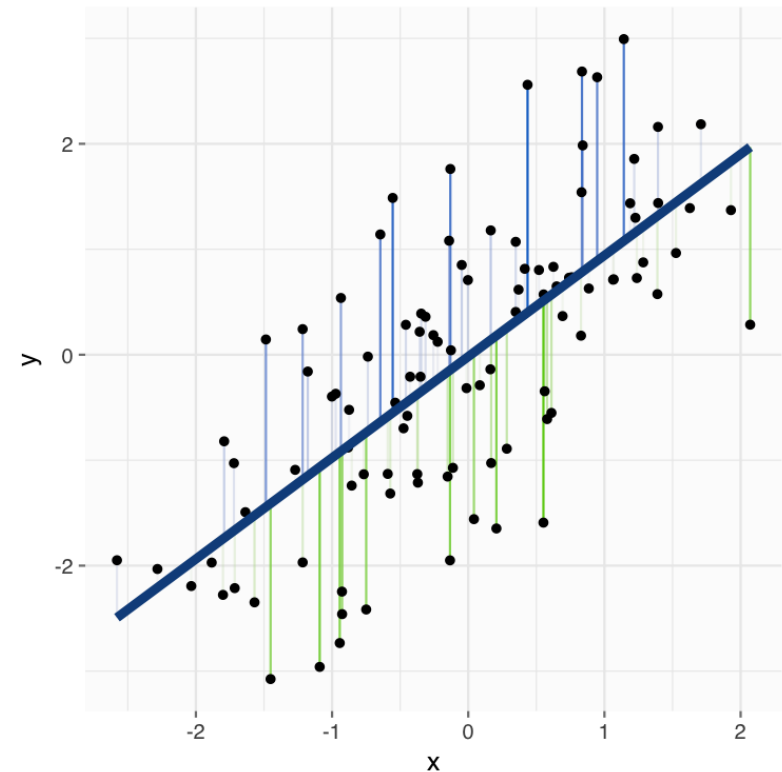
The relationship (prediction) is usually not perfect so regression coefficients ( $b_0, b_1$ ) computed to minimize error as much as possible

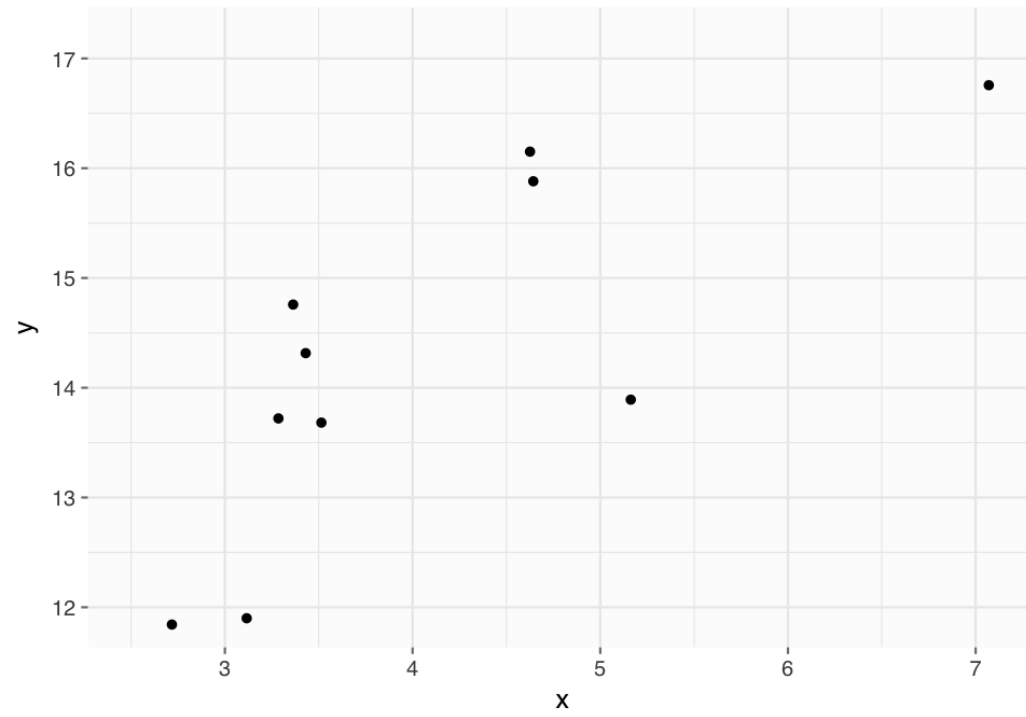
**Error of Residuals:** difference between observed  $Y$  and  $\hat{Y}$  -->  $e_i = Y_i - \hat{Y}_i$

