COMP 680
Statistics for Computing and Data Science
Week 11: Linear Regression

Su Chen, Assistant Teaching Professor,
Rice D2K Lab

Outline

Regression

- Regression is the general label for investigating relationship between (two) variables
  - the term first coined by Francis Galton (1822 - 1911)
  - study of human differences and inheritance of intelligence

Regression

- Regression is the general label for investigating relationship between (two) variables
  - the term first coined by Francis Galton (1822 - 1911)
  - study of human differences and inheritance of intelligence
- Start with two (numerical) variables:
  - dependent variable Y: outcome, response, label ...
  - independent variable X: covariate, predictor, feature ...
  - regress Y on X to understand how Y depends on X

A Regression Model

- A pair of variables X and Y with multiple observations (data):

$$(X, Y) = (x_i, y_i), \quad i = 1, 2, \cdots n$$

A Regression Model

- A pair of variables X and Y with multiple observations (data):

$$(X, Y) = (x_i, y_i), \quad i = 1, 2, \cdots n$$

- Mathematical model and the regression function $f(x)$:

$$y_i = f(x_i) + \epsilon_i, \quad i = 1, 2, \cdots n$$

  - the regression function: quantifies the relationship between X and Y
  - the error terms ϵ_i: captures measurement errors and other discrepancies

Assumptions about the Errors

$$y_i = f(x_i) + \epsilon_i, \quad i = 1, 2, \cdots n$$

- Captures measurement errors and other discrepancies:

Assumptions about the Errors

$$y_i = f(x_i) + \epsilon_i, \quad i = 1, 2, \cdots n$$

- Captures measurement errors and other discrepancies:
- Most common is Gaussian errors:

$$\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma^2)$$

- $\boldsymbol{\epsilon}$ are independent of the regression function f
- $\boldsymbol{\epsilon}$ are independent of each other
- $\mathbb{E}[\boldsymbol{\epsilon}] = 0, \quad \mathbb{V}[\boldsymbol{\epsilon}] = \sigma^2$

Parametric vs. Nonparametric

$$y_i = f(x_i) + \epsilon_i, \quad i = 1, 2, \cdots n$$

- Key assumptions about $f(x)$: some trade-off

Parametric vs. Nonparametric

$$y_i = f(x_i) + \epsilon_i, \quad i = 1, 2, \cdots n$$

- Key assumptions about $f(x)$: some trade-off
- Parametric models: restrict $f(x)$ to be "simple" form

Parametric vs. Nonparametric

$$y_i = f(x_i) + \epsilon_i, \quad i = 1, 2, \cdots n$$

- Key assumptions about $f(x)$: some trade-off
- Parametric models: restrict $f(x)$ to be "simple" form
  - finite and fixed number of parameters
  - less flexible but more interpretable
  - what is the simplest f?
- Nonparametric models:

Parametric vs. Nonparametric

$$y_i = f(x_i) + \epsilon_i, \quad i = 1, 2, \cdots n$$

- Key assumptions about $f(x)$: some trade-off
- Parametric models: restrict $f(x)$ to be "simple" form
  - finite and fixed number of parameters
  - less flexible but more interpretable
  - what is the simplest f?
- Nonparametric models:
  - no fixed number of parameters
  - more flexible but less interpretable
  - need more data to estimate

Outline

Simple Linear Regression

- The regression function $f(x)$ is linear!

Simple Linear Regression

- The regression function $f(x)$ is linear!
  - $y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \cdots n$
  - a linear function is a straight line with slope β_1 and intercept β_0
  - β_1 and β_0 are the unknown parameters of the model

Simple Linear Regression

- The regression function $f(x)$ is linear!
  - $y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \cdots n$
  - a linear function is a straight line with slope β_1 and intercept β_0
  - β_1 and β_0 are the unknown parameters of the model
- The relationship between Y and X:
  - scatter around a straight line with slope β_1 and intercept β_0
  - the conditional distribution of Y conditioned on X is normal!!!

Simple Linear Regression

- The regression function $f(x)$ is linear!
  - $y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \cdots n$
  - a linear function is a straight line with slope β_1 and intercept β_0
  - β_1 and β_0 are the unknown parameters of the model
- The relationship between Y and X:
  - scatter around a straight line with slope β_1 and intercept β_0
  - the conditional distribution of Y conditioned on X is normal!!!

$$\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma^2) \Rightarrow Y|X = x \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\beta_0 + \beta_1 x, \sigma^2)$$

The Conditional Distribution

$$Y|X \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\beta_0 + \beta_1 X, \sigma^2)$$

The Conditional Distribution

$$Y|X \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\beta_0 + \beta_1 X, \sigma^2)$$

