Problem Set 1

MOEC0021 Empirical Methods

Fenqi Guo Wenjie Tu

Fall Semester 2020

The Classical Linear Regression Model

1. Theory - Using the CLRM to Make Predictions

$$y_i = x_i'\beta + \varepsilon_i$$

- x_i and β are vectors of dimensions $K \times 1$.
- y_i and ε_i are scalars.

1(a)

Economic context:

- y_i: the score individual i achieves for the *Empirical Methods* course.
- x_i: how many hours invested in this course per week, course attendence, prior knowledge in Econometrics (binary variable).
- ε_i: error term.

1(b)

For the rest of the exercise, CLRM assumptions hold. In particular, $\varepsilon|\mathbf{X} \sim \mathcal{N}(0, \sigma^2 I_{n+1})$, where we define $\varepsilon = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_{n+1})'$ and $\mathbf{X}' = (x_1, x_2, \cdots, x_{n+1})$.

$$\begin{aligned}
\mathbb{E}(y_i|x_i) &= \mathbb{E}(x_i'\beta + \varepsilon_i|x_i) \\
&= x_i'\beta + \mathbb{E}(\varepsilon_i|x_i) \\
&= x_i'\beta
\end{aligned}$$

$$\begin{aligned}
\text{Var}(y_i|x_i) &= \mathbb{E}\left((y_i - \mathbb{E}[y_i|x_i])^2|x_i\right) \\
&= \mathbb{E}\left((y_i - x_i'\beta)^2|x_i\right) \\
&= \mathbb{E}(\varepsilon_i^2|x_i) \\
&= \text{Var}(\varepsilon_i|x_i) \\
&= \sigma^2
\end{aligned}$$

1(c)

$$\mathbb{E}(y_i|x_i) = x_i'\beta$$

The conditional expectation of individual i' score given x_i is a linear function of how many hours invested in this course per week, course attendence, and prior knowledge in Econometrics.

$$\text{Var}(y_i|x_i) = \sigma^2$$

The conditional variance of individual i' score given x_i remains constant.

1(d)

$$\begin{aligned}
\hat{\beta}_n &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}y \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \varepsilon) \\
&= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon \\
\mathbb{E}(\hat{\beta}_n|\mathbf{X}) &= \mathbb{E}(\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon|\mathbf{X}) \\
&= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}(\varepsilon|\mathbf{X}) \\
&= \beta
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}(\hat{e}_{n+1}|\mathbf{X}) &= \mathbb{E}(y_{n+1} - \hat{y}_{n+1}|\mathbf{X}) \\
&= \mathbb{E}\left(x_{n+1}'\beta + \varepsilon_{n+1} - x_{n+1}'\hat{\beta}_n|\mathbf{X}\right) \\
&= \mathbb{E}(x_{n+1}'\beta|\text{X}) + \mathbb{E}(\varepsilon_{n+1}|\mathbf{X}) - \mathbb{E}(x_{n+1}'\hat{\beta}_n|\mathbf{X}) \\
&= x_{n+1}'\beta - x_{n+1}'\mathbb{E}(\hat{\beta}_n|\mathbf{X}) \\
&= x_{n+1}'\beta - x_{n+1}'\beta \\
&= 0
\end{aligned}$$

- Conditional expectation function is correctly specified.
- Estimate of β is unbiased.

1(e)

$$\begin{aligned}
\text{Var}(\hat{e}_{n+1}|\mathbf{X}) &= \text{Var}(y_{n+1} - \hat{y}_{n+1}|\mathbf{X}) \\
&= \text{Var}(x_{n+1}'\beta + \varepsilon_{n+1} - x_{n+1}'\hat{\beta}_n|\mathbf{X}) \\
&= \text{Var}(\varepsilon_{n+1} - x_{n+1}'\hat{\beta}_n|\mathbf{X}) \\
&= \text{Var}(\varepsilon_{n+1}|\mathbf{X}) + \text{Var}(x_{n+1}'\hat{\beta}_n|\mathbf{X}) - 2\text{Cov}(\varepsilon_{n+1}, x_{n+1}'\hat{\beta}_n|\mathbf{X}) \\
&= \sigma^2 + x_{n+1}'\text{Var}(\hat{\beta}_n|\mathbf{X})x_{n+1} \\
&= \sigma^2 + x_{n+1}'\left(\sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right)x_{n+1} \\
&= \sigma^2\left(1 + x_{n+1}'(\mathbf{X}'\mathbf{X})^{-1}x_{n+1}\right) \\
&= \text{Var}(y_i|x_i)\left(1 + x_{n+1}'(\mathbf{X}'\mathbf{X})^{-1}x_{n+1}\right) \\
\frac{\text{Var}(\hat{e}_{n+1}|\mathbf{X})}{\text{Var}(y_i|x_i)} &= 1 + \underbrace{x_{n+1}'(\mathbf{X}'\mathbf{X})^{-1}x_{n+1}}_{\text{positive}} \\
\text{Var}(\hat{e}_{n+1}|\mathbf{X}) &> \text{Var}(y_i|x_i)
\end{aligned}$$