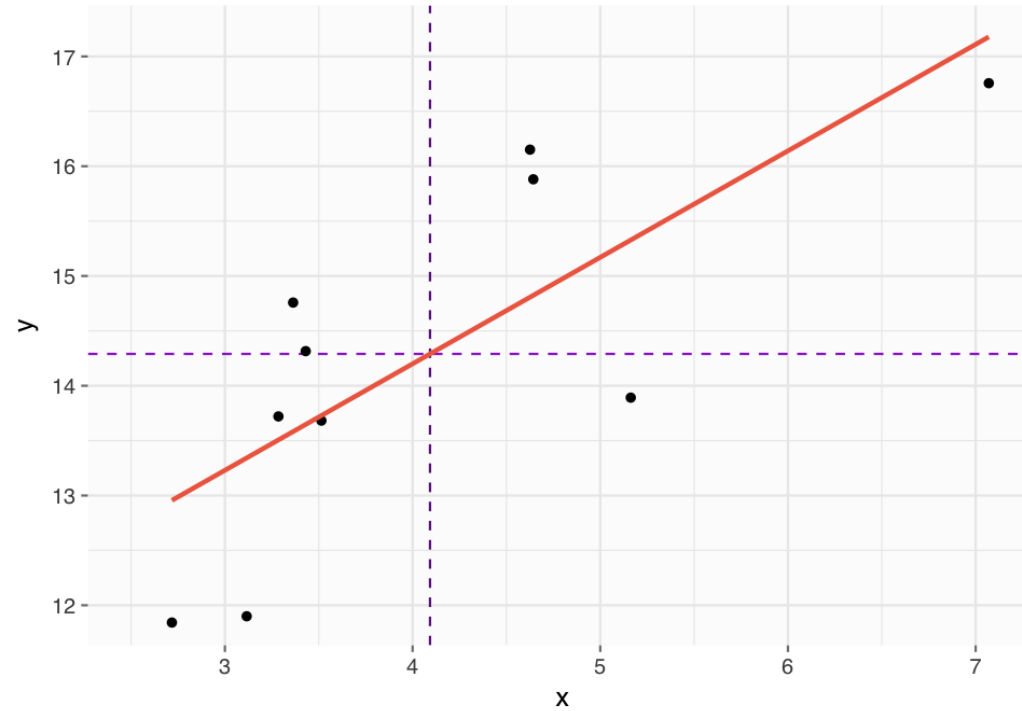
**Technique:** Ordinary Least Squares (OLS) regression

Goal: minimize  $SS_{error}$  ( $SS_{residuals}$ )