- The regression function $\beta_0 + \beta_1 X$ is the **conditional mean**!
  - the only randomness of the model comes from ϵ
  - the error distribution of ϵ is important
  - X is usually considered deterministic

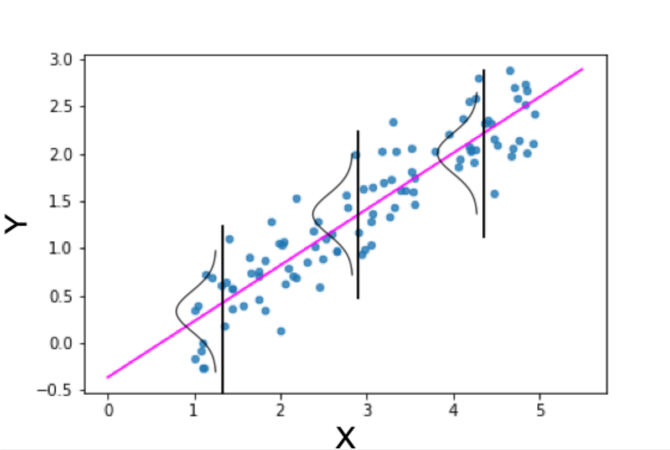
The Conditional Distribution

$$Y|X \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\beta_0 + \beta_1 X, \sigma^2)$$

- The regression function $\beta_0 + \beta_1 X$ is the **conditional mean**!
  - the only randomness of the model comes from ϵ
  - the error distribution of ϵ is important
  - X is usually considered deterministic
- The "true" regression function $f(x) = \beta_0 + \beta_1 x$
  - aka population regression line
  - unknown population parameters β_0 and β_1

Simple Linear Regression

The population regression line:



Simple Linear Regression

- Fitting the models means estimate β_0 and β_1 from the observed data
  - use $\hat{\beta}_0$ and $\hat{\beta}_1$ as notations of point estimates
  - $\hat{\beta}_0$ and $\hat{\beta}_1$ are statistics with sampling distribution!

Simple Linear Regression

- Fitting the models means estimate β_0 and β_1 from the observed data
  - use $\hat{\beta}_0$ and $\hat{\beta}_1$ as notations of point estimates
  - $\hat{\beta}_0$ and $\hat{\beta}_1$ are statistics with sampling distribution!
- What can a regression model do?
  - prediction: for a given value of X : $\hat{y}_i = \mathbb{E}[Y|X = x_i] = \hat{\beta}_0 + \hat{\beta}_1 x_i$
  - inference: sampling distributions and CI for β
  - interpretation: understand how Y depends on X
- Recall the 3 types of questions:

The Ordinary Least Square Line

- The notion of “the best” line:
  - the point estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ that best fit observed data

The Ordinary Least Square Line

- The notion of "the best" line:
  - the point estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ that best fit observed data
- Idea: minimize difference between $\hat{y}_i$ and y_i, i.e., the Residual Sum of Square (RSS):

$$\hat{\beta}_0^{\text{OLS}}, \hat{\beta}_1^{\text{OLS}} = \arg\min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \arg\min \sum_{i=1}^{n} (\beta_0 + \beta_1 x_i - y_i)^2$$

The Ordinary Least Square Line

- The notion of "the best" line:
  - the point estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ that best fit observed data
- Idea: minimize difference between $\hat{y}_i$ and y_i, i.e., the Residual Sum of Square (RSS):

$$\hat{\beta}_0^{\text{OLS}}, \hat{\beta}_1^{\text{OLS}} = \arg\min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \arg\min \sum_{i=1}^{n} (\beta_0 + \beta_1 x_i - y_i)^2$$

- Residual $r_i = y_i - \hat{y}_i$ represents the estimate of the error term ϵ_i
  - the OLS line minimize RSS

The Maximum Likelihood Estimate

- OLS is an optimization problem:
  - minimize loss function in ML: mean square loss = RSS / n
  - the distribution of ϵ or $Y|X$ is not important
  - however, no inference!

The Maximum Likelihood Estimate

- OLS is an optimization problem:
  - minimize loss function in ML: mean square loss = RSS / n
  - the distribution of ϵ or $Y|X$ is not important
  - however, no inference!
- Approach as a statistical problem:
  - maximum likelihood estimates

The Maximum Likelihood Estimate

- OLS is an optimization problem:
  - minimize loss function in ML: mean square loss = RSS / n
  - the distribution of ϵ or $Y|X$ is not important
  - however, no inference!
- Approach as a statistical problem:
  - maximum likelihood estimates
  - leads to equivalent solutions - why?

$$\hat{\beta}_0^{\text{MLE}}, \hat{\beta}_1^{\text{MLE}} \iff \hat{\beta}_0^{\text{OLS}}, \hat{\beta}_1^{\text{OLS}}$$

