Advanced Applications of Probability & Statistics

Linear Regression

Sudarsan N.S. Acharya (sudarsan.acharya@manipal.edu)

Questions from Data

Questions from Data

- Output and Input variables.

Questions from Data

- Output and Input variables.
- Is there a relationship between output and Input variables?

Questions from Data

- Output and Input variables.
- Is there a relationship between output and Input variables?
- How strong is the relationship?

Questions from Data

- Output and Input variables.
- Is there a relationship between output and Input variables?
- How strong is the relationship? *Accurate prediction.*

Questions from Data

- Output and Input variables.
- Is there a relationship between output and Input variables?
- How strong is the relationship? *Accurate prediction.*
- Can we quantify the effect of the relationship?

Questions from Data

- Output and Input variables.
- Is there a relationship between output and Input variables?
- How strong is the relationship? *Accurate prediction.*
- Can we quantify the effect of the relationship?
- Is the relationship approximately linear?

Questions from Data

- Output and Input variables.
- Is there a relationship between output and Input variables?
- How strong is the relationship? *Accurate prediction.*
- Can we quantify the effect of the relationship?
- Is the relationship approximately linear? *Linear Regression.*

Output & Input Variables in Linear Regression

Output & Input Variables in Linear Regression

- Output variables have other names:

Output & Input Variables in Linear Regression

- Output variables have other names: dependent variables,

Output & Input Variables in Linear Regression

- Output variables have other names: dependent variables, outcomes,

Output & Input Variables in Linear Regression

- Output variables have other names: dependent variables, outcomes, response variables,

Output & Input Variables in Linear Regression

- Output variables have other names: dependent variables, outcomes, response variables, target variables.

Output & Input Variables in Linear Regression

- Output variables have other names: dependent variables, outcomes, response variables, target variables.
- Output variables in linear regression should be *continuous*.

Output & Input Variables in Linear Regression

- Output variables have other names: dependent variables, outcomes, response variables, target variables.
- Output variables in linear regression should be *continuous*.
- Examples:

Output & Input Variables in Linear Regression

- Output variables have other names: dependent variables, outcomes, response variables, target variables.
- Output variables in linear regression should be *continuous*.
- Examples: sales in number of units,

Output & Input Variables in Linear Regression

- Output variables have other names: dependent variables, outcomes, response variables, target variables.
- Output variables in linear regression should be *continuous*.
- Examples: sales in number of units, fuel economy in Miles per gallon (mpg),

Output & Input Variables in Linear Regression

- Output variables have other names: dependent variables, outcomes, response variables, target variables.
- Output variables in linear regression should be *continuous*.
- Examples: sales in number of units, fuel economy in Miles per gallon (mpg), Body mass index (BMI) etc.

Output & Input Variables in Linear Regression

- Output variables have other names: dependent variables, outcomes, response variables, target variables.
- Output variables in linear regression should be *continuous*.
- Examples: sales in number of units, fuel economy in Miles per gallon (mpg), Body mass index (BMI) etc.
- Input variables also have other names:

Output & Input Variables in Linear Regression

- Output variables have other names: dependent variables, outcomes, response variables, target variables.
- Output variables in linear regression should be *continuous*.
- Examples: sales in number of units, fuel economy in Miles per gallon (mpg), Body mass index (BMI) etc.
- Input variables also have other names: independent variables,

Output & Input Variables in Linear Regression

- Output variables have other names: dependent variables, outcomes, response variables, target variables.
- Output variables in linear regression should be *continuous*.
- Examples: sales in number of units, fuel economy in Miles per gallon (mpg), Body mass index (BMI) etc.
- Input variables also have other names: independent variables, features,

Output & Input Variables in Linear Regression

- Output variables have other names: dependent variables, outcomes, response variables, target variables.
- Output variables in linear regression should be *continuous*.
- Examples: sales in number of units, fuel economy in Miles per gallon (mpg), Body mass index (BMI) etc.
- Input variables also have other names: independent variables, features, predictors,

Output & Input Variables in Linear Regression

- Output variables have other names: dependent variables, outcomes, response variables, target variables.
- Output variables in linear regression should be *continuous*.
- Examples: sales in number of units, fuel economy in Miles per gallon (mpg), Body mass index (BMI) etc.
- Input variables also have other names: independent variables, features, predictors, covariates.

Output & Input Variables in Linear Regression

- Output variables have other names: dependent variables, outcomes, response variables, target variables.
- Output variables in linear regression should be *continuous*.
- Examples: sales in number of units, fuel economy in Miles per gallon (mpg), Body mass index (BMI) etc.
- Input variables also have other names: independent variables, features, predictors, covariates.
- Input variables in linear regression can be a mix of *continuous* and *categorical* variables.

Output & Input Variables in Linear Regression

- Output variables have other names: dependent variables, outcomes, response variables, target variables.
- Output variables in linear regression should be *continuous*.
- Examples: sales in number of units, fuel economy in Miles per gallon (mpg), Body mass index (BMI) etc.
- Input variables also have other names: independent variables, features, predictors, covariates.
- Input variables in linear regression can be a mix of *continuous* and *categorical* variables.
- Examples:

Output & Input Variables in Linear Regression

- Output variables have other names: dependent variables, outcomes, response variables, target variables.
- Output variables in linear regression should be *continuous*.
- Examples: sales in number of units, fuel economy in Miles per gallon (mpg), Body mass index (BMI) etc.
- Input variables also have other names: independent variables, features, predictors, covariates.
- Input variables in linear regression can be a mix of *continuous* and *categorical* variables.
- Examples: advertisement budget in Dollars,

Output & Input Variables in Linear Regression

- Output variables have other names: dependent variables, outcomes, response variables, target variables.
- Output variables in linear regression should be *continuous*.
- Examples: sales in number of units, fuel economy in Miles per gallon (mpg), Body mass index (BMI) etc.
- Input variables also have other names: independent variables, features, predictors, covariates.
- Input variables in linear regression can be a mix of *continuous* and *categorical* variables.
- Examples: advertisement budget in Dollars, horse power of vehicle,

Output & Input Variables in Linear Regression

- Output variables have other names: dependent variables, outcomes, response variables, target variables.
- Output variables in linear regression should be *continuous*.
- Examples: sales in number of units, fuel economy in Miles per gallon (mpg), Body mass index (BMI) etc.
- Input variables also have other names: independent variables, features, predictors, covariates.
- Input variables in linear regression can be a mix of *continuous* and *categorical* variables.
- Examples: advertisement budget in Dollars, horse power of vehicle, individual's height, weight, education level, gender etc.

Population & Sample

Population & Sample

- In data science, it is important to distinguish between population and sample.

Population & Sample

- In data science, it is important to distinguish between population and sample.
- Example of a population parameter:

Population & Sample

- In data science, it is important to distinguish between population and sample.
- Example of a population parameter: the average height of all biological females in a city.

Population & Sample

- In data science, it is important to distinguish between population and sample.
- Example of a population parameter: the average height of all biological females in a city.
- Example of a sample statistic:

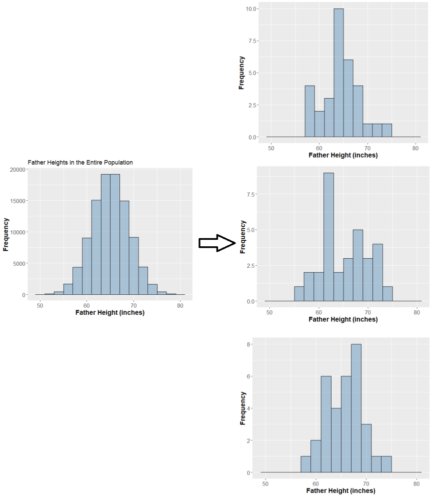
Population & Sample

- In data science, it is important to distinguish between population and sample.
- Example of a population parameter: the average height of all biological females in a city.
- Example of a sample statistic: the average height of n randomly chosen biological females in a city.

Population & Sample

- In data science, it is important to distinguish between population and sample.
- Example of a population parameter: the average height of all biological females in a city.
- Example of a sample statistic: the average height of n randomly chosen biological females in a city.
- Note that sample statistic (or just statistic) is a *random variable*.

Population & Sample - Example with Sample Size = 32



Population Model

Population Model

- We can use a *probabilistic model* for understanding the population.

Population Model

- We can use a *probabilistic model* for understanding the population.
- Example:

Population Model

- We can use a *probabilistic model* for understanding the population.
- Example: Let fuel efficiency (mpg) be the response variable and horse power (hp) be the only input variable.

Population Model

- We can use a *probabilistic model* for understanding the population.
- Example: Let fuel efficiency (mpg) be the response variable and horse power (hp) be the only input variable.
- For a random car from the population, let Y represent the mpg and X represent the hp.

Population Model

- We can use a *probabilistic model* for understanding the population.
- Example: Let fuel efficiency (mpg) be the response variable and horse power (hp) be the only input variable.
- For a random car from the population, let Y represent the mpg and X represent the hp.
- There is a probability distribution of X in the population.

Population Model

- We can use a *probabilistic model* for understanding the population.
- Example: Let fuel efficiency (mpg) be the response variable and horse power (hp) be the only input variable.
- For a random car from the population, let Y represent the mpg and X represent the hp.
- There is a probability distribution of X in the population.
- Y has a conditional probability distribution given X.

Population Model

- We can use a *probabilistic model* for understanding the population.
- Example: Let fuel efficiency (mpg) be the response variable and horse power (hp) be the only input variable.
- For a random car from the population, let Y represent the mpg and X represent the hp.
- There is a probability distribution of X in the population.
- Y has a conditional probability distribution given X.
- Population models are typically nonlinear:

Population Model

- We can use a *probabilistic model* for understanding the population.
- Example: Let fuel efficiency (mpg) be the response variable and horse power (hp) be the only input variable.
- For a random car from the population, let Y represent the mpg and X represent the hp.
- There is a probability distribution of X in the population.
- Y has a conditional probability distribution given X.
- Population models are typically nonlinear: $Y = f(X) + \epsilon$ for an unknown nonlinear function f,

Population Model

- We can use a *probabilistic model* for understanding the population.
- Example: Let fuel efficiency (mpg) be the response variable and horse power (hp) be the only input variable.
- For a random car from the population, let Y represent the mpg and X represent the hp.
- There is a probability distribution of X in the population.
- Y has a conditional probability distribution given X.
- Population models are typically nonlinear: $Y = f(X) + \epsilon$ for an unknown nonlinear function f, where ϵ is a *random error term*.

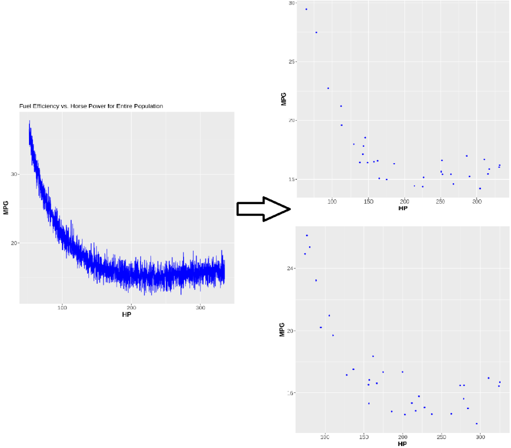
Population Model

- We can use a *probabilistic model* for understanding the population.
- Example: Let fuel efficiency (mpg) be the response variable and horse power (hp) be the only input variable.
- For a random car from the population, let Y represent the mpg and X represent the hp.
- There is a probability distribution of X in the population.
- Y has a conditional probability distribution given X.
- Population models are typically nonlinear: $Y = f(X) + \epsilon$ for an unknown nonlinear function f, where ϵ is a *random error term*.
- Example of a population model for mpg and hp:

Population Model

- We can use a *probabilistic model* for understanding the population.
- Example: Let fuel efficiency (mpg) be the response variable and horse power (hp) be the only input variable.
- For a random car from the population, let Y represent the mpg and X represent the hp.
- There is a probability distribution of X in the population.
- Y has a conditional probability distribution given X.
- Population models are typically nonlinear: $Y = f(X) + \epsilon$ for an unknown nonlinear function f, where ϵ is a *random error term*.
- Example of a population model for mpg and hp: $Y = \frac{1.8}{X} - 0.03X + \epsilon$.

Population & Sample - Another Example with Sample Size = 32



A Linear Population Model

A Linear Population Model

- For a random father-son pair in a population, let Y represent the son's height and X represent the father's height.

A Linear Population Model

- For a random father-son pair in a population, let Y represent the son's height and X represent the father's height.
- Suppose that in the population, father's heights are normally distributed with mean 65 inches and standard deviation 4 inches:

A Linear Population Model

- For a random father-son pair in a population, let Y represent the son's height and X represent the father's height.
- Suppose that in the population, father's heights are normally distributed with mean 65 inches and standard deviation 4 inches: $X \sim N(\mu = 65, \sigma^2 = 16)$.

A Linear Population Model

- For a random father-son pair in a population, let Y represent the son's height and X represent the father's height.
- Suppose that in the population, father's heights are normally distributed with mean 65 inches and standard deviation 4 inches: $X \sim N(\mu = 65, \sigma^2 = 16)$.
- Given the father's height $X = x$, suppose the son's height Y is also normally distributed with mean $42 + 0.4 \times x$ and standard deviation 3 inches:

A Linear Population Model

- For a random father-son pair in a population, let Y represent the son's height and X represent the father's height.
- Suppose that in the population, father's heights are normally distributed with mean 65 inches and standard deviation 4 inches: $X \sim N(\mu = 65, \sigma^2 = 16)$.
- Given the father's height $X = x$, suppose the son's height Y is also normally distributed with mean $42 + 0.4 \times x$ and standard deviation 3 inches: $\boxed{Y \mid (X = x)} \sim N(\mu = 42 + 0.4x, \sigma^2 = 9)$.

A Linear Population Model

- For a random father-son pair in a population, let Y represent the son's height and X represent the father's height.
- Suppose that in the population, father's heights are normally distributed with mean 65 inches and standard deviation 4 inches: $X \sim N(\mu = 65, \sigma^2 = 16)$.
- Given the father's height $X = x$, suppose the son's height Y is also normally distributed with mean $42 + 0.4 \times x$ and standard deviation 3 inches: $\boxed{Y \mid (X = x)} \sim N(\mu = 42 + 0.4x, \sigma^2 = 9)$.
- The population model for Y as a function of X is a linear one:

A Linear Population Model

- For a random father-son pair in a population, let Y represent the son's height and X represent the father's height.
- Suppose that in the population, father's heights are normally distributed with mean 65 inches and standard deviation 4 inches: $X \sim N(\mu = 65, \sigma^2 = 16)$.
- Given the father's height $X = x$, suppose the son's height Y is also normally distributed with mean $42 + 0.4 \times x$ and standard deviation 3 inches: $\boxed{Y \mid (X = x)} \sim N(\mu = 42 + 0.4x, \sigma^2 = 9)$.
- The population model for Y as a function of X is a linear one: $Y = 42 + 0.4X + \epsilon$,

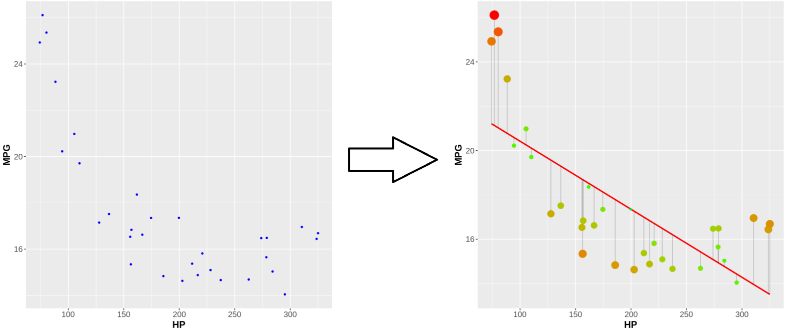
A Linear Population Model

- For a random father-son pair in a population, let Y represent the son's height and X represent the father's height.
- Suppose that in the population, father's heights are normally distributed with mean 65 inches and standard deviation 4 inches: $X \sim N(\mu = 65, \sigma^2 = 16)$.
- Given the father's height $X = x$, suppose the son's height Y is also normally distributed with mean $42 + 0.4 \times x$ and standard deviation 3 inches: $\boxed{Y \mid (X = x)} \sim N(\mu = 42 + 0.4x, \sigma^2 = 9)$.
- The population model for Y as a function of X is a linear one: $Y = 42 + 0.4X + \epsilon$, where $\epsilon \sim N(\mu = 0, \sigma^2 = 9)$.

The Geometric Idea Behind Simple Linear Regression Model (SLRM)

The Geometric Idea Behind Simple Linear Regression Model (SLRM)

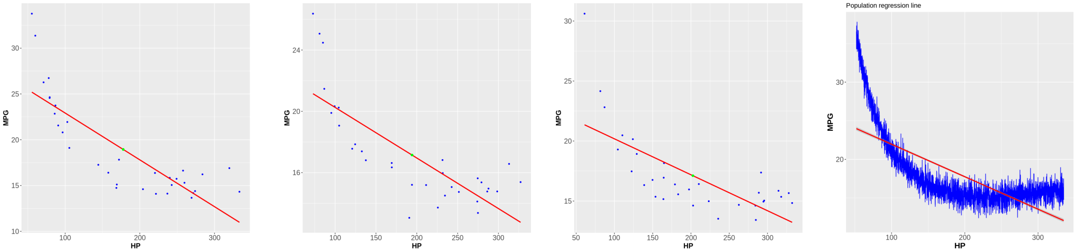
Given a dataset (*random samples from the population*), find a straight line that *fits* the data (*response variable and a single predictor*) well in an *average sense*:



Population & Sample - Revisited in the Context of SLRM

Population & Sample - Revisited in the Context of SLRM

Note that the straight line of best fit will depend on the dataset but there is only one unique straight line of best fit for the entire population data:



Ordinary Least Squares Estimation for an SLRM

Ordinary Least Squares Estimation for an SLRM

- Suppose Y is the response variable and we are interested in studying its relationship with a single predictor X_1.

Ordinary Least Squares Estimation for an SLRM

- Suppose Y is the response variable and we are interested in studying its relationship with a single predictor X_1.
- The true relationship (*real population model*) is $Y = f(X_1) + \varepsilon$,

Ordinary Least Squares Estimation for an SLRM

- Suppose Y is the response variable and we are interested in studying its relationship with a single predictor X_1.
- The true relationship (*real population model*) is $Y = f(X_1) + \varepsilon$, where f is an unknown and a possibly nonlinear function,

Ordinary Least Squares Estimation for an SLRM

- Suppose Y is the response variable and we are interested in studying its relationship with a single predictor X_1.
- The true relationship (*real population model*) is $Y = f(X_1) + \varepsilon$, where f is an unknown and a possibly nonlinear function, and ϵ is a random error term capturing the effects of noise inherent in the data, predictors that we may have missed, and other effects.

Ordinary Least Squares Estimation for an SLRM

- Suppose Y is the response variable and we are interested in studying its relationship with a single predictor X_1.
- The true relationship (*real population model*) is $Y = f(X_1) + \varepsilon$, where f is an unknown and a possibly nonlinear function, and ϵ is a random error term capturing the effects of noise inherent in the data, predictors that we may have missed, and other effects.
- In SLRM, we model the true population relationship using a linear function:

Ordinary Least Squares Estimation for an SLRM

- Suppose Y is the response variable and we are interested in studying its relationship with a single predictor X_1.
- The true relationship (*real population model*) is $Y = f(X_1) + \varepsilon$, where f is an unknown and a possibly nonlinear function, and ϵ is a random error term capturing the effects of noise inherent in the data, predictors that we may have missed, and other effects.
- In SLRM, we model the true population relationship using a linear function: $Y = \beta_0 + \beta_1 X_1 + \varepsilon$.

Ordinary Least Squares Estimation for an SLRM

- Suppose Y is the response variable and we are interested in studying its relationship with a single predictor X_1.
- The true relationship (*real population model*) is $Y = f(X_1) + \varepsilon$, where f is an unknown and a possibly nonlinear function, and ϵ is a random error term capturing the effects of noise inherent in the data, predictors that we may have missed, and other effects.
- In SLRM, we model the true population relationship using a linear function: $Y = \beta_0 + \beta_1 X_1 + \varepsilon$.
- Note that in the SLRM above, we use the same symbol ε for the random error term which now additionally includes the effect of missing out a possibly nonlinear relationship between Y and X_1.

Ordinary Least Squares Estimation for an SLRM - Continued

Ordinary Least Squares Estimation for an SLRM - Continued

- The SLRM model predicts Y as an approximation $\hat{Y} = \beta_0 + \beta_1 X_1$.

Ordinary Least Squares Estimation for an SLRM - Continued

- The SLRM model predicts Y as an approximation $\hat{Y} = \beta_0 + \beta_1 X_1$.
- The prediction error, also referred to as residual, is $R = Y - \hat{Y} = Y - (\beta_0 + \beta_1 X_1)$.

Ordinary Least Squares Estimation for an SLRM - Continued

- The SLRM model predicts Y as an approximation $\hat{Y} = \beta_0 + \beta_1 X_1$.
- The prediction error, also referred to as residual, is $R = Y - \hat{Y} = Y - (\beta_0 + \beta_1 X_1)$.
- The goal now is to compute *estimates* for the population parameters β_0 and β_1.

Ordinary Least Squares Estimation for an SLRM - Continued

- The SLRM model predicts Y as an approximation $\hat{Y} = \beta_0 + \beta_1 X_1$.
- The prediction error, also referred to as residual, is $R = Y - \hat{Y} = Y - (\beta_0 + \beta_1 X_1)$.
- The goal now is to compute *estimates* for the population parameters β_0 and β_1.
- To that end, we use a randomly sampled dataset comprising n samples with the ith sample's predictor and response values denoted as $x_1^{(i)}$ and $y^{(i)}$, respectively.

Ordinary Least Squares Estimation for an SLRM - Continued

- The SLRM model predicts Y as an approximation $\hat{Y} = \beta_0 + \beta_1 X_1$.
- The prediction error, also referred to as residual, is $R = Y - \hat{Y} = Y - (\beta_0 + \beta_1 X_1)$.
- The goal now is to compute *estimates* for the population parameters β_0 and β_1.
- To that end, we use a randomly sampled dataset comprising n samples with the ith sample's predictor and response values denoted as $x_1^{(i)}$ and $y^{(i)}$, respectively.
- The ordinary least squares estimates of β_0 and β_1 is obtained by minimizing the sum of the squares of the residuals (RSS) for all samples in the dataset:

Ordinary Least Squares Estimation for an SLRM - Continued

- The SLRM model predicts Y as an approximation $\hat{Y} = \beta_0 + \beta_1 X_1$.
- The prediction error, also referred to as residual, is $R = Y - \hat{Y} = Y - (\beta_0 + \beta_1 X_1)$.
- The goal now is to compute *estimates* for the population parameters β_0 and β_1.
- To that end, we use a randomly sampled dataset comprising n samples with the ith sample's predictor and response values denoted as $x_1^{(i)}$ and $y^{(i)}$, respectively.
- The ordinary least squares estimates of β_0 and β_1 is obtained by minimizing the sum of the squares of the residuals (RSS) for all samples in the dataset: $\min \sum_{i=1}^{n} \left(r^{(i)}\right)^2 = \sum_{i=1}^{n} \left(y^{(i)} - \left(\beta_0 + \beta_1 x_1^{(i)}\right)\right)^2.$

Ordinary Least Squares Estimation for an SLRM - Continued

Ordinary Least Squares Estimation for an SLRM - Continued

- We minimize the RSS by calculating its partial derivative w.r.t. β_0 and β_1, and set them equal to zero:

$$\begin{cases} \dfrac{\partial(\text{RSS})}{\partial \beta_0} & = 0 \Rightarrow -2\sum_{i=1}^{n} \left(y^{(i)} - \left(\beta_0 + \beta_1 x_1^{(i)} \right) \right) = 0, \\ \dfrac{\partial(\text{RSS})}{\partial \beta_1} & = 0 \Rightarrow -2\sum_{i=1}^{n} \left(y^{(i)} - \left(\beta_0 + \beta_1 x_1^{(i)} \right) \right) x_1^{(i)} = 0. \end{cases}$$

Ordinary Least Squares Estimation for an SLRM - Continued

- We minimize the RSS by calculating its partial derivative w.r.t. β_0 and β_1, and set them equal to zero:

$$\begin{cases} \dfrac{\partial(\text{RSS})}{\partial \beta_0} & = 0 \Rightarrow -2\sum_{i=1}^{n}\left(y^{(i)} - \left(\beta_0 + \beta_1 x_1^{(i)}\right)\right) = 0, \\ \dfrac{\partial(\text{RSS})}{\partial \beta_1} & = 0 \Rightarrow -2\sum_{i=1}^{n}\left(y^{(i)} - \left(\beta_0 + \beta_1 x_1^{(i)}\right)\right) x_1^{(i)} = 0. \end{cases}$$

- Solving this results in the estimates

$$\begin{aligned} \hat{\beta}_0 &= \bar{y_n} - \hat{\beta}_1 \bar{x_n}, \\ \hat{\beta}_1 &= \frac{s_{xy}}{s_{xx}}, \end{aligned}$$

Ordinary Least Squares Estimation for an SLRM - Continued

where

$$s_{xy} = \underbrace{\sum_{i=1}^{n} \left(x_1^{(i)} - \bar{x}_n\right)\left(y^{(i)} - \bar{y}_n\right)}_{\text{sample covariance-like measure}}$$

$$s_{xx} = \underbrace{\sum_{i=1}^{n} \left(x_1^{(i)} - \bar{x}_n\right)^2}_{\text{sample variance-like measure in the predictor}},$$

$$\bar{x}_n = \underbrace{\frac{1}{n}\sum_{i=1}^{n} x_1^{(i)}}_{\text{sample mean of predictors}} \quad \text{and } \bar{y}_n = \underbrace{\frac{1}{n}\sum_{i=1}^{n} y^{(i)}}_{\text{sample mean of responses}}.$$

Assumptions in SLRM

Assumptions in SLRM

- For a random ith sample, note that in the linear approximation $Y^{(i)} = \beta_0 + \beta_1 X_1^{(i)} + \varepsilon^{(i)}$, the random error term $\varepsilon^{(i)}$ for the ith sample is the same as its residual $R^{(i)}$.

