Lab 4: More on Regression

1 \( t \) and F distribution

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

\[
X = Z_1^2 + Z_2^2 + \cdots + Z_d^2 \sim \chi^2_d
\]
\[
Z_i \sim N(0, 1), \text{ i.i.d}
\]
\[
E(X) = d
\]  

Student’s t distribution \( t_d \) with degree of freedom \( d \) is defined as the probability distribution of

\[
W := \frac{Z}{\sqrt{V/d}}, \text{ where}
\]
\[
Z \sim N(0, 1)
\]
\[
V \sim \chi^2_d
\]
\[
Z \perp V
\]  

![Figure 1: From Wikipedia](image)

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{align*}
V_1 & \sim \chi^2_{d_1} \\
V_2 & \sim \chi^2_{d_2} \\
V_1 & \perp V_2
\end{align*}

(3)

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

\[2\] Generalized Least Squares (GLS)

Suppose $\text{cov}(Y) = \Sigma$, where $\Sigma$ 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{align*}
Y^* &= \Sigma^{-1/2}Y \\
&= \Sigma^{-1/2}X\beta + \Sigma^{-1/2}\epsilon \\
&= X^*\beta + \epsilon^*
\end{align*}

(5)

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

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

(6)

$\beta^*$ is the BLUE.
3 Generalized Linear Model

\[ Y \sim EF(\theta, \phi) \]
\[ f(y, \theta, \phi) = \exp\left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\} \]
\[ \mu = E(y) \]
\[ \eta = g(\mu) \]
\[ \eta = \beta_0 + \sum_j \beta_j x_{ij} \quad (7) \]

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

- If \( y \)'s are \( 0-1 \) data following Bernoulli(\( p \)). \( a(\phi) = 1, \theta = \text{logit} p \). If we let \( g(\mu) = \text{logit} \mu \), we have the logistic regression model. \( g \) is the canonical link.

- If \( y \)'s are counts distributed as Poisson(\( \lambda \)), e.g. number of events per unit time. \( \theta = \text{log} \lambda \). If we let \( g(\mu) = \text{log} \lambda \), we have the log-linear model. The coefficient \( \beta_1 \) can be interpreted as the relative risk of disease caused with one unit increase of \( X_1 \) with other covariates fixed.