The Coefficients

- Closed form solutions:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2} = r_{X,Y} \cdot \frac{s_Y}{s_X}$$
$$\hat{\beta}_0 = \bar{y}_n - \hat{\beta}_1 \bar{x}_n$$

The Coefficients

- Closed form solutions:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2} = r_{X,Y} \cdot \frac{s_Y}{s_X}$$

$$\hat{\beta}_0 = \bar{y}_n - \hat{\beta}_1 \bar{x}_n$$

- slope $\hat{\beta}_1$ is rescaled Pearson correlation coefficient $r_{X,Y}$
  - describe the linear association between X and Y
  - depends on unit of X and Y

Inference

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased!

Inference

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased!
- The sampling distributions of them are normal, why?

Inference

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased!
- The sampling distributions of them are normal, why?

$$\hat{\beta}_1 \sim \mathrm{N}\left(\beta_1, \quad \sigma^2 \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}\right)$$

$$\hat{\beta}_0 \sim \mathrm{N}\left(\beta_0, \quad \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}_n^2}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}\right]\right)$$

Inference

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased!
- The sampling distributions of them are normal, why?

$$\hat{\beta}_1 \sim \mathrm{N}\left(\beta_1, \quad \sigma^2 \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}\right)$$
$$\hat{\beta}_0 \sim \mathrm{N}\left(\beta_0, \quad \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}_n^2}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}\right]\right)$$

- What about σ^2? usually unknown...
  - nuisance parameter, can be estimated from data
  - replace σ^2 with $\hat{\sigma}^2 = \mathrm{RSS}/(n-2)$: the above Normal $\rightarrow$ t distribution!

Inference

- Confidence intervals for β_j and $j = 0, 1$
  - $[\hat{\beta}_j + t_{\alpha/2, df=n-2} \cdot \mathrm{SE}(\hat{\beta}_j), \quad \hat{\beta}_j + t_{1-\alpha/2, df=n-2} \cdot \mathrm{SE}(\hat{\beta}_j)]$
  - standard error estimated with $\hat{\sigma}^2$

Inference

- Confidence intervals for β_j and $j = 0, 1$
  - $[\hat{\beta}_j + t_{\alpha/2, df=n-2} \cdot \mathrm{SE}(\hat{\beta}_j), \quad \hat{\beta}_j + t_{1-\alpha/2, df=n-2} \cdot \mathrm{SE}(\hat{\beta}_j)]$
  - standard error estimated with $\hat{\sigma}^2$
- Hypothesis testing on regression slope is standard:
  - H_0: $\beta_1 = 0$ vs. H_1: $\beta_1 \neq 0$: one-sample t-test
  - reject H_0 means β_1 is significantly different from 0
  - which means X and Y are correlated!

Inference

- Confidence intervals for β_j and $j = 0, 1$
  - $[\hat{\beta}_j + t_{\alpha/2, df=n-2} \cdot \mathrm{SE}(\hat{\beta}_j), \quad \hat{\beta}_j + t_{1-\alpha/2, df=n-2} \cdot \mathrm{SE}(\hat{\beta}_j)]$
  - standard error estimated with $\hat{\sigma}^2$
- Hypothesis testing on regression slope is standard:
  - H_0: $\beta_1 = 0$ vs. H_1: $\beta_1 \neq 0$: one-sample t-test
  - reject H_0 means β_1 is significantly different from 0
  - which means X and Y are correlated!
- for two-sided test, the following are equivalent:
  - reject H_0: $\beta_1 = 0$ at α level
  - $1 - \alpha$ CI for β_1 does NOT include 0

Prediction

For any given value of x^*, how do we predict corresponding y^*?

Prediction

For any given value of x^*, how do we predict corresponding y^*?

- Point estimate: conditional mean!
  - $\hat{y^*} = \mathbb{E}[Y|X = x^*] = \hat{\beta}_0 + \hat{\beta}_1 x^*$

Prediction

For any given value of x^*, how do we predict corresponding y^*?

- Point estimate: conditional mean!
  - $\hat{y^*} = \mathbb{E}[Y|X = x^*] = \hat{\beta}_0 + \hat{\beta}_1 x^*$
- Sampling distribution and CI:
  - $\hat{y^*} \sim \mathrm{N}\left(\beta_0 + \beta_1 x^*, \quad \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \bar{x}_n)^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}\right]\right)$
  - $[\hat{y^*} + t_{\alpha/2, df=n-2} \cdot \mathrm{SE}(\hat{y^*}), \quad \hat{y^*} + t_{1-\alpha/2, df=n-2} \cdot \mathrm{SE}(\hat{y^*})]$
  - standard error estimated with $\hat{\sigma}^2$

Prediction

For any given value of x^*, how do we predict corresponding y^*?

