Econometrics

Shinjo Hiroki

November 13, 2024

Contents

1 Linear Regression **2**
1.1 OLS 2
1.1.1 Normal Regression model 2
1.1.2 Linear Regression model 2
1.1.3 Influence function 3
1.1.4 Linear Predictor 4
1.2 Endogeneity 5
1.2.1 Omitted variable Bias 6
1.2.2 Measurement Error 6
1.3 IV method and GMM 7
1.3.1 IV estimator 7
1.3.2 GMM 8
1.3.3 2SLS 8
1.3.4 Efficient GMM 8
1.4 Quantile Regression model 8

2 Non-linear MLE estimation **10**
2.1 Binary Choice model 10
2.1.1 Probit model 10
2.1.2 Random Utility Model 11
2.2 Censored Tobit model 12
2.2.1 Parametric assumption 12
2.2.2 Censored LAD 14
2.3 Truncated Tobit model 15
2.4 Sample Selection model 15
2.5 Duration model 17
2.6 Dynamic Programming model 17

3 Nonparametric model **20**
3.1 Local approach(Kernel Regression) 20
3.1.1 Local constant estimation 20
3.1.2 Local linear estimation 21
3.1.3 Local Polynomial estimation 22

1 Linear Regression

1.1 OLS

1.1.1 Normal Regression model

We consider the following model and derive the likelihood.

$$y_i = x_i^T \beta + u_i$$
$$u_i \sim^{iid} N(0, \sigma^2)$$

We assume that Y given X is normal. Then we have

$$f(y_i \mid x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{1}{2\sigma^2}(y - x^T\beta)^2 \right)$$

Therefore, likelihood function is

$$\begin{aligned} L_n(\beta, \sigma) &:= f(y_1, \cdots, y_n \mid x_1, \cdots, x_n) \\ &= \prod_{i \in I} f(y_i \mid x_i) \because \text{ samples are mutually independent} \\ &= \prod_{i \in I} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{1}{2\sigma^2}(y_i - x_i^T\beta)^2 \right) \end{aligned}$$

Then, Log-Likelihood function is

$$\begin{aligned} logL_n(\beta, \sigma) &= f(y_1, \cdots, y_n \mid x_1, \cdots, x_n) \\ &= -nlog\sigma - \frac{1}{2\sigma^2} \sum_{i \in I}(y_i - x_i^T\beta)^2 - \frac{1}{2} log(2\pi) \end{aligned}$$

FOC implies

$$0 = \frac{\partial logL_n(\beta, \sigma)}{\partial \beta} = \frac{1}{\hat{\sigma}_{mle}^2} \sum_{i \in I} x_i (y_i - x_i^T \hat{\beta}_{mle})$$
$$0 = \frac{\partial logL_n(\beta, \sigma)}{\partial \sigma} = -\frac{n}{\hat{\sigma}_{mle}} + \frac{1}{\hat{\sigma}_{mle}^3} \sum_{i \in I}(y_i - x_i^T \hat{\beta}_{mle})^2$$

Hence,

$$\hat{\beta}_{mle} = \left(\frac{1}{n} \sum_{i \in I} x_i x_i^T \right)^{-1} \frac{1}{n} \sum_{i \in I} x_i y_i = \hat{\beta}_{ols}$$
$$\hat{\sigma}_{mle}^2 = \frac{1}{n} \sum_{i \in I}(y_i - x_i^T \hat{\beta}_{mle})^2 = \frac{1}{n} \sum_{i \in I}(y_i - x_i^T \hat{\beta}_{ols})^2 = \frac{1}{n} \sum_{i \in I} \hat{e_i}^2 = \hat{\sigma}_{ols}^2$$

1.1.2 Linear Regression model

Asymptotic Property

$$y_i = x_i^T \beta + u_i$$

,where $\mathbb{E}[u_i|x_i] = 0$.

$$\hat{\beta}_{ols} = \left(\frac{1}{n}\sum_{i\in I} x_i x_i^T\right)^{-1} \frac{1}{n}\sum_{i\in I} x_i y_i = \beta + \left(\frac{1}{n}\sum_{i\in I} x_i x_i^T\right)^{-1} \frac{1}{n}\sum_{i\in I} x_i u_i$$

By assuming

1. $E[u_i \mid x_i] = 0$
2. (x_i, y_i) are iid for $i = 1, \cdots, n$.
3. $\left(\frac{1}{n}\sum_{i\in I} x_i x_i^T\right)$ is invertible.
4. x_i and u_i have at least 4th moment.

, central limit theorem implies

$$\frac{1}{\sqrt{n}}\sum_{i\in I} x_i u_i \rightarrow^d N(0, E[u_i^2 x_i x_i^T]).$$

Hence, the continuous mapping theorem implies

$$\begin{aligned}\frac{1}{\sqrt{n}}(\hat{\beta}_{ols} - \beta) &= \left(\frac{1}{n}\sum_{i\in I} x_i x_i^T\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i\in I} x_i u_i \\ &\rightarrow^d N(0, E[x_i x_i^T]^{-1} E[u_i^2 x_i x_i^T] E[x_i x_i^T]^{-1})\end{aligned}$$

1.1.3 Influence function

We derive influence function.

$$\begin{aligned}\hat{Q}_n(\beta) &= \frac{1}{n}\sum_{i\in I}(y_i - x_i^T\beta)^2 \\ \nabla_b \hat{Q}_n(\beta) &= -\frac{2}{n}\sum_{i\in I}(y_i - x_i^T\beta)x_i \\ \nabla_b^2 \hat{Q}_n(\beta) &= \frac{2}{n}\sum_{i\in I} x_i x_i^T\end{aligned}$$

Then influence function is

$$\mathbb{E}[x_i x_i^T]^{-1}(y_i - x_i^T\beta)x_i = \mathbb{E}[x_i x_i^T]^{-1} u_i x_i$$

Exercise

Consider the data for $i = 1, \cdots, 1000$, $y_i = 1\{1 + x_i + u_i \geq 0\}$,where $u_i \sim N(0, \sigma^2)$, α_i and $v_i \sim N(0, 1)$,and $x_i = \alpha_i + v_i$. Solve the following OLS model

$$y_i = \alpha + x_i\beta + u_i$$

1. Estimate parameters without package.

2. Estimate with lm.

```r
#Data generation
set.seed(123)
n <- 1000; u <- rnorm(n);alpha <- rnorm(n); v <- rnorm(n)
x <- alpha + v
y <- ifelse(1 + x + u >=0,1,0)
data <- data.frame(x,y)

#(1)
LinearReg <- function(y_in,x_in){
  x <- cbind(1,x_in)
  beta <- c(1,1)

  f_ols <- function(beta,y_in,x_in){
    xb <- x_in %*% beta
    residuals <- y_in -xb
    return(sum(residuals^2))
  }

  lm_ols <- optim(par = beta,fn = f_ols,y_in =y_in,x_in = x)

  return(lm_ols$par)
}
LinearReg(y_in = data$y,data$x)

#(2)
lm(y~x,data=data)
```