1(f)

$$\lim_{n\to\infty} \text{Var}(y_i|x_i) = \sigma^2$$

$$\lim_{n\to\infty} \text{Var}(\hat{e}_{n+1}|\mathbf{X}) = \lim_{n\to\infty} \sigma^2 \left(1 + x'_{n+1}(\mathbf{X}'\mathbf{X})^{-1}x_{n+1}\right) = \sigma^2$$

$x'_{n+1}(\mathbf{X}'\mathbf{X})^{-1}x_{n+1}$ quantifies the prediction value for a data point to exert influence on the regression. As the sample increases, $x'_{n+1}(\mathbf{X}'\mathbf{X})^{-1}x_{n+1}$ will approach to zero and $\text{Var}(\hat{e}_{n+1}|\mathbf{X})$ gets closer to $\text{Var}(y_i|x_i)$.

2. Empirical Application- No Risk, No Steak? Interpreting Regressions in the CLRM

```
library(stargazer)
library(ggplot2)
```

2(a)

```
# read data
d.steak <- fivethirtyeight::steak_survey
```

```
str(d.steak)
```

```
## tibble [550 x 15] (S3: tbl_df/tbl/data.frame)
##  $ respondent_id: num [1:550] 3.24e+09 3.23e+09 3.23e+09 3.23e+09 3.23e+09 ...
##  $ lottery_a    : logi [1:550] FALSE TRUE TRUE FALSE FALSE TRUE ...
##  $ smoke        : logi [1:550] NA FALSE FALSE TRUE FALSE FALSE ...
##  $ alcohol      : logi [1:550] NA TRUE TRUE TRUE TRUE FALSE ...
##  $ gamble       : logi [1:550] NA FALSE TRUE TRUE FALSE FALSE ...
##  $ skydiving    : logi [1:550] NA FALSE FALSE FALSE FALSE FALSE ...
##  $ speed        : logi [1:550] NA FALSE TRUE TRUE TRUE TRUE ...
##  $ cheated      : logi [1:550] NA FALSE TRUE TRUE TRUE FALSE ...
##  $ steak        : logi [1:550] NA TRUE TRUE TRUE TRUE TRUE ...
##  $ steak_prep   : Ord.factor w/ 5 levels "Rare"<"Medium rare"<..: NA 2 1 3 3 2 NA 2 3 2 ...
##  $ female       : logi [1:550] NA FALSE FALSE FALSE FALSE FALSE ...
##  $ age          : Ord.factor w/ 4 levels "18-29"<"30-44"<..: NA 4 4 4 4 1 4 1 1 4 ...
##  $ hhold_income : Factor w/ 6 levels "$0 - $24,999",..: NA 3 5 3 3 1 5 2 3 2 ...
##  $ educ         : Ord.factor w/ 5 levels "Less than high school degree"<..: NA 3 5 4 5 3 5 3 4 3 ...
##  $ region       : chr [1:550] NA "East North Central" "South Atlantic" "New England" ...
```

```
# missing value summary
sapply(d.steak, function(x) sum(is.na(x)))
```

```
## respondent_id     lottery_a         smoke       alcohol        gamble
##             0             4            13             9            13
##     skydiving         speed       cheated         steak    steak_prep
##            12            11            11            11           118
##        female           age  hhold_income          educ        region
##            36            36           120            38            38
```