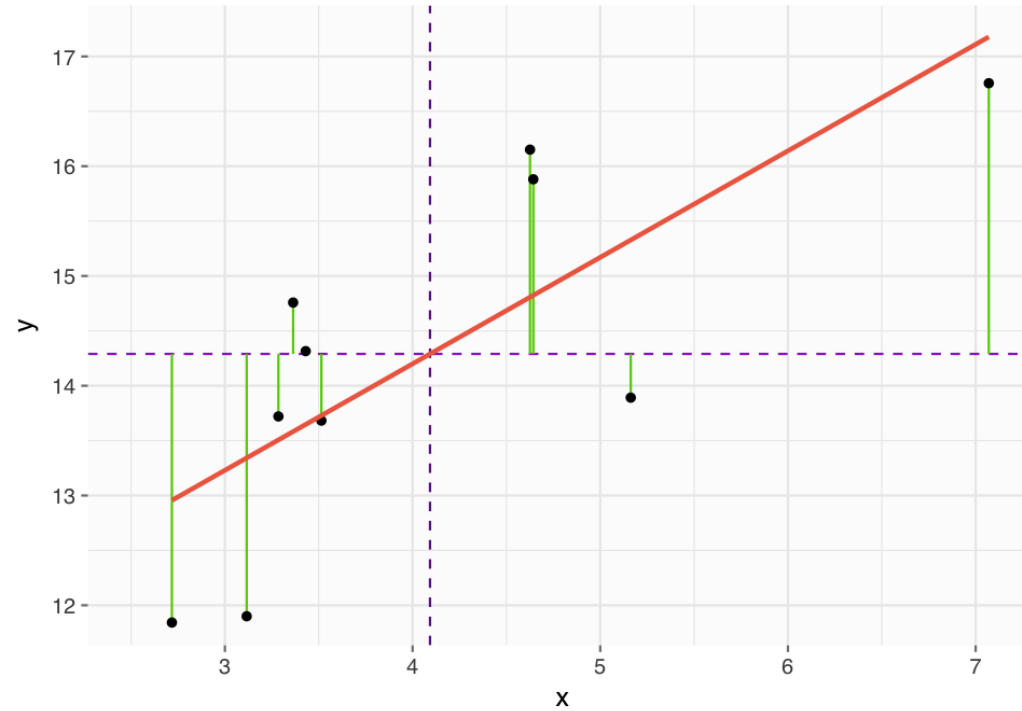
$$SS_{residuals} = \sum_{i=1}^n (Y_i - \hat{Y}_i)$$





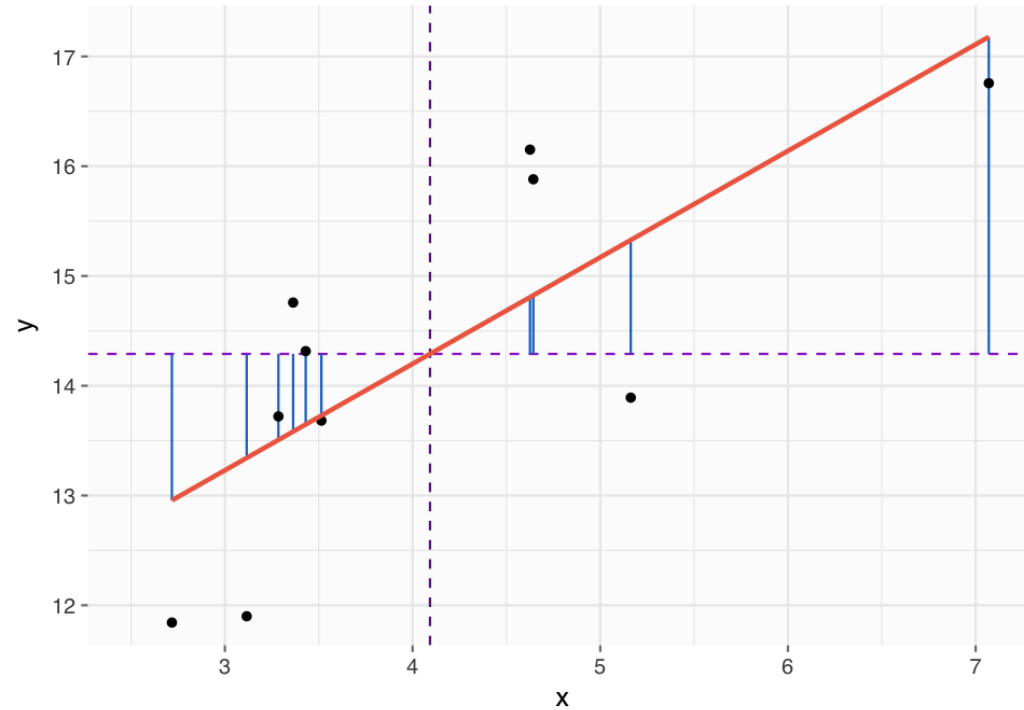


- Correlation = 0.764
- Slope =  $b_1 = r \frac{s_y}{s_x} = .764 \frac{1.66}{1.31} = .968$
- Intercept =  $b_0 = \bar{Y} - b_1 \bar{X} = 14.290 - (.968 * 4.093) = 10.328$



