Welcome to DSCI 101

Introduction to Data Science

Week 9-10 Recap

- Population vs. sample and parameters vs. statistics
  - sampling distribution of a statistic
  - point estimate and confidence interval

- Simulation based statistical inference:
  - generate many random samples from population
  - generate Bootstrap resamples from one original random sample

- Hypothesis testing – apply statistical inference to decision making
  - general framework and rational
  - Permutation test – A/B testing

Week 11 Preview

- Linear regression with numerical response
  - correlation and simple linear regression
  - multiple linear regression
  - categorical predictors
  - interpret regression coefficients

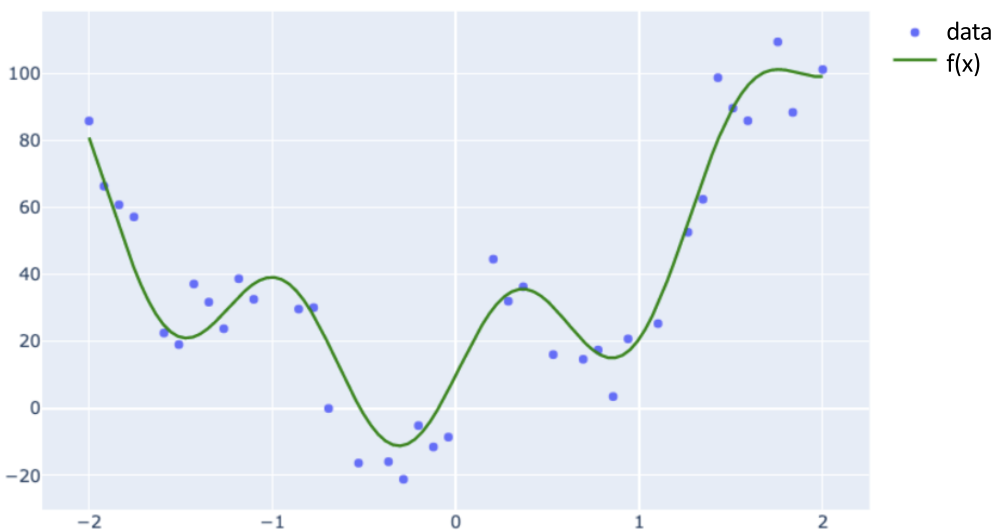
- Logistic regression with binary response
  - logit and logistic function
  - interpret logistic coefficients as log odds ratio

What is Regression?

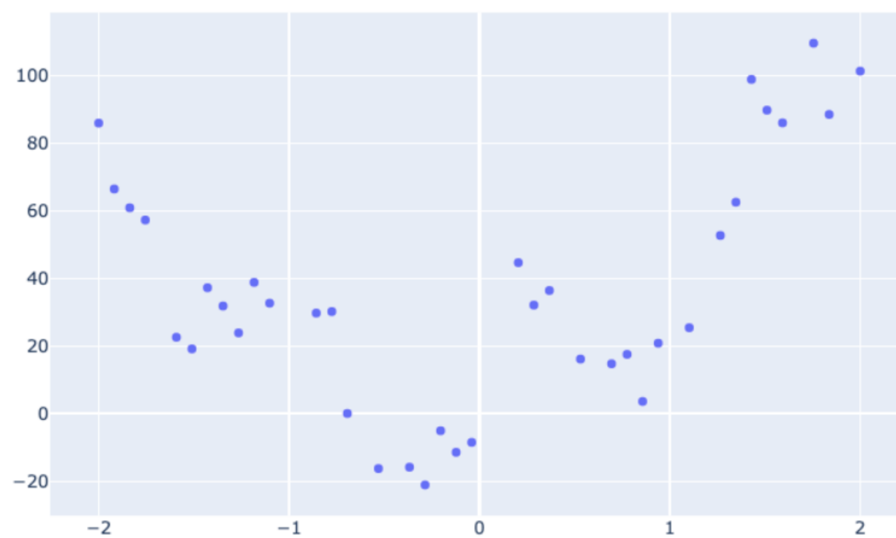
- A set of statistical processes for **estimating the relation between a dependent variable Y** (outcome or response) **and one or more independent variables X** (predictors, covariates, or features).
- The goal is to estimate the relation using a regression function f(X) from observed data:
  - Y = f(X) + **a random error**
- What is a regression function good for?
  - Prediction: use X to predict Y
  - Inference: generalize relation between X and Y to a larger population

