

# CORRELATION & LINEAR REGRESSION

---

Analytics Primer

# Regression Analysis

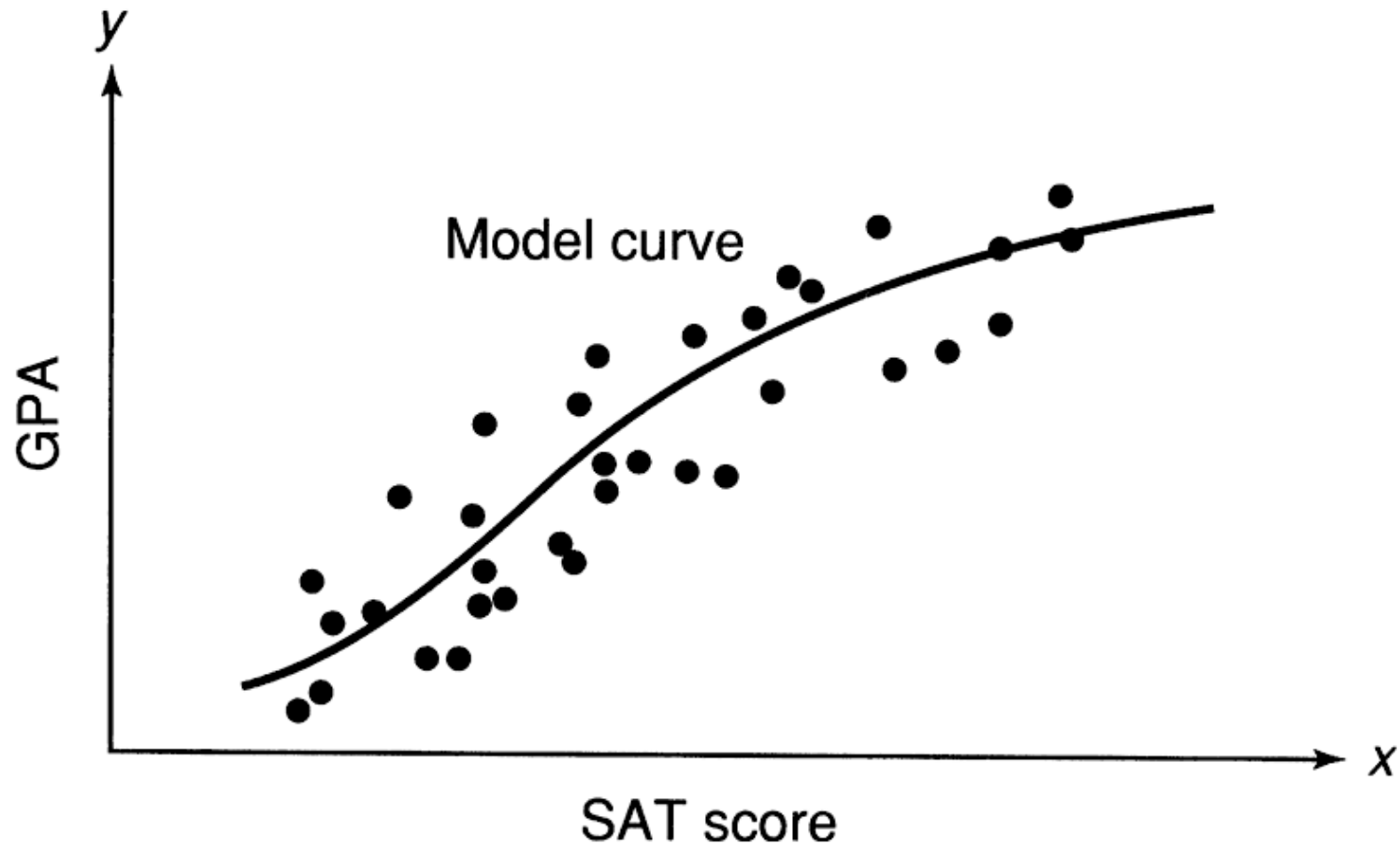
- Typically, the desire is to predict values of a variable.
- Process of finding an equation to predict our variable is called **regression analysis**.
- The variable of interest to be predicted is called the **response** ( or **dependent**) variable.

# Example

- Try to estimate the GPA at the end of freshman year.
- You could use the average GPA of all freshman, but that would only be so accurate.
- What about SAT score, IQ, major, etc.?
- Variables used to try and predict  $y$  are called **independent** (or **explanatory**) variables.
- Models using independent variables potentially have smaller errors.

# Example

- Try to estimate the GPA at the end of freshman year using SAT score.



# Visualizing Relationships

- Viewing the relationship between a response variable and an explanatory variable is very beneficial.
- **Linear relationship** – relationship between variables that exhibits a fairly straight / linear pattern
- **Nonlinear relationship** – relationship between variables that exhibits a pattern that is nonlinear in nature

# Visualizing Relationships

- Viewing the relationship between a response variable and an explanatory variable is very beneficial.
- **Linear relationship** – relationship between variables that exhibits a fairly straight / linear pattern
- **Nonlinear relationship** – relationship between variables that exhibits a pattern that is nonlinear in nature



Not the same thing as  
linear vs. nonlinear models.  
Will discuss later in course!

# Visualizing Relationships

- Viewing the relationship between a response variable and an explanatory variable is very beneficial.
- **Positive relationship** – as one variable increases (or decreases) the other has a *tendency* to do the same
- **Negative relationship** – as one variable increases (or decreases) the other has a *tendency* to do the opposite

# CORRELATION COEFFICIENT

---



# Correlation

- Correlation is a popular term that is thrown around by people who may not understand the implications (or lack there of) of what they are saying.
- The **Pearson correlation coefficient**,  $r$ , is a measure of strength of the *linear* relationship between two variables.

# Correlation Coefficient

- The Pearson correlation coefficient,  $r$ , is calculated as follows:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{[\sum_{i=1}^n (x_i - \bar{x})^2][\sum_{i=1}^n (y_i - \bar{y})^2]}}$$
$$= \frac{1}{n-1} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)$$

# Correlation Coefficient

- The Pearson correlation coefficient,  $r$ , is calculated as follows:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{[\sum_{i=1}^n (x_i - \bar{x})^2][\sum_{i=1}^n (y_i - \bar{y})^2]}}$$

$$= \frac{1}{n-1} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)$$

Standardizing the variables x and y

# Correlation Coefficient

- The Pearson correlation coefficient is unit less due to the standardization of  $x$  and  $y$ .

$$-1 \leq r \leq 1$$

# Correlation Coefficient

- The Pearson correlation coefficient is unit less due to the standardization of  $x$  and  $y$ .

$$-1 \leq r \leq 1$$

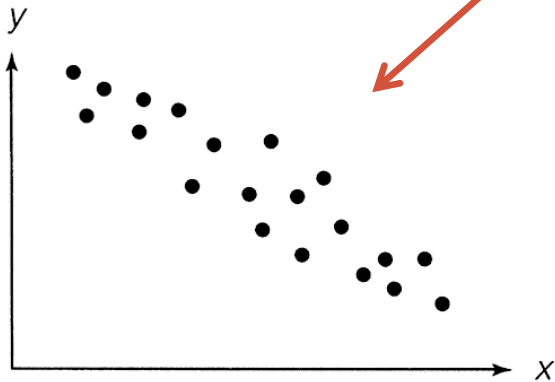
Property 1

# Correlation Coefficient

- The Pearson correlation coefficient is unit less due to the standardization of  $x$  and  $y$ .

$$-1 \leq r \leq 1$$

Property 2



# Correlation Coefficient

- The Pearson correlation coefficient is unit less due to the standardization of  $x$  and  $y$ .

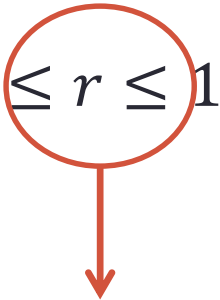
$$-1 \leq r \leq 1$$

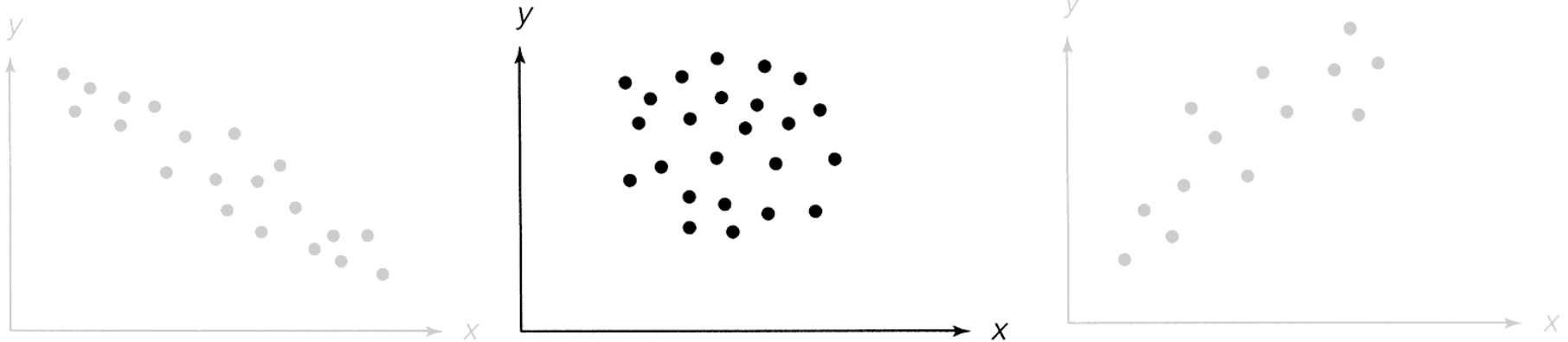
Property 3



# Correlation Coefficient

- The Pearson correlation coefficient is unit less due to the standardization of  $x$  and  $y$ .

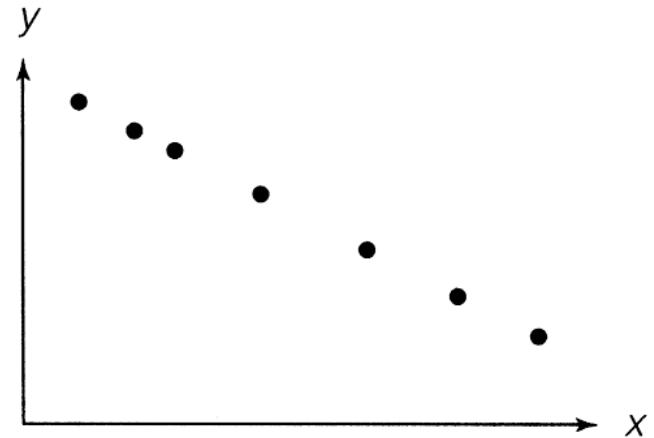
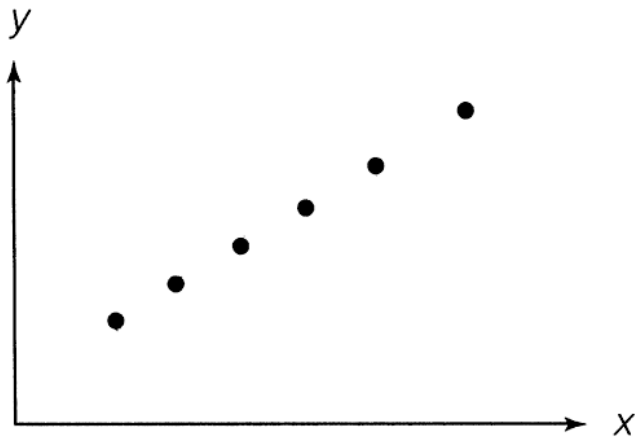
$$-1 \leq r \leq 1$$




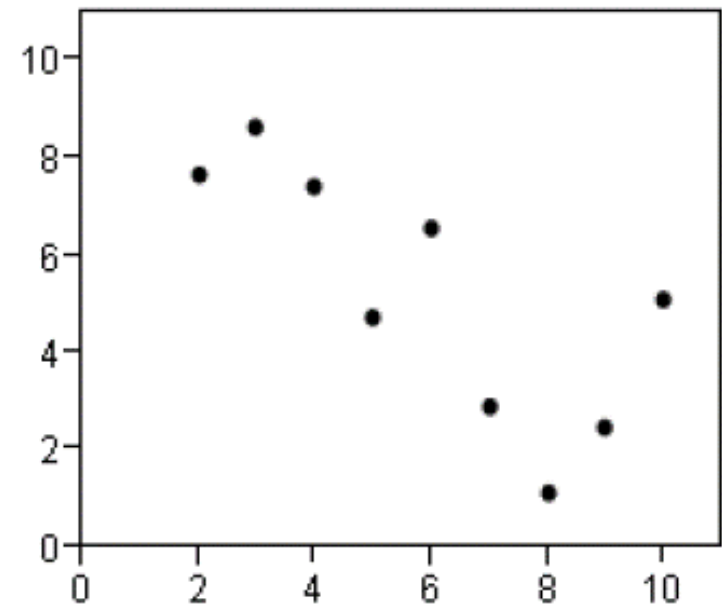
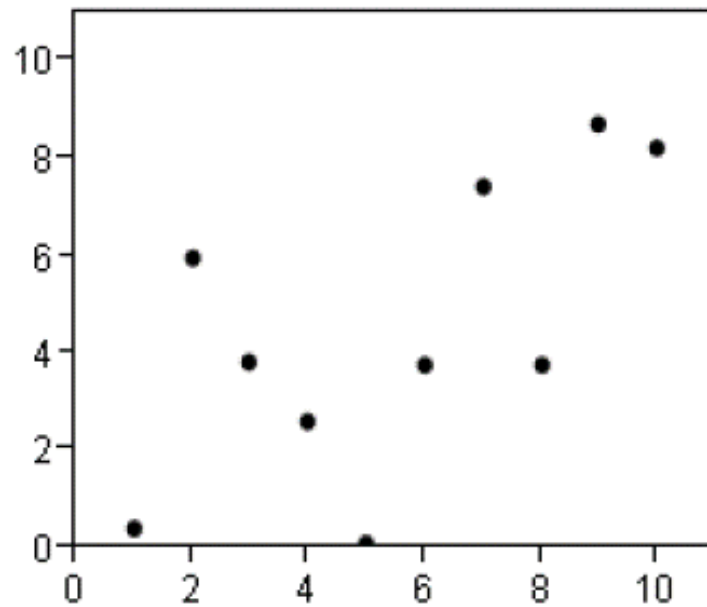


# Correlation Coefficient

- Values of 1 or -1 imply a **perfect** linear relationship between  $y$  and  $x$ .

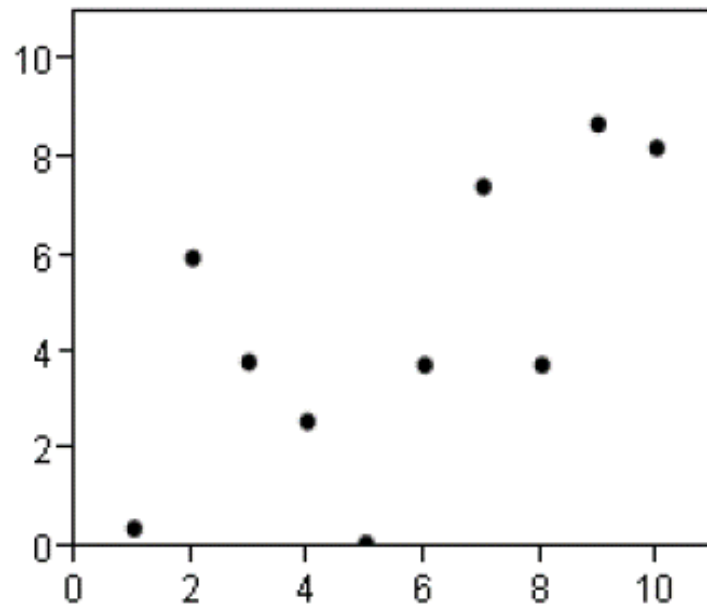


# Examples

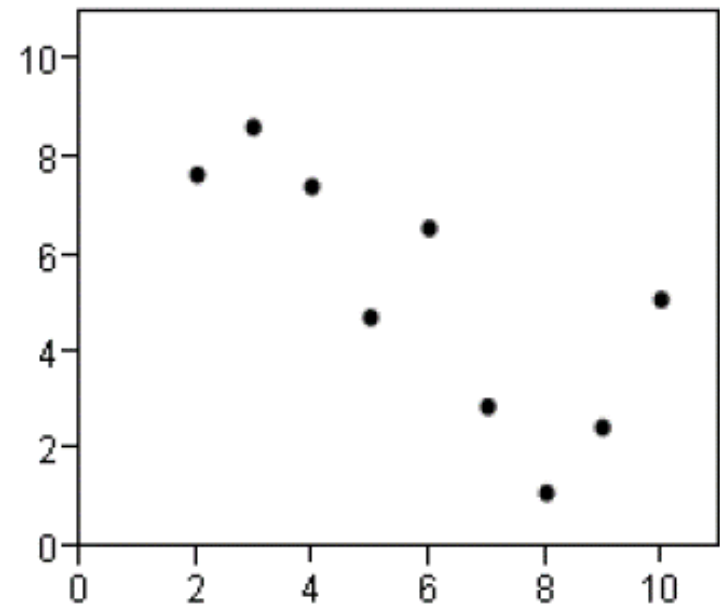


# Examples

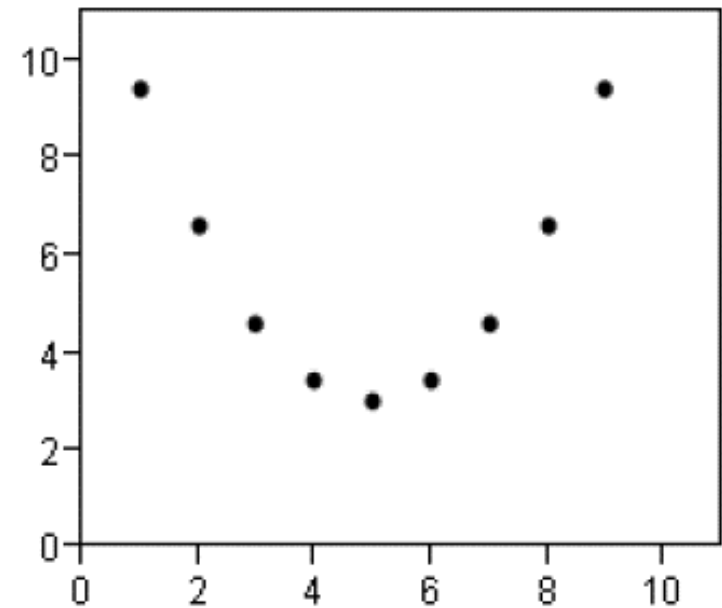
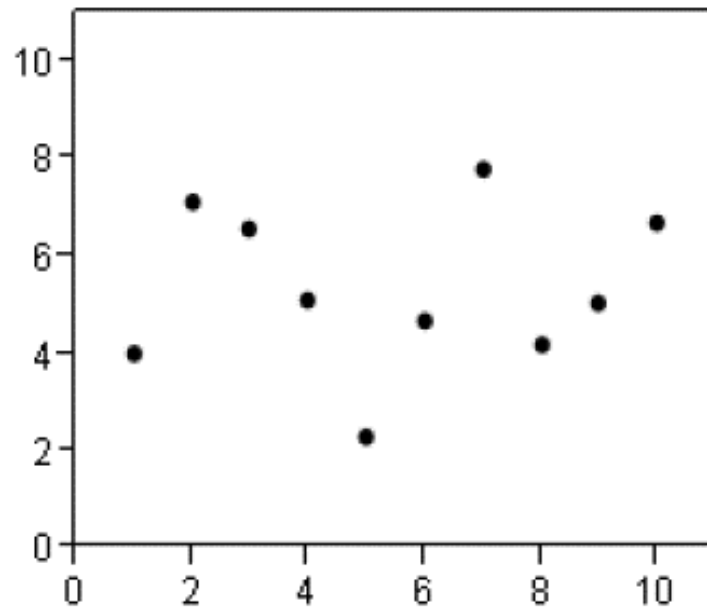
$r = 0.65$



$r = -0.83$

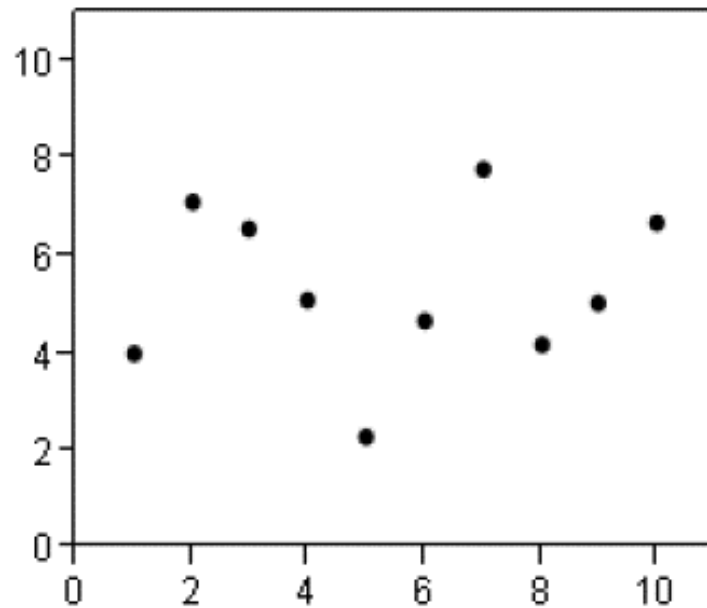


# Examples

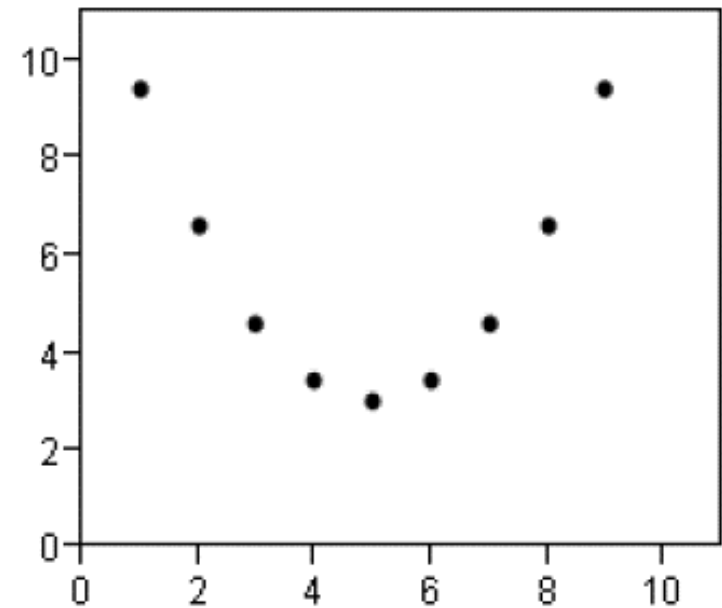


# Examples

$r = 0$



$r = 0$

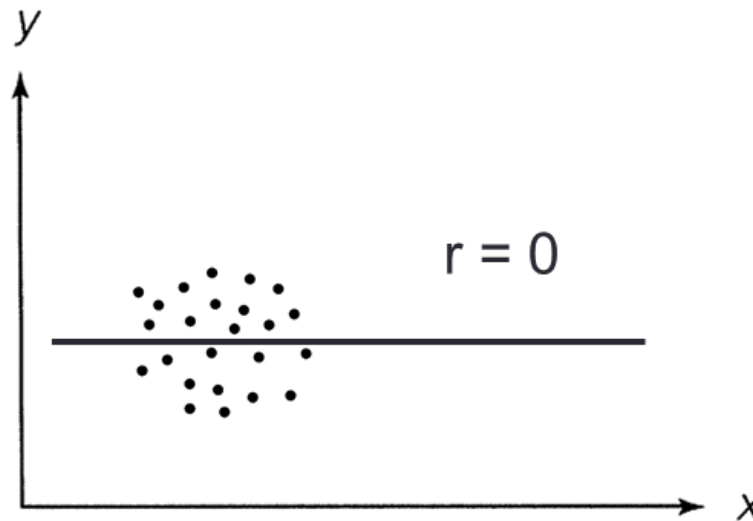


# Potential Issues with Correlation

- Two of the biggest problems with correlation are the following:
  1. Outliers
  2. Causation

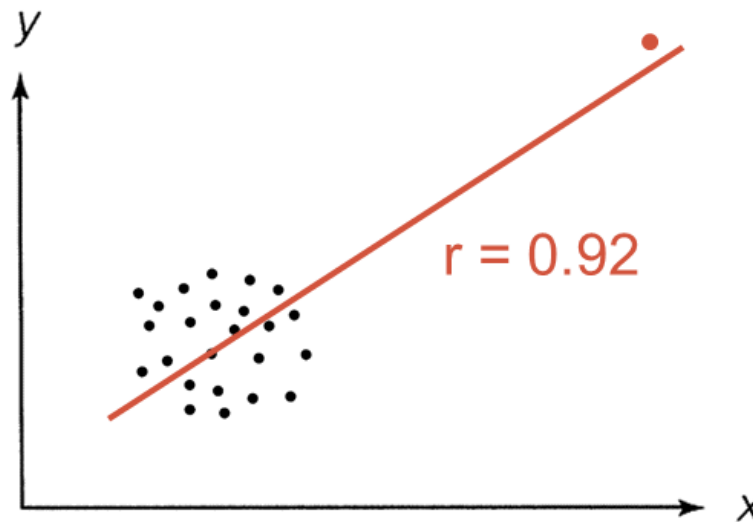
# Outliers in Correlation

- Outliers can lead to false conclusions about correlation if you don't visualize the data to help us see what might be going on.
- Outliers can make relationships that aren't really there.



# Outliers in Correlation

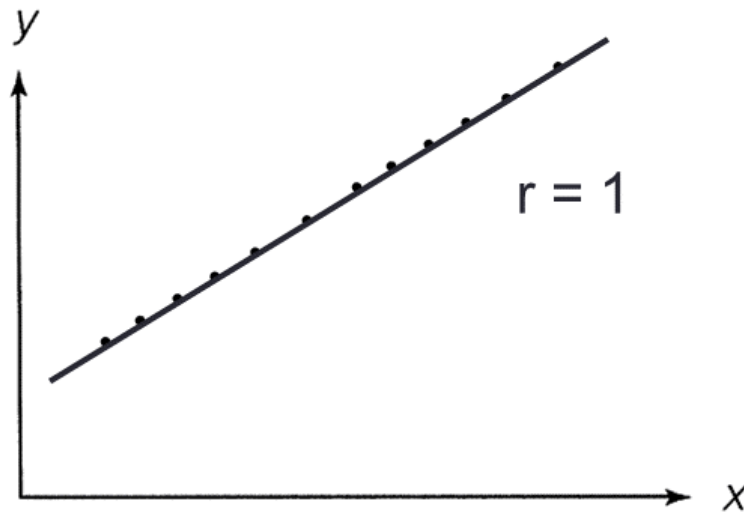
- Outliers can lead to false conclusions about correlation if you don't visualize the data to help us see what might be going on.
- Outliers can **make** relationships that **aren't** really there.





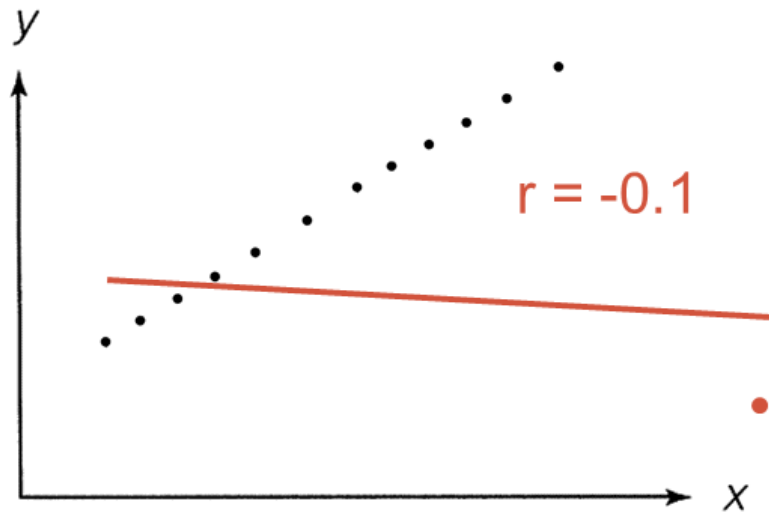
# Outliers in Correlation

- Outliers can lead to false conclusions about correlation if you don't visualize the data to help us see what might be going on.
- Outliers can hide relationships that are really there.



# Outliers in Correlation

- Outliers can lead to false conclusions about correlation if you don't visualize the data to help us see what might be going on.
- Outliers can **hide** relationships that **are** really there.



# Causation?

- Legalized gambling is available on different riverboats casinos operating in a city in Mississippi. The mayor wants to know the correlation between casino employees and crime rate.

# Causation?

- Legalized gambling is available on different riverboats casinos operating in a city in Mississippi. The mayor wants to know the correlation between casino employees and crime rate.
- Correlation **does not imply** causation.

# Causation?

- Legalized gambling is available on different riverboats casinos operating in a city in Mississippi. The mayor wants to know the correlation between casino employees and crime rate.
- Correlation does not imply causation.
- Just because the crime rate and number of casino employees are correlated, this doesn't imply that more casino employee hires causes more crime.

# Correlation vs. Causation

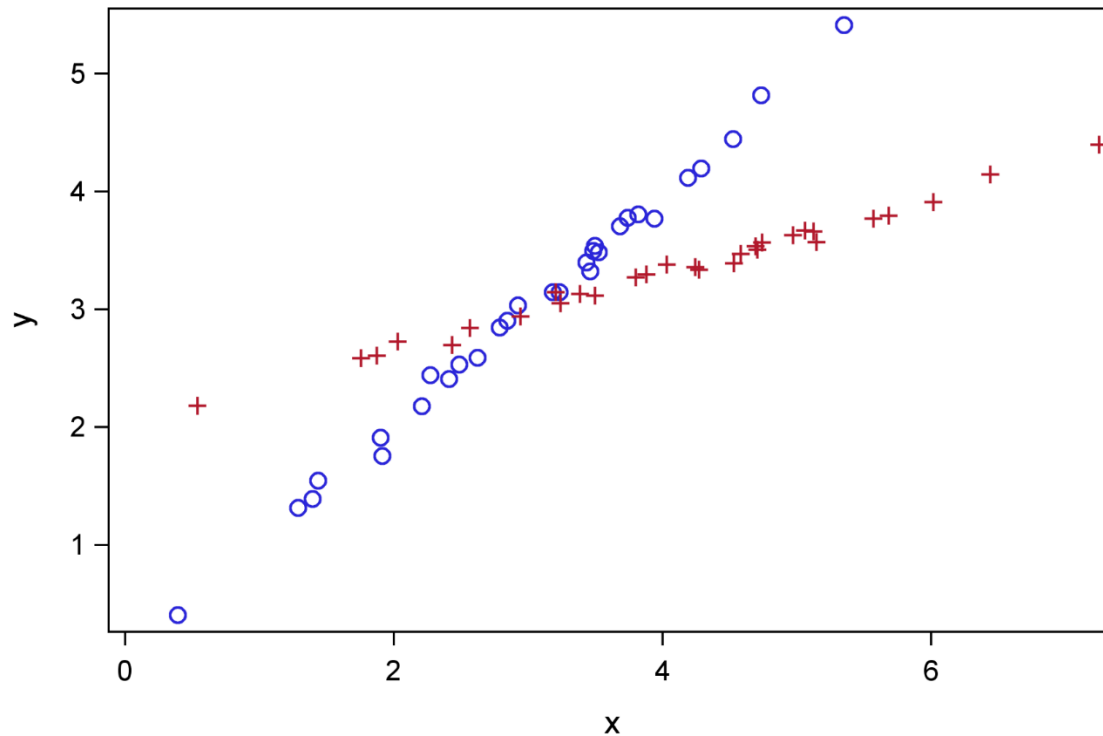
- Confusing correlation and causation is a common phenomena.
- All correlation implies is a linear trend may exist between two variables of interest.
- Many famous examples of correlations that are **not** causations.