Assumptions in SLRM

- For a random ith sample, note that in the linear approximation $Y^{(i)} = \beta_0 + \beta_1 X_1^{(i)} + \varepsilon^{(i)}$, the random error term $\varepsilon^{(i)}$ for the ith sample is the same as its residual $R^{(i)}$.
- For deriving the ordinary least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, no assumptions about $\varepsilon^{(i)}$ are needed.

Assumptions in SLRM

- For a random ith sample, note that in the linear approximation $Y^{(i)} = \beta_0 + \beta_1 X_1^{(i)} + \varepsilon^{(i)}$, the random error term $\varepsilon^{(i)}$ for the ith sample is the same as its residual $R^{(i)}$.
- For deriving the ordinary least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, no assumptions about $\varepsilon^{(i)}$ are needed.
- For the purpose of deriving statistical inferences (mean, variance etc.) about the least square estimates, we will assume that $\varepsilon^{(i)}$ will have zero mean, constant variance, and uncorrelated across the samples that will be chosen from the population.

Assumptions in SLRM

- For a random ith sample, note that in the linear approximation $Y^{(i)} = \beta_0 + \beta_1 X_1^{(i)} + \varepsilon^{(i)}$, the random error term $\varepsilon^{(i)}$ for the ith sample is the same as its residual $R^{(i)}$.
- For deriving the ordinary least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, no assumptions about $\varepsilon^{(i)}$ are needed.
- For the purpose of deriving statistical inferences (mean, variance etc.) about the least square estimates, we will assume that $\varepsilon^{(i)}$ will have zero mean, constant variance, and uncorrelated across the samples that will be chosen from the population.
- Later, for the purpose of constructing hypotheses tests and confidence intervals for the least squares estimates, we will also assume that $\varepsilon^{(i)}$ is normally distributed.

Estimator and Estimate

Estimator and Estimate

- Suppose that we want to build an SLRM for response Y and a single predictor predictor X_1:

Estimator and Estimate

- Suppose that we want to build an SLRM for response Y and a single predictor predictor X_1: $\hat{Y} = \beta_0 + \beta_1 X_1$.

Estimator and Estimate

- Suppose that we want to build an SLRM for response Y and a single predictor predictor X_1: $\hat{Y} = \beta_0 + \beta_1 X_1$.
- To that end, we want to use a randomly sampled dataset of size n

Estimator and Estimate

- Suppose that we want to build an SLRM for response Y and a single predictor predictor X_1: $\hat{Y} = \beta_0 + \beta_1 X_1$.
- To that end, we want to use a randomly sampled dataset of size n *but the samples are not identified yet.*

Estimator and Estimate

- Suppose that we want to build an SLRM for response Y and a single predictor predictor X_1: $\hat{Y} = \beta_0 + \beta_1 X_1$.
- To that end, we want to use a randomly sampled dataset of size n *but the samples are not identified yet.*
- Recall the OLS estimators of β_0 and β_1:

Estimator and Estimate

- Suppose that we want to build an SLRM for response Y and a single predictor predictor X_1: $\hat{Y} = \beta_0 + \beta_1 X_1$.
- To that end, we want to use a randomly sampled dataset of size n *but the samples are not identified yet.*
- Recall the OLS estimators of β_0 and β_1: $\hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_1 \bar{X}_n$,

Estimator and Estimate

- Suppose that we want to build an SLRM for response Y and a single predictor predictor X_1: $\hat{Y} = \beta_0 + \beta_1 X_1$.
- To that end, we want to use a randomly sampled dataset of size n *but the samples are not identified yet.*
- Recall the OLS estimators of β_0 and β_1: $\hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_1 \bar{X}_n$, and $\hat{\beta}_1 = \dfrac{S_{XY}}{S_{XX}}$.

Estimator and Estimate

- Suppose that we want to build an SLRM for response Y and a single predictor predictor X_1: $\hat{Y} = \beta_0 + \beta_1 X_1$.
- To that end, we want to use a randomly sampled dataset of size n *but the samples are not identified yet.*
- Recall the OLS estimators of β_0 and β_1: $\hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_1 \bar{X}_n$, and $\hat{\beta}_1 = \dfrac{S_{XY}}{S_{XX}}$.
- The OLS estimators are random variables because they depend on the dataset.

Estimator and Estimate

- Suppose that we want to build an SLRM for response Y and a single predictor predictor X_1: $\hat{Y} = \beta_0 + \beta_1 X_1$.
- To that end, we want to use a randomly sampled dataset of size n *but the samples are not identified yet.*
- Recall the OLS estimators of β_0 and β_1: $\hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_1 \bar{X}_n$, and $\hat{\beta}_1 = \dfrac{S_{XY}}{S_{XX}}$.
- The OLS estimators are random variables because they depend on the dataset.
- Suppose we identify, say, $n = 32$ samples;

Estimator and Estimate

- Suppose that we want to build an SLRM for response Y and a single predictor predictor X_1: $\hat{Y} = \beta_0 + \beta_1 X_1$.
- To that end, we want to use a randomly sampled dataset of size n *but the samples are not identified yet.*
- Recall the OLS estimators of β_0 and β_1: $\hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_1 \bar{X}_n$, and $\hat{\beta}_1 = \dfrac{S_{XY}}{S_{XX}}$.
- The OLS estimators are random variables because they depend on the dataset.
- Suppose we identify, say, $n = 32$ samples; then we have that dataset-specific estimates:

Estimator and Estimate

- Suppose that we want to build an SLRM for response Y and a single predictor predictor X_1: $\hat{Y} = \beta_0 + \beta_1 X_1$.
- To that end, we want to use a randomly sampled dataset of size n *but the samples are not identified yet.*
- Recall the OLS estimators of β_0 and β_1: $\hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_1 \bar{X}_n$, and $\hat{\beta}_1 = \dfrac{S_{XY}}{S_{XX}}$.
- The OLS estimators are random variables because they depend on the dataset.
- Suppose we identify, say, $n = 32$ samples; then we have that dataset-specific estimates: $\hat{\beta}_0 = \bar{y}_n - \hat{\beta}_1 \bar{x}_n$,

Estimator and Estimate

- Suppose that we want to build an SLRM for response Y and a single predictor predictor X_1: $\hat{Y} = \beta_0 + \beta_1 X_1$.
- To that end, we want to use a randomly sampled dataset of size n *but the samples are not identified yet.*
- Recall the OLS estimators of β_0 and β_1: $\hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_1 \bar{X}_n$, and $\hat{\beta}_1 = \dfrac{S_{XY}}{S_{XX}}$.
- The OLS estimators are random variables because they depend on the dataset.
- Suppose we identify, say, $n = 32$ samples; then we have that dataset-specific estimates: $\hat{\beta}_0 = \bar{y}_n - \hat{\beta}_1 \bar{x}_n$, and $\hat{\beta}_1 = \dfrac{s_{xy}}{s_{xx}}$.

Interpreting OLS Estimators for a Continuous Predictor

Interpreting OLS Estimators for a Continuous Predictor

- Suppose that we consider the *mtcars* dataset with mpg as the response and hp as the predictor.

Interpreting OLS Estimators for a Continuous Predictor

- Suppose that we consider the *mtcars* dataset with mpg as the response and hp as the predictor.
- SLRM predicts $\widehat{mpg} = \hat{\beta}_0 + \hat{\beta}_1 hp$.

Interpreting OLS Estimators for a Continuous Predictor

- Suppose that we consider the *mtcars* dataset with mpg as the response and hp as the predictor.
- SLRM predicts $\widehat{mpg} = \hat{\beta}_0 + \hat{\beta}_1 hp$.
- $\hat{\beta}_0$ is the predicted mpg when hp is 0.

Interpreting OLS Estimators for a Continuous Predictor

- Suppose that we consider the *mtcars* dataset with mpg as the response and hp as the predictor.
- SLRM predicts $\widehat{mpg} = \hat{\beta}_0 + \hat{\beta}_1 hp$.
- $\hat{\beta}_0$ is the predicted mpg when hp is 0.
- What about $\hat{\beta}_1$?

Interpreting OLS Estimators for a Continuous Predictor

- Suppose that we consider the *mtcars* dataset with mpg as the response and hp as the predictor.
- SLRM predicts $\widehat{mpg} = \hat{\beta}_0 + \hat{\beta}_1 hp$.
- $\hat{\beta}_0$ is the predicted mpg when hp is 0.
- What about $\hat{\beta}_1$? It is the change in the predicted mpg for a 1 unit increase in hp:

Interpreting OLS Estimators for a Continuous Predictor

- Suppose that we consider the *mtcars* dataset with mpg as the response and hp as the predictor.
- SLRM predicts $\widehat{mpg} = \hat{\beta}_0 + \hat{\beta}_1 hp$.
- $\hat{\beta}_0$ is the predicted mpg when hp is 0.
- What about $\hat{\beta}_1$? It is the change in the predicted mpg for a 1 unit increase in hp:

$$\begin{cases} \widehat{mpg}_{\text{old}} & = \hat{\beta}_0 + \hat{\beta}_1 hp \\ \widehat{mpg}_{\text{new}} & = \hat{\beta}_0 + \hat{\beta}_1 (hp + 1) \end{cases} \Rightarrow \widehat{mpg}_{\text{new}} - \widehat{mpg}_{\text{old}} = \hat{\beta}_1.$$

Interpreting OLS Estimators for a Categorical Predictor

Interpreting OLS Estimators for a Categorical Predictor

- Suppose that in the *mtcars* dataset we add a new column called $heavy$ with *yes* or *no* entries indicating whether a car is heavy or not.

Interpreting OLS Estimators for a Categorical Predictor

- Suppose that in the *mtcars* dataset we add a new column called $heavy$ with *yes* or *no* entries indicating whether a car is heavy or not.
- Suppose that mpg is the response and that $heavy$ is the predictor.

Interpreting OLS Estimators for a Categorical Predictor

- Suppose that in the *mtcars* dataset we add a new column called $heavy$ with *yes* or *no* entries indicating whether a car is heavy or not.
- Suppose that mpg is the response and that $heavy$ is the predictor.
- As there are 2 levels (*yes*, *no*) for the categorical variable $heavy$,

Interpreting OLS Estimators for a Categorical Predictor

- Suppose that in the *mtcars* dataset we add a new column called $heavy$ with *yes* or *no* entries indicating whether a car is heavy or not.
- Suppose that mpg is the response and that $heavy$ is the predictor.
- As there are 2 levels (*yes*, *no*) for the categorical variable $heavy$, 1 new *dummy variable* called $heavyyes$ is created.

Interpreting OLS Estimators for a Categorical Predictor

- Suppose that in the *mtcars* dataset we add a new column called $heavy$ with *yes* or *no* entries indicating whether a car is heavy or not.
- Suppose that mpg is the response and that $heavy$ is the predictor.
- As there are 2 levels (*yes*, *no*) for the categorical variable $heavy$, 1 new *dummy variable* called $heavyyes$ is created.
- Note that the level *no* is the reference level per alphabetical order;

Interpreting OLS Estimators for a Categorical Predictor

- Suppose that in the *mtcars* dataset we add a new column called $heavy$ with *yes* or *no* entries indicating whether a car is heavy or not.
- Suppose that mpg is the response and that $heavy$ is the predictor.
- As there are 2 levels (*yes*, *no*) for the categorical variable $heavy$, 1 new *dummy variable* called $heavyyes$ is created.
- Note that the level *no* is the reference level per alphabetical order; that is, $heavyyes$ is equal to 0 if car is not heavy and 1 if it is.

Interpreting OLS Estimators for a Categorical Predictor

- Suppose that in the *mtcars* dataset we add a new column called $heavy$ with *yes* or *no* entries indicating whether a car is heavy or not.
- Suppose that mpg is the response and that $heavy$ is the predictor.
- As there are 2 levels (*yes*, *no*) for the categorical variable $heavy$, 1 new *dummy variable* called $heavyyes$ is created.
- Note that the level *no* is the reference level per alphabetical order; that is, $heavyyes$ is equal to 0 if car is <u>not</u> heavy and 1 if it is.
- The SLRM is $\widehat{mpg} = \hat{\beta}_0 + \hat{\beta}_1 heavyyes$.

Interpreting OLS Estimators for a Categorical Predictor

- Suppose that in the *mtcars* dataset we add a new column called $heavy$ with *yes* or *no* entries indicating whether a car is heavy or not.
- Suppose that mpg is the response and that $heavy$ is the predictor.
- As there are 2 levels (*yes*, *no*) for the categorical variable $heavy$, 1 new *dummy variable* called $heavyyes$ is created.
- Note that the level *no* is the reference level per alphabetical order; that is, $heavyyes$ is equal to 0 if car is <u>not</u> heavy and 1 if it is.
- The SLRM is $\widehat{mpg} = \hat{\beta}_0 + \hat{\beta}_1 heavyyes$.
- $\hat{\beta}_0$ is the average mpg of the not heavy cars (reference level).

Interpreting OLS Estimators for a Categorical Predictor

- Suppose that in the *mtcars* dataset we add a new column called $heavy$ with *yes* or *no* entries indicating whether a car is heavy or not.
- Suppose that mpg is the response and that $heavy$ is the predictor.
- As there are 2 levels (*yes*, *no*) for the categorical variable $heavy$, 1 new *dummy variable* called $heavyyes$ is created.
- Note that the level *no* is the reference level per alphabetical order; that is, $heavyyes$ is equal to 0 if car is <u>not</u> heavy and 1 if it is.
- The SLRM is $\widehat{mpg} = \hat{\beta}_0 + \hat{\beta}_1 heavyyes.$
- $\hat{\beta}_0$ is the average mpg of the not heavy cars (reference level).
- $\hat{\beta}_1$ is the difference between the average mpg of the heavy cars and the average mpg of the not heavy cars (reference level).

Properties of OLS estimators

Properties of OLS estimators

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators of β_0 and β_1, respectively:

Properties of OLS estimators

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators of β_0 and β_1, respectively: $E[\hat{\beta}_0] = \beta_0$ and $E[\hat{\beta}_1] = \beta_1$.

Properties of OLS estimators

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators of β_0 and β_1, respectively: $E[\hat{\beta}_0] = \beta_0$ and $E[\hat{\beta}_1] = \beta_1$.
- In the linear approximation $Y^{(i)} = \beta_0 + \beta_1 X_1^{(i)} + \varepsilon^{(i)}$, recall the assumptions about the random error term $\varepsilon^{(i)}$:

Properties of OLS estimators

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators of β_0 and β_1, respectively: $E[\hat{\beta}_0] = \beta_0$ and $E[\hat{\beta}_1] = \beta_1$.
- In the linear approximation $Y^{(i)} = \beta_0 + \beta_1 X_1^{(i)} + \varepsilon^{(i)}$, recall the assumptions about the random error term $\varepsilon^{(i)}$: zero mean, constant variance = σ^2, and uncorrelated across the samples.

Properties of OLS estimators

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators of β_0 and β_1, respectively: $E[\hat{\beta}_0] = \beta_0$ and $E[\hat{\beta}_1] = \beta_1$.
- In the linear approximation $Y^{(i)} = \beta_0 + \beta_1 X_1^{(i)} + \varepsilon^{(i)}$, recall the assumptions about the random error term $\varepsilon^{(i)}$: zero mean, constant variance = σ^2, and uncorrelated across the samples.
- Variance of $\hat{\beta}_0$ and $\hat{\beta}_1$:

Properties of OLS estimators

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators of β_0 and β_1, respectively: $E[\hat{\beta}_0] = \beta_0$ and $E[\hat{\beta}_1] = \beta_1$.
- In the linear approximation $Y^{(i)} = \beta_0 + \beta_1 X_1^{(i)} + \varepsilon^{(i)}$, recall the assumptions about the random error term $\varepsilon^{(i)}$: zero mean, constant variance = σ^2, and uncorrelated across the samples.
- Variance of $\hat{\beta}_0$ and $\hat{\beta}_1$: $Var[\hat{\beta}_0] = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}_n^2}{s_{xx}}\right)$,

Properties of OLS estimators

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators of β_0 and β_1, respectively: $E[\hat{\beta}_0] = \beta_0$ and $E[\hat{\beta}_1] = \beta_1$.
- In the linear approximation $Y^{(i)} = \beta_0 + \beta_1 X_1^{(i)} + \varepsilon^{(i)}$, recall the assumptions about the random error term $\varepsilon^{(i)}$: zero mean, constant variance = σ^2, and uncorrelated across the samples.
- Variance of $\hat{\beta}_0$ and $\hat{\beta}_1$: $Var[\hat{\beta}_0] = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}_n^2}{s_{xx}} \right)$, and $Var[\hat{\beta}_1] = \frac{\sigma^2}{s_{xx}}$.

Properties of OLS estimators

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators of β_0 and β_1, respectively: $E[\hat{\beta}_0] = \beta_0$ and $E[\hat{\beta}_1] = \beta_1$.
- In the linear approximation $Y^{(i)} = \beta_0 + \beta_1 X_1^{(i)} + \varepsilon^{(i)}$, recall the assumptions about the random error term $\varepsilon^{(i)}$: zero mean, constant variance = σ^2, and uncorrelated across the samples.
- Variance of $\hat{\beta}_0$ and $\hat{\beta}_1$: $Var[\hat{\beta}_0] = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}_n^2}{s_{xx}}\right)$, and $Var[\hat{\beta}_1] = \frac{\sigma^2}{s_{xx}}$.
- σ^2 is typically unknown:

Properties of OLS estimators

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators of β_0 and β_1, respectively: $E[\hat{\beta}_0] = \beta_0$ and $E[\hat{\beta}_1] = \beta_1$.
- In the linear approximation $Y^{(i)} = \beta_0 + \beta_1 X_1^{(i)} + \varepsilon^{(i)}$, recall the assumptions about the random error term $\varepsilon^{(i)}$: zero mean, constant variance = σ^2, and uncorrelated across the samples.
- Variance of $\hat{\beta}_0$ and $\hat{\beta}_1$: $Var[\hat{\beta}_0] = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}_n^2}{s_{xx}} \right)$, and $Var[\hat{\beta}_1] = \frac{\sigma^2}{s_{xx}}$.
- σ^2 is typically unknown: use the in-sample approximation $\sigma^2 \approx \frac{1}{n-2} \sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)} \right)^2$.

Properties of OLS estimators

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators of β_0 and β_1, respectively: $E[\hat{\beta}_0] = \beta_0$ and $E[\hat{\beta}_1] = \beta_1$.
- In the linear approximation $Y^{(i)} = \beta_0 + \beta_1 X_1^{(i)} + \varepsilon^{(i)}$, recall the assumptions about the random error term $\varepsilon^{(i)}$: zero mean, constant variance = σ^2, and uncorrelated across the samples.
- Variance of $\hat{\beta}_0$ and $\hat{\beta}_1$: $Var[\hat{\beta}_0] = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}_n^2}{s_{xx}} \right)$, and $Var[\hat{\beta}_1] = \frac{\sigma^2}{s_{xx}}$.
- σ^2 is typically unknown: use the in-sample approximation $\sigma^2 \approx \frac{1}{n-2} \sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)} \right)^2$.
- Recall that the summation term is the RSS.

Prediction Problem

Prediction Problem

- Let $x_1^{(i)}$ and $y^{(i)}$ represent the ith sample's predictor and response values in a dataset with, say, n samples.

Prediction Problem

- Let $x_1^{(i)}$ and $y^{(i)}$ represent the ith sample's predictor and response values in a dataset with, say, n samples.
- The true population relationship is given by $Y = f(X) + \varepsilon$ for an unknown function f.

Prediction Problem

- Let $x_1^{(i)}$ and $y^{(i)}$ represent the ith sample's predictor and response values in a dataset with, say, n samples.
- The true population relationship is given by $Y = f(X) + \varepsilon$ for an unknown function f.
- The prediction problem is to build an approximation $\hat{f}$ of f using the dataset such that

Prediction Problem

- Let $x_1^{(i)}$ and $y^{(i)}$ represent the ith sample's predictor and response values in a dataset with, say, n samples.
- The true population relationship is given by $Y = f(X) + \varepsilon$ for an unknown function f.
- The prediction problem is to build an approximation $\hat{f}$ of f using the dataset such that for a new predictor value X from the population, the *prediction error* between $\hat{f}(X)$ and the corresponding response value Y is minimized.

Prediction Error

Prediction Error

- How do we measure the prediction error?

Prediction Error

- How do we measure the prediction error?
- For regression, we have, for example,

Prediction Error

- How do we measure the prediction error?
- For regression, we have, for example, squared error $\left(Y - \hat{f}(X)\right)^2$,

Prediction Error

- How do we measure the prediction error?
- For regression, we have, for example, <u>squared error</u> $\left(Y - \hat{f}(X)\right)^2$, and <u>absolute deviation</u> $\left|Y - \hat{f}(X)\right|$.

Prediction Error

- How do we measure the prediction error?
- For regression, we have, for example, squared error $\left(Y - \hat{f}(X)\right)^2$, and absolute deviation $\left|Y - \hat{f}(X)\right|$.
- We will use the squared error as the measure of prediction error.

Prediction Error

- How do we measure the prediction error?
- For regression, we have, for example, squared error $\left(Y - \hat{f}(X)\right)^2$, and absolute deviation $\left|Y - \hat{f}(X)\right|$.
- We will use the squared error as the measure of prediction error.
- How do we build a good "fitted" model?

Prediction Error

- How do we measure the prediction error?
- For regression, we have, for example, squared error $\left(Y - \hat{f}(X)\right)^2$, and absolute deviation $\left|Y - \hat{f}(X)\right|$.
- We will use the squared error as the measure of prediction error.
- How do we build a good “fitted” model? Minimizes the prediction error on *unseen* data.

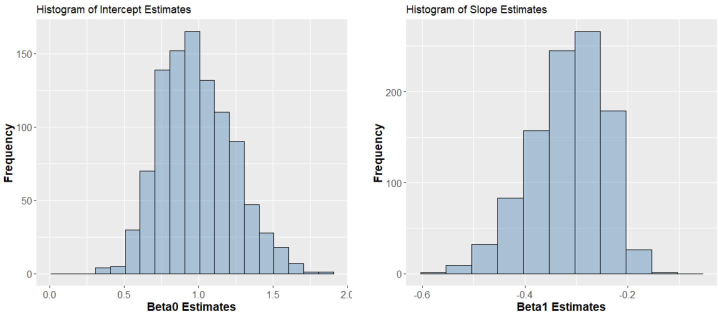
Prediction Error

- How do we measure the prediction error?
- For regression, we have, for example, squared error $\left(Y - \hat{f}(X)\right)^2$, and absolute deviation $\left|Y - \hat{f}(X)\right|$.
- We will use the squared error as the measure of prediction error.
- How do we build a good "fitted" model? Minimizes the prediction error on *unseen* data.
- Train-validation-test split of data.

Accuracy of the Coefficient Estimates: Standard Errors

Accuracy of the Coefficient Estimates: Standard Errors

How can we assess the accuracy of the SLRM coefficient estimates $\hat{\beta}_0 = \bar{y}_n - \hat{\beta}_1 \bar{x}_n$, and $\hat{\beta}_1 = \dfrac{s_{xy}}{s_{xx}}$ derived from a dataset?



