Course Materials May Not Be Distributed or Posted Electronically

These course materials are the sole property of Dr. Todd M. Gross. They are strictly for use by students enrolled in a course taught by Dr. Gross. They may not be altered, excerpted, distributed, or posted to any website or other document-sharing service without express permission.

PSTAT 126
Regression Analysis

Dr. Todd Gross
Department of Statistics and Applied Probability
UCSB

Lecture 5

Lecture Outline

- ANOVA in R output
- Issues in Testing β_1
- Confidence Intervals (see pages 81-82 in ISL text)
  - Regression Coefficients
  - Predicting Mean Response of Y Given X
  - Predicting Future Observation of Y Given X
- Major Assumptions of the Normal Error Regression Model

ANOVA Output in R

R Commands – Units Taken vs. Hours of Study

```
units<-c(9,10,12,13,14,16,17,18,20,21)
hours<-c(18,25,32,34,38,36,42,46,48,46)
model<-lm(hours~units)
plot(units,hours,main="Course Units versus Hours of
Study",xlab="Units Taken",ylab="Weekly Hours of
Study")
abline(model)
summary(model)
anova(model)
```

R Plot - Example #2

R Summary Output

```
Call:
lm(formula = hours ~ units)

Residuals:
   Min     1Q Median     3Q    Max
-4.940 -2.120  0.590  2.215  3.760

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.6000     4.0090   0.649    0.535
units         2.2600     0.2588   8.733 2.31e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.169 on 8 degrees of freedom
Multiple R-squared:  0.9051,  Adjusted R-squared:  0.8932
F-statistic: 76.27 on 1 and 8 DF,  p-value: 2.311e-05
```

ANOVA Output in R

```
> anova(model2)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
units      1 766.14  766.14  76.271 2.311e-05 ***
Residuals  8  80.36   10.04
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- $F(1,8) = 76.271$, $p < 0.05$, therefore REJECT H_O
- This is same value of F and p as in the Summary output

Issues in Testing β_1

Issues in Testing β_1

- Two-tailed vs one-tailed test

- Specifying the hypothesized value of β_1

Two-tailed vs. One-tailed Test of B_1

Two-tailed test

- H_0: $\beta_1 = \beta_{10}$, H_1: $\beta_1 \neq \beta_{10}$
- Tests whether there is a linear relationship between X and Y
- We are interested in finding any relationship, positive or negative

One-tailed test

- H_0: $\beta_1 \geq \beta_{10}$, H_1: $\beta_1 < \beta_{10}$

-OR-

- H_0: $\beta_1 \leq \beta_{10}$, H_1: $\beta_1 > \beta_{10}$
- Tests whether there is specifically a positive <or negative> relationship between X and Y
- Any relationship in the opposite direction is treated the same as no relationship at all

Examples

Which of the following could be tested as one-tailed hypotheses?

- Can we predict first year post-college salary from college GPA?
- Does increasing R & D expenditure affect sales revenue?
- Does year in college predict number of units taken?
- Does hours of exercise predict weight loss?

There often is a different answer from a researcher vs. someone using the prediction

NOTE: It is the burden of the researcher to demonstrate that a relationship in the opposite direction is of NO interest

Year in College vs. Units Taken Example

Question: Do units taken increase **more** than one unit per year?

- Have we specified a value for β_1?
- Can we justify a one-tailed test?

A Worked Example Calculations

	Year in School (X)	Number of Units (Y)	X-Xbar	Y-Ybar	(X-Xbar)(Y-Ybar)	Yhat	Y - Yhat
	1	9	-2	-3	6	9.2	-0.2
	2	12	-1	0	0	10.6	1.4
	3	10	0	-2	0	12	-2
	4	14	1	2	2	13.4	0.6
	5	15	2	3	6	14.8	0.2
n	5	5	5	5	5	5	5
Sum	15	60	0	0	14	60	0
Mean	3	12	0	0	2.8	12	0
Sum of Squares	55	746	10	26	76	739.6	6.4
b1	1.4		SSE	6.4			
bo	7.8		n-2	3			
			MSE	2.13			
			SSX	10			
			se(b1)	0.46			
			t(b1)	3.03			

Specifying B_{10}

β_{10} is the value of β_1 specified under the null hypothesis (H_0)

- Examples:
  - H_O: $\beta_1 = 0$
  - H_O: $\beta_1 = 1$

Where do we obtain β_{10}?

- Theory
- Prior Research
- Utility

Year in College vs. Units Taken Example

- Statistical Hypotheses
  - $H_O: \beta_1 \leq 1$ & $H_1: \beta_1 > 1$
- Given Data
  - $n = 5$, $b_0 = 7.8$, $b_1 = 1.4$, $SSE = 6.4$, $SSX = 10$
- Calculations
  - $dfe = n - 2 = 3$
  - $MSE = SSE/dfe = 6.4/3 = 2.133$
  - $s\{b_1\} = \text{sqrt}(MSE/SSX) = \surd(2.133/10) = 0.46$
  - $t^* = b_1 - \beta_{10}/s\{b_1\} = (1.4 - 1)/0.46 = \boxed{0.87}$
  - t(alpha = .05, df = 3) = 2.353 – from Table B.2
- Statistical Conclusion
  - Fail to Reject

Hypothesized value of B_1

Value of t is now smaller

R Example for Testing $H_0: \beta_1 = 1$