If there is a Data God

- for each value of x_i
- true regression function = f(x_i)
- random noise = ε_i
- response y_i = f(x_i) +ε_i

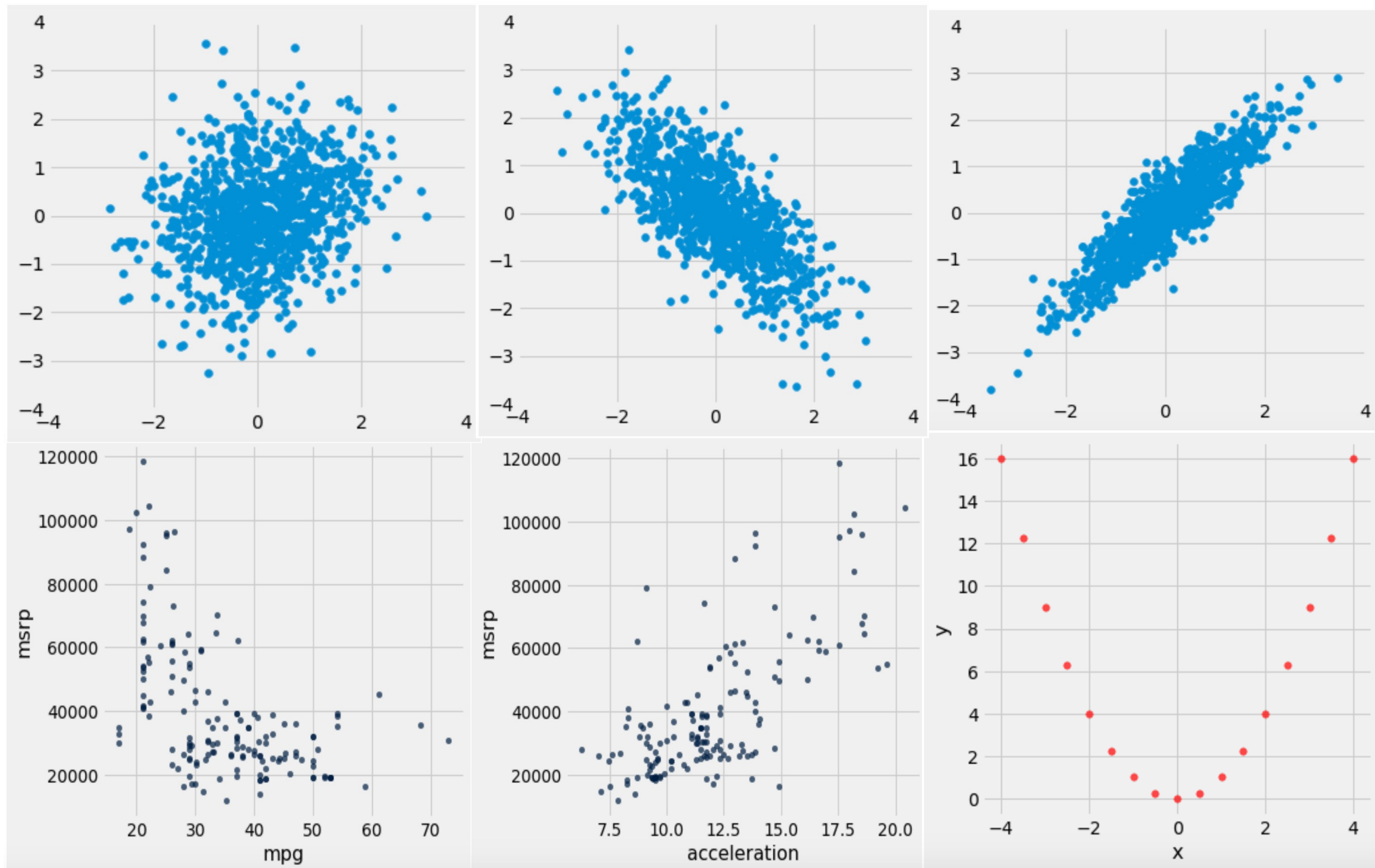


we only see data (x_i, y_i)