# SIMPLE LINEAR REGRESSION

---

# Correlation Isn't Everything

- Plot below has two sets of data with exact same correlation.





# Regression Modeling

- Many people across industries devote research funding to discover how variables are related (modeling).
- The simplest graphical technique to relating a response variable,  $y$ , to an independent variable,  $x$ , is through a **straight-line relationship**.
- This section will focus on the **simple linear regression (SLR) model**.
- Most models are more extensive and complicated than SLR models, but SLR models form a good foundation.

# Example

- Suppose you want to model monthly sales revenue,  $y$ , of an appliance store as a function of advertising expenditure,  $x$ .
  - Will advertising expenditure perfectly predict sales revenue?
  - Is a perfect prediction of sales revenue even possible?
  - What other factors might we consider?

# Example

- Suppose you want to model monthly sales revenue,  $y$ , of an appliance store as a function of advertising expenditure,  $x$ .
  - Will advertising expenditure perfectly predict sales revenue?
  - Is a perfect prediction of sales revenue even possible?
  - What other factors might we consider?

Focus of Next Section



# Simple Linear Regression Model

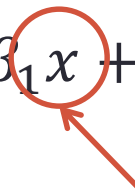
- Population Simple Linear Regression:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Response variable

# Simple Linear Regression Model

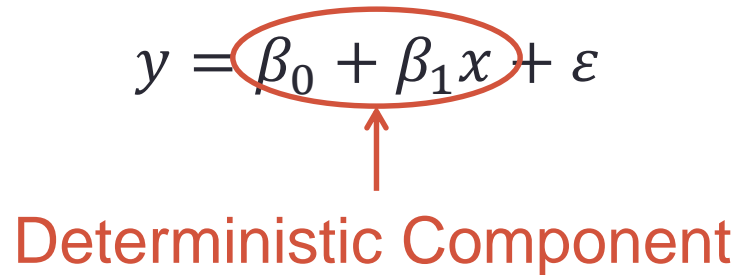
- Population Simple Linear Regression:

$$y = \beta_0 + \beta_1 x + \varepsilon$$


Independent variable

# Simple Linear Regression Model


- Population Simple Linear Regression:

$$y = \beta_0 + \beta_1 x + \varepsilon$$


Deterministic Component

# Simple Linear Regression Model

- Population Simple Linear Regression:

$$y = \beta_0 + \beta_1 x + \varepsilon$$


Random Component

# Simple Linear Regression Model

- Population Simple Linear Regression:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

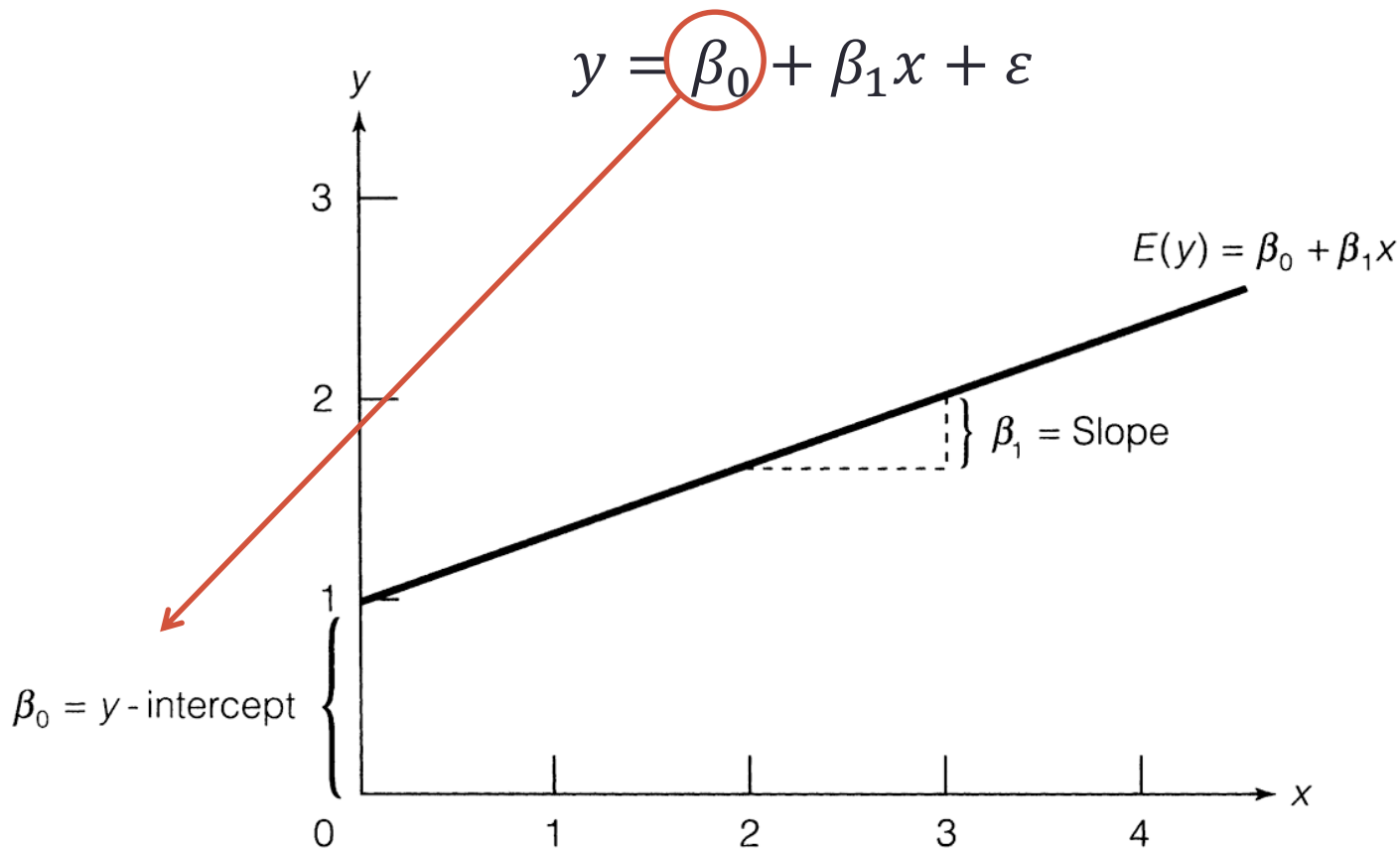
Intercept

Slope



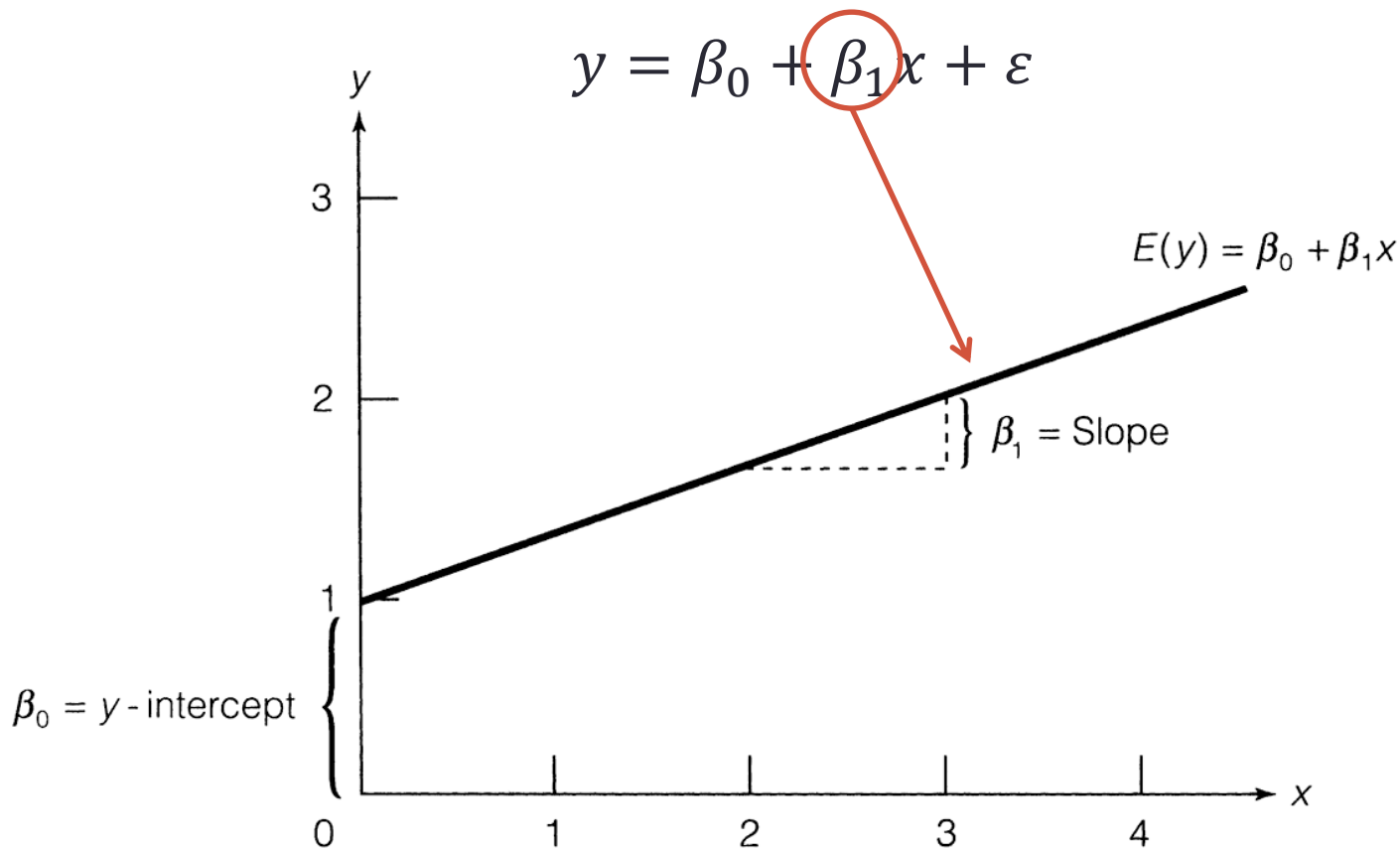
# Simple Linear Regression Model

- Population Simple Linear Regression:



# Simple Linear Regression Model

- Population Simple Linear Regression:



# Simple Linear Regression Model

- The intercept is the value of the average of the response variable,  $y$ , when the independent variable,  $x$ , equals zero.
- The slope is the **average** increase in the value of the response variable,  $y$ , with a one-unit increase in the independent variable,  $x$ .

# Assumptions

- There are four main assumptions:

# Assumptions

- There are four main assumptions:
  1. Linearity of the mean.

# Assumptions

- There are four main assumptions:
  1. Linearity of the mean.
  2. The variance of the probability distribution of  $\varepsilon$  (usually denoted  $\sigma^2$ ) is constant.

# Assumptions

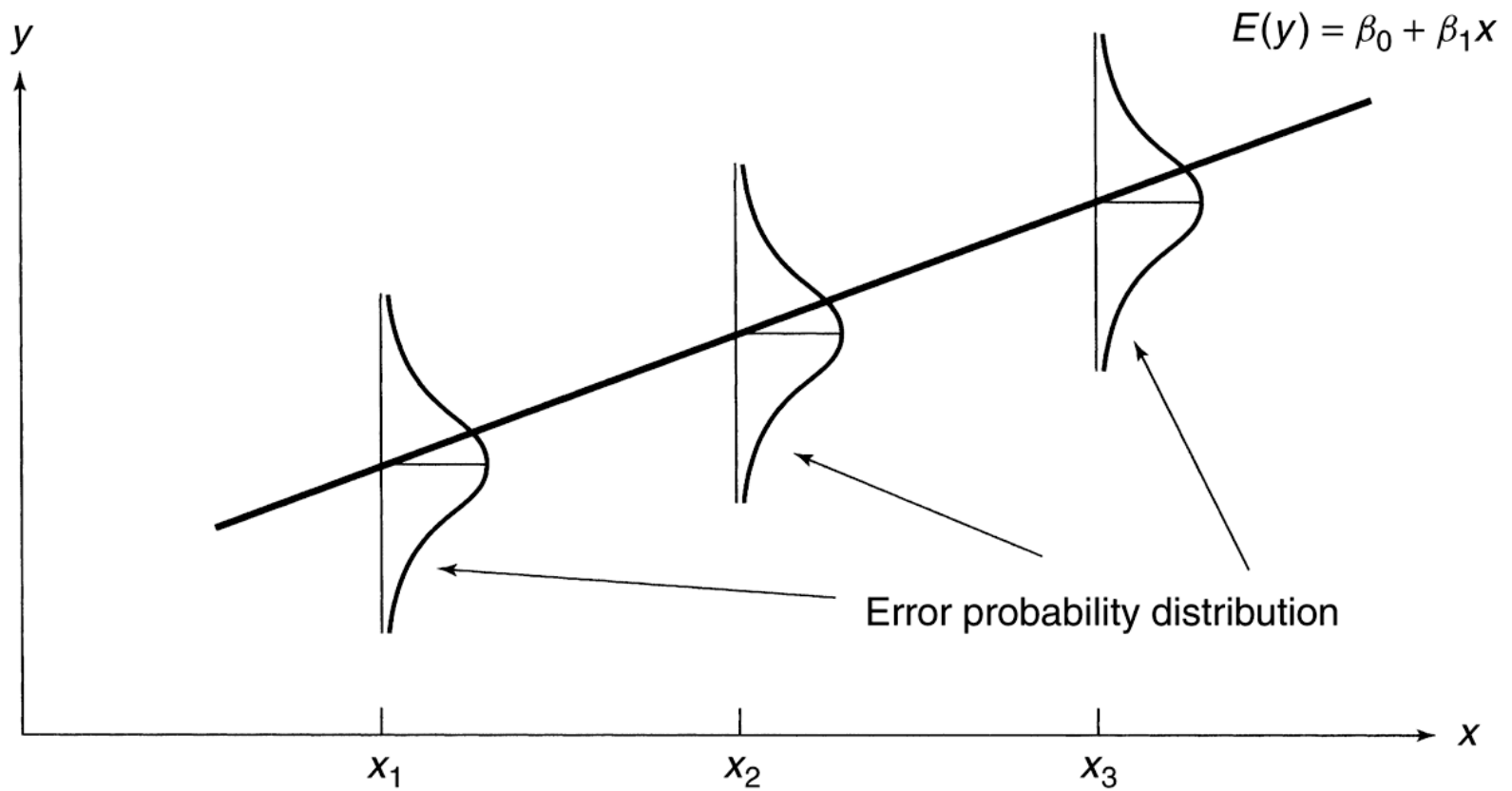
- There are four main assumptions:
  1. Linearity of the mean.
  2. The variance of the probability distribution of  $\varepsilon$  (usually denoted  $\sigma^2$ ) is constant.
  3. The probability distribution of  $\varepsilon$  is Normal.

# Assumptions

- There are four main assumptions:
  1. Linearity of the mean.
  2. The variance of the probability distribution of  $\varepsilon$  (usually denoted  $\sigma^2$ ) is constant.
  3. The probability distribution of  $\varepsilon$  is Normal.
  4. The errors associated with any two different observations are independent of each other.



# Assumptions



# Assumptions

- These assumptions should be tested and diagnosed in **every** regression model.
- The techniques used to diagnose the validity of the assumptions will be discussed later.

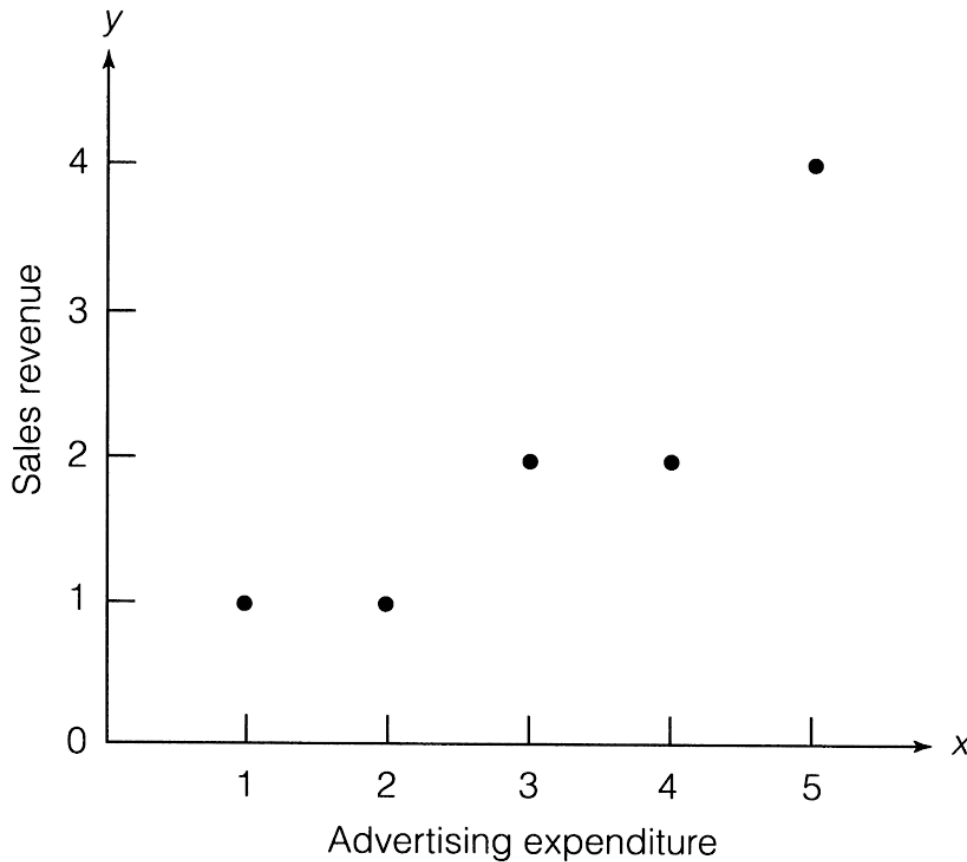
# SIMPLE LINEAR REGRESSION

---

Least Squares Method

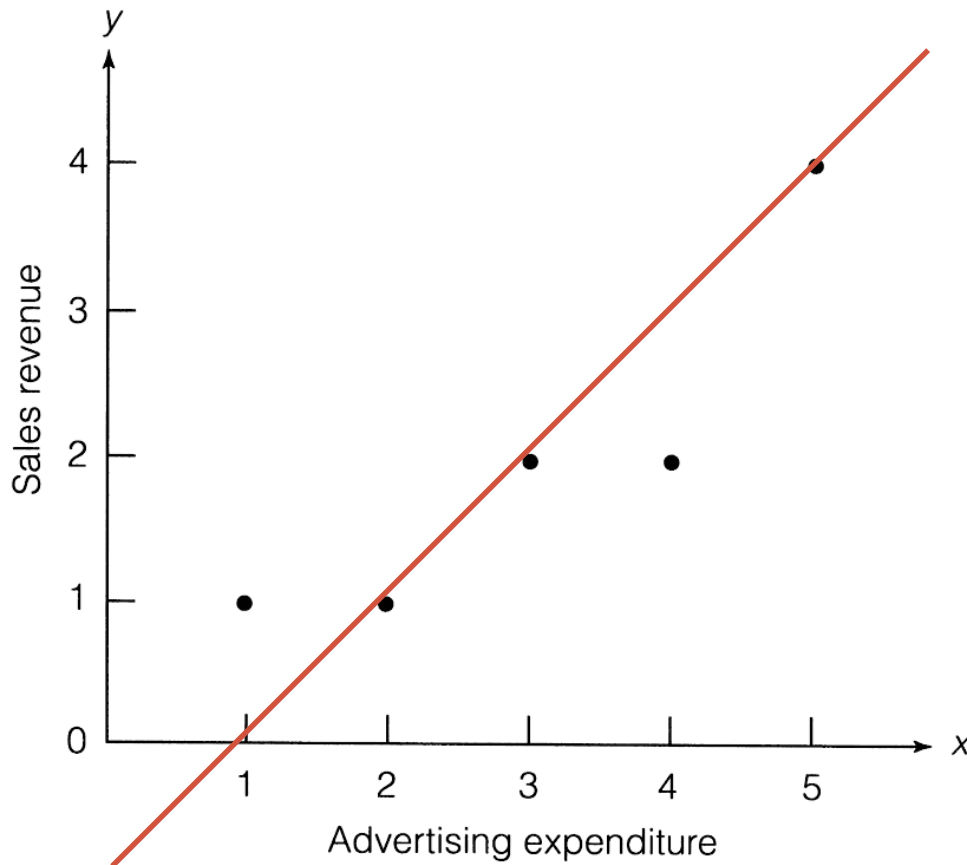
# Example

- Predicting sales revenue (thousands of \$) with advertising expenditure (hundreds of \$).



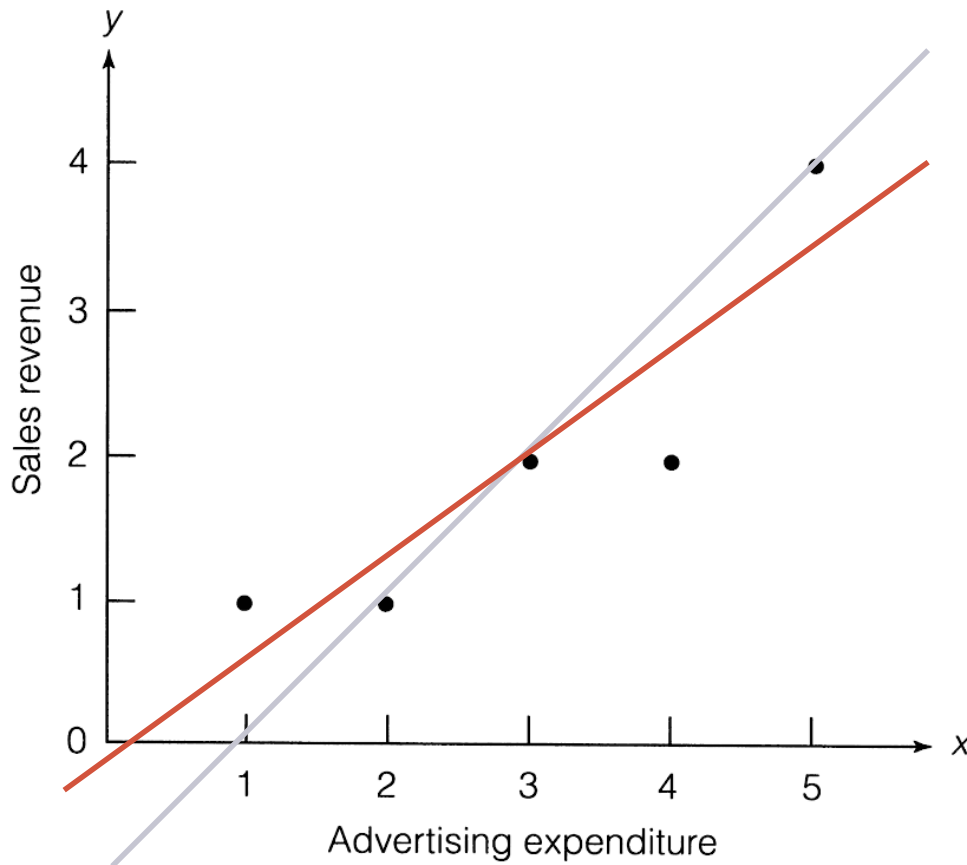
# Example

- Predicting sales revenue (thousands of \$) with advertising expenditure (hundreds of \$).



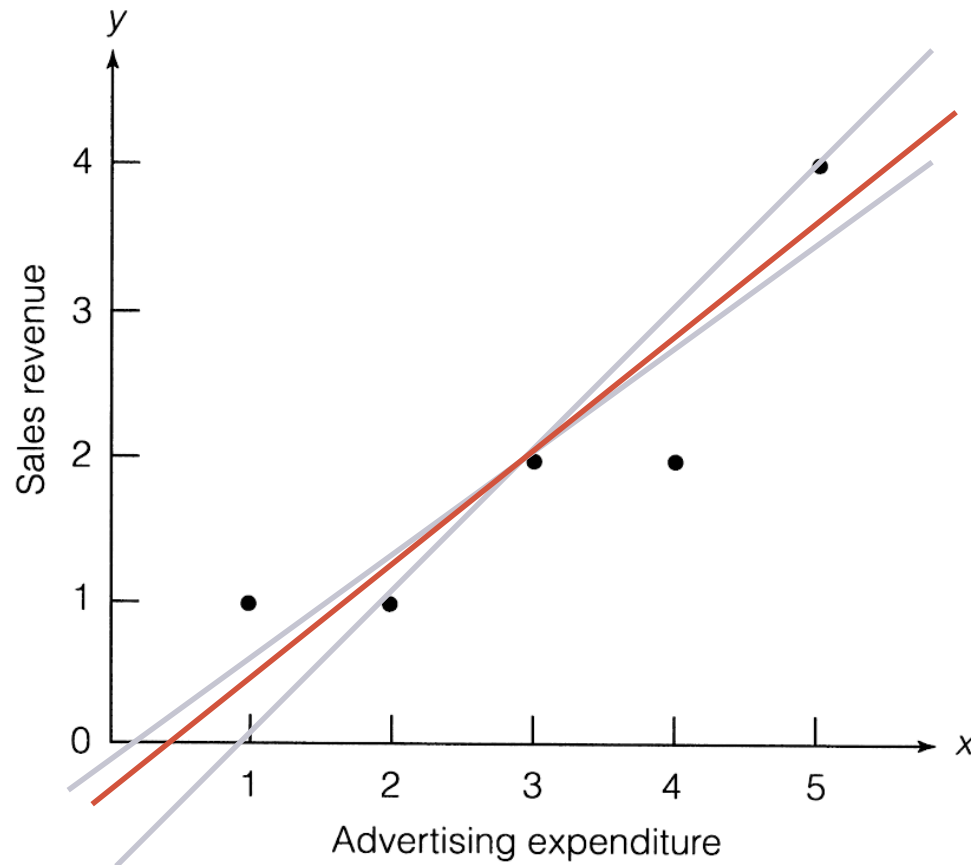
# Example

- Predicting sales revenue (thousands of \$) with advertising expenditure (hundreds of \$).



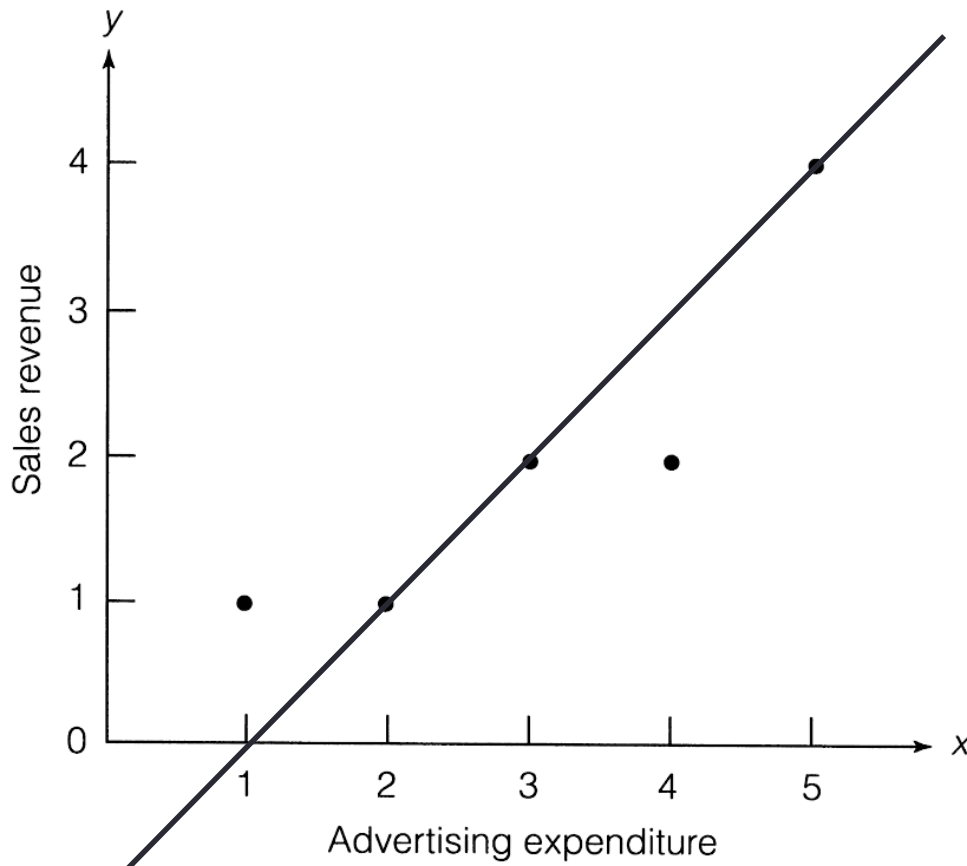
# Example

- Predicting sales revenue (thousands of \$) with advertising expenditure (hundreds of \$).



# Example

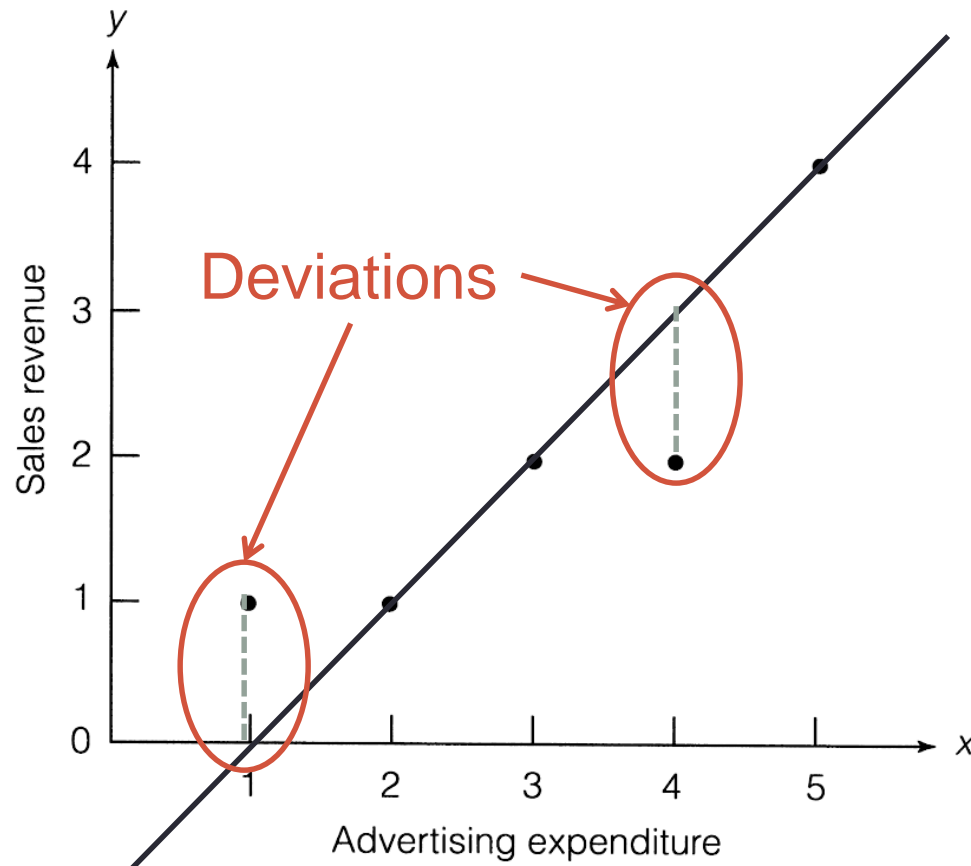
- Predicting sales revenue (thousands of \$) with advertising expenditure (hundreds of \$).





