Lab 4: More on Regression

1 t and F distribution

χ^2_d can be generated by the sum square of d independent variables with standard normal distribution. i.e.

$$\begin{aligned} X &= Z_1^2 + Z_2^2 + \cdots + Z_d^2 \sim \chi^2_d \\ Z_i &\sim N(0,1), \text{ i.i.d} \\ E(X) &= d \end{aligned} \tag{1}$$

Student's t distribution t_d with degree of freedom d is defined as the probability distribution of $W := \frac{Z}{\sqrt{V/d}}$, where

$$\begin{aligned} Z &\sim N(0,1) \\ V &\sim \chi^2_d \\ Z &\perp V \end{aligned} \tag{2}$$

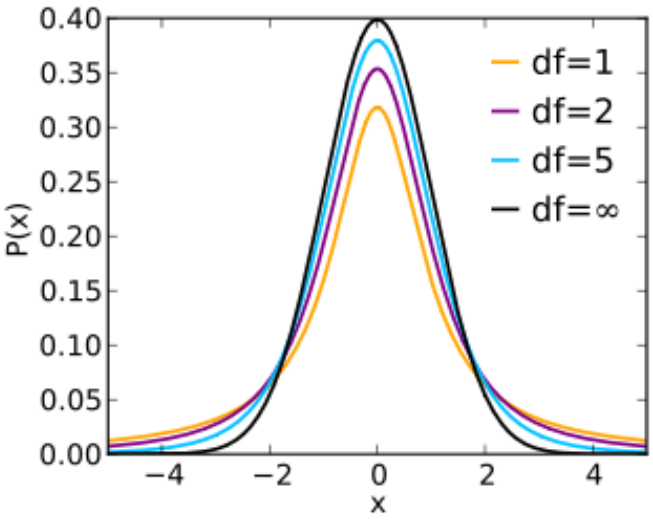

Figure 1: From Wikipedia

The curve of t distribution is symmetric around the origin. When $d \to \infty$, $W \to N(0,1)$ in distribution by law of large numbers.

F distribution $F(d_1, d_2)$ with degrees of freedom d_1, d_2 is defined as the distribution of $\frac{V_1/d_1}{V_2/d_2}$, where

$$\begin{aligned} V_1 &\sim \chi^2_{d_1} \\ V_2 &\sim \chi^2_{d_2} \\ V_1 &\perp V_2 \end{aligned} \tag{3}$$

F distribution is nonnegative and $F(1, d) = t^2_d$.

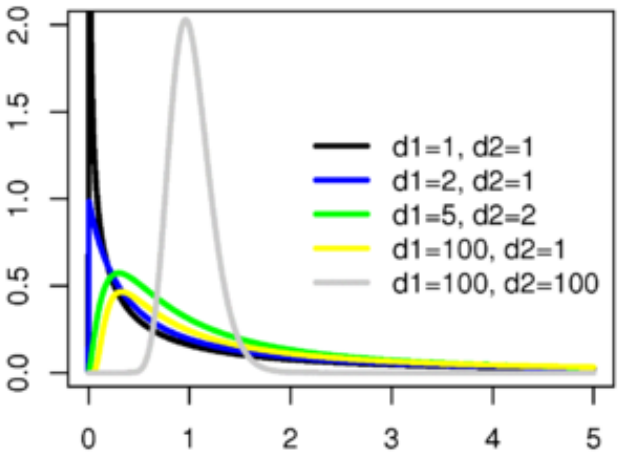

Figure 2: From Wikipedia

In regression, the residual sum of square

$$\begin{aligned} RSS &= (\mathbf{Y} - \hat{\mathbf{Y}})^T(\mathbf{Y} - \hat{\mathbf{Y}}) \\ &= \mathbf{Y}^T(\mathbf{I} - \mathbf{H})\mathbf{Y} \\ &\sim \sigma^2\chi^2_{n-p} \end{aligned} \tag{4}$$

2 Generalized Least Squares (GLS)

Suppose $cov(\mathbf{Y}) = \Sigma$, where Σ could be independent, exchangeable, autoregressive or unstructured. Decompose the covariance matrix to be $\Sigma = \Sigma^{1/2}\Sigma^{1/2}$. If we transform the model to be

$$\begin{aligned} \mathbf{Y}^* &= \Sigma^{-1/2}\mathbf{Y} \\ &= \Sigma^{-1/2}\mathbf{X}\beta + \Sigma^{-1/2}\epsilon \\ &= \mathbf{X}^*\beta + \epsilon^* \end{aligned} \tag{5}$$

Then $cov(\epsilon^*) = I$ and the model goes back to the ordinary least square(OLS) problem. The GLS gives

$$\beta^* = (\mathbf{X^T\Sigma^{-1}X})^{-1}(\mathbf{X^T\Sigma^{-1}Y}) \tag{6}$$

β^* is the BLUE.

3 Generalized Linear Model

$$\begin{aligned} Y &\sim EF(\theta, \phi) \\ f(y, \theta, \phi) &= exp\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\} \\ \mu &= E(y) \\ \eta &= g(\mu) \\ \eta &= \beta_0 + \sum_j^p \beta_j x_{ij} \end{aligned} \tag{7}$$

$g(\cdot)$ is called the link function. Normal, Exponential, Bernoulli,$\cdots$ etc. belong to the exponential family.

- If $y's$ are $0-1$ data following Bernoulli(p). $a(\phi) = 1$, $\theta = logitp$. If we let $g(\mu) = logit\mu$, we have the logistic regression model. g is the canonical link.
- If $y's$ are counts distributed as Poisson(λ), e.g, number of events per unit time. $\theta = log\lambda$. If we let $g(\mu) = log\lambda$, we have the log-linear model. The coefficient β_1 can be interpreted as the relative risk of disease caused with one unit increase of X_1 with other covariates fixed.