1.1.4 Linear Predictor

We want to predict Y with X. Let predictor be $P(X)$. $P(X)$ possibly takes any function of X. We evaluate the predictive accuracy with mean squared error (MSE) criteria.

$$\mathbb{E}\left[(Y - P(X))^2\right]$$

I show $P(X) = \mathbb{E}[Y \mid X]$ minimizes MSE.

$$\begin{aligned}
\mathbb{E}\left[(Y - P(X))^2\right] =& \mathbb{E}\left[(Y - \mathbb{E}[Y \mid X] + \mathbb{E}[Y \mid X] - P(X))^2\right] \\
=& \mathbb{E}\left[(Y - \mathbb{E}[Y \mid X])^2\right] + 2\mathbb{E}[(Y - \mathbb{E}[Y \mid X])(\mathbb{E}[Y \mid X] - P(X))] \\
&+ \mathbb{E}\left[(\mathbb{E}[Y \mid X] - P(X))^2\right] \\
=& \mathbb{E}\left[(Y - \mathbb{E}[Y \mid X])^2\right] + 2\mathbb{E}[(\mathbb{E}[Y \mid X] - P(X))(\mathbb{E}[Y \mid X] - \mathbb{E}[Y \mid X]) \mid X] \\
&+ \mathbb{E}\left[(\mathbb{E}[Y \mid X] - P(X))^2\right] \because \text{ law of iterated expectation} \\
=& \mathbb{E}\left[(Y - \mathbb{E}[Y \mid X])^2\right] + \mathbb{E}\left[(\mathbb{E}[Y \mid X] - P(X))^2\right]
\end{aligned}$$

$P(X) = \mathbb{E}[Y \mid X]$ minimizes MSE. The implication of the above argument is that if we have data X and want to predict Y, then $E[Y \mid X]$ provide the best prediction of Y.

How we obtain $\mathbb{E}[Y \mid X]$ from data? We can consider a lot of way to approximate $\mathbb{E}[Y \mid X]$.

We often employ linear approximation. It implies we predict $\mathbb{E}[Y \mid X]$ by $X^T\beta$. We also evaluate via MSE criteria.

$$\min_{\beta} \mathbb{E}\left[\left(\mathbb{E}[Y \mid X] - X^T\beta\right)^2\right]$$

FOC implies

$$\begin{aligned} &\mathbb{E}\left[X(\mathbb{E}[Y \mid X] - X^T\hat{\beta})\right] = 0 \\ \iff &\hat{\beta} = \mathbb{E}\left[XX^T\right]^{-1}\mathbb{E}[X\mathbb{E}[Y \mid X]] = \mathbb{E}\left[XX^T\right]^{-1}\mathbb{E}[XY] \end{aligned}$$

This is the probability limit of the OLS estimator under regularity assumptions. This result is one of the validation of why we predict outcome Y by means of OLS.

The above discussion is cumbersome, so I demonstrate the validity of OLS in a more straightforward way. We want to minimize

$$\mathbb{E}\left[(Y - X^T\beta)^2\right]$$

FOC implies

$$\begin{aligned} &\mathbb{E}\left[X(XY - X^T\hat{\beta})\right] = 0 \\ \iff &\hat{\beta} = \mathbb{E}\left[XX^T\right]^{-1}\mathbb{E}[XY] \end{aligned}$$

Summary

We considered how to predict Y via MSE criteria. We checked $P(X) = E[Y \mid X]$ gives the best prediction of Y within MSE. We have many way to approximate $E[Y \mid X]$ (If we know ture $E[Y \mid X]$, it is best!).

If we think linear prediction is reasonable, the OLS estimator achieves best prediction in the class of linear prediction. However, other approximation of $E[Y \mid X]$ may achieve smaller value of MSE than linear prediction.

$$\exists \hat{P}(X) \forall \beta \left(E\left[\left(Y - X^T\beta\right)^2\right] \geq \mathbb{E}\left[\left(Y - X^T\hat{\beta}\right)^2\right] \geq \mathbb{E}\left[\left(Y - \hat{P}(X)^2\right)^2\right] \geq \mathbb{E}\left[(Y - \mathbb{E}[Y \mid X])^2\right]\right)$$

1.2 Endogeneity

Consider the model

$$Y = X^T\beta + u.$$

suppose $\mathbb{E}[Xu] \neq 0$. Then we find

$$\hat{\beta} \to_p \beta + \mathbb{E}[XX^T]^{-1}\mathbb{E}[Xu].$$

It implies $\hat{\beta}$ is inconsistent under the endogeneity.

I provide some examples which endogeneity is violated.

- Omitted Variables
- Measurement Error
- Simultaneous Equation

1.2.1 Omitted variable Bias

Let the model

$$Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + u.$$

Suppose $\mathbb{E}[Xu] = 0(\mathbb{E}[u] = 0)$ and X_2 is unobservable. We have the model

$$Y = \beta_0^* + \beta_1^* X_1 + u^*,$$

with

$$\begin{aligned}
&\beta_0^* = \beta_0 + \beta_2\mathbb{E}[X_2] \\
&\beta_1^* = \beta_1 \\
&u^* = \beta_2(X_2 - \mathbb{E}[X_2]) + u
\end{aligned}$$

Note we normalized β_0^* to obtain $\mathbb{E}[u^*] = \mathbb{E}[\beta_2(X_2 - \mathbb{E}[X_2]) + u] = 0$. Then we have $\mathbb{E}[X_1u^*] = \mathbb{E}[X_1(\beta_2(X_2 - \mathbb{E}[X_2]) + u)] = \beta_2 Cov(X_1, X_2)$. Therefore, if $\beta_2 \neq 0$ and $Cov(X_1, X_2) \neq 0$, X_1 is endogenous.

$$\begin{aligned}
\hat{\beta} \to_p &\beta_1 + \mathbb{E}[X_1^2]^{-1}\mathbb{E}[X_1u^*] \\
= &\beta_1 + \beta_2\frac{Cov(X_1, X_2)}{Var(X_1)}
\end{aligned}$$

1.2.2 Measurement Error

Let

$$Y = \beta_0 + X^T\beta_1 + u.$$

Suppose $\mathbb{E}[Xu] = 0$. Consider the case X is not observable and only X^* is observed, where $X^* = X + v \in \mathbb{R}^K$. Assume $\mathbb{E}[V] = 0, Cov(X_1, \cdot, X_K, V) = 0$ and $Cov(u, v) = 0$. We have the model

$$Y = \beta_0^* + X^{T*}\beta_1 + u^*,$$

with

$$\begin{aligned}
&\beta_0^* = \beta_0 \\
&\beta_1^* = \beta_1 \\
&u^* = -v^T\beta_1 + u
\end{aligned}$$

Note

$$\mathbb{E}[X^* u^*] = -\mathbb{E}[X^* v^T]\beta_1 = -\mathbb{E}[vv^T]\beta_1$$

Therefore, if $Var(v) \neq 0$ and $\beta_1 \neq 0$, X^* is endogenous.

$$\hat{\beta^*} \to_p \left(1 - \mathbb{E}[X^* X^{T*}]^{-1}\mathbb{E}[vv^T]\right)\beta_1$$

If we regard X as a scalar, we obtain

$$\hat{\beta^*} \to_p \left(1 - \frac{Var(v)}{Var(X^*)}\right)\beta_1 < \beta_1$$

This bias is called the attenuation bias.