- Point estimate: conditional mean!
  - $\hat{y^*} = \mathbb{E}[Y|X = x^*] = \hat{\beta}_0 + \hat{\beta}_1 x^*$
- Sampling distribution and CI:
  - $\hat{y^*} \sim \mathrm{N}\left(\beta_0 + \beta_1 x^*, \quad \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \bar{x}_n)^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}\right]\right)$
  - $[\hat{y^*} + t_{\alpha/2, df=n-2} \cdot \mathrm{SE}(\hat{y^*}), \quad \hat{y^*} + t_{1-\alpha/2, df=n-2} \cdot \mathrm{SE}(\hat{y^*})]$
  - standard error estimated with $\hat{\sigma}^2$
- Prediction interval: interval estimate for single data
  - add extra uncertainty from ϵ!
  - variance increased by $\Longrightarrow \sigma^2 \left[1 + \frac{1}{n} + \frac{(x^* - \bar{x}_n)^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}\right]$

Interpretation

- $\hat{\beta}_1$: the average change in Y when X increase by 1 unit
  - only if we reject H_0: $\beta_1 = 0$!
  - the point estimate $\hat{\beta}_1$ can be huge but still fail to reject, why?

Interpretation

- $\hat{\beta}_1$: the average change in Y when X increase by 1 unit
  - only if we reject H_0: $\beta_1 = 0$!
  - the point estimate $\hat{\beta}_1$ can be huge but still fail to reject, why?
  - $\hat{\beta}_1$ depends on the units!
  - the t-statistic is the "standardized" slope

Interpretation

- $\hat{\beta}_1$: the average change in Y when X increase by 1 unit
  - only if we reject H_0: $\beta_1 = 0$!
  - the point estimate $\hat{\beta}_1$ can be huge but still fail to reject, why?
  - $\hat{\beta}_1$ depends on the units!
  - the t-statistic is the “standardized” slope
- $\hat{\beta}_0$: the average value of Y when X is 0
  - not meaningful if X cannot be 0 - extrapolate is dangerous!
  - can choose model with no intercept

Outline

Multiple Covariates

- A set of p covariates $X_1, X_2, \cdots X_p$ and a response variable Y

Multiple Covariates

- A set of p covariates $X_1, X_2, \cdots X_p$ and a response variable Y

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p} + \epsilon_i, \quad i = 1, 2, \cdots n$$

  - design matrix $X_{n,p}$ with n observations and p variables
  - independent Gaussian error $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma^2)$

Multiple Covariates

- A set of p covariates $X_1, X_2, \cdots X_p$ and a response variable Y

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p} + \epsilon_i, \quad i = 1, 2, \cdots n$$

  - design matrix $X_{n,p}$ with n observations and p variables
  - independent Gaussian error $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma^2)$
- The regression function is linear in the parameters $\boldsymbol{\beta}$
  - for 2 covariates, the linear function is a plane
  - for more than 2 covariates, a hyperplane

Matrix Notation

$$
\begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{pmatrix} = \begin{pmatrix} \beta_0 \\ \beta_0 \\ \cdots \\ \beta_0 \end{pmatrix} + \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots x_{1,p} \\ x_{2,1} & x_{2,2} & \cdots x_{2,p} \\ \cdots & \cdots & \cdots \\ x_{n,1} & x_{n,2} & \cdots x_{n,p} \end{pmatrix} \times \begin{pmatrix} \beta_1 \\ \beta_2 \\ \cdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \cdots \\ \epsilon_n \end{pmatrix}
$$

$$
= \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \cdots x_{1,p} \\ 1 & x_{2,1} & x_{2,2} & \cdots x_{2,p} \\ \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n,1} & x_{n,2} & \cdots x_{n,p} \end{pmatrix} \times \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \cdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \cdots \\ \epsilon_n \end{pmatrix}
$$

$$
\boldsymbol{y}_{n\times 1} = \boldsymbol{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\epsilon}_{n\times 1}
$$

Use p for number of parameters from now on.

The Coefficients

- Closed form OLS/MLE solution for β when $\boldsymbol{n} > \boldsymbol{p}$:

$$\hat{\beta} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}$$

The Coefficients

- Closed form OLS/MLE solution for β when $\boldsymbol{n} > \boldsymbol{p}$:

$$\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

- Unbiased solution for σ^2:

$$\hat{\sigma}^2 = \frac{\text{RSS}}{n - p}$$

The Coefficients

- Closed form OLS/MLE solution for β when $\boldsymbol{n} > \boldsymbol{p}$:

$$\hat{\beta} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}$$

- Unbiased solution for σ^2:

$$\hat{\sigma}^2 = \frac{\text{RSS}}{n - p}$$

- Recall multivariate normal - Gaussian distribution
  - p-dimensional random vector: mean vector and $p \times p$ covariate matrix
  - marginal of Gaussian: each dimension is normal!

Inference

- Sampling distribution for $\hat{\boldsymbol{\beta}}$ is Gaussian:

$$\hat{\boldsymbol{\beta}} \sim \text{MVN}\left(\boldsymbol{\beta}, \quad (\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^2\right)$$

Inference

- Sampling distribution for $\hat{\boldsymbol{\beta}}$ is Gaussian:

$$\hat{\boldsymbol{\beta}} \sim \text{MVN}\left(\boldsymbol{\beta}, \quad (\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^2\right)$$

- Hypothesis testing and CI for each $\hat{\beta}_j$
    - use $\hat{\sigma}^2$
    - t-test
    - t-CI

Prediction

$$\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \implies \hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\beta} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}$$

- Point estimate: conditional mean!
  - $\hat{y^*} = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x^*}] = \boldsymbol{x}^*\hat{\beta}$

Prediction

$$\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \implies \hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\beta} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}$$

- Point estimate: conditional mean!
  - $\hat{y^*} = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x^*}] = \boldsymbol{x}^*\hat{\beta}$
- Sampling distribution and CI:
  - $\hat{y^*} \sim \mathrm{N}\left(\boldsymbol{x}^*\beta, \quad \sigma^2\boldsymbol{x}^*(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}^{*T}\right)$
  - $[\hat{y^*} + t_{\alpha/2, df=n-p} \cdot \mathrm{SE}(\hat{y^*}), \quad \hat{y^*} + t_{1-\alpha/2, df=n-p} \cdot \mathrm{SE}(\hat{y^*})]$
  - standard error estimated with $\hat{\sigma}^2$

Prediction

$$\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \implies \hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\beta} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}$$

- Point estimate: conditional mean!
  - $\hat{y^*} = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x^*}] = \boldsymbol{x}^*\hat{\beta}$
- Sampling distribution and CI:
  - $\hat{y^*} \sim \mathrm{N}\left(\boldsymbol{x}^*\beta, \quad \sigma^2\boldsymbol{x}^*(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}^{*T}\right)$
  - $[\hat{y^*} + t_{\alpha/2, df=n-p} \cdot \mathrm{SE}(\hat{y^*}), \quad \hat{y^*} + t_{1-\alpha/2, df=n-p} \cdot \mathrm{SE}(\hat{y^*})]$
  - standard error estimated with $\hat{\sigma}^2$
- Prediction interval: interval estimate for single data
  - add extra uncertainty from ϵ!
  - variance increased by $\implies \sigma^2\left(\sigma^2(1 + \boldsymbol{x}^*(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}^{*T})\right)$

Interpretation

- The slopes:
  - $\hat{\beta}_j$: the average change in Y when X_j increase by 1 unit while **holding all the other covariates constant**

Interpretation

- The slopes:
  - $\hat{\beta}_j$: the average change in Y when X_j increase by 1 unit while **holding all the other covariates constant**
  - if some covariates are highly correlated, this is problematic!

Interpretation

- The slopes:
  - $\hat{\beta}_j$: the average change in Y when X_j increase by 1 unit while **holding all the other covariates constant**
  - if some covariates are highly correlated, this is problematic!
- The intercept:
  - the average value of Y when all covariates are 0
  - may not be interpretable if any covariate cannot be 0
- More covariates $\Rightarrow$ more complex and less interpretable model
  - is the goal making inference or making prediction?
  - variable/model selection

Interpretation

- Correlation vs. partial correlation
  - slope in simple regression $\propto$ correlation
  - slope in multiple regression $\propto$ partial correlation

Interpretation

- Correlation vs. partial correlation
  - slope in simple regression $\propto$ correlation
  - slope in multiple regression $\propto$ partial correlation
- The "correlation" between X_j and Y after taking out the effects of other covariates

Interpretation

- Correlation vs. partial correlation
  - slope in simple regression $\propto$ correlation
  - slope in multiple regression $\propto$ partial correlation
- The "correlation" between X_j and Y after taking out the effects of other covariates
- It is possible that X_j and Y are positively correlated but $\hat{\beta}_j$ is negative and significant!
  - example?

Model Assumptions

1. The regression function is linear!
    - linear in parameters

Model Assumptions

1. The regression function is linear!
    - linear in parameters
2. Covariates are not correlated: no perfect multicollinearity
    - interpretation!
    - X need to be full rank

Model Assumptions

1. The regression function is linear!
   - linear in parameters
2. Covariates are not correlated: no perfect multicollinearity
   - interpretation!
   - X need to be full rank
3. Gaussian errors:
   - errors have 0 mean
   - errors are independent: no auto-correlation
   - errors have constant variance: homoscedasticity
   - implies the conditional distribution of Y given X is Gaussian!

Gauss-Markov Theorem

Under the previous assumptions, the OLS solution $\hat{\beta}$ is the Best Linear Unbiased Estimate (BLUE)!

- smallest variance among unbiased estimates
  - $\mathrm{SE}(\hat{\beta}^{\mathrm{OLS}}) \leq \mathrm{SE}(\hat{\beta})$
  - in general MLE is efficient

Gauss-Markov Theorem

Under the previous assumptions, the OLS solution $\hat{\beta}$ is the Best Linear Unbiased Estimate (BLUE)!

- smallest variance among unbiased estimates
  - $\text{SE}(\hat{\beta}^{\text{OLS}}) \leq \text{SE}(\hat{\beta})$
  - in general MLE is efficient
- OLS does not necessarily have the smallest MSE!
  - recall MSE $=$ bias^2 $+$ variance
  - Ridge and Lasso estimate are both biased

Outline

Categorical Covariates

- Design matrix X is numerical, what about categorical variables?
- Example: gender, ethnicity, education level...
- Common mistake: code categories with 1, 2, 3, ...
  - what is wrong with that?

Categorical Covariates

- Design matrix X is numerical, what about categorical variables?
- Example: gender, ethnicity, education level...
- Common mistake: code categories with 1, 2, 3, ...
  - what is wrong with that?
- Correct way: dummy variable coding
  - K categories $\rightarrow K - 1$ dummy variables
  - adds $K - 1$ columns to $\boldsymbol{X}$
  - why not K dummy variables?

Categorical Covariates

Example: investigate difference in credit card balance between males and females. We create 1 dummy variable:

$$x_i = \begin{cases} 1 & \text{if ith person is female} \\ 0 & \text{if ith person is male} \end{cases}$$

Categorical Covariates

Example: investigate difference in credit card balance between males and females. We create 1 dummy variable:

$$x_i = \begin{cases} 1 & \text{if ith person is female} \\ 0 & \text{if ith person is male} \end{cases}$$

Resulting model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if ith person is female} \\ \beta_0 + \epsilon_i & \text{if ith person is male} \end{cases}$$

Interpretation?

Categorical Covariates

Results for gender model:

	Coefficient	Std. Error	t-statistic	p-value
Intercept	509.80	33.13	15.389	< 0.0001
gender [female]	19.73	46.05	0.429	0.6690

Categorical Covariates

- For the ethnicity variable we create 2 dummy variables:
- The first one could be:

$$x_{i1} = \begin{cases} 1 & \text{if ith person is Asian} \\ 0 & \text{if ith person is not Asian} \end{cases}$$

- The second one could be:

$$x_{i2} = \begin{cases} 1 & \text{if ith person is Caucasian} \\ 0 & \text{if ith person is not Caucasian} \end{cases}$$

Categorical Covariates

- Then both of these variables can be used in the regression equation:
-

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if ith person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if ith person is Caucasia} \\ \beta_0 + \epsilon_i & \text{if ith person is African} \end{cases}$$

- The category with no dummy variable – African American in this example – is the baseline .

Categorical Covariates

	Coefficient	Std. Error	t-statistic	p-value
Intercept	531.00	46.32	11.464	< 0.0001
Ethnicity [Asian]	-18.69	65.02	-0.287	0.7740
Ethnicity [Caucasian]	-12.50	56.68	-0.221	0.8260

Interaction

- Relax the additive assumption: interactions and nonlinearity
- Example: predict credit card balance using income and student status (binary)
- Model without interaction:

$$
\begin{aligned}
\text{balance}_i &= \beta_0 + \beta_1 \cdot \text{income}_i + \begin{cases} \beta_2 & \text{if } i\text{th person is a student} \\ 0 & \text{if } i\text{th person is not a student} \end{cases} \\
&= \beta_1 \cdot \text{income}_i + \begin{cases} \beta_0 + \beta_2 & \text{if } i\text{th person is a student} \\ \beta_0 & \text{if } i\text{th person is not a student} \end{cases}
\end{aligned}
$$

- Different intercept, same slope!

Interaction Term

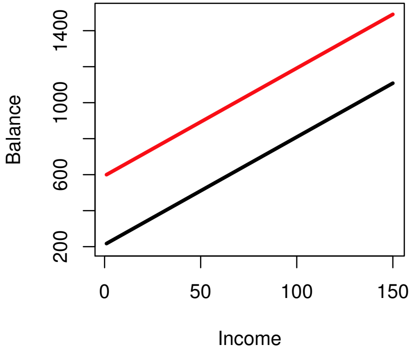
- Model with interaction:

$$\begin{aligned} \text{balance}_i &= \beta_0 + \beta_1 \cdot \text{income}_i + \beta_2 \cdot \text{student}_i + \beta_3 \cdot \text{income}_i \times \text{student}_i \\ &= \beta_0 + \beta_1 \cdot \text{income}_i + \begin{cases} \beta_2 + \beta_3 \cdot \text{income}_i & \text{if student} \\ 0 & \text{if not student} \end{cases} \\ &= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \cdot \text{income}_i & \text{if student} \\ \beta_0 + \beta_1 \cdot \text{income}_i & \text{if not student} \end{cases} \end{aligned}$$

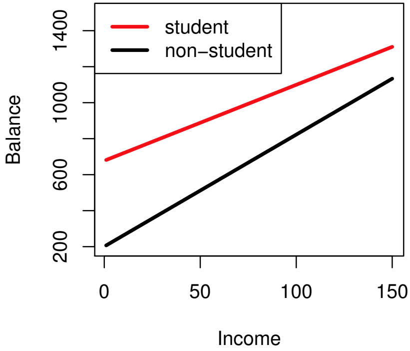
- Different intercept, different slope!

Interaction Term

- no interaction between income and student



- with an interaction between income and student



Polynomial Term

- Polynomial regression on Auto data:

$$mpg = \beta_0 + \beta_1 \cdot horsepower + \beta_2 \cdot horsepower^2 + \epsilon$$

Polynomial Term

- Polynomial regression on Auto data:

$$mpg = \beta_0 + \beta_1 \cdot horsepower + \beta_2 \cdot horsepower^2 + \epsilon$$

- Results may provide a better fit than just linear terms:

	Coefficient	Std. Error	t-statistic	p-value
Intercept	56.9001	1.8004	31.6	< 0.0001
horsepower	-0.4662	0.0311	-15.0	< 0.0001
horsepower^2	0.0012	0.0001	10.1	< 0.0001

Polynomial Term

- Polynomial regression on Auto data:

$$mpg = \beta_0 + \beta_1 \cdot horsepower + \beta_2 \cdot horsepower^2 + \epsilon$$

- Results may provide a better fit than just linear terms:

	Coefficient	Std. Error	t-statistic	p-value
Intercept	56.9001	1.8004	31.6	< 0.0001
horsepower	-0.4662	0.0311	-15.0	< 0.0001
horsepower^2	0.0012	0.0001	10.1	< 0.0001

- But coefficients lose interpretability - why?

Model Fit

How do we check if the model fits the data well?

- RSS is unit dependent and minimized by definition of OLS
    - more for diagnostic

Model Fit

How do we check if the model fits the data well?

- RSS is unit dependent and minimized by definition of OLS
  - more for diagnostic
- R^2: percentage of variation in response Y that is explained by the model
  - in simple linear regression, $R^2 = r^2_{X,Y}$!
  - always increase with more covariates, why?

Model Fit

How do we check if the model fits the data well?

- RSS is unit dependent and minimized by definition of OLS
  - more for diagnostic
- R^2: percentage of variation in response Y that is explained by the model
  - in simple linear regression, $R^2 = r^2_{X,Y}$!
  - always increase with more covariates, why?
- F-statistic and F-test
  - is at least one covariate useful? (better than null model)
  - $F = \frac{(TSS-RSS)/p}{RSS/(n-p-1)} \sim F_{p,n-p-1}$

Model Fit

How do we check if the model fits the data well?

- RSS is unit dependent and minimized by definition of OLS
  - more for diagnostic
- R^2: percentage of variation in response Y that is explained by the model
  - in simple linear regression, $R^2 = r^2_{X,Y}$!
  - always increase with more covariates, why?
- F-statistic and F-test
  - is at least one covariate useful? (better than null model)
  - $F = \frac{(TSS-RSS)/p}{RSS/(n-p-1)} \sim F_{p,n-p-1}$

Model Comparison

- Adjusted R^2
  - R^2 + model complexity penalty (number of covariates)
  - no longer blindly favor large model like R^2

Model Comparison

- Adjusted R^2
  - R^2 + model complexity penalty (number of covariates)
  - no longer blindly favor large model like R^2
- Akaike information criterion (AIC)
  - -2 log likelihood + $2p$

Model Comparison

- Adjusted R^2
  - R^2 + model complexity penalty (number of covariates)
  - no longer blindly favor large model like R^2
- Akaike information criterion (AIC)
  - -2 log likelihood + $2p$
- Bayesian information criterion (BIC)
  - -2 log likelihood + $p(\log n)$
  - penalize large model more

Model Diagnostic - Residual Plot

Always check residual plot!!!

Model Diagnostic - Residual Plot

Always check residual plot!!!

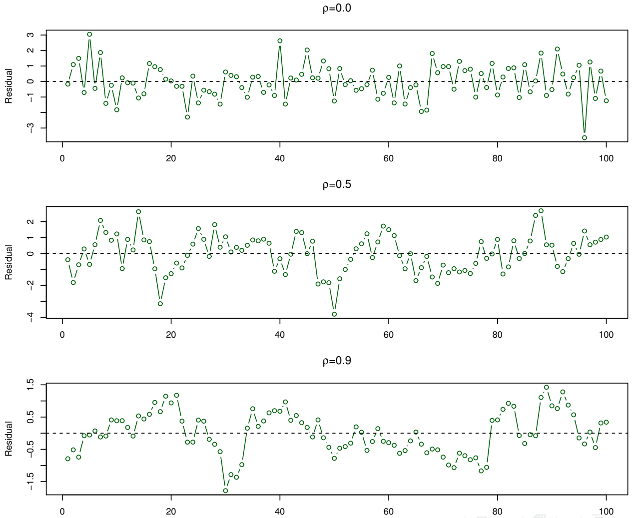
Good:

- scattered around 0
- no clear pattern
- normal distribution

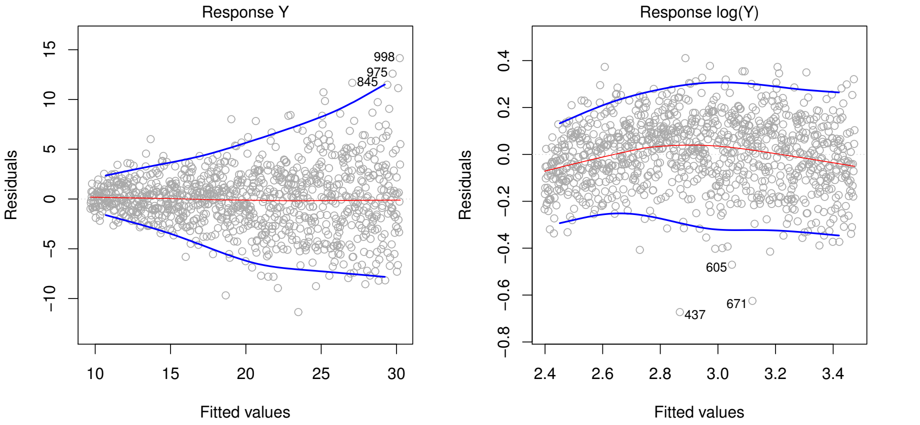
Bad:

- auto-correlation
- heteroscedasticity
- non-linearity

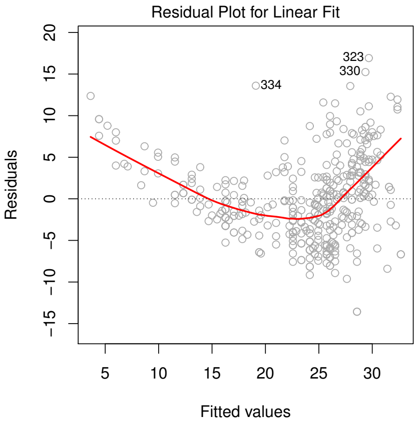
Residual Plot

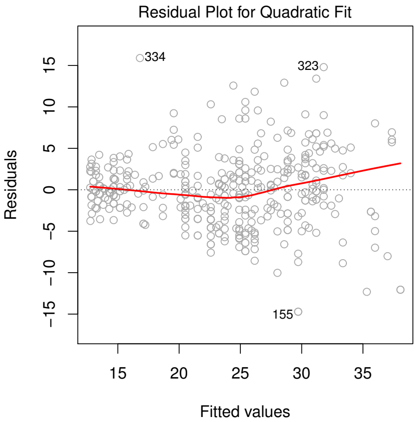


Residual Plot



Residual Plot





Model Diagnostic - Multicollinearity

- Refers to two or more covariates highly correlated
- Cause issues
  - numerical stability
  - coefficients SE inflated
  - interpretability
- Identify:
  - scatter plot, correlation matrix
  - condition number, variance inflation factor (VIF)

Model Diagnostic - Multicollinearity

- Refers to two or more covariates highly correlated
- Cause issues
  - numerical stability
  - coefficients SE inflated
  - interpretability
- Identify:
  - scatter plot, correlation matrix
  - condition number, variance inflation factor (VIF)
- Solution:
  - variable selection
  - feature engineering

Outline