# Linear Regression and Logistic Regression

Yun (Renee) Zhang, PhD

Email: zhangy@jcvi.org

Monthly Biostatistics Discussion

08-15-2018

# Linear Regression

Regressing CONTINUOUS outcome on predictors

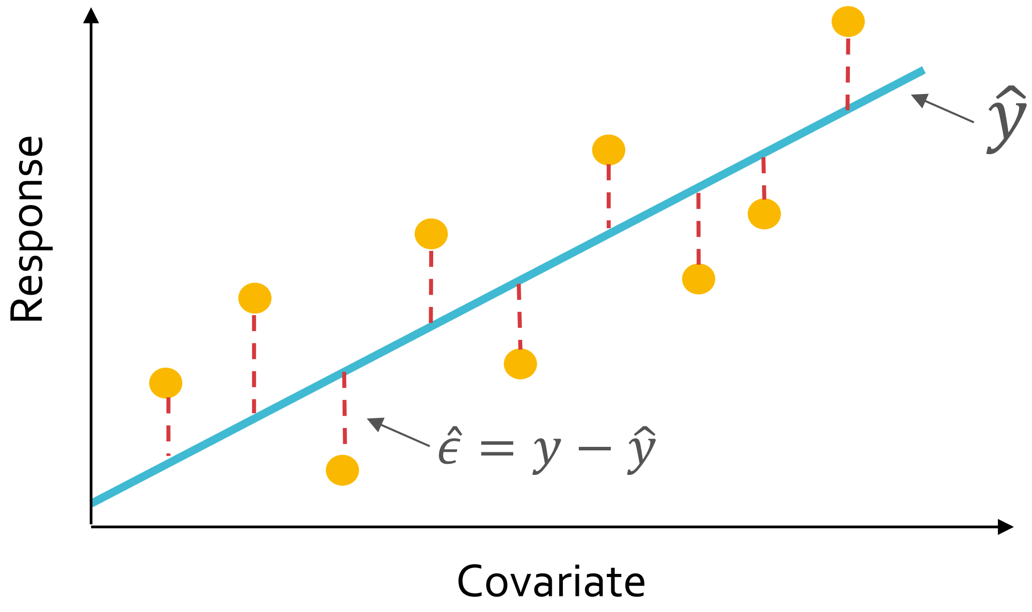
# Terminology and Assumptions

$$Y = \beta_0 + \beta_1 x + \epsilon$$

| Symbol | Term |
|---|---|
| $Y$ | Response / outcome / dependent variable |
| $x$ | Covariate / predictor / explanatory / independent variable |
| $\beta_0, \beta_1$ | Regression coefficients |
| $\epsilon$ | Random error |

- Assumptions: LINE
  - The response $Y$ and covariate $x$ have Linear relationship
  - The errors are Independent
  - The errors are Normally distributed
  - The errors have Equal variances ("homoscedasticity")

# Best Fitted Line and Coefficient Estimation




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

| | Terminology |
|---|---|
| $\hat{y}$ | Fitted value |
| $\hat{\beta}_0$ | Intercept |
| $\hat{\beta}_1$ | Slope |
| $\hat{\epsilon}$ | Residual |

- Residual sum of squares

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

- Least-squares estimation (OLS)

$$\hat{\beta}_0, \hat{\beta}_1 \text{ such that } RSS \text{ is minimized}$$

# Multiple Regression

predictor, 'x-variable', independent variable, explanatory variable

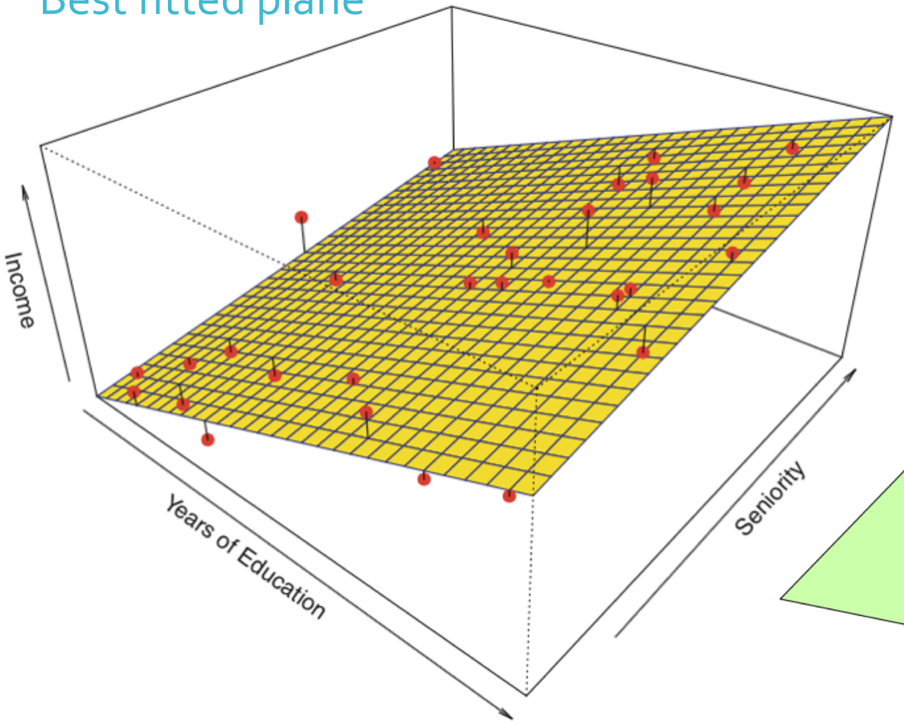
coefficient

$$Y = \underbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p}_{\text{Linear predictor}} + \varepsilon$$