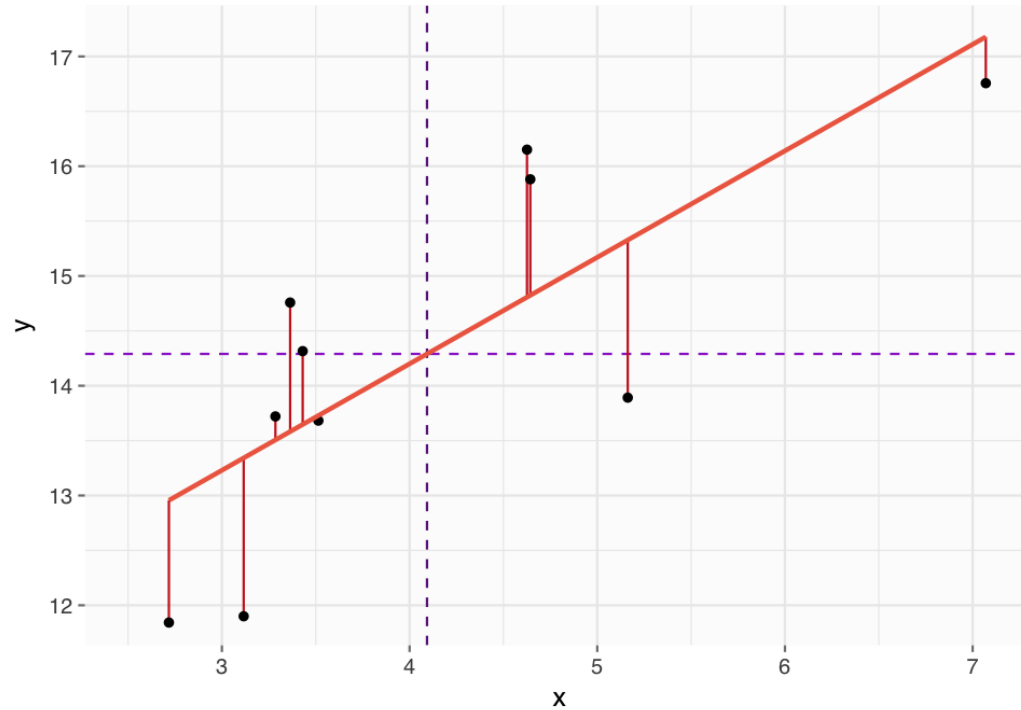
- Correlation = 0.764
- Slope =  $b_1 = r \frac{s_y}{s_x} = .764 \frac{1.66}{1.31} = .968$
- Intercept =  $b_0 = \bar{Y} - b_1 \bar{X} = 14.290 - (.968 * 4.093) = 10.328$

*SS<sub>total</sub>*



- Correlation = 0.764
- Slope =  $b_1 = r \frac{s_y}{s_x} = .764 \frac{1.66}{1.31} = .968$
- Intercept =  
 $b_0 = \bar{Y} - b_1 \bar{X} = 14.290 - (.968 * 4.093) = 10.328$

$$SS_{total} = SS_{explained}$$



- Correlation = 0.764
- Slope =  $b_1 = r \frac{s_y}{s_x} = .764 \frac{1.66}{1.31} = .968$
- Intercept =  
 $b_0 = \bar{Y} - b_1 \bar{X} = 14.290 - (.968 * 4.093) = 10.328$

$$SS_{total} = SS_{explained} + SS_{unexplained}$$



# Explaining Variance

$$SS_{total} = SS_{explained} + SS_{unexplained}$$

- Synonyms: Explained = Regression, Unexplained = Residual or Error

# Explaining Variance

$$SS_{total} = SS_{explained} + SS_{unexplained}$$

- Synonyms: Explained = Regression, Unexplained = Residual or Error

## Coefficient of Determination ( $r^2$ )

$$r^2 = \frac{\text{Explained Variation}}{\text{Total Variation}} = \frac{SS_{regression}}{SS_{total}}$$

