(7) $Bin(12, .3)$ is exponential tilt of $Bin(12, .6)$

- Strategy: We'll write the Binomial pmf in exponential family form.
- Then, we'll find the exponential tilt that takes you from one Binomial to the other.

Binomial pmf

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

we'll write

$$f_1(x) = \binom{n}{x} (p_1)^x (1-p_1)^{n-x} \quad \text{and}$$

$$f_2(x) = \binom{n}{x} (p_2)^x (1-p_2)^{n-x}$$

$$\text{Logit}(p) = \log\left(\frac{p}{1-p}\right)$$

Goal

$$f_\alpha(x) = e^{n[\alpha^T \bar{y} - \psi(\alpha)]} \cdot f_0(x)$$

$x = y$

$$f_\alpha(x) = e^{n[\alpha \bar{x} - \psi(\alpha)]} \cdot f_0(x)$$

let's take logits to get expressions...

$$f_1(x) = \binom{n}{x} \cdot \exp\left\{ x \log(p_1) + (n-x) \log(1-p_1) \right\}$$

$$= \binom{n}{x} \cdot \exp\left\{ x\left[\log(p_1) - \log(1-p_1)\right] + n\log(1-p_1) \right\}$$

$$= \binom{n}{x} \cdot \exp\left\{ x \log\left(\frac{p_1}{1-p_1}\right) + n \log(1-p_1) \right\}$$

Likewise ...

$$f_2(x) = \binom{n}{x} \exp\left\{ x \log\left(\frac{p_2}{1-p_2}\right) + n\log(1-p_2) \right\}$$

↓ Following the same strategy as for the Poisson ...

$$\frac{f_1(x)}{f_2(x)} = \exp\left\{ x\left[\log\left(\frac{p_1}{1-p_1}\right) - \log\left(\frac{p_2}{1-p_2}\right)\right] + n\left[\log(1-p_1) - \log(1-p_2)\right] \right\}$$

Recall $\log(a) - \log(b) = \log\left(\frac{a}{b}\right)$

$$= \exp\left[x \cdot \log\left(\frac{p_1(1-p_2)}{p_2(1-p_1)}\right) + n \cdot \log\left(\frac{1-p_1}{1-p_2}\right) \right]$$

So we let $\alpha = \log\left(\frac{p_1(1-p_2)}{p_2(1-p_1)}\right)$

and $\psi(\alpha) = -n\log\left(\frac{1-p_1}{1-p_2}\right)$ ← $= -\psi(\alpha)$

$$\Rightarrow \frac{f_1(x)}{f_2(x)} = e^{\alpha x - \psi(\alpha)}$$

$p_1 = .3$
$p_2 = .6$
$n = 12$

$$\rightarrow f_1(x) = e^{\alpha x - \psi(\alpha)} \cdot f_2(x)$$

$\Rightarrow$ we have written $f_1(x)$ as an exponential tilt of $f_2(x)$.

"Binomial is the number of successes in n bernoulli trials".

$$\text{Bernoulli}(p) = \begin{cases} 1 \text{ with prob } p. \\ 0 \text{ otherwise} \end{cases}$$

Chapter 7 : Ridge Regression

- Linear regression is based on a version of $\hat{\mu}^{MLE}$. We usually assume an n-dimensional vector $Y = (y_1, \ldots, y_n)^T$ from linear model

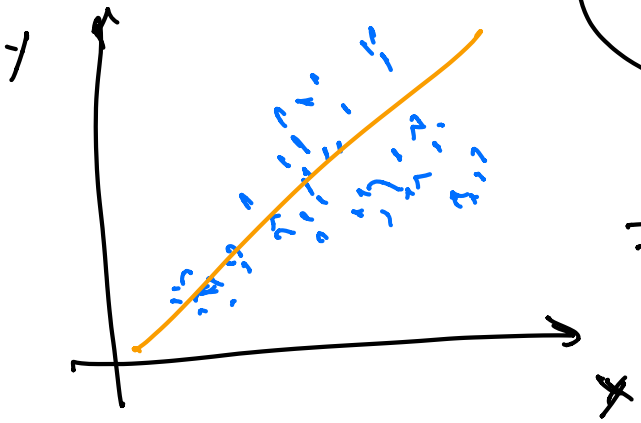
$$\underline{Y} = X \cdot \beta + \underline{\varepsilon}$$

$X^{n \times p}$ is known as the design/structure matrix;