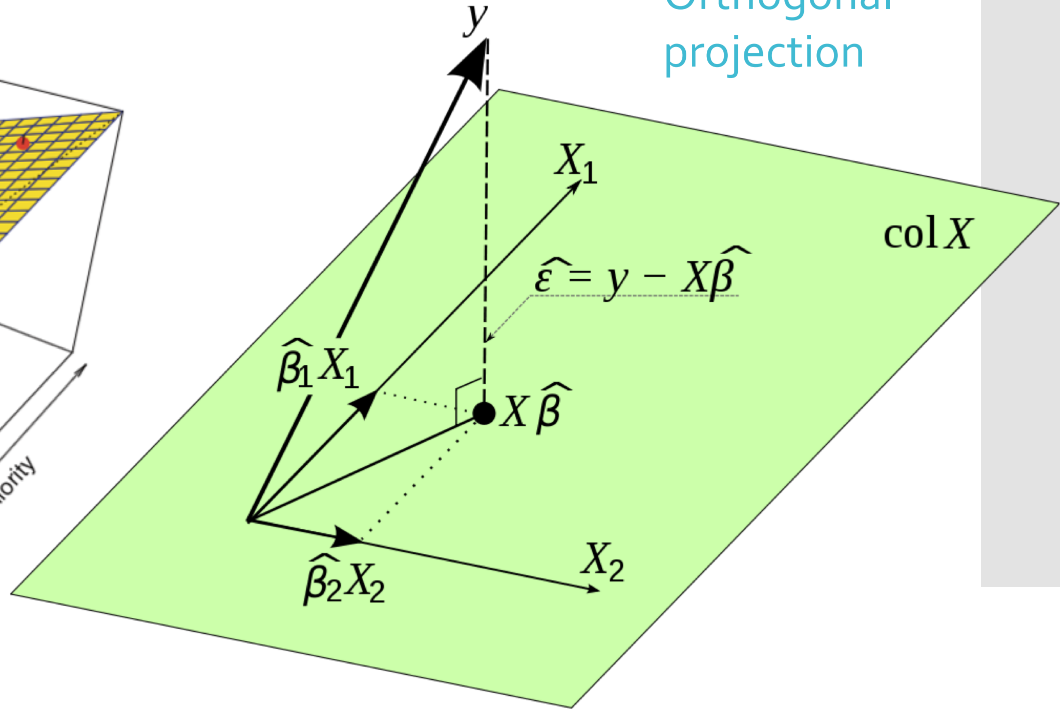
response, dependent variable, observation, 'y-variable'

random error, "noise"

Best fitted plane

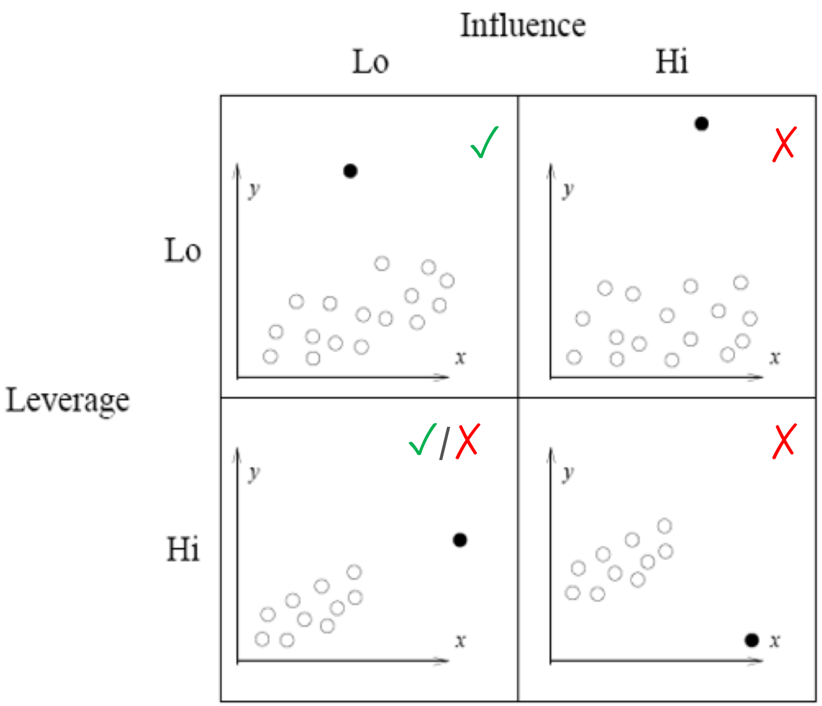



Orthogonal projection

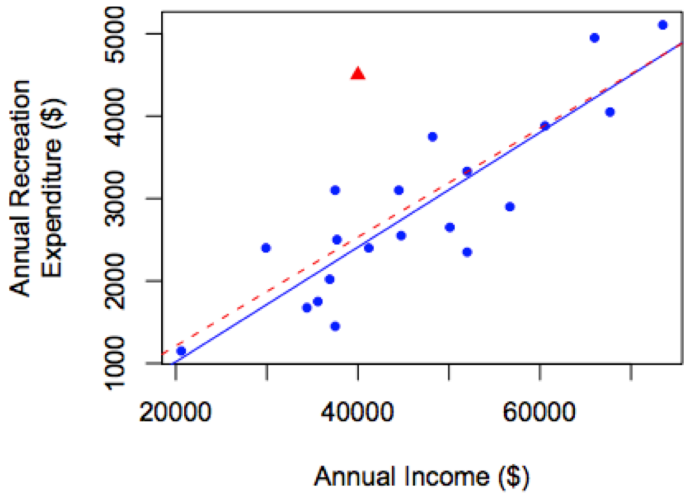


# Data Sanity Check

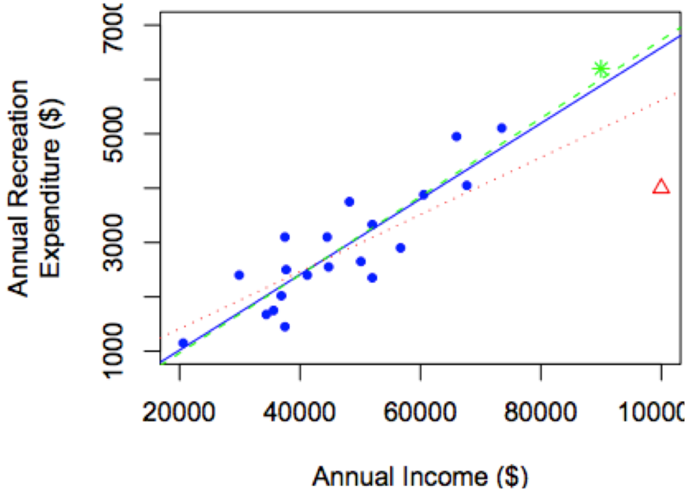
- Scatter plot
- Outliers?
  - High influential point
  - High leverage point




Example 1




Example 2

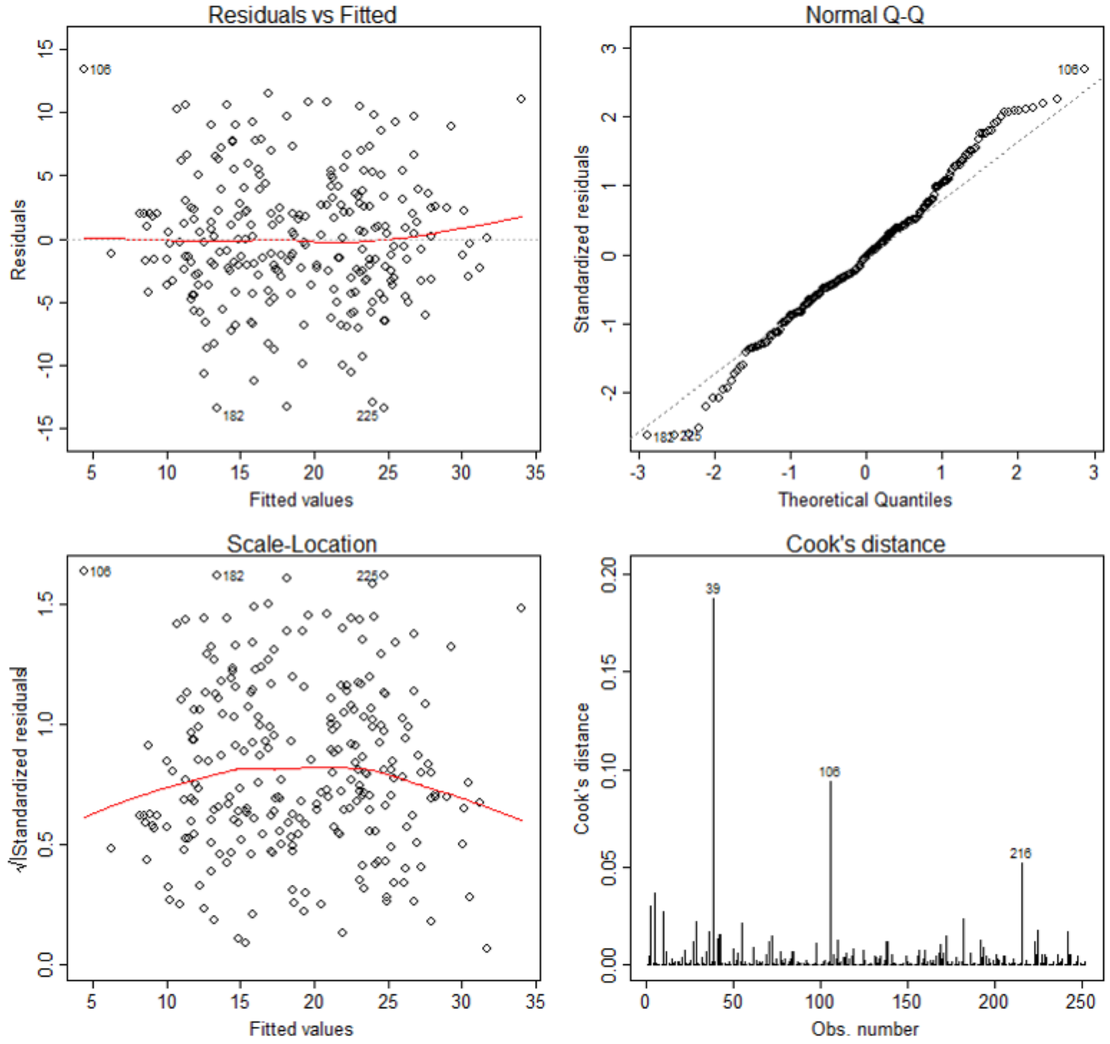


# Model Diagnostics

A good example of diagnostic plots

Linearity

Normality

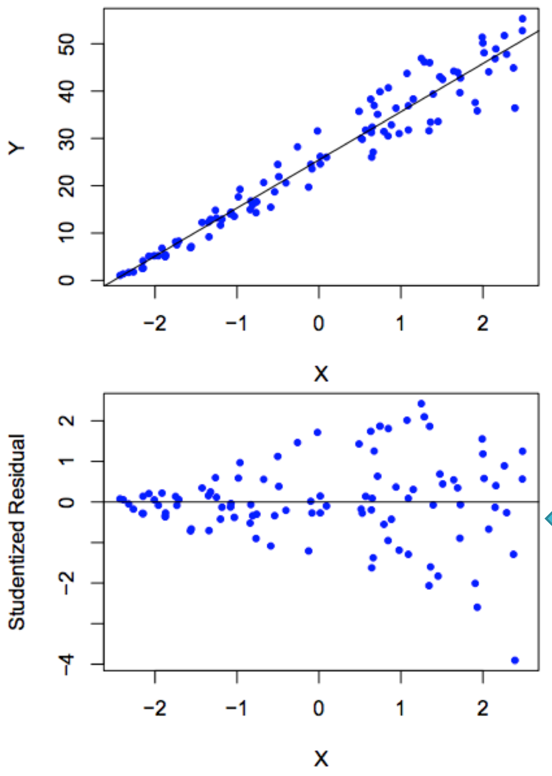



Equal variance

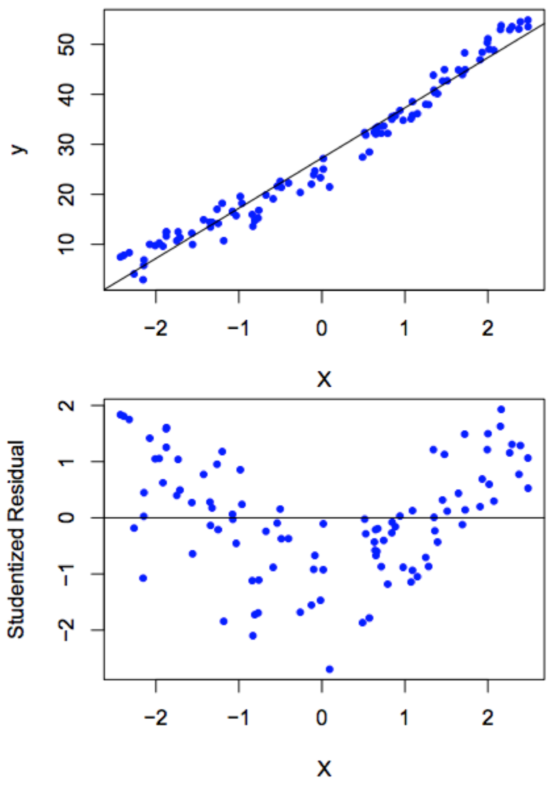
Influential point

# Model Diagnostics – Violation of Assumptions

Example of non-constant variance




Example of non-linearity




Pattern in residual plot

X

- Rule of thumb:

| standardized/Studentized residuals | < 2

# Goodness-of-Fit

- Analysis of Variance Table

| Source | df | SS | MS | F |
|---|---|---|---|---|
| Regression | $p$ | $SSR = \sum_i (\hat{y}_i - \bar{y})^2$ | $MSR = SSR/p$ | $MSR/MSE$ |
| Residual | $n-p-1$ | $SSE = \sum_i (y_i - \hat{y}_i)^2$ | $MSE = SSE/(n-p-1)$ | |
| Total | $n-1$ | $SST = \sum_i (y_i - \bar{y})^2$ | | |

- Partitioning sum of squares

$$\underbrace{Y_i - \bar{Y}}_{\text{Total deviation}} = \underbrace{\hat{Y}_i - \bar{Y}}_{\text{Deviation of fitted regression value around mean}} + \underbrace{Y_i - \hat{Y}_i}_{\text{Deviation around fitted regression line}}$$

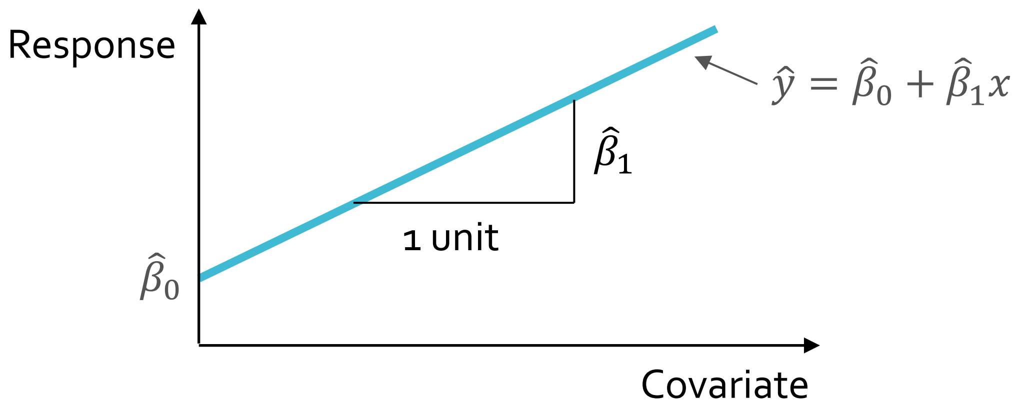
$$SST = SSR + SSE$$

- R-squared

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST} = \frac{\text{Explained variation}}{\text{Total variation}}$$

# Interpretation of Regression Coefficients

- One covariate




- Multiple covariates
  - $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$
  - ***Controlling all other covariates constant***, for one unit change of covariate $x_j$, $\hat{y}$ changes $\hat{\beta}_j$ units
- E.g. $\text{weight} \sim \hat{\beta}_0 + \hat{\beta}_1 \text{height} + \hat{\beta}_2 \text{sex}$
  - For one unit change of height, weight changes $\hat{\beta}_1$ units ✗
  - Fixing sex as the same, for one unit change of height, weight changes $\hat{\beta}_1$ units ✓

# Categorical Covariate / Factor

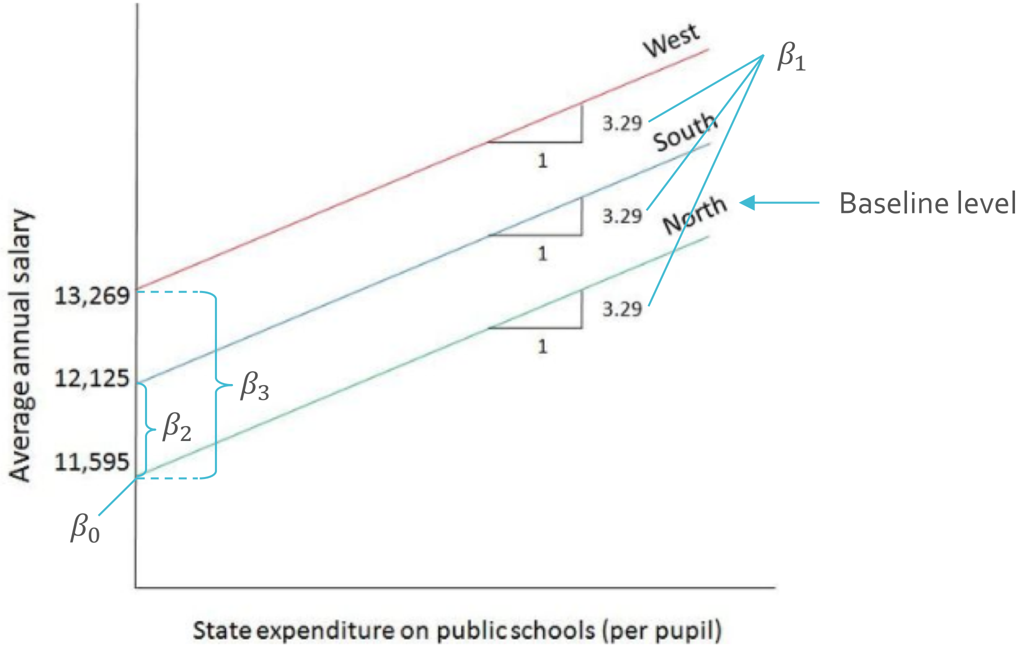
- E.g. sex (M / F), region (East Coast / West Coast / Midwest / South)
- Use of dummy variables
  - E.g. $D_M = \begin{cases} 1, & \text{sex} = \text{M} \\ 0, & \text{otherwise} \end{cases}$ , $D_F = \begin{cases} 1, & \text{sex} = \text{F} \\ 0, & \text{otherwise} \end{cases}$
- Regression model with *only* one categorical covariate

| Model | What does it actually fit? | Baseline level |
|---|---|---|
| $Y \sim \text{sex}$ | $Y = \beta_0 + \beta_1 D_F + \epsilon$ | Male |
| $Y \sim \text{region}$ | $Y = \beta_0 + \beta_1 D_{EC} + \beta_2 D_{WC} + \beta_3 D_{MW} + \epsilon$ | South |

- Interpretation of coefficients
  - E.g. $\text{height} = \beta_0 + \beta_1 D_F + \epsilon$
  - $\beta_0$ is the average height of male (baseline level)
  - $\beta_1$ is the difference between average height of female and that of male
- Regression on a binary covariate is equivalent to two-sample $t$-test
- Regression on a factor with more than two levels is equivalent to one-way ANOVA

# Continuous and Categorical Covariates

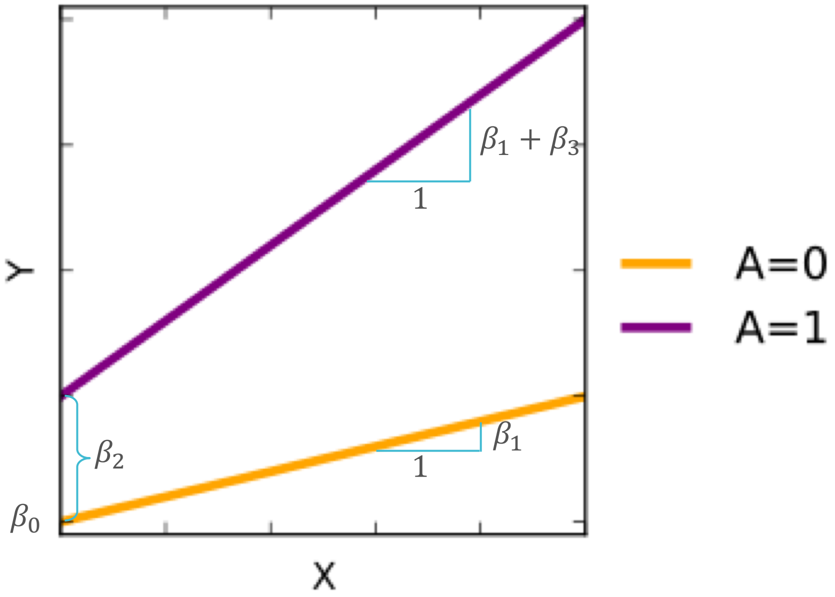
- Regression model with *both* continuous and categorical covariates
- E.g. $\text{salary} = \beta_0 + \beta_1 \text{StateExpenditure} + \beta_2 D_S + \beta_3 D_W + \epsilon$




- No interaction between covariates
  - Constant slope for all levels
  - Shifted intercepts for each level

# Interaction Term

- E.g. $Y = \beta_0 + \beta_1 X + \beta_2 D_1 + \overbrace{\beta_3 X D_1}^{\text{Interaction}} + \epsilon$

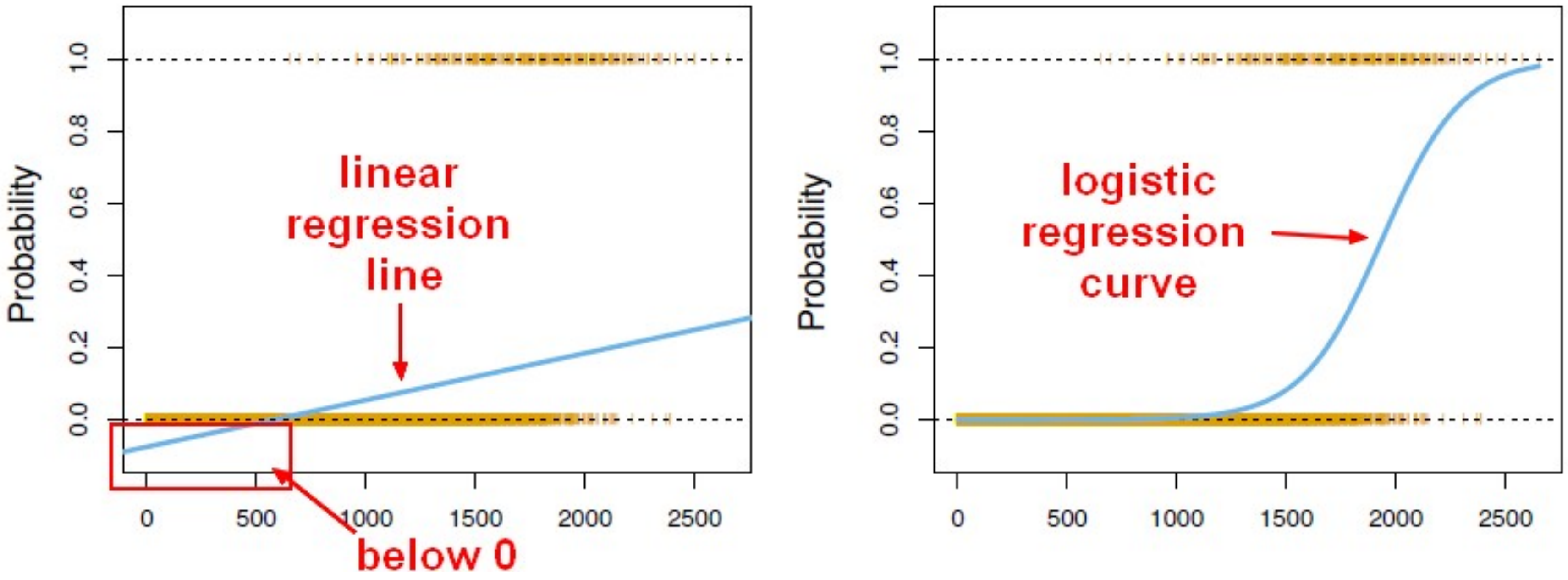



- The interaction term quantifies the slope change between different levels
- Modeling the interaction term between two categorical covariates is equivalent to two-way ANOVA

# Logistic Regression

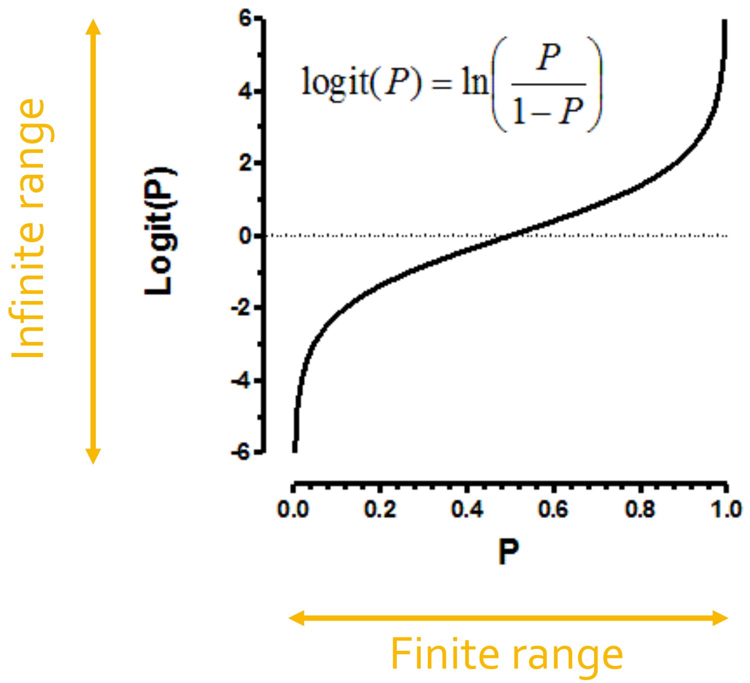
Regressing BINARY outcome on predictors

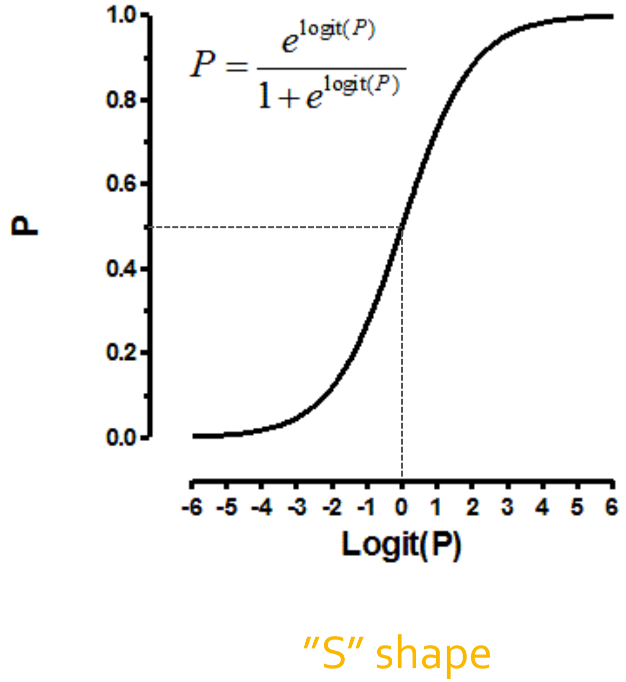
# Why do we need logistic regression?



# Logit Function

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) \quad \text{for} \quad p \in (0,1)$$




- The inverse of the logit function is called the logistic function or sigmoid curve
- Property: monotonically increasing

# Philosophy

**Logistic Regression**

| | | |
|---|---|---|
| Binary values $Y \in \{0,1\}$ | $Y \sim x$ | Observed data |

$p = \text{Probability}(Y = 1)$

| | | |
|---|---|---|
| Finite range $p \in (0,1)$ | $p \sim x$ | Nonlinear relationship |

Logit function

| | | |
|---|---|---|
| Infinite range $\text{logit}(p) \in \mathbb{R}$ | $\text{logit}(p) \sim x$ | Linear model |

$$\text{logit}(p) = \beta_0 + \beta_1 x$$

# Bernoulli Distribution

- A probabilistic view of logistic regression

$$
\begin{aligned}
Y_i \mid x_{1,i}, \ldots, x_{m,i} \ &\sim \operatorname{Bernoulli}(p_i) \\
\mathbb{E}[Y_i \mid x_{1,i}, \ldots, x_{m,i}] &= p_i \\
\Pr(Y_i = y \mid x_{1,i}, \ldots, x_{m,i}) &= \begin{cases} p_i & \text{if } y = 1 \\ 1 - p_i & \text{if } y = 0 \end{cases} \\
\Pr(Y_i = y \mid x_{1,i}, \ldots, x_{m,i}) &= p_i^y (1 - p_i)^{(1-y)}
\end{aligned}
$$

- Line 1 is the probability distribution of the outcome variable
- Line 2 means regression models are essentially conditional expectations (Also true for linear regression)
- Line 3 and 4 are two ways of expressing the probability mass function of Bernoulli distribution
- Bernoulli distribution is the simplest Binomial distribution

# Coefficient Estimation

- Observed data: $(y_i, x_i), i = 1, \cdots, n$
- Density function for each observation: $f(y_i|p_i) = p_i^{y_i}(1-p_i)^{1-y_i}$
- Likelihood function: $L(y_1, \cdots, y_n|p_i) = \prod_i f(y_i|p_i)$
- Log-likelihood function: $l(y_1, \cdots, y_n|p_i) = \log(L) = \sum_i f(y_i|p_i)$
- Maximum likelihood estimation (MLE)

$$l(y_1, \cdots, y_n|p_1, \cdots, p_n) \text{ is maximized}$$

- Conventionally,

$$-l(y_1, \cdots, y_n|p_1, \cdots, p_n) \text{ is minimized}$$

- Logit function and its inverse

$$\text{logit}(p_i) = \beta_0 + \beta_1 x_i \quad \Leftrightarrow \quad p_i = \text{logit}^{-1}(\beta_0 + \beta_1 x_i) = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$$

- Estimation of logistic regression coefficients

$$\hat{\beta}_0, \hat{\beta}_1 \text{ such that } -l(y_1, \cdots, y_n|\beta_0, \beta_1; x_1, \cdots, x_n) \text{ is minimized}$$

# Goodness-of-Fit

- R-squared is not valid (*since logistic regression is heteroscedastic*)
- Likelihood ratio test

$$-2\ln\frac{\text{likelihood of the null model}}{\text{likelihood of the fitted model}} \sim \chi^2_{\text{df}}$$

- Various pseudo $R^2$'s

# Odds and Odds Ratio

- Logistic regression model

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

- Odds

$$\text{odds} = \frac{p}{1-p} = e^{\beta_0+\beta_1 x}$$

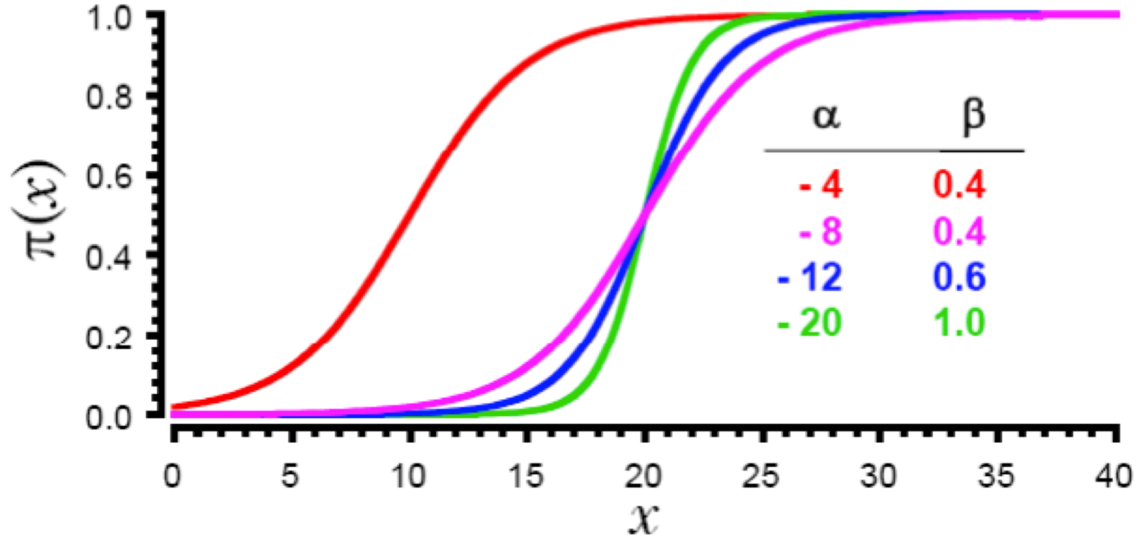
- Odds ratio for one unit increase in the covariate: $x$ -> $x + 1$

$$\text{OR} = \frac{e^{\beta_0+\beta_1(x+1)}}{e^{\beta_0+\beta_1 x}} = e^{\beta_1}$$

- Same for multiple covariates, but ***controlling all other covariates constant***