Generate variables:

```
# initialize columns
d.steak$cooking_temp <- NA
d.steak$yrs_ed <- NA
d.steak$rand_age <- NA
```

```
attach(d.steak)

# cooking_temp
d.steak$cooking_temp[steak_prep == "Rare"] <- 120
d.steak$cooking_temp[steak_prep == "Medium rare"] <- 130
d.steak$cooking_temp[steak_prep == "Medium"] <- 135
d.steak$cooking_temp[steak_prep == "Medium Well"] <- 140
d.steak$cooking_temp[steak_prep == "Well"] <- 150

# cheat
d.steak$cheat <- ifelse(cheated, 1, 0)

# riskaverse
d.steak$riskaverse <- ifelse(lottery_a == F, 1, 0)

# yrs_ed
d.steak$yrs_ed[educ == "Less than high school degree"] <- 8
d.steak$yrs_ed[educ == "High school degree"] <- 12
d.steak$yrs_ed[educ == "Some college or Associate degree"] <- 14
d.steak$yrs_ed[educ == "Bachelor degree"] <- 16
d.steak$yrs_ed[educ == "Graduate degree"] <- 18

# rand_age
set.seed(123)

n1 <- length(age[age == "18-29" & is.na(age) == F])
n2 <- length(age[age == "30-44" & is.na(age) == F])
n3 <- length(age[age == "45-60" & is.na(age) == F])
n4 <- length(age[age == "> 60" & is.na(age) == F])

d.steak$rand_age[age == "18-29" & is.na(age) == F] <- sample(18:29, n1, replace = T)
d.steak$rand_age[age == "30-44" & is.na(age) == F] <- sample(30:44, n2, replace = T)
d.steak$rand_age[age == "45-60" & is.na(age) == F] <- sample(45:60, n3, replace = T)
d.steak$rand_age[age == "> 60" & is.na(age) == F] <- sample(61:90, n4, replace = T)

detach(d.steak)
```

2(b)

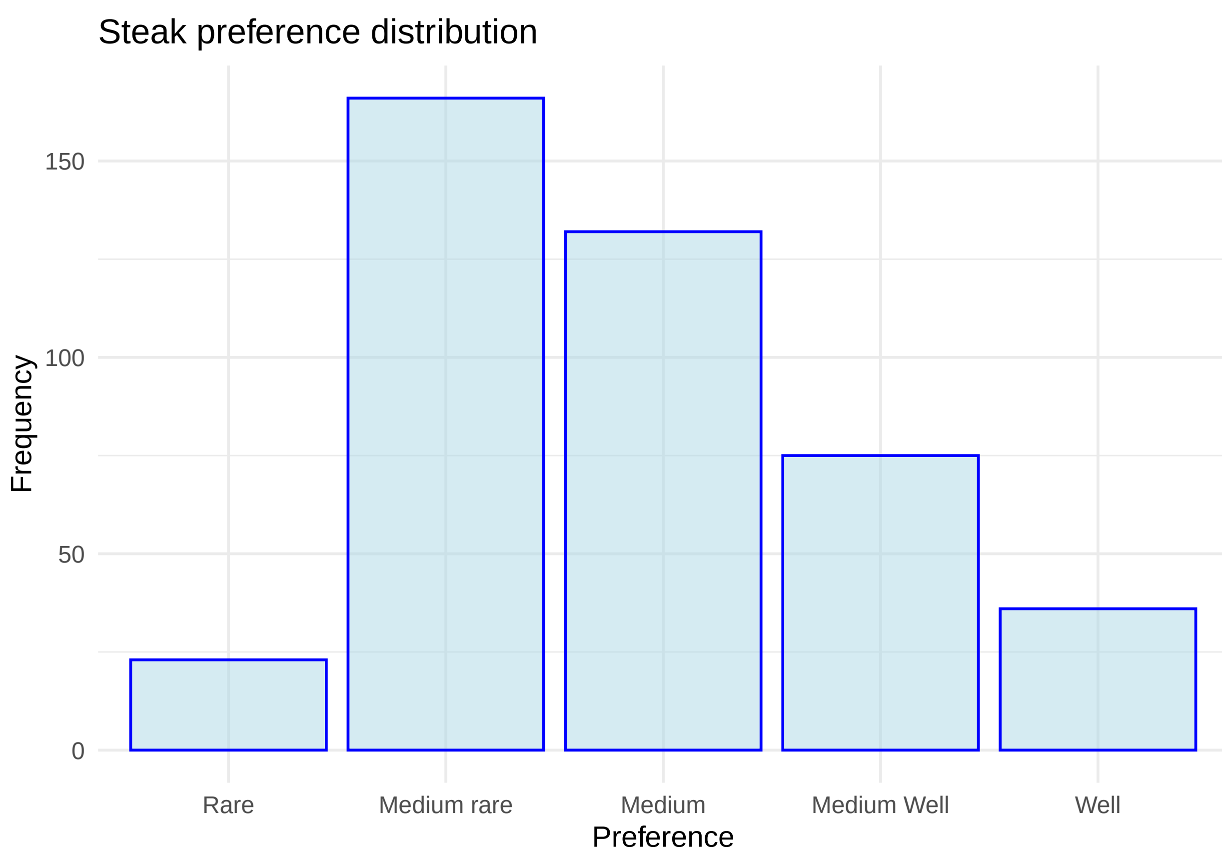
```
subdata <- as.data.frame(
  subset(d.steak,
         select=c("cooking_temp",
                  "cheat",
                  "riskaverse",
                  "yrs_ed",
```

```
                  "rand_age"))
)

stargazer(subdata, header = F, title = "Summary statistics")
```