# Example

- Predicting sales revenue (thousands of \$) with advertising expenditure (hundreds of \$).



# Sum of Squared Errors

- The deviations (or **errors of prediction**) can be summed up to calculate a total error.
- These errors are also called **residuals**:  $\hat{\varepsilon} = y_i - \hat{y}_i$

# Sum of Squared Errors

- The deviations (or **errors of prediction**) can be summed up to calculate a total error.
- These errors are also called **residuals**:  $\hat{\varepsilon} = y_i - \hat{y}_i$
- However, these errors have both positive and negative values that could cancel each other out.
- Summing the squared errors (**SSE**) puts greater emphasis on the deviations of the data points.

# Least Squares Regression

- It can be shown that there is **only** one line for which the SSE is minimized.
- This line is called the **sample simple linear regression line** (or **line of best fit**).

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

# Least Squares Regression

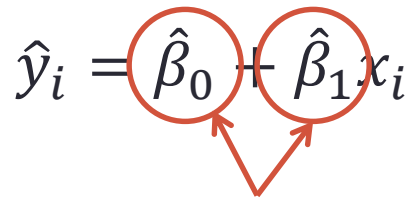
- It can be shown that there is **only** one line for which the SSE is minimized.
- This line is called the **sample simple linear regression line** (or **line of best fit**).

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Predicted value for  $y$  (estimate of  $E(y)$ )

# Least Squares Regression

- It can be shown that there is **only** one line for which the SSE is minimized.
- This line is called the **sample simple linear regression line** (or **line of best fit**).

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$


Estimates for intercept and slope of line

# Least Squares Regression

- The minimization of the SSE leads us to calculations for the estimators  $\hat{\beta}_0$  and  $\hat{\beta}_1$ .

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$$

# Least Squares Regression

- The minimization of the SSE leads us to calculations for the estimators  $\hat{\beta}_0$  and  $\hat{\beta}_1$ .

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$$

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

$$SS_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$$



# Least Squares Regression

- The minimization of the SSE leads us to calculations for the estimators  $\hat{\beta}_0$  and  $\hat{\beta}_1$ .

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = r \times \frac{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} = r \times \frac{s_y}{s_x}$$

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

$$SS_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$$

# Least Squares Regression

- The minimization of the SSE leads us to calculations for the estimators  $\hat{\beta}_0$  and  $\hat{\beta}_1$ .

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = r \times \frac{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} = r \times \left( \frac{s_y}{s_x} \right)$$

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

$$SS_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$$

Always positive!

# Least Squares Regression

- The minimization of the SSE leads us to calculations for the estimators  $\hat{\beta}_0$  and  $\hat{\beta}_1$ .

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = r \times \frac{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} = r \times \frac{s_y}{s_x}$$

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

$$SS_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$$

**MUST have  
same sign!**

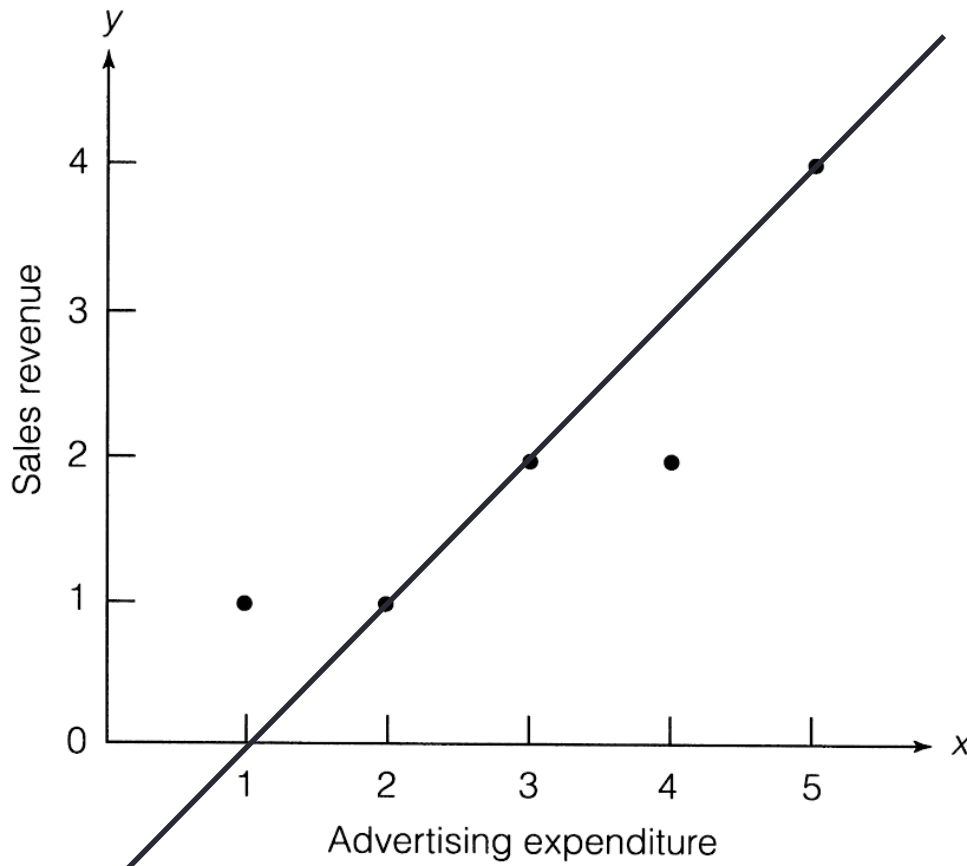
# Least Squares Regression

- The minimization of the SSE leads us to calculations for the estimators  $\hat{\beta}_0$  and  $\hat{\beta}_1$ .

$$\hat{\beta}_1 = r \times \frac{s_y}{s_x} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

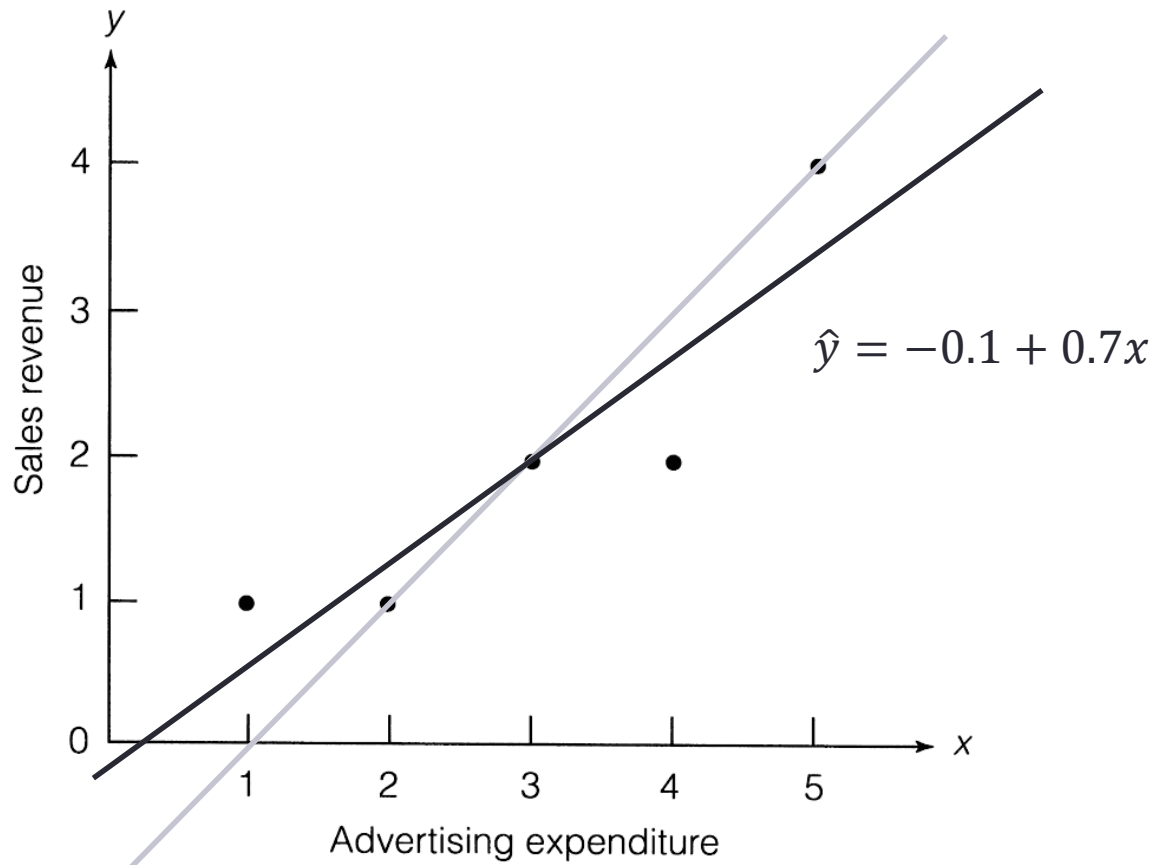
# Example

- Predicting sales revenue (thousands of \$) with advertising expenditure (hundreds of \$).



# Example

- Predicting sales revenue (thousands of \$) with advertising expenditure (hundreds of \$).



# Least Squares Regression

- These predicted values help determine the deviations seen earlier in the graph of the data.
- These deviations are more often called **residuals**:

$$y_i - \hat{y}_i$$

- The goal is to minimize the sum of the squared residuals (or SSE):

$$\text{SSE} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n \left( y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right)^2$$

# SIMPLE LINEAR REGRESSION

---

Coefficient of Determination



# Utility of a Model

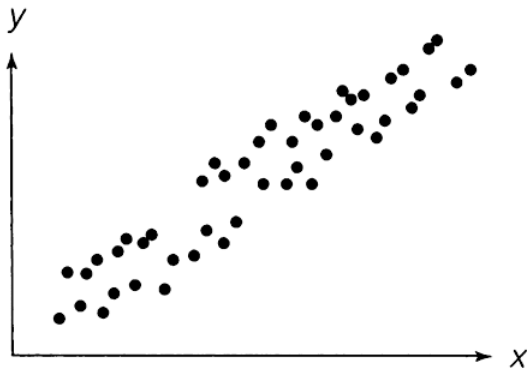
- One method to determine the utility of a model is to compute the reduction in the prediction errors of  $y$  due to the additional information provided by  $x$ .
- If  $x$  provides no additional information in predicting  $y$ , then the basic model,  $\hat{y} = \bar{y}$ , is the best prediction of  $y$ .

$$TSS = SS_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2$$

# Utility of a Model

- This can be visually seen through the following charts:

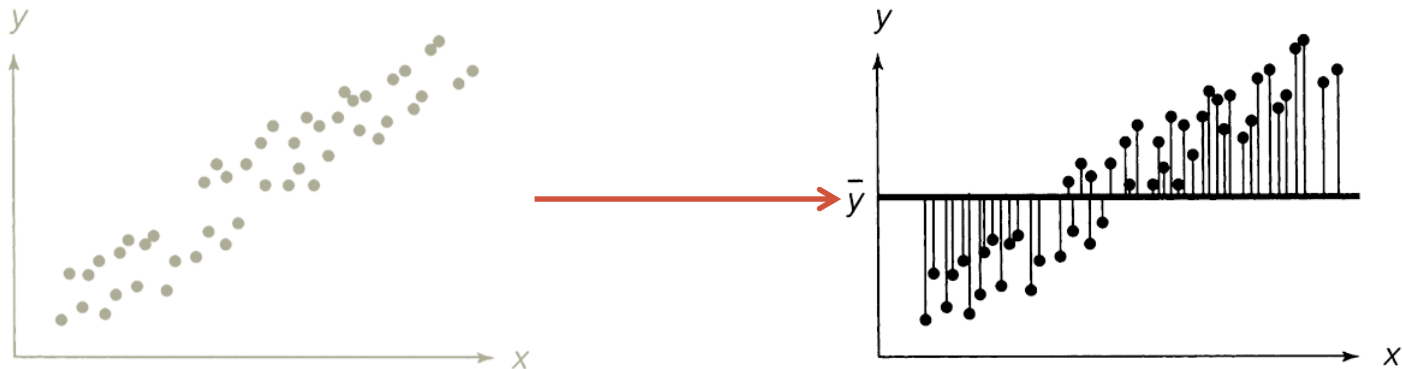
$$TSS = \sum_{i=1}^n (y_i - \bar{y})^2$$



# Utility of a Model

- This can be visually seen through the following charts:

$$TSS = \sum_{i=1}^n (y_i - \bar{y})^2$$



# Utility of a Model

- If a least squares line is fit to the data, the  $SSE$  measures the deviations of the new predicted values ( $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ ) with the actual values of  $y$ .

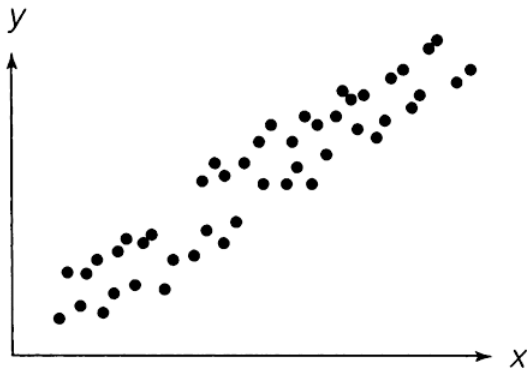
$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

- If  $x$  provides no information to the prediction of  $y$ , then  $TSS$  and  $SSE$  are approximately equal.
- If  $x$  contributes information then  $SSE < TSS$ .

# Utility of a Model

- This can be visually seen through the following charts:

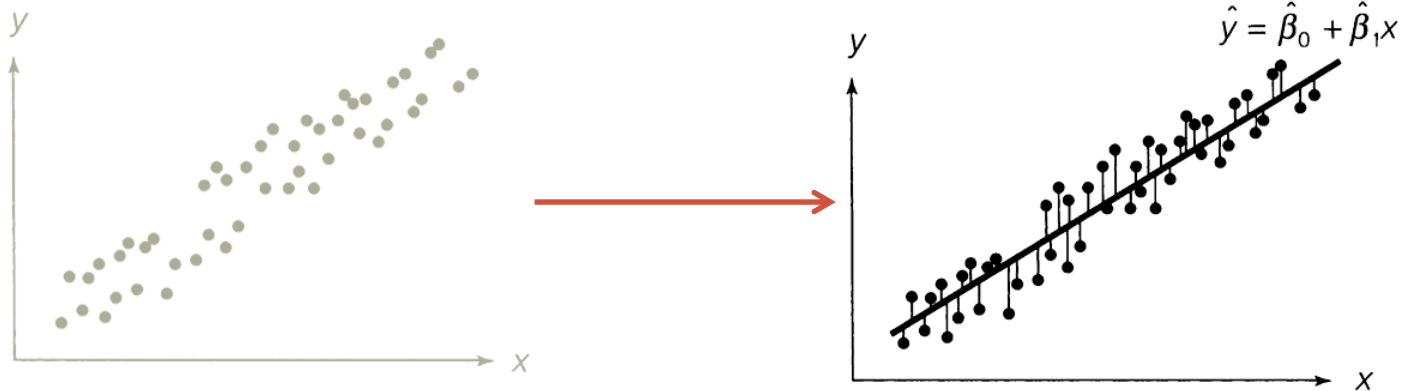
$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$



# Utility of a Model

- This can be visually seen through the following charts:

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



# Utility of a Model

- The difference between the SSE and TSS is the amount of variability that is explained by the model – typically denoted as SSM or SSR:

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

# Coefficient of Determination

- The reduction of  $TSS$  by  $SSE$  expressed as a proportion of  $SSE$  reveals the utility of having the straight-line model.
- This proportion is called the **coefficient of determination**:

$$r^2 = R^2 = \frac{TSS - SSE}{TSS} = 1 - \frac{SSE}{TSS} = \frac{SSR}{TSS}$$



# Coefficient of Determination

- The reduction of  $TSS$  by  $SSE$  expressed as a proportion of  $SSE$  reveals the utility of having the straight-line model.
- This proportion is called the **coefficient of determination**:

$$r^2 = R^2 = \frac{TSS - SSE}{TSS} = 1 - \frac{SSE}{TSS}$$

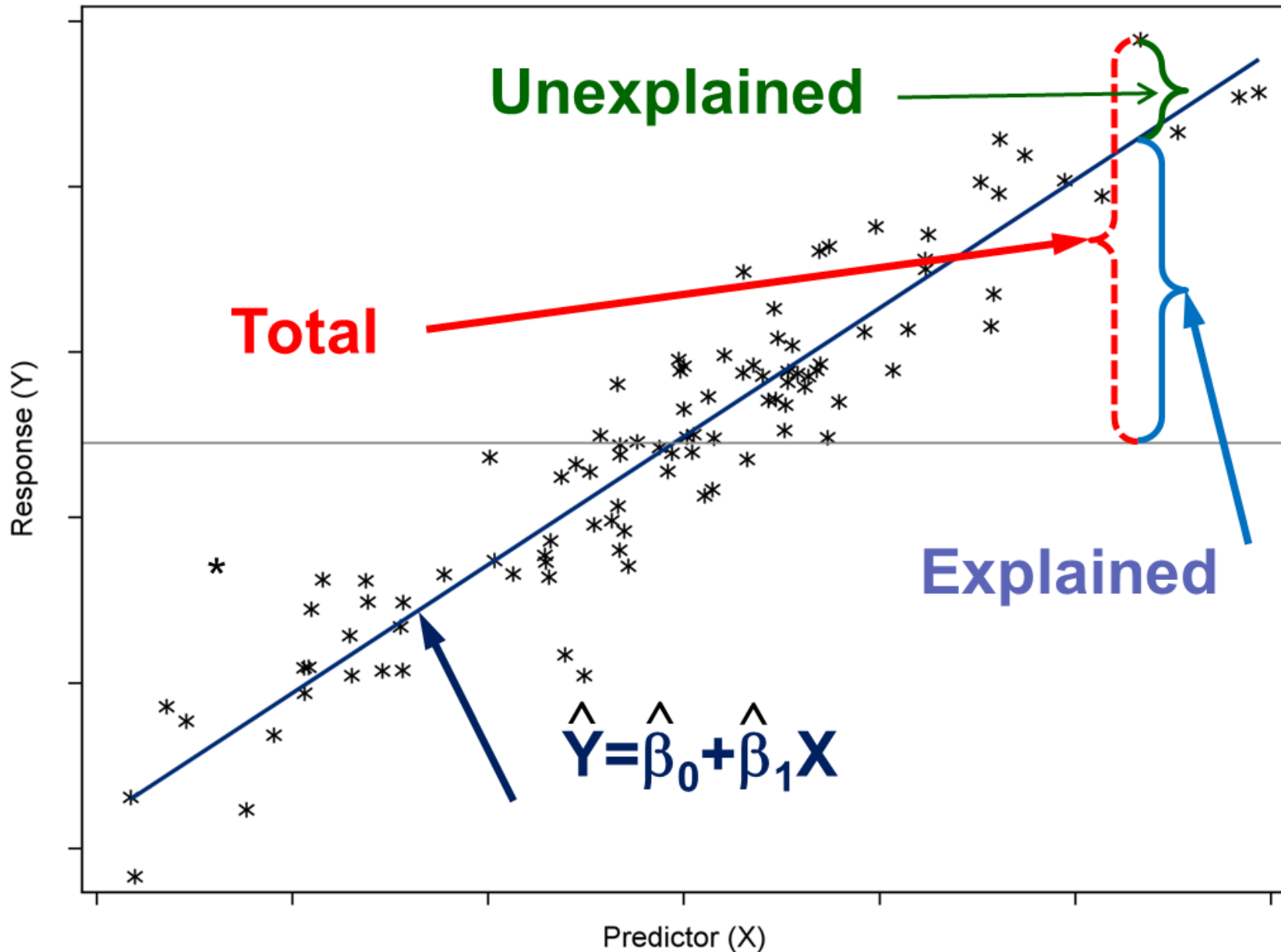
Square of correlation coefficient between  
x and y only in SLR.

# Coefficient of Determination

- Due to  $SSE \leq TSS$ ,  $0 \leq r^2 \leq 1$ .
- There is a useful interpretation to  $r^2$ .

$$R^2 = \frac{TSS - SSE}{TSS} = \frac{\text{Explained sample variability}}{\text{Total sample variability}}$$

- About  $100(R^2)\%$  of the variation in  $y$  can be attributed to using  $x$  as a predictor in a linear model.



# Example

- An analyst for the State of North Carolina has collected data on salaries and years of education with the intention of finding out how education affects the salary an individual makes monthly (in dollars). The correlation between the two variables is 0.327. The simple linear regression line between these two variables is the following:

$$\hat{y} = 66.271 + 66.054x$$

# Example

$$\hat{y} = 66.271 + 66.054x$$

1. Interpret the values of the coefficients.
2. How much salary per month do you expect someone with only a high school education to earn monthly (13 years of education)?
3. How much more salary can a person earn monthly if they go to school 4 more years after high school?
4. Calculate and interpret  $R^2$ .

# Example

$$\hat{y} = 66.271 + 66.054x$$

1. Interpret the values of the coefficients.
  - The average salary of people with no education is \$66.27/month.
  - The average increase in salary for every additional year of education is \$66.05/month.
2. How much salary per month do you expect someone with only a high school education to earn monthly (13 years of education)?

$$\hat{y} = 66.271 + 66.05(13) = \$924.97$$

# Example

$$\hat{y} = 66.271 + 66.054x$$

3. How much more salary can a person earn monthly if they go to school 4 more years after high school?

$$66.054(4) = \$264.22$$

4. Calculate and interpret  $R^2$ .

$$R^2 = r^2 = 0.327^2 = 0.107$$

- 10.7% of the variation in monthly salary can be explained by the relationship with years of education.

# SIMPLE LINEAR REGRESSION

---

Regression Inference



# Example

- Predicting sales revenue (thousands of \$) with advertising expenditure (hundreds of \$).
- What would the theorized regression line be for a model where advertising expenditure is not related to sales revenue?

# Example

- Predicting sales revenue (thousands of \$) with advertising expenditure (hundreds of \$).
- What would the theorized regression line be for a model where advertising expenditure is not related to sales revenue?

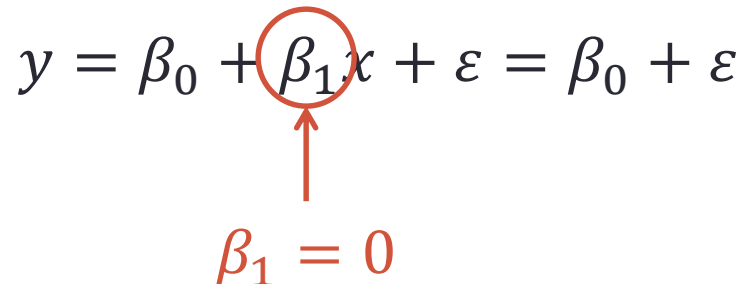
$$y = \beta_0 + \beta_1 x + \varepsilon$$

# Example

- Predicting sales revenue (thousands of \$) with advertising expenditure (hundreds of \$).
- What would the theorized regression line be for a model where advertising expenditure is not related to sales revenue?

$$y = \beta_0 + \beta_1 x + \varepsilon = \beta_0 + \varepsilon$$

$\beta_1 = 0$



# Hypothesis Test for $\beta_1$

- The parameter  $\beta_1$  is not directly observed.
- Need to test if the value of the coefficient is zero to determine if a relationship exists between the response variable  $y$  and the explanatory variable  $x$ .

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0$$

# Hypothesis Test for $\beta_1$

- Able to do hypothesis tests for mean  $\mu$  because the sampling distribution of  $\bar{x}$  is known (based on CLT).
- To test  $\beta_1$  need to know the sampling distribution of  $\hat{\beta}_1$ .

<http://www.rossmanchance.com/applets/RegSim/RegCoeff.html>

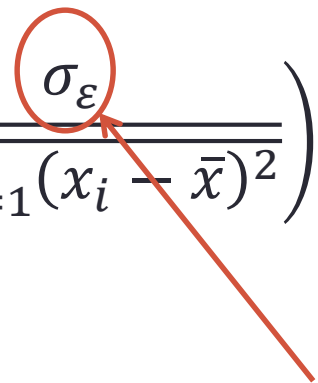
# Hypothesis Test for $\beta_1$

- The sampling distribution of  $\hat{\beta}_1$  is the following normal distribution **IF** we make the assumptions discussed previously.

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma_\varepsilon}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}\right)$$

# Hypothesis Test for $\beta_1$

- The sampling distribution of  $\hat{\beta}_1$  is the following normal distribution **IF** we make the assumptions discussed previously.

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma_\varepsilon}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}\right)$$


Don't know true  
variance of errors!

# Estimation of $\sigma_\varepsilon^2$

- The population parameter  $\sigma_\varepsilon^2$  can not be directly measured.
- Estimate  $\sigma_\varepsilon^2$  with  $s_\varepsilon^2$  where:

$$s_\varepsilon^2 = \frac{SSE}{n - k - 1} = \frac{\sum_{i=1}^n (y_i - \hat{y})^2}{n - k - 1}$$

$$s_\varepsilon = \sqrt{s_\varepsilon^2}$$



# Estimation of $\sigma_\varepsilon^2$

- The population parameter  $\sigma_\varepsilon^2$  can not be directly measured.
- Estimate  $\sigma_\varepsilon^2$  with  $s_\varepsilon^2$  where:

$$s_\varepsilon^2 = \frac{SSE}{n - k - 1} = \frac{\sum_{i=1}^n (y_i - \hat{y})^2}{n - k - 1}$$

$$s_\varepsilon = \sqrt{s_\varepsilon^2}$$

Number of  
independent  
variables

# Hypothesis Test for $\beta_1$

- The sampling distribution of  $\hat{\beta}_1$  is the following normal distribution **IF** we make the assumptions discussed previously.

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma_\varepsilon}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}\right)$$

Estimated as  $s_{\hat{\beta}_1}$

$$s_{\hat{\beta}_1} = \frac{s_\varepsilon}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

# Hypothesis Test for $\beta_1$

- The test statistic for the hypothesis test is the following:

$$t = \frac{\hat{\beta} - 0}{s_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{\left( \frac{s_{\varepsilon}}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} \right)}$$

$$d.f. = n - 2$$

# Hypothesis Test for $\beta_1$

- The test statistic for the hypothesis test is the following:

$$t = \frac{\hat{\beta} - 0}{s_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{\left( \frac{s_{\varepsilon}}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} \right)}$$

$$d.f. = n - 2$$

Hypothesized value of  $\beta_1$

# Hypothesis Test for $\beta_1$

- The test statistic for the hypothesis test is the following:

$$t = \frac{\hat{\beta} - 0}{s_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{\left( \frac{s_{\varepsilon}}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} \right)}$$

$$d.f. = n - 2$$

- The remaining steps of the hypothesis test are the same as any other.

# Confidence Interval for $\beta_1$

- If we can have a hypothesis test, then we can also create a confidence interval:

$$\hat{\beta}_1 \pm t^* \times s_{\hat{\beta}_1}$$

# Example

- A clothing manufacturer wants to understand the relationship between height and weight of American men between the ages of 18 and 24 when designing their new line of clothing. They randomly samples 101 American males of that age group to collect their data. From this data they determined the following:

$$\hat{y} = -167 + 4.7x \quad \sum(x - \bar{x})^2 = 900 \quad SSE = 259320$$

# Example

$$\hat{y} = -167 + 4.7x \quad \sum(x - \bar{x})^2 = 900 \quad SSE = 259320$$

1. Calculate a hypothesis test to determine if the slope of the true regression line equals zero with  $\alpha = 0.05$ .



# Example

$$\hat{y} = -167 + 4.7x \quad \sum(x - \bar{x})^2 = 900 \quad SSE = 259320$$

1. Calculate a hypothesis test to determine if the slope of the true regression line equals zero with  $\alpha = 0.05$ .

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$
$$t = \frac{4.7 - 0}{\left( \frac{51.18}{\sqrt{900}} \right)} = 2.75$$

$$s_\varepsilon = \sqrt{\frac{259320}{101 - 1 - 1}} = 51.18$$

$$\text{P-value} = (0.002, 0.01)$$

REJECT  $H_0$

# Example

$$\hat{y} = -167 + 4.7x \quad \sum(x - \bar{x})^2 = 900 \quad SSE = 259320$$

2. Create a 90% confidence interval for the slope.
3. Can you compare these results?

# Example

$$\hat{y} = -167 + 4.7x \quad \sum(x - \bar{x})^2 = 900 \quad SSE = 259320$$

2. Create a 90% confidence interval for the slope.

$$\hat{\beta}_1 \pm t^* \times s_{\hat{\beta}_1}$$

$$4.7 \pm 1.662 \times 1.706$$

$$4.7 \pm 2.835$$

3. Can you compare these results?

$$\text{NO! } C \neq 1 - \alpha$$

# SIMPLE LINEAR REGRESSION

---

Complete Example

# Example

- The director of admissions of a small college in the Midwest has hired you as an analyst to administer a newly designed entrance test. This test ranges from a score of 1-7. You administer the test to 213 students selected randomly from the new freshman class in a study to determine whether a student's GPA at the end of their freshman year can be predicted from the entrance test score. The sample's average GPA was 2.67 with a s.d. of 0.72. The average entrance test score was 4.81 with a s.d. of 0.69. The correlation between these two variables is 0.735. Use this information to answer the following questions...

# Example

1. Create the sample regression line for predicting GPA from the entrance test score.
2. What would a predicted GPA at the end of freshman year be for a student who scored a 6.1? How about a 2.9?
3. What would the expected increase in GPA at the end of freshman year be for an increase of 1.5 points on the entrance test?

# Example

1. Create the sample regression line for predicting GPA from the entrance test score.

$$\hat{\beta}_1 = 0.735 \times \frac{0.72}{0.69} = 0.767 \quad \hat{\beta}_0 = 2.67 - 0.767(4.81) = -1.019$$

2. What would a predicted GPA at the end of freshman year be for a student who scored a 6.1? How about a 2.9?

$$\hat{y}_{6.1} = -1.019 + 0.767(6.1) = 3.66 \quad \hat{y}_{2.9} = 1.21$$

3. What would the expected increase in GPA at the end of freshman year be for an increase of 1.5 points on the entrance test?

$$0.767(1.5) = 1.1505$$

# Example

4. State the hypothesis for a test to determine whether the slope of the true regression line is equal to zero.

5. Fill in the blanks for the following table:

Parameter	Estimate	Std. Error	t Value	P-value
Intercept		0.238997		
Slope		0.049238		

6. Summarize the results of the hypothesis test of the slope.



# Example

4. State the hypothesis for a test to determine whether the slope of the true regression line is equal to zero.

$$H_0: \beta_1 = 0 \quad H_a: \beta_1 \neq 0$$

5. Fill in the blanks for the following table:

Parameter	Estimate	Std. Error	t Value	P-value
Intercept	-1.019	0.238997	-4.264	< 0.001
Slope	0.767	0.049238	15.577	< 0.001

6. Summarize the results of the hypothesis test of the slope. Enough evidence to say there is a relationship between entrance test score and GPA.

# Example

7. What is the value and meaning of  $R^2$  in this problem?

# Example

7. What is the value and meaning of  $R^2$  in this problem?

$$R^2 = r^2 = 0.735^2 = 0.54$$

54% of the variation in freshman year GPA is explained by the entrance test score.

# MULTIPLE LINEAR REGRESSION MODEL

---

# Regression Modeling

- Most practical applications of regression modeling involve using more complicated models than the simple linear regression model.
- Typically it is better to have more than one variable in a regression model.
- Models with more than one predictor variable are called **multiple regression models**.