Accuracy of the Coefficient Estimates: Standard Errors - Continued

Accuracy of the Coefficient Estimates: Standard Errors - Continued

- Standard deviations of the estimators are also referred to as their standard errors:

Accuracy of the Coefficient Estimates: Standard Errors - Continued

- Standard deviations of the estimators are also referred to as their standard errors: $SE[\hat{\beta}_0] = \sqrt{\sigma^2 \left(\frac{1}{n} + \frac{\bar{x}_n^2}{s_{xx}}\right)}$,

Accuracy of the Coefficient Estimates: Standard Errors - Continued

- Standard deviations of the estimators are also referred to as their standard errors: $SE[\hat{\beta}_0] = \sqrt{\sigma^2 \left(\frac{1}{n} + \frac{\bar{x}_n^2}{s_{xx}} \right)}$, and $SE[\hat{\beta}_1] = \sqrt{\frac{\sigma^2}{s_{xx}}}$ with

Accuracy of the Coefficient Estimates: Standard Errors - Continued

- Standard deviations of the estimators are also referred to as their standard errors: $SE[\hat{\beta}_0] = \sqrt{\sigma^2 \left(\frac{1}{n} + \frac{\bar{x}_n^2}{s_{xx}}\right)}$, and $SE[\hat{\beta}_1] = \sqrt{\frac{\sigma^2}{s_{xx}}}$ with $\sigma^2 \approx \frac{1}{n-2} \underbrace{\sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2}_{RSS}$.

Accuracy of the Coefficient Estimates: Standard Errors - Continued

- Standard deviations of the estimators are also referred to as their standard errors: $SE[\hat{\beta}_0] = \sqrt{\sigma^2 \left(\frac{1}{n} + \frac{\bar{x}_n^2}{s_{xx}}\right)}$, and $SE[\hat{\beta}_1] = \sqrt{\frac{\sigma^2}{s_{xx}}}$ with $\sigma^2 \approx \frac{1}{n-2} \underbrace{\sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2}_{RSS}$.
- Note that the above are *in-sample* estimates;

Accuracy of the Coefficient Estimates: Standard Errors - Continued

- Standard deviations of the estimators are also referred to as their standard errors: $SE[\hat{\beta}_0] = \sqrt{\sigma^2\left(\frac{1}{n} + \frac{\bar{x}_n^2}{s_{xx}}\right)}$, and $SE[\hat{\beta}_1] = \sqrt{\frac{\sigma^2}{s_{xx}}}$ with $\sigma^2 \approx \frac{1}{n-2}\underbrace{\sum_{i=1}^{n}\left(y^{(i)} - \hat{y}^{(i)}\right)^2}_{RSS}$.
- Note that the above are *in-sample* estimates; that is, calculated using the dataset.

Accuracy of the Coefficient Estimates: Standard Errors - Continued

- Standard deviations of the estimators are also referred to as their standard errors: $SE[\hat{\beta}_0] = \sqrt{\sigma^2\left(\frac{1}{n} + \frac{\bar{x}_n^2}{s_{xx}}\right)}$, and $SE[\hat{\beta}_1] = \sqrt{\frac{\sigma^2}{s_{xx}}}$ with $\sigma^2 \approx \frac{1}{n-2}\underbrace{\sum_{i=1}^{n}\left(y^{(i)} - \hat{y}^{(i)}\right)^2}_{RSS}$.
- Note that the above are *in-sample* estimates; that is, calculated using the dataset.
- Random error term $\varepsilon^{(i)}$ is assumed to have

Accuracy of the Coefficient Estimates: Standard Errors - Continued

- Standard deviations of the estimators are also referred to as their standard errors: $SE[\hat{\beta}_0] = \sqrt{\sigma^2\left(\frac{1}{n} + \frac{\bar{x}_n^2}{s_{xx}}\right)}$, and $SE[\hat{\beta}_1] = \sqrt{\frac{\sigma^2}{s_{xx}}}$ with $\sigma^2 \approx \frac{1}{n-2} \underbrace{\sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2}_{RSS}.$
- Note that the above are *in-sample* estimates; that is, calculated using the dataset.
- Random error term $\varepsilon^{(i)}$ is assumed to have zero mean,

Accuracy of the Coefficient Estimates: Standard Errors - Continued

- Standard deviations of the estimators are also referred to as their standard errors: $SE[\hat{\beta}_0] = \sqrt{\sigma^2\left(\frac{1}{n} + \frac{\bar{x}_n^2}{s_{xx}}\right)}$, and $SE[\hat{\beta}_1] = \sqrt{\frac{\sigma^2}{s_{xx}}}$ with $\sigma^2 \approx \frac{1}{n-2} \underbrace{\sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2}_{RSS}$.
- Note that the above are *in-sample* estimates; that is, calculated using the dataset.
- Random error term $\varepsilon^{(i)}$ is assumed to have zero mean, constant variance, and

Accuracy of the Coefficient Estimates: Standard Errors - Continued

- Standard deviations of the estimators are also referred to as their standard errors: $SE[\hat{\beta}_0] = \sqrt{\sigma^2\left(\frac{1}{n} + \frac{\bar{x}_n^2}{s_{xx}}\right)}$, and $SE[\hat{\beta}_1] = \sqrt{\frac{\sigma^2}{s_{xx}}}$ with $\sigma^2 \approx \frac{1}{n-2}\underbrace{\sum_{i=1}^{n}\left(y^{(i)} - \hat{y}^{(i)}\right)^2}_{RSS}$.
- Note that the above are *in-sample* estimates; that is, calculated using the dataset.
- Random error term $\varepsilon^{(i)}$ is assumed to have zero mean, constant variance, and uncorrelated across the samples in the dataset.

Accuracy of the Coefficient Estimates: Confidence Intervals

Accuracy of the Coefficient Estimates: Confidence Intervals

- The standard errors can be extended to calculate confidence intervals (CI) for population coefficients β_0 and β_1.

Accuracy of the Coefficient Estimates: Confidence Intervals

- The standard errors can be extended to calculate confidence intervals (CI) for population coefficients β_0 and β_1.
- A 95% CI for β_0 is

Accuracy of the Coefficient Estimates: Confidence Intervals

- The standard errors can be extended to calculate confidence intervals (CI) for population coefficients β_0 and β_1.
- A 95% CI for β_0 is $\left[\hat{\beta}_0 - 1.96 \times SE(\hat{\beta}_0),\ \hat{\beta}_0 + 1.96 \times SE(\hat{\beta}_0)\right]$.

Accuracy of the Coefficient Estimates: Confidence Intervals

- The standard errors can be extended to calculate confidence intervals (CI) for population coefficients β_0 and β_1.
- A 95% CI for β_0 is $\left[\hat{\beta}_0 - 1.96 \times SE(\hat{\beta}_0),\ \hat{\beta}_0 + 1.96 \times SE(\hat{\beta}_0)\right]$.
- How to interpret this?

Accuracy of the Coefficient Estimates: Confidence Intervals

- The standard errors can be extended to calculate confidence intervals (CI) for population coefficients β_0 and β_1.
- A 95% CI for β_0 is $\left[\hat{\beta}_0 - 1.96 \times SE(\hat{\beta}_0),\ \hat{\beta}_0 + 1.96 \times SE(\hat{\beta}_0)\right]$.
- How to interpret this? Each dataset gives us one such CI;

Accuracy of the Coefficient Estimates: Confidence Intervals

- The standard errors can be extended to calculate confidence intervals (CI) for population coefficients β_0 and β_1.
- A 95% CI for β_0 is $\left[\hat{\beta}_0 - 1.96 \times SE(\hat{\beta}_0),\ \hat{\beta}_0 + 1.96 \times SE(\hat{\beta}_0)\right]$.
- How to interpret this? Each dataset gives us one such CI; approximately 95% of those intervals will contain the true population parameter β_0.

Accuracy of the Coefficient Estimates: Confidence Intervals

- The standard errors can be extended to calculate confidence intervals (CI) for population coefficients β_0 and β_1.
- A 95% CI for β_0 is $\left[\hat{\beta}_0 - 1.96 \times SE(\hat{\beta}_0),\ \hat{\beta}_0 + 1.96 \times SE(\hat{\beta}_0)\right]$.
- How to interpret this? Each dataset gives us one such CI; approximately 95% of those intervals will contain the true population parameter β_0.
- In practice, however, we have one dataset and one CI that comes out of it.

Accuracy of the Coefficient Estimates: Confidence Intervals

- The standard errors can be extended to calculate confidence intervals (CI) for population coefficients β_0 and β_1.
- A 95% CI for β_0 is $\left[\hat{\beta}_0 - 1.96 \times SE(\hat{\beta}_0),\ \hat{\beta}_0 + 1.96 \times SE(\hat{\beta}_0)\right]$.
- How to interpret this? Each dataset gives us one such CI; approximately 95% of those intervals will contain the true population parameter β_0.
- In practice, however, we have one dataset and one CI that comes out of it.
- For calculating CI as above, additional assumption on the random error term $\varepsilon^{(i)}$ that it is *normally distributed* is needed.

Accuracy of the Coefficient Estimates: Hypothesis Tests

Accuracy of the Coefficient Estimates: Hypothesis Tests

- Standard errors can be used to perform hypothesis tests on the population coefficients.

Accuracy of the Coefficient Estimates: Hypothesis Tests

- Standard errors can be used to perform hypothesis tests on the population coefficients.
- Suppose that we build an SLRM using the *mtcars* dataset with mpg as the response and hp as the predictor:

Accuracy of the Coefficient Estimates: Hypothesis Tests

- Standard errors can be used to perform hypothesis tests on the population coefficients.
- Suppose that we build an SLRM using the *mtcars* dataset with mpg as the response and hp as the predictor: $\widehat{mpg} = \hat{\beta}_0 + \hat{\beta}_1 hp$.

Accuracy of the Coefficient Estimates: Hypothesis Tests

- Standard errors can be used to perform hypothesis tests on the population coefficients.
- Suppose that we build an SLRM using the *mtcars* dataset with mpg as the response and hp as the predictor: $\widehat{mpg} = \hat{\beta}_0 + \hat{\beta}_1 hp$.
- We want to check if there is indeed a relationship between mpg and hp;

Accuracy of the Coefficient Estimates: Hypothesis Tests

- Standard errors can be used to perform hypothesis tests on the population coefficients.
- Suppose that we build an SLRM using the *mtcars* dataset with mpg as the response and hp as the predictor: $\widehat{mpg} = \hat{\beta}_0 + \hat{\beta}_1 hp$.
- We want to check if there is indeed a relationship between mpg and hp; how close should $\hat{\beta}_1$ be equal to 0 to conclude that there is no relationship?

Accuracy of the Coefficient Estimates: Hypothesis Tests

- Standard errors can be used to perform hypothesis tests on the population coefficients.
- Suppose that we build an SLRM using the *mtcars* dataset with mpg as the response and hp as the predictor: $\widehat{mpg} = \hat{\beta}_0 + \hat{\beta}_1 hp$.
- We want to check if there is indeed a relationship between mpg and hp; how close should $\hat{\beta}_1$ be equal to 0 to conclude that there is no relationship?
- We consider a *null hypothesis* $\beta_1 = 0$;

Accuracy of the Coefficient Estimates: Hypothesis Tests

- Standard errors can be used to perform hypothesis tests on the population coefficients.
- Suppose that we build an SLRM using the *mtcars* dataset with mpg as the response and hp as the predictor: $\widehat{mpg} = \hat{\beta}_0 + \hat{\beta}_1 hp$.
- We want to check if there is indeed a relationship between mpg and hp; how close should $\hat{\beta}_1$ be equal to 0 to conclude that there is no relationship?
- We consider a *null hypothesis* $\beta_1 = 0$; that is, we hypothesize that there is no relationship between mpg and hp.

Accuracy of the Coefficient Estimates: Hypothesis Tests

- Standard errors can be used to perform hypothesis tests on the population coefficients.
- Suppose that we build an SLRM using the *mtcars* dataset with mpg as the response and hp as the predictor: $\widehat{mpg} = \hat{\beta}_0 + \hat{\beta}_1 hp$.
- We want to check if there is indeed a relationship between mpg and hp; how close should $\hat{\beta}_1$ be equal to 0 to conclude that there is no relationship?
- We consider a *null hypothesis* $\beta_1 = 0$; that is, we hypothesize that there is no relationship between mpg and hp.

- *Assuming that the null hypothesis is true,*

Accuracy of the Coefficient Estimates: Hypothesis Tests

- Standard errors can be used to perform hypothesis tests on the population coefficients.
- Suppose that we build an SLRM using the *mtcars* dataset with mpg as the response and hp as the predictor: $\widehat{mpg} = \hat{\beta}_0 + \hat{\beta}_1 hp$.
- We want to check if there is indeed a relationship between mpg and hp; how close should $\hat{\beta}_1$ be equal to 0 to conclude that there is no relationship?
- We consider a *null hypothesis* $\beta_1 = 0$; that is, we hypothesize that there is no relationship between mpg and hp.
- *Assuming that the null hypothesis is true,* $T = \frac{\hat{\beta}_1 - E[\hat{\beta}_1]}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \underbrace{\beta_1}_{=0}}{SE(\hat{\beta}_1)}$

Accuracy of the Coefficient Estimates: Hypothesis Tests

- Standard errors can be used to perform hypothesis tests on the population coefficients.
- Suppose that we build an SLRM using the *mtcars* dataset with mpg as the response and hp as the predictor: $\widehat{mpg} = \hat{\beta}_0 + \hat{\beta}_1 hp$.
- We want to check if there is indeed a relationship between mpg and hp; how close should $\hat{\beta}_1$ be equal to 0 to conclude that there is no relationship?
- We consider a *null hypothesis* $\beta_1 = 0$; that is, we hypothesize that there is no relationship between mpg and hp.
- *Assuming that the null hypothesis is true,* $T = \frac{\hat{\beta}_1 - E[\hat{\beta}_1]}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \underbrace{\beta_1}_{=0}}{SE(\hat{\beta}_1)}$ follows a t-distribution with $n-2$ degrees of freedom.

Accuracy of the Coefficient Estimates: Hypothesis Tests - Continued

Accuracy of the Coefficient Estimates: Hypothesis Tests - Continued

- From a particular dataset, we have the estimate $\hat{\beta}_1$ and the realized value of T denoted as t.

Accuracy of the Coefficient Estimates: Hypothesis Tests - Continued

- From a particular dataset, we have the estimate $\hat{\beta}_1$ and the realized value of T denoted as t.
- We calculate $P(T \geq |t|)$, which is the probability of observing a realization of T that is more extreme than what we observed from the dataset given that the null hypothesis is true.

Accuracy of the Coefficient Estimates: Hypothesis Tests - Continued

- From a particular dataset, we have the estimate $\hat{\beta}_1$ and the realized value of T denoted as t.
- We calculate $P(T \geq |t|)$, which is the probability of observing a realization of T that is more extreme than what we observed from the dataset given that the null hypothesis is true.
- This probability is referred to as the p-value;

Accuracy of the Coefficient Estimates: Hypothesis Tests - Continued

- From a particular dataset, we have the estimate $\hat{\beta}_1$ and the realized value of T denoted as t.
- We calculate $P(T \geq |t|)$, which is the probability of observing a realization of T that is more extreme than what we observed from the dataset given that the null hypothesis is true.
- This probability is referred to as the p-value; we reject the null hypothesis if the p-value is smaller than a threshold, typically 0.05.

Accuracy of the Coefficient Estimates: Hypothesis Tests - Continued

- From a particular dataset, we have the estimate $\hat{\beta}_1$ and the realized value of T denoted as t.
- We calculate $P(T \geq |t|)$, which is the probability of observing a realization of T that is more extreme than what we observed from the dataset given that the null hypothesis is true.
- This probability is referred to as the p-value; we reject the null hypothesis if the p-value is smaller than a threshold, typically 0.05.
- Rejecting the null hypothesis is equivalent to accepting the alternative hypothesis that $\beta_1 \neq 0$, and therefore mpg and hp are indeed related.

Accuracy of the Coefficient Estimates: Hypothesis Tests - Continued

- From a particular dataset, we have the estimate $\hat{\beta}_1$ and the realized value of T denoted as t.
- We calculate $P(T \geq |t|)$, which is the probability of observing a realization of T that is more extreme than what we observed from the dataset given that the null hypothesis is true.
- This probability is referred to as the p-value; we reject the null hypothesis if the p-value is smaller than a threshold, typically 0.05.
- Rejecting the null hypothesis is equivalent to accepting the alternative hypothesis that $\beta_1 \neq 0$, and therefore mpg and hp are indeed related.
- If the p-value is greater than the threshold, we fail to reject the null hypothesis;

Accuracy of the Coefficient Estimates: Hypothesis Tests - Continued

- From a particular dataset, we have the estimate $\hat{\beta}_1$ and the realized value of T denoted as t.
- We calculate $P(T \geq |t|)$, which is the probability of observing a realization of T that is more extreme than what we observed from the dataset given that the null hypothesis is true.
- This probability is referred to as the p-value; we reject the null hypothesis if the p-value is smaller than a threshold, typically 0.05.
- Rejecting the null hypothesis is equivalent to accepting the alternative hypothesis that $\beta_1 \neq 0$, and therefore mpg and hp are indeed related.
- If the p-value is greater than the threshold, we fail to reject the null hypothesis; this means that there possibly is no relationship between mpg and hp.

Residual Standard Error

Residual Standard Error

- The residual standard error (RSE) is a measure of lack of fit of the model to the data.

Residual Standard Error

- The residual standard error (RSE) is a measure of lack of fit of the model to the data.
- In an SLRM $Y = \beta_0 + \beta_1 X_1 + \varepsilon$, the random error term prevents a perfect prediction even if the population coefficients are exactly known.

Residual Standard Error

- The residual standard error (RSE) is a measure of lack of fit of the model to the data.
- In an SLRM $Y = \beta_0 + \beta_1 X_1 + \varepsilon$, the random error term prevents a perfect prediction even if the population coefficients are exactly known.
- The RSE is an in-sample estimate of the standard deviation σ of the random error term ε:

Residual Standard Error

- The residual standard error (RSE) is a measure of lack of fit of the model to the data.
- In an SLRM $Y = \beta_0 + \beta_1 X_1 + \varepsilon$, the random error term prevents a perfect prediction even if the population coefficients are exactly known.
- The RSE is an in-sample estimate of the standard deviation σ of the random error term ε: RSE $= \sqrt{\frac{1}{n-2} \underbrace{\sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2}_{RSS}}$.

Residual Standard Error

- The residual standard error (RSE) is a measure of lack of fit of the model to the data.
- In an SLRM $Y = \beta_0 + \beta_1 X_1 + \varepsilon$, the random error term prevents a perfect prediction even if the population coefficients are exactly known.
- The RSE is an in-sample estimate of the standard deviation σ of the random error term ε: RSE $= \sqrt{\frac{1}{n-2} \underbrace{\sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2}_{RSS}}.$
- The RSE is the amount by which the response will deviate from the true regression line on an average.

R^2 Statistic

R^2 Statistic

- Suppose we have a dataset with n samples and want a simple model to predict the response Y from a single predictor X_1:

R^2 Statistic

- Suppose we have a dataset with n samples and want a simple model to predict the response Y from a single predictor X_1: the average model $\hat{y}^{(i)} = \bar{y}_n$.

R^2 Statistic

- Suppose we have a dataset with n samples and want a simple model to predict the response Y from a single predictor X_1: the average model $\hat{y}^{(i)} = \bar{y}_n$.
- What is the error associated with this model?

R^2 Statistic

- Suppose we have a dataset with n samples and want a simple model to predict the response Y from a single predictor X_1: the average model $\hat{y}^{(i)} = \bar{y}_n$.
- What is the error associated with this model? Total sum of squares (TSS) = $\sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2 = \sum_{i=1}^{n} \left(y^{(i)} - \bar{y}_n\right)^2$.

R^2 Statistic

- Suppose we have a dataset with n samples and want a simple model to predict the response Y from a single predictor X_1: the average model $\hat{y}^{(i)} = \bar{y}_n$.
- What is the error associated with this model? Total sum of squares (TSS) = $\sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2 = \sum_{i=1}^{n} \left(y^{(i)} - \bar{y}_n\right)^2$.
- Suppose now we build an SLRM;

R^2 Statistic

- Suppose we have a dataset with n samples and want a simple model to predict the response Y from a single predictor X_1: the average model $\hat{y}^{(i)} = \bar{y}_n$.
- What is the error associated with this model? Total sum of squares (TSS) = $\sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2 = \sum_{i=1}^{n} \left(y^{(i)} - \bar{y}_n\right)^2$.
- Suppose now we build an SLRM; the associated error is

R^2 Statistic

- Suppose we have a dataset with n samples and want a simple model to predict the response Y from a single predictor X_1: the average model $\hat{y}^{(i)} = \bar{y}_n$.
- What is the error associated with this model? Total sum of squares (TSS) = $\sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2 = \sum_{i=1}^{n} \left(y^{(i)} - \bar{y}_n\right)^2$.
- Suppose now we build an SLRM; the associated error is residual sum of squares (RSS) = $\sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2 = \sum_{i=1}^{n} \left(y^{(i)} - \left(\hat{\beta}_0 + \hat{\beta}_1 x_1^{(i)}\right)\right)^2$.

R^2 Statistic

- Suppose we have a dataset with n samples and want a simple model to predict the response Y from a single predictor X_1: the average model $\hat{y}^{(i)} = \bar{y}_n$.
- What is the error associated with this model? Total sum of squares (TSS) = $\sum_{i=1}^{n}\left(y^{(i)} - \hat{y}^{(i)}\right)^2 = \sum_{i=1}^{n}\left(y^{(i)} - \bar{y}_n\right)^2$.
- Suppose now we build an SLRM; the associated error is residual sum of squares (RSS) = $\sum_{i=1}^{n}\left(y^{(i)} - \hat{y}^{(i)}\right)^2 = \sum_{i=1}^{n}\left(y^{(i)} - \left(\hat{\beta}_0 + \hat{\beta}_1 x_1^{(i)}\right)\right)^2$.
- The R^2 statistic is defined as

R^2 Statistic

- Suppose we have a dataset with n samples and want a simple model to predict the response Y from a single predictor X_1: the average model $\hat{y}^{(i)} = \bar{y}_n$.
- What is the error associated with this model? Total sum of squares (TSS) = $\sum_{i=1}^{n}\left(y^{(i)} - \hat{y}^{(i)}\right)^2 = \sum_{i=1}^{n}\left(y^{(i)} - \bar{y}_n\right)^2$.
- Suppose now we build an SLRM; the associated error is residual sum of squares (RSS) = $\sum_{i=1}^{n}\left(y^{(i)} - \hat{y}^{(i)}\right)^2 = \sum_{i=1}^{n}\left(y^{(i)} - \left(\hat{\beta}_0 + \hat{\beta}_1 x_1^{(i)}\right)\right)^2$.
- The R^2 statistic is defined as $\frac{\text{TSS}-\text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$.

R^2 Statistic

- Suppose we have a dataset with n samples and want a simple model to predict the response Y from a single predictor X_1: the average model $\hat{y}^{(i)} = \bar{y}_n$.
- What is the error associated with this model? Total sum of squares (TSS) = $\sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2 = \sum_{i=1}^{n} \left(y^{(i)} - \bar{y}_n\right)^2$.
- Suppose now we build an SLRM; the associated error is residual sum of squares (RSS) = $\sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2 = \sum_{i=1}^{n} \left(y^{(i)} - \left(\hat{\beta}_0 + \hat{\beta}_1 x_1^{(i)}\right)\right)^2$.
- The R^2 statistic is defined as $\frac{\text{TSS}-\text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$.
- The R^2 statistic varies between 0 & 1 is and is a measure of the variability in the response Y that the SLRM (built using the predictor X_1) is able to explain.

Centering

Centering

- Suppose we have a dataset with n samples and build an SLRM to predict the response Y from a single predictor X_1:

Centering

- Suppose we have a dataset with n samples and build an SLRM to predict the response Y from a single predictor X_1: $\hat{y}^{(i)} = \hat{\beta}_0 + \hat{\beta}_1 x_1^{(i)}$.

Centering

- Suppose we have a dataset with n samples and build an SLRM to predict the response Y from a single predictor X_1: $\hat{y}^{(i)} = \hat{\beta}_0 + \hat{\beta}_1 x_1^{(i)}$.
- We can make the coefficient estimates more interpretable by centering the predictor values:

Centering

- Suppose we have a dataset with n samples and build an SLRM to predict the response Y from a single predictor X_1: $\hat{y}^{(i)} = \hat{\beta}_0 + \hat{\beta}_1 x_1^{(i)}$.
- We can make the coefficient estimates more interpretable by centering the predictor values: $\tilde{x}_1^{(i)} = x_1^{(i)} - \bar{x}_n$.

Centering

- Suppose we have a dataset with n samples and build an SLRM to predict the response Y from a single predictor X_1: $\hat{y}^{(i)} = \hat{\beta}_0 + \hat{\beta}_1 x_1^{(i)}$.
- We can make the coefficient estimates more interpretable by centering the predictor values: $\tilde{x}_1^{(i)} = x_1^{(i)} - \bar{x}_n$.
- The resulting SLRM is

Centering

- Suppose we have a dataset with n samples and build an SLRM to predict the response Y from a single predictor X_1: $\hat{y}^{(i)} = \hat{\beta}_0 + \hat{\beta}_1 x_1^{(i)}$.
- We can make the coefficient estimates more interpretable by centering the predictor values: $\tilde{x}_1^{(i)} = x_1^{(i)} - \bar{x}_n$.
- The resulting SLRM is $\hat{y}^{(i)} = \hat{\beta}_0 + \hat{\beta}_1 \tilde{x}_1^{(i)}$.