Describe scatter plot relations

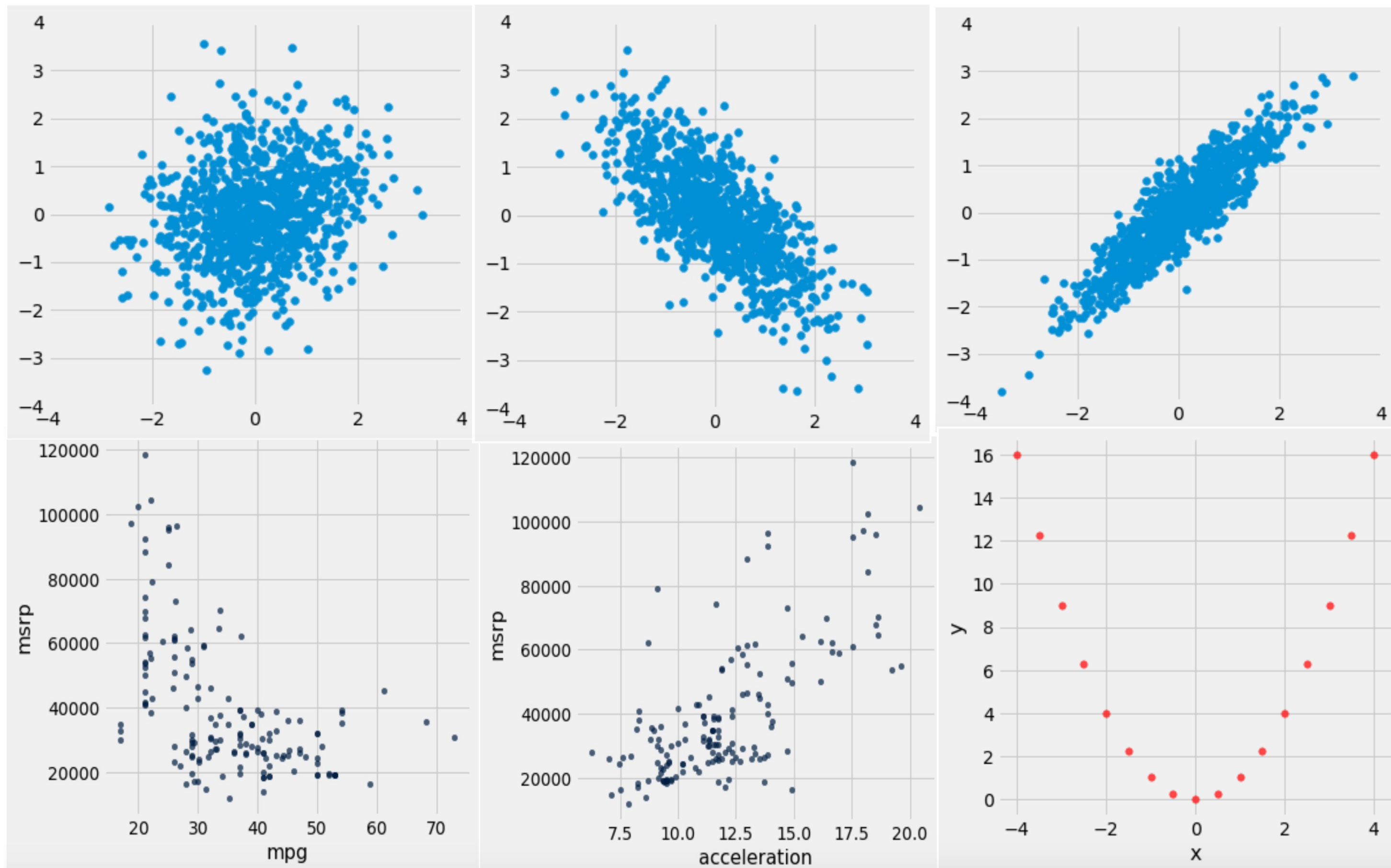
- **Direction**
  - positive vs. negative
- **Form**
  - Linear vs. non-linear
- **Strength**
  - weak - moderate - strong