# Multiple Linear Regression (MLR)

- Models with more than one predictor variable are called **multiple regression models**.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

# Multiple Linear Regression (MLR)

- Models with more than one predictor variable are called **multiple regression models**.

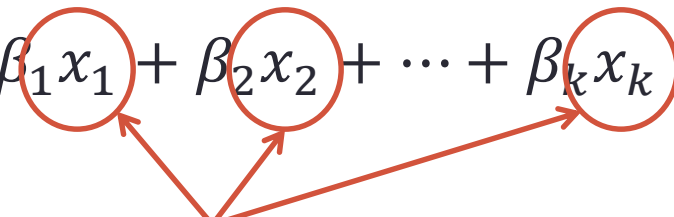
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$



Response variable

# Multiple Linear Regression (MLR)

- Models with more than one predictor variable are called **multiple regression models**.

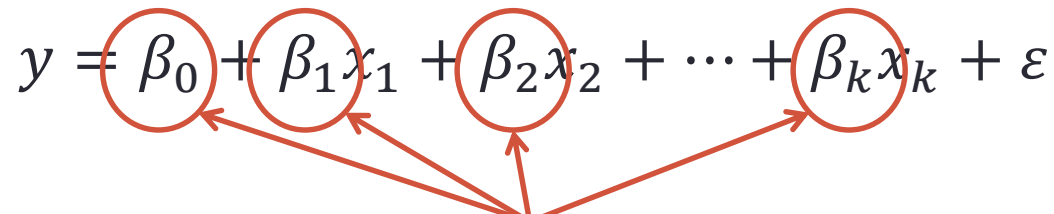
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$
A diagram consisting of three red circles highlighting the terms  $\beta_1 x_1$ ,  $\beta_2 x_2$ , and  $\beta_k x_k$  in the equation above. Three red arrows originate from a single point below the text 'Independent variables' and point to the center of each of the three circles.

Independent variables



# Multiple Linear Regression (MLR)

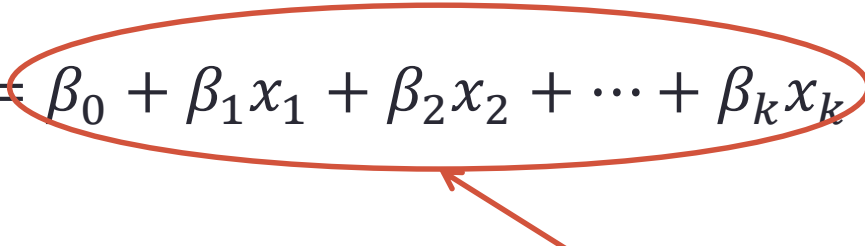
- Models with more than one predictor variable are called **multiple regression models**.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$


Coefficients on variables

# Multiple Linear Regression (MLR)

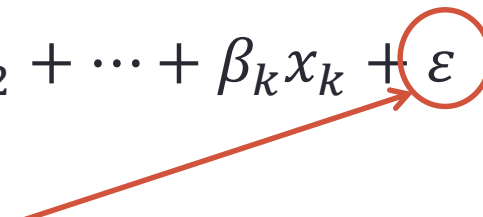
- Models with more than one predictor variable are called **multiple regression models**.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$


Deterministic component of model

# Multiple Linear Regression (MLR)

- Models with more than one predictor variable are called **multiple regression models**.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$


Random component of model

# MULTIPLE LINEAR REGRESSION MODEL

---

Model Assumptions

# MLR Assumptions

- The **random** portion of the model is the following:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

# Assumptions

- There are four main assumptions:

# Assumptions

- There are four main assumptions:
  1. The mean of the probability distribution of  $\varepsilon$  is 0.

# Assumptions

- There are four main assumptions:
  1. The mean of the probability distribution of  $\varepsilon$  is 0.
  2. The variance of the probability distribution of  $\varepsilon$  (usually denoted  $\sigma^2$ ) is constant.



# Assumptions

- There are four main assumptions:
  1. The mean of the probability distribution of  $\varepsilon$  is 0.
  2. The variance of the probability distribution of  $\varepsilon$  (usually denoted  $\sigma^2$ ) is constant.
  3. The probability distribution of  $\varepsilon$  is Normal.

# Assumptions

- There are four main assumptions:
  1. The mean of the probability distribution of  $\varepsilon$  is 0.
  2. The variance of the probability distribution of  $\varepsilon$  (usually denoted  $\sigma^2$ ) is constant.
  3. The probability distribution of  $\varepsilon$  is Normal.
  4. The errors associated with any two different observations are independent of each other.

# Assumptions

- There are four main assumptions:
  1. The mean of the probability distribution of  $\varepsilon$  is 0.
  2. The variance of the probability distribution of  $\varepsilon$  (usually denoted  $\sigma^2$ ) is constant.
  3. The probability distribution of  $\varepsilon$  is Normal.
  4. The errors associated with any two different observations are independent of each other.
- 5. No perfect collinearity.

# Assumptions

- There are four main assumptions:
  1. The mean of the probability distribution of  $\varepsilon$  is 0.
  2. The variance of the probability distribution of  $\varepsilon$  (usually denoted  $\sigma^2$ ) is constant.
  3. The probability distribution of  $\varepsilon$  is Normal.
  4. The errors associated with any two different observations are independent of each other.

5. No perfect collinearity.

Extra assumption  
sometimes seen.

# Sample MLR Model

- Just like in simple linear regression, we do not observe the true population regression line and instead have to estimate the response,  $\hat{y}_i$ , with the **sample multiple linear regression model**:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \cdots + \hat{\beta}_k x_{k,i}$$

# Fitting the Model

- The method for finding the line of best fit for multiple linear regression is the exact same for simple linear regression – the least squares method.
- The only thing that has changed is the predicted value of the response,  $\hat{y}_i$ .

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \cdots + \hat{\beta}_k x_{k,i}$$

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

# Fitting the Model

- The method for finding the line of best fit for multiple linear regression is the exact same for simple linear regression – the least squares method.
- The only thing that has changed is the predicted value of the response,  $\hat{y}_i$ .

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \cdots + \hat{\beta}_k x_{k,i}$$

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

Minimum is difficult to find by hand, or even represent in equations.

# Fitting the Model

- The method for finding the line of best fit for multiple linear regression is the exact same for simple linear regression – the least squares method.
- The only thing that has changed is the predicted value of the response,  $\hat{y}_i$ .

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \cdots + \hat{\beta}_k x_{k,i}$$

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

Easier to represent in matrix form which will be covered in linear algebra piece of primer.



# Interpretation Adjustment

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$

- With multiple variables in the model, the interpretation of  $\hat{\beta}_j$  changes slightly.
- The estimate  $\hat{\beta}_j$  is the predicted (or expected or average) change in  $y$  with a one unit increase in  $x_j$  **given all other variables are held constant.**

# Example

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$


$$\hat{y} = 1 + 2x_1 + x_2$$

# Example

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

$$\hat{y} = 1 + 2x_1 + x_2$$

Let  $x_2 = 0$



$$\hat{y} = 1 + 2x_1$$

# Example

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

$$\hat{y} = 1 + 2x_1 + x_2$$

Let  $x_2 = 1$



$$\hat{y} = 1 + 2x_1 + 1 = 2 + 2x_1$$

# Example

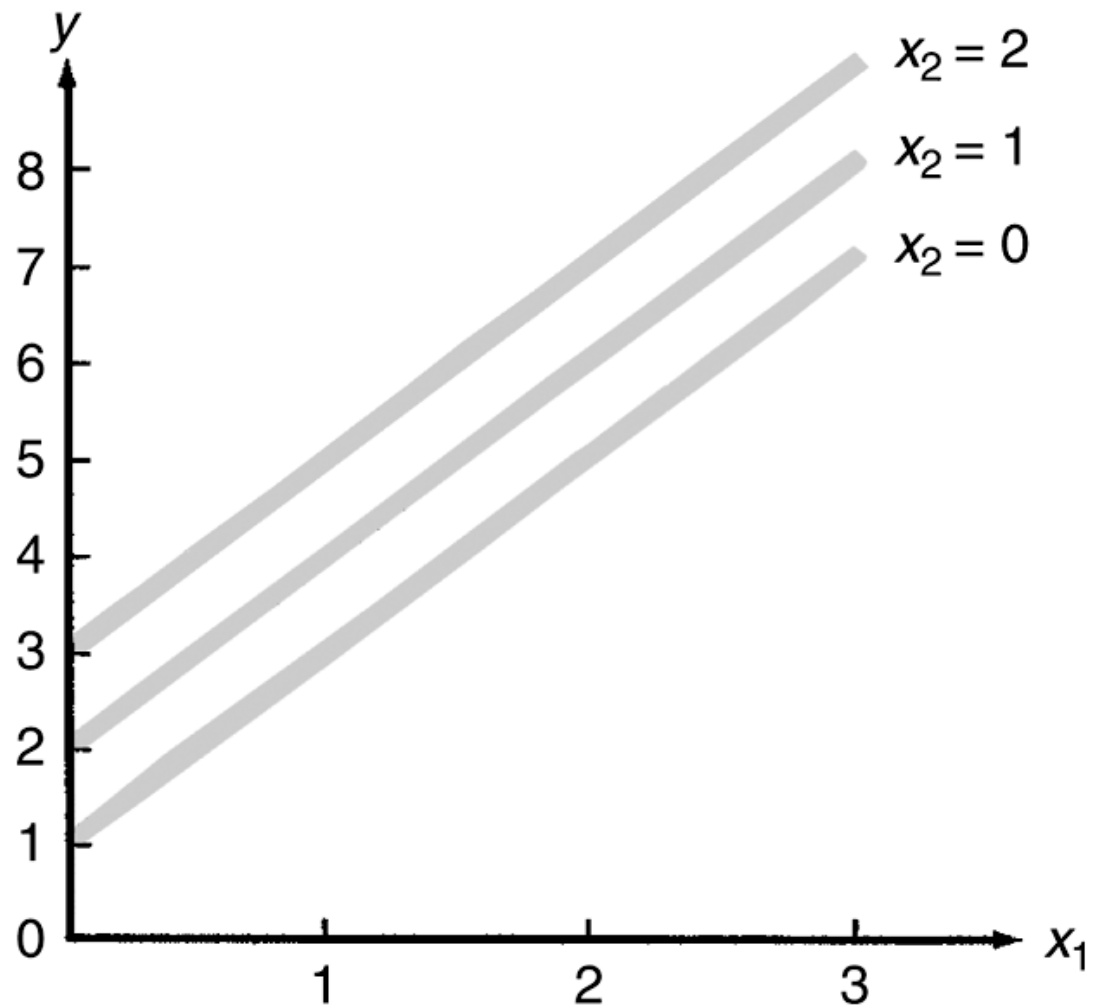
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

$$\hat{y} = 1 + 2x_1 + x_2$$

Let  $x_2 = 2$


$$\hat{y} = 1 + 2x_1 + 2 = 3 + 2x_1$$

# Example



# MULTIPLE LINEAR REGRESSION MODEL

---

Multiple Coefficients of Determination

# Coefficient of Determination

- Similar to simple linear regression, multiple linear regression also has a coefficient of determination,  $R^2$ .
- The calculation is the same as before:

$$R^2 = 1 - \frac{SSE}{TSS}$$



# Coefficient of Determination

- Similar to simple linear regression, multiple linear regression also has a coefficient of determination,  $R^2$ .
- The calculation is the same as before:

$$R^2 = 1 - \frac{SSE}{TSS}$$

- The interpretation is that  $(100)R^2\%$  is the percentage of the variation in  $y$  explained by the linear model using  $x_1, \dots, x_k$  as predictors.

# Problem with $R^2$

- The problem with the calculation of  $R^2$  in a multiple linear regression is that the addition of any variable (good or bad) will make the  $R^2$  value increase if even slightly.

$$R^2 = 1 - \frac{SSE}{TSS}$$

Will never increase with the addition of a variable.

# Adjusted Coefficient of Determination

- To account for this problem, most people use the **adjusted coefficient of determination**,  $R_a^2$ .
- The calculations are as follows:

$$R_a^2 = 1 - \left( \frac{n - 1}{n - (k + 1)} \right) \left( \frac{SSE}{TSS} \right)$$

OR

$$R_a^2 = 1 - (1 - R^2) \left( \frac{n - 1}{n - (k + 1)} \right)$$

# Adjusted Coefficient of Determination

- The  $R_a^2$  penalizes a model for adding a variable that does not provide any useful information.

$$R_a^2 \leq R^2$$

- The adjusted coefficient of determination loses its interpretation (because it could be negative!), but is better at determining utility of a model.

# Example

- A real estate company is trying to model housing prices (in dollars) of their customers with the variables:
  - $x_1$ : Size of Home (square feet)
  - $x_2$ : Age of Home (years)
  - $x_3$ : Acreage of Land (acres)
  - $x_4$ : Number of Bedrooms
- Using a sample of 105 houses they derive the following model:

$$\hat{y} = 24,312 + 86.5x_1 - 324x_2 + 9,610x_3 + 3,617x_4$$

$$SSE = 27695831$$

$$SSR = 45963293$$

$$TSS = 73659124$$

# Example

$$\hat{y} = 24,312 + 86.5x_1 - 324x_2 + 9,610x_3 + 3,617x_4$$

$$SSE = 27695831 \quad SSR = 45963293 \quad TSS = 73659124$$

1. Interpret the coefficient on  $x_2$  in the model.
2. Calculate  $R^2$  and  $R_a^2$ .

# Example

$$\hat{y} = 24,312 + 86.5x_1 - 324x_2 + 9,610x_3 + 3,617x_4$$

$$SSE = 27695831 \quad SSR = 45963293 \quad TSS = 73659124$$

1. Interpret the coefficient on  $x_2$  in the model.
  - Every extra year older, decreases the price of a home by an **average** of \$324, **all else equal**.
2. Calculate  $R^2$  and  $R_a^2$ .

$$R^2 = \frac{45963293}{73659124} = 0.624$$

$$R_a^2 = 1 - (1 - 0.624) \times \left( \frac{105 - 1}{105 - 4 - 1} \right) = 0.609$$

# MULTIPLE LINEAR REGRESSION MODEL

---

Inference for Multiple Regression



# Utility of the Model

- In simple linear regression we could just look at the t-test for our slope parameter estimate to determine the utility of our model.
- With multiple parameter estimates comes multiple t-tests.
- Ideally there should be a way of determining whether the model is adequate for predicting  $y$  overall, instead of looking at every individual parameter estimate.

# Global F-test

- The utility of a multiple regression model can be tested with a test that encompasses all the  $\beta$  parameters – a **global test**.

$$H_0: \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_k = 0$$

$H_a$ : At least one coefficient is nonzero

# Global F-test

- The utility of a multiple regression model can be tested with a test that encompasses all the  $\beta$  parameters – a **global test**.

$$H_0: \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_k = 0$$

$H_a$ : At least one coefficient is nonzero

None of the variables are good.

# Global F-test

- The utility of a multiple regression model can be tested with a test that encompasses all the  $\beta$  parameters – a **global test**.

$$H_0: \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_k = 0$$

$$H_a: \text{At least one coefficient is nonzero}$$



At least one of the variables is good

# Global F-test

- The utility of a multiple regression model can be tested with a test that encompasses all the  $\beta$  parameters – a **global test**.

$$H_0: \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_k = 0$$

$$H_a: \text{At least one coefficient is nonzero}$$

- The test statistic for this hypothesis test follows an  $F$ -distribution and is calculated as follows:

$$F = \frac{\left(\frac{SSR}{k}\right)}{\left(\frac{SSE}{n - (k + 1)}\right)}$$

# Global F-test

- The utility of a multiple regression model can be tested with a test that encompasses all the  $\beta$  parameters – a **global test**.

$$H_0: \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_k = 0$$

$H_a$ : At least one coefficient is nonzero

- The test statistic for this hypothesis test follows an  $F$ -distribution and is calculated as follows:

$$F = \frac{\left(\frac{SSR}{k}\right)}{\left(\frac{SSE}{n - (k + 1)}\right)}$$

Average amount of variation each variable explains (i.e. Mean Square Regression)

# Global F-test

- The utility of a multiple regression model can be tested with a test that encompasses all the  $\beta$  parameters – a **global test**.

$$H_0: \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_k = 0$$

$H_a$ : At least one coefficient is nonzero

- The test statistic for this hypothesis test follows an  $F$ -distribution and is calculated as follows:

$$F = \frac{\left(\frac{SSR}{k}\right)}{\left(\frac{SSE}{n - (k + 1)}\right)}$$

“Average” amount of variation left per data point (i.e. Mean Square Error)

# Global F-test

- The utility of a multiple regression model can be tested with a test that encompasses all the  $\beta$  parameters – a **global test**.

$$H_0: \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_k = 0$$

$$H_a: \text{At least one coefficient is nonzero}$$

- The test statistic for this hypothesis test follows an  $F$ -distribution and is calculated as follows:

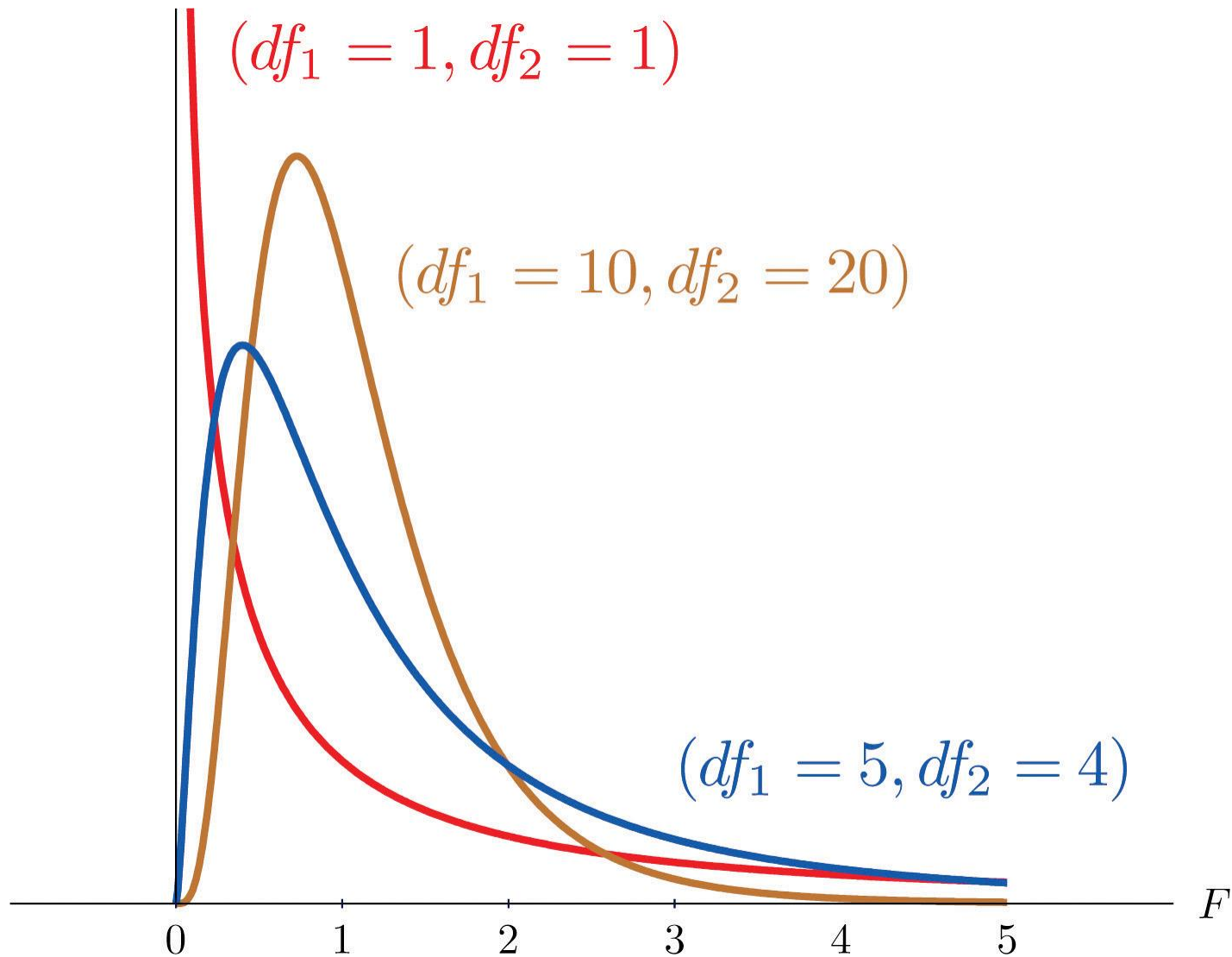
$$F = \frac{\left(\frac{SSR}{k}\right)}{\left(\frac{SSE}{n - (k + 1)}\right)} = \frac{MSR}{MSE}$$



# F-Distribution

- The F-test comes from the **F-distribution**.
- Characteristics of the F-distribution:
  1. Bounded Below By Zero
  2. Right Skewed
  3. Numerator **and** Denominator Degrees of Freedom

# F-Distribution



# Hypothesis Test for $\beta_j$

- Just like in simple linear regression, the values of the parameters  $\beta_1, \dots, \beta_k$  are not directly observed.
- Need to test if the values of each of these coefficients are zero to determine if a relationship exists between the response variable  $y$  and that **specific** explanatory variable  $x$ .

$$H_0: \beta_j = 0, \quad \text{for } j = 1, \dots, k$$

$$H_a: \beta_j \neq 0, \quad \text{for } j = 1, \dots, k$$

# Hypothesis Test for $\beta_j$

- The distribution of these are the same as in the simple linear regression case.
- The test statistic for the hypothesis test is the following:

$$t = \frac{\hat{\beta}_j - 0}{s\hat{\beta}_j} = \frac{\hat{\beta}_j}{\left( \frac{s}{\left(1 - R_j^2\right) \sqrt{SS_{x_j x_j}}} \right)}$$

$$d.f. = n - (k + 1)$$

# Hypothesis Test for $\beta_j$

- The distribution of these are the same as in the simple linear regression case.
- The test statistic for the hypothesis test is the following:

$$t = \frac{\hat{\beta}_j - 0}{s\hat{\beta}_j} = \frac{\hat{\beta}_j}{\left( \frac{s}{\left(1 - R_j^2\right) \sqrt{SS_{x_j x_j}}} \right)}$$

$$d.f. = n - (k + 1)$$

$$SS_{x_j x_j} = \sum_{i=1}^n (x_{j,i} - \bar{x}_j)^2$$

# Hypothesis Test for $\beta_j$

- The distribution of these are the same as in the simple linear regression case.
- The test statistic for the hypothesis test is the following:

$$t = \frac{\hat{\beta}_j - 0}{s_{\hat{\beta}_j}} = \frac{\hat{\beta}_j}{\left( \frac{s}{(1 - R_j^2)} \sqrt{SS_{x_j x_j}} \right)}$$

$$d.f. = n - (k + 1)$$

Coefficient of determination  
of regression of  $x_1, \dots, x_k$  on  $x_j$

# Hypothesis Test for $\beta_j$

- The distribution of these are the same as in the simple linear regression case.
- The test statistic for the hypothesis test is the following:

$$t = \frac{\hat{\beta}_j - 0}{s_{\hat{\beta}_j}} = \frac{\hat{\beta}_j}{\left( \frac{s}{\left(1 - R_j^2\right) \sqrt{SS_{x_j x_j}}} \right)}$$

$$d.f. = n - (k + 1)$$

- All of the other steps in the hypothesis test are the same as any other.

# Confidence Interval for $\beta_j$

- The confidence interval form is the same as for simple linear regression.
- The confidence interval is the following:

$$\hat{\beta}_j \pm (t_{\alpha/2})s_{\hat{\beta}_j}$$



# Utility of a Model

- Best practice in multiple linear regression is to take the following steps:
  1. Conduct the F global test.
  2. If you reject the null hypothesis in step 1, conduct individual t-tests on the parameter estimates.

# Example

- A real estate company is trying to model housing prices (in dollars) of their customers with the variables:
  - $x_1$ : Size of Home (square feet)
  - $x_2$ : Age of Home (years)
  - $x_3$ : Acreage of Land (acres)
  - $x_4$ : Number of Bedrooms
- Using a sample of 105 houses they derive the following model:

$$\hat{y} = 24,312 + 86.5x_1 - 324x_2 + 9,610x_3 + 3,617x_4$$

$$SSE = 27695831$$

$$SSR = 45963293$$

$$TSS = 73659124$$

# Example

$$\hat{y} = 24,312 + 86.5x_1 - 324x_2 + 9,610x_3 + 3,617x_4$$

$$SSE = 27695831 \quad SSR = 45963293 \quad TSS = 73659124$$

1. Test the overall significance of the model.

# Example

$$\hat{y} = 24,312 + 86.5x_1 - 324x_2 + 9,610x_3 + 3,617x_4$$

$$SSE = 27695831 \quad SSR = 45963293 \quad TSS = 73659124$$

1. Test the overall significance of the model.

$$H_0: \beta_1 = \beta_2 = \beta_3 = \dots = \beta_k = 0$$

$H_a$ : At least one coefficient is nonzero

$$MSR = \frac{45963293}{4} = 11490823.25$$

$$MSE = \frac{27695831}{105 - 4 - 1} = 276958.31$$

$$F = \frac{MSR}{MSE} = 41.49 \quad \text{P-value} < 0.05 \rightarrow \text{REJECT } H_0$$

# Example

$$\hat{y} = 24,312 + 86.5x_1 - 324x_2 + 9,610x_3 + 3,617x_4$$

$$SSE = 27695831 \quad SSR = 45963293 \quad TSS = 73659124$$

2. Test the individual significance of the variable  $x_3$ .

$$s_{\hat{\beta}_3} = 3313$$

# Example

$$\hat{y} = 24,312 + 86.5x_1 - 324x_2 + 9,610x_3 + 3,617x_4$$

$$SSE = 27695831 \quad SSR = 45963293 \quad TSS = 73659124$$

2. Test the individual significance of the variable  $x_3$ .

$$s_{\hat{\beta}_3} = 3313$$

$$H_0: \beta_3 = 0$$

$$H_a: \beta_3 \neq 0$$

$$t = \frac{9610 - 0}{3313} = 2.9$$

P-value = (0.002, 0.01)  $\rightarrow$  REJECT  $H_0$

# Example

$$\hat{y} = 24,312 + 86.5x_1 - 324x_2 + 9,610x_3 + 3,617x_4$$

$$SSE = 27695831 \quad SSR = 45963293 \quad TSS = 73659124$$

3. Test the individual significance of the remaining variables. Should any be removed from the model?

$$s_{\hat{\beta}_1} = 7109$$

$$s_{\hat{\beta}_2} = 15$$

$$s_{\hat{\beta}_4} = 3480$$

# Example

$$\hat{y} = 24,312 + 86.5x_1 - 324x_2 + 9,610x_3 + 3,617x_4$$

$$SSE = 27695831 \quad SSR = 45963293 \quad TSS = 73659124$$

3. Test the individual significance of the remaining variables. Should any be removed from the model?

$$s_{\hat{\beta}_1} = 7109 \quad \text{P-value} < 0.001 \rightarrow \text{REJECT } H_0$$

$$s_{\hat{\beta}_2} = 15 \quad \text{P-value} < 0.001 \rightarrow \text{REJECT } H_0$$

$$s_{\hat{\beta}_4} = 3480 \quad \text{P-value} = (0.3, 0.4) \rightarrow \text{DON'T REJECT } H_0$$



# MULTIPLE LINEAR REGRESSION MODEL

---

Categorical Independent Variables

# Types of Variables

- There are two types of variables used in regression modeling:
  - Quantitative – numeric
  - Qualitative – categorical
- Different values of independent variables are referred to as **levels**.

# Dummy Variables

- Categorical variables need to be coded differently because they are not numerical in nature.
- **Dummy variables** are a common way to code categorical variables.
- 2 Category Example (A, B):
$$x = \begin{cases} 1 & \text{if A} \\ 0 & \text{if B} \end{cases}$$

# Dummy Variables

- Categorical variables need to be coded differently because they are not numerical in nature.
- **Dummy variables** are a common way to code categorical variables.
- 2 Category Example (A, B):
$$x = \begin{cases} 1 & \text{if A} \\ 0 & \text{if B} \end{cases}$$
- 3 Category Example (A, B, C):
$$x_1 = \begin{cases} 1 & \text{if A} \\ 0 & \text{if not} \end{cases}$$
$$x_2 = \begin{cases} 1 & \text{if B} \\ 0 & \text{if not} \end{cases}$$

# Dummy Variables – k Levels

- The following is a summary of a k level qualitative variable:

$$x_1 = \begin{cases} 1 & \text{if category 1} \\ 0 & \text{if not} \end{cases}$$

$$x_2 = \begin{cases} 1 & \text{if category 2} \\ 0 & \text{if not} \end{cases}$$

$\vdots$

$$x_{k-1} = \begin{cases} 1 & \text{if category } k - 1 \\ 0 & \text{if not} \end{cases}$$

# Dummy Variable Trap

- Having too many dummy variables can lead to perfect multicollinearity!
- Example:

$$x_1 = \begin{cases} 1 & \text{if category 1} \\ 0 & \text{if category 2} \end{cases}$$

$$x_2 = \begin{cases} 1 & \text{if category 2} \\ 0 & \text{if category 1} \end{cases}$$

# Dummy Variable Trap

- Having too many dummy variables can lead to perfect multicollinearity!
- Example:

$$x_1 = \begin{cases} 1 & \text{if category 1} \\ 0 & \text{if category 2} \end{cases}$$

$$x_2 = \begin{cases} 1 & \text{if category 2} \\ 0 & \text{if category 1} \end{cases}$$

SAME  
INFO!

# Dummy Variables

- Categorical variables need to be coded differently because they are not numerical in nature.
- **Dummy variables** are a common way to code categorical variables.
- 2 Category Example (A, B):
$$x = \begin{cases} 1 & \text{if A} \\ 0 & \text{if B} \end{cases}$$
- 3 Category Example (A, B, C):
$$x_1 = \begin{cases} 1 & \text{if A} \\ 0 & \text{if not} \end{cases}$$
$$x_2 = \begin{cases} 1 & \text{if B} \\ 0 & \text{if not} \end{cases}$$



# Dummy Variables

- Categorical variables need to be coded differently because they are not numerical in nature.
- **Dummy variables** are a common way to code categorical variables.
- 2 Category Example (A, B):
$$x = \begin{cases} 1 & \text{if A} \\ 0 & \text{if B} \end{cases}$$
- 3 Category Example (A, B, C):

	$x_1$	$x_2$
A	1	0
B	0	1
C	0	0

# Dummy Variables

- Categorical variables need to be coded differently because they are not numerical in nature.
- **Dummy variables** are a common way to code categorical variables.
- 3 Category Example (A, B, C):

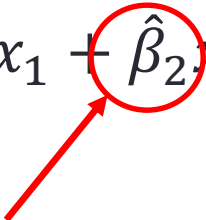
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

Average difference between category A and C.

	$x_1$	$x_2$
A	1	0
B	0	1
C	0	0

# Dummy Variables

- Categorical variables need to be coded differently because they are not numerical in nature.
- **Dummy variables** are a common way to code categorical variables.
- 3 Category Example (A, B, C):

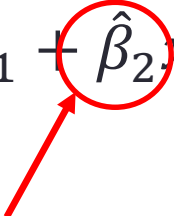
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$


Average difference between category B and C.