Centering

- Suppose we have a dataset with n samples and build an SLRM to predict the response Y from a single predictor X_1: $\hat{y}^{(i)} = \hat{\beta}_0 + \hat{\beta}_1 x_1^{(i)}$.
- We can make the coefficient estimates more interpretable by centering the predictor values: $\tilde{x}_1^{(i)} = x_1^{(i)} - \bar{x}_n$.
- The resulting SLRM is $\hat{y}^{(i)} = \hat{\beta}_0 + \hat{\beta}_1 \tilde{x}_1^{(i)}$.
- The intercept estimate $\hat{\beta}_0$ can now be interpreted as approximately the average response value for an average predictor input.

Centering

- Suppose we have a dataset with n samples and build an SLRM to predict the response Y from a single predictor X_1: $\hat{y}^{(i)} = \hat{\beta}_0 + \hat{\beta}_1 x_1^{(i)}$.
- We can make the coefficient estimates more interpretable by centering the predictor values: $\tilde{x}_1^{(i)} = x_1^{(i)} - \bar{x}_n$.
- The resulting SLRM is $\hat{y}^{(i)} = \hat{\beta}_0 + \hat{\beta}_1 \tilde{x}_1^{(i)}$.
- The intercept estimate $\hat{\beta}_0$ can now be interpreted as approximately the average response value for an average predictor input.
- The response values can also be centered:

Centering

- Suppose we have a dataset with n samples and build an SLRM to predict the response Y from a single predictor X_1: $\hat{y}^{(i)} = \hat{\beta}_0 + \hat{\beta}_1 x_1^{(i)}$.
- We can make the coefficient estimates more interpretable by centering the predictor values: $\tilde{x}_1^{(i)} = x_1^{(i)} - \bar{x}_n$.
- The resulting SLRM is $\hat{y}^{(i)} = \hat{\beta}_0 + \hat{\beta}_1 \tilde{x}_1^{(i)}$.
- The intercept estimate $\hat{\beta}_0$ can now be interpreted as approximately the average response value for an average predictor input.
- The response values can also be centered: $\tilde{y}^{(i)} = y^{(i)} - \bar{y}_n$.

Centering

- Suppose we have a dataset with n samples and build an SLRM to predict the response Y from a single predictor X_1: $\hat{y}^{(i)} = \hat{\beta}_0 + \hat{\beta}_1 x_1^{(i)}$.
- We can make the coefficient estimates more interpretable by centering the predictor values: $\tilde{x}_1^{(i)} = x_1^{(i)} - \bar{x}_n$.
- The resulting SLRM is $\hat{y}^{(i)} = \hat{\beta}_0 + \hat{\beta}_1 \tilde{x}_1^{(i)}$.
- The intercept estimate $\hat{\beta}_0$ can now be interpreted as approximately the average response value for an average predictor input.
- The response values can also be centered: $\tilde{y}^{(i)} = y^{(i)} - \bar{y}_n$.
- The resulting SLRM will have zero intercept:

Centering

- Suppose we have a dataset with n samples and build an SLRM to predict the response Y from a single predictor X_1: $\hat{y}^{(i)} = \hat{\beta}_0 + \hat{\beta}_1 x_1^{(i)}$.
- We can make the coefficient estimates more interpretable by centering the predictor values: $\tilde{x}_1^{(i)} = x_1^{(i)} - \bar{x}_n$.
- The resulting SLRM is $\hat{y}^{(i)} = \hat{\beta}_0 + \hat{\beta}_1 \tilde{x}_1^{(i)}$.
- The intercept estimate $\hat{\beta}_0$ can now be interpreted as approximately the average response value for an average predictor input.
- The response values can also be centered: $\tilde{y}^{(i)} = y^{(i)} - \bar{y}_n$.
- The resulting SLRM will have zero intercept: $\hat{\tilde{y}}^{(i)} = \hat{\beta}_1 \tilde{x}_1^{(i)}$.

Standardizing

Standardizing

- Sometimes, it is also helpful to standardize the predictor:

Standardizing

- Sometimes, it is also helpful to standardize the predictor: $\tilde{x}_1^{(i)} = \frac{x_1^{(i)} - \bar{x}_n}{\hat{\sigma}_{x_1}}$,

Standardizing

- Sometimes, it is also helpful to standardize the predictor: $\tilde{x}_1^{(i)} = \frac{x_1^{(i)} - \bar{x}_n}{\hat{\sigma}_{x_1}}$, where $\hat{\sigma}_{x_1}$ is the sample standard deviation of the predictor.

Standardizing

- Sometimes, it is also helpful to standardize the predictor: $\tilde{x}_1^{(i)} = \frac{x_1^{(i)} - \bar{x}_n}{\hat{\sigma}_{x_1}}$, where $\hat{\sigma}_{x_1}$ is the sample standard deviation of the predictor.
- This is helpful typically in the multiple linear regression setup where different scales may be present in the data.

Logarithmic Transformation

Logarithmic Transformation

- We may have response variables such as height, weight etc., which cannot be predicted to be negative values using an SLRM.

Logarithmic Transformation

- We may have response variables such as height, weight etc., which cannot be predicted to be negative values using an SLRM.
- One way to overcome that issue is to build an SLRM for logarithmically transformed response values:

Logarithmic Transformation

- We may have response variables such as height, weight etc., which cannot be predicted to be negative values using an SLRM.
- One way to overcome that issue is to build an SLRM for logarithmically transformed response values: $\log\left(\hat{y}^{(i)}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_1^{(i)}$.

Logarithmic Transformation

- We may have response variables such as height, weight etc., which cannot be predicted to be negative values using an SLRM.
- One way to overcome that issue is to build an SLRM for logarithmically transformed response values: $\log\left(\hat{y}^{(i)}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_1^{(i)}$.
- This can be seen as a *multiplicative* model by exponentiating:

Logarithmic Transformation

- We may have response variables such as height, weight etc., which cannot be predicted to be negative values using an SLRM.
- One way to overcome that issue is to build an SLRM for logarithmically transformed response values: $\log\left(\hat{y}^{(i)}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_1^{(i)}$.
- This can be seen as a *multiplicative* model by exponentiating: $\hat{y}^{(i)} = e^{\hat{\beta}_0} \times e^{\hat{\beta}_1 x_1^{(i)}}$.

Logarithmic Transformation

- We may have response variables such as height, weight etc., which cannot be predicted to be negative values using an SLRM.
- One way to overcome that issue is to build an SLRM for logarithmically transformed response values: $\log\left(\hat{y}^{(i)}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_1^{(i)}$.
- This can be seen as a *multiplicative* model by exponentiating: $\hat{y}^{(i)} = e^{\hat{\beta}_0} \times e^{\hat{\beta}_1 x_1^{(i)}}$.
- What is the interpretation of the estimate $\hat{\beta}_1$ now?

Logarithmic Transformation

- We may have response variables such as height, weight etc., which cannot be predicted to be negative values using an SLRM.
- One way to overcome that issue is to build an SLRM for logarithmically transformed response values: $\log\left(\hat{y}^{(i)}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_1^{(i)}$.
- This can be seen as a *multiplicative* model by exponentiating: $\hat{y}^{(i)} = e^{\hat{\beta}_0} \times e^{\hat{\beta}_1 x_1^{(i)}}$.
- What is the interpretation of the estimate $\hat{\beta}_1$ now?
- Suppose there is a 1 unit increase in the predictor value x_1:

Logarithmic Transformation

- We may have response variables such as height, weight etc., which cannot be predicted to be negative values using an SLRM.
- One way to overcome that issue is to build an SLRM for logarithmically transformed response values: $\log\left(\hat{y}^{(i)}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_1^{(i)}$.
- This can be seen as a *multiplicative* model by exponentiating: $\hat{y}^{(i)} = e^{\hat{\beta}_0} \times e^{\hat{\beta}_1 x_1^{(i)}}$.
- What is the interpretation of the estimate $\hat{\beta}_1$ now?
- Suppose there is a 1 unit increase in the predictor value x_1: $\frac{\hat{y}_{\text{new}}}{\hat{y}_{\text{old}}} = e^{\hat{\beta}_1} \approx 1 + \hat{\beta}_1$ for small $\hat{\beta}_1$.

Logarithmic Transformation

- We may have response variables such as height, weight etc., which cannot be predicted to be negative values using an SLRM.
- One way to overcome that issue is to build an SLRM for logarithmically transformed response values: $\log\left(\hat{y}^{(i)}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_1^{(i)}$.
- This can be seen as a *multiplicative* model by exponentiating: $\hat{y}^{(i)} = e^{\hat{\beta}_0} \times e^{\hat{\beta}_1 x_1^{(i)}}$.
- What is the interpretation of the estimate $\hat{\beta}_1$ now?
- Suppose there is a 1 unit increase in the predictor value x_1: $\frac{\hat{y}_{\text{new}}}{\hat{y}_{\text{old}}} = e^{\hat{\beta}_1} \approx 1 + \hat{\beta}_1$ for small $\hat{\beta}_1$.
- This means, $\hat{\beta}_1$ is the proportionate change in the response value for a unit increase in the predictor value.

Matrix Notations for Data: Design Matrix

Matrix Notations for Data: Design Matrix

- Suppose Y is the response variable and we are interested in studying its relationship with multiple predictors $X_1, X_2, \ldots, X_p$.

Matrix Notations for Data: Design Matrix

- Suppose Y is the response variable and we are interested in studying its relationship with multiple predictors $X_1, X_2, \ldots, X_p$.
- Example:

Matrix Notations for Data: Design Matrix

- Suppose Y is the response variable and we are interested in studying its relationship with multiple predictors $X_1, X_2, \ldots, X_p$.
- Example: Let fuel efficiency (mpg) be the response variable and horse power, weight, and transmission type (automatic or manual) be the predictors.

Matrix Notations for Data: Design Matrix

- Suppose Y is the response variable and we are interested in studying its relationship with multiple predictors $X_1, X_2, \ldots, X_p$.
- Example: Let fuel efficiency (mpg) be the response variable and horse power, weight, and transmission type (automatic or manual) be the predictors.
- In a multiple linear regression model (MLRM), the true relationship (*real population model*):

Matrix Notations for Data: Design Matrix

- Suppose Y is the response variable and we are interested in studying its relationship with multiple predictors $X_1, X_2, \ldots, X_p$.
- Example: Let fuel efficiency (mpg) be the response variable and horse power, weight, and transmission type (automatic or manual) be the predictors.
- In a multiple linear regression model (MLRM), the true relationship (*real population model*): $Y = f(X_1, X_2, \ldots, X_p) + \epsilon$,

Matrix Notations for Data: Design Matrix

- Suppose Y is the response variable and we are interested in studying its relationship with multiple predictors $X_1, X_2, \ldots, X_p$.
- Example: Let fuel efficiency (mpg) be the response variable and horse power, weight, and transmission type (automatic or manual) be the predictors.
- In a multiple linear regression model (MLRM), the true relationship (*real population model*): $Y = f(X_1, X_2, \ldots, X_p) + \epsilon$, for an unknown nonlinear function f is modeled using a *linear approximation* of the function f:

Matrix Notations for Data: Design Matrix

- Suppose Y is the response variable and we are interested in studying its relationship with multiple predictors $X_1, X_2, \ldots, X_p$.
- Example: Let fuel efficiency (mpg) be the response variable and horse power, weight, and transmission type (automatic or manual) be the predictors.
- In a multiple linear regression model (MLRM), the true relationship (*real population model*): $Y = f(X_1, X_2, \ldots, X_p) + \epsilon$, for an unknown nonlinear function f is modeled using a *linear approximation* of the function f: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$.

Matrix Notations for Data: Design Matrix

- Suppose Y is the response variable and we are interested in studying its relationship with multiple predictors $X_1, X_2, \ldots, X_p$.
- Example: Let fuel efficiency (mpg) be the response variable and horse power, weight, and transmission type (automatic or manual) be the predictors.
- In a multiple linear regression model (MLRM), the true relationship (*real population model*): $Y = f(X_1, X_2, \ldots, X_p) + \epsilon$, for an unknown nonlinear function f is modeled using a *linear approximation* of the function f: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$.
- To that end, we collect *sample* data from the *population* and use the following notation:

Matrix Notations for Data: Design Matrix

- Suppose Y is the response variable and we are interested in studying its relationship with multiple predictors $X_1, X_2, \ldots, X_p$.
- Example: Let fuel efficiency (mpg) be the response variable and horse power, weight, and transmission type (automatic or manual) be the predictors.
- In a multiple linear regression model (MLRM), the true relationship (*real population model*): $Y = f(X_1, X_2, \ldots, X_p) + \epsilon$, for an unknown nonlinear function f is modeled using a *linear approximation* of the function f: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$.
- To that end, we collect *sample* data from the *population* and use the following notation:

$y^{(i)}$ = ith sample's response value,

Matrix Notations for Data: Design Matrix

- Suppose Y is the response variable and we are interested in studying its relationship with multiple predictors $X_1, X_2, \ldots, X_p$.
- Example: Let fuel efficiency (mpg) be the response variable and horse power, weight, and transmission type (automatic or manual) be the predictors.
- In a multiple linear regression model (MLRM), the true relationship (*real population model*): $Y = f(X_1, X_2, \ldots, X_p) + \epsilon$, for an unknown nonlinear function f is modeled using a *linear approximation* of the function f: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$.
- To that end, we collect *sample* data from the *population* and use the following notation:

$y^{(i)} =$ *i*th sample's response value, $x_j^{(i)} =$ *i*th sample's *j*th predictor value.

Matrix Notations for Data: Design Matrix

Matrix Notations for Data: Design Matrix

- The MLRM model predicts Y as an approximation

Matrix Notations for Data: Design Matrix

- The MLRM model predicts Y as an approximation $\hat{Y} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$.

Matrix Notations for Data: Design Matrix

- The MLRM model predicts Y as an approximation $\hat{Y} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$.
- Residual $R = Y - \hat{Y} = Y - (\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)$.

Matrix Notations for Data: Design Matrix

- The MLRM model predicts Y as an approximation $\hat{Y} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$.
- Residual $R = Y - \hat{Y} = Y - (\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)$.
- After sampling data, we have

Matrix Notations for Data: Design Matrix

- The MLRM model predicts Y as an approximation $\hat{Y} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$.
- Residual $R = Y - \hat{Y} = Y - (\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)$.
- After sampling data, we have

$$\begin{bmatrix} r^{(1)} \\ \vdots \\ r^{(n)} \end{bmatrix}$$

Matrix Notations for Data: Design Matrix

- The MLRM model predicts Y as an approximation $\hat{Y} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$.
- Residual $R = Y - \hat{Y} = Y - (\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)$.
- After sampling data, we have

$$\begin{bmatrix} r^{(1)} \\ \vdots \\ r^{(n)} \end{bmatrix} = \begin{bmatrix} y^{(1)} - \hat{y}^{(1)} \\ \vdots \\ y^{(n)} - \hat{y}^{(n)} \end{bmatrix}$$

Matrix Notations for Data: Design Matrix

- The MLRM model predicts Y as an approximation $\hat{Y} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$.
- Residual $R = Y - \hat{Y} = Y - (\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)$.
- After sampling data, we have

$$\begin{bmatrix} r^{(1)} \\ \vdots \\ r^{(n)} \end{bmatrix} = \begin{bmatrix} y^{(1)} - \hat{y}^{(1)} \\ \vdots \\ y^{(n)} - \hat{y}^{(n)} \end{bmatrix} = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix} - \begin{bmatrix} \beta_0 + \beta_1 x_1^{(1)} + \cdots + \beta_p x_p^{(1)} \\ \vdots \\ \beta_0 + \beta_1 x_1^{(n)} + \cdots + \beta_p x_p^{(n)} \end{bmatrix}$$

Matrix Notations for Data: Design Matrix

- The MLRM model predicts Y as an approximation $\hat{Y} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$.
- Residual $R = Y - \hat{Y} = Y - (\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)$.
- After sampling data, we have

$$\begin{bmatrix} r^{(1)} \\ \vdots \\ r^{(n)} \end{bmatrix} = \begin{bmatrix} y^{(1)} - \hat{y}^{(1)} \\ \vdots \\ y^{(n)} - \hat{y}^{(n)} \end{bmatrix} = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix} - \begin{bmatrix} \beta_0 + \beta_1 x_1^{(1)} + \cdots + \beta_p x_p^{(1)} \\ \vdots \\ \beta_0 + \beta_1 x_1^{(n)} + \cdots + \beta_p x_p^{(n)} \end{bmatrix}$$

$$= \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix} - \begin{bmatrix} 1 & x_1^{(1)} & \ldots & x_p^{(1)} \\ \vdots & \vdots & \ldots & \vdots \\ 1 & x_1^{(n)} & \ldots & x_p^{(n)} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}.$$

Matrix Notations for Data: Design Matrix

Matrix Notations for Data: Design Matrix

$$\underbrace{\begin{bmatrix} r^{(1)} \\ r^{(2)} \\ \vdots \\ r^{(n)} \end{bmatrix}}_{\text{residual vector } \mathbf{r}} = \underbrace{\begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{bmatrix}}_{\text{true response vector } \mathbf{y}} - \underbrace{\begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \dots & x_p^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \dots & x^{(2)} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_1^{(n)} & x_2^{(n)} & \dots & x_p^{(n)} \end{bmatrix}}_{\text{design matrix } \mathbf{X}} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}}_{\text{unknown coefficients vector } \boldsymbol{\beta}}$$

Matrix Notations for Data: Design Matrix

$$\underbrace{\begin{bmatrix} r^{(1)} \\ r^{(2)} \\ \vdots \\ r^{(n)} \end{bmatrix}}_{\text{residual vector } \mathbf{r}} = \underbrace{\begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{bmatrix}}_{\text{true response vector } \mathbf{y}} - \underbrace{\begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \dots & x_p^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \dots & x^{(2)} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_1^{(n)} & x_2^{(n)} & \dots & x_p^{(n)} \end{bmatrix}}_{\text{design matrix } \mathbf{X}} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}}_{\text{unknown coefficients vector } \boldsymbol{\beta}}$$

$$\Rightarrow \mathbf{r} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}.$$

Dealing with Categorical Covariates

Dealing with Categorical Covariates

- Suppose we have a categorical predictor `heating` which can take three possible values:

Dealing with Categorical Covariates

- Suppose we have a categorical predictor `heating` which can take three possible values: (1) `electric` (2) `hot water/steam` (3) `hot air`.

Dealing with Categorical Covariates

- Suppose we have a categorical predictor `heating` which can take three possible values: (1) `electric` (2) `hot water/steam` (3) `hot air`.
- Based on alphabetical order, the categorical level `electric` is chosen as the *reference*.

Dealing with Categorical Covariates

- Suppose we have a categorical predictor `heating` which can take three possible values: (1) `electric` (2) `hot water/steam` (3) `hot air`.
- Based on alphabetical order, the categorical level `electric` is chosen as the *reference*.
- Two new dummy predictors are introduced:

Dealing with Categorical Covariates

- Suppose we have a categorical predictor `heating` which can take three possible values: (1) `electric` (2) `hot water/steam` (3) `hot air`.
- Based on alphabetical order, the categorical level `electric` is chosen as the *reference*.
- Two new dummy predictors are introduced: (1) `heatinghot air`, (2) `heatinghot water/steam`.

Dealing with Categorical Covariates

- Suppose we have a categorical predictor `heating` which can take three possible values: (1) `electric` (2) `hot water/steam` (3) `hot air`.
- Based on alphabetical order, the categorical level `electric` is chosen as the *reference*.
- Two new dummy predictors are introduced: (1) `heatinghot air`, (2) `heatinghot water/steam`.
- The dummy encoding for building the model is as follows:

Dealing with Categorical Covariates

- Suppose we have a categorical predictor heating which can take three possible values: (1) electric (2) hot water/steam (3) hot air.
- Based on alphabetical order, the categorical level electric is chosen as the *reference*.
- Two new dummy predictors are introduced: (1) heatinghot air, (2) heatinghot water/steam.
- The dummy encoding for building the model is as follows:

	heatinghot air	heatinghot water/steam
electric	0	0
hot air	1	0
hot water/steam	0	1

Multiple Linear Regression Models (MLRM) and assumptions

Multiple Linear Regression Models (MLRM) and assumptions

- The random errors for yet to be decided samples $i = 1, 2, \ldots, n$ in the MLRM

$$Y^{(i)} = \beta_0 + \beta_1 X_1^{(i)} + \beta_2 X_2^{(i)} + \cdots + \beta_p X_p^{(i)} + \epsilon^{(i)}$$

Multiple Linear Regression Models (MLRM) and assumptions

- The random errors for yet to be decided samples $i = 1, 2, \ldots, n$ in the MLRM

$$Y^{(i)} = \beta_0 + \beta_1 X_1^{(i)} + \beta_2 X_2^{(i)} + \cdots + \beta_p X_p^{(i)} + \epsilon^{(i)}$$

can be put into a random vector $\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon^{(1)} \\ \vdots \\ \epsilon^{(n)} \end{bmatrix}$.

Multiple Linear Regression Models (MLRM) and assumptions

- The random errors for yet to be decided samples $i = 1, 2, \ldots, n$ in the MLRM

$$Y^{(i)} = \beta_0 + \beta_1 X_1^{(i)} + \beta_2 X_2^{(i)} + \cdots + \beta_p X_p^{(i)} + \epsilon^{(i)}$$

can be put into a <u>random vector</u> $\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon^{(1)} \\ \vdots \\ \epsilon^{(n)} \end{bmatrix}.$

- For drawing statistical inferences about the coefficients estimates, we will assume that:

Multiple Linear Regression Models (MLRM) and assumptions

- The random errors for yet to be decided samples $i = 1, 2, \ldots, n$ in the MLRM
$$Y^{(i)} = \beta_0 + \beta_1 X_1^{(i)} + \beta_2 X_2^{(i)} + \cdots + \beta_p X_p^{(i)} + \epsilon^{(i)}$$
can be put into a random vector $\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon^{(1)} \\ \vdots \\ \epsilon^{(n)} \end{bmatrix}$.

- For drawing statistical inferences about the coefficients estimates, we will assume that:
  1. random error has zero mean:

Multiple Linear Regression Models (MLRM) and assumptions

- The random errors for yet to be decided samples $i = 1, 2, \ldots, n$ in the MLRM
$$Y^{(i)} = \beta_0 + \beta_1 X_1^{(i)} + \beta_2 X_2^{(i)} + \cdots + \beta_p X_p^{(i)} + \epsilon^{(i)}$$
can be put into a random vector $\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon^{(1)} \\ \vdots \\ \epsilon^{(n)} \end{bmatrix}$.
- For drawing statistical inferences about the coefficients estimates, we will assume that:
  1. random error has zero mean: $E[\boldsymbol{\epsilon}] = \mathbf{0}$

Multiple Linear Regression Models (MLRM) and assumptions

- The random errors for yet to be decided samples $i = 1, 2, \ldots, n$ in the MLRM
$$Y^{(i)} = \beta_0 + \beta_1 X_1^{(i)} + \beta_2 X_2^{(i)} + \cdots + \beta_p X_p^{(i)} + \epsilon^{(i)}$$
can be put into a random vector $\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon^{(1)} \\ \vdots \\ \epsilon^{(n)} \end{bmatrix}$.

- For drawing statistical inferences about the coefficients estimates, we will assume that:
  1. random error has zero mean: $E[\boldsymbol{\epsilon}] = \mathbf{0}$
  2. random errors across samples are uncorrelated with constant variance:

Multiple Linear Regression Models (MLRM) and assumptions

- The random errors for yet to be decided samples $i = 1, 2, \ldots, n$ in the MLRM
$$Y^{(i)} = \beta_0 + \beta_1 X_1^{(i)} + \beta_2 X_2^{(i)} + \cdots + \beta_p X_p^{(i)} + \epsilon^{(i)}$$
can be put into a random vector $\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon^{(1)} \\ \vdots \\ \epsilon^{(n)} \end{bmatrix}$.
- For drawing statistical inferences about the coefficients estimates, we will assume that:
  1. random error has zero mean: $E[\boldsymbol{\epsilon}] = \mathbf{0}$
  2. random errors across samples are uncorrelated with constant variance: $\text{Cov}(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}$

Multiple Linear Regression Models (MLRM) and assumptions

- The random errors for yet to be decided samples $i = 1, 2, \ldots, n$ in the MLRM
$$Y^{(i)} = \beta_0 + \beta_1 X_1^{(i)} + \beta_2 X_2^{(i)} + \cdots + \beta_p X_p^{(i)} + \epsilon^{(i)}$$
can be put into a <u>random vector</u> $\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon^{(1)} \\ \vdots \\ \epsilon^{(n)} \end{bmatrix}$.