Correlation

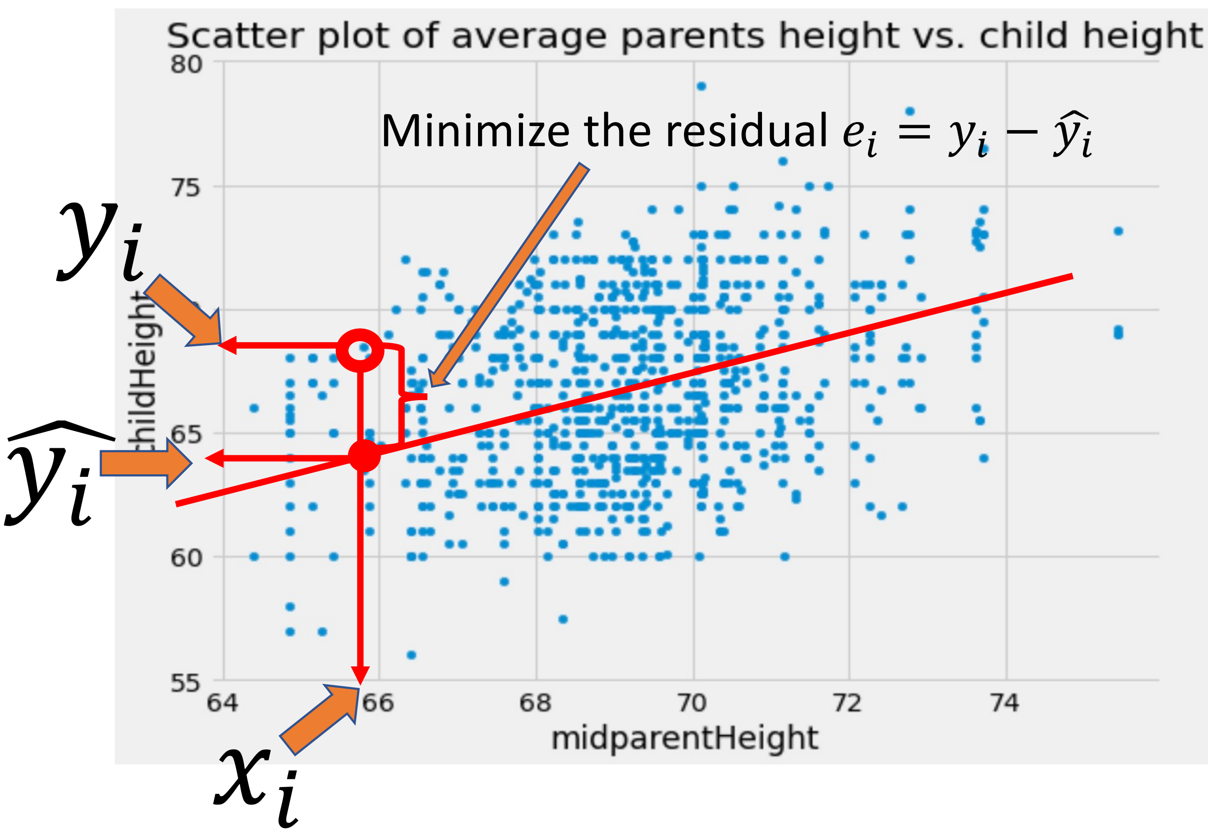
- [Pearson's correlation]
- **Liner association** between 2 **numerical** variables
- $-1 \leq r \leq 1$
- [r = 0]: uncorrelated (not linear associated)

r=0.3 r=-0.7 r=0.9



r=-0.53 r=0.69 r=0

Simple Linear Regression



- Want to predict Y value for any given X.
- Model this relation using a linear function of x
  - $y_i = f(x_i) + \epsilon_i$
  - $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
  - $\widehat{y_i} = \hat{f}(x_i) = \widehat{\beta_0} + \widehat{\beta_1} x_i$
  - what’s the best line?