# Graphical Interpretation

$$\pi(x) = \exp(\alpha + \beta x) / (1 + \exp(\alpha + \beta x))$$

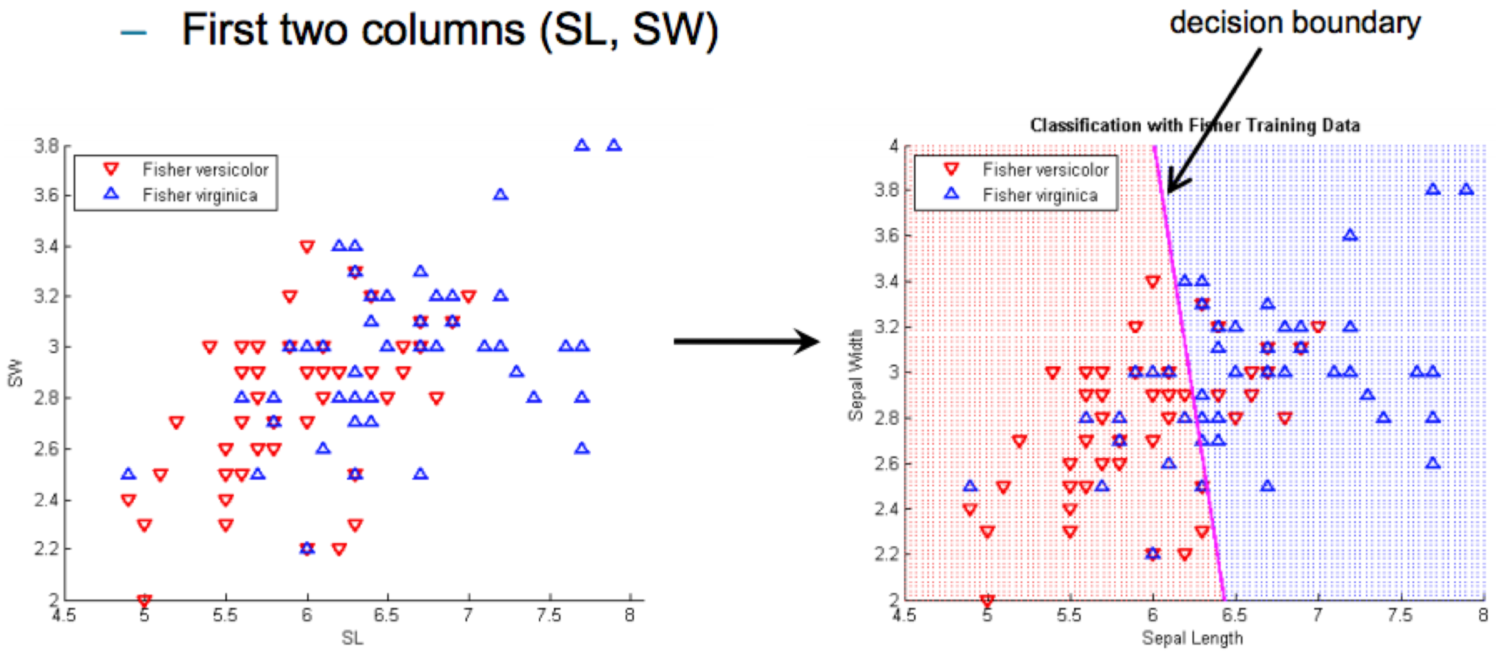



When $x = -\alpha / \beta$, $\alpha + \beta x = 0$ and hence $\pi(x) = 1/(1+1) = 0.5$

- Midpoint is where the probability is 0.5
- $\beta$ controls the slope at the midpoint
- $-\frac{\alpha}{\beta}$ controls the location of the midpoint

# Classification Problem

- A machine learning view of logistic regression
- E.g. **Subset of Fisher iris dataset**
  - Two classes
  - First two columns (SL, SW)



# Logistic Regression vs. Deep Learning

Supplemental Table 1: Prediction accuracy of each task of deep learning model compared to baselines

| | Hospital A | Hospital B |
|---|---|---|
| **Inpatient Mortality, AUROC[1](95% CI)** | | |
| Deep learning 24 hours after admission | **0.95**(0.94-0.96) | **0.93**(0.92-0.94) |
| Full feature enhanced baseline at 24 hours after admission | 0.93 (0.92-0.95) | 0.91 (0.89-0.92) |
| Full feature simple baseline at 24 hours after admission | 0.93 (0.91-0.94) | 0.90 (0.88-0.92) |
| Baseline (aEWS[2]) at 24 hours after admission | 0.85 (0.81-0.89) | 0.86 (0.83-0.88) |
| **30-day Readmission, AUROC (95% CI)** | | |
| Deep learning at discharge | **0.77**(0.75-0.78) | **0.76**(0.75-0.77) |
| Full feature enhanced baseline at discharge | 0.75 (0.73-0.76) | 0.75 (0.74-0.76) |
| Full feature simple baseline at discharge | 0.74 (0.73-0.76) | 0.73 (0.72-0.74) |
| Baseline (mHOSPITAL[3]) at discharge | 0.70 (0.68-0.72) | 0.68 (0.67-0.69) |
| **Length of Stay at least 7 days AUROC (95% CI)** | | |
| Deep learning 24 hours after admission | **0.86**(0.86-0.87) | **0.85**(0.85-0.86) |
| Full feature enhanced baseline at 24 hours after admission | 0.85 (0.84-0.85) | 0.83 (0.83-0.84) |
| Full feature simple baseline at 24 hours after admission | 0.83 (0.82-0.84) | 0.81 (0.80-0.82) |
| Baseline (mLiu[4]) at 24 hours after admission | 0.76 (0.75-0.77) | 0.74 (0.73-0.75) |

Rajkomar, Alvin, et al. "Scalable and accurate deep learning with electronic health records." *npj Digital Medicine* 1.1 (2018): 18.

# Some Comments

# Computational Solutions

- In linear regression, OLS solution is equivalent to MLE solution (i.e. under normal distribution)
- Closed form solution for linear regression
$$\widehat{\boldsymbol{\beta}} = (X'X)^{-1}X'\boldsymbol{y}$$
- The closed form solution involves inversion of (large) matrix, so computationally the solution is calculated by numerical approximation algorithm, e.g. Newton-Raphson, gradient descent

- No closed form solution for logistic regression
- Logistic regression coefficients are estimated by iteratively reweighted least squares (IRLS), which is equivalent to the MLE solution under Bernoulli distribution estimated by Newton-Raphson

# More Topics in Regression

- Generalized linear model
  - Probability distribution
  - Linear predictor
  - Link function
- Variable selection
  - Stepwise: backward-, forward-selection
  - Information criterion: AIC, BIC
- High-dimensional regression
  - The "large *p*, small *n*" problem / Curse of dimensionality
  - Regularization / penalized regression techniques
  - Ridge, LASSO, elastic-net regressions
  - Principal component regression