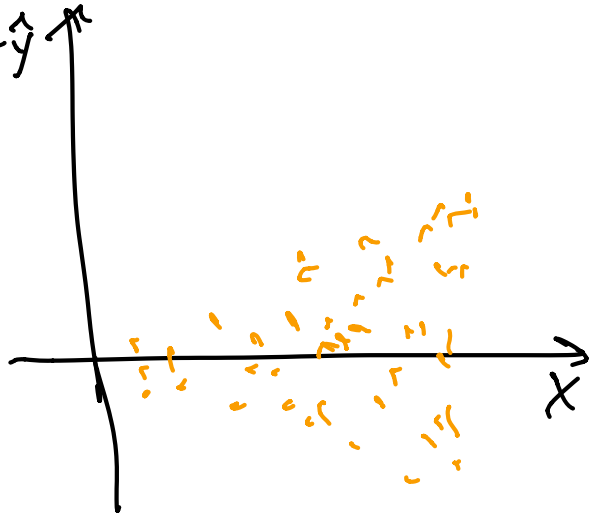
β is an unknown p-dimensional vector. ε is the noise vector and it assumed to have uncorrelated components. Mathematically stated: $\varepsilon \sim (0, \sigma^2 I_n)$

Variance is constant: Scalar

Identity matrix of size n.



$\Rightarrow$



Then, your regression coefs are not valid b/c variance is not constant.

The Least Squares Estimator (Gauss & Legendre in the early 1800's!) minimizes Sum of Squared errors

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \|y - X\beta\|^2 \right\}$$

(Minimizing $\|y - X\beta\|^2$)

$$2(-1)(y - X\beta) \cdot X = 0$$

$$(y - X\beta) \cdot X = 0$$

$$yX - X^T X\beta = 0$$

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

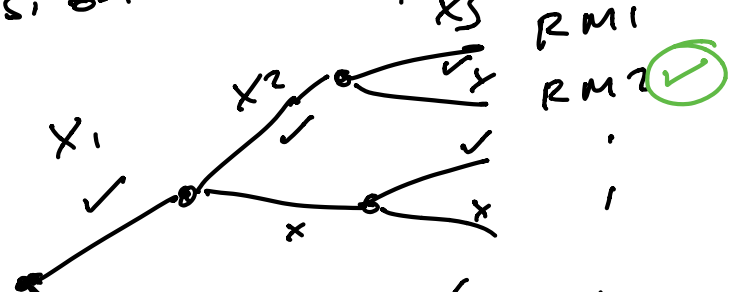
$$\hat{y} = X\beta$$

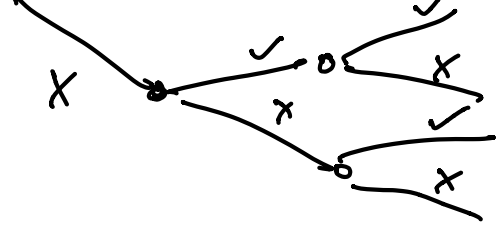
Let $S = (X^T X) \Rightarrow \hat{\beta} = S^{-1} X^T y$

Awesome for a moderate amount of dimensions (say in the 10's if you have a few thousand observations).

In higher dimensions, we often need to limit the # of important variables. To do this we'd ideally like to solve the subset selection problem.

Example: Say you had 3 variables in your X. Question is: which variables to consider in our regression model?





If have p variables, the subset selection problem explodes exponentially (2^p different models).

→ Because of this, the subset-selection problem has been solved through approximations.

historically

- One of the earliest ones was Ridge regression
  - Ridge considers the square of the sum of your regression coeffs and penalizes it! $\|\beta\|^2$

- In short, Ridge regression will attempt to shrink your coefficients towards zero.

(S7) Sum of our coeffs.

Ridge obj. function is:

$$\hat{\beta}(\lambda) = \underset{\beta}{\operatorname{argmin}} \left\{ \|y - X\beta\|^2 + \lambda \|\beta\|^2 \right\}.$$

sum of squares

λ Large → More β_j's are small

λ small → Less β_j's are small

$$\|\beta\|^2 = \sum_{j=1}^{p} \beta_j^2$$

Penalty weight / Regularization Parameter.

Following same minimization logic...

$$-2 \cdot (y - X\beta)X + 2\lambda\beta = 0$$

$$-X^T y + X^T X\beta + \lambda\beta = 0$$

$$X^T X\beta + \lambda\beta = X^T y$$

$$\beta(X^T X + \lambda I) = X^T y$$

$$\Rightarrow \hat{\beta}^{ridge} = (X^T X + \lambda I)^{-1} \cdot X^T y$$

$$\hat{\beta}(\lambda) = \underset{\beta}{\arg\min} \left(\frac{1}{n}\|y - X\beta\|^2 + \frac{\lambda}{n}\|\beta\|^{2'} \right).$$

MSE: Mean squared error.

Sum of squared residuals

Penalty weight / Regularization Parameter.