Least Square Line

- RSS $= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - (\widehat{\beta_0} + \widehat{\beta_1} x_i))^2$
  - RSS: residual sum of squares
  - find $\widehat{\beta_0}$ and $\widehat{\beta_1}$ such that RSS is minimized → least square line
  - closed form solutions: $\widehat{\beta_1} = \mathrm{r}\frac{s_y}{s_x}$ where r is the correlation, s is sample sd

- What do $\widehat{\beta_0}$ and $\widehat{\beta_1}$ mean? → interpretation
  - intercept β_0: average value of y when $\mathrm{x} = 0$
  - slope β_1: average change in y when x increase by 1 unit
  - slope is the key parameter of interest: quantify relation

Multiple Linear Regression

- Data: X n by p matrix, Y n by 1 vector
- Regression model: $\boldsymbol{y} = f\left(\boldsymbol{x_1}, \boldsymbol{x_2}, \cdots \boldsymbol{x_p}\right) + \boldsymbol{\epsilon}$
- Linear regression model:
  - still assume f to be a linear function of X!
  - $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$ for $i = 1, 2, \cdots n$
  - fitting the model means estimate coefficients $\beta_0, \beta_1, \cdots \beta_p$
- Why linear function?
  - a good approximation of “true” f
  - useful both conceptually and practically

Multiple Linear Regression

- RSS $= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - (\widehat{\beta_0} + \widehat{\beta_1}x_{i1} + \widehat{\beta_2}x_{i2} + \cdots))^2$
- A linear function in higher dimension is a hyperplane

- Interpretation of model coefficients
  - Intercept: average value of y when **all x = 0**
  - Slope for x_1 : average change in y when x_1 increase by 1 unit while **holding $x_2 \ldots x_p$ constant**
  - interpretation of one slope means **holding all the others constant**!

Regression with categorical variable

- Y: cumulative GPA = $\beta_0 + \beta_1 \cdot X_1 + \cdots$ + possible other $X's$
  - X_1 : categorical variable school at Rice, 7 categories;

- Dummy variable encoding: K-1 dummy variables for K categories
  - Why it is wrong to just use number coding for categories?
  - interpretation of model coefficients!!!

Student	School	Expand into 6 Dummy Variables	Natural Science	Social Science	Humanity	Business	Architecture	Music
Claire	Engineering	→	0	0	0	0	0	0
Karla	Humanity	→	0	0	1	0	0	0
Dhilani	Social Sci	→	0	1	0	0	0	0
Priya	Natural Sci	→	1	0	0	0	0	0
Isabella	Business	→	0	0	0	1	0	0
Annelie	Social Sci	→	0	1	0	0	0	0

Regression with categorical variable

Student	School	Expand into 6 Dummy Variables	Natural Science	Social Science	Humanity	Business	Architecture	Music
Claire	Engineering	→	0	0	0	0	0	0
Karla	Humanity	→	0	0	1	0	0	0
Dhilani	Social Sci	→	0	1	0	0	0	0
Priya	Natural Sci	→	1	0	0	0	0	0
Isabella	Business	→	0	0	0	1	0	0
Annelie	Social Sci	→	0	1	0	0	0	0

- GPA = $\beta_0 + \beta_1$*NaSci + β_2*SoSci + β_3*Hum + β_4*Busi + β_5*Arch + β_6*Musi
- **Interpretation**
  - β_0: **average GPA in school of Engineering – the baseline category!!!**
  - β_1: **average GPA in NaSci - average GPA in SoE**
  - $\beta_2 \cdots \beta_5$: **???**

Regression with categorical variable

Student	School	Natural Science	Social Science	Humanity	Business	Architecture	Music	Study hour
Claire	Engineering	0	0	0	0	0	0	5.5
Karla	Humanity	0	0	1	0	0	0	8
Dhilani	Social Sci	0	1	0	0	0	0	7.5
Priya	Natural Sci	1	0	0	0	0	0	6
Isabella	Business	0	0	0	1	0	0	6.5
Annelie	Social Sci	0	1	0	0	0	0	4

- GPA = $\beta_0 + \beta_1$*NaSci + β_2*SoSci + β_3*Hum + β_4*Busi + β_5*Arch + β_6*Musi + β_7*study_hour
- **Interpretation**
  - β_0**: average GPA for school of Engineering when study hour is 0!!!**
  - β_1**: difference in average GPA between NaSci and SoE for same study hour**
  - $\beta_2 \cdots \beta_5$**: ???**