- Computed to determine how well regression equation predicts Y from X
- Range from 0 to 1
- SS divided by corresponding df gives us the Mean Square (Regression or Error)
- The proportion of variance in the outcome "accounted for" or "attributable to" or "predictable from" or "explained by" the predictor

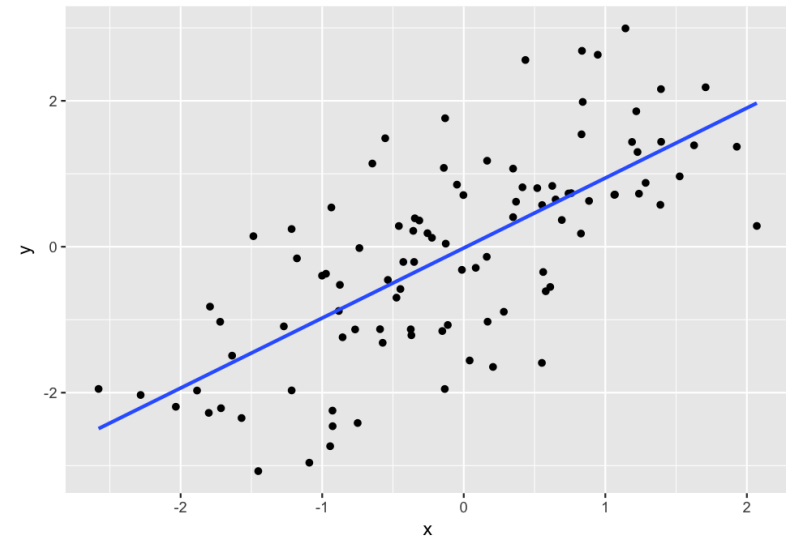
# Standardized Coefficients (i.e. Beta weights)

- 1 SD-unit change in X represents a  $\beta$  SD change in Y
- Intercept = 0 and is not reported when using  $\beta$
- For simple regression only -->  $r = \beta$  and  $r^2 = \beta^2$ 
  - When raw scores transformed into z-scores:  $r = b = \beta$
- Useful for variables with abstract unit of measure

# Again, Always Visualize Data First

## Scatterplots

```
library(tidyverse)
df %>%
  ggplot(aes(x, y)) +
  geom_point() +
  geom_smooth(se = FALSE,
             method = "lm")
```



# R Code: Regression

```
df %>%  
  lm(y ~ x,  
     data = .) %>%  
  summary()
```

Call:

```
lm(formula = y ~ x, data = .)
```

Residuals:

Min	1Q	Median	3Q	Max
-2.10376	-0.56125	0.05069	0.65004	2.15932

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-0.01762	0.09888	-0.178	0.859
x	0.95964	0.09696	9.897	<2e-16 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9849 on 98 degrees of freedom

Multiple R-squared: 0.4999, Adjusted R-squared: 0.4948

F-statistic: 97.95 on 1 and 98 DF, p-value: < 2.2e-16

# R Code: Regression

```
df %>%  
  lm(y ~ x,  
     data = .) %>%  
  confint()
```

```
                2.5 %    97.5 %  
(Intercept) -0.2138558  0.1786119  
x             0.7672237  1.1520547
```

# R Code: Regression

```
df %>%  
  lm(y ~ x,  
     data = .) %>%  
  coef()
```

```
(Intercept)          x  
-0.01762194  0.95963917
```

# R Code: Regression

```
coef1 <- df %>%  
  lm(y ~ x,  
     data = .) %>%  
  coef()  
confint1 <- df %>%  
  lm(y ~ x,  
     data = .) %>%  
  confint()  
cbind(coef1, confint1)
```

	coef1	2.5 %	97.5 %
(Intercept)	-0.01762194	-0.2138558	0.1786119
x	0.95963917	0.7672237	1.1520547