1.3 IV method and GMM

If we find the "instrumental variable (IV)", you can couple with a endogeneity problem. Instrumental variable satisfies

1. $\mathbb{E}[u \mid Z] = 0$ (Exogeneity)
2. $\mathbb{E}\left[ZX^T\right]$ is of full rank (Relevance).

1.3.1 IV estimator

Let

$$Y = X^T\beta + u.$$

Suppose X and Z are $K \times 1$ vector and Z satisfies the condtion of Instrumental variable.

We have the moment condition

$$\mathbb{E}[Z_i u_i] = 0 \iff \mathbb{E}\left[Z_i\left(Y_i - X_i^T\beta\right)\right].$$

Then we obtain IV estimator by minimizing following objective function.

$$\frac{1}{n}\sum_{i=1}^{N} Z_i\left(Y_i - X_i^T\beta\right)$$

Thus

$$\beta_{IV} = \left(\frac{1}{n}\sum_{i=1}^{N} Z_i X_i^T\right)^{-1}\frac{1}{n}\sum_{i=1}^{N} Z_i Y_i$$

1.3.2 GMM

When the dimension of Z is greater than or equal to X, we can consider the GMM estimator. The GMM estimator is defined as the minimizor of the following objective funciton.

$$\left(\frac{1}{n}\sum_{i=1}^{n} Z_i\left(Y_i - X_i^T\beta\right)\right)^T W\left(\frac{1}{n}\sum_{i=1}^{n} Z_i\left(Y_i - X_i^T\beta\right)\right)$$

,where W is a some positive definite matrix. The GMM estimator is

$$\hat{\beta}_{GMM} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i Z_i^T W \frac{1}{n}\sum_{i=1}^{n} Z_i X_i^T\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} X_i Z_i^T W \frac{1}{n}\sum_{i=1}^{n} Z_i Y_i$$

1.3.3 2SLS

Let $W = \left(\frac{1}{n}\sum_{i=1}^{n} Z_i Z_i^T\right)^{-1}$. Then we obtain "2SLS" (two step least square) estimator.

$$\hat{\beta}_{2SLS} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i Z_i^T \left(\frac{1}{n}\sum_{i=1}^{n} Z_i Z_i^T\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} Z_i X_i^T\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} X_i Z_i^T \left(\frac{1}{n}\sum_{i=1}^{n} Z_i Z_i^T\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} Z_i Y_i$$

2SLS asuumes that $var(e_i \mid Z_i) = var(e_i)$.

1.3.4 Efficient GMM

2 step least square estimation is widly used ,but 2SLS estimator is not efficient. By substituting W for $(var(Z_i e_i))^{-1}$ or an element of $(var(Z_i e_i))^{-1}$, we obtain the efficient estimator.

$$\hat{\beta}_{eGMM} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i Z_i^T \left(\hat{\mathbb{E}}\left[e_i^2 Z_i Z_i^T\right]\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} Z_i X_i^T\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} X_i Z_i^T \left(\hat{\mathbb{E}}\left[e_i^2 Z_i Z_i^T\right]\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} Z_i Y_i$$

1.4 Quantile Regression model

By using least absolute deviation estimator, we can estimate $Med(y|x)$.

$$\begin{aligned}
\min_a \mathbb{E}[|y-a|] &= -\int_{-\infty}^{a} (y-a)f(y)dy + \int_{a}^{\infty} (y-a)f(y)dy \\
&= -\int_{-\infty}^{a} yf(y)dy + \int_{-\infty}^{a} af(y)dy + \int_{a}^{\infty} yf(y)dy - \int_{a}^{\infty} af(y)dy
\end{aligned}$$

FOC implies

$$\begin{aligned}
0 &= -af(a) + \int_{-\infty}^{a} f(y)dy + af(a) - af(a) - \int_{a}^{\infty} f(y)dy + af(a) \\
&= F(a) - (1 - F(a)) \\
\iff F(a) &= \frac{1}{2}
\end{aligned}$$

It implies $a^* = \hat{m}(x)$ gives a median of distribution of y.
SOC is

$$f(a) + f(a) = 2f(a)$$

By assuming $f(a) > 0$, FOC gives a global minimizer. We can consider the more general case.

$$\mathbb{E}[|\phi_\theta(y - a)|]$$

,where $\phi_\theta(s) = -\theta s \mathbb{1}\{s \leq 0\} + (1 - \theta) s \mathbb{1}\{s > 0\}$

Exercise

Consider the data for $i = 1, \cdots, 1000$, $y_i = 1\{1 + x_i + u_i \geq 0\}$,where $u_i \sim N(0, \sigma^2)$, α_i and $v_i \sim N(0, 1)$,and $x_i = \alpha_i + v_i$. Solve the following quantile regression model for 50 percent quantile.

$$y_i = \alpha + x_i \beta + u_i$$

1. Estimate parameters without package.

2. Estimate with glm.

```
library(texreg)
library(quantreg)

#(1)
quantile_reg <- function(y_in, x_in, tau) {
  n <- length(y_in)
  fit <- lm(y ~ x)
  beta <- fit$coef

  f_quantile <- function(beta,y_in,x_in,tau) {
    xb <- cbind(1,x_in) %*% beta
    residuals <- y_in-xb
    quantile_loss <- sum(residuals[residuals>0])*(1-tau) - sum(residuals[
        residuals<=0])*tau
    return ((1/n)*quantile_loss)
  }

  result <- optim(par = beta, fn=f_quantile,y_in=y_in,x_in=x_in,tau = tau)
  return(result)
}
quantile_coef <- quantile_reg(y_in = data$y, x_in = data$x, tau = 0.5)

formatted_coef <- format(round(quantile_coef$par, digits = 4), scientific = FALSE
  )
print(formatted_coef)

#(2)
rq <- rq(y ~ 1 + x, tau = 0.5, data = data)
screenreg(rq)
```

I need to write how to construct confidence interval for quantile regression.

2 Non-linear MLE estimation

We overviwe some non-linear estimation method.