More on Linear Regression Models

- ML vs. Stats approaches
  - probabilistic models and maximum likelihood estimate
- Hypothesis testing
  - is the slope statistically different than 0?
- Check goodness of fit
  - R^2, adjusted R^2, Cp, AIC, BIC
- Residual plots and diagnostics
- Interactions, collinearity, heteroscedasticity…

Logistic Regression

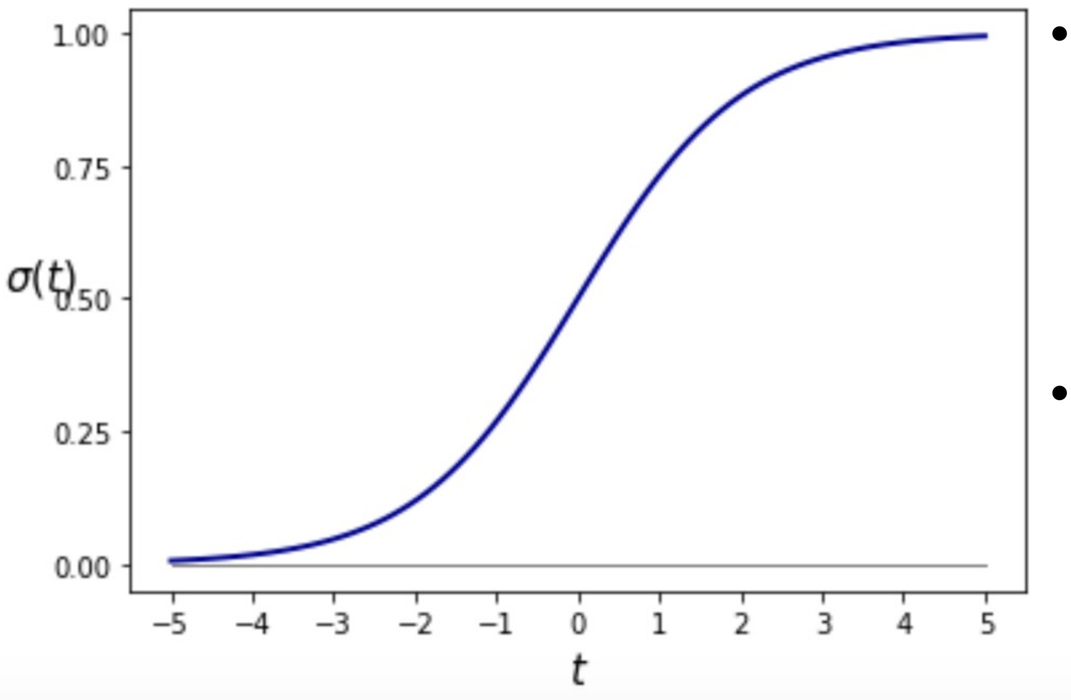
- Suppose now the response variable is categorical - **binary**

- A classification problem → predict y label
  - binary Y usually coded as 0 and 1
  - A Bernoulli random variable!

- A logistic regression model → predict $\mathbb{P}(Y = 1)$
  - A linear function of x can be arbitrary values
  - Logistic function to ensure output between 0 and 1

The Logistic Function

- logistic (t) $= \sigma(t) = \frac{1}{1+e^{-t}}$, also called sigmoid function



- Properties of $\sigma(t)$:
  - $t \to -\infty \Rightarrow \sigma(t) \to 0$
  - $t \to +\infty \Rightarrow \sigma(t) \to 1$
  - $t = 0 \Rightarrow \sigma(t) = 0.5$

- More properties:
  - inverse function
  - derivative

Logistic Regression – model a probability

- Assumption of logistic regression:

  log-odds of y=1 is a linear function of X

- Let $p = \mathbb{P}(y = 1 \mid X = x)$:
  - odds = $\frac{p}{1-p}$ ⟺ log-odds (p) = $\log\left(\frac{p}{1-p}\right)$
  - $\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$ ⟺ $p = \frac{e^{(\beta_0+\beta_1 x)}}{1+e^{(\beta_0+\beta_1 x)}} = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$

    this is the logistic function!