	$x_1$	$x_2$
A	1	0
B	0	1
C	0	0

# How Does the Math Work?

- 3 Category Example (A, B, C):

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$


Average difference between  
category B and C. **BUT WHY?**

	$x_1$	$x_2$
A	1	0
B	0	1
C	0	0

# How Does the Math Work?

- 3 Category Example (A, B, C):

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

	$x_1$	$x_2$
A	1	0
B	0	1
C	0	0

$$\hat{y}_B = \hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot 1 = \hat{\beta}_0 + \hat{\beta}_2$$

$$\hat{y}_C = \hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot 0 = \hat{\beta}_0$$

$$\hat{y}_{B-C} = (\hat{\beta}_0 + \hat{\beta}_2) - \hat{\beta}_0 = \hat{\beta}_2$$

# Effects Coding

- Categorical variables need to be coded differently because they are not numerical in nature.
- **Effects coding** is another common way to code categorical variables.

- 2 Category Example (A, B):
$$x = \begin{cases} 1 & \text{if A} \\ -1 & \text{if B} \end{cases}$$

- 3 Category Example (A, B, C):

	$x_1$	$x_2$
A	1	0
B	0	1
C	-1	-1

# Effects Coding

- Categorical variables need to be coded differently because they are not numerical in nature.
- **Effects coding** is another common way to code categorical variables.
- 3 Category Example (A, B, C):

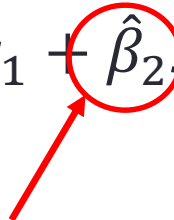
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$


Average difference between category A and the overall average of categories **A, B, & C**.

	$x_1$	$x_2$
A	1	0
B	0	1
C	-1	-1

# Effects Coding

- Categorical variables need to be coded differently because they are not numerical in nature.
- **Effects coding** is another common way to code categorical variables.
- 3 Category Example (A, B, C):

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$


Average difference between category B and the overall average of categories **A, B, & C**.

	$x_1$	$x_2$
A	1	0
B	0	1
C	-1	-1



# How Does the Math Work?

- 3 Category Example (A, B, C):

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

	$x_1$	$x_2$
A	1	0
B	0	1
C	-1	-1

$$\hat{y}_A = \hat{\beta}_0 + \hat{\beta}_1 \cdot 1 + \hat{\beta}_2 \cdot 0 = \hat{\beta}_0 + \hat{\beta}_1$$

$$\hat{y}_B = \hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot 1 = \hat{\beta}_0 + \hat{\beta}_2$$

$$\hat{y}_C = \hat{\beta}_0 + \hat{\beta}_1 \cdot (-1) + \hat{\beta}_2 \cdot (-1) = \hat{\beta}_0 - \hat{\beta}_1 - \hat{\beta}_2$$

$$\hat{y}_{Avg.} = \frac{((\hat{\beta}_0 + \hat{\beta}_1) + (\hat{\beta}_0 + \hat{\beta}_2) + (\hat{\beta}_0 - \hat{\beta}_1 - \hat{\beta}_2))}{3} = \hat{\beta}_0$$

# How Does the Math Work?

- 3 Category Example (A, B, C):

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

	$x_1$	$x_2$
A	1	0
B	0	1
C	-1	-1

$$\hat{y}_B = \hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot 1 = \hat{\beta}_0 + \hat{\beta}_2$$

$$\hat{y}_{Avg.} = \frac{((\hat{\beta}_0 + \hat{\beta}_1) + (\hat{\beta}_0 + \hat{\beta}_2) + (\hat{\beta}_0 - \hat{\beta}_1 - \hat{\beta}_2))}{3} = \hat{\beta}_0$$

$$\hat{y}_{B-Avg} = (\hat{\beta}_0 + \hat{\beta}_2) - \hat{\beta}_0 = \hat{\beta}_2$$

# Example

- Develop both effects coding and dummy / reference coding for a categorical variable with 4 categories.

# Example

- Develop both **effects** coding and dummy / reference coding for a categorical variable with 4 categories.

	$x_1$	$x_2$	$x_3$
A	1	0	0
B	0	1	0
C	0	0	1
D	-1	-1	-1

# Example

- Develop both effects coding and **dummy / reference** coding for a categorical variable with 4 categories.

	$x_1$	$x_2$	$x_3$
A	1	0	0
B	0	1	0
C	0	0	1
D	0	0	0

# Example

- Develop both effects coding and dummy / reference coding for a categorical variable with 4 categories.

	$x_1$	$x_2$	$x_3$
A	1	0	0
B	0	1	0
C	0	0	1
D	-1	-1	-1

	$x_1$	$x_2$	$x_3$
A	1	0	0
B	0	1	0
C	0	0	1
D	0	0	0

# Example

- A real estate company is trying to model housing prices (in dollars) of their customers with the variables:
  - $x_1$ : Size of Home (square feet)
  - $x_2$ : Age of Home (years)
  - $x_3$ : Acreage of Land (acres)
  - $x_4$ : Number of Bedrooms
  - $x_5$ : Located on golf course
- Using a sample of 105 houses they derive the following model:

$$\hat{y} = 24,312 + 86.5x_1 - 324x_2 + 9,610x_3 + 3,617x_4 + 12,550x_5$$

$$s_{\hat{\beta}_5} = 4532$$

# Example

$$\hat{y} = 24,312 + 86.5x_1 - 324x_2 + 9,610x_3 + 3,617x_4 + 12,550x_5$$

$$s_{\hat{\beta}_5} = 4532$$

1. How would you code the variable summarizing whether a house was on the golf course?
2. What is the interpretation of the coefficient on the variable  $x_5$ ?



# Example

$$\hat{y} = 24,312 + 86.5x_1 - 324x_2 + 9,610x_3 + 3,617x_4 + 12,550x_5$$

$$s_{\hat{\beta}_5} = 4532$$

1. How would you code the variable summarizing whether a house was on the golf course?

$$x_5 = \begin{cases} 1 & \text{if on golf course} \\ 0 & \text{if not on golf course} \end{cases}$$

2. What is the interpretation of the coefficient on the variable  $x_5$ ?
  - The **average** increase in home price for home on a golf course compared to not is \$12,550, **all else equal**.

# Example

$$\hat{y} = 24,312 + 86.5x_1 - 324x_2 + 9,610x_3 + 3,617x_4 + 12,550x_5$$

$$s_{\hat{\beta}_5} = 4532$$

3. Calculate the test of significance for the variable  $x_5$ .

# Example

$$\hat{y} = 24,312 + 86.5x_1 - 324x_2 + 9,610x_3 + 3,617x_4 + 12,550x_5$$

$$s_{\hat{\beta}_5} = 4532$$

3. Calculate the test of significance for the variable  $x_5$ .

$$H_0: \beta_5 = 0$$

$$H_a: \beta_5 \neq 0$$

$$t = \frac{12550 - 0}{4532} = 2.77$$

P-value = (0.002, 0.01)  $\rightarrow$  REJECT  $H_0$

# MULTIPLE LINEAR REGRESSION MODEL

---

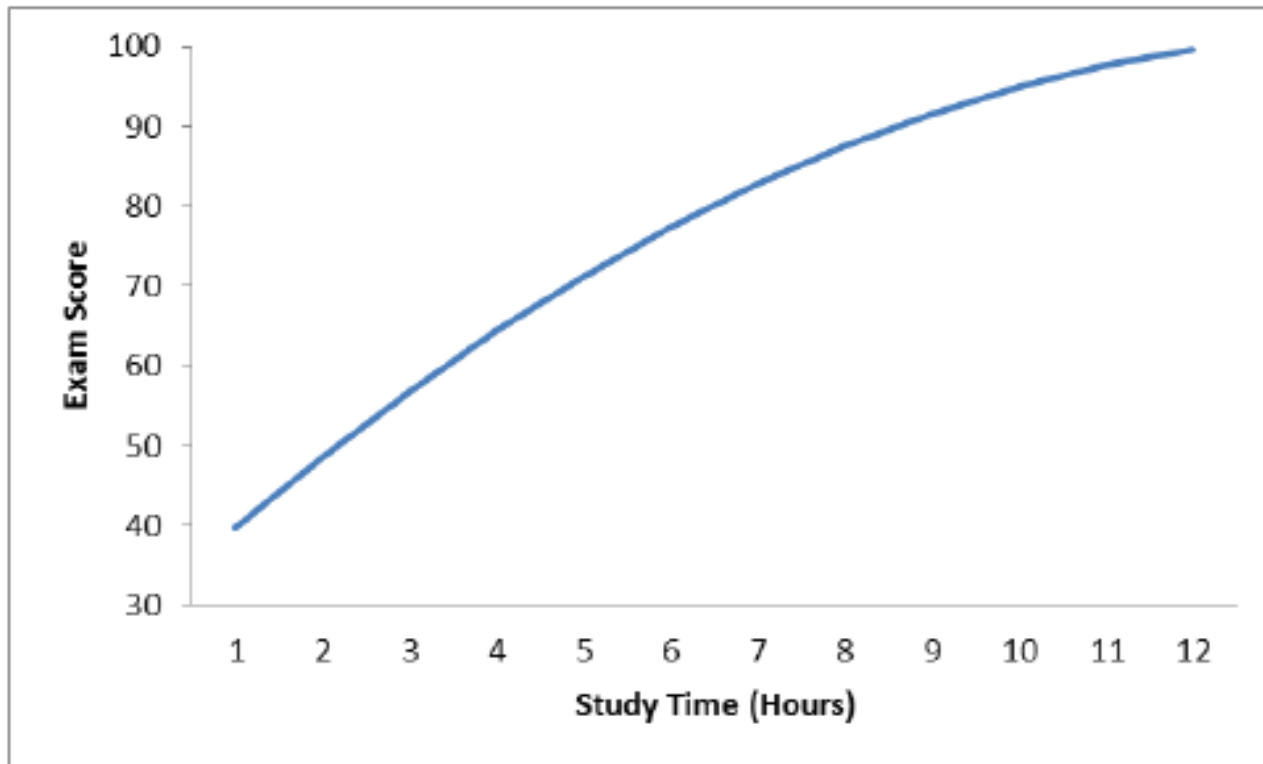
Polynomial Regression

# Higher Order Terms

- Up until this point, only straight-line relationships between the target variable and the predictor variables have been discussed.
- Other, more complicated, relationships can exist between predictor variables and target variables as well.
- One example is when the relationship between a target and predictor variable **do not** remain constant across all the levels of the predictor variable.

# Higher Order Terms

- One example is when the relationship between a target and predictor variable **do not** remain constant across all the levels of the predictor variable.



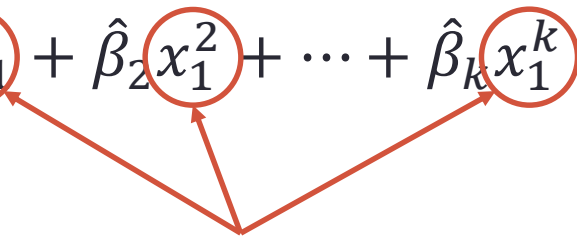
# Polynomial Regression

- These types of relationships are typically modelled with polynomial regression terms.
- The following is the **sample polynomial regression model of order k**:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_1^2 + \cdots + \hat{\beta}_k x_1^k$$

# Polynomial Regression

- These types of relationships are typically modelled with polynomial regression terms.
- The following is the **sample polynomial regression model of order k**:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_1^2 + \cdots + \hat{\beta}_k x_1^k$$


Same variable being raised to different powers.



# Quadratic Model

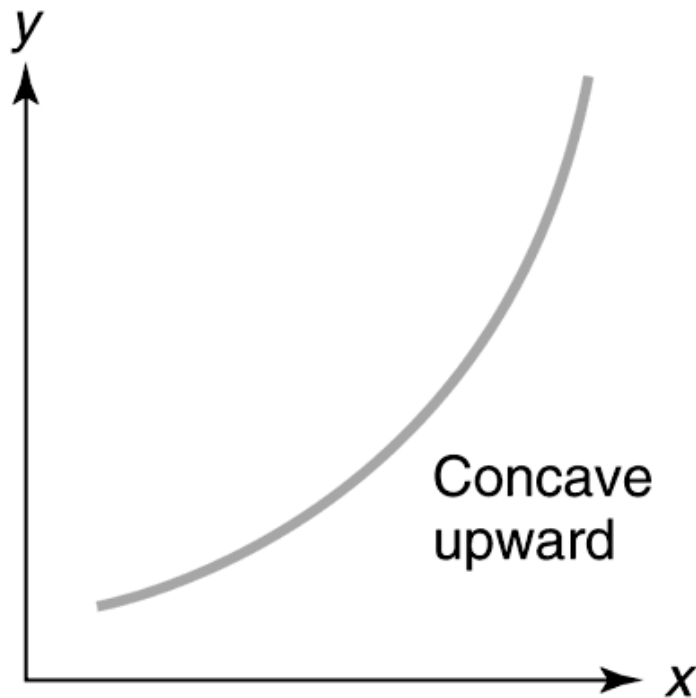
- The most common polynomial regression model goes up to the second power.
- A **quadratic model** will include not only a predictor variable  $x$  but it's quadratic term  $x^2$  as well.

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

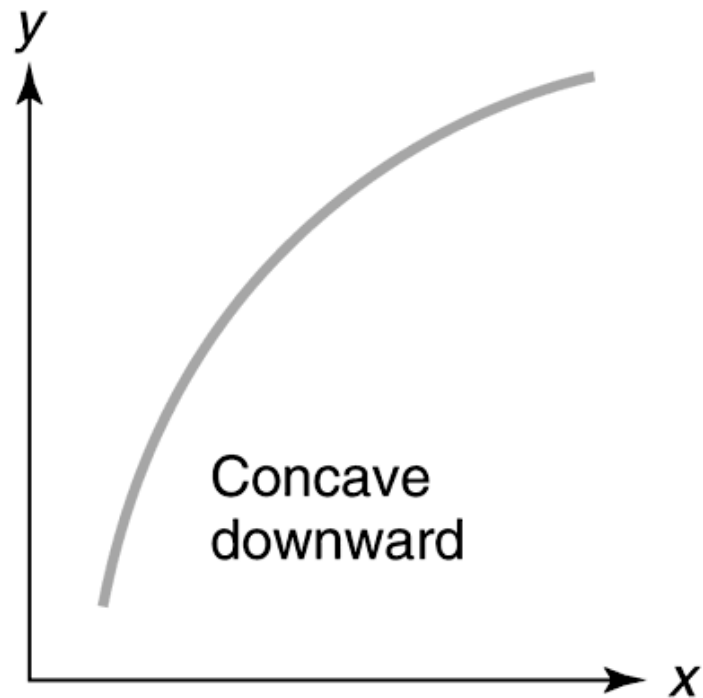
# Quadratic Model

- The value of  $\beta_2$  in the quadratic model will influence the curvature of the relationship between  $x$  and  $y$ .

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$



(a)  $\beta_2 > 0$



(b)  $\beta_2 < 0$

# Deterministic Portion

- The following is the deterministic portion of the second-order model:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2$$

# Deterministic Portion

- The following is the deterministic portion of the second-order model:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2$$



Shift Parameter

# Deterministic Portion

- The following is the deterministic portion of the second-order model:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2$$



Rate of Curvature Parameter

# Caution

- Polynomial regression models are still considered **LINEAR** regression models.
- The parameters are **LINEARLY** related to the response variable – that is what is meant by **LINEAR** regression.

# Caution

- Polynomial regression models are still considered **LINEAR** regression models.
- The parameters are **LINEARLY** related to the response variable – that is what is meant by **LINEAR** regression.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2$$

Linear Combination!

- Multiplying each term by constant and adding

# MULTIPLE LINEAR REGRESSION MODEL

---

Interaction Terms



# Interpretation Adjustment

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$

- With multiple variables in the model, the interpretation of  $\hat{\beta}_j$  changes slightly.
- The estimate  $\hat{\beta}_j$  is the predicted (or expected or average) change in  $y$  with a one unit increase in  $x_j$  **given all other variables are held constant.**

# Example

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$


$$\hat{y} = 1 + 2x_1 + x_2$$

# Example

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

$$\hat{y} = 1 + 2x_1 + x_2$$

Let  $x_2 = 0$



$$\hat{y} = 1 + 2x_1$$

# Example

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

$$\hat{y} = 1 + 2x_1 + x_2$$

Let  $x_2 = 1$



$$\hat{y} = 1 + 2x_1 + 1 = 2 + 2x_1$$

# Example

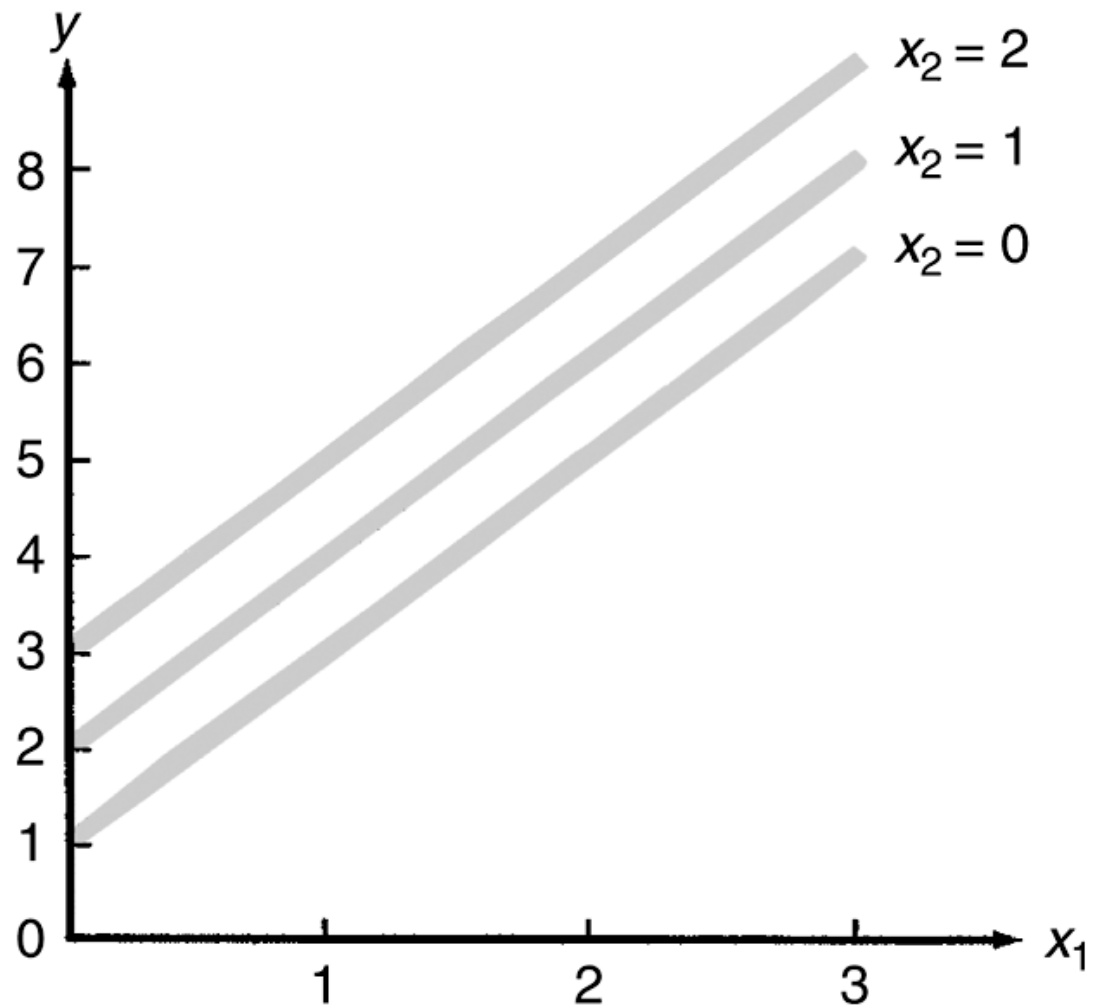
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

$$\hat{y} = 1 + 2x_1 + x_2$$

Let  $x_2 = 2$


$$\hat{y} = 1 + 2x_1 + 2 = 3 + 2x_1$$

# Example




# Interaction

$$\hat{y} = 1 + 2x_1 + x_2$$

- What if the relationship between  $x_1$  and  $\hat{y}$  changed depending on the value of  $x_2$ ?
- In other words, what if the coefficient in front of  $x_1$  didn't remain at 2?
- The model could look like the following:

$$\hat{y} = 1 + 2x_1 + x_2 + x_1x_2$$

# Example



The diagram consists of three elements arranged vertically. At the top is the equation  $\hat{y} = 1 + 2x_1 + x_2 + x_1x_2$  in black text. In the middle is the text "Let  $x_2 = 0$ " in red text. At the bottom is the equation  $\hat{y} = 1 + 2x_1$  in red text. A red curved arrow starts from the left side of the top equation and points down to the bottom equation. A second red curved arrow starts from the left side of the middle text and points down to the bottom equation.

$$\hat{y} = 1 + 2x_1 + x_2 + x_1x_2$$

Let  $x_2 = 0$

$$\hat{y} = 1 + 2x_1$$



# Example

$$\hat{y} = 1 + 2x_1 + x_2 + x_1x_2$$

Let  $x_2 = 1$

$$\hat{y} = 1 + 2x_1 + 1 + x_1 = 2 + 3x_1$$

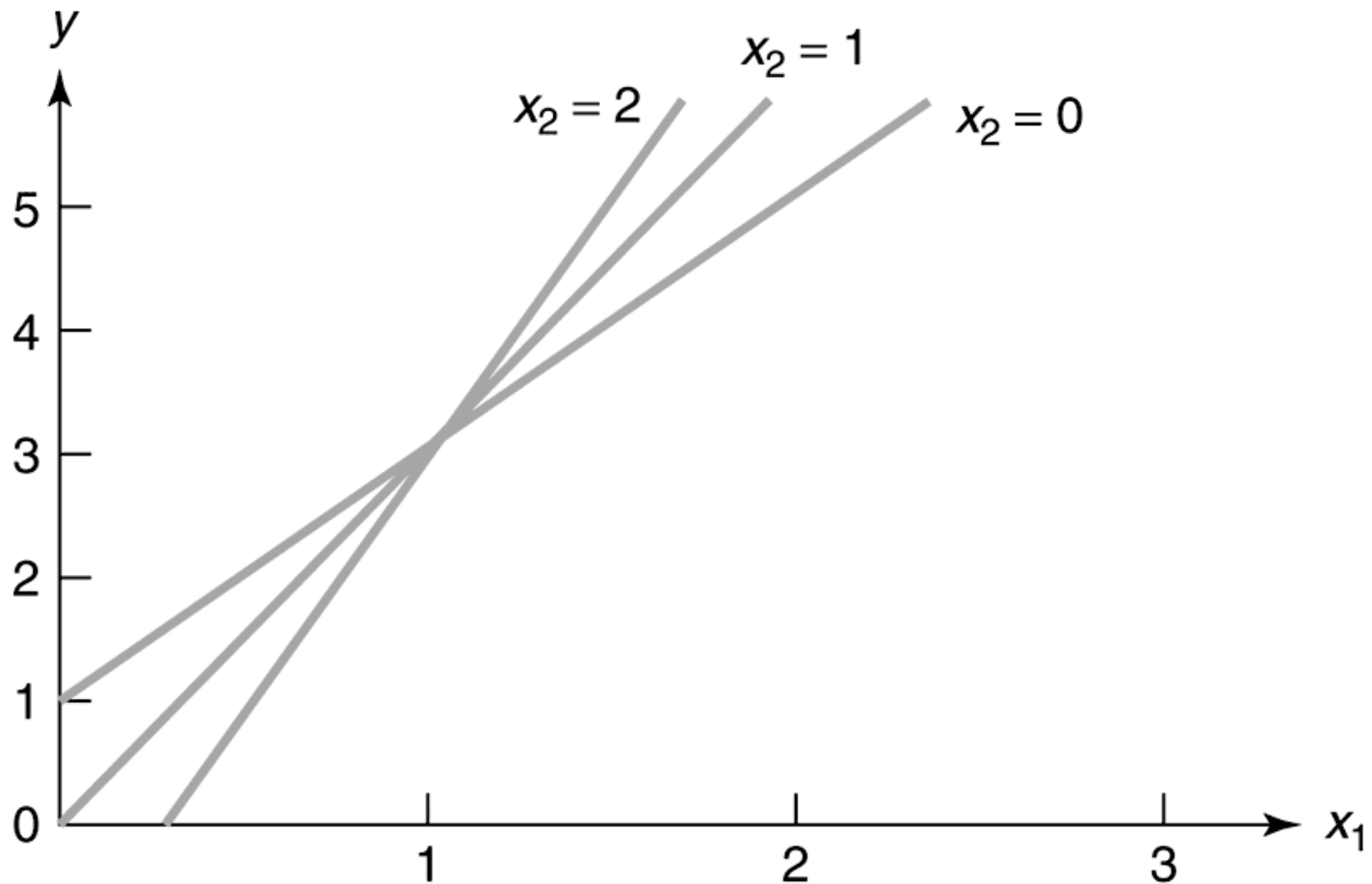
# Example

$$\hat{y} = 1 + 2x_1 + x_2 + x_1x_2$$

Let  $x_2 = 2$

$$\hat{y} = 1 + 2x_1 + 2 + 2x_1 = 3 + 4x_1$$

# Example



# Interaction

- **Interaction** – the relationship between an explanatory variable and  $\hat{y}$  changes depending on the value of another explanatory variable.
- **Independence** – the relationship between an explanatory variable and  $\hat{y}$  **doesn't change** across the values of another explanatory variable.

# Interaction Model

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_1 x_2$$

- Now the value of the “slope” between  $x_1$  and  $y$  depends on the value of  $x_2$ .
- The change in  $\hat{y}$  for every 1-unit increase in  $x_1$  is now represented by  $\hat{\beta}_1 + \hat{\beta}_3 x_2$ .
- The change in  $\hat{y}$  for every 1-unit increase in  $x_2$  is now represented by  $\hat{\beta}_2 + \hat{\beta}_3 x_1$ .

# REGRESSION CAUTIONS

---

# Cautions

- Regressions are not perfect.
- For example, there are several reasons that could lead to not rejecting the null hypothesis in an individual t-test:
  1. There is no relationship between  $x_j$  and  $y$ .
  2. A straight-line relationship actually exists, but a Type II error has occurred.
  3. A more complex relationship (other than linear) exists between  $x_j$  and  $y$ .

# Cautions

- Regressions are not perfect.
- For example, there are several reasons that could lead to not rejecting the null hypothesis in an individual t-test:
  1. There is no relationship between  $x_j$  and  $y$ .
  2. A straight-line relationship actually exists, but a Type II error has occurred.
  3. A more complex relationship (other than linear) exists between  $x_j$  and  $y$ .
  4. Problems in the model lead to incorrect calculation of the standard errors of the t-tests.



# REGRESSION CAUTIONS

---

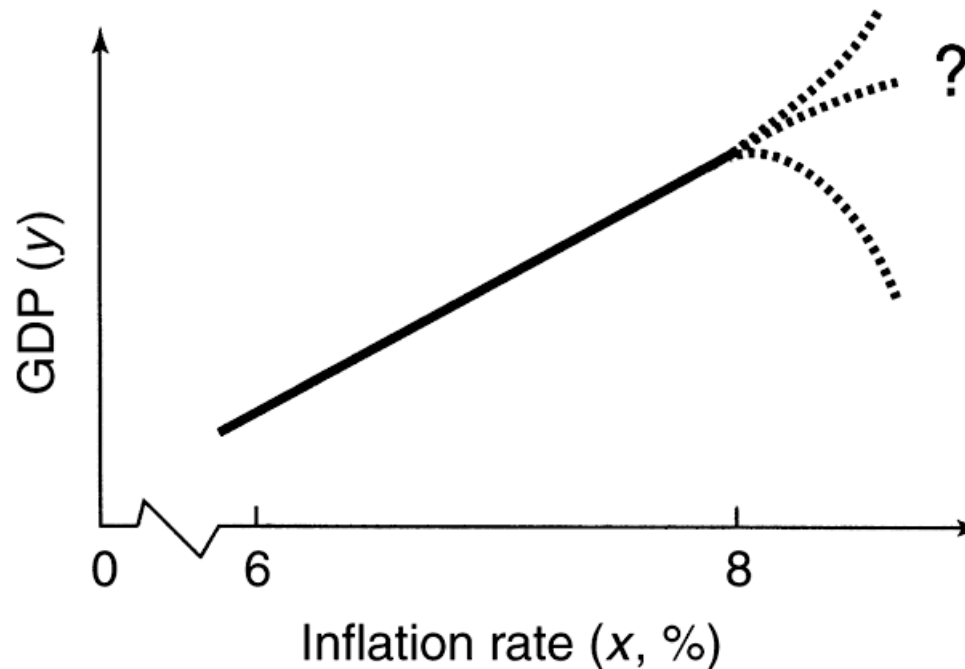
Extrapolation

# Extrapolation

- **Extrapolation** – the prediction of the response variable with inputs outside the range of the data.
- The relationship between the explanatory variable and the response variable is **unknown** outside of the range of the data.

# Example

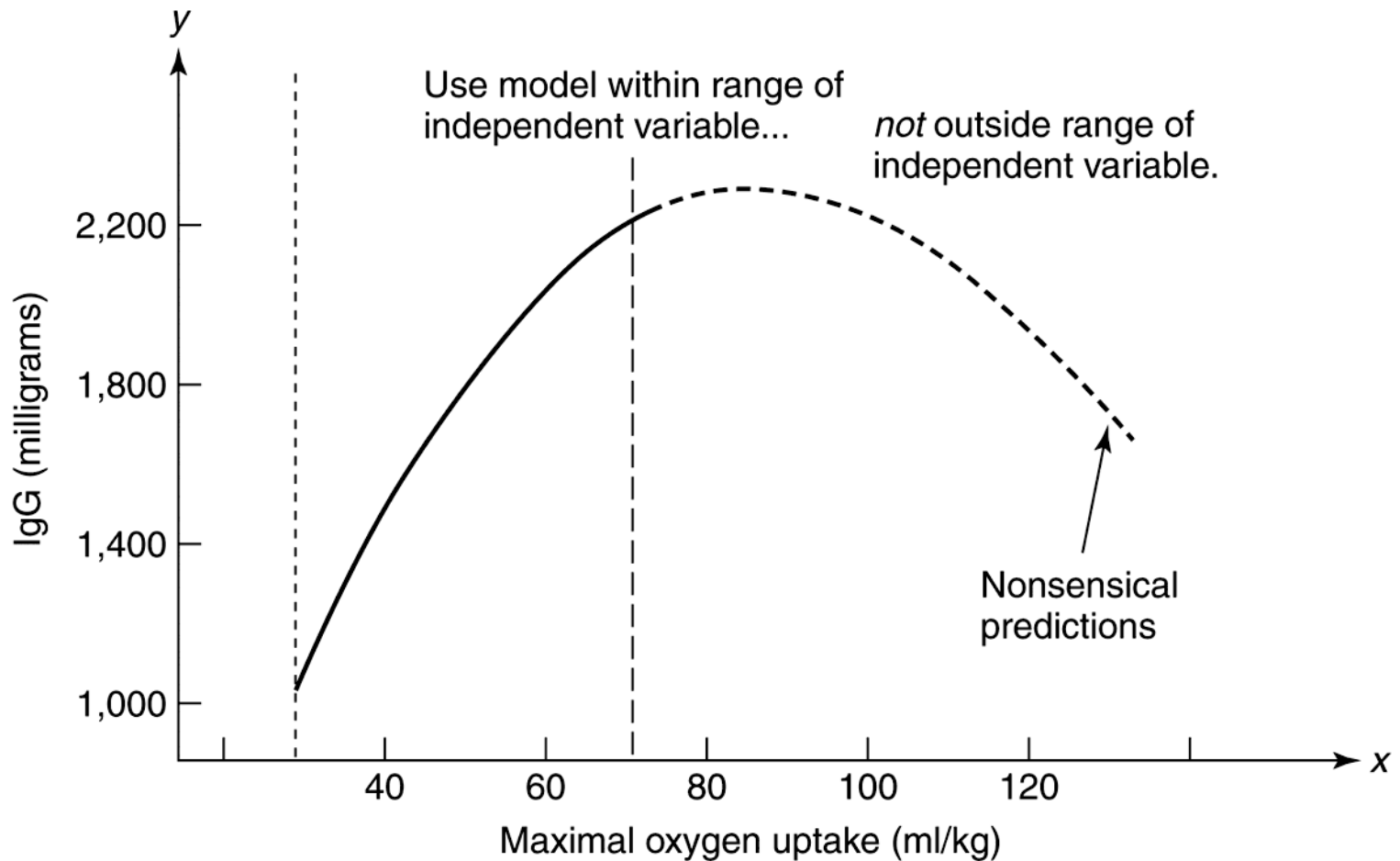
- Want to try and predict GDP with the inflation rate. In the 1960's the inflation rate was between 6-8%. In the 1970's inflation increased, yet researchers tried using the same model.