2.1 Binary Choice model

We consider the model $y_i^* = x_i^T \beta + u_i$. We only observe whether y_i^* is greate than 1 or not, i.e.

$$y_i = \begin{cases} 1 & \text{if } y_i^* \geq 0 \\ 0 & \text{if } y_i^* < 0 \end{cases}$$

This type of model is called as Binary Choice model.

2.1.1 Probit model

We assume the normality of the error term, i.e. $u_i \sim_{iid} N(0, \sigma_u^2)$. Then we obtain

$$\begin{aligned} Pr(y_i = 1|x) = Pr(y_i^* \geq 0|x) &= Pr(x_i^T \beta + u_i \geq 0|x) \\ &= Pr\left(\frac{x_i^T \beta}{\sigma_u} \geq -\frac{u_i}{\sigma_u} \mid x\right) \\ &= \Phi\left(\frac{x_i^T \beta}{\sigma_u}\right) \end{aligned}$$

,where $\Phi(\cdot)$ is the distribution function of the standard normal distribution. Note $Pr(y_i = 0|x) = 1 - Pr(y_i = 1|x)$. Then likelihood function is

$$\begin{aligned} L(\beta, \sigma_u) &= \prod_{i \in I} Pr(y_i = 1|x)^{\mathbb{1}\{y_i=1\}} \times Pr(y_i = 0|x)^{\mathbb{1}\{y_i=0\}} \\ &= \prod_{i \in I} \Phi\left(\frac{x_i^T \beta}{\sigma_u}\right)^{\mathbb{1}\{y_i=1\}} \times \left(1 - \Phi\left(\frac{x_i^T \beta}{\sigma_u}\right)\right)^{\mathbb{1}\{y_i=0\}} \end{aligned}$$

Log-likelihood function is

$$logL(\beta, \sigma_u) = \sum_{i \in I} \left[\mathbb{1}\{y_i = 1\} log\Phi\left(\frac{x_i^T \beta}{\sigma_u}\right) + \mathbb{1}\{y_i = 0\} log\left(1 - \Phi\left(\frac{x_i^T \beta}{\sigma_u}\right)\right)\right]$$

Exercise

Consider the model for $i = 1, \cdots, 1000$,

$$y_i = 1\{1 + x_i + u_i \geq 0\}$$

,where $u_i \sim N(0, \sigma^2)$, α_i and $v_i \sim N(0, 1)$,and $x_i = \alpha_i + v_i$.

1. Estimate parameters without glm package.

2. Estimate with glm.

```r
# Data generation#
set.seed(123)
n <- 1000; u <- rnorm(n);alpha <- rnorm(n); v <- rnorm(n)
x <- alpha + v
y <- ifelse(1 + x + u >=0,1,0)
data <- data.frame(x,y)

# (1)#
lm_1 <- lm(y~x)
f_probit <- function(beta, y, x){
  xb <- cbind(1,x) %*% beta
  lik <- pnorm(xb)^y*(1-pnorm(xb))^(1-y)
  return(-sum(log(lik)))
  }
optim(par=lm_1$coefficients,fn=f_probit,y=y,x=x)$par

# (2)#
model_2 <- glm(y ~ 1 + x, family =
 binomial("probit"), data =data)
model_2$coef
```

2.1.2 Random Utility Model

Suppose there are J goods and the man n wants to choose one good he buys within J alternatives. For example, in the restaurant, he chooses to drink beer or wine. In the case, the choice set C is $\{beer, wine\}$. Now we consider the case he chooses only one good.

We express the utility that person n chooses a option i within a choice set as U_{ni}. If n chooses i, then it implies for all $j \neq i, U_{ni} > U_{nj}$ (Assume $Pr(ties) = 0$). In the example, we observed that he chosen beer. Then we guess $U_{n,beer} > U_{n,wine}$ holds.

U_{ni} is often decomposed to $V(x_{ni}, s_n) + \varepsilon_{ni}$, where

- V is a representative utility
- x_{ni} is observed attributes of the good n.
- s_n is observed n's characteristics.
- ε_{ni} is unobserved idiosyncratic taste of person n for the good i.

Let $\varepsilon_n = (\varepsilon_{n1}, \cdots, \varepsilon_{nJ})$ and joint density function is $f(\varepsilon_n)$. We want to know the probability n chooses i.

To obtain closed form solution late, suppose ε_{ni} and ε_{nj} are iid Type 1 Extreme Value distribution. Then $\varepsilon_{nj} - \varepsilon_{ni}$ follows Logisitic distribution.

$$
\begin{aligned}
P_{ni} &= Pr(\forall j \neq i, U_{ni} > U_{nSj}) \\
&= Pr(\forall j \neq i, \varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj}) \\
&= \int_{\varepsilon} 1(\forall j \neq i, \varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj}) f(\varepsilon_n) d\varepsilon_n \\
&= \int \left(\prod_{j \neq i} exp(-exp(-\varepsilon_{ni} + V_{ni} - V_{nj})) exp(-\varepsilon_{ni}) \right) exp(-exp(-\varepsilon_{ni})) d\varepsilon_{ni} \\
&= \frac{exp(V_{ni})}{\sum_j exp(V_{nj})}
\end{aligned}
$$

2.2 Censored Tobit model

2.2.1 Parametric assumption

We firstly assume the normality of the error term. In the next section, we relax this normality assumption. Censored Tobit model is characterized as

$$
\begin{aligned}
y_i^* &= x_i^T \beta + u_i \\
y_i &= \begin{cases} y_i^* & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* \leq 0 \end{cases} \\
u_i &\sim_{iid} N(0, \sigma_u^2)
\end{aligned}
$$

Under the normality of the error term, we have

$$
\begin{aligned}
F_y(t \mid x) &= Pr(y_i \leq t \mid x_i) \\
&= Pr\left(\frac{u_i}{\sigma_u} \leq \frac{t - x_i^T \beta}{\sigma_u} \mid x \right) = \Phi\left(\frac{t - x_i^T \beta}{\sigma_u} \right)
\end{aligned}
$$

Thus we obtain

$$
f_y(t \mid x) = \frac{1}{\sigma_u} \phi\left(\frac{t - x_i^T \beta}{\sigma_u} \right)
$$

The likelihood funtion is

$$
\begin{aligned}
L(\beta, \sigma_u) &= \prod_{i \in I} Pr(y_i > 0 | x)^{\mathbb{1}\{y_i > 0\}} \times Pr(y_i = 0 | x)^{\mathbb{1}\{y_i = 0\}} \\
&= \prod_{i \in I} \left[\frac{1}{\sigma_u} \phi\left(\frac{y_i - x_i^T \beta}{\sigma_u} \right) \right]^{\mathbb{1}\{y_i > 0\}} \times \left[1 - \Phi\left(\frac{x_i^T \beta}{\sigma_u} \right) \right]^{\mathbb{1}\{y_i = 0\}}
\end{aligned}
$$

Log-likelihood funtion is

$$
logL(\beta, \sigma_u) = \sum_{i \in I} \left[\mathbb{1}\{y_i > 0\} \left[log\phi\left(\frac{y_i - x_i^T \beta}{\sigma_u} \right) - log(\sigma_u) \right] + \mathbb{1}\{y_i = 0\} log\left(1 - \Phi\left(\frac{x_i^T \beta}{\sigma_u} \right) \right) \right]
$$

Exercise

Consider the model for $i = 1, \cdots, 1000$,

$$
\begin{aligned}
y_i^* &= 1 + x_{1i} + x_{2i} + u_i \\
y_i &= \begin{cases} y_i^* & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* \leq 0 \end{cases} \\
u_i &\sim_{iid} N(0,1)
\end{aligned}
$$

and the generate the data x_{1i} and $x_{2i} \sim N(0,1)$.

1. Estimate parameters without glm package.
2. Estimate with censReg package.
3. Change to $u_i \sim N(0,2)$. What happens on coefficients.
4. Change the initial value and confirm it does not affect the solution.

```
# Data Generation#
set.seed(123)
n <- 1000; u <- rnorm(n,0,1); x1 <- rnorm(n,0,1); x2 <- rnorm(n,0,1)
y_star <- 1 + x1 + 2*x2 + u
y <- ifelse(y_star > 0, y_star, 0)
df_cen <- data.frame(y,y_star,x1,x2)

#(1)
Censored_Reg <- function(y_in, x_in) {
  x <- cbind(1,as.matrix(x_in))
  parameters <- c(rep(1, ncol(x)+1))

  f_censored <- function(parameters, y_in, x_in) {
    beta <- parameters[2:4]
    sigma <- parameters[1]
    xb <- cbind(1,as.matrix(x_in)) %*% beta
    z <- (y_in - xb) / sigma

    y.indic <- ifelse(y_in > 0, 1, 0)

    log_lik <- sum(y.indic*(log(dnorm(z))-log(sigma)) + (1 - y.indic)*log(1 -
       pnorm(xb/sigma)))
    return(-log_lik)
  }

  result <- optim(par = parameters, fn = f_censored, y_in = y_in, x_in = x_in)
  return(result)
}

# Results # -
result <- Censored_Reg(y_in = df$y, x_in = df[,c("x1","x2")])
print(result)

# (2) #
model_cr = censReg(y~1 + x1 + x2, left = 0, data = df)
stargazer(type = "text",model_cr)
```

2.2 Application

$$
\begin{aligned}
y_i^* &= x_i^T\beta + u_i \\
y_i &= \begin{cases} y_i^* & \text{if } T\, y_i^* > 0 \\ T & \text{if } y_i^* \geq T \\ 0 & \text{if } y_i^* \leq 0 \end{cases} \\
u_i &\sim_{iid} N(0, \sigma_u^2)
\end{aligned}
$$

We find

$$
\begin{aligned}
Pr(y_i = T \mid x_i) &= Pr(y_i^* \geq T \mid x_i) \\
&= Pr\left(\frac{u_i}{\sigma_u} \geq \frac{T - x_i^T\beta}{\sigma_u} \mid x\right) = 1 - \Phi\left(\frac{T - x_i^T\beta}{\sigma_u}\right)
\end{aligned}
$$

Thus the likelihood funtion is

$$
\begin{aligned}
L(\beta, \sigma_u) &= \prod_{i\in I} Pr(y_i = T|x)^{\mathbb{1}\{y_i=T\}} \times \prod_{i\in I} Pr(T > y_i > 0|x)^{\mathbb{1}\{T>y_i>0\}} \times Pr(y_i = 0|x)^{\mathbb{1}\{y_i=0\}} \\
&= \prod_{i\in I} \left[1 - \Phi\left(\frac{T - x_i^T\beta}{\sigma_u}\right)\right]^{\mathbb{1}\{y_i=T\}} \left[\frac{1}{\sigma_u}\phi\left(\frac{y_i - x_i^T\beta}{\sigma_u}\right)\right]^{\mathbb{1}\{T>y_i>0\}} \times \left[1 - \Phi\left(\frac{x_i^T\beta}{\sigma_u}\right)\right]^{\mathbb{1}\{y_i=0\}}
\end{aligned}
$$

2.2.2 Censored LAD

We relax the normality assumption.

$$
\begin{aligned}
y_i^* &= x_i^T\beta + u_i \\
y_i &= \begin{cases} y_i^* & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* \leq 0 \end{cases} \\
Med[u_i|x] &= x_i^T\beta
\end{aligned}
$$

Thus, we have $Med[y_i|x] = max\{0, x_i^T\beta\}$.
Objective function is

$$
\frac{1}{n}\sum_{i\in I} |y_i - max\{0, x_i^T\beta\}|
$$

The minimizer of this function estimates $Med[y_i|x]$ consistently. The asymptotoic distribution is

$$
\sqrt{n}(\hat{\beta} - \beta) \to^d N\left(0, \lim_{N\to\infty} C_T^{-1} M_T C_T^{-1}\right)
$$

,where

$$
\begin{aligned}
C_T &= E\left[\frac{2}{N}\sum_{i=1}^{N} f_u(u_i = 0|x) \cdot 1(x_i^T\beta > 0) x_i x_i^T\right] \\
M_T &= E\left[\frac{1}{N}\sum_{i=1}^{N} 1(x_i^T\beta > 0) x_i x_i^T\right]
\end{aligned}
$$

2.3 Truncated Tobit model

$$
\begin{aligned}
y_i^* &= x_i^T \beta + u_i \\
y_i &= \begin{cases} y_i^* & \text{if } y_i^* > 0 \\ missing & \text{if } y_i^* \leq 0 \end{cases} \\
u_i &\sim_{iid} N(0, \sigma_u^2)
\end{aligned}
$$

$$
\begin{aligned}
Pr(0 \leq y_i \leq t \mid x, y_i > 0) &= \frac{Pr(0 \leq y_i \leq t \mid x)}{Pr(y_i > 0 \mid x)} \\
&= \frac{\Phi\left(\frac{t - x_i^T \beta}{\sigma_u}\right) - \Phi\left(\frac{x_i^T \beta}{\sigma_u}\right)}{\Phi\left(\frac{x_i^T \beta}{\sigma_u}\right)}
\end{aligned}
$$

$$
f_{Y|X,Y>0}(t \mid x, y_i > 0) = \frac{1}{\sigma_u} \frac{\phi\left(\frac{t - x_i^T \beta}{\sigma_u}\right)}{\Phi\left(\frac{x_i^T \beta}{\sigma_u}\right)}
$$

The likelihood funtion is

$$
\begin{aligned}
L(\beta, \sigma_u) &= \prod_{i \in I} Pr(y_i > 0 | x) \\
&= \prod_{i \in I} \left(\frac{1}{\sigma_u} \frac{\phi\left(\frac{y_i - x_i^T \beta}{\sigma_u}\right)}{\Phi\left(\frac{x_i^T \beta}{\sigma_u}\right)} \right)
\end{aligned}
$$

Log-likelihood funtion is

$$
logL(\beta, \sigma_u) = \sum_{i \in I} \left(-log(\sigma_u) + log\phi\left(\frac{y_i - x_i^T \beta}{\sigma_u}\right) - log\Phi\left(\frac{x_i^T \beta}{\sigma_u}\right) \right)
$$

2.4 Sample Selection model

We consider the model for $i = 1, \cdots, n$,

$$
\begin{aligned}
y_i^* &= x_i^T \beta + u_i \\
S_i^* &= z_i^T \theta + v_i \\
S_i &= \begin{cases} 1 & \text{if } z_i^T \theta + v_i > 0 \\ 0 & \text{if } z_i^T \theta + v_i \leq 0 \end{cases} \\
y_i &= \begin{cases} y_i^* & \text{if } z_i^T \theta + v_i > 0 \iff S_i = 1 \\ 0 & \text{if } z_i^T \theta + v_i \leq 0 \iff S_i = 0 \end{cases} \\
\begin{pmatrix} u_i \\ v_i \end{pmatrix} &\sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_u^2 & \rho\sigma_u\sigma_v \\ \rho\sigma_u\sigma_v & \sigma_v^2 \end{pmatrix} \right)
\end{aligned}
$$

The likelihood function is

$$
\begin{aligned}
&logL(\beta^T, \theta^T, \sigma_u, \sigma_v, \rho) \\
&= \prod_{i \in I} f(y, S \mid x_i, z_i) \\
&= \prod_{i \in I} f(y, S = 1 \mid x_i, z_i)^{\mathbb{1}\{D_i\}} \times Pr(S_i = 0 \mid x_i, z_i)^{\mathbb{1}\{1-S_i\}}
\end{aligned}
$$

Note

$$
\begin{aligned}
f(y, S = 1 \mid x_i, z_i) &= \int_0^\infty f(y_i, s_i^* \mid x_i, z_i) ds^* \\
&= f(y_i \mid x_i, z_i) \int_0^\infty f(s_i^* \mid y_i, x_i, z_i) ds^*
\end{aligned}
$$

Also, we know

$$
\begin{aligned}
v_i \mid y_i, x_i, z_i &= z_i^T \theta + v_i \mid y_i, x_i, z_i \\
&= z_i^T \theta + v_i \mid y_i, x_i, z_i, u_i
\end{aligned}
$$

u_i and v_i follow to bivariate normal distribution, then there exists $\varepsilon_i \sim N(\mu_\varepsilon, \sigma_\varepsilon)$. Therefore

$$
v_i \mid y_i, x_i, z_i, u_i = \frac{\rho \sigma_u \sigma_v}{\sigma_v^2} u_i + \varepsilon_i \mid y_i, x_i, z_i, u_i
$$

, note that the coefficient of the u_i is the just that of OLS. $\mathbb{E}[u_i] = 0$ and $\mathbb{E}[v_i] = 0$ implies $\mathbb{E}[\varepsilon_i] = 0$. It leads to $\text{Var}[\varepsilon_i] = \mathbb{E}[(\varepsilon_i - \mathbb{E}[\varepsilon_i])^2] = \mathbb{E}[\varepsilon_i^2] = \mathbb{E}[(v_i - \frac{\rho\sigma_u\sigma_v}{\sigma_v^2} u_i)^2] = \sigma_v^2 - 2\rho^2\sigma_u^2 + \rho^2\sigma_u^2 = (1-\rho^2)\sigma_u^2$.

Hence we obtain

$$
\begin{aligned}
S_i^* \mid y_i, x_i, z_i &= z_i^T \theta + v_i \mid y_i, x_i, z_i \\
&= z_i^T \theta + \frac{\rho \sigma_u \sigma_v}{\sigma_v^2} u_i + \varepsilon_i \mid y_i, x_i, z_i, u_i \\
&\sim N(z_i^T \theta + \frac{\rho \sigma_u \sigma_v}{\sigma_v^2}(y_i - x_i^T \beta), (1-\rho^2)\sigma_u^2)
\end{aligned}
$$

,then it implies

$$
f(y, S = 1 \mid x_i, z_i) = \frac{1}{\sigma_u} \phi\left(\frac{y_i - x_i^T \beta}{\sigma_u}\right) \Phi\left(\frac{z_i^T \theta + \frac{\rho\sigma_u\sigma_v}{\sigma_v^2}(y_i - x_i^T \beta)}{\sqrt{(1-\rho^2)\sigma_u^2}}\right)
$$

Therefore the likelihood function is

$$
\begin{aligned}
&logL(\beta^T, \theta^T, \sigma_u, \sigma_v, \rho) \\
&= \prod_{i \in I} \left[\frac{1}{\sigma_u} \phi\left(\frac{y_i - x_i^T \beta}{\sigma_u}\right) \Phi\left(\frac{z_i^T \theta + \frac{\rho\sigma_u\sigma_v}{\sigma_v^2}(y_i - x_i^T \beta)}{\sqrt{(1-\rho^2)\sigma_u^2}}\right)\right]^{\mathbb{1}\{S_i\}} \times \left[1 - \Phi\left(\frac{z_i^T \theta}{\sigma_v}\right)\right]^{\mathbb{1}\{1-S_i\}}
\end{aligned}
$$

Log-likelihood function is

$$
\begin{aligned}
&logL(\beta^T, \theta^T, \sigma_u, \sigma_v, \rho) \\
&= \sum_{S_i=1} \left[-log\sigma_u + log\phi\left(\frac{y_i - x_i^T\beta}{\sigma_u} \right) + log\Phi\left(\frac{z_i^T\theta + \frac{\rho\sigma_u\sigma_v}{\sigma_v^2}(y_i - x_i^T\beta)}{\sqrt{(1-\rho^2)\sigma_u^2}} \right) \right] \\
&+ \sum_{S_i=0} log\left(1 - \Phi\left(\frac{z_i^T\theta}{\sigma_v} \right) \right)
\end{aligned}
$$

2.5 Duration model

Hazard function and distribution function of dependent variable have the following relationship.

$$
\begin{aligned}
\lambda(t_i) &:= \frac{f(t_i)}{1-F(t_i)} = -\frac{\partial}{\partial t}[log(1-F(t_i)] \\
F(t_i) &= 1 - exp\left(-\int_0^{t_i} \lambda(s_i)ds \right)
\end{aligned}
$$

Therefore, if $\lambda(t_i) = exp(x_i^T\beta)$,

$$
\begin{aligned}
\lambda(t_i) &= exp(x_i^T\beta) \\
F(t_i) &= 1 - exp\left(-\int_0^{t} exp(x_i^T\beta)ds \right) = 1 - exp(-exp(x_i^T\beta)t_i) \\
f(t_i) &= \lambda(t_i) \times (1 - F(t_i)) = exp(x_i^T\beta) \times exp(-exp(x_i^T\beta)t_i)
\end{aligned}
$$

2.6 Dynamic Programming model

`http://www.its.caltech.edu/~mshum/gradio/zurcher.pdf`
`https://github.com/QuentinAndre/John-Rust-1987-Python`
We consider Optimal Stopping problem introduced in Rust(1987). x_t donates the mileage of the engine, and i_t is the dummy variable if the engine is replaced, it takes 1, otherwise 0.

$$
i_t = \begin{cases} 1 \text{ if the engine is replaced} \\ 0 \text{ if not} \end{cases}
$$

The flow cost is characterized as

$$
RC \times i_t + c(x_t \mid i_t, \theta_1) + \varepsilon(i_t)
$$

Rust(1987) considered several form of $c(x_t \mid i_t, \theta_1)$

$$
c(x_t \mid i_t, \theta_1) = \begin{cases} \theta_{11}x_t + \theta_{12}x_t^2 + \theta_{13}x_t^3 \\ \theta_{11}exp(\theta_{12}x_t) \\ \frac{\theta_{11}}{91-x_t} \\ \theta_{11}\sqrt{x_t}. \end{cases}
$$

We use $c(x_t \mid i_t, \theta_1) = exp(\theta_{12} x_t (1 - i_t))$ as cost function.

Then, we can define utility function as

$$u(x_t, i_t \mid \theta) = -RC \times i_t - \theta_{11} exp(\theta_{12} x_t (1 - i_t)) + \varepsilon(i_t)$$

The stochastic process of $\{x_t\}$ is

$$p(x_{t+1} \mid x_t, i_t, \theta_2) = \begin{cases} \theta_2 exp(\theta_2 (x_{t+1} - x_t)) \text{ if } i_t = 0 \text{ and } x_{t+1} \geq x_t \\ \theta_2 exp(\theta_2 x_{t+1}) \text{ if } i_t = 1 \text{ and } x_{t+1} \geq 0 \\ 0 \text{ otherwise.} \end{cases}$$

Let $\theta = (\theta_1^T, \theta_2, RC, \beta)$. The value function is

$$\begin{aligned} V(x, i) &= max\{u(x, i \mid \theta) + \beta \mathbb{E}_{x'|x} V(x')\} \\ &= max\{u(x, 1 \mid \theta) + \beta \mathbb{E}_{x'|x} V(0, i'), u(x, 0 \mid \theta) + \beta \mathbb{E}_{x'|x} V(x', i')\}\} \end{aligned}$$

Likelihood function is

$$\begin{aligned} L(x_1, \cdots, x_T, i_1, \cdots, i_T \mid x_0, i_0, \theta) &= \prod_{t=1}^{T} Pr(i_t, x_t \mid x_0, \cdots, x_{t-1}, i_0, \cdots, i_{t-1}; \theta) \\ &= \prod_{t=1}^{T} Pr(i_t, x_t \mid x_{t-1}, i_{t-1}, ; \theta) \\ &= \prod_{t=1}^{T} Pr(i_t, \mid x_t; \theta) \times Pr(x_t, \mid x_{t-1}, i_{t-1}; \theta) \end{aligned}$$

Estimation

We estimate parameters by following 2 step.
Step1: Estimate θ_2, which gorverns stochastic process of x.
Step2: Estimate $(\theta_{11}, \theta_{12}, RC, \beta)$ by ML.

Step1

Let

$$\Delta x_{t+1} = \begin{cases} x_{t+1} - x_t & \text{if } D_t = 0 \\ x_{t+1} & \text{if } D_t = 1 \end{cases}$$

Likelihood function is

$$\begin{aligned} L(\Delta x_{i,t-1}, \cdots, \Delta x_{i,1}, i_{i,t}, \cdots, i_{i,1}) &= \prod_{i \in I} \prod_{s=1}^{t_{i-1}} p(\Delta x_{i,t} \mid \Delta x_{i,s-1}, \cdots, \Delta x_{i,1}, i_{i,s}, \cdots, i_{i,1}) \\ &= \prod_{i \in I} \prod_{s=1}^{t_{i-1}} p(\Delta x_{i,t} \mid x_{i,s-1}, i_{i,s}) \\ &= \prod_{i \in I} \prod_{s=1}^{t_{i-1}} \theta_2 exp(-\theta_2 \Delta x_{i,s}) \end{aligned}$$

Step2

$$
\begin{aligned}
&Pr(i_t = 1 \mid x_t m\theta) \\
&= Pr(V(x, i_t = 1) > V(x, i_t = 0) \mid x_t\theta) \\
&= Pr(u(x, 1 \mid \theta) + \beta\mathbb{E}_{x'|x}V(0, i') > u(x, 0 \mid \theta) + \beta\mathbb{E}_{x'|x}V(x', i') \mid x_t\theta) \\
&= Pr(-RC - \theta_{11} + \varepsilon_t(1) + \beta\mathbb{E}_{x'|x}V(0, i') > -\theta_{11}exp(\theta_{12}x_t) + \varepsilon_t(0) + \beta\mathbb{E}_{x'|x}V(x', i') \mid x_t\theta) \\
&= Pr(\varepsilon_t(1) - \varepsilon_t(0) > RC + \theta_{11} - \beta\mathbb{E}_{x'|x}V(0, i') - \theta_{11}exp(\theta_{12}x_t) + \beta\mathbb{E}_{x'|x}V(x', i') \mid x_t\theta)
\end{aligned}
$$

3 Nonparametric model

Let $X_i = (X_{1i}, \cdots, X_{Ki})^T \in \mathbb{R}^K$. We consider the model

$$Y_i = m(X_i) + \varepsilon_i$$

We assume $E[\varepsilon_i | X_i] = 0$ and $E[\varepsilon_i^2 | X_i] \leq N < \infty$.
There are two approaches to estimate $m(X)$. They are local and global approach.

3.1 Local approach(Kernel Regression)

We introduce kernel function to conduct Local approach. Therefore, local approach include the kernel regression.