- From probability to label:
  - predict $\hat{y} = 1$ if $\mathbb{P}(y = 1 \mid x) > \mathbb{P}(y = 0 \mid x)$, i.e., if $\mathbb{P}(y = 1 \mid x) > 0.5$

Fit a Logistic Regression Model

- Recall in linear regression we minimize RSS:
- Similar measure in logistic regression is Cross-entropy:
  - $\sum_{i=1}^{n}[-y_i \log(\hat{p}_i) - (1-y_i)\log(1-\hat{p}_i)]$
- What is cross-entropy???
  - Information theory: info gain
  - Statistics: negative log likelihood function
  - Machine learning: a loss/cost function
- Model fitting by minimizing cross-entropy
  - estimate those parameters β_0 and β_1
  - how to interpret them?

Parameter Interpretation

- $\mathbb{P}(Y = 1 \mid X) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$ with estimated $\hat{\beta}_0$ and $\hat{\beta}_1$
- $\log\left(\frac{\mathbb{P}(Y=1 \mid X=x)}{\mathbb{P}(Y=0 \mid X=x)}\right) = \hat{\beta}_0 + \hat{\beta}_1 x$
- When x increases by 1 unit, log odds for y=1 change by $\hat{\beta}_1$
  - **if $\widehat{\boldsymbol{\beta}}_1$>0: increase X means $\mathbb{P}(\boldsymbol{Y} = \mathbf{1})$ increases**
  - **if $\widehat{\boldsymbol{\beta}}_1$<0: increase X means $\mathbb{P}(\boldsymbol{Y} = \mathbf{1})$ decreases**
- Log odds equals $\hat{\beta}_0$ when X=0
- Decision boundary: when predicted $\hat{y}$ changes label
  - odds = 1 ⇔ log odds = 0 ⇔ $\hat{\beta}_0 + \hat{\beta}_1 x = 0$ ⇔ linear function of x!!!

Multiple Logistic Regression

- Multiple predictors: $\mathbb{P}(y = 1 \mid X) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots)}}$
- Interpretation:
  - increase X_1 by 1 unit while **holding all other predictors constant**
  - **log odds** change by β_1 -> **odds** multiply by e^{β_1}
  - interpret the sign:
    - $\boldsymbol{\beta_1}$**>0 means increase** $\boldsymbol{X_1}$ (while holding…) will **increase** $\mathbb{P}(Y = 1)$
    - $\boldsymbol{\beta_1}$**>0 means increase** $\boldsymbol{X_1}$ (while holding…) will **decrease** $\mathbb{P}(Y = 1)$
- Categorical predictors
  - same dummy coding trick!

Assess a Logistic Regression Model

- The Confusion Matrix – a pivot table!

	predicted y 0	Predicted y 1
true y 0	**True-Negatives**	**False-Positives**
true y 1	**False-Negatives**	**True-Positives**

Evaluation metrics:

- accuracy:
  - (TN+TP) / total
- Misclassification:
  - (FN+FP) / total
- TPR = TP / (TP+FN)
  - sensitivity/recall
- TNR = TN / (TN+FP)
  - specificity
- PPV = TP / (TP+FP)
  - precision
- F1 score...

More on Logistic Regression

- General decision rule for binary classification:
  - predict $\hat{y} = 1$ if $\mathbb{P}(y = 1 \mid x) > \tau$ where τ is some threshold
  - treat τ as hyperparameter to maximize accuracy
- Class imbalance:
  - weighted loss function
  - resampling
- Extension to multi-class logistic regression
  - soft-max function instead of sigmoid (logistic) function
- Generalized liner models
  - linear, logistic, multinomial, poison regression…

Take away message

- Understand these basic models well
  - extremely important and widely used
  - more in STAT courses
- Lots of extension to advanced models
  - add non-linear features $x^2, x^3, \ldots$ and interaction terms
  - regularized linear regression, kernel regression, Generalized Additive Models…
- Interpretation is the key!
  - black box models are good at making predictions
  - interpretable models help us understand the world