- For drawing statistical inferences about the coefficients estimates, we will assume that:
  1. random error has zero mean: $E[\boldsymbol{\epsilon}] = \mathbf{0}$
  2. random errors across samples are uncorrelated with constant variance: $\text{Cov}(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}$
  3. the design matrix has full rank:

Multiple Linear Regression Models (MLRM) and assumptions

- The random errors for yet to be decided samples $i = 1, 2, \ldots, n$ in the MLRM

$$Y^{(i)} = \beta_0 + \beta_1 X_1^{(i)} + \beta_2 X_2^{(i)} + \cdots + \beta_p X_p^{(i)} + \epsilon^{(i)}$$

can be put into a random vector $\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon^{(1)} \\ \vdots \\ \epsilon^{(n)} \end{bmatrix}$.

- For drawing statistical inferences about the coefficients estimates, we will assume that:
  1. random error has zero mean: $E[\boldsymbol{\epsilon}] = \mathbf{0}$
  2. random errors across samples are uncorrelated with constant variance: $\text{Cov}(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}$
  3. the design matrix has full rank: $\text{rank}(\mathbf{X}) = p + 1$

Multiple Linear Regression Models (MLRM) and assumptions

- The random errors for yet to be decided samples $i = 1, 2, \ldots, n$ in the MLRM
$$Y^{(i)} = \beta_0 + \beta_1 X_1^{(i)} + \beta_2 X_2^{(i)} + \cdots + \beta_p X_p^{(i)} + \epsilon^{(i)}$$
can be put into a random vector $\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon^{(1)} \\ \vdots \\ \epsilon^{(n)} \end{bmatrix}.$
- For drawing statistical inferences about the coefficients estimates, we will assume that:
  1. random error has zero mean: $E[\boldsymbol{\epsilon}] = \mathbf{0}$
  2. random errors across samples are uncorrelated with constant variance: $\text{Cov}(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}$
  3. the design matrix has full rank: $\text{rank}(\mathbf{X}) = p + 1$
  4. the random error vector is (multivariate) normally distributed:

Multiple Linear Regression Models (MLRM) and assumptions

- The random errors for yet to be decided samples $i = 1, 2, \ldots, n$ in the MLRM
$$Y^{(i)} = \beta_0 + \beta_1 X_1^{(i)} + \beta_2 X_2^{(i)} + \cdots + \beta_p X_p^{(i)} + \epsilon^{(i)}$$
can be put into a random vector $\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon^{(1)} \\ \vdots \\ \epsilon^{(n)} \end{bmatrix}$.
- For drawing statistical inferences about the coefficients estimates, we will assume that:
  1. random error has zero mean: $E[\boldsymbol{\epsilon}] = \mathbf{0}$
  2. random errors across samples are uncorrelated with constant variance: $\text{Cov}(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}$
  3. the design matrix has full rank: $\text{rank}(\mathbf{X}) = p + 1$
  4. the random error vector is (multivariate) normally distributed: $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$.

Ordinary Least Squares (OLS) Solution: Intuition, Geometry, & Algebraic Proof

Ordinary Least Squares (OLS) Solution: Intuition, Geometry, & Algebraic Proof

- To find the coefficient estimates,

Ordinary Least Squares (OLS) Solution: Intuition, Geometry, & Algebraic Proof

- To find the coefficient estimates, just as in SLRM, we minimize the the sum of the squares of the residuals (RSS) for all samples in the dataset:

Ordinary Least Squares (OLS) Solution: Intuition, Geometry, & Algebraic Proof

- To find the coefficient estimates, just as in SLRM, we minimize the the sum of the squares of the residuals (RSS) for all samples in the dataset: $\min \sum_{i=1}^{n} \left(r^{(i)}\right)^2 = \sum_{i=1}^{n} \left(y^{(i)} - \left(\beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)}\right)\right)^2.$

Ordinary Least Squares (OLS) Solution: Intuition, Geometry, & Algebraic Proof

- To find the coefficient estimates, just as in SLRM, we minimize the the sum of the squares of the residuals (RSS) for all samples in the dataset: $\min \sum_{i=1}^{n} \left(r^{(i)}\right)^2 = \sum_{i=1}^{n} \left(y^{(i)} - \left(\beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)}\right)\right)^2.$
- Note that $\sum_{i=1}^{n} \left(r^{(i)}\right)^2 = \underbrace{\left\| \begin{bmatrix} r^{(1)} \\ \vdots \\ r^{(n)} \end{bmatrix} \right\|^2}_{\text{norm of vector squared}} = \|\mathbf{r}\|^2.$

Ordinary Least Squares (OLS) Solution: Intuition, Geometry, & Algebraic Proof

- To find the coefficient estimates, just as in SLRM, we minimize the the sum of the squares of the residuals (RSS) for all samples in the dataset: $\min \sum_{i=1}^{n} \left(r^{(i)}\right)^2 = \sum_{i=1}^{n} \left(y^{(i)} - \left(\beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)}\right)\right)^2.$
- Note that $\sum_{i=1}^{n} \left(r^{(i)}\right)^2 = \underbrace{\left\| \begin{bmatrix} r^{(1)} \\ \vdots \\ r^{(n)} \end{bmatrix} \right\|^2}_{\text{norm of vector squared}} = \|\mathbf{r}\|^2.$
- Using the equation $\mathbf{r} = \mathbf{y} - \mathbf{X}\beta$, minimizing the RSS corresponds to

Ordinary Least Squares (OLS) Solution: Intuition, Geometry, & Algebraic Proof

- To find the coefficient estimates, just as in SLRM, we minimize the the sum of the squares of the residuals (RSS) for all samples in the dataset: $\min \sum_{i=1}^{n} \left(r^{(i)}\right)^2 = \sum_{i=1}^{n} \left(y^{(i)} - \left(\beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)}\right)\right)^2.$

- Note that $\sum_{i=1}^{n} \left(r^{(i)}\right)^2 = \underbrace{\left\| \begin{bmatrix} r^{(1)} \\ \vdots \\ r^{(n)} \end{bmatrix} \right\|^2}_{\text{norm of vector squared}} = \|\mathbf{r}\|^2.$

- Using the equation $\mathbf{r} = \mathbf{y} - \mathbf{X}\beta$, minimizing the RSS corresponds to minimizing $\|\mathbf{r}\|^2$ = minimizing $\|\mathbf{y} - \mathbf{X}\beta\|^2$.

Ordinary Least Squares (OLS) Solution: Intuition, Geometry, & Algebraic Proof

- To find the coefficient estimates, just as in SLRM, we minimize the the sum of the squares of the residuals (RSS) for all samples in the dataset: $\min \sum_{i=1}^{n} \left(r^{(i)}\right)^2 = \sum_{i=1}^{n} \left(y^{(i)} - \left(\beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)}\right)\right)^2.$
- Note that $\sum_{i=1}^{n} \left(r^{(i)}\right)^2 = \underbrace{\left\| \begin{bmatrix} r^{(1)} \\ \vdots \\ r^{(n)} \end{bmatrix} \right\|^2}_{\text{norm of vector squared}} = \|\mathbf{r}\|^2.$
- Using the equation $\mathbf{r} = \mathbf{y} - \mathbf{X}\beta$, minimizing the RSS corresponds to minimizing $\|\mathbf{r}\|^2$ = minimizing $\|\mathbf{y} - \mathbf{X}\beta\|^2$.
- The resulting solution is the OLS solution: $\hat{\beta} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}.$

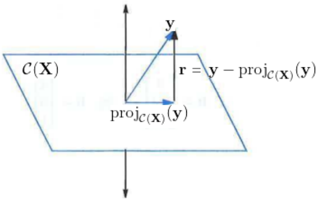
Ordinary Least Squares (OLS) Solution: Intuition, Geometry, & Algebraic Proof

- To find the coefficient estimates, just as in SLRM, we minimize the the sum of the squares of the residuals (RSS) for all samples in the dataset: $\min \sum_{i=1}^{n} \left(r^{(i)}\right)^2 = \sum_{i=1}^{n} \left(y^{(i)} - \left(\beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)}\right)\right)^2$.
- Note that $\sum_{i=1}^{n} \left(r^{(i)}\right)^2 = \underbrace{\left\| \begin{bmatrix} r^{(1)} \\ \vdots \\ r^{(n)} \end{bmatrix} \right\|^2}_{\text{norm of vector squared}} = \|\mathbf{r}\|^2$.
- Using the equation $\mathbf{r} = \mathbf{y} - \mathbf{X}\beta$, minimizing the RSS corresponds to minimizing $\|\mathbf{r}\|^2$ = minimizing $\|\mathbf{y} - \mathbf{X}\beta\|^2$.
- The resulting solution is the OLS solution: $\hat{\beta} = \left(\mathbf{X}^\mathrm{T}\mathbf{X}\right)^{-1} \mathbf{X}^\mathrm{T}\mathbf{y}$.
- Full rank of the design matrix $\mathbf{X}$ ensures the existence of $\left(\mathbf{X}^\mathrm{T}\mathbf{X}\right)^{-1}$.

Ordinary Least Squares (OLS) Solution: Intuition, Geometry, & Algebraic Proof

Ordinary Least Squares (OLS) Solution: Intuition, Geometry, & Algebraic Proof

Minimizing $\|\mathbf{r}\|^2 = \|\mathbf{y} - \mathbf{X}\beta\|^2 = \left\|\mathbf{y} - \left(\underbrace{\beta_0\mathbf{x}_1 + \beta_1\mathbf{x}_2 + \cdots + \beta_p\mathbf{x}_{p+1}}_{\text{linear combination of columns of } \mathbf{X}}\right)\right\|^2$ is equivalent to solving the equation $\mathbf{X}\hat{\beta} = \text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y})$ which represents the *orthogonal projection* of $\mathbf{y}$ on to the column space of the design matrix $\mathcal{C}(\mathbf{X})$ (set of all possible linear combinations of the columns of $\mathbf{X}$):



Ordinary Least Squares (OLS) Solution: Intuition, Geometry, & Algebraic Proof

Ordinary Least Squares (OLS) Solution: Intuition, Geometry, & Algebraic Proof

- Let $\text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{X}\mathbf{z}$

Ordinary Least Squares (OLS) Solution: Intuition, Geometry, & Algebraic Proof

- Let $\text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{X}\mathbf{z} \Rightarrow \mathbf{r} = \mathbf{y} - \text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{y} - \mathbf{X}\mathbf{z}$.

Ordinary Least Squares (OLS) Solution: Intuition, Geometry, & Algebraic Proof

- Let $\text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{X}\mathbf{z} \Rightarrow \mathbf{r} = \mathbf{y} - \text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{y} - \mathbf{X}\mathbf{z}$.
- Residual vector orthogonal to the column space of $\mathbf{X} \Rightarrow \mathbf{r} \perp \mathcal{C}(\mathbf{X})$

Ordinary Least Squares (OLS) Solution: Intuition, Geometry, & Algebraic Proof

- Let $\text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{X}\mathbf{z} \Rightarrow \mathbf{r} = \mathbf{y} - \text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{y} - \mathbf{X}\mathbf{z}$.
- Residual vector orthogonal to the column space of $\mathbf{X} \Rightarrow \mathbf{r} \perp \mathcal{C}(\mathbf{X})$ $\Rightarrow \mathbf{X}^{\mathrm{T}}\mathbf{r} = \mathbf{0}$.

Ordinary Least Squares (OLS) Solution: Intuition, Geometry, & Algebraic Proof

- Let $\text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{X}\mathbf{z} \Rightarrow \mathbf{r} = \mathbf{y} - \text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{y} - \mathbf{X}\mathbf{z}$.
- Residual vector orthogonal to the column space of $\mathbf{X} \Rightarrow \mathbf{r} \perp \mathcal{C}(\mathbf{X})$ $\Rightarrow \mathbf{X}^{\mathrm{T}}\mathbf{r} = \mathbf{0}$.
- This implies $\mathbf{X}^{\mathrm{T}}\left(\mathbf{y} - \mathbf{X}\mathbf{z}\right) = \mathbf{0}$

Ordinary Least Squares (OLS) Solution: Intuition, Geometry, & Algebraic Proof

- Let $\text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{X}\mathbf{z} \Rightarrow \mathbf{r} = \mathbf{y} - \text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{y} - \mathbf{X}\mathbf{z}$.
- Residual vector orthogonal to the column space of $\mathbf{X} \Rightarrow \mathbf{r} \perp \mathcal{C}(\mathbf{X})$ $\Rightarrow \mathbf{X}^{\mathrm{T}}\mathbf{r} = \mathbf{0}$.
- This implies $\mathbf{X}^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\mathbf{z}) = \mathbf{0} \Rightarrow \mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{z} = \mathbf{X}^{\mathrm{T}}\mathbf{y}$

Ordinary Least Squares (OLS) Solution: Intuition, Geometry, & Algebraic Proof

- Let $\text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{X}\mathbf{z} \Rightarrow \mathbf{r} = \mathbf{y} - \text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{y} - \mathbf{X}\mathbf{z}$.
- Residual vector orthogonal to the column space of $\mathbf{X} \Rightarrow \mathbf{r} \perp \mathcal{C}(\mathbf{X})$ $\Rightarrow \mathbf{X}^{\mathrm{T}}\mathbf{r} = \mathbf{0}$.
- This implies $\mathbf{X}^{\mathrm{T}}\left(\mathbf{y} - \mathbf{X}\mathbf{z}\right) = \mathbf{0} \Rightarrow \mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{z} = \mathbf{X}^{\mathrm{T}}\mathbf{y} \Rightarrow \mathbf{z} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$.

Ordinary Least Squares (OLS) Solution: Intuition, Geometry, & Algebraic Proof

- Let $\text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{X}\mathbf{z} \Rightarrow \mathbf{r} = \mathbf{y} - \text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{y} - \mathbf{X}\mathbf{z}$.
- Residual vector orthogonal to the column space of $\mathbf{X} \Rightarrow \mathbf{r} \perp \mathcal{C}(\mathbf{X})$ $\Rightarrow \mathbf{X}^{\mathrm{T}}\mathbf{r} = \mathbf{0}$.
- This implies $\mathbf{X}^{\mathrm{T}}\left(\mathbf{y} - \mathbf{X}\mathbf{z}\right) = \mathbf{0} \Rightarrow \mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{z} = \mathbf{X}^{\mathrm{T}}\mathbf{y} \Rightarrow \mathbf{z} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$.
- This leads to $\text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{X}\mathbf{z} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$.

Ordinary Least Squares (OLS) Solution: Intuition, Geometry, & Algebraic Proof

- Let $\text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{Xz} \Rightarrow \mathbf{r} = \mathbf{y} - \text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{y} - \mathbf{Xz}$.
- Residual vector orthogonal to the column space of $\mathbf{X} \Rightarrow \mathbf{r} \perp \mathcal{C}(\mathbf{X})$ $\Rightarrow \mathbf{X}^{\mathrm{T}}\mathbf{r} = \mathbf{0}$.
- This implies $\mathbf{X}^{\mathrm{T}}\left(\mathbf{y} - \mathbf{Xz}\right) = \mathbf{0} \Rightarrow \mathbf{X}^{\mathrm{T}}\mathbf{Xz} = \mathbf{X}^{\mathrm{T}}\mathbf{y} \Rightarrow \mathbf{z} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$.
- This leads to $\text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{Xz} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$.
- Therefore, the equation $\mathbf{X}\hat{\beta} = \text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y})$ can be written as

Ordinary Least Squares (OLS) Solution: Intuition, Geometry, & Algebraic Proof

- Let $\text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{X}\mathbf{z} \Rightarrow \mathbf{r} = \mathbf{y} - \text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{y} - \mathbf{X}\mathbf{z}$.
- Residual vector orthogonal to the column space of $\mathbf{X} \Rightarrow \mathbf{r} \perp \mathcal{C}(\mathbf{X})$ $\Rightarrow \mathbf{X}^{\mathrm{T}}\mathbf{r} = \mathbf{0}$.
- This implies $\mathbf{X}^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\mathbf{z}) = \mathbf{0} \Rightarrow \mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{z} = \mathbf{X}^{\mathrm{T}}\mathbf{y} \Rightarrow \mathbf{z} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$.
- This leads to $\text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{X}\mathbf{z} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$.
- Therefore, the equation $\mathbf{X}\hat{\beta} = \text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y})$ can be written as
  $\mathbf{X}\hat{\beta} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$

Ordinary Least Squares (OLS) Solution: Intuition, Geometry, & Algebraic Proof

- Let $\text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{X}\mathbf{z} \Rightarrow \mathbf{r} = \mathbf{y} - \text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{y} - \mathbf{X}\mathbf{z}$.
- Residual vector orthogonal to the column space of $\mathbf{X} \Rightarrow \mathbf{r} \perp \mathcal{C}(\mathbf{X})$ $\Rightarrow \mathbf{X}^{\mathrm{T}}\mathbf{r} = \mathbf{0}$.
- This implies $\mathbf{X}^{\mathrm{T}}\left(\mathbf{y} - \mathbf{X}\mathbf{z}\right) = \mathbf{0} \Rightarrow \mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{z} = \mathbf{X}^{\mathrm{T}}\mathbf{y} \Rightarrow \mathbf{z} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$.
- This leads to $\text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{X}\mathbf{z} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$.
- Therefore, the equation $\mathbf{X}\hat{\beta} = \text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y})$ can be written as $\mathbf{X}\hat{\beta} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \Rightarrow \mathbf{X}\left(\hat{\beta} - \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}\right) = \mathbf{0}$.

Ordinary Least Squares (OLS) Solution: Intuition, Geometry, & Algebraic Proof

- Let $\text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{X}\mathbf{z} \Rightarrow \mathbf{r} = \mathbf{y} - \text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{y} - \mathbf{X}\mathbf{z}$.
- Residual vector orthogonal to the column space of $\mathbf{X} \Rightarrow \mathbf{r} \perp \mathcal{C}(\mathbf{X})$ $\Rightarrow \mathbf{X}^{\mathrm{T}}\mathbf{r} = \mathbf{0}$.
- This implies $\mathbf{X}^{\mathrm{T}}\left(\mathbf{y} - \mathbf{X}\mathbf{z}\right) = \mathbf{0} \Rightarrow \mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{z} = \mathbf{X}^{\mathrm{T}}\mathbf{y} \Rightarrow \mathbf{z} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$.
- This leads to $\text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{X}\mathbf{z} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$.
- Therefore, the equation $\mathbf{X}\hat{\beta} = \text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y})$ can be written as $\mathbf{X}\hat{\beta} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \Rightarrow \mathbf{X}\left(\hat{\beta} - \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}\right) = \mathbf{0}$.
- Using the fact that the design matrix $\mathbf{X}$ has full rank (that is, its columns are linearly independent), we arrive at the unique OLS solution $\hat{\beta} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$.

Residual Vector and its Properties

Residual Vector and its Properties

- By construction, the residual vector $\mathbf{r}$ is orthogonal to the columns of the design matrix $\mathbf{X}$.

Residual Vector and its Properties

- By construction, the residual vector $\mathbf{r}$ is orthogonal to the columns of the design matrix $\mathbf{X}$.
- In particular, the residual vector is orthogonal to the first column of $\mathbf{X}$ which is the column full of ones or the ones-vector $\mathbf{1}$.

Residual Vector and its Properties

- By construction, the residual vector $\mathbf{r}$ is orthogonal to the columns of the design matrix $\mathbf{X}$.
- In particular, the residual vector is orthogonal to the first column of $\mathbf{X}$ which is the column full of ones or the ones-vector **1**.
- This implies that $\mathbf{1}^{\mathrm{T}}\mathbf{r} = 0$

Residual Vector and its Properties

- By construction, the residual vector $\mathbf{r}$ is orthogonal to the columns of the design matrix $\mathbf{X}$.
- In particular, the residual vector is orthogonal to the first column of $\mathbf{X}$ which is the column full of ones or the ones-vector **1**.
- This implies that $\mathbf{1}^{\mathrm{T}}\mathbf{r} = 0$ which leads to the fact that sum of the residuals is always equal to 0.

Residual Vector and its Properties

- By construction, the residual vector $\mathbf{r}$ is orthogonal to the columns of the design matrix $\mathbf{X}$.
- In particular, the residual vector is orthogonal to the first column of $\mathbf{X}$ which is the column full of ones or the ones-vector **1**.
- This implies that $\mathbf{1}^{\mathrm{T}}\mathbf{r} = 0$ which leads to the fact that sum of the residuals is always equal to 0.
- This further implies that $\sum_{i=1}^{n} \left[\mathbf{y}^{(i)} - \hat{\mathbf{y}}^{(i)}\right] = 0$

Residual Vector and its Properties

- By construction, the residual vector $\mathbf{r}$ is orthogonal to the columns of the design matrix $\mathbf{X}$.
- In particular, the residual vector is orthogonal to the first column of $\mathbf{X}$ which is the column full of ones or the ones-vector **1**.
- This implies that $\mathbf{1}^{\mathrm{T}}\mathbf{r} = 0$ which leads to the fact that sum of the residuals is always equal to 0.
- This further implies that $\sum_{i=1}^{n} \left[\mathbf{y}^{(i)} - \hat{\mathbf{y}}^{(i)}\right] = 0$ $\Rightarrow \frac{1}{n}\sum_{i=1}^{n} \mathbf{y}^{(i)} = \frac{1}{n}\sum_{i=1}^{n} \hat{\mathbf{y}}^{(i)}$.

Residual Vector and its Properties

- By construction, the residual vector $\mathbf{r}$ is orthogonal to the columns of the design matrix $\mathbf{X}$.
- In particular, the residual vector is orthogonal to the first column of $\mathbf{X}$ which is the column full of ones or the ones-vector **1**.
- This implies that $\mathbf{1}^{\mathrm{T}}\mathbf{r} = 0$ which leads to the fact that sum of the residuals is always equal to 0.
- This further implies that $\sum_{i=1}^{n} \left[\mathbf{y}^{(i)} - \hat{\mathbf{y}}^{(i)}\right] = 0$ $\Rightarrow \frac{1}{n}\sum_{i=1}^{n} \mathbf{y}^{(i)} = \frac{1}{n}\sum_{i=1}^{n} \hat{\mathbf{y}}^{(i)}$.
- This means the true and fitted response values always have the same sample mean.

Residual Vector and its Properties

- By construction, the residual vector $\mathbf{r}$ is orthogonal to the columns of the design matrix $\mathbf{X}$.
- In particular, the residual vector is orthogonal to the first column of $\mathbf{X}$ which is the column full of ones or the ones-vector **1**.
- This implies that $\mathbf{1}^{\mathrm{T}}\mathbf{r} = 0$ which leads to the fact that sum of the residuals is always equal to 0.
- This further implies that $\sum_{i=1}^{n}\left[\mathbf{y}^{(i)} - \hat{\mathbf{y}}^{(i)}\right] = 0$ $\Rightarrow \frac{1}{n}\sum_{i=1}^{n}\mathbf{y}^{(i)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{y}}^{(i)}.$
- This means the true and fitted response values always have the same sample mean.
- This is a reiteration of the fact that linear regression works best on an average.

Interpretation of OLS Estimators

Interpretation of OLS Estimators

- Suppose that we consider the *SaratogaHouses* dataset with $price$ as the response and $livingArea$ and age as the predictors.

Interpretation of OLS Estimators

- Suppose that we consider the *SaratogaHouses* dataset with $price$ as the response and $livingArea$ and age as the predictors.
- Multiple linear regression model (MLRM) predicts
$\widehat{price} = \hat{\beta}_0 + \hat{\beta}_1 livingArea + \hat{\beta}_2 age.$

Interpretation of OLS Estimators

- Suppose that we consider the *SaratogaHouses* dataset with $price$ as the response and $livingArea$ and age as the predictors.
- Multiple linear regression model (MLRM) predicts $\widehat{price} = \hat{\beta}_0 + \hat{\beta}_1 livingArea + \hat{\beta}_2 age.$
- $\hat{\beta}_0$ is the predicted $price$ when both predictors $livingArea$ and age are equal to 0.

Interpretation of OLS Estimators

- Suppose that we consider the *SaratogaHouses* dataset with $price$ as the response and $livingArea$ and age as the predictors.
- Multiple linear regression model (MLRM) predicts $\widehat{price} = \hat{\beta}_0 + \hat{\beta}_1 livingArea + \hat{\beta}_2 age$.
- $\hat{\beta}_0$ is the predicted $price$ when both predictors $livingArea$ and age are equal to 0.
- What about $\hat{\beta}_1$?