```
> x<-c(1,2,3,4,5)
> y<-c(9,12,10,14,15)
> fit<-lm(y~x)
> summary(fit)

Call:
lm(formula = y ~ x)

Residuals:
   1    2    3    4    5 
-0.2  1.4 -2.0  0.6  0.2 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)   7.8000     1.5319   5.092   0.0146 *
x             1.4000     0.4619   3.031   0.0563 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.461 on 3 degrees of freedom
Multiple R-squared:  0.7538,	Adjusted R-squared:  0.6718 
F-statistic: 9.187 on 1 and 3 DF,  p-value: 0.05626
```

We can use these results to hand-calculate
$t = (1.4-1)/0.46 = 0.87$

Confidence Intervals in Regression

The Concept of a Confidence Interval

- A confidence interval is a range of values that we are "confident" includes the population parameter of interest
- Example: Hours of sleep
  - How many hours did <u>you</u> sleep last night?
  - What do you think the average hours of sleep was for students in this class?
  - What range would capture 90% of the students (9 out of 10)? How about 99% (99 out of 100)?
- A <u>wider</u> confidence interval yields <u>higher confidence</u> that it contains the population parameter, but a wider interval is also <u>less exact</u>.

Types of Confidence Intervals in Regression

- Confidence interval for regression coefficients
  - Slope and intercept
- Confidence Interval for predicting the <u>mean response</u> of Y for a given value of X
- Confidence Interval when predicting <u>an individual value</u> of Y (i.e., a future observation) for a given value of X

- We will use the example predicting exam score from hours of study

R Commands and Output
Hours of Study and Exam Score

```
> hours<-c(1, 1.2, 1.4, 1.6, 1.9, 2.8, 3.8, 4.2, 5, 5.2, 5.8, 6.2, 6.8, 7.4, 8.1, 8.4, 10, 10.7, 11.1)
> score<-c(29, 13, 31, 27, 45, 48, 44, 50, 44, 61, 73, 50, 78, 54, 83, 61, 74, 67, 80)
> plot(hours,score,xlab="Hours of Study",ylab="Exam Score")
> model1=lm(score~hours)
> abline(model1)
> summary(model1)

Call:
lm(formula = score ~ hours)

Residuals:
    Min      1Q  Median      3Q     Max 
-18.909  -7.280  -1.925   8.355  17.703 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  25.8073     4.8378   5.335 5.48e-05 ***
hours         5.0844     0.7703   6.601 4.50e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.77 on 17 degrees of freedom
Multiple R-squared:  0.7193,      Adjusted R-squared:  0.7028 
F-statistic: 43.57 on 1 and 17 DF,  p-value: 4.497e-06
```

R Example: Confidence Intervals for Regression Coefficients (slope and intercept)

- A 95% confidence interval for the slope and intercept

```
> confint(model1)
                2.5 %    97.5 %
(Intercept) 15.600409 36.014118
hours        3.459293  6.709557
```

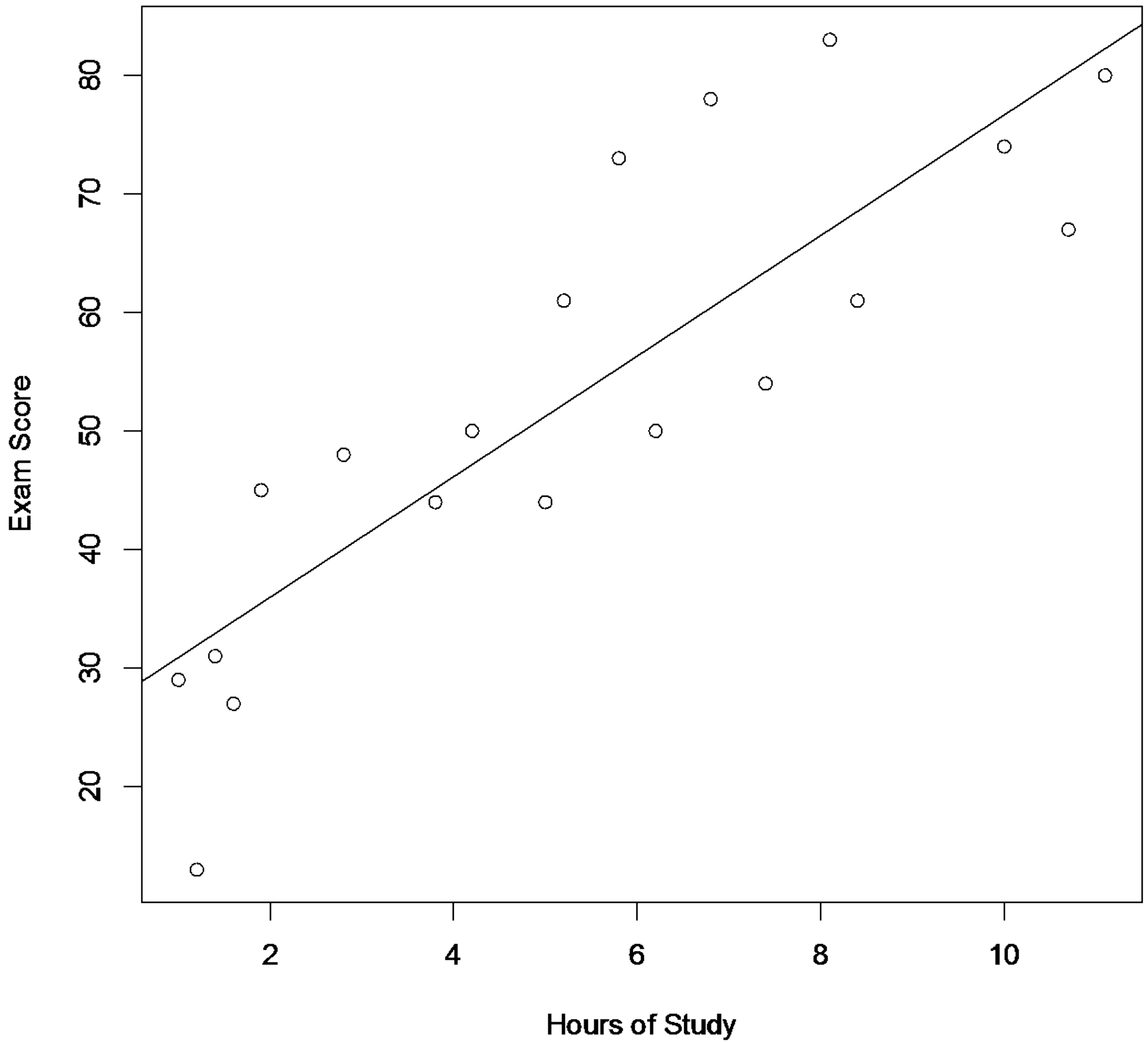
“We are 95% confident that the population slope is between 3.46 and 6.71”

- A 99% confidence interval for the slope and intercept

```
> confint(model1,level=.99)
                0.5 %    99.5 %
(Intercept) 11.786219 39.828308
hours        2.851999  7.316851
```

“We are 99% confident that the population slope is between 2.85 and 7.32”

Plot of Study Hours and Exam Score



Y', Predicted Values of Y

- We can get all of the predicted values of Y' in R

```
> hours
 [1]  1.0  1.2  1.4  1.6  1.9  2.8  3.8  4.2  5.0  5.2  5.8  6.2  6.8  7.4  8.1  8.4
10.0 10.7 11.1
> predict(model1)
1        2        3        4        5        6
30.89169 31.90857 32.92546 33.94234 35.46767 40.04365
       7        8        9       10       11       12
45.12808 47.16185 51.22939 52.24627 55.29693 57.33070
      13       14       15       16       17       18
60.38135 63.43201 66.99111 68.51643 76.65151 80.21061
      19
82.24438
```

- For example, if a student studies 1.9 hours (the 5th value of X), we predict an exam score of 35.47
- But these are only for values of X that appear in the dataset
- What score would we predict for a student who studied 7 hours?

Predicting a Future Outcome

- To predict an outcome, we apply the regression equation:

  $$Y' = b_0 + b_1(X)$$

- For X = 7, we have

  $$Y' = 25.8 + 5.1\ (7) = 61.4$$

- We predict that a student who studies 7 hours will score 61.37

- In R, we use a dataframe to set the value of X=7

```
> predict(model1,data.frame(hours=7))
       1
61.39824
```

Confidence Intervals for Predicted Response

- We can calculate two different confidence intervals for a predicted response
  - Confidence interval for predicting the mean response for a given value of X
  - Prediction interval for predicting an individual response for a given value of X

Predicting the Mean Response

- What is the <u>mean</u> exam score we predict for <u>all</u> students who study 7 hours?
  - Y' = 61.4
  - We can calculate a confidence inteval around this value using:

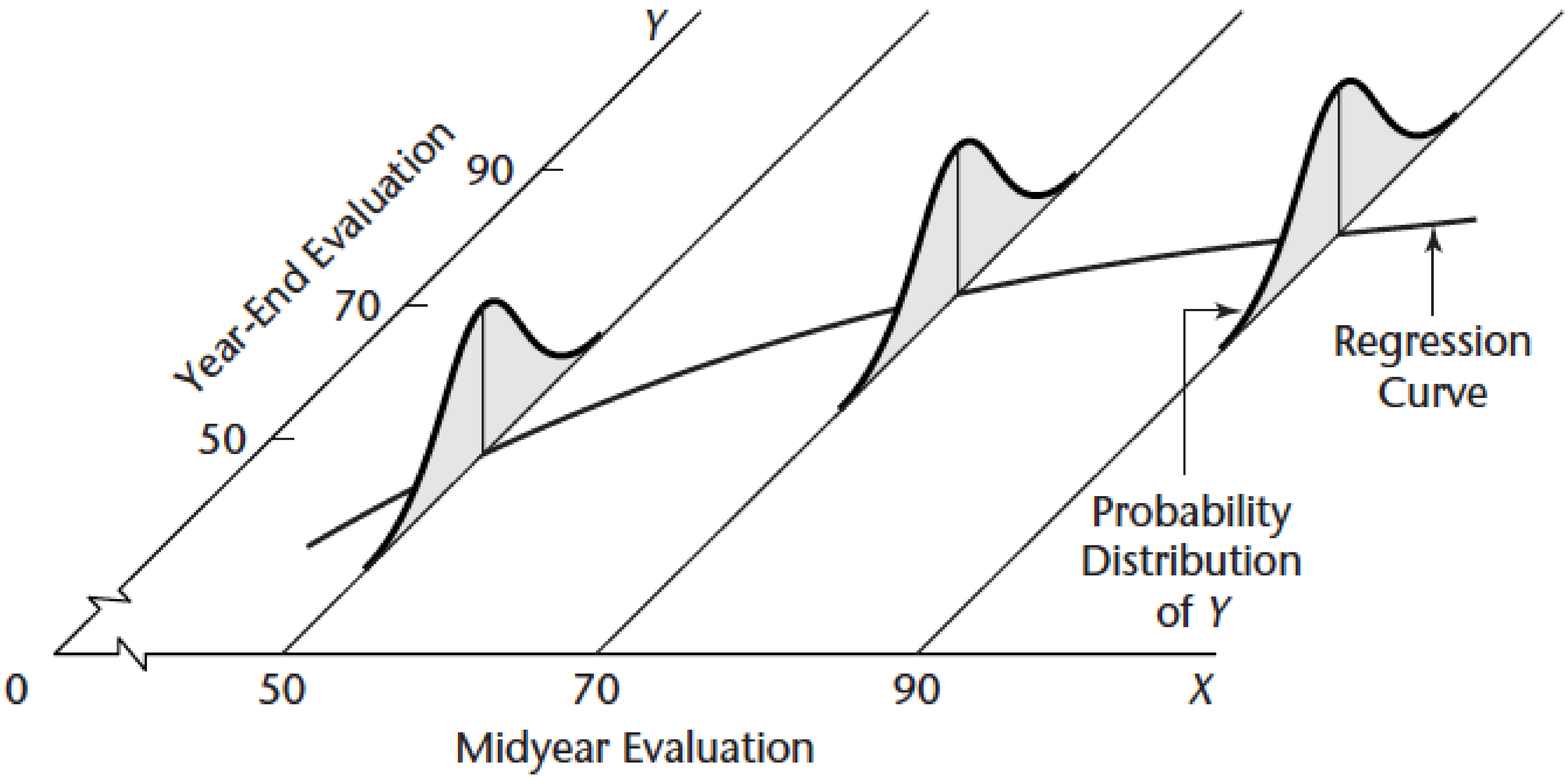
$$\beta_0 + \beta_1 X_h \pm t(1-\alpha/2; n-2)\sqrt{MSE\left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)}$$

- In R, we simply add a confidence interval to our prediction command:

```
> predict(model1,data.frame(hours=7),interval="confidence")
       fit      lwr    upr
1 61.39824 55.57348 67.223
```

"We are 95% confident that the mean exam score for students who study 7 hours is between 55.6 and 67.2"

Error around the Regression Line



- There is a distribution of observed values of Y around each predicted value of Y'

Predicting a Future Observation of Y

Example: What exam score would we predict for a specific student who studies 7 hours?

- This is different than predicting the mean exam score for all students who study 7 hours
  - The predicted value is same ($Y'=b_0 + b_1X$)
  - The variance of an individual score is greater than the variance of the mean (central slimit theorem)

- The confidence interval is known as a Prediction Interval for Y'

Predicting an Individual Response

- What is the exam score we predict for an <u>individual</u> student who studies 7 hours?
  - Y' = 61.4 (same as the predicted <u>mean</u> response)
  - We can calculate a confidence interval around this value using:

$$\beta_0 + \beta_1 X_h \pm t(1-\alpha/2; n-2)\sqrt{MSE\left(1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)}$$

- In R, we simply add a <u>prediction</u> interval to our predict command:

```
> predict(model1,data.frame(hours=7),interval="prediction")
 fit        lwr        upr
1 61.39824 37.94412 84.85235
```

"We are 95% confident that the individual exam score for a student who studies 7 hours is between 37.9 and 84.9"

Prediction Interval for Individual Response

- We are predicting a single score, instead of the mean of a set of scores
- The single prediction has more variability

$$\beta_0 + \beta_1 X_h \pm t(1-\alpha/2; n-2)\sqrt{MSE\left(1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)}$$

Additional Variability of Individual Score

- Due to this additional variability, a prediction interval will be wider than a confidence interval

Confidence vs Prediction Intervals

- Confidence Intervals
  - For intercept, β_0
  - For slope, β_1
  - For estimating the mean response at a given value of X

- Prediction Interval
  - For an individual future observation

Major Assumptions of the Normal Error Regression Model

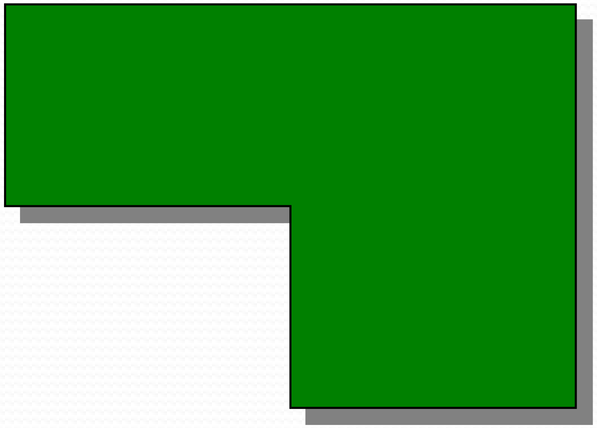
"All models are wrong, some are useful."

"Statisticians, like artists, have the bad habit of falling in love with their models."

George Box (1919-2013)

The Importance of Model Accuracy

- I need to fertilize my lawn. How much fertilizer do I need?



Area = height X width – .25(h*w)

- My estimate is only as accurate as the model I use to calculate the area
- If I choose the wrong model, I will get an inaccurate estimate

The Need for Assumptions

- In order for the linear model to provide <u>valid</u> estimates, its assumptions must hold true
- If the assumptions are <u>violated</u>, then the estimates of regression coefficients, and/or the corresponding confidence intervals and p-values, may be distorted.
- We must <u>diagnose</u> possible violations (i.e., deviations) from the model before we accept the results of fitting and testing the model
- We may be able to correct, or <u>remediate</u>, violations to produce a valid regression model

Major Assumptions of the Normal Error Regression Model

The normal error regression model is:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, where\ \epsilon_i \sim iid\ N(0, \sigma^2)$$

This equation implies the following assumptions:

- Linearity – Y is a linear function of X
- Independence – the responses (and their errors) are independent of each other
- Identically Distributed – all observations come from the same population
- Normality – the responses follow a Normal distribution
- Constant Variance – σ^2 is the same for all values of X

Linearity

$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Y is a linear function of X}} + \epsilon_i, where\ \epsilon_i \sim iid\ N(0, \sigma^2)$$

- Linear regression is valid when there is a linear relationship between X and Y
- If the relationship between X and Y is non-linear, then the linear model <u>cannot</u> accurately describe the shape and strength of the relationship

Independence

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

Errors are Independent and Identically Distributed

- *iid* stands for Independent and Identically Distributed
- This first part of this assumption (Independence) means that each observation is independent of every other observation
- If the observations are related somehow, the resulting regression model may be distorted

Identically Distributed

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, where\ \epsilon_i \sim iid\ N(0, \sigma^2)$$

Errors are Independent and Identically Distributed

- *iid* stands for Independent and Identically Distributed
- This second part of this assumption (Identically Distributed) means that all of the observations come from the same population
- An outlier, or extreme observation, may be from a different population

Normality

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, where\ \epsilon_i \sim iid\ \underbrace{N}_{\text{Errors are Normally Distributed}}(0, \sigma^2)$$

- Y values, and their corresponding errors, are normally distributed
- If Y values are not normally distributed (e.g., skewed) then the sampling distribution of b_1 will be incorrect

Constant Variance

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, where\ \epsilon_i \sim iid\ N(0, \underbrace{\sigma^2}_{})$$

Variance is constant across range of X values

- Variance of the Y values, and their corresponding errors, is constant across all values of X
- If variance is not constant (i.e., variance is smaller at low values of X and higher at high values of X), then the standard error of b_1 will be incorrect