# Extrapolation

- Extrapolation is an even more severe mistake in a quadratic model.
- A quadratic function eventually begins to turn downward, which can lead to nonsensical results when extrapolation takes place.

# Extrapolation



# Hidden Extrapolation

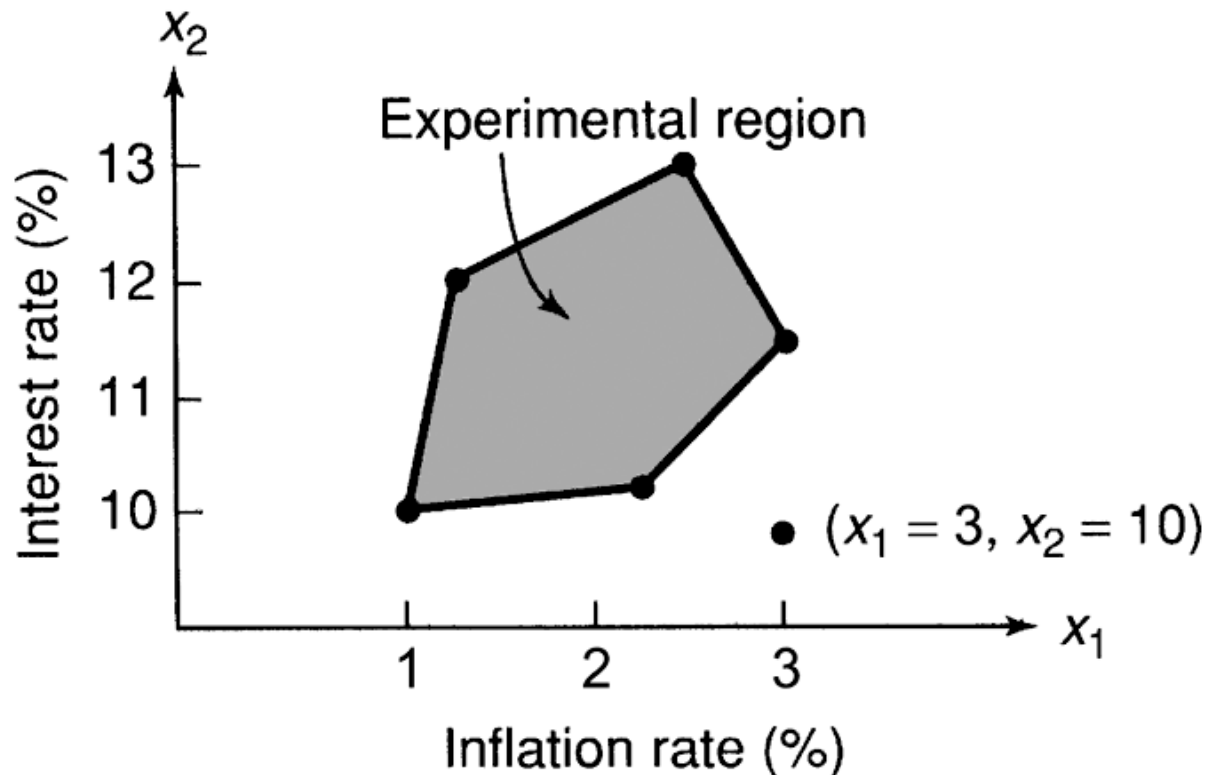
- Extrapolation with multiple variables can be slightly different than with a single variable.
- The variables do not define the range of inputs independently, but jointly.
- Going outside this joint range, but staying inside the ranges individually is called **hidden extrapolation**.

# Example

- Imagine we want to add another variable to the model – prime interest rate.
- The inflation rate is between 1-3% and the prime interest rate is between 10-13%, however, the **joint** values do not cover this whole range.

# Example

- The inflation rate is between 1-3% and the prime interest rate is between 10-13%, however, the **joint** values do not cover this whole range.





# REGRESSION CAUTIONS

---

Model Misspecification

# Regression Models Always Correct?

- There is a real possibility that the final model using the sample is not the same as the true model from the population – even the variables may be different!
- What would happen to the model if we include **irrelevant** variables in the model?
- What would happen to the model if we **omit** important variables from the model?

# Overspecification

- What would happen to the model if we include **irrelevant** variables in the model?

TRUE MODEL:  $y = \beta_0 + \beta_1 x_1 + \varepsilon$

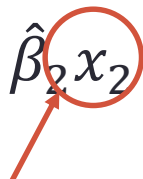
OUR MODEL:  $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$

# Overspecification

- What would happen to the model if we include **irrelevant** variables in the model?

TRUE MODEL:  $y = \beta_0 + \beta_1 x_1 + \varepsilon$

OUR MODEL:  $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$



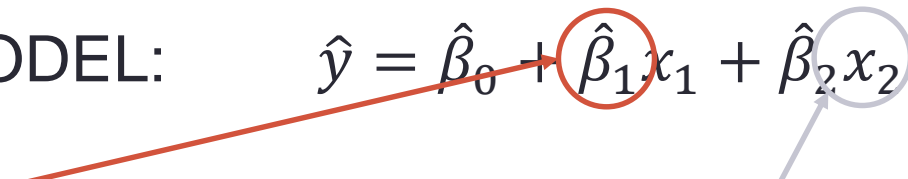
Shouldn't be there!  
True value of  $\beta_2 = 0$ .

# Overspecification

- What would happen to the model if we include **irrelevant** variables in the model?

TRUE MODEL:  $y = \beta_0 + \beta_1 x_1 + \varepsilon$

OUR MODEL:  $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$



Still an **unbiased** estimate of  $\beta_1$ .

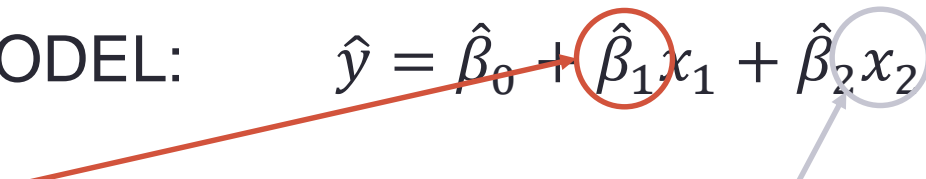
Shouldn't be there!  
True value of  $\beta_2 = 0$ .

# Overspecification

- What would happen to the model if we include **irrelevant** variables in the model?

TRUE MODEL:  $y = \beta_0 + \beta_1 x_1 + \varepsilon$

OUR MODEL:  $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$



Standard error  
estimate of  $\beta_1$  is  
**too high!**

Shouldn't be there!  
True value of  $\beta_2 = 0$ .

# Overspecification

- What would happen to the model if we include **irrelevant** variables in the model?

TRUE MODEL:  $y = \beta_0 + \beta_1 x_1 + \varepsilon$

OUR MODEL:  $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$

Standard error  
estimate of  $\beta_1$  is  
**too high!**

Shouldn't be there!  
True value of  $\beta_2 = 0$ .

$$\frac{S_\varepsilon}{(1 - R_1^2) \sqrt{\sum_{i=1}^n (x_{1,i} - \bar{x}_1)^2}} \geq \frac{S_\varepsilon}{\sqrt{\sum_{i=1}^n (x_{1,i} - \bar{x}_1)^2}}$$

# Underspecification

- What would happen to the model if we **omit** important variables from the model?

TRUE MODEL:  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

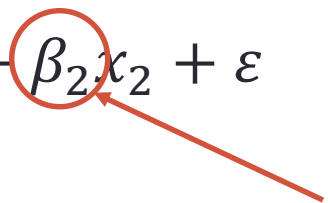
OUR MODEL:  $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$



# Underspecification

- What would happen to the model if we **omit** important variables from the model?

TRUE MODEL:  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$



OUR MODEL:  $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$  Should be in our model!  
True value of  $\beta_2 \neq 0$ .

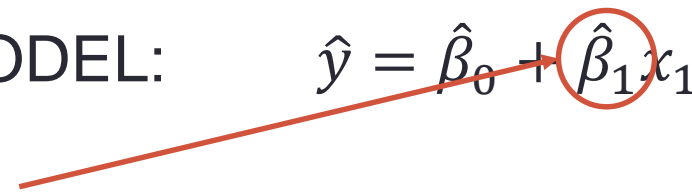
# Underspecification

- What would happen to the model if we **omit** important variables from the model?

TRUE MODEL:  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$



OUR MODEL:  $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$



Should be in our model!  
True value of  $\beta_2 \neq 0$ .

Potentially **biased**  
estimate of  $\beta_1$ .

# Underspecification

- What would happen to the model if we **omit** important variables from the model?

TRUE MODEL:  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

OUR MODEL:  $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$

Should be in our model!  
True value of  $\beta_2 \neq 0$ .

Potentially **biased**  
estimate of  $\beta_1$ .

Bias of  $\hat{\beta}_1 = \beta_2 \times r_{1,2} \times \frac{s_{x_1}}{s_{x_2}}$

# Underspecification

- What would happen to the model if we **omit** important variables from the model?

TRUE MODEL:  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

OUR MODEL:  $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$

Should be in our model!  
True value of  $\beta_2 \neq 0$ .

Potentially **biased**  
estimate of  $\beta_1$ .

Bias of  $\hat{\beta}_1 = \beta_2 \times r_{1,2} \times \frac{s_{x_1}}{s_{x_2}}$

No bias if  $x_1$  and  $x_2$  aren't correlated.

# Underspecification

- What would happen to the model if we **omit** important variables from the model?

TRUE MODEL:  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$



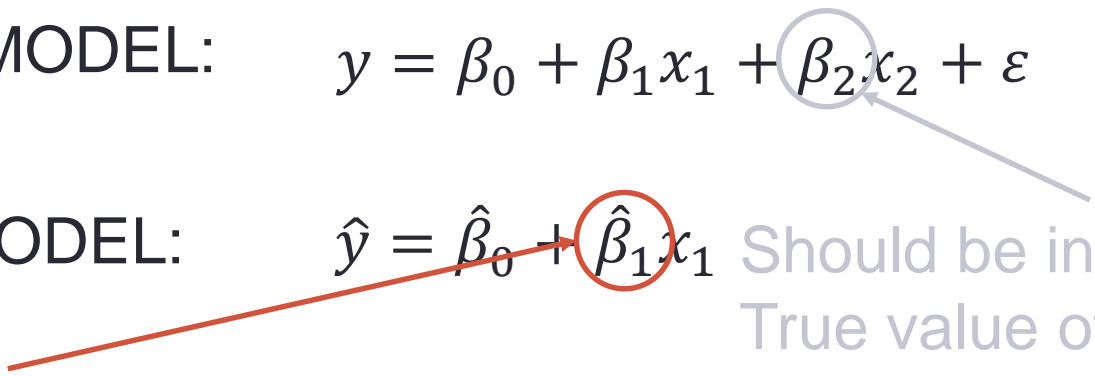
OUR MODEL:  $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$  Should be in our model!  
True value of  $\beta_2 \neq 0$ .

Standard error  
estimate of  $\beta_1$  is  
**too low!**

# Underspecification

- What would happen to the model if we **omit** important variables from the model?

TRUE MODEL:  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$



OUR MODEL:  $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$  Should be in our model!  
True value of  $\beta_2 \neq 0$ .

Standard error  
estimate of  $\beta_1$  is  
**too low!**

$$\frac{S_\varepsilon}{\sqrt{\sum_{i=1}^n (x_{1,i} - \bar{x}_1)^2}} \leq \frac{S_\varepsilon}{(1 - R_1^2) \sqrt{\sum_{i=1}^n (x_{1,i} - \bar{x}_1)^2}}$$

# REGRESSION CAUTIONS

---

Multicollinearity

# Correlation Between Inputs

- **Multicollinearity** occurs when two or more of the independent variables in a regression model are correlated with each other.
- If two inputs are correlated, they could be bringing similar information to the prediction of the response variable.
- It is difficult to find independent variables that are not correlated with each other.



# Signs of Multicollinearity

- There are some easy signs/tests for the presence of severe multicollinearity in a regression model:
  1. Incorrect signs of coefficients.

# Signs of Multicollinearity

- There are some easy signs/tests for the presence of severe multicollinearity in a regression model:
  1. Incorrect signs of coefficients.
  2. Extreme differences in coefficients after addition (or deletion) of variable.

# Signs of Multicollinearity

- There are some easy signs/tests for the presence of severe multicollinearity in a regression model:
  1. Incorrect signs of coefficients.
  2. Extreme differences in coefficients after addition (or deletion) of variable.
  3. Switches in significance of variables.

# Signs of Multicollinearity

- There are some easy signs/tests for the presence of severe multicollinearity in a regression model:
  1. Incorrect signs of coefficients.
  2. Extreme differences in coefficients after addition (or deletion) of variable.
  3. Switches in significance of variables.
  4. Standard deviation of model error increases after addition of variable.

# Problems

- Problems occur if the correlation between multiple input variables is high.
  - Errors in the calculation of the parameter estimates.
  - Errors in the calculation of the standard errors.
  - **Results that are counterintuitive.**

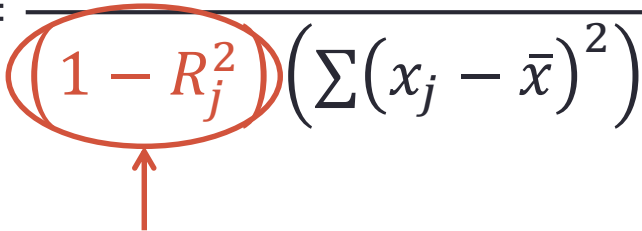
# Variance Inflation Factor

- The **variance inflation factor** is the amount of inflation that the standard error of the parameter estimates have due to multicollinearity.
- Recall the equation for standard errors of parameter estimates in multiple linear regression:

$$s_{\hat{\beta}_j}^2 = \frac{s^2}{(1 - R_j^2) \left( \sum (x_j - \bar{x})^2 \right)}$$

# Variance Inflation Factor

- The **variance inflation factor** is the amount of inflation that the standard error of the parameter estimates have due to multicollinearity.
- Recall the equation for standard errors of parameter estimates in multiple linear regression:

$$s_{\hat{\beta}_j}^2 = \frac{s^2}{(1 - R_j^2) \left( \sum (x_j - \bar{x})^2 \right)}$$


Tolerance

# Variance Inflation Factor

- The **variance inflation factor** is the amount of inflation that the standard error of the parameter estimates have due to multicollinearity.
- Recall the equation for standard errors of parameter estimates in multiple linear regression:

$$s_{\hat{\beta}_j}^2 = \frac{s^2}{(1 - R_j^2) \left( \sum (x_j - \bar{x})^2 \right)} = \frac{1}{(1 - R_j^2)} \times \frac{s^2}{\sum (x_j - \bar{x})^2}$$



# Variance Inflation Factor

- The **variance inflation factor** is the amount of inflation that the standard error of the parameter estimates have due to multicollinearity.
- Recall the equation for standard errors of parameter estimates in multiple linear regression:

$$s_{\hat{\beta}_j}^2 = \frac{s^2}{(1 - R_j^2) \left( \sum (x_j - \bar{x})^2 \right)} = \frac{1}{(1 - R_j^2)} \times \frac{s^2}{\sum (x_j - \bar{x})^2}$$

$VIF_j = \frac{1}{\text{Tolerance}_j}$

# Variance Inflation Factor

- The **variance inflation factor** is the amount of inflation that the standard error of the parameter estimates have due to multicollinearity.
- Recall the equation for standard errors of parameter estimates in multiple linear regression:

$$s_{\hat{\beta}_j}^2 = \frac{s^2}{(1 - R_j^2) \left( \sum (x_j - \bar{x})^2 \right)} = \frac{1}{(1 - R_j^2)} \times \frac{s^2}{\sum (x_j - \bar{x})^2}$$

- Variance inflation factors greater than 10 are typically considered too high.

# Solutions to Multicollinearity

- The following are some common solutions to multicollinearity:
  1. Drop one of the correlated variables.
  2. Avoid making inferences about the parameter estimates.
  3. Develop and design an experiment instead.
  4. Use coded variables in polynomial regression.
  5. Biased regression techniques.

# Solutions to Multicollinearity

- The following are some common solutions to multicollinearity:
  1. Drop one of the correlated variables.
  2. Avoid making inferences about the parameter estimates.
  3. Develop and design an experiment instead.
  4. Use coded variables in polynomial regression.
  5. Biased regression techniques.

# Higher-Order Modeling

- It is good practice to **code** higher-order variables in any linear regression model.
- To code a variable means to transform a set of independent variables into a new set of independent variables.

# Multicollinearity

- In higher-order models, multicollinearity is unavoidable because a variable will always be correlated with its own higher-order counterparts.

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2$$

# Quantitative Variable Coding

- The following is how to code an observational quantitative variable:

$$u_i = \frac{x_i - \bar{x}}{s_x}$$

- Another name for this process is called standardizing.

# Example

- A real estate company is trying to model housing prices (in dollars) of their customers with the variables:
  - $x_1$ : Size of Home (square feet)
  - $x_2$ : Age of Home (years)
  - $x_3$ : Acreage of Land (acres)
  - $x_4$ : Number of Bedrooms
- Using a sample of 105 houses they derive the following model:

$$\hat{y} = 24,312 + 86.5x_1 - 324x_2 + 9,610x_3 + 3,617x_4$$

$$s_{\hat{\beta}_4} = 3480$$



# Example

- A real estate company is trying to model housing prices (in dollars) of their customers with the variables:
  - $x_1$ : Size of Home (square feet)
  - $x_2$ : Age of Home (years)
  - $x_3$ : Acreage of Land (acres)
  - $x_4$ : Number of Bedrooms
  - $x_5$ : Number of Bathrooms
- Using a sample of 105 houses they derive the following model:

$$\hat{y} = 28,438 + 62.9x_1 - 336x_2 + 9,610x_3 - 2,102x_4 + 3,498x_5$$

$$s_{\hat{\beta}_4} = 4002$$

# Example

1. Without examining the numbers, is there any potential for multicollinearity?
2. Calculate the hypothesis test for significance on the variable  $x_4$ . In the 4 variable model, the p-value was between (0.3, 0.4).

# Example

1. Without examining the numbers, is there any potential for multicollinearity?
  - Yes! Number of bathrooms could easily be correlated with number of bedrooms, and square footage.
2. Calculate the hypothesis test for significance on the variable  $x_4$ . In the 4 variable model, the p-value was between (0.3, 0.4).

$$H_0: \beta_4 = 0$$

$$H_a: \beta_4 \neq 0$$

P-value > 0.5 → DON'T REJECT  $H_0$

$$t = \frac{-2102 - 0}{4002} = -0.53$$

# Example

3. What signs do you see that might signal multicollinearity?

# Example

3. What signs do you see that might signal multicollinearity?
  - Change in signs of coefficients
  - Shifts in values of coefficients

# RESIDUAL ANALYSIS

---

# Assumptions

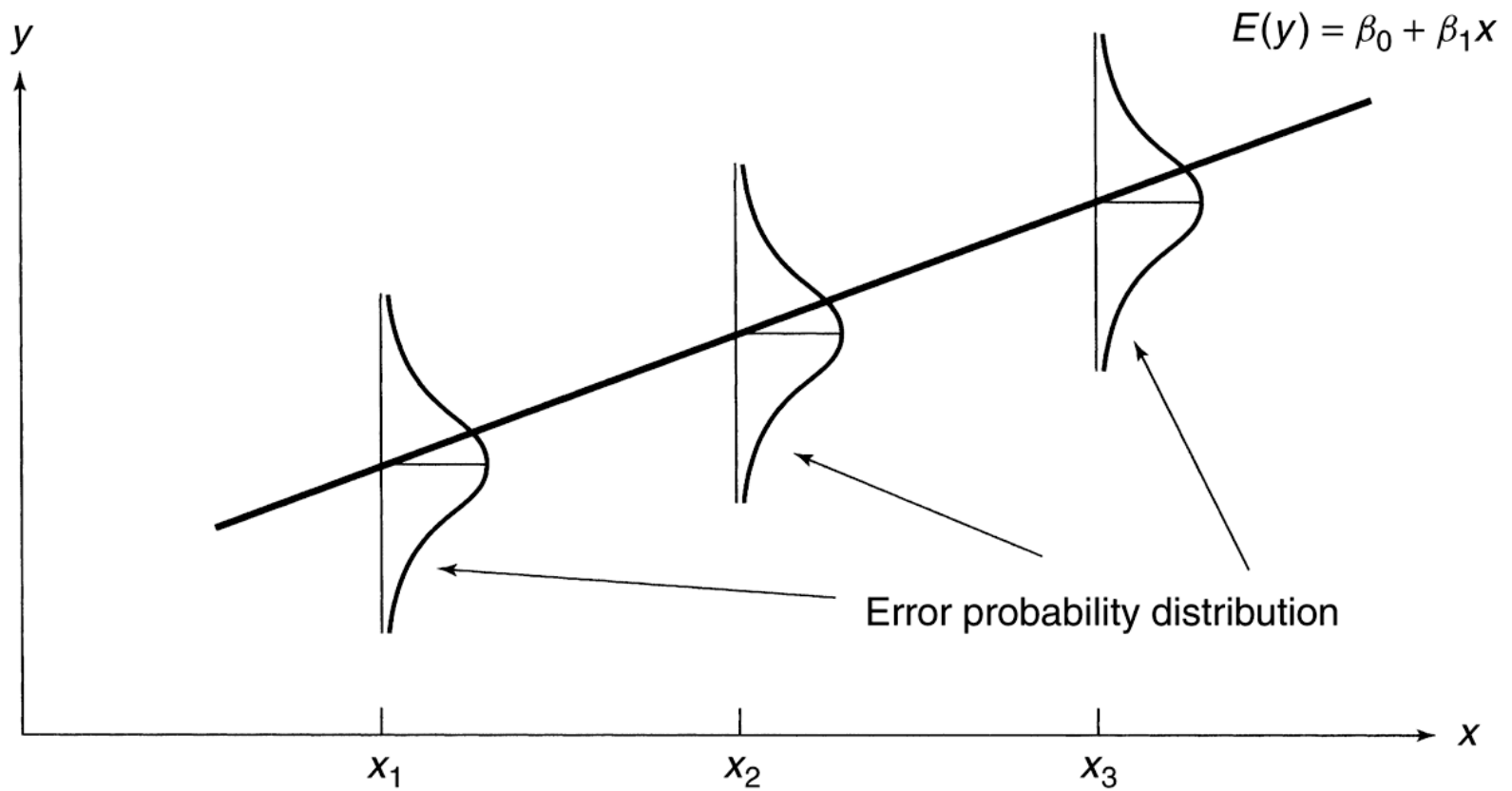
- There are four main assumptions:
  1. The mean of the probability distribution of  $\varepsilon$  is 0.
  2. The variance of the probability distribution of  $\varepsilon$  (usually denoted  $\sigma^2$ ) is constant.
  3. The probability distribution of  $\varepsilon$  is Normal.
  4. The errors associated with any two different observations are independent of each other.

# Assumptions

- There are four main assumptions:
  1. The mean of the probability distribution of  $\varepsilon$  is 0.
  2. The variance of the probability distribution of  $\varepsilon$  (usually denoted  $\sigma^2$ ) is constant.
  3. The probability distribution of  $\varepsilon$  is Normal.
  4. The errors associated with any two different observations are independent of each other.



# Assumptions




# Regression Residuals

- The true error of our model,  $\varepsilon$ , is not observed in practice.

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$$

# Regression Residuals

- The true error of our model,  $\varepsilon$ , is not observed in practice.


$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$$
$$\varepsilon = y - (\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)$$

# Regression Residuals

- The true error of our model,  $\varepsilon$ , is not observed in practice.

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$$

$$\varepsilon = y - (\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)$$

- This is due to the estimation of the  $\beta$  coefficients.
- Instead we have an estimate of the error – called a residual:

$$\hat{\varepsilon} = y - (\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k)$$


# Regression Residuals

- The true error of our model,  $\varepsilon$ , is not observed in practice.

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$$

$$\varepsilon = y - (\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)$$

- This is due to the estimation of the  $\beta$  coefficients.
- Instead we have an estimate of the error – called a residual:


$$\hat{\varepsilon} = y - (\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k)$$
$$\hat{\varepsilon}_i = y_i - \hat{y}_i$$

# RESIDUAL ANALYSIS

---

Linearity / Lack-of-fit

# Misspecified Models

- Previously discussed the problems with misspecifying the model – **lack of fit**.
- Residuals may reveal any misspecification in modeling if the residuals have patterns.
- Residuals, just like the error they estimate, should be random.

# Detecting Lack of Fit

- Here are the steps for detecting a lack of fit in a model:
  1. Plot the residuals against all of the independent variables separately.
  2. Plot the residuals against the predicted value of the response variable.
  3. Look for patterns in these plots.



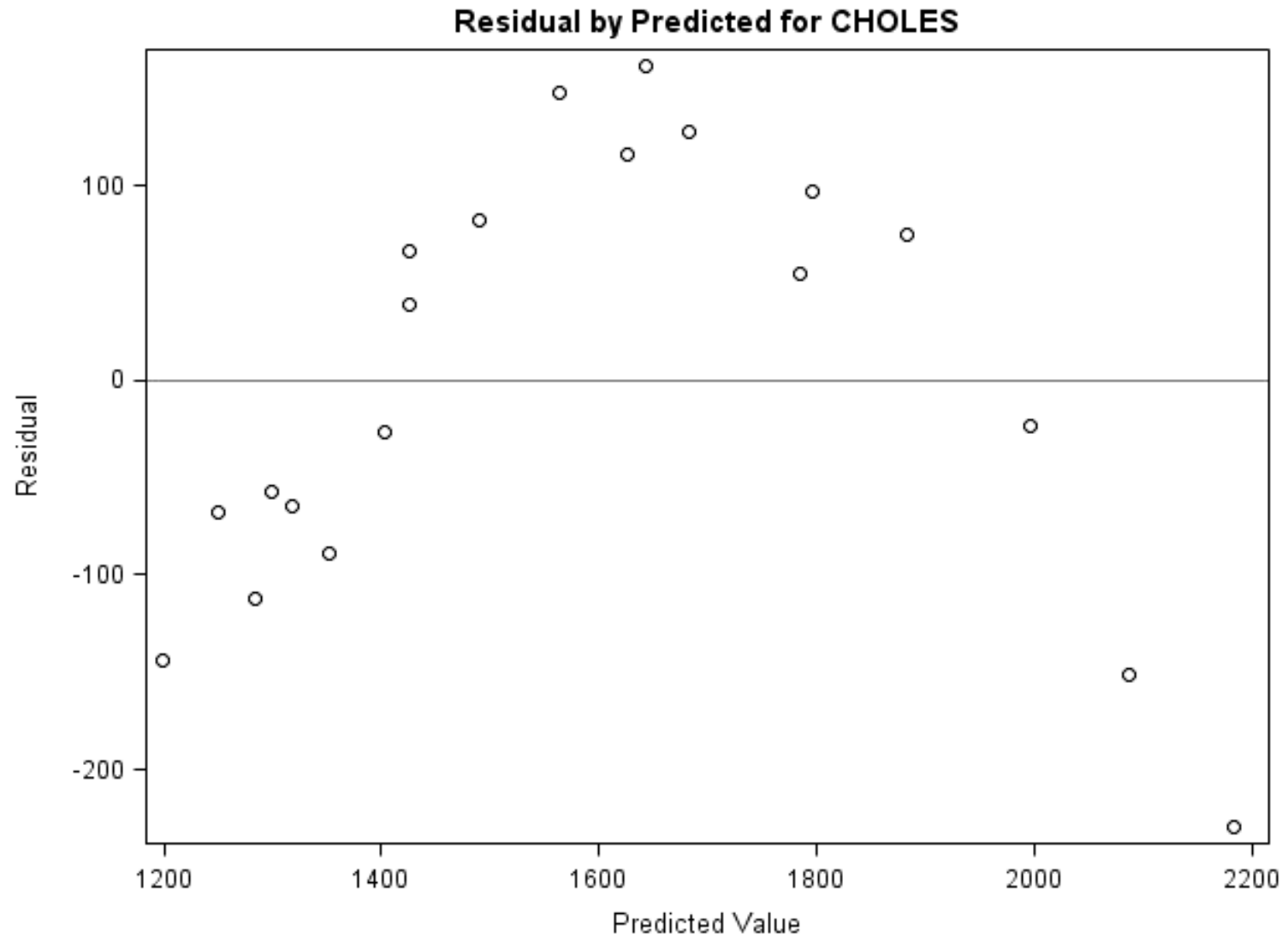
# Detecting Lack of Fit

- Here are the steps for detecting a lack of fit in a model:
  1. Plot the residuals against all of the independent variables separately.
  2. Plot the residuals against the predicted value of the response variable.
  3. Look for patterns in these plots.
    - Trends
    - Changes in variation
    - Isolated, extreme observations

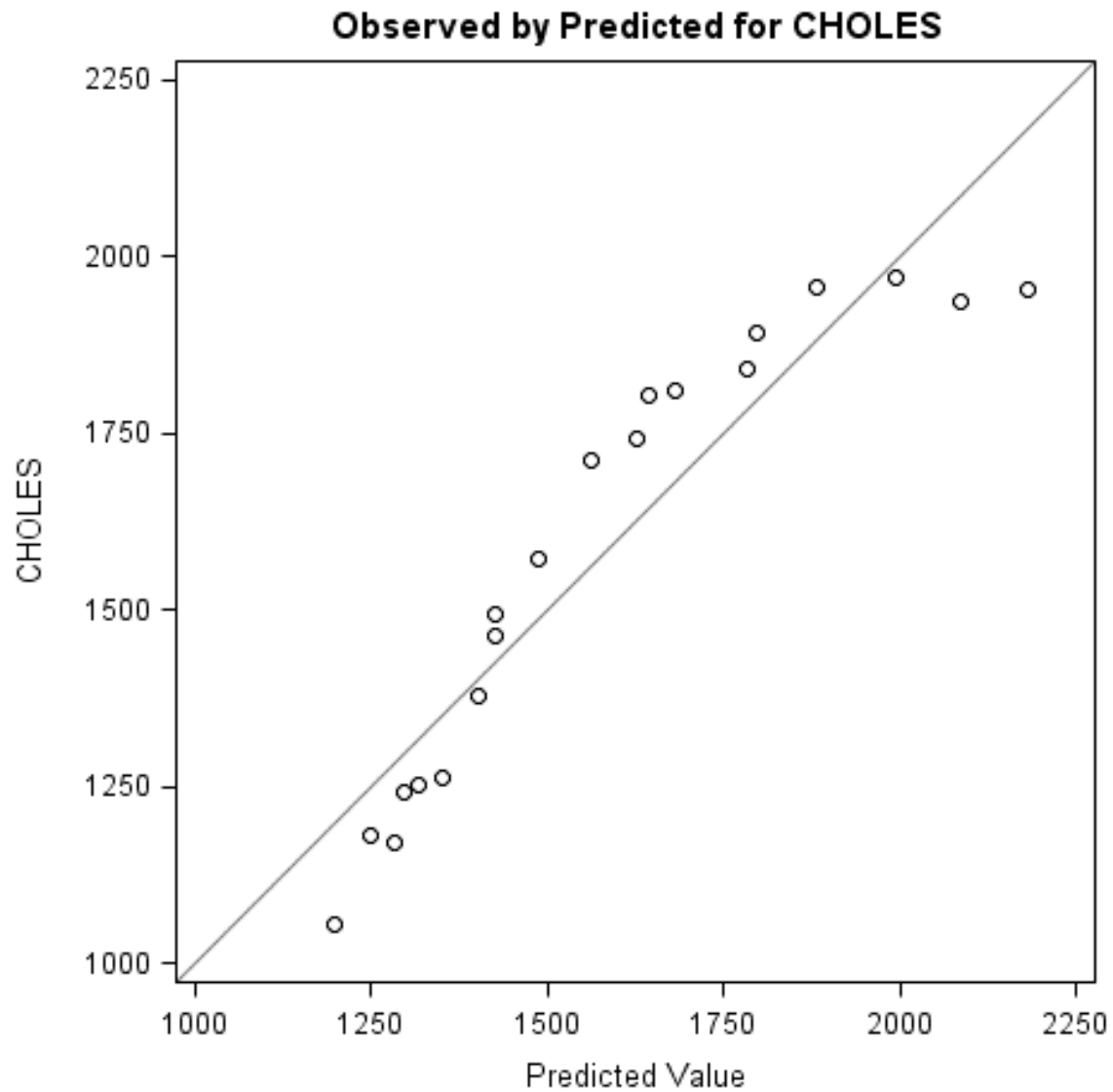
# Example

- Collected data on the level of cholesterol (mg/L) and average daily intake of saturated fat (mg) for a sample of 20 Olympic athletes. Build a model trying to predict cholesterol based on saturated fat intake.

# Example



# Example

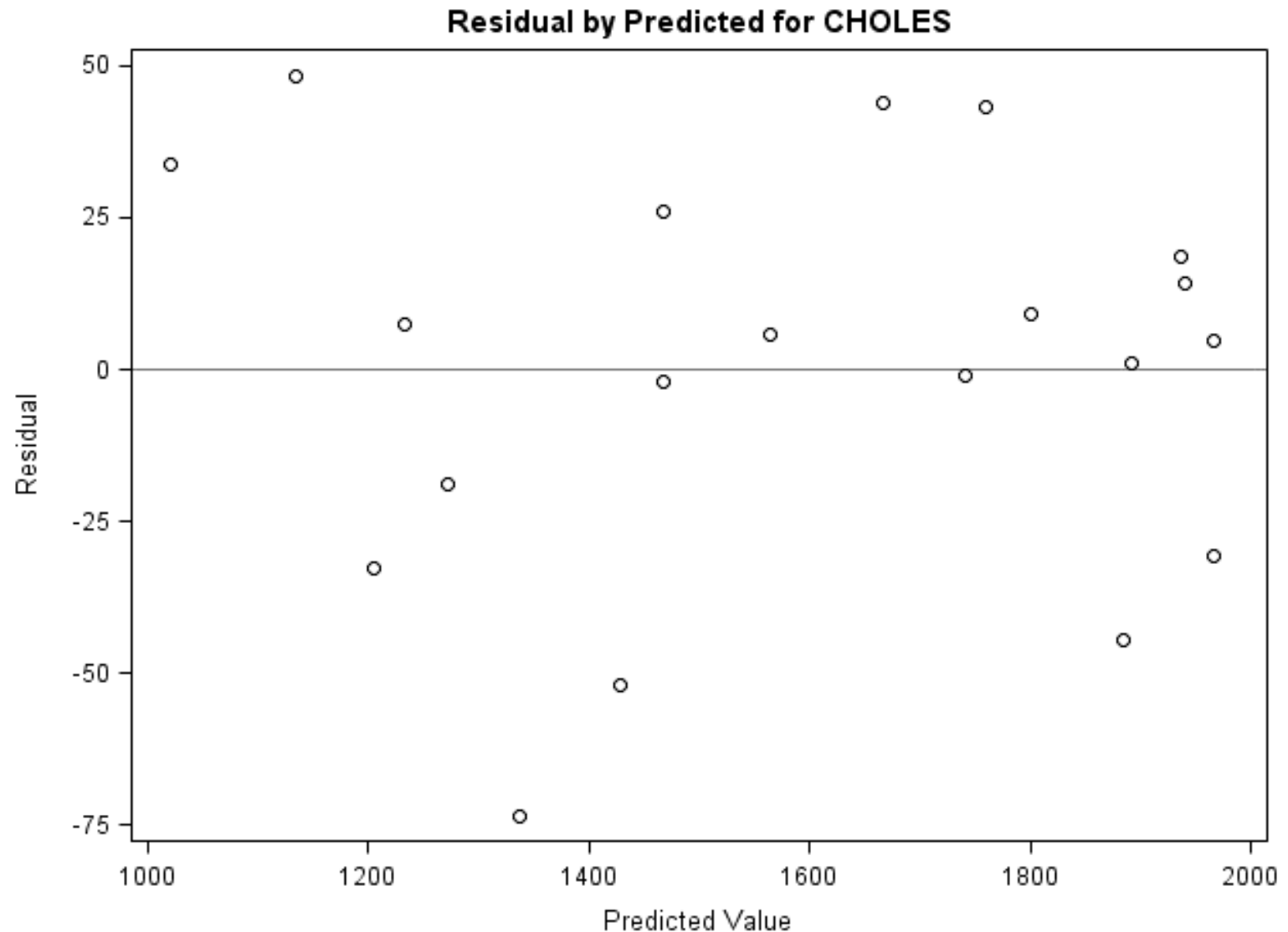


# Example

- Collected data on the level of cholesterol (mg/L) and average daily intake of saturated fat (mg) for a sample of 20 Olympic athletes. Build a model trying to predict cholesterol based on saturated fat intake.

QUADRATIC MODEL!

# Example – After Fitting Quadratic



# RESIDUAL ANALYSIS

---

Normality

# Normality

- One of the assumptions of regression is that the probability distribution of  $\varepsilon$  is Normal.
- This is a very hard assumption to meet in practice.
- Does not alter results very much if the assumption is not met on a small scale.



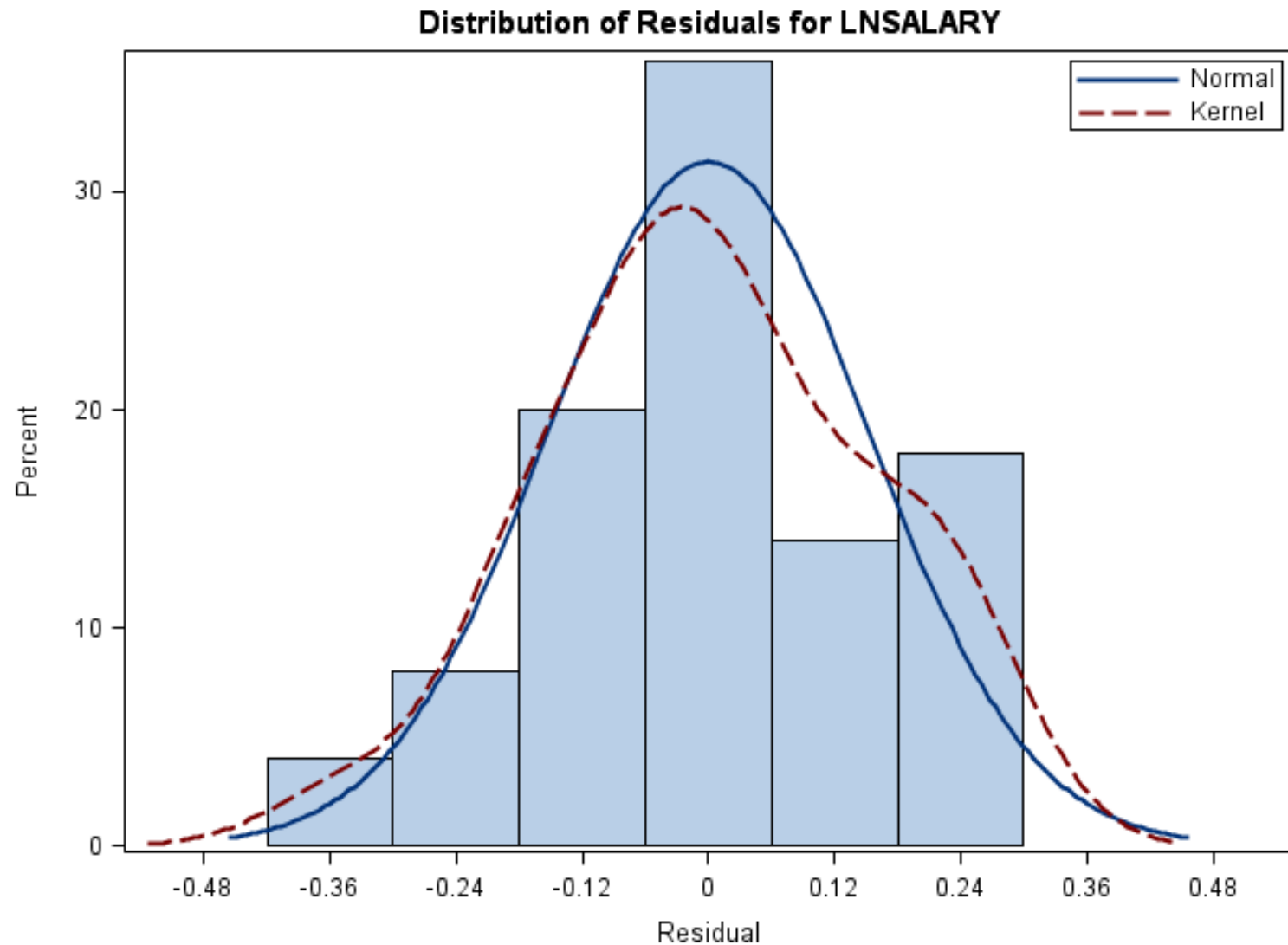
# Visualization

- Common techniques for checking for Normality of the error distribution are visual.
- Two common techniques are to look at:
  1. Histograms of the residuals
  2. Normal probability plots (QQ-plots) of the residuals.

# Example

- Want to investigate the relationship between experience and salary for social workers. Believe there might be a quadratic relationship in the data.

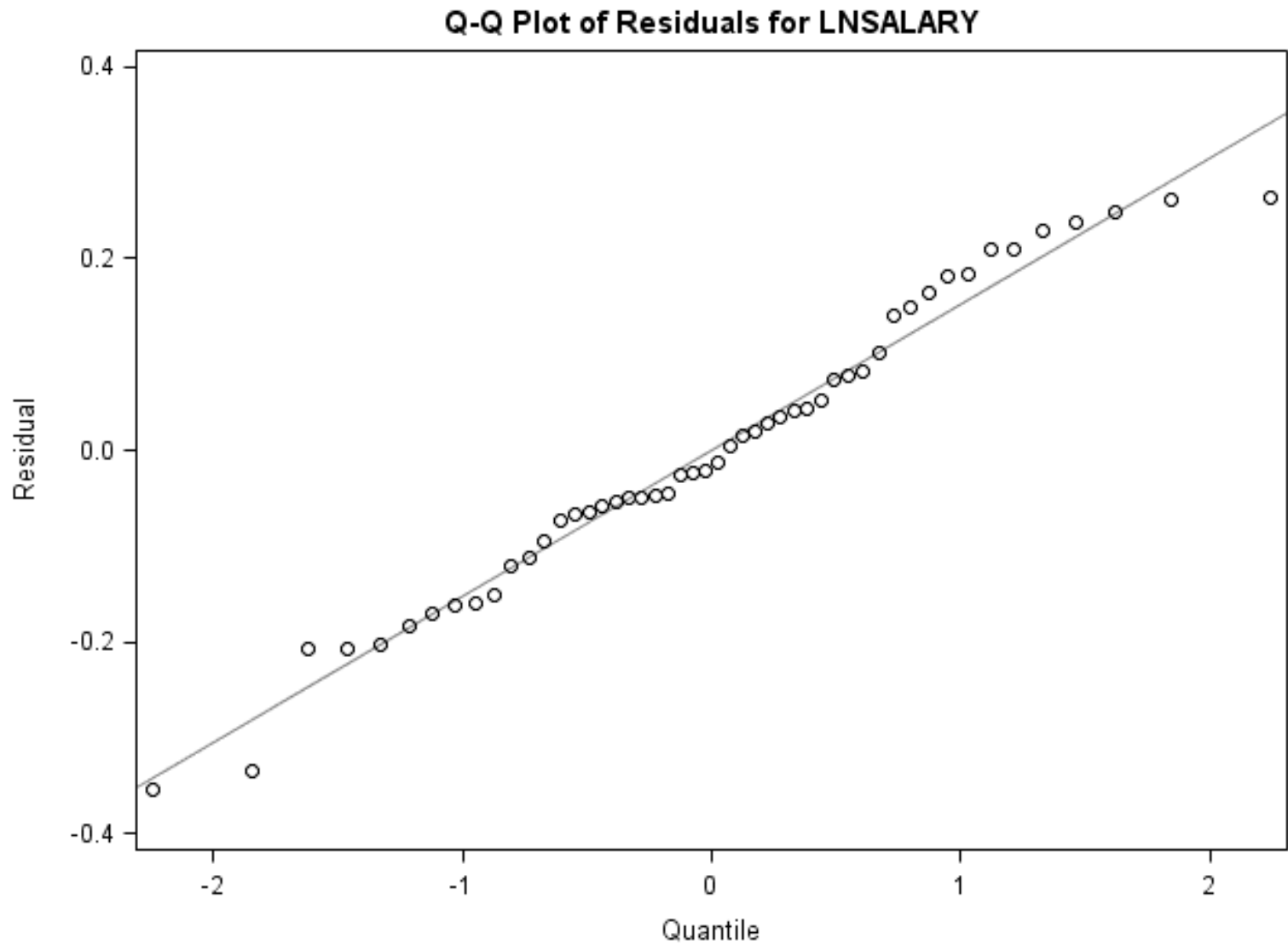
# Example



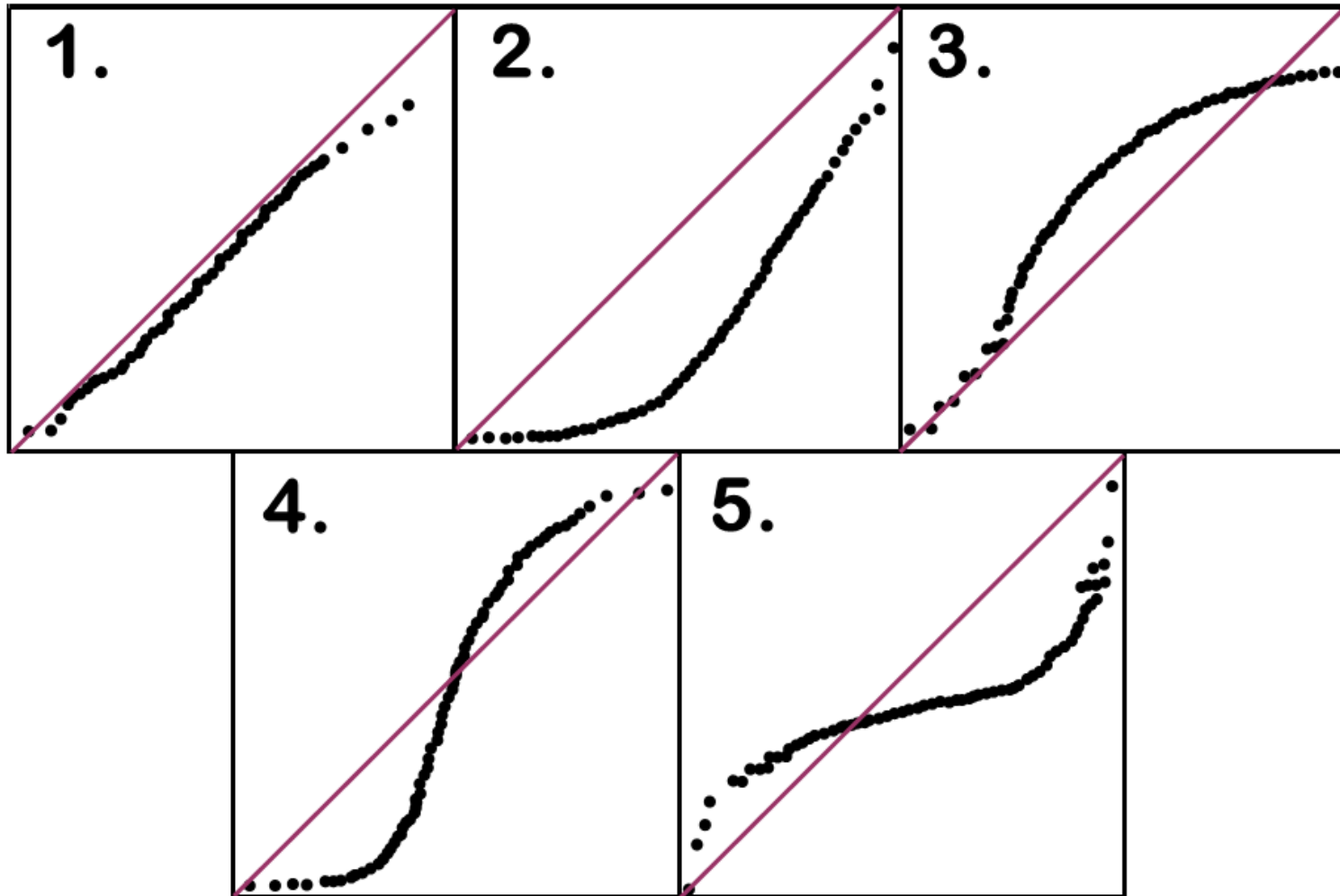
# Visualization

- Normal probability plots of residuals against expected quantiles from a Normal distribution with the same mean and standard deviation as the residuals.
- If the residuals are approximately equal to their expected place on the Normal distribution, a straight diagonal line is formed.
- Departures from a straight line are signs of the assumption not being met.

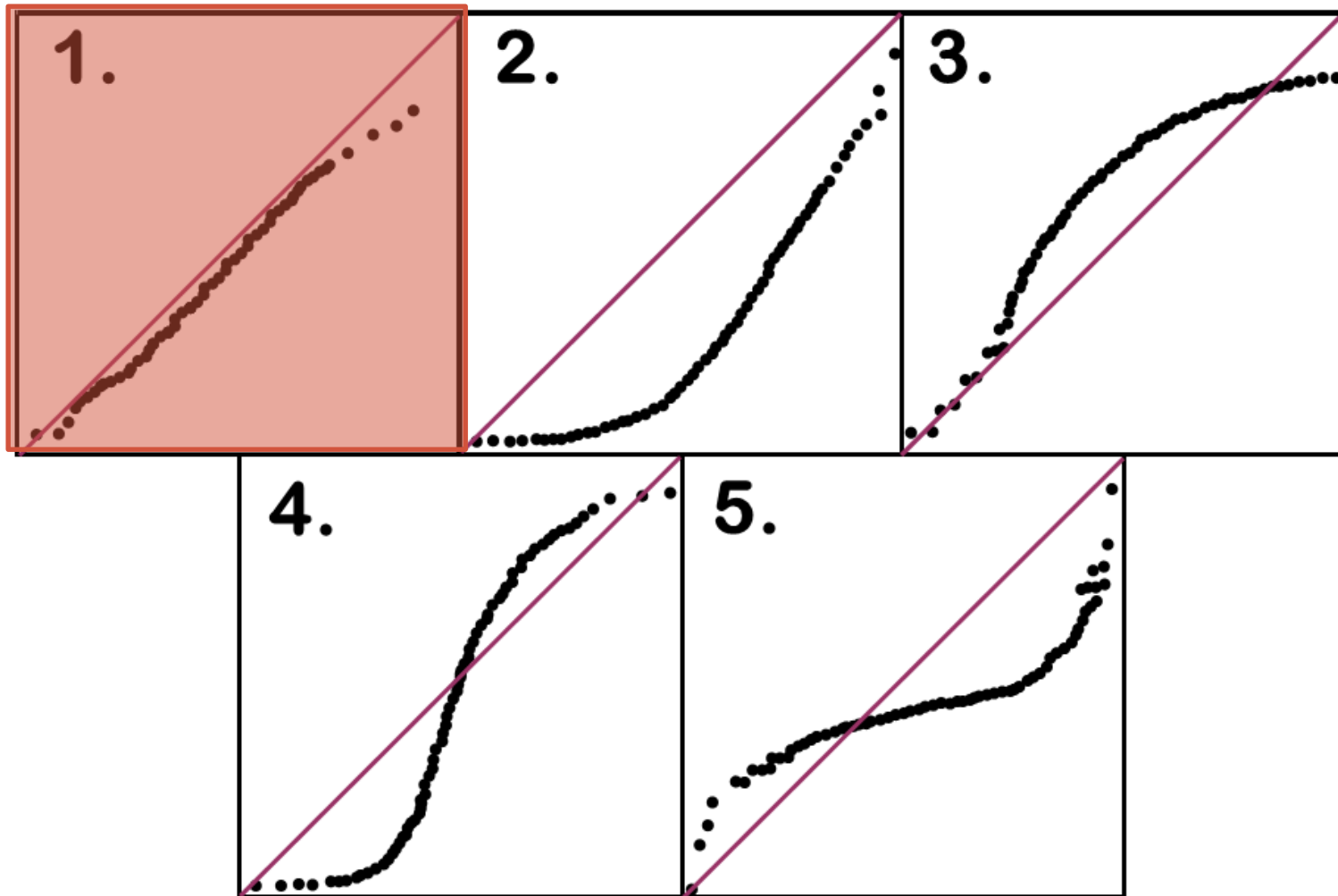
# Example



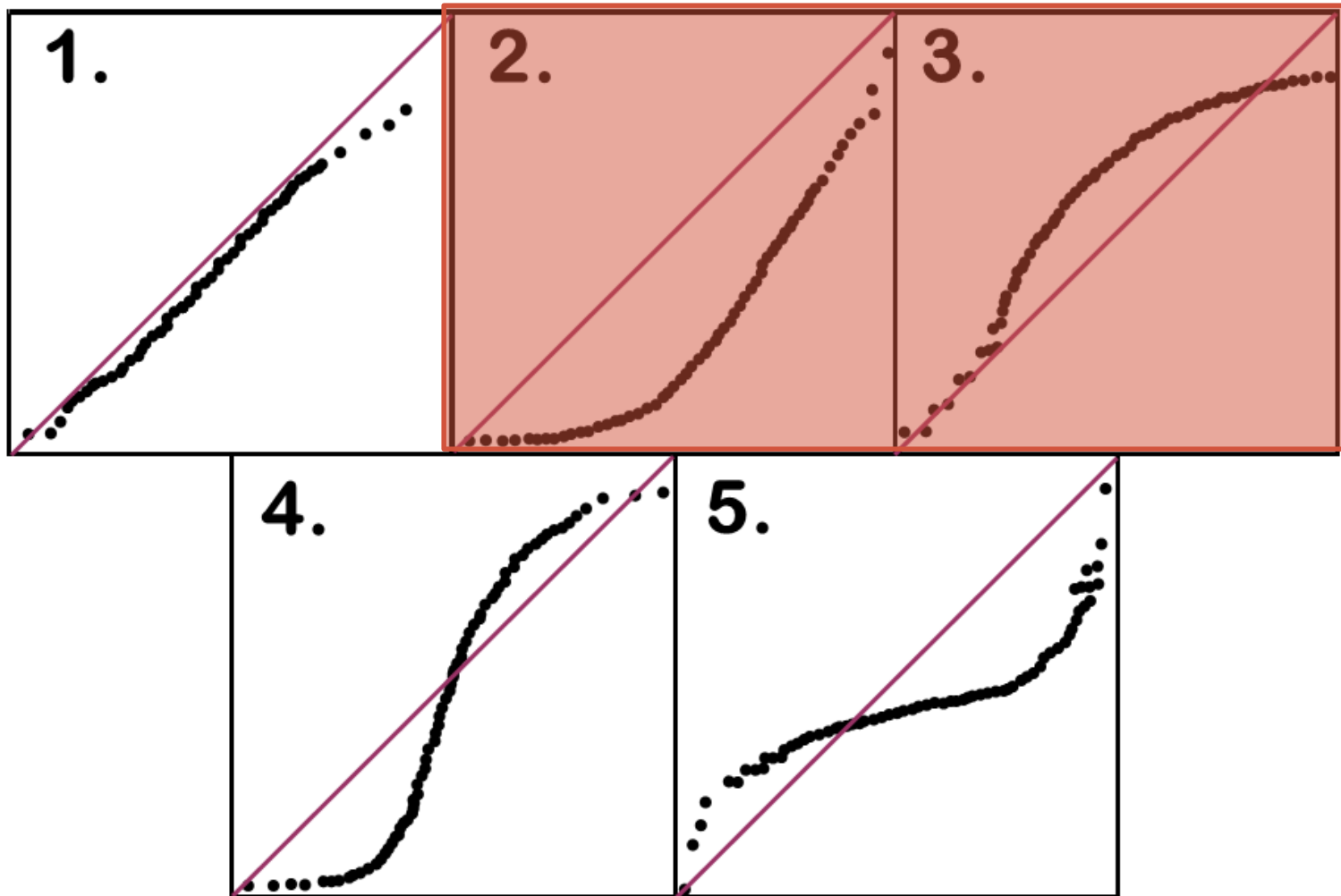
# Example



# Example – Normality Met!

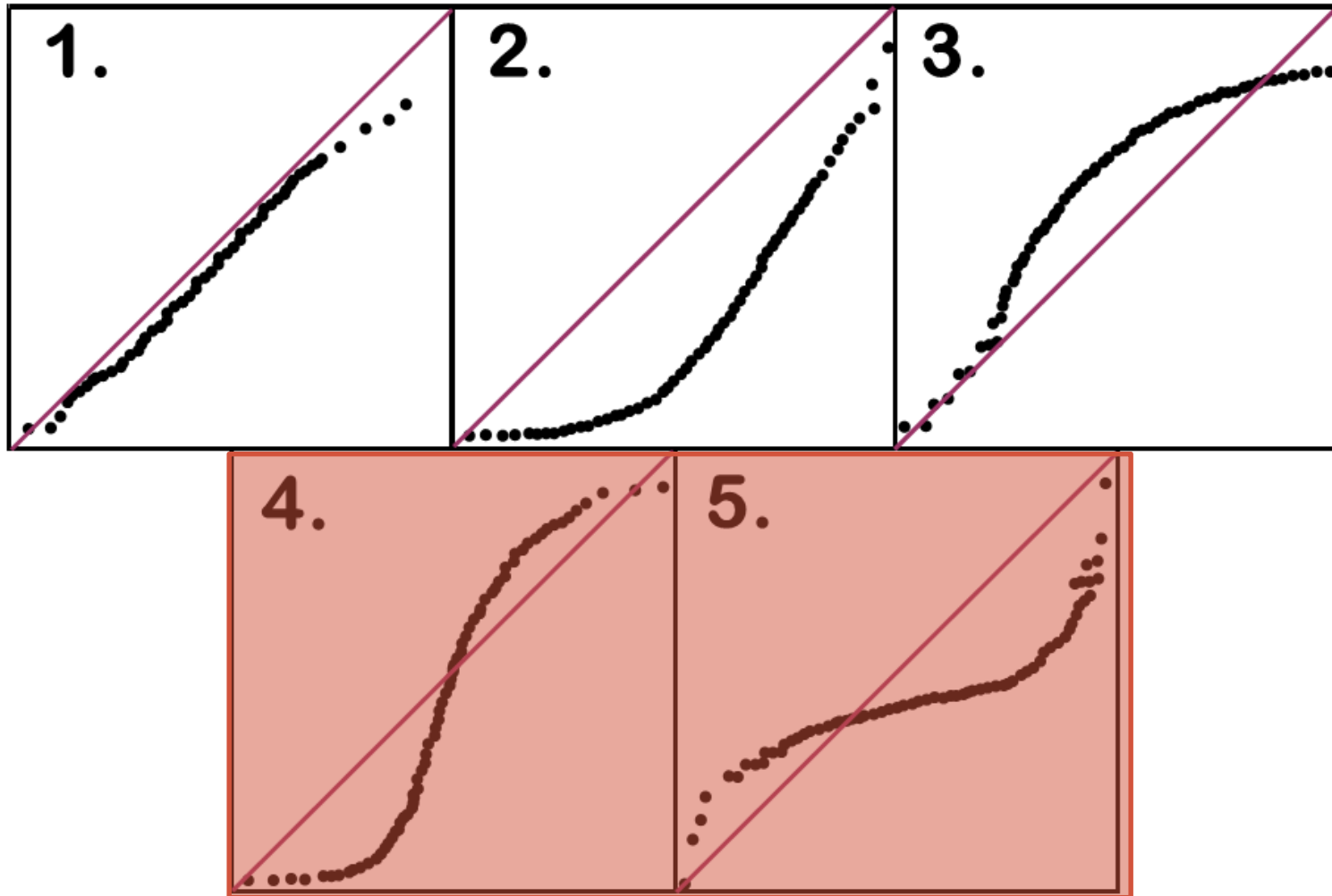


# Example – Skewness Problem





# Example – Kurtosis Problem



# Formal Test for Normality

- Visualizations can be difficult at times.
- Formal tests for normality can come in handy to put more of a statistical feel on the question of normality.
- Two popular tests for normality:
  1. Kolmogorov-Smirnov (K-S)
  2. Anderson-Darling

# Formal Test for Normality

- Visualizations can be difficult at times.
- Formal tests for normality can come in handy to put more of a statistical feel on the question of normality.
- Two popular tests for normality:
  1. Kolmogorov-Smirnov (K-S)
  2. Anderson-Darling
- Both have the same **hypotheses!**

$H_0$ : Normality

$H_a$ : Not Normal

# Possible Solution

- One possible solution for having residuals that don't follow a normal distribution is to **transform the dependent variable**.
- A common statistical transformation is the **Box-Cox transformation**:

$$y' = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log(y), & \lambda = 0 \end{cases}$$

- The value of  $\lambda$  is optimized to best fit the data and typically done with the help of computers.

# RESIDUAL ANALYSIS

---

Detecting Unequal Variance

# Heteroscedasticity

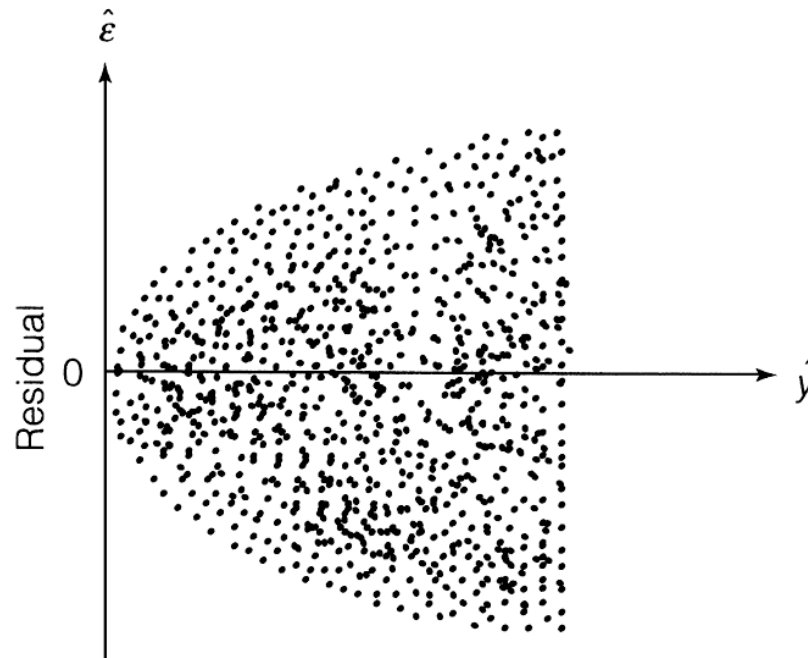
- One of the assumptions of linear regression is that the variance of the probability distribution of  $\varepsilon$  (usually denoted  $\sigma^2$ ) is constant.
- This property is called **homoscedasticity**.
- The opposite of this property is called **heteroscedasticity**.