Interpretation of OLS Estimators

- Suppose that we consider the *SaratogaHouses* dataset with $price$ as the response and $livingArea$ and age as the predictors.
- Multiple linear regression model (MLRM) predicts $\widehat{price} = \hat{\beta}_0 + \hat{\beta}_1 livingArea + \hat{\beta}_2 age.$
- $\hat{\beta}_0$ is the predicted $price$ when both predictors $livingArea$ and age are equal to 0.
- What about $\hat{\beta}_1$? It is the change in the predicted $price$ for a 1 unit increase in $livingArea$ while keeping the remaining predictor age fixed:

Interpretation of OLS Estimators

- Suppose that we consider the *SaratogaHouses* dataset with $price$ as the response and $livingArea$ and age as the predictors.
- Multiple linear regression model (MLRM) predicts $\widehat{price} = \hat{\beta}_0 + \hat{\beta}_1 livingArea + \hat{\beta}_2 age$.
- $\hat{\beta}_0$ is the predicted $price$ when both predictors $livingArea$ and age are equal to 0.
- What about $\hat{\beta}_1$? It is the change in the predicted $price$ for a 1 unit increase in $livingArea$ while keeping the remaining predictor age fixed:

$$\begin{cases} \widehat{price}_{\text{old}} & = \hat{\beta}_0 + \hat{\beta}_1 livingArea + \hat{\beta}_2 age \\ \widehat{price}_{\text{new}} & = \hat{\beta}_0 + \hat{\beta}_1 \left(livingArea + 1\right) + \hat{\beta}_2 age \end{cases} \Rightarrow \widehat{price}_{\text{new}} - \widehat{price}_{\text{old}} = \hat{\beta}_1.$$

Interpretation of OLS Estimators

Interpretation of OLS Estimators

- Suppose now that we consider $price$ as the response and $livingArea$ and $heating$ as the predictors.

Interpretation of OLS Estimators

- Suppose now that we consider $price$ as the response and $livingArea$ and $heating$ as the predictors.
- Note that the predictor $heating$ is categorical with three levels:

Interpretation of OLS Estimators

- Suppose now that we consider $price$ as the response and $livingArea$ and $heating$ as the predictors.
- Note that the predictor $heating$ is categorical with three levels: (1) `electric` (2) `hot air` (3) `hot water/steam`.

Interpretation of OLS Estimators

- Suppose now that we consider $price$ as the response and $livingArea$ and $heating$ as the predictors.
- Note that the predictor $heating$ is categorical with three levels: (1) `electric` (2) `hot air` (3) `hot water/steam`.
- Multiple linear regression model (MLRM) predicts
  $\widehat{price} = \hat{\beta}_0 + \hat{\beta}_1 livingArea + \hat{\beta}_2 heatinghotair + \hat{\beta}_3 heatinghotwater/steam.$

Interpretation of OLS Estimators

- Suppose now that we consider $price$ as the response and $livingArea$ and $heating$ as the predictors.
- Note that the predictor $heating$ is categorical with three levels: (1) `electric` (2) `hot air` (3) `hot water/steam`.
- Multiple linear regression model (MLRM) predicts $\widehat{price} = \hat{\beta}_0 + \hat{\beta}_1 livingArea + \hat{\beta}_2 heatinghotair + \hat{\beta}_3 heatinghotwater/steam.$
- $\hat{\beta}_0$ and $\hat{\beta}_1$ have the same interpretation as before.

Interpretation of OLS Estimators

- Suppose now that we consider $price$ as the response and $livingArea$ and $heating$ as the predictors.
- Note that the predictor $heating$ is categorical with three levels: (1) `electric` (2) `hot air` (3) `hot water/steam`.
- Multiple linear regression model (MLRM) predicts $\widehat{price} = \hat{\beta}_0 + \hat{\beta}_1 livingArea + \hat{\beta}_2 heatinghotair + \hat{\beta}_3 heatinghotwater/steam.$
- $\hat{\beta}_0$ and $\hat{\beta}_1$ have the same interpretation as before.
- What about $\hat{\beta}_2$?

Interpretation of OLS Estimators

- Suppose now that we consider $price$ as the response and $livingArea$ and $heating$ as the predictors.
- Note that the predictor $heating$ is categorical with three levels: (1) `electric` (2) `hot air` (3) `hot water/steam`.
- Multiple linear regression model (MLRM) predicts $\widehat{price} = \hat{\beta}_0 + \hat{\beta}_1 livingArea + \hat{\beta}_2 heatinghotair + \hat{\beta}_3 heatinghotwater/steam.$
- $\hat{\beta}_0$ and $\hat{\beta}_1$ have the same interpretation as before.
- What about $\hat{\beta}_2$? it is the difference between the predicted price of a hot air-heated house and the predicted price of an electric-heated house (reference level) with the same living area:

Interpretation of OLS Estimators

- Suppose now that we consider $price$ as the response and $livingArea$ and $heating$ as the predictors.
- Note that the predictor $heating$ is categorical with three levels: (1) `electric` (2) `hot air` (3) `hot water/steam`.
- Multiple linear regression model (MLRM) predicts
$\widehat{price} = \hat{\beta}_0 + \hat{\beta}_1 livingArea + \hat{\beta}_2 heatinghotair + \hat{\beta}_3 heatinghotwater/steam.$
- $\hat{\beta}_0$ and $\hat{\beta}_1$ have the same interpretation as before.
- What about $\hat{\beta}_2$? it is the difference between the predicted price of a hot air-heated house and the predicted price of an electric-heated house (reference level) with the same living area:

$$\underbrace{[\hat{\beta}_0 + \hat{\beta}_1 livingArea + \hat{\beta}_2]}_{\widehat{price}_{\text{hot air}}} - \underbrace{[\hat{\beta}_0 + \hat{\beta}_1 livingArea]}_{\widehat{price}_{\text{electric}}} = \hat{\beta}_2.$$

Accuracy of the Coefficient Estimates

Accuracy of the Coefficient Estimates

- It can be shown that the OLS coefficient estimates for an MLRM are unbiased:

Accuracy of the Coefficient Estimates

- It can be shown that the OLS coefficient estimates for an MLRM are unbiased: $E[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}] = \mathbf{0}$.

Accuracy of the Coefficient Estimates

- It can be shown that the OLS coefficient estimates for an MLRM are unbiased: $E[\hat{\beta} - \beta] = \mathbf{0}$.
- The variances of the OLS coefficient estimates are the diagonal elements of the covariance matrix of the random vector $\hat{\beta}$:

Accuracy of the Coefficient Estimates

- It can be shown that the OLS coefficient estimates for an MLRM are unbiased: $E[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}] = \mathbf{0}$.
- The variances of the OLS coefficient estimates are the diagonal elements of the covariance matrix of the random vector $\hat{\boldsymbol{\beta}}$:

$$\mathsf{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2 \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1},$$

Accuracy of the Coefficient Estimates

- It can be shown that the OLS coefficient estimates for an MLRM are unbiased: $E[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}] = \mathbf{0}$.
- The variances of the OLS coefficient estimates are the diagonal elements of the covariance matrix of the random vector $\hat{\boldsymbol{\beta}}$:

$$\text{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2 \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}, \text{ where } \sigma^2 \approx \frac{1}{n-(p+1)} \underbrace{\sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2}_{RSS}.$$

Accuracy of the Coefficient Estimates

- It can be shown that the OLS coefficient estimates for an MLRM are unbiased: $E[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}] = \mathbf{0}$.
- The variances of the OLS coefficient estimates are the diagonal elements of the covariance matrix of the random vector $\hat{\boldsymbol{\beta}}$:

$$\text{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2 \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}, \text{ where } \sigma^2 \approx \frac{1}{n-(p+1)} \underbrace{\sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2}_{RSS}.$$

- The standard deviations of the OLS coefficient estimates,

Accuracy of the Coefficient Estimates

- It can be shown that the OLS coefficient estimates for an MLRM are unbiased: $E[\hat{\beta} - \beta] = \mathbf{0}$.
- The variances of the OLS coefficient estimates are the diagonal elements of the covariance matrix of the random vector $\hat{\beta}$:

$$\text{cov}(\hat{\beta}) = \sigma^2 \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}, \text{ where } \sigma^2 \approx \frac{1}{n-(p+1)} \underbrace{\sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2}_{RSS}.$$

- The standard deviations of the OLS coefficient estimates, also called their standard errors (SE),

Accuracy of the Coefficient Estimates

- It can be shown that the OLS coefficient estimates for an MLRM are unbiased: $E[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}] = \mathbf{0}$.
- The variances of the OLS coefficient estimates are the diagonal elements of the covariance matrix of the random vector $\hat{\boldsymbol{\beta}}$:

$$\text{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2 \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}, \text{ where } \sigma^2 \approx \frac{1}{n-(p+1)} \underbrace{\sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2}_{RSS}.$$

- The standard deviations of the OLS coefficient estimates, also called their standard errors (SE), can be used to calculate confidence intervals (CI) for population coefficient parameters β_j:

Accuracy of the Coefficient Estimates

- It can be shown that the OLS coefficient estimates for an MLRM are unbiased: $E[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}] = \mathbf{0}$.
- The variances of the OLS coefficient estimates are the diagonal elements of the covariance matrix of the random vector $\hat{\boldsymbol{\beta}}$:

$$\text{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2 \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}, \text{ where } \sigma^2 \approx \frac{1}{n-(p+1)} \underbrace{\sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2}_{RSS}.$$

- The standard deviations of the OLS coefficient estimates, also called their standard errors (SE), can be used to calculate confidence intervals (CI) for population coefficient parameters β_j: a 95% CI for β_j is

Accuracy of the Coefficient Estimates

- It can be shown that the OLS coefficient estimates for an MLRM are unbiased: $E[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}] = \mathbf{0}$.
- The variances of the OLS coefficient estimates are the diagonal elements of the covariance matrix of the random vector $\hat{\boldsymbol{\beta}}$:

$$\text{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2 \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}, \text{ where } \sigma^2 \approx \frac{1}{n-(p+1)} \underbrace{\sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2}_{RSS}.$$

- The standard deviations of the OLS coefficient estimates, also called their standard errors (SE), can be used to calculate confidence intervals (CI) for population coefficient parameters β_j: a 95% CI for β_j is $\left[\hat{\beta}_j - 1.96 \times SE(\hat{\beta}_j),\ \hat{\beta}_j + 1.96 \times SE(\hat{\beta}_j)\right]$.

Accuracy of the Coefficient Estimates

Accuracy of the Coefficient Estimates

- In order to check if there is a relationship between the response and a particular predictor, standard errors can be used to perform hypothesis tests on the population coefficient parameters.

Accuracy of the Coefficient Estimates

- In order to check if there is a relationship between the response and a particular predictor, standard errors can be used to perform hypothesis tests on the population coefficient parameters.
- Recall that the p-value associated with the coefficient estimate $\hat{\beta}_j$ is a measure of how likely it is to observe that particular value of the estimate assuming that the null hypothesis about the corresponding population coefficient parameter ($\beta_j = 0$) is true.

Accuracy of the Coefficient Estimates

- In order to check if there is a relationship between the response and a particular predictor, standard errors can be used to perform hypothesis tests on the population coefficient parameters.
- Recall that the p-value associated with the coefficient estimate $\hat{\beta}_j$ is a measure of how likely it is to observe that particular value of the estimate assuming that the null hypothesis about the corresponding population coefficient parameter ($\beta_j = 0$) is true.
- If the p-value is smaller than a threshold, typically 0.05, then

Accuracy of the Coefficient Estimates

- In order to check if there is a relationship between the response and a particular predictor, standard errors can be used to perform hypothesis tests on the population coefficient parameters.
- Recall that the p-value associated with the coefficient estimate $\hat{\beta}_j$ is a measure of how likely it is to observe that particular value of the estimate assuming that the null hypothesis about the corresponding population coefficient parameter ($\beta_j = 0$) is true.
- If the p-value is smaller than a threshold, typically 0.05, then we reject the null hypothesis, and conclude that the jth predictor contributes to the linear model.

Accuracy of the Coefficient Estimates

- In order to check if there is a relationship between the response and a particular predictor, standard errors can be used to perform hypothesis tests on the population coefficient parameters.
- Recall that the p-value associated with the coefficient estimate $\hat{\beta}_j$ is a measure of how likely it is to observe that particular value of the estimate assuming that the null hypothesis about the corresponding population coefficient parameter ($\beta_j = 0$) is true.
- If the p-value is smaller than a threshold, typically 0.05, then we reject the null hypothesis, and conclude that the jth predictor contributes to the linear model.
- A better hypothesis test, when the number of predictors is large, is the F test,

Accuracy of the Coefficient Estimates

- In order to check if there is a relationship between the response and a particular predictor, standard errors can be used to perform hypothesis tests on the population coefficient parameters.
- Recall that the p-value associated with the coefficient estimate $\hat{\beta}_j$ is a measure of how likely it is to observe that particular value of the estimate assuming that the null hypothesis about the corresponding population coefficient parameter ($\beta_j = 0$) is true.
- If the p-value is smaller than a threshold, typically 0.05, then we reject the null hypothesis, and conclude that the jth predictor contributes to the linear model.
- A better hypothesis test, when the number of predictors is large, is the F test, in which the null hypothesis is that all population coefficients are zeros except the intercept.

Accuracy of the Model: R^2 and Adjusted R^2 Statistic

Accuracy of the Model: R^2 and Adjusted R^2 Statistic

- The R^2 statistic varies between 0 and 1, and is a measure of the variability in the response Y that the MLRM (built using the predictors $X_1, X_2, \ldots, X_p$) is able to explain.

Accuracy of the Model: R^2 and Adjusted R^2 Statistic

- The R^2 statistic varies between 0 and 1, and is a measure of the variability in the response Y that the MLRM (built using the predictors $X_1, X_2, \ldots, X_p$) is able to explain.
- However, the R^2 statistic of a model can always be increased by adding more yet insignificant predictors.

Accuracy of the Model: R^2 and Adjusted R^2 Statistic

- The R^2 statistic varies between 0 and 1, and is a measure of the variability in the response Y that the MLRM (built using the predictors $X_1, X_2, \ldots, X_p$) is able to explain.
- However, the R^2 statistic of a model can always be increased by adding more yet insignificant predictors.
- Models with too many predictors over fit the data and typically do not perform well on unseen data.

Accuracy of the Model: R^2 and Adjusted R^2 Statistic

- The R^2 statistic varies between 0 and 1, and is a measure of the variability in the response Y that the MLRM (built using the predictors $X_1, X_2, \ldots, X_p$) is able to explain.
- However, the R^2 statistic of a model can always be increased by adding more yet insignificant predictors.
- Models with too many predictors over fit the data and typically do not perform well on unseen data.
- Adjusted R^2 statistic is a measure which penalizes the addition of predictors. It is the proportion of variance in the response explained by the linear model built using predictors that *actually* affect the response:

Accuracy of the Model: R^2 and Adjusted R^2 Statistic

- The R^2 statistic varies between 0 and 1, and is a measure of the variability in the response Y that the MLRM (built using the predictors $X_1, X_2, \ldots, X_p$) is able to explain.
- However, the R^2 statistic of a model can always be increased by adding more yet insignificant predictors.
- Models with too many predictors over fit the data and typically do not perform well on unseen data.
- Adjusted R^2 statistic is a measure which penalizes the addition of predictors. It is the proportion of variance in the response explained by the linear model built using predictors that *actually* affect the response: $R^2_{\text{adj}} = 1 - \left[\frac{(1-R^2)(n-1)}{n-(p+1)}\right].$

Feature Engineering: Centering, Standardization, and Logarithmic Transformation

Feature Engineering: Centering, Standardization, and Logarithmic Transformation

- Suppose that we consider the *SaratogaHouses* dataset with $price$ as the response and $livingArea$ and age as the predictors.

Feature Engineering: Centering, Standardization, and Logarithmic Transformation

- Suppose that we consider the *SaratogaHouses* dataset with $price$ as the response and $livingArea$ and age as the predictors.
- If we center the predictor values (subtract the sample mean), then the coefficient estimates become mote interpretable.

Feature Engineering: Centering, Standardization, and Logarithmic Transformation

- Suppose that we consider the *SaratogaHouses* dataset with $price$ as the response and $livingArea$ and age as the predictors.
- If we center the predictor values (subtract the sample mean), then the coefficient estimates become mote interpretable.
- With centering of the predictors, the MLRM predicts

Feature Engineering: Centering, Standardization, and Logarithmic Transformation

- Suppose that we consider the *SaratogaHouses* dataset with $price$ as the response and $livingArea$ and age as the predictors.
- If we center the predictor values (subtract the sample mean), then the coefficient estimates become mote interpretable.
- With centering of the predictors, the MLRM predicts

$$\widehat{price} = \hat{\beta}_0 + \hat{\beta}_1 \left(livingArea - \text{mean}(livingArea)\right) + \hat{\beta}_2 \left(age - \text{mean}(age)\right).$$

Feature Engineering: Centering, Standardization, and Logarithmic Transformation

- Suppose that we consider the *SaratogaHouses* dataset with $price$ as the response and $livingArea$ and age as the predictors.
- If we center the predictor values (subtract the sample mean), then the coefficient estimates become mote interpretable.
- With centering of the predictors, the MLRM predicts

$$\widehat{price} = \hat{\beta}_0 + \hat{\beta}_1 \left(livingArea - \text{mean}(livingArea) \right) + \hat{\beta}_2 \left(age - \text{mean}(age) \right).$$

- The intercept estimate can now be interpreted as the (approximate) average value of $price$ around the average value of $livingArea$ and average value of age.

Feature Engineering: Centering, Standardization, and Logarithmic Transformation

Feature Engineering: Centering, Standardization, and Logarithmic Transformation

- In MLRMs, it is helpful to standardize the predictors (subtract the sample mean and divide by sample standard deviation) if there are different scales (units) present in the features.

Feature Engineering: Centering, Standardization, and Logarithmic Transformation

- In MLRMs, it is helpful to standardize the predictors (subtract the sample mean and divide by sample standard deviation) if there are different scales (units) present in the features.
- For example, in the *SaratogaHouses* dataset with $price$ as the response and $livingArea$ and $lotSize$ as the predictors, it would be helpful to standardize the predictors.

Feature Engineering: Centering, Standardization, and Logarithmic Transformation

- In MLRMs, it is helpful to standardize the predictors (subtract the sample mean and divide by sample standard deviation) if there are different scales (units) present in the features.
- For example, in the *SaratogaHouses* dataset with $price$ as the response and $livingArea$ and $lotSize$ as the predictors, it would be helpful to standardize the predictors.
- The interpretations of the coefficient estimates have to be made with respect to the scaled predictor values.

Feature Engineering: Centering, Standardization, and Logarithmic Transformation

- In MLRMs, it is helpful to standardize the predictors (subtract the sample mean and divide by sample standard deviation) if there are different scales (units) present in the features.
- For example, in the *SaratogaHouses* dataset with $price$ as the response and $livingArea$ and $lotSize$ as the predictors, it would be helpful to standardize the predictors.
- The interpretations of the coefficient estimates have to be made with respect to the scaled predictor values.
- The response variable may be logarithmically transformed if it cannot be predicted to be negative (for example, height, weight etc.).

Regularization: Ridge and Lasso

Regularization: Ridge and Lasso

- How can we choose predictors that *explain* the response variable the most?

Regularization: Ridge and Lasso

- How can we choose predictors that *explain* the response variable the most?
- We achieve this by adding a regularization term to the residual sum of squares (RSS) and minimize it:

Regularization: Ridge and Lasso

- How can we choose predictors that *explain* the response variable the most?
- We achieve this by adding a regularization term to the residual sum of squares (RSS) and minimize it:

$$\underbrace{\sum_{i=1}^{n}\left(y^{(i)} - \left(\beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)}\right)\right)^2}_{\text{RSS}} +$$

where $\lambda > 0$ is the strength of regularization.

Regularization: Ridge and Lasso

- How can we choose predictors that *explain* the response variable the most?
- We achieve this by adding a regularization term to the residual sum of squares (RSS) and minimize it:

$$\underbrace{\sum_{i=1}^{n}\left(y^{(i)} - \left(\beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)}\right)\right)^2}_{\text{RSS}} + \begin{cases} \lambda \sum_{j=1}^{p} |\beta_j|^2 & \text{(ridge)} \\ \quad \text{or} & \\ \lambda \sum_{j=1}^{p} |\beta_j| & \text{(lasso)}, \end{cases}$$

where $\lambda > 0$ is the strength of regularization.

Regularization: Ridge and Lasso

- How can we choose predictors that *explain* the response variable the most?
- We achieve this by adding a regularization term to the residual sum of squares (RSS) and minimize it:

$$\underbrace{\sum_{i=1}^{n}\left(y^{(i)} - \left(\beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)}\right)\right)^2}_{\text{RSS}} + \begin{cases} \lambda \sum_{j=1}^{p} |\beta_j|^2 & \text{(ridge)} \\ \quad\text{or} & \\ \lambda \sum_{j=1}^{p} |\beta_j| & \text{(lasso)}, \end{cases}$$

  where $\lambda > 0$ is the strength of regularization.
- Both ridge and lasso approaches for regularization shrink the coefficient estimates towards 0 but lasso typically yields a much smaller subset of nonzero coefficient estimates.

Confounding and Collinearity: Basic Ideas

Confounding and Collinearity: Basic Ideas

- Confounding:

Confounding and Collinearity: Basic Ideas

- Confounding: occurs when a third variable that distorts the observed relationship between the predictor and response.

Confounding and Collinearity: Basic Ideas

- Confounding: occurs when a third variable that distorts the observed relationship between the predictor and response.
- Example:

Confounding and Collinearity: Basic Ideas

- Confounding: occurs when a third variable that distorts the observed relationship between the predictor and response.
- Example: Japan's death rate is higher than that of countries like Vietnam;

Confounding and Collinearity: Basic Ideas

- Confounding: occurs when a third variable that distorts the observed relationship between the predictor and response.
- Example: Japan's death rate is higher than that of countries like Vietnam; before concluding that Japan is a riskier place to live, confounding factors such as *age* need to be accounted for;

Confounding and Collinearity: Basic Ideas

- Confounding: occurs when a third variable that distorts the observed relationship between the predictor and response.
- Example: Japan's death rate is higher than that of countries like Vietnam; before concluding that Japan is a riskier place to live, confounding factors such as *age* need to be accounted for; median age of Japan is much higher than that of, say, Vietnam.

Confounding and Collinearity: Basic Ideas

- Confounding: occurs when a third variable that distorts the observed relationship between the predictor and response.
- Example: Japan's death rate is higher than that of countries like Vietnam; before concluding that Japan is a riskier place to live, confounding factors such as *age* need to be accounted for; median age of Japan is much higher than that of, say, Vietnam.
- Collinearity:

Confounding and Collinearity: Basic Ideas

- Confounding: occurs when a third variable that distorts the observed relationship between the predictor and response.
- Example: Japan's death rate is higher than that of countries like Vietnam; before concluding that Japan is a riskier place to live, confounding factors such as *age* need to be accounted for; median age of Japan is much higher than that of, say, Vietnam.
- Collinearity: occurs when predictors are highly correlated such that it is difficult to distinguish their effect on the response.

Confounding and Collinearity: Basic Ideas

- Confounding: occurs when a third variable that distorts the observed relationship between the predictor and response.
- Example: Japan's death rate is higher than that of countries like Vietnam; before concluding that Japan is a riskier place to live, confounding factors such as *age* need to be accounted for; median age of Japan is much higher than that of, say, Vietnam.
- Collinearity: occurs when predictors are highly correlated such that it is difficult to distinguish their effect on the response.
- Also referred to as multicollinearity or ill-conditioning.

Confounding and Collinearity: Basic Ideas

- Confounding: occurs when a third variable that distorts the observed relationship between the predictor and response.
- Example: Japan's death rate is higher than that of countries like Vietnam; before concluding that Japan is a riskier place to live, confounding factors such as *age* need to be accounted for; median age of Japan is much higher than that of, say, Vietnam.
- Collinearity: occurs when predictors are highly correlated such that it is difficult to distinguish their effect on the response.
- Also referred to as multicollinearity or ill-conditioning.
- Structural collinearity:

Confounding and Collinearity: Basic Ideas

- Confounding: occurs when a third variable that distorts the observed relationship between the predictor and response.
- Example: Japan's death rate is higher than that of countries like Vietnam; before concluding that Japan is a riskier place to live, confounding factors such as *age* need to be accounted for; median age of Japan is much higher than that of, say, Vietnam.
- Collinearity: occurs when predictors are highly correlated such that it is difficult to distinguish their effect on the response.
- Also referred to as multicollinearity or ill-conditioning.
- Structural collinearity: when model is created with correlated predictors.