# R Code: Predicted Values

```
df %>%  
  lm(y ~ x,  
     data = .) %>%  
  predict()
```

1	2	3	4	5	6
-1.66331253	-1.58805266	-0.37685641	-0.36001934	-1.82554446	1.96902590
7	8	9	10	11	12
-1.44361263	-2.20795037	-1.52382088	0.13823564	0.40028777	1.32040382
13	14	15	16	17	18
1.44610197	1.17018122	-1.18462186	-0.31876293	0.14390364	-0.85728422
19	20	21	22	23	24
0.83163117	-1.23725243	-0.44710577	0.31680345	0.02232455	0.52088462
25	26	27	28	29	30
0.58236193	-0.26353990	-0.42729936	-0.75393890	0.77690375	0.51344384
31	32	33	34	35	36
-0.06357724	-0.45745486	-1.74608438	-2.49312908	0.33677392	0.78885811
37	38	39	40	41	42
0.71086918	1.21521941	0.51198239	1.54369860	-0.12583856	-0.53196921
43	44	45	46	47	48
-0.47371349	0.78368856	-0.23333494	0.69249078	-0.58503655	1.15183741
49	50	51	52	53	54

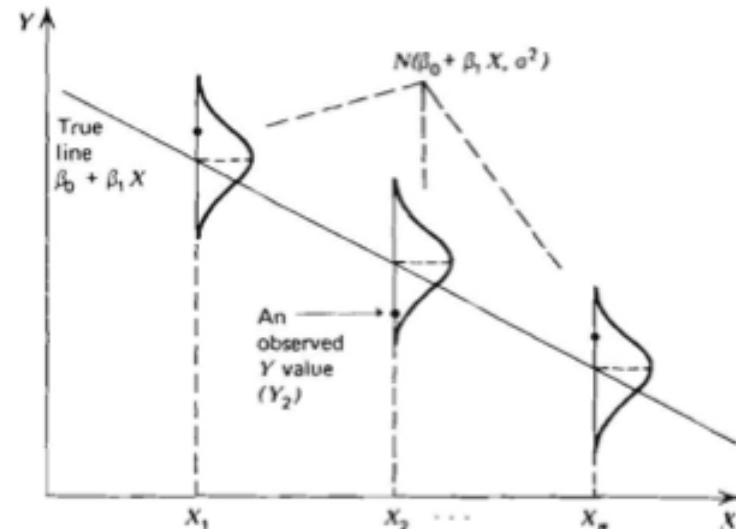
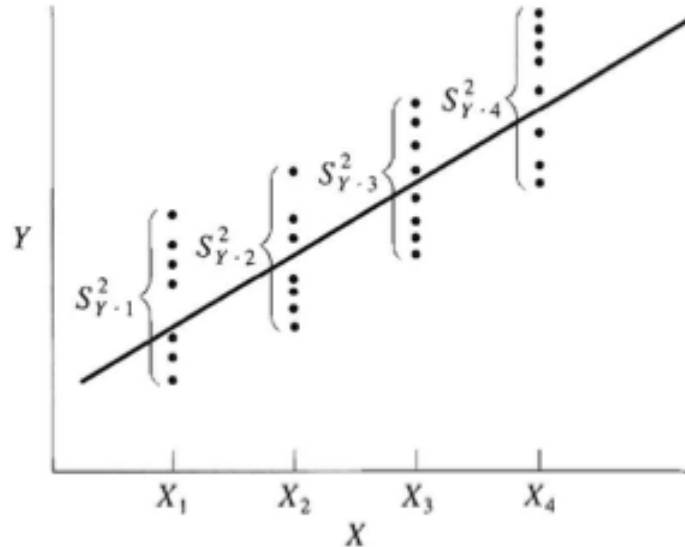
# Assumptions

- **Independence** of observations
- **Y normally distributed**
  - Does NOT apply to predictor variable(s) X --> Can be categorical or continuous
- Sampling **distribution of the slope** ( $b_1$ ) assumed normally distributed
- **Straight line** best fits data

# Assumptions

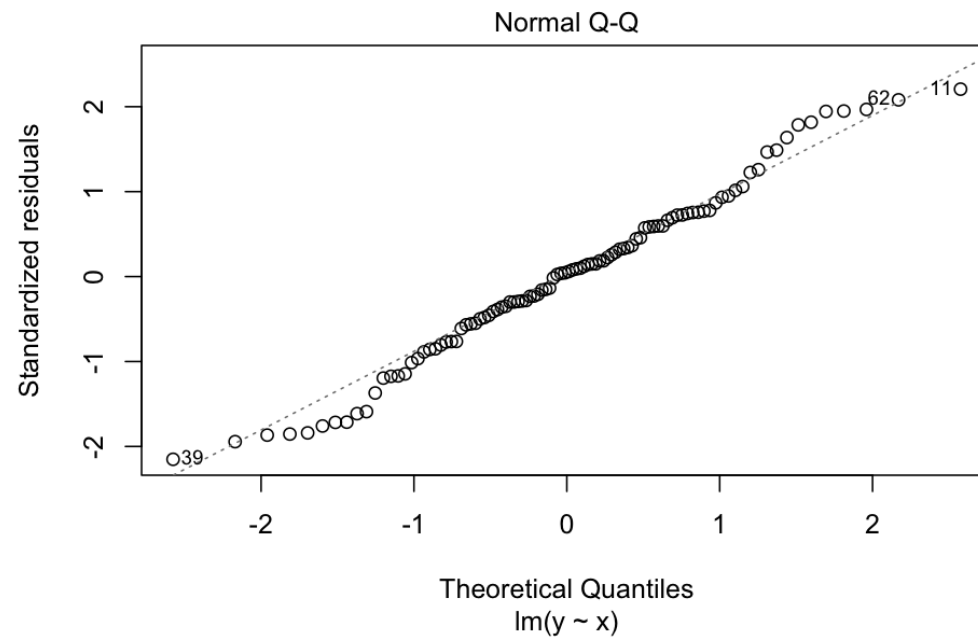
Homogeneity of variance of  $Y$  for all values of  $X$

- Computed error variance ( $s^2_{Y.X}$  or  $MSE$ ) is representative of all individual conditional error variances (for each value of  $X$ )
- ‘Homoscedasticity’
  - Violation = ‘Heteroscedasticity’



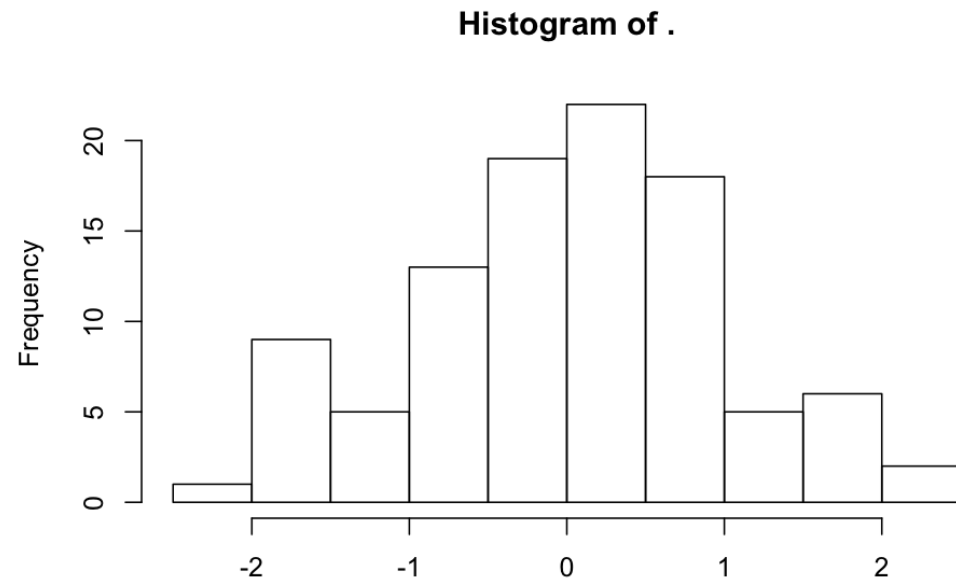
# R Code: Assumptions

```
df %>%  
  lm(y ~ x,  
     data = .) %>%  
  plot(which = 2)
```