3.1.1 Local constant estimation

Let $X_i = (X_{1i}, \cdots, X_{Ki})^T \in \mathbb{R}^K$.

$$\begin{aligned} Y_i &= m(X_i) + \varepsilon_i \\ &= m(x) + \varepsilon_i + m(X_i) - m(x) \end{aligned}$$

,where $m(x) = E[Y_i | X_i = x]$.
If $X_i \approx x$ and $m(\cdot)$ is smooth, then $m(X_i) - m(x) \approx 0$ hold. In the case, we have $Y_i \approx m(x) + \varepsilon_i$.
Objective function is

$$\sum_{i \in I} (Y_i - m(x))^2 K\left(\frac{X_i - x}{h}\right)$$

,where $m(x) = E[Y_i | X_i = x]$ is constant. $K(\cdot)$ is a **Kernel function** and $h \in \mathbb{R}^K$ is called as **bandwith**, **window width**,or **smoothing parameter**. $K(\cdot)$ is caluculated as follow

$$K\left(\frac{X_i - x}{h}\right) = K\left(\frac{X_{1i} - x_1}{h_1}\right) \times \cdots \times K\left(\frac{X_{Ki} - x_K}{h_K}\right).$$

Note this is the weighted least square estimation. By minimizing the Objective funtion, we obtain

$$\begin{aligned} \sum_{i \in I} Y_i K\left(\frac{X_i - x}{h}\right) &= \sum_{i \in I} \hat{m}_C(x, h) K\left(\frac{X_i - x}{h}\right) \\ \iff \hat{m}_C(x, h) &= \frac{\sum_{i = \in I} Y_i K\left(\frac{X_i - x}{h}\right)}{\sum_{i = \in I} K\left(\frac{X_i - x}{h}\right)} \end{aligned}$$

$\hat{m}_C(x, h)$ is called **Nadayara-Watson(NW) estimator**, **Kernel regression estimator** or **local constant estimator**.
Let $w_{ni}^C(x) = \frac{K\left(\frac{X_i - x}{h}\right)}{\sum_{j \in I} K\left(\frac{X_j - x}{h}\right)}$. Note $\sum_{i = \in I} w_{ni}^C(x) = 1$. Then we have

$$\hat{m}_C(x, h) = \sum_{i = \in I} w_{ni}^C(x) Y_i$$

From this notation, we can easily find $\hat{m}_C(x, h)$ is a linear estimator of Y_i, for $i = 1, \cdots, n$.
In practice, Gaussian kernel or Epaniechnikov kernel is commonly used. Especially, Gaussian kernel is recommended because the derivatives of estimator have any orders (Hansen(2019).

Gaussian	$K\Big(\frac{X_i - x}{h}\Big) = \frac{1}{\sqrt{2\pi}} exp\Big(-\frac{1}{2}(\frac{X_i - x}{h})^2\Big)$
Epanechnikov	$K\Big(\frac{X_i - x}{h}\Big) = \frac{3}{4\sqrt{5}}\Big(1 - \frac{1}{5}(\frac{X_i - x}{h})^2\Big)$ if $\lvert\frac{X_i - x}{h}\rvert < \sqrt{5}$, 0 otherwise.

I provide R code for one dimension case. Since local approach is WLS, we can follow same steps as we conduct OLS.

```
---Data generation---
n <- 1000; x <- runif(n,0,1); epsilon_i <- rnorm(n,0,1)
y <- 1 + x + exp(-x) + epsilon_i
data <- data.frame(x,y)


---Create Kernel---
GaussianK <- function(x_data, x, h) {
  u <- (x_data - x) / h
  K <- (1/sqrt(2*pi))*exp(-(u^2)/2)
  return(K)
}
GK <- GaussianK(x, 0.5, 0.1)

EpanechnikovK <- function(x_data, x, h) {
  u <- (x_data - x) / h
  K <- numeric(length(u))
  condition <- which(abs(u) < sqrt(5))
  K[condition] <- (3/(4*sqrt(5))*(1-(u[condition]^2)/5))
  return(K)
}
EK <- EpanechnikovK(x,0.5,0.1)


---Estimation---
lm_robust(y ~ 1, weights = GK, se_type = "HC0",data = data)
lm_robust(y ~ 1, weights = EK, se_type = "HC0",data = data)
```

	Estimate	Std. Error	t value	$\Pr(>\lvert t\rvert)$	CI Lower	CI Upper
Gaussian	2.086	0.051	40.91	1.065×10^{-215}	1.986	2.186
Epanechnikov	2.077	0.04996	41.57	4.553×10^{-220}	1.979	2.175

Table 1: NW estimator

3.1.2 Local linear estimation

Let $X_i = (X_{1i}, \cdots, X_{Ki})^T \in \mathbb{R}^K$, $\alpha = m(x)$ and $\beta = \nabla m(x) = \Big(\frac{\partial m(x)}{\partial x_1}, \cdots, \frac{\partial m(x)}{\partial x_K}\Big)^T$. We consider the model
Objective function is

$$\sum_{i \in I}(Y_i - \alpha - (X_i - x)^T\beta)^2 K\Big(\frac{X_i - x}{h}\Big)$$

It is convenient to write down as WLS.

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, X = \begin{pmatrix} 1 & (X_1 - x)^T \\ 1 & (X_2 - x)^T \\ \vdots & \vdots \\ 1 & (X_n - x)^T \end{pmatrix}, W = \begin{pmatrix} K\left(\frac{X_1 - x}{h}\right) & 0 & \cdots & 0 \\ 0 & K\left(\frac{X_2 - x}{h}\right) & \ddots & 0 \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & (X_n - x)^T \end{pmatrix}, \beta(x) = \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{pmatrix}$$

Hence, we obtain

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = (X^T W X)^{-1} X^T W Y$$

3.1.3 Local Polynomial estimation

Fix the degree p. Then

$$\begin{aligned} Y_i &= m(X_i) + \varepsilon_i \\ &\approx m(x) + (X_i - x)^T \nabla m(x) + \cdots + \frac{((X_i - x)^p)^T}{p!} \nabla^p m(x) + \varepsilon_i \end{aligned}$$

By following same argument to Local linear estimation, Let

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, X = \begin{pmatrix} 1 & (X_1 - x)^T & \cdots & \frac{((X_1 - x)^p)^T}{p!} \\ 1 & (X_2 - x)^T & \cdots & \frac{((X_2 - x)^p)^T}{p!} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & (X_n - x)^T & \cdots & \frac{((X_p - x)^p)^T}{p!} \end{pmatrix}, \beta(x) = \begin{pmatrix} \alpha \\ \nabla m(x) \\ \nabla^2 m(x) \\ \vdots \\ \nabla^p m(x) \end{pmatrix}$$

We obtain

$$\begin{pmatrix} \hat{\alpha} \\ \nabla \hat{m}(x) \\ \nabla^2 \hat{m}(x) \\ \vdots \\ \nabla^p \hat{m}(x) \end{pmatrix} = (X^T W X)^{-1} X^T W Y$$

In R, dnorm represents PDF.
pnorm represents CDF.

$$pnorm(x) = \Phi(x)$$
$$dnorm(x) = \phi(x)$$

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i z_i + u_i$$