Confounding and Collinearity: Basic Ideas

- Confounding: occurs when a third variable that distorts the observed relationship between the predictor and response.
- Example: Japan's death rate is higher than that of countries like Vietnam; before concluding that Japan is a riskier place to live, confounding factors such as *age* need to be accounted for; median age of Japan is much higher than that of, say, Vietnam.
- Collinearity: occurs when predictors are highly correlated such that it is difficult to distinguish their effect on the response.
- Also referred to as multicollinearity or ill-conditioning.
- Structural collinearity: when model is created with correlated predictors.
- Data collinearity:

Confounding and Collinearity: Basic Ideas

- Confounding: occurs when a third variable that distorts the observed relationship between the predictor and response.
- Example: Japan's death rate is higher than that of countries like Vietnam; before concluding that Japan is a riskier place to live, confounding factors such as *age* need to be accounted for; median age of Japan is much higher than that of, say, Vietnam.
- Collinearity: occurs when predictors are highly correlated such that it is difficult to distinguish their effect on the response.
- Also referred to as multicollinearity or ill-conditioning.
- Structural collinearity: when model is created with correlated predictors.
- Data collinearity: when data comprises correlated predictors.

When Does Confounding Arise?

When Does Confounding Arise?

- Indication bias:

When Does Confounding Arise?

- Indication bias: the effect of a trial drug for treating a particular medical condition may differ substantially between those who have the condition and those who do not.

When Does Confounding Arise?

- Indication bias: the effect of a trial drug for treating a particular medical condition may differ substantially between those who have the condition and those who do not.
- Selection bias:

When Does Confounding Arise?

- Indication bias: the effect of a trial drug for treating a particular medical condition may differ substantially between those who have the condition and those who do not.
- Selection bias: the effect of a trial drug for treating a particular medical condition may be affected by the imbalance between the groups.

When Does Confounding Arise?

- Indication bias: the effect of a trial drug for treating a particular medical condition may differ substantially between those who have the condition and those who do not.
- Selection bias: the effect of a trial drug for treating a particular medical condition may be affected by the imbalance between the groups.
- Recall bias:

When Does Confounding Arise?

- Indication bias: the effect of a trial drug for treating a particular medical condition may differ substantially between those who have the condition and those who do not.
- Selection bias: the effect of a trial drug for treating a particular medical condition may be affected by the imbalance between the groups.
- Recall bias: study participants who have cancer may be more likely to recall being a smoker.

Confounding Example

Which of the following are likely to be confounding factors for the hypothesis that high cholesterol food is associated with heart disease?

Factor	Low cholesterol food	High cholesterol food

Confounding Example

Which of the following are likely to be confounding factors for the hypothesis that high cholesterol food is associated with heart disease?

Factor	Low cholesterol food	High cholesterol food
Smoker (%)	10%	30%

Confounding Example

Which of the following are likely to be confounding factors for the hypothesis that high cholesterol food is associated with heart disease?

Factor	Low cholesterol food	High cholesterol food
Smoker (%)	10%	30%
Age (mean years)	42	41

Confounding Example

Which of the following are likely to be confounding factors for the hypothesis that high cholesterol food is associated with heart disease?

Factor	Low cholesterol food	High cholesterol food
Smoker (%)	10%	30%
Age (mean years)	42	41
Daily exercise (%)	25%	28%

Confounding Example

Which of the following are likely to be confounding factors for the hypothesis that high cholesterol food is associated with heart disease?

Factor	Low cholesterol food	High cholesterol food
Smoker (%)	10%	30%
Age (mean years)	42	41
Daily exercise (%)	25%	28%
Diabetes (%)	12%	32%

Confounding Example

Which of the following are likely to be confounding factors for the hypothesis that high cholesterol food is associated with heart disease?

Factor	Low cholesterol food	High cholesterol food
Smoker (%)	10%	30%
Age (mean years)	42	41
Daily exercise (%)	25%	28%
Diabetes (%)	12%	32%
BMI (mean)	24	26

Confounding Example

Which of the following are likely to be confounding factors for the hypothesis that high cholesterol food is associated with heart disease?

Factor	Low cholesterol food	High cholesterol food
Smoker (%)	10%	30%
Age (mean years)	42	41
Daily exercise (%)	25%	28%
Diabetes (%)	12%	32%
BMI (mean)	24	26

Predictors that are *imbalanced* among the two groups:

Confounding Example

Which of the following are likely to be confounding factors for the hypothesis that high cholesterol food is associated with heart disease?

Factor	**Low cholesterol food**	**High cholesterol food**
Smoker (%)	10%	30%
Age (mean years)	42	41
Daily exercise (%)	25%	28%
Diabetes (%)	12%	32%
BMI (mean)	24	26

Predictors that are *imbalanced* among the two groups: Smoker, Diabetes are potential confounders.

Collinearity

Collinearity

- Collinearity, an extreme case of confounding, implies an exact linear relationship between predictors.

Collinearity

- Collinearity, an extreme case of confounding, implies an exact linear relationship between predictors.
- Example:

Collinearity

- Collinearity, an extreme case of confounding, implies an exact linear relationship between predictors.
- Example: predictor $age1$ in years is collinear with the predictor $age2$ in months because $age1 = 12 \times age2 \Rightarrow$

Collinearity

- Collinearity, an extreme case of confounding, implies an exact linear relationship between predictors.
- Example: predictor $age1$ in years is collinear with the predictor $age2$ in months because $age1 = 12 \times age2 \Rightarrow$ columns of design matrix $\mathbf{X} = \begin{bmatrix} age1 & age2 \end{bmatrix}$ are linearly dependent

Collinearity

- Collinearity, an extreme case of confounding, implies an exact linear relationship between predictors.
- Example: predictor $age1$ in years is collinear with the predictor $age2$ in months because $age1 = 12 \times age2 \Rightarrow$ columns of design matrix $\mathbf{X} = \begin{bmatrix} age1 & age2 \end{bmatrix}$ are linearly dependent $\Rightarrow (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ does not exist.

Collinearity

- Collinearity, an extreme case of confounding, implies an exact linear relationship between predictors.
- Example: predictor $age1$ in years is collinear with the predictor $age2$ in months because $age1 = 12 \times age2 \Rightarrow$ columns of design matrix $\mathbf{X} = \begin{bmatrix} age1 & age2 \end{bmatrix}$ are linearly dependent $\Rightarrow (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ does not exist.
- Collinearity poses no problem for prediction:

Collinearity

- Collinearity, an extreme case of confounding, implies an exact linear relationship between predictors.
- Example: predictor $age1$ in years is collinear with the predictor $age2$ in months because $age1 = 12 \times age2 \Rightarrow$ columns of design matrix $\mathbf{X} = \begin{bmatrix} age1 & age2 \end{bmatrix}$ are linearly dependent $\Rightarrow (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ does not exist.
- Collinearity poses no problem for prediction: the model $\widehat{height} = \hat{\beta}_0 + \hat{\beta}_1 \times age1 + \hat{\beta}_2 \times age2$

Collinearity

- Collinearity, an extreme case of confounding, implies an exact linear relationship between predictors.
- Example: predictor $age1$ in years is collinear with the predictor $age2$ in months because $age1 = 12 \times age2 \Rightarrow$ columns of design matrix $\mathbf{X} = \begin{bmatrix} age1 & age2 \end{bmatrix}$ are linearly dependent $\Rightarrow (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ does not exist.
- Collinearity poses no problem for prediction: the model $\widehat{height} = \hat{\beta}_0 + \hat{\beta}_1 \times age1 + \hat{\beta}_2 \times age2$ has theoretically infinitely many solutions but all result in the same predicted height.

Collinearity

- Collinearity, an extreme case of confounding, implies an exact linear relationship between predictors.
- Example: predictor $age1$ in years is collinear with the predictor $age2$ in months because $age1 = 12 \times age2 \Rightarrow$ columns of design matrix $\mathbf{X} = \begin{bmatrix} age1 & age2 \end{bmatrix}$ are linearly dependent $\Rightarrow (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ does not exist.
- Collinearity poses no problem for prediction: the model $\widehat{height} = \hat{\beta}_0 + \hat{\beta}_1 \times age1 + \hat{\beta}_2 \times age2$ has theoretically infinitely many solutions but all result in the same predicted height.
- For example, the following solutions are all equivalent:

Collinearity

- Collinearity, an extreme case of confounding, implies an exact linear relationship between predictors.
- Example: predictor $age1$ in years is collinear with the predictor $age2$ in months because $age1 = 12 \times age2 \Rightarrow$ columns of design matrix $\mathbf{X} = \begin{bmatrix} age1 & age2 \end{bmatrix}$ are linearly dependent $\Rightarrow (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ does not exist.
- Collinearity poses no problem for prediction: the model $\widehat{height} = \hat{\beta}_0 + \hat{\beta}_1 \times age1 + \hat{\beta}_2 \times age2$ has theoretically infinitely many solutions but all result in the same predicted height.
- For example, the following solutions are all equivalent: $\widehat{height} = 30 + 3 \times age1 + 0 \times age2 =$

Collinearity

- Collinearity, an extreme case of confounding, implies an exact linear relationship between predictors.
- Example: predictor $age1$ in years is collinear with the predictor $age2$ in months because $age1 = 12 \times age2 \Rightarrow$ columns of design matrix $\mathbf{X} = \begin{bmatrix} age1 & age2 \end{bmatrix}$ are linearly dependent $\Rightarrow (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ does not exist.
- Collinearity poses no problem for prediction: the model $\widehat{height} = \hat{\beta}_0 + \hat{\beta}_1 \times age1 + \hat{\beta}_2 \times age2$ has theoretically infinitely many solutions but all result in the same predicted height.
- For example, the following solutions are all equivalent: $\widehat{height} = 30 + 3 \times age1 + 0 \times age2 = 30 + 2 \times age1 + 12 \times age2 =$

Collinearity

- Collinearity, an extreme case of confounding, implies an exact linear relationship between predictors.
- Example: predictor $age1$ in years is collinear with the predictor $age2$ in months because $age1 = 12 \times age2 \Rightarrow$ columns of design matrix $\mathbf{X} = \begin{bmatrix} age1 & age2 \end{bmatrix}$ are linearly dependent $\Rightarrow (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ does not exist.
- Collinearity poses no problem for prediction: the model $\widehat{height} = \hat{\beta}_0 + \hat{\beta}_1 \times age1 + \hat{\beta}_2 \times age2$ has theoretically infinitely many solutions but all result in the same predicted height.
- For example, the following solutions are all equivalent: $\widehat{height} = 30 + 3 \times age1 + 0 \times age2 = 30 + 2 \times age1 + 12 \times age2 = 30 + 1 \times age1 + 24 \times age2.$

Collinearity

- Collinearity, an extreme case of confounding, implies an exact linear relationship between predictors.
- Example: predictor $age1$ in years is collinear with the predictor $age2$ in months because $age1 = 12 \times age2 \Rightarrow$ columns of design matrix $\mathbf{X} = \begin{bmatrix} age1 & age2 \end{bmatrix}$ are linearly dependent $\Rightarrow (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ does not exist.
- Collinearity poses no problem for prediction: the model $\widehat{height} = \hat{\beta}_0 + \hat{\beta}_1 \times age1 + \hat{\beta}_2 \times age2$ has theoretically infinitely many solutions but all result in the same predicted height.
- For example, the following solutions are all equivalent: $\widehat{height} = 30 + 3 \times age1 + 0 \times age2 = 30 + 2 \times age1 + 12 \times age2 = 30 + 1 \times age1 + 24 \times age2.$
- Quantifying individual effects of collinear predictors is a problem.

Detecting Collinearity: Correlation Matrix

Detecting Collinearity: Correlation Matrix

- Collinearity can be detected by studying the sample correlation matrix of continuous predictors.

Detecting Collinearity: Correlation Matrix

- Collinearity can be detected by studying the sample correlation matrix of continuous predictors.
- Given a dataset, sample correlation measure between two predictors $\mathbf{x}_1$ and $\mathbf{x_2} \Rightarrow$

Detecting Collinearity: Correlation Matrix

- Collinearity can be detected by studying the sample correlation matrix of continuous predictors.
- Given a dataset, sample correlation measure between two predictors $\mathbf{x}_1$ and $\mathbf{x_2} \Rightarrow$ mean-centering them $\tilde{\mathbf{x}}_1$ and $\tilde{\mathbf{x}}_2 \Rightarrow$

Detecting Collinearity: Correlation Matrix

- Collinearity can be detected by studying the sample correlation matrix of continuous predictors.
- Given a dataset, sample correlation measure between two predictors $\mathbf{x}_1$ and $\mathbf{x_2}$ $\Rightarrow$ mean-centering them $\tilde{\mathbf{x}}_1$ and $\tilde{\mathbf{x}}_2$ $\Rightarrow \rho = \frac{\tilde{\mathbf{x}}_1^{\mathrm{T}}\tilde{\mathbf{x}}_2}{\|\tilde{\mathbf{x}}_1\|\|\tilde{\mathbf{x}}_2\|}$ is in between -1 and 1.

Detecting Collinearity: Correlation Matrix

- Collinearity can be detected by studying the sample correlation matrix of continuous predictors.
- Given a dataset, sample correlation measure between two predictors $\mathbf{x}_1$ and $\mathbf{x_2}$ $\Rightarrow$ mean-centering them $\tilde{\mathbf{x}}_1$ and $\tilde{\mathbf{x}}_2$ $\Rightarrow \rho = \dfrac{\tilde{\mathbf{x}}_1^{\mathrm{T}}\tilde{\mathbf{x}}_2}{\|\tilde{\mathbf{x}}_1\|\|\tilde{\mathbf{x}}_2\|}$ is in between -1 and 1.
- Correlation matrix for some continuous predictors from the `saratogaHouses` dataset:

	livingArea	lotSize	age	landValue	bedrooms	rooms
livingArea	1.00	0.16	-0.17	0.42	0.66	0.73
lotSize	0.16	1.00	-0.02	0.06	0.11	0.14
age	-0.17	-0.02	1.00	-0.02	0.03	-0.08
landValue	0.42	0.06	-0.02	1.00	0.20	0.30
bedrooms	0.66	0.11	0.03	0.20	1.00	0.67
rooms	0.73	0.14	-0.08	0.30	0.67	1.00

Consequences of Correlated Predictors

In model built with correlated predictors:

Consequences of Correlated Predictors

In model built with correlated predictors:

- regression coefficients estimates will change dramatically depending on which correlated predictors are included or not;

Consequences of Correlated Predictors

In model built with correlated predictors:

- regression coefficients estimates will change dramatically depending on which correlated predictors are included or not;
- coefficient estimates for predictors with known strong relationships with the response will not be accurate;

Consequences of Correlated Predictors

In model built with correlated predictors:

- regression coefficients estimates will change dramatically depending on which correlated predictors are included or not;
- coefficient estimates for predictors with known strong relationships with the response will not be accurate;
- standard errors of the coefficients estimates will be (relatively) large;

Consequences of Correlated Predictors

In model built with correlated predictors:

- regression coefficients estimates will change dramatically depending on which correlated predictors are included or not;
- coefficient estimates for predictors with known strong relationships with the response will not be accurate;
- standard errors of the coefficients estimates will be (relatively) large;
- wider confidence intervals for coefficients.

Quantifying Collinearity: Variance Inflation Factor (VIF)

Quantifying Collinearity: Variance Inflation Factor (VIF)

- If all the predictors are perfectly collinear,

Quantifying Collinearity: Variance Inflation Factor (VIF)

- If all the predictors are perfectly collinear, then the R^2 metric when one predictor is regressed upon the others will be exactly 1.

Quantifying Collinearity: Variance Inflation Factor (VIF)

- If all the predictors are perfectly collinear, then the R^2 metric when one predictor is regressed upon the others will be exactly 1.
- The *tolerance* of the predictor which is regressed upon the others is $1 - R^2$.

Quantifying Collinearity: Variance Inflation Factor (VIF)

- If all the predictors are perfectly collinear, then the R^2 metric when one predictor is regressed upon the others will be exactly 1.
- The *tolerance* of the predictor which is regressed upon the others is $1 - R^2$.
- A small value of the *tolerance* (< 0.1, for example) indicates that the predictor under consideration is highly correlated with the other predictors.

Quantifying Collinearity: Variance Inflation Factor (VIF)

- If all the predictors are perfectly collinear, then the R^2 metric when one predictor is regressed upon the others will be exactly 1.
- The *tolerance* of the predictor which is regressed upon the others is $1 - R^2$.
- A small value of the *tolerance* (< 0.1, for example) indicates that the predictor under consideration is highly correlated with the other predictors.
- The variance inflation factor (VIF) of the predictor under consideration is $1/tolerance = 1/(1 - R^2)$.

Quantifying Collinearity: Variance Inflation Factor (VIF)

- If all the predictors are perfectly collinear, then the R^2 metric when one predictor is regressed upon the others will be exactly 1.
- The *tolerance* of the predictor which is regressed upon the others is $1 - R^2$.
- A small value of the *tolerance* (< 0.1, for example) indicates that the predictor under consideration is highly correlated with the other predictors.
- The variance inflation factor (VIF) of the predictor under consideration is $1/tolerance = 1/(1 - R^2)$.
- A large vale of VIF (> 10, for example) indicates additional study about the correlation between predictors.

Residual plots: Introduction

Residual plots: Introduction

- Recall that the residual $R = Y - \hat{Y} = Y - \mathbf{X}\beta$ is a catch-all quantifier for everything that the linear model does not capture.

Residual plots: Introduction

- Recall that the residual $R = Y - \hat{Y} = Y - \mathbf{X}\beta$ is a catch-all quantifier for everything that the linear model does not capture.
- Residual is uncorrelated with the fitted response value:

Residual plots: Introduction

- Recall that the residual $R = Y - \hat{Y} = Y - \mathbf{X}\beta$ is a catch-all quantifier for everything that the linear model does not capture.
- Residual is uncorrelated with the fitted response value:

cov$(R, \hat{Y})$

Residual plots: Introduction

- Recall that the residual $R = Y - \hat{Y} = Y - \mathbf{X}\beta$ is a catch-all quantifier for everything that the linear model does not capture.
- Residual is uncorrelated with the fitted response value:

$$\text{cov}(R, \hat{Y}) = E\left[\left(R - \underbrace{E[R]}_{\approx \frac{1}{n}\mathbf{1}^{\mathrm{T}}\mathbf{r}=0}\right)\left(\hat{Y} - E\left[\hat{Y}\right]\right)\right]$$

Residual plots: Introduction

- Recall that the residual $R = Y - \hat{Y} = Y - \mathbf{X}\beta$ is a catch-all quantifier for everything that the linear model does not capture.
- Residual is <u>uncorrelated</u> with the fitted response value:

$$\text{cov}(R, \hat{Y}) = E\left[\left(R - \underbrace{E[R]}_{\approx \frac{1}{n}\mathbf{1}^{\mathrm{T}}\mathbf{r} = 0}\right)\left(\hat{Y} - E\left[\hat{Y}\right]\right)\right] = E\left[R\hat{Y} - RE\left[\hat{Y}\right]\right]$$

Residual plots: Introduction

- Recall that the residual $R = Y - \hat{Y} = Y - \mathbf{X}\beta$ is a catch-all quantifier for everything that the linear model does not capture.
- Residual is <u>uncorrelated</u> with the fitted response value:

$$\text{cov}(R, \hat{Y}) = E\left[\left(R - \underbrace{E[R]}_{\approx \frac{1}{n}\mathbf{1}^{\mathrm{T}}\mathbf{r}=0}\right)\left(\hat{Y} - E\left[\hat{Y}\right]\right)\right] = E\left[R\hat{Y} - RE\left[\hat{Y}\right]\right]$$

$$= \underbrace{E\left[R\hat{Y}\right]}_{\approx \frac{1}{n}\mathbf{r}^{\mathrm{T}}\hat{\mathbf{y}}} - \underbrace{E\left[R\right]}_{=0} E\left[\hat{Y}\right]$$

Residual plots: Introduction

- Recall that the residual $R = Y - \hat{Y} = Y - \mathbf{X}\beta$ is a catch-all quantifier for everything that the linear model does not capture.
- Residual is uncorrelated with the fitted response value:

$$\mathrm{cov}(R, \hat{Y}) = E\left[\left(R - \underbrace{E[R]}_{\approx \frac{1}{n}\mathbf{1}^{\mathrm{T}}\mathbf{r}=0}\right)\left(\hat{Y} - E\left[\hat{Y}\right]\right)\right] = E\left[R\hat{Y} - RE\left[\hat{Y}\right]\right]$$

$$= \underbrace{E\left[R\hat{Y}\right]}_{\approx \frac{1}{n}\mathbf{r}^{\mathrm{T}}\hat{\mathbf{y}}} - \underbrace{E\left[R\right]}_{=0} E\left[\hat{Y}\right] \approx \frac{1}{n}\mathbf{r}^{\mathrm{T}}\left(\underbrace{\mathbf{X}\hat{\beta}}_{=\hat{\mathbf{y}}}\right)$$

Residual plots: Introduction

- Recall that the residual $R = Y - \hat{Y} = Y - \mathbf{X}\beta$ is a catch-all quantifier for everything that the linear model does not capture.
- Residual is <u>uncorrelated</u> with the fitted response value:

$$\text{cov}(R, \hat{Y}) = E\left[\left(R - \underbrace{E[R]}_{\approx \frac{1}{n}\mathbf{1}^{\mathrm{T}}\mathbf{r}=0}\right)\left(\hat{Y} - E\left[\hat{Y}\right]\right)\right] = E\left[R\hat{Y} - RE\left[\hat{Y}\right]\right]$$

$$= \underbrace{E\left[R\hat{Y}\right]}_{\approx \frac{1}{n}\mathbf{r}^{\mathrm{T}}\hat{\mathbf{y}}} - \underbrace{E\left[R\right]}_{=0} E\left[\hat{Y}\right] \approx \frac{1}{n}\mathbf{r}^{\mathrm{T}}\left(\underbrace{\mathbf{X}\hat{\beta}}_{=\hat{\mathbf{y}}}\right) = \frac{1}{n}\left(\underbrace{\mathbf{X}^{\mathrm{T}}\mathbf{r}}_{=0}\right)^{\mathrm{T}}\hat{\beta} = 0.$$

Residual plots: Introduction

- Recall that the residual $R = Y - \hat{Y} = Y - \mathbf{X}\beta$ is a catch-all quantifier for everything that the linear model does not capture.
- Residual is <u>uncorrelated</u> with the fitted response value:

$$\text{cov}(R, \hat{Y}) = E\left[\left(R - \underbrace{E[R]}_{\approx \frac{1}{n}\mathbf{1}^{\mathrm{T}}\mathbf{r}=0}\right)\left(\hat{Y} - E\left[\hat{Y}\right]\right)\right] = E\left[R\hat{Y} - RE\left[\hat{Y}\right]\right]$$

$$= \underbrace{E\left[R\hat{Y}\right]}_{\approx \frac{1}{n}\mathbf{r}^{\mathrm{T}}\hat{\mathbf{y}}} - \underbrace{E\left[R\right]}_{=0} E\left[\hat{Y}\right] \approx \frac{1}{n}\mathbf{r}^{\mathrm{T}}\left(\underbrace{\mathbf{X}\hat{\beta}}_{=\hat{\mathbf{y}}}\right) = \frac{1}{n}\left(\underbrace{\mathbf{X}^{\mathrm{T}}\mathbf{r}}_{=0}\right)^{\mathrm{T}}\hat{\beta} = 0.$$

- Residual is <u>+vely correlated</u> with true response value:

Residual plots: Introduction

- Recall that the residual $R = Y - \hat{Y} = Y - \mathbf{X}\beta$ is a catch-all quantifier for everything that the linear model does not capture.
- Residual is uncorrelated with the fitted response value:

$$\text{cov}(R, \hat{Y}) = E\left[\left(R - \underbrace{E[R]}_{\approx \frac{1}{n}\mathbf{1}^{\mathrm{T}}\mathbf{r}=0}\right)\left(\hat{Y} - E\left[\hat{Y}\right]\right)\right] = E\left[R\hat{Y} - RE\left[\hat{Y}\right]\right]$$

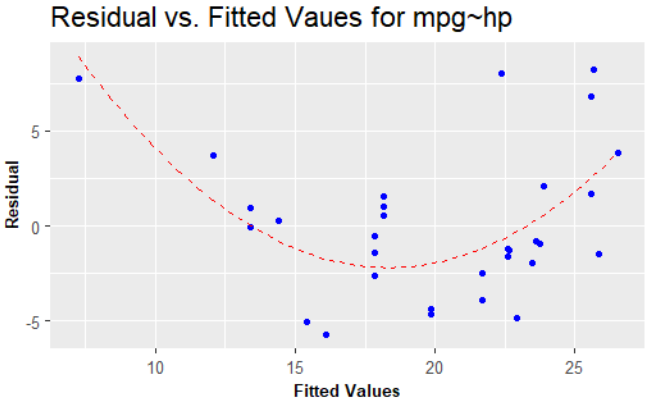
$$= \underbrace{E\left[R\hat{Y}\right]}_{\approx \frac{1}{n}\mathbf{r}^{\mathrm{T}}\hat{\mathbf{y}}} - \underbrace{E\left[R\right]}_{=0} E\left[\hat{Y}\right] \approx \frac{1}{n}\mathbf{r}^{\mathrm{T}}\left(\underbrace{\mathbf{X}\hat{\beta}}_{=\hat{\mathbf{y}}}\right) = \frac{1}{n}\left(\underbrace{\mathbf{X}^{\mathrm{T}}\mathbf{r}}_{=0}\right)^{\mathrm{T}}\hat{\beta} = 0.$$

- Residual is +vely correlated with true response value: $\text{cov}(R, Y) = \sigma^2$.