# Heteroscedasticity

- Here are three common examples of when the assumption of homoscedasticity is broken:

# Heteroscedasticity

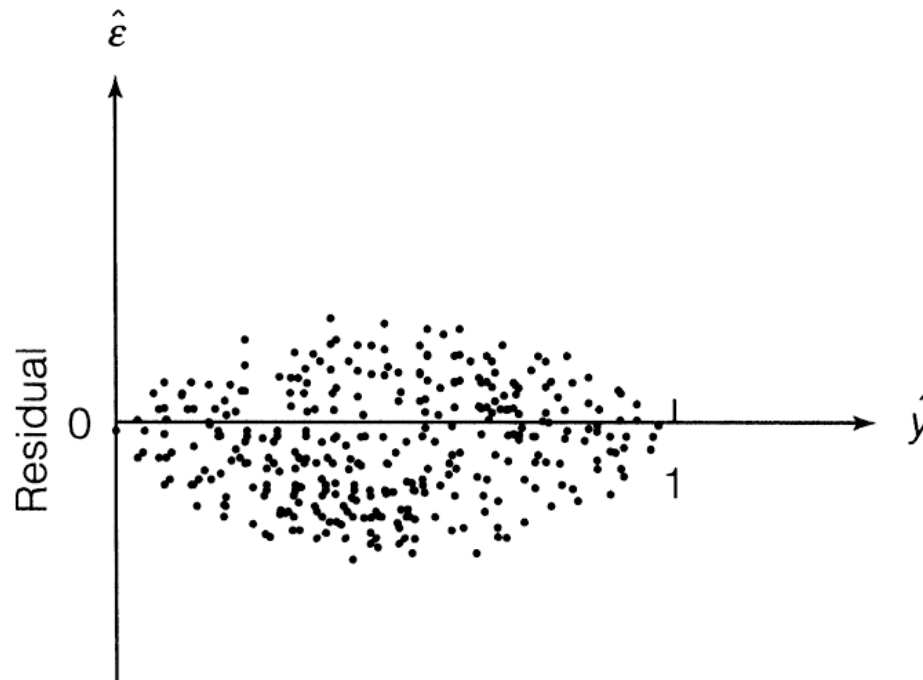
- Here are three common examples of when the assumption of homoscedasticity is broken:
  1. If the response variable is a count variable that follows a Poisson distribution.





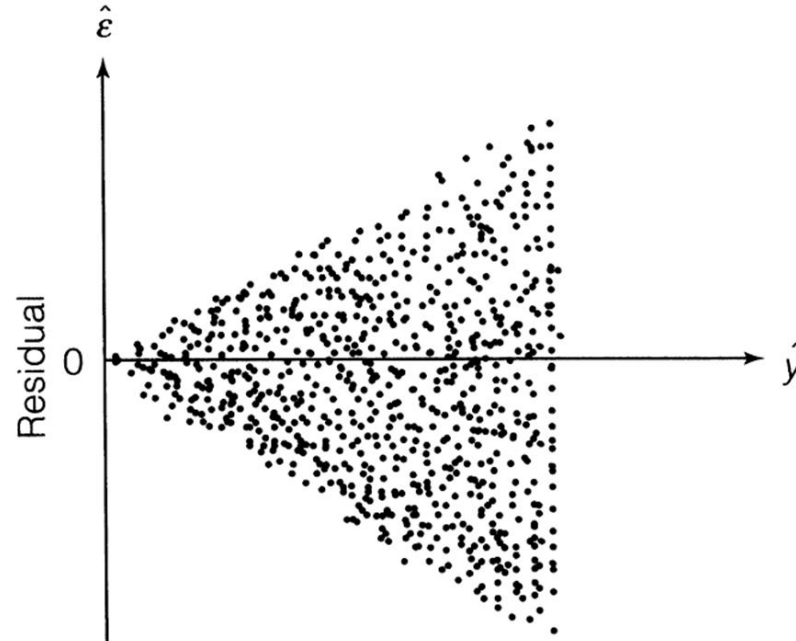
# Heteroscedasticity

- Here are three common examples of when the assumption of homoscedasticity is broken:
  2. The response variable is a proportion generated from a binomial distribution.



# Heteroscedasticity

- Here are three common examples of when the assumption of homoscedasticity is broken:
  3. The error is modeled as additive, when it is actually multiplicative.



# Heteroscedasticity

- Here are three common examples of when the assumption of homoscedasticity is broken:
  1. If the response variable is a count variable that follows a Poisson distribution.
  2. The response variable is a proportion generated from a binomial distribution.
  3. The error is modeled as additive, when it is actually multiplicative.

# Stabilizing Variance

- How do you get around heteroscedasticity?
- One possibility is to transform your response variable with a **variance-stabilizing transformation**.
- A variance-stabilizing transformation is a transformation that converts a heteroscedastic model into a homoscedastic one.

# Transformations

- Here are three common transformations to fix the three common causes of heteroscedasticity:
  1. Poisson: Transformation  $\rightarrow \sqrt{y}$
  2. Binomial: Transformation  $\rightarrow \sin^{-1} \sqrt{y}$
  3. Multiplicative: Transformation  $\rightarrow \log(y)$

# Transformations

- Here are three common transformations to fix the three common causes of heteroscedasticity:
  1. Poisson: Transformation  $\rightarrow \sqrt{y}$
  2. Binomial: Transformation  $\rightarrow \sin^{-1} \sqrt{y}$
  3. Multiplicative: Transformation  $\rightarrow \log(y)$
- Another possible solution to heteroscedasticity is **weighted least squares (WLS)**, which will be discussed later in the program.

# RESIDUAL ANALYSIS

---

Independence of Errors

# Time Series

- Cross-sectional data is collected across different individuals at a specific point in time.
- **Time series** data is collected for one individual across consecutive points in time.



# Error Correlation

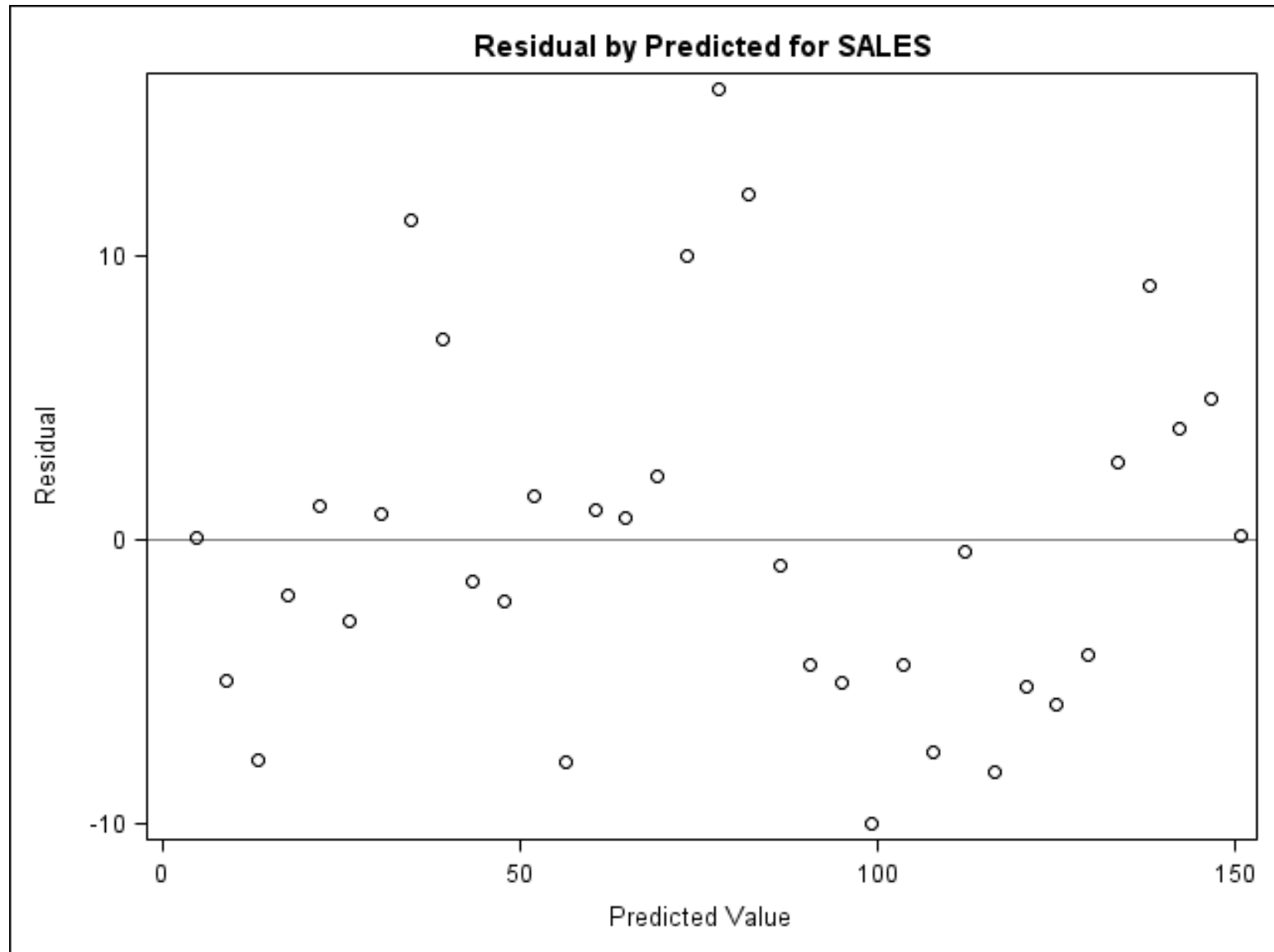
- Regression models of time series data can lead to problems in the modeling process.
- The value of a times series at the point in time  $t$  is often related to the value at the point in time  $t + 1$ .
- Errors are now correlated with each other – **underestimated  $\beta$  coefficients.**

# Example

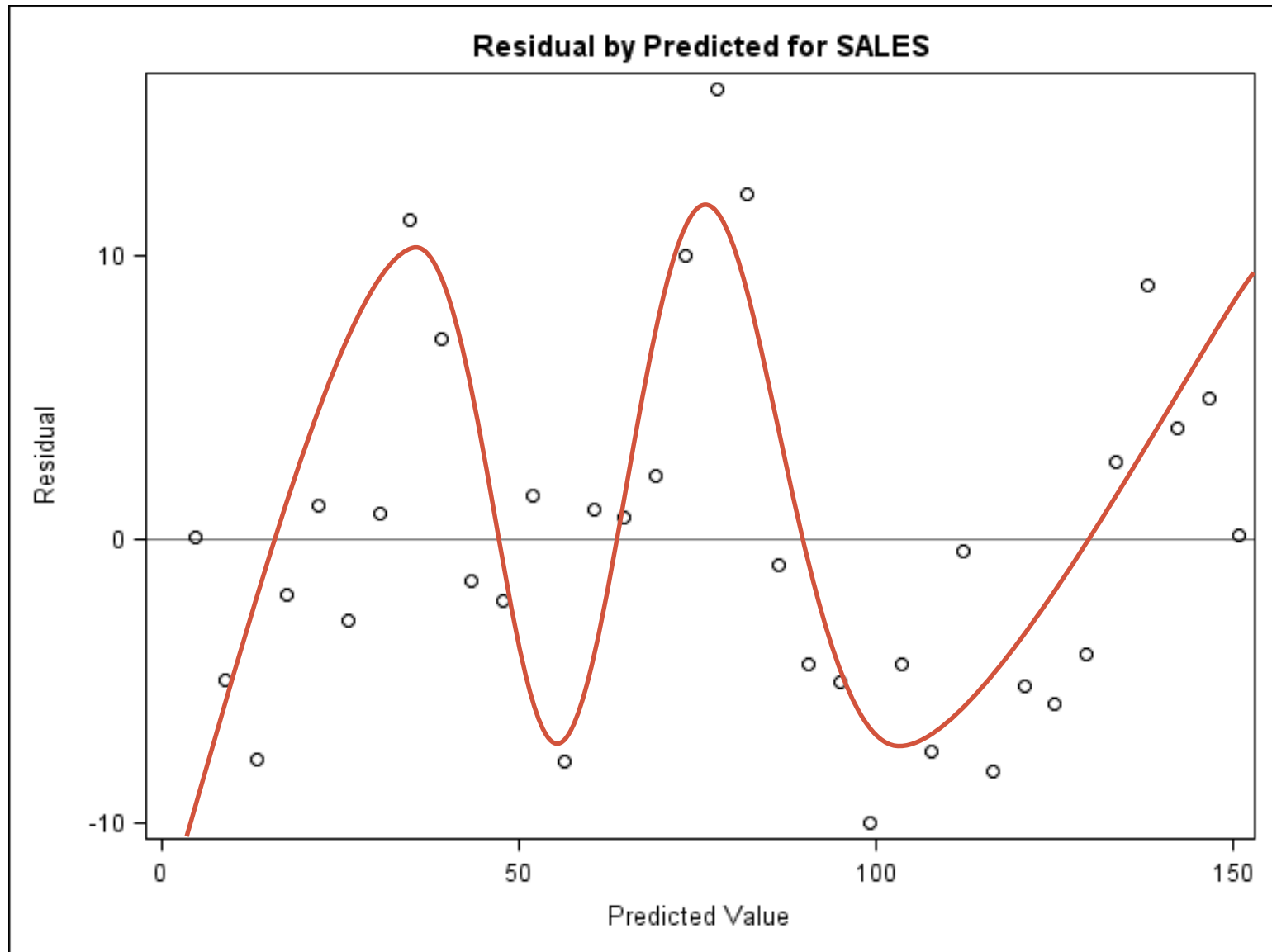
- Measure annual sales revenue data for a company across 35 years. Try to predict sales based on time.

$$Sales = \beta_0 + \beta_1 t + \varepsilon$$

# Example



# Example



# Durbin-Watson Test

- Want to test if there is possible correlation in our residuals.

$H_0$ : No residual correlation

$H_a$ : Residual correlation

- The Durbin-Watson  $d$  statistic tests for residual correlation:

$$d = \frac{\sum_{t=2}^n (\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{t=1}^n \hat{\varepsilon}_t^2}$$

# Durbin-Watson Test

- Want to test if there is possible correlation in our residuals.

$H_0$ : No residual correlation

$H_a$ : Residual correlation

- The Durbin-Watson  $d$  statistic tests for residual correlation:

$$d = \frac{\sum_{t=2}^n (\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{t=1}^n \hat{\varepsilon}_t^2}$$

Difference between errors at successive points in time.

# Properties of Durbin-Watson Test

- The Durbin-Watson  $d$  statistic tests for residual correlation:

$$d = \frac{\sum_{t=2}^n (\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{t=1}^n \hat{\varepsilon}_t^2}$$

- The  $d$  statistic has the following properties:
  1.  $0 \leq d \leq 4$
  2.  $d \approx 2$ : uncorrelated
  3.  $d < 2$ : positively correlated
  4.  $d > 2$ : negatively correlated

# Properties of Durbin-Watson Test

- The  $d$  statistic has the following properties:
  1.  $0 \leq d \leq 4$
  2.  $d \approx 2$ : uncorrelated
  3.  $d < 2$ : positively correlated
  4.  $d > 2$ : negatively correlated
- Sampling distribution for  $d$  is very complex, so no direct cut-off can be calculated for the statistic.
- Have to use approximations.



# RESIDUAL ANALYSIS

---

Outliers and Influential Observations

# Outliers

- Observations with **residuals** that are extremely large are called **outliers**.
- How large is extremely large?
  - 3 standard deviations away from 0 (the mean).
- To find if these residuals are 3 standard deviations away, **standardized residuals** need to be calculated.

# Standardized Residuals

- To find if these residuals are 3 standard deviations away, **standardized residuals** need to be calculated:

$$\hat{\varepsilon}_i^* = \frac{\hat{\varepsilon}_i}{s} = \frac{(y_i - \hat{y}_i)}{s}$$

# Studentized Residuals

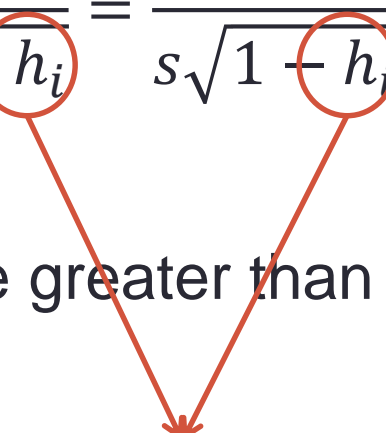
- Instead of calculating standardized residuals, **studentized residuals** are also popular to calculate:

$$\hat{\varepsilon}_i^{**} = \frac{\hat{\varepsilon}_i}{s\sqrt{1 - h_i}} = \frac{(y_i - \hat{y}_i)}{s\sqrt{1 - h_i}}$$

- Still looking for values that are greater than 3.

# Studentized Residuals

- Instead of calculating standardized residuals, **studentized residuals** are also popular to calculate:

$$\hat{\varepsilon}_i^{**} = \frac{\hat{\varepsilon}_i}{s\sqrt{1-h_i}} = \frac{(y_i - \hat{y}_i)}{s\sqrt{1-h_i}}$$


- Still looking for values that are greater than 3.

Leverage

# Leverage

- The **leverage** of an observation,  $h_i$ , is the influence of that particular observation on the respective predicted value.
- In other words, how do the respective values of the independent variables for the  $i^{th}$  observation affect the prediction  $\hat{y}_i$ .
- Equation for calculating leverage is extremely complicated, therefore, computers are needed for calculation.

# Leverage

- Observations with large values of leverage are referred to as **influential observations**.
- How large is large?

$$h_i > \frac{2(k + 1)}{n}$$

# Cook's D

- Cook's D (distance) is another way to determine if an observation is an anomaly.
- Cook's D measures the influence of each observation on the estimated  $\beta$  coefficients:

$$D_i = \frac{(y_i - \hat{y}_i)^2}{(k + 1)MSE} \left( \frac{h_i}{(1 - h_i)^2} \right)$$



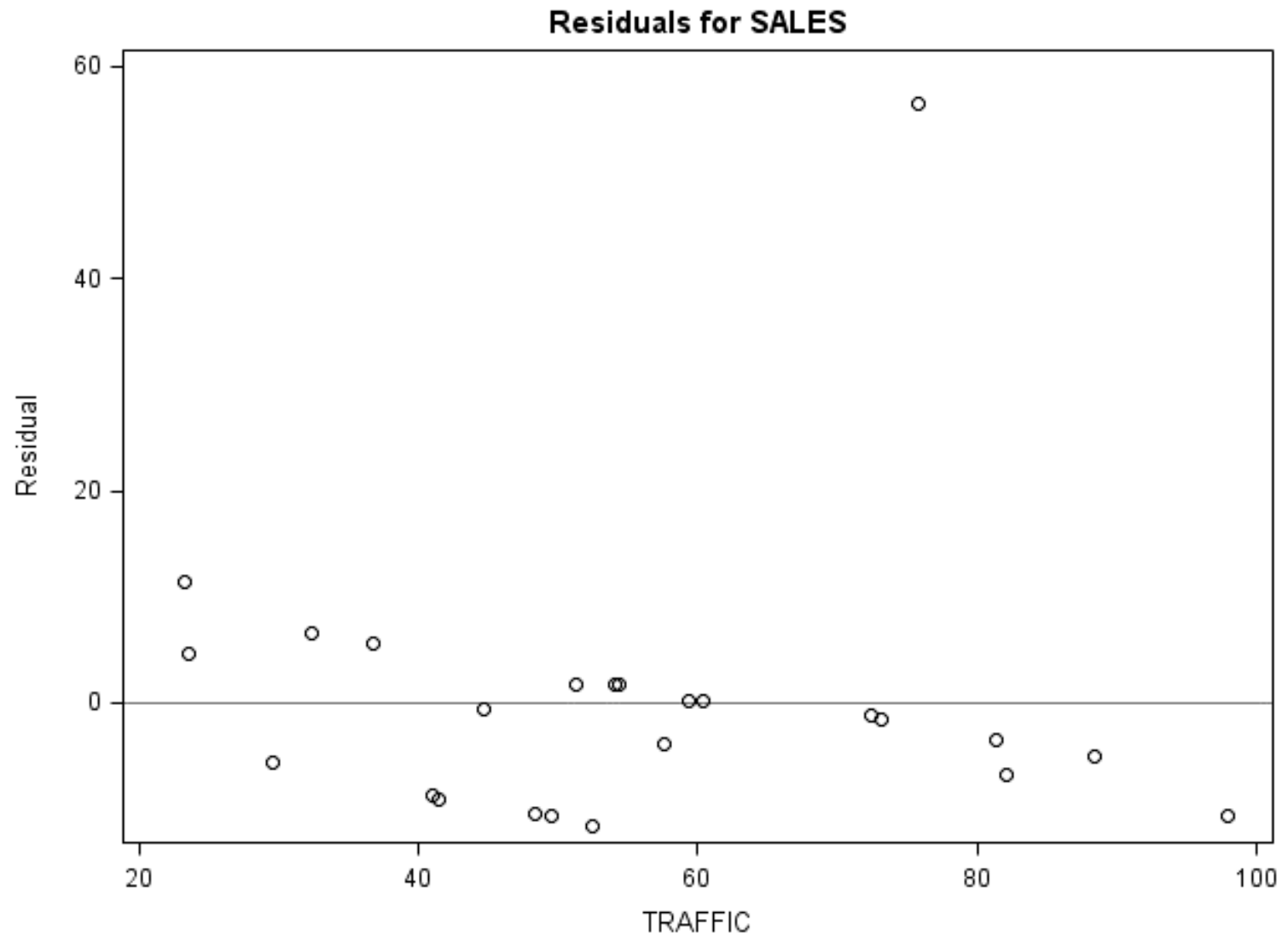
# Dffits and Dfbetas

- Two other common measures of the influence of one observation on the regression model are **Dffits** and **Dfbetas**.
- Dffits measures the difference between the predicted value of  $y$  with and without the observation in the regression.
- Dfbetas measures the difference between the estimated coefficient for each variable with and without the observation in the regression.

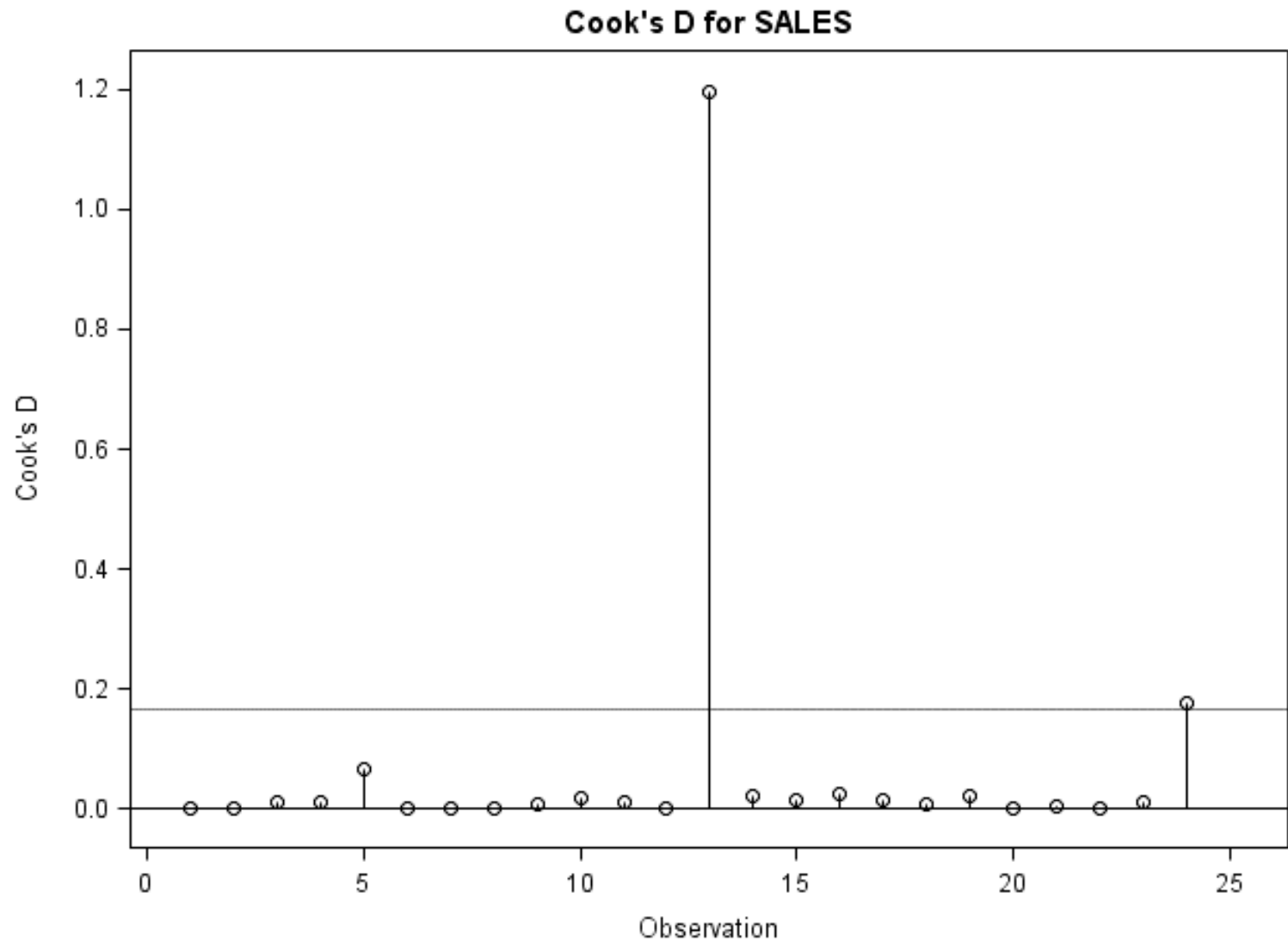
# Example

- A fast food chain measured the weekly sales of the fast food outlets in each of 4 cities along with the traffic flow of each city for a given week. Want to see if the city and traffic flow are predictors of sales.

# Example



# Example



# COMPREHENSIVE EXAMPLE

---

# Example

- The director of a company's human resources department is trying to develop a model for predicting starting salary (in dollars) of incoming employees based on the following variables:
  - $x_1$ : Years of Professional Work Experience
  - $x_2$ : Age in Years
  - $x_3$ : Gender (1 if Male, 0 if Female)
  - $x_4$ : Previous Work Salary
  - $x_5$ : Communication Test (assigns scores from 0 to 100)
- Using a sample of 108 employees they derive the following model:

$$\hat{y} = 15,006.2 + 1,365.5x_1 + 309.3x_2 + 1,822.7x_3 + 0.7x_4 + 46.2x_5$$

# Example

1. Fill in the blanks to the following table:

Source	DF	SS	MS	F-Value	P-Value
Model		89,816.28			
Error		45,671.59			
Total		135,487.87			

Parameter	Estimate	Std. Error	T-Value	P-Value
Intercept	15,006.2	5,983.5	2.508	0.0138
$x_1$	1,365.5	204.8		
$x_2$	309.9	174.0		
$x_3$	1,822.7	1,368.1		
$x_4$	0.7	0.2		
$x_5$	46.2	14.3		

# Example

1. Fill in the blanks to the following table:

Source	DF	SS	MS	F-Value	P-Value
Model	5	89,816.28	17,963.26	40.12	< 0.05
Error	102	45,671.59	447.76		
Total	107	135,487.87			

Parameter	Estimate	Std. Error	T-Value	P-Value
Intercept	15,006.2	5,983.5	2.508	0.0138
$x_1$	1,365.5	204.8	6.67	< 0.001
$x_2$	309.9	174.0	1.78	(0.01, 0.05)
$x_3$	1,822.7	1,368.1	1.33	(0.1, 0.2)
$x_4$	0.7	0.2	3.50	< 0.001
$x_5$	46.2	14.3	3.23	(0.001, 0.002)



# Example

2. Is the overall model significant? Explain.
3. What would be a predicted starting salary for a 35 years old male with 10 years of previous work experience at a previous salary of \$48,000 that scored an 89 on the communication test? How about a female with the same qualifications?

# Example

2. Is the overall model significant? Explain.
  - Yes, the test of overall significance is rejected and therefore implies that at least one variable is useful.
3. What would be a predicted starting salary for a 35 years old male with 10 years of previous work experience at a previous salary of \$48,000 that scored an 89 on the communication test? How about a female with the same qualifications?

$$\hat{y}_M = 15006.2 + 1365.5(10) + 309.3(35) + 1822.7(1) + 0.7(48000) + 46.2(89)$$

$$\hat{y}_M = \$79,021.20$$

$$\hat{y}_F = \$77,198.50$$

# Example

4. Regardless of significance, interpret the coefficient of the variable for gender.
5. The director of human resources says they are offended that the model predicts the company pays men more than women. Would the director be statistically correct in making this statement? Explain.

# Example

4. Regardless of significance, interpret the coefficient of the variable for gender.
  - **All else equal**, males earn \$1,822.70 more than females on **average**.
5. The director of human resources says they are offended that the model predicts the company pays men more than women. Would the director be statistically correct in making this statement? Explain.
  4. **No, based on the t-test, we don't have enough evidence to say gender is significant in predicting income.**

# Example

6. Calculate the  $R^2$  and  $R_a^2$ .
7. Based solely on the variables in the model, is there any potential for multicollinearity? Which variables?

# Example

6. Calculate the  $R^2$  and  $R_a^2$ .

$$R^2 = \frac{89816.28}{135487.87} = 0.663$$

$$R_a^2 = 1 - (1 - 0.663) \times \left( \frac{107}{102} \right) = 0.646$$

7. Based solely on the variables in the model, is there any potential for multicollinearity? Which variables?
- Age and work experience
  - Previous work salary with age, work experience, and communication.

# Example

8. Fill in the blanks for the following table.

Parameter	$R_j^2$	VIF
$x_1$	0.54	
$x_2$	0.63	
$x_3$	0.88	
$x_4$	0.98	
$x_5$	0.21	

9. Are there variables you would suggest dropping from the model to account for multicollinearity? Explain.

# Example

8. Fill in the blanks for the following table.

Parameter	$R_j^2$	VIF
$x_1$	0.54	2.174
$x_2$	0.63	2.703
$x_3$	0.88	8.333
$x_4$	0.98	50.0
$x_5$	0.21	1.266

9. Are there variables you would suggest dropping from the model to account for multicollinearity? Explain.
- Previous work salary has a high value of VIF ( $> 10$ ).



# Example

10. What potential problems arise from this multicollinearity?
11. Now which variables should we focus on dropping?

# Example

10. What potential problems arise from this multicollinearity?
  - The values of the other tests could be incorrect, leading to misconceptions about variable relationships.
11. Now which variables should we focus on dropping?
  - Traditionally people just start dropping insignificant variables from the model one at a time based solely on p-value.
  - However, multicollinearity may lead to inaccurate t-tests, so that should be solved first!