# R Code: Assumptions

```
df %>%  
  lm(y ~ x,  
     data = .) %>%  
  resid %>%  
  hist
```



Let's Apply This to the Cancer Dataset

# Read in the Data

```
library(tidyverse)    # Loads several very helpful 'tidy' packages
library(haven)        # Read in SPSS datasets
library(furniture)    # for tableC()
```

```
cancer_raw <- haven::read_spss("cancer.sav")
```

## And Clean It

```
cancer_clean <- cancer_raw %>%
  dplyr::rename_all(tolower) %>%
  dplyr::mutate(id = factor(id)) %>%
  dplyr::mutate(trt = factor(trt,
                            labels = c("Placebo",
                                       "Aloe Juice"))) %>%
  dplyr::mutate(stage = factor(stage))
```

# R Code: Regression

```
cancer_clean %>%  
  lm(totalcin ~ age,  
     data = .) %>%  
  summary()
```

Call:

```
lm(formula = totalcin ~ age, data = .)
```

Residuals:

Min	1Q	Median	3Q	Max
-2.0463	-0.6825	-0.4097	0.6510	5.2266

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	4.71197	1.45471	3.239	0.00362 **
age	0.03032	0.02386	1.271	0.21657

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.'

Residual standard error: 1.512 on 23 degrees of freedom

Multiple R-squared: 0.06559, Adjusted R-squared:

F-statistic: 1.614 on 1 and 23 DF, p-value: 0.2166



# R Code: Standardized

```
cancer_clean %>%  
  mutate(totalcinZ = scale(totalcin),  
         ageZ = scale(age)) %>%  
  lm(totalcinZ ~ ageZ,  
     data = .) %>%  
  summary()
```

Call:

```
lm(formula = totalcinZ ~ ageZ, data = .)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.3367	-0.4458	-0.2676	0.4253	3.4143

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2.442e-16	1.975e-01	0.000	1.000
ageZ	2.561e-01	2.016e-01	1.271	0.217

Residual standard error: 0.9874 on 23 degrees of freedom

Multiple R-squared: 0.06559, Adjusted R-squared: 0.02496

F-statistic: 1.614 on 1 and 23 DF, p-value: 0.2166

# R Code: Correlation vs. Standardized

```
cancer_clean %>%  
  cor.test(~ totalcinZ + ageZ,  
           data = .)
```

```
cancer_clean %>%  
  mutate(totalcinZ = scale(totalcin),  
         ageZ = scale(age)) %>%  
  lm(totalcinZ ~ ageZ,  
     data = .) %>%  
  summary()
```

# R Code: Correlation vs. Standardized

```
Pearson's product-moment correlation
```

```
data: totalcin and age
```

```
t = 1.2706, df = 23, p-value = 0.2166
```

```
alternative hypothesis: true correlation is not equal to 0
```

```
95 percent confidence interval:
```

```
-0.1546769  0.5913913
```

```
sample estimates:
```

```
cor
```

```
0.2561066
```

```
Call:
```

```
lm(formula = totalcinZ ~ ageZ, data = .)
```

```
Residuals:
```

```
Min      1Q  Median      3Q      Max
-1.3367 -0.4458 -0.2676  0.4253  3.4143
```

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2.442e-16	1.975e-01	0.000	1.000
ageZ	2.561e-01	2.016e-01	1.271	0.217

```
Residual standard error: 0.9874 on 23 degrees of freedom
```

```
Multiple R-squared:  0.06559,    Adjusted R-squared:
```

```
F-statistic: 1.614 on 1 and 23 DF,  p-value: 0.2166
```

Questions?

# Next Topic

Matched T-Test