Residual plots: Continued

Residual plots: Continued

A residual plot shows the relationship between the residuals and the fitted values.



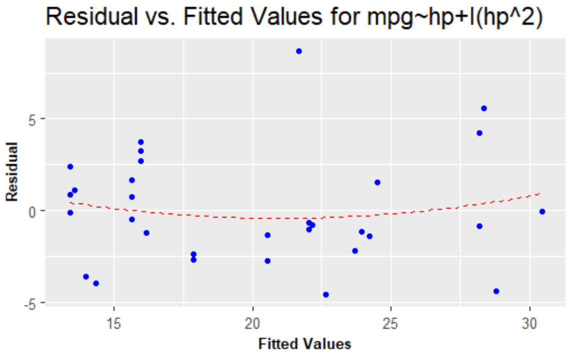
Interpreting Residual Plots

Interpreting Residual Plots

A residual plot with a discernible pattern is an indication of *nonlinearity*;

Interpreting Residual Plots

A residual plot with a discernible pattern is an indication of *nonlinearity*; apply nonlinear transformation for the predictor such as $X^2, \sqrt{X}$, etc.



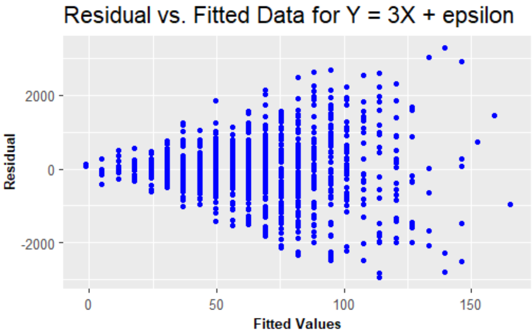
Interpreting Residual Plots: Continued

Interpreting Residual Plots: Continued

A funnel-shaped residual plot is an indication of heteroskedasticity,

Interpreting Residual Plots: Continued

A funnel-shaped residual plot is an indication of heteroskedasticity, which means that the random error term <u>does not</u> have a constant variance impacting the *standard error*, *confidence interval*, and hypothesis test calculations.



Heteroskedasticity: Non-constant Variance of Error & Weighted Least Squares

Heteroskedasticity: Non-constant Variance of Error & Weighted Least Squares

- In weighted least squares, we minimize the the sum of the squares of the *weighted* residuals for all samples in the dataset:

Heteroskedasticity: Non-constant Variance of Error & Weighted Least Squares

- In weighted least squares, we minimize the the sum of the squares of the *weighted* residuals for all samples in the dataset:

$$\min \sum_{i=1}^{n} \left(w_i r^{(i)}\right)^2 = \sum_{i=1}^{n} w_i^2 \left(y^{(i)} - \left(\beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)}\right)\right)^2.$$

Heteroskedasticity: Non-constant Variance of Error & Weighted Least Squares

- In weighted least squares, we minimize the the sum of the squares of the *weighted* residuals for all samples in the dataset:
$$\min \sum_{i=1}^{n} \left(w_i r^{(i)}\right)^2 = \sum_{i=1}^{n} w_i^2 \left(y^{(i)} - \left(\beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)}\right)\right)^2.$$
- The weights are chosen such that samples with large error variances contribute less to the summation above,

Heteroskedasticity: Non-constant Variance of Error & Weighted Least Squares

- In weighted least squares, we minimize the the sum of the squares of the *weighted* residuals for all samples in the dataset:
$$\min \sum_{i=1}^{n} \left(w_i r^{(i)}\right)^2 = \sum_{i=1}^{n} w_i^2 \left(y^{(i)} - \left(\beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)}\right)\right)^2.$$
- The weights are chosen such that samples with large error variances contribute less to the summation above, thus to the linear model.

Heteroskedasticity: Non-constant Variance of Error & Weighted Least Squares

- In weighted least squares, we minimize the the sum of the squares of the *weighted* residuals for all samples in the dataset: $\min \sum_{i=1}^{n} \left(w_i r^{(i)}\right)^2 = \sum_{i=1}^{n} w_i^2 \left(y^{(i)} - \left(\beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)}\right)\right)^2.$
- The weights are chosen such that samples with large error variances contribute less to the summation above, thus to the linear model.
- Typically, the recorded response for the ith sample is calculated as an average of n_i raw observations.

Heteroskedasticity: Non-constant Variance of Error & Weighted Least Squares

- In weighted least squares, we minimize the the sum of the squares of the *weighted* residuals for all samples in the dataset:
  $\min \sum_{i=1}^{n} \left(w_i r^{(i)}\right)^2 = \sum_{i=1}^{n} w_i^2 \left(y^{(i)} - \left(\beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)}\right)\right)^2.$
- The weights are chosen such that samples with large error variances contribute less to the summation above, thus to the linear model.
- Typically, the recorded response for the ith sample is calculated as an average of n_i raw observations.
- If each of those raw observations are uncorrelated with constant variance σ^2,

Heteroskedasticity: Non-constant Variance of Error & Weighted Least Squares

- In weighted least squares, we minimize the the sum of the squares of the *weighted* residuals for all samples in the dataset:
  $\min \sum_{i=1}^{n} \left(w_i r^{(i)}\right)^2 = \sum_{i=1}^{n} w_i^2 \left(y^{(i)} - \left(\beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)}\right)\right)^2.$
- The weights are chosen such that samples with large error variances contribute less to the summation above, thus to the linear model.
- Typically, the recorded response for the ith sample is calculated as an average of n_i raw observations.
- If each of those raw observations are uncorrelated with constant variance σ^2, then the variance in the ith sample is σ^2/n_i.

Heteroskedasticity: Non-constant Variance of Error & Weighted Least Squares

- In weighted least squares, we minimize the the sum of the squares of the *weighted* residuals for all samples in the dataset:
$\min \sum_{i=1}^{n} \left(w_i r^{(i)}\right)^2 = \sum_{i=1}^{n} w_i^2 \left(y^{(i)} - \left(\beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)}\right)\right)^2.$
- The weights are chosen such that samples with large error variances contribute less to the summation above, thus to the linear model.
- Typically, the recorded response for the ith sample is calculated as an average of n_i raw observations.
- If each of those raw observations are uncorrelated with constant variance σ^2, then the variance in the ith sample is σ^2/n_i.
- The weights for each sample can now be chosen to be proportional to the inverse of the associated variances;

Heteroskedasticity: Non-constant Variance of Error & Weighted Least Squares

- In weighted least squares, we minimize the the sum of the squares of the *weighted* residuals for all samples in the dataset:
  $\min \sum_{i=1}^{n} \left(w_i r^{(i)}\right)^2 = \sum_{i=1}^{n} w_i^2 \left(y^{(i)} - \left(\beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)}\right)\right)^2.$
- The weights are chosen such that samples with large error variances contribute less to the summation above, thus to the linear model.
- Typically, the recorded response for the ith sample is calculated as an average of n_i raw observations.
- If each of those raw observations are uncorrelated with constant variance σ^2, then the variance in the ith sample is σ^2/n_i.
- The weights for each sample can now be chosen to be proportional to the inverse of the associated variances; that is, $w_i = n_i$.

Interaction: Basic Ideas

Interaction: Basic Ideas

- An interaction occurs when the effect of a predictor on the response variable depends on the value of another predictor.

Interaction: Basic Ideas

- An interaction occurs when the effect of a predictor on the response variable depends on the value of another predictor.
- Example: suppose that we consider the *SaratogaHouses* dataset with $price$ as the response and $livingArea$ and $newConstruction$ as the predictors;

Interaction: Basic Ideas

- An interaction occurs when the effect of a predictor on the response variable depends on the value of another predictor.
- Example: suppose that we consider the *SaratogaHouses* dataset with $price$ as the response and $livingArea$ and $newConstruction$ as the predictors; the predictor $newConstruction$ is categorical with two levels: (1) No (2) Yes.

Interaction: Basic Ideas

- An interaction occurs when the effect of a predictor on the response variable depends on the value of another predictor.
- Example: suppose that we consider the *SaratogaHouses* dataset with $price$ as the response and $livingArea$ and $newConstruction$ as the predictors; the predictor $newConstruction$ is categorical with two levels: (1) No (2) Yes.
- An MLRM with interaction between $livingArea$ and $newConstruction$ is

Interaction: Basic Ideas

- An interaction occurs when the effect of a predictor on the response variable depends on the value of another predictor.
- Example: suppose that we consider the *SaratogaHouses* dataset with $price$ as the response and $livingArea$ and $newConstruction$ as the predictors; the predictor $newConstruction$ is categorical with two levels: (1) No (2) Yes.
- An MLRM with interaction between $livingArea$ and $newConstruction$ is $\widehat{price} = \hat{\beta}_0 + \hat{\beta}_1 \times livingArea + \hat{\beta}_2 \times newConstructionYes + \hat{\beta}_3 \times livingArea \times newConstructionYes$.

Interaction: Basic Ideas

- An interaction occurs when the effect of a predictor on the response variable depends on the value of another predictor.
- Example: suppose that we consider the *SaratogaHouses* dataset with $price$ as the response and $livingArea$ and $newConstruction$ as the predictors; the predictor $newConstruction$ is categorical with two levels: (1) No (2) Yes.
- An MLRM with interaction between $livingArea$ and $newConstruction$ is $\widehat{price} = \hat{\beta}_0 + \hat{\beta}_1 \times livingArea + \hat{\beta}_2 \times newConstructionYes + \hat{\beta}_3 \times livingArea \times newConstructionYes.$
- Interaction to be considered

Interaction: Basic Ideas

- An interaction occurs when the effect of a predictor on the response variable depends on the value of another predictor.
- Example: suppose that we consider the *SaratogaHouses* dataset with $price$ as the response and $livingArea$ and $newConstruction$ as the predictors; the predictor $newConstruction$ is categorical with two levels: (1) No (2) Yes.
- An MLRM with interaction between $livingArea$ and $newConstruction$ is $\widehat{price} = \hat{\beta}_0 + \hat{\beta}_1 \times livingArea + \hat{\beta}_2 \times newConstructionYes + \hat{\beta}_3 \times livingArea \times newConstructionYes.$
- Interaction to be considered (1) if a particular predictor has a large effect on the response (large coefficient estimate)

Interaction: Basic Ideas

- An interaction occurs when the effect of a predictor on the response variable depends on the value of another predictor.
- Example: suppose that we consider the *SaratogaHouses* dataset with $price$ as the response and $livingArea$ and $newConstruction$ as the predictors; the predictor $newConstruction$ is categorical with two levels: (1) No (2) Yes.
- An MLRM with interaction between $livingArea$ and $newConstruction$ is $\widehat{price} = \hat{\beta}_0 + \hat{\beta}_1 \times livingArea + \hat{\beta}_2 \times newConstructionYes + \hat{\beta}_3 \times livingArea \times newConstructionYes.$
- Interaction to be considered (1) if a particular predictor has a large effect on the response (large coefficient estimate) and/or (2) the presence of categorical predictors as coefficients of other predictors may vary across groups.

Interaction Between Two Continuous Predictors

Interaction Between Two Continuous Predictors

- Example: suppose that we consider the *SaratogaHouses* dataset with $price$ as the response and $livingArea$ and $rooms$ as the predictors.

Interaction Between Two Continuous Predictors

- Example: suppose that we consider the *SaratogaHouses* dataset with $price$ as the response and $livingArea$ and $rooms$ as the predictors.
- An MLRM with interaction between $livingArea$ and $rooms$ is

Interaction Between Two Continuous Predictors

- Example: suppose that we consider the *SaratogaHouses* dataset with $price$ as the response and $livingArea$ and $rooms$ as the predictors.
- An MLRM with interaction between $livingArea$ and $rooms$ is $\widehat{price} = \hat{\beta}_0 + \hat{\beta}_1 \times livingArea + \hat{\beta}_2 \times rooms + \hat{\beta}_3 \times livingArea \times rooms.$

Interaction Between Two Continuous Predictors

- Example: suppose that we consider the *SaratogaHouses* dataset with $price$ as the response and $livingArea$ and $rooms$ as the predictors.
- An MLRM with interaction between $livingArea$ and $rooms$ is $\widehat{price} = \hat{\beta}_0 + \hat{\beta}_1 \times livingArea + \hat{\beta}_2 \times rooms + \hat{\beta}_3 \times livingArea \times rooms.$
- How to interpret the coefficient estimates?

Interaction Between Two Continuous Predictors

- Example: suppose that we consider the *SaratogaHouses* dataset with $price$ as the response and $livingArea$ and $rooms$ as the predictors.
- An MLRM with interaction between $livingArea$ and $rooms$ is $\widehat{price} = \hat{\beta}_0 + \hat{\beta}_1 \times livingArea + \hat{\beta}_2 \times rooms + \hat{\beta}_3 \times livingArea \times rooms.$
- How to interpret the coefficient estimates? Suppose we increase the living area by 1 unit while keeping the number of rooms fixed.

Interaction Between Two Continuous Predictors

- Example: suppose that we consider the *SaratogaHouses* dataset with $price$ as the response and $livingArea$ and $rooms$ as the predictors.
- An MLRM with interaction between $livingArea$ and $rooms$ is $\widehat{price} = \hat{\beta}_0 + \hat{\beta}_1 \times livingArea + \hat{\beta}_2 \times rooms + \hat{\beta}_3 \times livingArea \times rooms.$
- How to interpret the coefficient estimates? Suppose we increase the living area by 1 unit while keeping the number of rooms fixed.

$$\underbrace{[\hat{\beta}_0 + \hat{\beta}_1 \times (livingArea + 1) + \hat{\beta}_2 \times rooms + \hat{\beta}_3 \times (livingArea + 1) \times rooms]}_{\widehat{price}_{\text{new}}}$$

$$-\underbrace{[\hat{\beta}_0 + \hat{\beta}_1 \times livingArea + \hat{\beta}_2 \times rooms + \hat{\beta}_3 \times livingArea \times rooms]}_{\widehat{price}_{\text{old}}}$$

$$= \hat{\beta}_1 + \hat{\beta}_3 \times rooms.$$

Interaction Between Two Categorical Predictors

Interaction Between Two Categorical Predictors

- Example: suppose that we consider the *SaratogaHouses* dataset with $price$ as the response and $heating$ and $centralAir$ as the predictors.

Interaction Between Two Categorical Predictors

- Example: suppose that we consider the *SaratogaHouses* dataset with $price$ as the response and $heating$ and $centralAir$ as the predictors.
- The predictor $heating$ has three levels: (1) electric (2) hot air (3) hot water/steam.

Interaction Between Two Categorical Predictors

- Example: suppose that we consider the *SaratogaHouses* dataset with $price$ as the response and $heating$ and $centralAir$ as the predictors.
- The predictor $heating$ has three levels: (1) electric (2) hot air (3) hot water/steam.
- The predictor $centralAir$ has two levels: (1) No (2) Yes.

Interaction Between Two Categorical Predictors

- Example: suppose that we consider the *SaratogaHouses* dataset with $price$ as the response and $heating$ and $centralAir$ as the predictors.
- The predictor $heating$ has three levels: (1) electric (2) hot air (3) hot water/steam.
- The predictor $centralAir$ has two levels: (1) No (2) Yes.
- An MLRM with interaction between $heating$ and $centralAir$ is

Interaction Between Two Categorical Predictors

- Example: suppose that we consider the *SaratogaHouses* dataset with $price$ as the response and $heating$ and $centralAir$ as the predictors.
- The predictor $heating$ has three levels: (1) electric (2) hot air (3) hot water/steam.
- The predictor $centralAir$ has two levels: (1) No (2) Yes.
- An MLRM with interaction between $heating$ and $centralAir$ is

$$\begin{aligned}\widehat{price} = \hat{\beta}_0 + \hat{\beta}_1 \times heatinghotair + \hat{\beta}_2 \times heatinghot\ water \\ + \hat{\beta}_3 \times centralAirYes \\ + \hat{\beta}_4 \times heatinghotair \times centralAirYes \\ + \hat{\beta}_5 \times heatinghot\ water/steam \times centralAirYes.\end{aligned}$$

Interaction Between Two Categorical Predictors: Continued

Interaction Between Two Categorical Predictors: Continued

- For electric-heated houses, predicted price

$$= \hat{\beta}_0 + \hat{\beta}_3 \times centralAirYes = \begin{cases} \hat{\beta}_0 & \text{if not centrally air-conditioned} \\ \hat{\beta}_0 + \hat{\beta}_3 & \text{if centrally air-conditioned.} \end{cases}$$

Interaction Between Two Categorical Predictors: Continued

- For electric-heated houses, predicted price
$$= \hat{\beta}_0 + \hat{\beta}_3 \times centralAirYes = \begin{cases} \hat{\beta}_0 & \text{if not centrally air-conditioned} \\ \hat{\beta}_0 + \hat{\beta}_3 & \text{if centrally air-conditioned.} \end{cases}$$
- Subtracting

Interaction Between Two Categorical Predictors: Continued

- For electric-heated houses, predicted price
$$= \hat{\beta}_0 + \hat{\beta}_3 \times centralAirYes = \begin{cases} \hat{\beta}_0 & \text{if not centrally air-conditioned} \\ \hat{\beta}_0 + \hat{\beta}_3 & \text{if centrally air-conditioned.} \end{cases}$$
- Subtracting $\Rightarrow \hat{\beta}_3 =$ difference between average prices of centrally air-conditioned and not centrally air-conditioned houses among electric-heated houses.

Interaction Between Two Categorical Predictors: Continued

- For electric-heated houses, predicted price
$$= \hat{\beta}_0 + \hat{\beta}_3 \times centralAirYes = \begin{cases} \hat{\beta}_0 & \text{if not centrally air-conditioned} \\ \hat{\beta}_0 + \hat{\beta}_3 & \text{if centrally air-conditioned.} \end{cases}$$
- Subtracting $\Rightarrow \hat{\beta}_3 =$ difference between average prices of centrally air-conditioned and not centrally air-conditioned houses among electric-heated houses.
- Suppose we have a house that is hot air heated and not centrally air-conditioned;

Interaction Between Two Categorical Predictors: Continued

- For electric-heated houses, predicted price
$$= \hat{\beta}_0 + \hat{\beta}_3 \times centralAirYes = \begin{cases} \hat{\beta}_0 & \text{if not centrally air-conditioned} \\ \hat{\beta}_0 + \hat{\beta}_3 & \text{if centrally air-conditioned.} \end{cases}$$
- Subtracting $\Rightarrow \hat{\beta}_3 =$ difference between average prices of centrally air-conditioned and not centrally air-conditioned houses among electric-heated houses.
- Suppose we have a house that is hot air heated and not centrally air-conditioned; the predicted price of the house is $\hat{\beta}_0 + \hat{\beta}_1$.

Interaction Between Two Categorical Predictors: Continued

- For electric-heated houses, predicted price
$$= \hat{\beta}_0 + \hat{\beta}_3 \times centralAirYes = \begin{cases} \hat{\beta}_0 & \text{if not centrally air-conditioned} \\ \hat{\beta}_0 + \hat{\beta}_3 & \text{if centrally air-conditioned.} \end{cases}$$
- Subtracting $\Rightarrow \hat{\beta}_3 =$ difference between average prices of centrally air-conditioned and not centrally air-conditioned houses among electric-heated houses.
- Suppose we have a house that is hot air heated and not centrally air-conditioned; the predicted price of the house is $\hat{\beta}_0 + \hat{\beta}_1$.
- Now consider another house that is hot air heated and centrally air-conditioned;

Interaction Between Two Categorical Predictors: Continued

- For electric-heated houses, predicted price
$$= \hat{\beta}_0 + \hat{\beta}_3 \times centralAirYes = \begin{cases} \hat{\beta}_0 & \text{if not centrally air-conditioned} \\ \hat{\beta}_0 + \hat{\beta}_3 & \text{if centrally air-conditioned.} \end{cases}$$
- Subtracting $\Rightarrow \hat{\beta}_3 =$ difference between average prices of centrally air-conditioned and not centrally air-conditioned houses among electric-heated houses.
- Suppose we have a house that is hot air heated and not centrally air-conditioned; the predicted price of the house is $\hat{\beta}_0 + \hat{\beta}_1$.
- Now consider another house that is hot air heated and centrally air-conditioned; the predicted price of this house is $\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_3 + \hat{\beta}_4$.

Interaction Between Two Categorical Predictors: Continued

- For electric-heated houses, predicted price
$$= \hat{\beta}_0 + \hat{\beta}_3 \times centralAirYes = \begin{cases} \hat{\beta}_0 & \text{if not centrally air-conditioned} \\ \hat{\beta}_0 + \hat{\beta}_3 & \text{if centrally air-conditioned.} \end{cases}$$
- Subtracting $\Rightarrow \hat{\beta}_3 =$ difference between average prices of centrally air-conditioned and not centrally air-conditioned houses among electric-heated houses.
- Suppose we have a house that is hot air heated and not centrally air-conditioned; the predicted price of the house is $\hat{\beta}_0 + \hat{\beta}_1$.
- Now consider another house that is hot air heated and centrally air-conditioned; the predicted price of this house is $\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_3 + \hat{\beta}_4$.
- Subtracting the two results above, we see that $\hat{\beta}_3 + \hat{\beta}_4 =$ difference between average prices of centrally air-conditioned and not centrally air-conditioned houses among hot air-heated houses.

Interaction Between Continuous & Categorical Predictors

Interaction Between Continuous & Categorical Predictors

- Example: suppose that we consider the *SaratogaHouses* dataset with $price$ as the response and $livingArea$ and $newConstruction$ as the predictors.

Interaction Between Continuous & Categorical Predictors

- Example: suppose that we consider the *SaratogaHouses* dataset with $price$ as the response and $livingArea$ and $newConstruction$ as the predictors.
- The predictor $newConstruction$ has two levels: (1) No (2) Yes.

Interaction Between Continuous & Categorical Predictors

- Example: suppose that we consider the *SaratogaHouses* dataset with $price$ as the response and $livingArea$ and $newConstruction$ as the predictors.
- The predictor $newConstruction$ has two levels: (1) No (2) Yes.
- An MLRM with interaction between $livingArea$ and $newConstruction$ is

Interaction Between Continuous & Categorical Predictors

- Example: suppose that we consider the *SaratogaHouses* dataset with $price$ as the response and $livingArea$ and $newConstruction$ as the predictors.
- The predictor $newConstruction$ has two levels: (1) No (2) Yes.
- An MLRM with interaction between $livingArea$ and $newConstruction$ is

$$\widehat{price} = \hat{\beta}_0 + \hat{\beta}_1 \times livingArea + \hat{\beta}_2 \times newConstructionYes \\ + \hat{\beta}_3 \times livingArea \times newConstructionYes.$$

Interaction Between Continuous & Categorical Predictors

- Example: suppose that we consider the *SaratogaHouses* dataset with $price$ as the response and $livingArea$ and $newConstruction$ as the predictors.
- The predictor $newConstruction$ has two levels: (1) No (2) Yes.
- An MLRM with interaction between $livingArea$ and $newConstruction$ is

$$\widehat{price} = \hat{\beta}_0 + \hat{\beta}_1 \times livingArea + \hat{\beta}_2 \times newConstructionYes$$
$$+ \hat{\beta}_3 \times livingArea \times newConstructionYes.$$

- The predicted house price

$$\widehat{price} = \begin{cases} \hat{\beta}_0 + \hat{\beta}_1 \times livingArea & \text{if old house,} \\ \left(\hat{\beta}_0 + \hat{\beta}_2\right) + \left(\hat{\beta}_1 + \hat{\beta}_3\right) \times livingArea & \text{if new house.} \end{cases}$$

Interaction Between Continuous & Categorical Predictors

- Example: suppose that we consider the *SaratogaHouses* dataset with $price$ as the response and $livingArea$ and $newConstruction$ as the predictors.
- The predictor $newConstruction$ has two levels: (1) No (2) Yes.
- An MLRM with interaction between $livingArea$ and $newConstruction$ is

$$\widehat{price} = \hat{\beta}_0 + \hat{\beta}_1 \times livingArea + \hat{\beta}_2 \times newConstructionYes \\ + \hat{\beta}_3 \times livingArea \times newConstructionYes.$$

- The predicted house price

$$\widehat{price} = \begin{cases} \hat{\beta}_0 + \hat{\beta}_1 \times livingArea & \text{if old house,} \\ \left(\hat{\beta}_0 + \hat{\beta}_2\right) + \left(\hat{\beta}_1 + \hat{\beta}_3\right) \times livingArea & \text{if new house.} \end{cases}$$

- Note the differences in both intercept and slope for new houses.

Interaction Between Continuous & Categorical Predictors: Continued

Interaction Between Continuous & Categorical Predictors: Continued

The scatter plot indicates that a higher slope is needed for new houses: