Linear Regression

FA590 Statistical Learning in Finance

Dr. Zonghao Yang

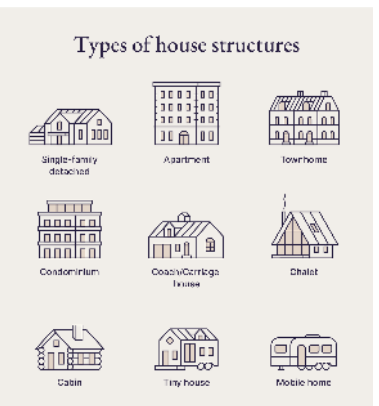
Stevens Institute of Technology

Learning Objectives

- Understand and apply linear regression to predict outcomes, including interpreting coefficients, residuals, and model performance metrics
- Perform hypothesis testing on regression models and evaluate the significance of predictors
- Extend linear regression with generalizations such as regularization, feature interaction, and nonlinear transformations
- Utilize model selection techniques to improve prediction accuracy and avoid overfitting

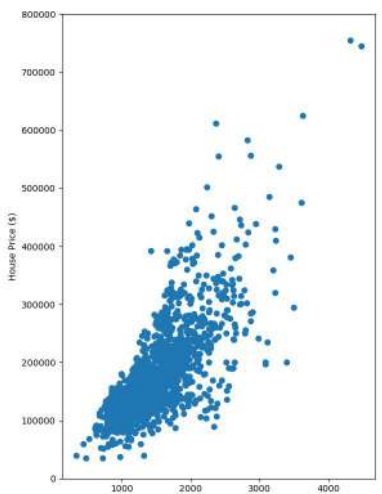
Motivation: House Prices

- House prices dataset from Kaggle
- House prices and 79 predictors
  - Structural and size: Total basement area, living area, full bathrooms, size of garage, etc.
  - Quality and condition features: Overall material and finish quality of the house, original construction date, kitchen quality, etc.
  - Location: The general zoning classification (e.g., agriculture, commercial, and residential), neighborhood, etc.
  - House styles: Single-family detached, townhouse, apartment, etc.

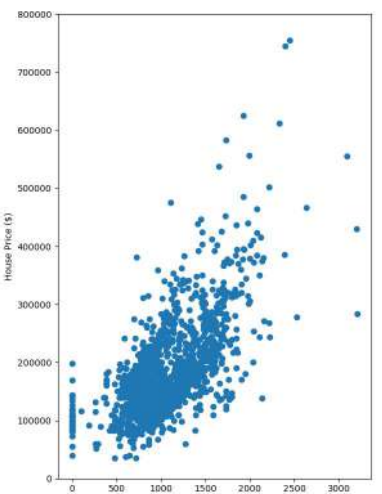


MOTIVATION: HOUSE PRICES

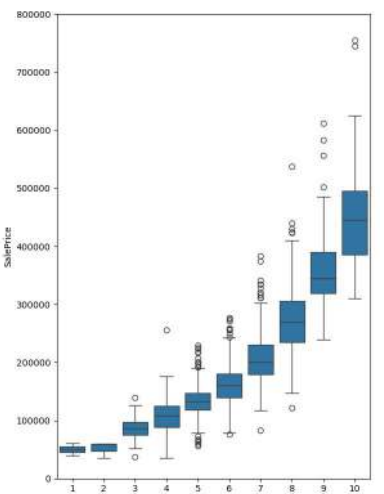
Figure. Predicting House Prices?



(a) Total Living Area



(b) Total Basement Area



(c) Overall Quality

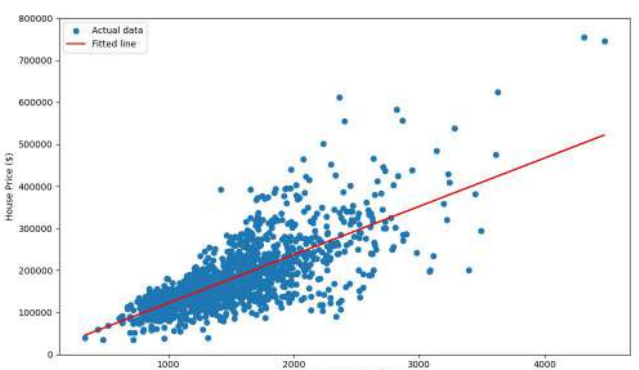
Motivation: House Prices

- Is there a relationship between the number of full bathrooms and house prices?
- How strong is the relationship between the number of full bathrooms and house prices?
- Which house styles are associated with house prices?
- How large is the association between each house style and house prices?
- How accurately can we predict house prices?

Linear regression can be used to answer each of these questions

Linear Regression

- Used for curve fitting: Tell a computer how to draw a line through a scatterplot
- A probabilistic framework for optimal prediction
- A starting point to build intuition before building more complex models



From Linear Regression to Statistical Learning

- Most core concepts of machine learning (e.g. in-sample vs. out-of-sample, performance evaluation) can be understood with the comparatively simple example of linear regression and logistic regression
- We thus first discuss linear regression, and then generalize the concepts
- Moreover, despite its simplicity (or especially because of it), linear regression is still a very potent and often used method

Part I

Simple Linear Regression

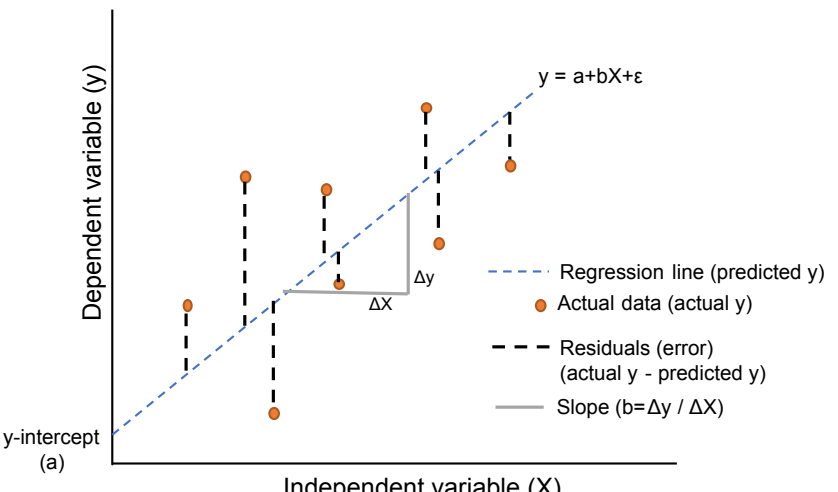
Simple Linear Regression

Fit a linear relationship between a numerical outcome variable Y and a predictor X

$$Y = \beta_0 + \beta_1 X + \epsilon, \quad \epsilon \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

- In statistics, we typically call X the independent variable or the covariate , and Y the dependent variable or the response
- In machine learning, we typically call X the feature , and Y the target
- β_0 and β_1 are two unknown constants, also known as coefficients or parameters
- β_0: Intercept ; β_1: slope
- ϵ: Error term, noise, residual

Simple Linear Regression: An Illustration



A Probalistic Model

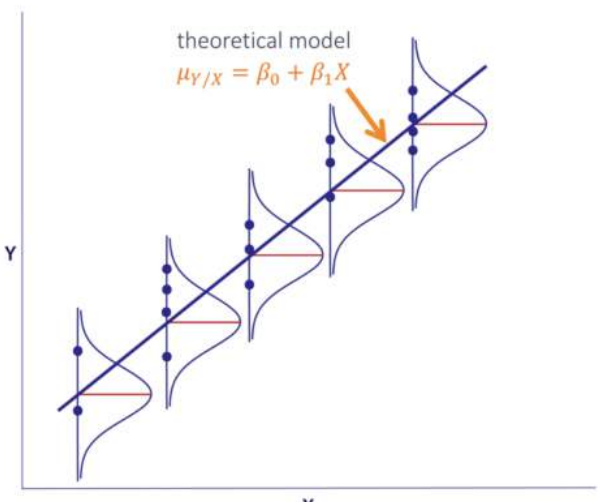
$$Y = \beta_0 + \beta_1 X + \epsilon, \quad \epsilon \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

- X is not random
- ϵ is random noise, and so is Y

$$\begin{aligned}\mathbb{E}[Y] &= \mathbb{E}[\beta_0 + \beta_1 X + \epsilon] \\ &= \mathbb{E}[\beta_0 + \beta_1 X] + \mathbb{E}[\epsilon] \\ &= \beta_0 + \beta_1 X\end{aligned}$$

$$\begin{aligned}Var[Y] &= Var[\beta_0 + \beta_1 X + \epsilon] \\ &= 0 + Var[\epsilon] \\ &= \sigma^2\end{aligned}$$

$$Y \sim N(\beta_0 + \beta_1 X, \sigma^2)$$



Curve Fitting

- N pairs of observations: (x_1, y_1), (x_2, y_2),...,(x_N, y_N)
- Fit a line

$$y_i = \beta_0 + \beta_1 x_i$$

- Minimize **residual sum of squares** (RSS)

$$RSS = \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i)^2$$

- **Least squares** , also known as, **ordinary least squares** (OLS)
- **Quadratic loss**

Curve Fitting

- Least-squares fitted parameter:

$$\hat{\beta}_0, \hat{\beta}_1$$

- Fitted values

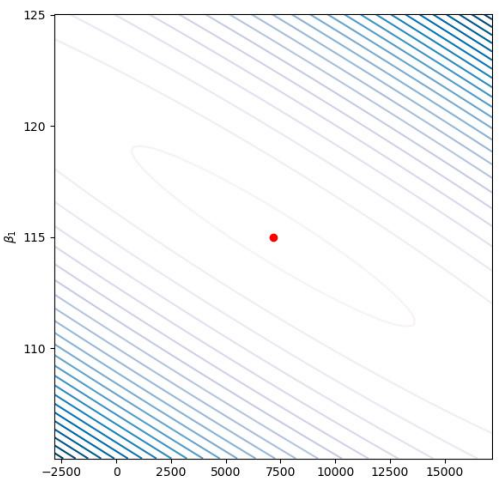
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, i = 1, \ldots, N$$

- **Residuals**

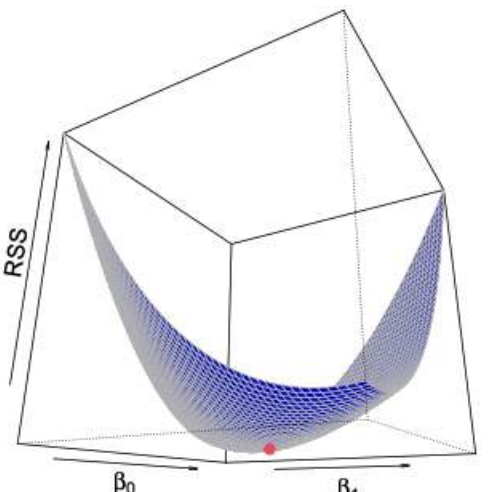
$$e_i = y_i - \hat{y}_i$$

- “Hats" denote fitted values

VISUALIZATION: OBJECTIVE FUNCTION



(a) Contour



(b) Three-dimensional

Estimating the Coefficients

$$\hat{\beta}_0, \hat{\beta}_1 = \underset{\beta_0, \beta_1}{\arg\min} \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i)^2$$

- Ordinary least squares (OLS) estimators

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2}$$

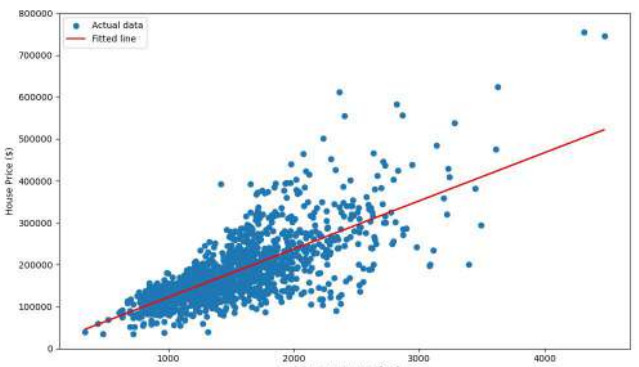
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- People prefer least squares over “least absolute values" because the square function is differentiable
- **Interpretation** of the slope β_1 (or $\hat{\beta}_1$): For every 1 unit increase in x, the average of the response variable increases (or decreases) by β_1 units

Derivation of OLS Estimators

EXAMPLE: HOUSE PRICES DATA

Figure. House Price ($) vs. Living Area (square feet)



$$\text{House Price} = 7169.01 + 115.04 \times \text{Living Area}$$

"For every 1 square feet increase in living area, the average of the house price increases by $115.04."

Accuracy of the Coefficient Estimates

- The standard error of an estimator reflects how it varies under repeated sampling

$$\text{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

$$\text{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{N} + \frac{\bar{x}^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \right]$$

  where $\sigma^2 = Var(\epsilon)$
- In practice, σ^2 is often unknown. Residual standard error (RSE) is an estimate of the standard deviation of ϵ

$$RSE = \sqrt{\frac{1}{N-2}RSS} = \sqrt{\frac{1}{N-2}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

- 2 is the degrees of freedom lost in estimating the regression coefficient (β_0 and β_1)

Accuracy of the Coefficient Estimates

- Confidence interval: e.g., 95% confidence interval

$$\hat{\beta}_0 : \quad \left[\hat{\beta}_0 - 2 \cdot SE(\hat{\beta}_0), \hat{\beta}_0 + 2 \cdot SE(\hat{\beta}_0)\right]$$

$$\hat{\beta}_1 : \quad \left[\hat{\beta}_1 - 2 \cdot SE(\hat{\beta}_1), \hat{\beta}_1 + 2 \cdot SE(\hat{\beta}_1)\right]$$

- **Interpretation** There is approximately a 95% chance that the interval will contain the true value of β_0 or β_1

Hypothesis Testing

- We are interested in β_1 in hypothesis testing

  H_0: There is no relationship between X and Y
  H_a: There is some relationship between X and Y

- In math notation

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

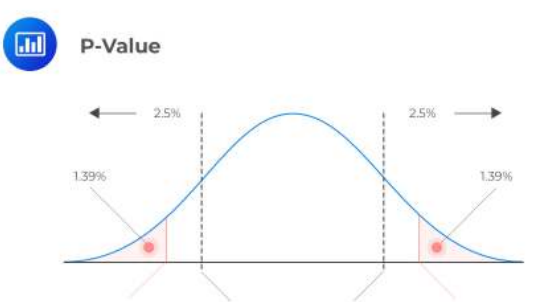
since if $\beta_1 = 0$ then the model reduces to $Y = \beta_0 + \epsilon$, and X is not associated with Y

Hypothesis Testing

- To test the null hypothesis, we compute a t-statistic:

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} \sim t(N-2)$$

  where $t(N-2)$ is the t-distribution with $N-2$ degrees of freedom
- Using statistical software, it is easy to compute the probability of observing any value equal to $|t|$ or larger. We call this probability the **p-value**



Results for the House Prices Data

Table. Regression Results

	Coefficient	Std. Error	t-statistic	p-value
Intercept	7169.01	4434.46	1.62	0.106
Living area	115.04	2.78	41.34	< 0.0001

- The intercept term is 7169.01, meaning fixing the living area at zero, the average house price is $7169.01
- The p-value for the t-test is 0.106, suggesting that we cannot reject the null hypothesis that the intercept term is zero
- The coefficient for the feature living area is 115.04, and statistically significant at the 1% level
- One square feet increase in living area leads to a $115.04 increase in house price

Model Evaluation

- **Residual sum of squares** (RSS): Optimization objective

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

- **Mean squared error** (MSE)

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

- **Residual standard of error** (RSE)

$$RSE = \sqrt{\frac{1}{N-2} RSS}$$

Model Evaluation

- R-squared (R^2): The proportion of variance explained

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

  where $TSS = \sum_{i=1}^{N}(y_i - \bar{y})^2$ is the total sum of squares
- $R^2 \in [0, 1]$ in sample
- **Interpretation** R^2 represents the proportion of variability in Y that can be explained using X

Results for the House Prices Data

$$\text{House Price} = 7169.01 + 115.04 \times \text{Living Area}$$

- The residual standard error is 53942.76
  - RSE is measured in the units of Y, so it is not always clear what constitutes a good RSE
  - Meaningful when comparing across models
- The R^2 is 54.0%
  - R^2 is independent of the scale of Y
  - 54.0% of the variability in house prices can be explained by the total living area of the house
  - In other words, there is still a large degree of variability in house prices that is unexplained

Part II

Multiple Linear Regression

Motivation

- We have shown that the total living area predicts house prices, which explains 54.0% of the variability in the response.
- What other variables should we include in the model?
  - Structural and size: Total basement area, living area, full bathrooms, size of garage, etc.
  - Quality and condition features: Overall material and finish quality of the house, original construction date, kitchen quality, etc.
  - Location: The general zoning classification (e.g., agriculture, commercial, and residential), neighborhood, etc.
  - House styles: Single-family detached, townhouse, apartment, etc.
- How to incorporate these variables in the regression model?

Multiple Linear Regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

- Optimization objective

$$\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p = \underset{\beta_0,\ldots,\beta_p}{\arg\min} \sum_{i=0}^{N} (y_i - \beta_0 - \beta_1 x_{i,1} - \cdots - \beta_p x_{i,p})^2$$

- Fitted hyperplane

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p$$

- Prediction given $X = (x_{i,1}, x_{i,2}, \ldots, x_{i,p})$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \hat{\beta}_2 x_{i,2} + \cdots + \hat{\beta}_p x_{i,p}$$

VISUALIZATION: CURVE FITTING

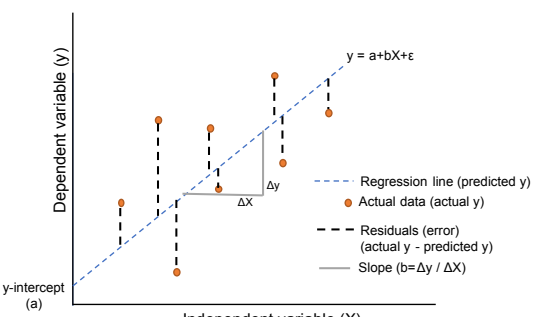


Figure. Simple Linear Regression

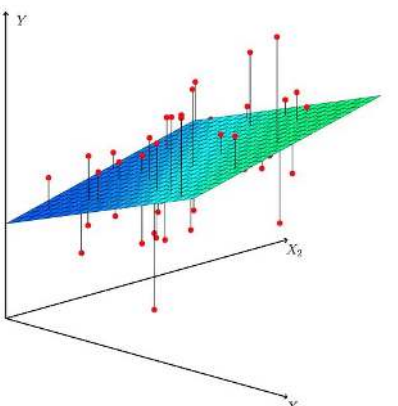


Figure. Multiple Linear Regression

Matrix Notation

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N,1} & x_{N,2} & \cdots & x_{N,p} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{pmatrix}$$

Matrix Review: Special Matrix

- The identity matrix

$$\mathbf{I}_n = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

- For any matrix $\mathbf{A}$,

$$\mathbf{I} \cdot \mathbf{A} = \mathbf{A} \text{ and } \mathbf{A} \cdot \mathbf{I} = \mathbf{A}$$

- The one-vector and one-matrix

$$\mathbf{1}_n = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \quad \mathbf{J}_n = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix}$$

Matrix Review: Matrix Operations

- Transpose

$$\mathbf{A}'_{i,j} = \mathbf{A}_{j,i}$$

- Addition: Matrices $\mathbf{A}$ and $\mathbf{B}$ with the same size $n \times m$

$$(\mathbf{A} + \mathbf{B})_{i,j} = \mathbf{A}_{i,j} + \mathbf{B}_{i,j}$$

- Multiplication: Matrix $\mathbf{A}$ of size $n \times m$; matrix $\mathbf{B}$ of size $m \times p$

$$(\mathbf{AB})_{i,j} = \sum_{k=1}^{m} \mathbf{A}_{i,k}\mathbf{B}_{k,j}$$

- Inversion: Non-singular $\mathbf{A}$ of size $n \times n$, $\mathbf{A}^{-1}$ satisfies

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}_n$$

DGP IN MATRIX FORM

- Original form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

- Matrix notation

$$\mathbf{y} = \mathbf{X}\beta + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I)$$

  where the matrix $\mathbf{X}$ is also called the design matrix

- Equivalently

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N,1} & x_{N,2} & \cdots & x_{N,p} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{pmatrix}$$

$$\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix} \right)$$

OLS Estimator in Matrix Form

- The least squares estimator solves

$$\min_{\beta_0,\ldots,\beta_p} \sum_{i=0}^{N} (y_i - \beta_0 - \beta_1 x_{i,1} - \cdots - \beta_p x_{i,p})^2$$

- In matrix notation

$$\min_{\beta} (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)$$

- The solution is then

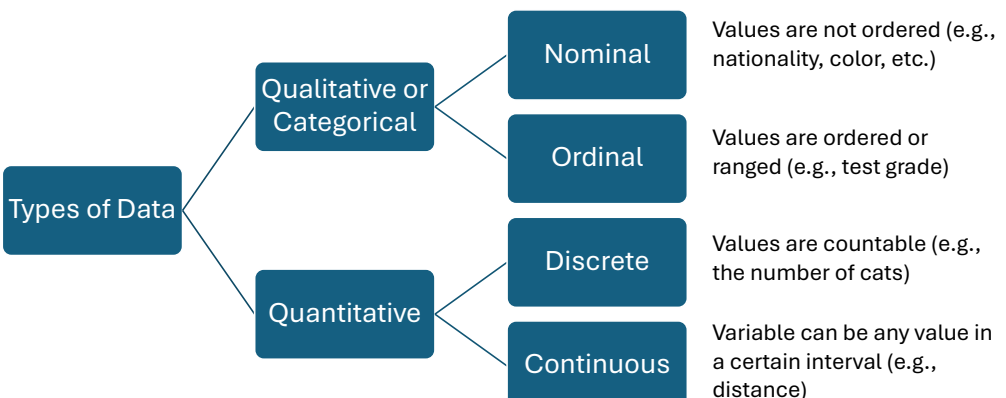
$$\hat{\beta}_{\text{OLS}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Interpretation of the Coefficients

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

- β_0 captures the average effect on Y
- β_j is the average effect on Y of a one unit increase in X_j, holding all other predictors fixed
- "For every one unit increase in covariate X_j, the mean of the response variable increases (or decreases) by β_j units when holding all other covariates constant"
- If the predictors are standardized, β_i are comparable
- Regression only captures the correlation between the predictive target and the predictors, not the causal relationship

Types of Predictors



Qualitative Predictors with Two Levels

- **Binary** or **dummy** variable
- Example: whether the house is remodeled
- 0/1 coding scheme

$$x_i = \begin{cases} 1 & \text{if } i\text{th house is remodeled} \\ 0 & \text{if } i\text{th house is not remodeled} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th house is remodeled} \\ \beta_0 + \epsilon_i & \text{if } i\text{th house is not remodeled} \end{cases}$$

Qualitative Predictors with More Than Two Levels

- Nominal variable: Values are not ordered
  - Example: House styles, including single-family detached, townhouse, apartment, etc.
- **One-hot Encoding** Suppose k categories, create $k-1$ dummy variables

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th house is single-family detached} \\ 0 & \text{if } i\text{th house is not single-family detached} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th house is townhouse} \\ 0 & \text{if } i\text{th house is not townhouse} \end{cases}$$

$$x_{i3} = \cdots$$

- Why $k-1$? Multicolinearity

Qualitative Predictors with More Than Two Levels

- Ordinal variable: Values are ordered or ranged
  - Example: General shape of property, including regular, slightly irregular, moderately irregular, and irregular
- **Label Encoding** Each unique category value is assigned a unique integer based on alphabetical or numerical ordering

$$x_i = \begin{cases} 1 & \text{Regular} \\ 2 & \text{Slightly irregular} \\ 3 & \text{Moderatedly irregular} \\ 4 & \text{Irregular} \end{cases}$$

Hypothesis Testing

- Is at least one of the predictors $X_1, X_2, \ldots, X_p$ useful in predicting the response?

$$\begin{aligned} H_0 &: \quad \beta_1 = \beta_2 = \cdots = \beta_p = 0 \\ H_a &: \quad \text{at least one } \beta_j \text{ is non-zero} \end{aligned}$$

- This hypothesis test compares the following two regressions

$$Y = \beta_0 + \epsilon$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

- **F-statistics** examines whether the regression model as a whole provides a better fit to the data than a simple model with no predictors

Analysis of Variance (ANOVA)

Table. ANOVA Table for Regression

Source of Variation	Sum of Squares	Degrees of Freedom
Total	$TSS = \sum(y_i - \bar{y})^2$	$N - 1$
Residual	$RSS = \sum(y_i - \hat{y}_i)^2$	$N - p - 1$
Model	$TSS - RSS$	$(N - 1) - (N - p - 1) = p$

- The **ANOVA table** summarizes the partitioning of the total variability (TSS) into variation explained by the model and the unexplained variation (RSS)
- **F-statistics** compares the variation explained by the regression model and the remaining variation

$$F = \frac{(TSS - RSS)/p}{RSS/(N - p - 1)} \sim F_{p,N-p-1}$$

F-STATISTICS

- **F-statistics**

$$F = \frac{(TSS - RSS)/p}{RSS/(N - p - 1)} \sim F_{p,N-p-1}$$

- Large F-statistic: This suggests that the model explains significantly more variance than a model with no predictors
- Small F-statistic: If the F-statistic is close to 1 or smaller, it suggests that the independent variables do not explain much more of the variability than random noise, and we fail to reject the null hypothesis
- How large does the F-statistic need to be before we can reject H_0?
  - Depends on N and p
  - *Python* and *R* provide predefined functions to compute the p-value

Model Evaluation

- Mean squared error: $MSE = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2/n$
- R-squared: $R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$
- The issues with the above metrics
  - Overestimate the predictive performance of the models; e.g., training MSE is likely lower than test MSE
  - Not account for the number of predictors used in the model
- We need metrics that adjust the error for the model size

Model Evaluation: Adjusted R-Squared

- R-squared

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

- **Adjusted R-squared**

$$\bar{R}^2 = 1 - \frac{\frac{1}{N-p}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

  where p is the number of predictors
- R^2 always increases as more predictors are added (regardless of whether the additional predictors actually improve the model)
- Adjusted R^2 adjusts for the number of predictors and penalizes overfitting by considering the degrees of freedom

In-Sample and Out-of-Sample R-Squared

- In-sample R-squared

$$\bar{R}^2 = 1 - \frac{\frac{1}{N-p}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

- Out-of-sample R-squared

$$\bar{R}^2 = 1 - \frac{\frac{1}{N-p}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y}_{\text{train}})^2}$$

  where $\bar{y}_{\text{train}}$ is the average if y's in the training set
- In stock return prediction, we often set $\bar{y}_{\text{train}} = 0$

Feature Selection: Forward Stepwise Selection

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors.
2. For $k = 0, \ldots, p-1$:
   (a) Consider all $p - k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
   (b) Choose the *best* among these $p - k$ models, and call it $\mathcal{M}_{k+1}$. Here *best* is defined as having smallest MSE or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using the prediction error, MSE or R^2 on a validation set.

Feature Selection: Backward Stepwise Selection

1. Let $\mathcal{M}_p$ denote the *full model*, which contains all p predictors.
2. For $k = p, p-1, \ldots, 1$:
   (a) Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$, for a total of $k-1$ predictors.
   (b) Choose the *best* among these k models, and call it $\mathcal{M}_{k-1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using the prediction error, MSE or R^2 on a validation set.

Return to Motivation Questions

- Is there a relationship between the number of full bathrooms and house prices?

Return to Motivation Questions

- Is there a relationship between the number of full bathrooms and house prices?
  - Fit a linear regression model where house price is the dependent variable and the number of full bathrooms is the independent variable
  - Hypothesis testing using t-test on the coefficient

Return to Motivation Questions

- Is there a relationship between the number of full bathrooms and house prices?
  - Fit a linear regression model where house price is the dependent variable and the number of full bathrooms is the independent variable
  - Hypothesis testing using t-test on the coefficient
- How strong is the relationship between the number of full bathrooms and house prices?

Return to Motivation Questions

- Is there a relationship between the number of full bathrooms and house prices?
  - Fit a linear regression model where house price is the dependent variable and the number of full bathrooms is the independent variable
  - Hypothesis testing using t-test on the coefficient
- How strong is the relationship between the number of full bathrooms and house prices?
  - R^2 of the regression

Return to Motivation Questions

- Which house styles are associated with house prices?

Return to Motivation Questions

- Which house styles are associated with house prices?
  - Create dummy variables to represent the house styles and run multiple linear regression
  - Hypothesis testing using t-test on the coefficient for each dummy variable

Return to Motivation Questions

- Which house styles are associated with house prices?
  - Create dummy variables to represent the house styles and run multiple linear regression
  - Hypothesis testing using t-test on the coefficient for each dummy variable
- How large is the association between each house style and house prices?

Return to Motivation Questions

- Which house styles are associated with house prices?
  - Create dummy variables to represent the house styles and run multiple linear regression
  - Hypothesis testing using t-test on the coefficient for each dummy variable
- How large is the association between each house style and house prices?
  - Compare the magnitude and sign of the coefficients

Return to Motivation Questions

- Which house styles are associated with house prices?
  - Create dummy variables to represent the house styles and run multiple linear regression
  - Hypothesis testing using t-test on the coefficient for each dummy variable
- How large is the association between each house style and house prices?
  - Compare the magnitude and sign of the coefficients
- How accurately can we predict house prices?

Return to Motivation Questions

- Which house styles are associated with house prices?
  - Create dummy variables to represent the house styles and run multiple linear regression
  - Hypothesis testing using t-test on the coefficient for each dummy variable
- How large is the association between each house style and house prices?
  - Compare the magnitude and sign of the coefficients
- How accurately can we predict house prices?
  - R^2, MSE, etc.

Part III

Extensions and Generalization

Extensions of Linear Regression

- Interaction terms: Capture the synergy effect between two predictors
  - Example: Overall Material and Finish Quality and Living Area Square Feet; Year the House was Built and Neighborhood
  - The standard regression with two variables X_1 and X_2

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

  - Extending this model by including a third interaction term

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + + \beta_3 X_1 X_2 + \epsilon$$

- Nonlinear relationships: e.g., quadratic terms
  - Example: Overall Material and Finish Quality; Living Area Square Feet;

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \epsilon$$

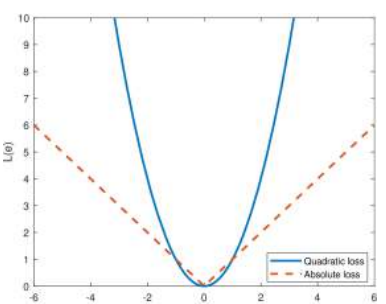
Extensions of Linear Regression: Loss Functions

- **Root mean squared error**

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} e_i^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

- **Mean absolute error**

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |e_i| = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i|$$



- RMSE punishes large errors more than MAE
- For certain objective functions, e.g., MSE, one can calcucate the optimal parameters analytically

Generalizations of the Linear Model

- Classification problems: The predictive target is not quantitative
  - Logistic regression, support vector machines
- Non-linearity and interaction: The relationship between independent variable and the predictors are non-linear
  - Kernel methods, tree-based methods, neural networks
- Regularized fitting: Ridge, Lasso, and Elastic Net

Regularization: Motivation

- **Multi-collinearity Problem** When two independent variables are highly correlated, the estimated coefficients can become highly variable
- Build intuition: Consider the extreme case where two variables are perfectly correlated
  - Regress a variable Y against two variables A and B: $Y = \alpha A + \beta B + \varepsilon$ where α, β are the regression coefficients to be determined and ε is residual noise
  - Assume that A and B are perfectly correlated: $A = \eta B$ for some constant η. Then, we can write

$$Y = (\alpha\eta + \beta)B + \varepsilon$$

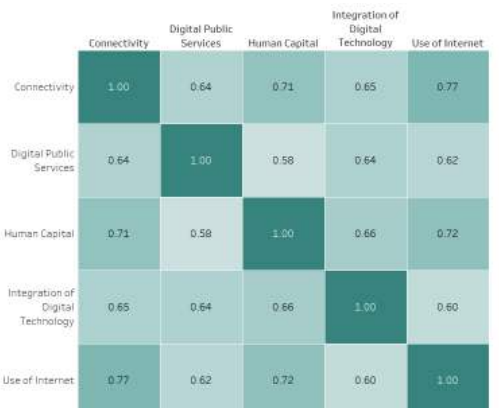
  - $\tilde{\beta} = \alpha\eta + \beta$ is uniquely determined via optimization
  - However, there is still an infinite number of solutions to α and β

Regularization: Motivation

- In practical situations, highly correlated regression variables lead to badly defined fitting problems
  - Small changes in the data may lead to large changes in the regression coefficients
  - Large estimated coefficients
- In other words, the regression result is unstable and does not generalize well
- In one-hot encoding , including all the dummy variables will lead to the multi-collinearity problem
  - Summing all the dummy variables for a categorical feature will result in a vector of ones, which is perfectly correlated with the intercept term

Address the Multi-Collinearity Problem

- Remove highly correlated predictors: Correlation matrix



- Regularization: Ridge, Lasso, and Elastic Net regression

Regularization

- Regularization: Address multi-colinearity by punishing large coefficients
- Without regularization:

$$\hat{\beta}_0, \ldots, \hat{\beta}_p = \arg\min \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

- With regularization:

$$\hat{\beta}_0, \ldots, \hat{\beta}_p = \arg\min \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

where λ is a constant specifying the relative weighting of the two objective, which controls the degree of punishment

- This is called **Ridge regression**

Regularization

- **Ridge regression** (or L2 regularization)

$$\hat{\beta}_0, \ldots, \hat{\beta}_p = \arg\min \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

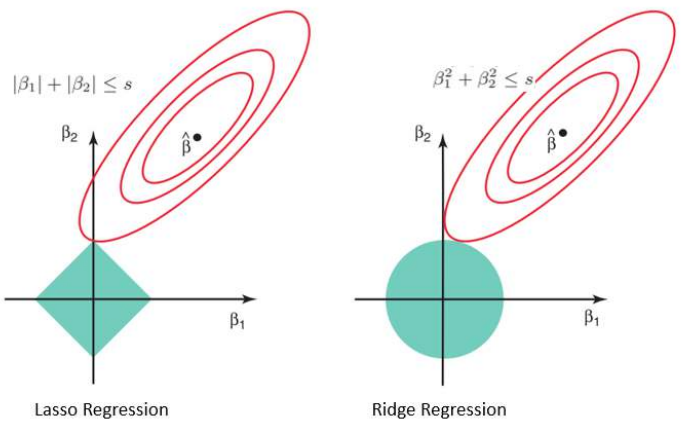
- **Lasso regression** (or L1 regularization)

$$\hat{\beta}_0, \ldots, \hat{\beta}_p = \arg\min \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 + \omega \sum_{j=1}^{p} |\beta_j|$$

- **Elastic net** A linear combination of the Ridge and Lasso regularization

$$\hat{\beta}_0, \ldots, \hat{\beta}_p = \arg\min \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 + \omega \sum_{j=1}^{p} |\beta_j|$$

LASSO VS. RIDGE



- Lasso actively encourages setting coefficients to zero
- For this reason, Lasso is commonly used for feature selection
- Ridge set the coefficients to arbitrarily small values, but not zero

Parsimony Principle

- **Parsimony Principle** Among competing models that explain the data equally well, the simplest model, i.e., the one with fewer parameters or assumptions, should be preferred
- Rationale
  - Simpler models are less likely to overfit the training data, making them more generalizable to unseen data
  - Overly complex models may capture noise and irrelevant details, leading to poor performance on new datasets
- The importance of simplicity in model selection