Table 1: Summary statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
cooking_temp	432	134.398	6.665	120.000	130.000	140.000	150.000
cheat	539	0.171	0.377	0.000	0.000	0.000	1.000
riskaverse	546	0.511	0.500	0.000	0.000	1.000	1.000
yrs_ed	512	15.543	1.894	8.000	14.000	18.000	18.000
rand_age	514	48.360	19.597	18.000	33.000	61.000	90.000

```
ggplot(subset(d.steak, !is.na(steak_prep)), aes(steak_prep)) +
  geom_bar(color="blue", fill="lightblue", alpha=0.5) +
  xlab("Preference") + ylab("Frequency") +
  theme_minimal() + ggtitle("Steak preference distribution")
```



2(c)

We guess that the sign of the coefficient should be positive. Our reasoning is that the more risk-averse an individual is, the less likely he or she is to risk eating rare steak.

```r
# restrict data to steakeaters
steakeaters <- subset(
  d.steak, steak=="TRUE",
  select = c("cooking_temp", "cheat", "steak", "riskaverse", "rand_age", "yrs_ed")
)
```

```r
# compute coefficient "by hand"
y1 <- steakeaters$cooking_temp
x1 <- steakeaters$riskaverse

cov_cr <- cov(y1, x1, use="pairwise")
var_r <- var(x1, na.rm = T)

beta_1 <- cov_cr/var_r

# beta_0 <- mean(y1) - beta_1*mean(x1, na.rm=T)

beta_0 <-
  mean(
    subset(steakeaters, is.na(riskaverse)==F, select = c("cooking_temp"))$cooking_temp
  )- beta_1 * mean(x1, na.rm=T)

cat(sprintf("The intercept is %.3f\nThe coefficient is %.3f", beta_0, beta_1))
```

```
## The intercept is 134.390
## The coefficient is -0.069
```

```r
# compute coefficient by running regression
model1 <- lm(cooking_temp ~ riskaverse, data = steakeaters)

stargazer(model1, header = F, title = "Model (1)",
          keep.stat = c("n", "rsq", "ser"), single.row = T)
```

Table 2: Model (1)

	Dependent variable:
	cooking_temp
riskaverse	−0.069 (0.646)
Constant	134.390*** (0.465)
Observations	426
R^2	0.00003
Residual Std. Error	6.658 (df = 424)
Note:	*p<0.1; **p<0.05; ***p<0.01

If the mean-zero-error assumption holds, the coefficient of risk aversion is interpreted as the marginal effect of risk aversion on cooking temperature. In this setting, if someone prefers lottery B, then his or her preference for cooking temperature would decrease by 0.161 degrees Fahrenheit.

2(d)

- It is important to include a constant in the regression model.

- Regression estimates will be biased if forced to go through the origin.
- No constant implies that preferred cooking temperature is zero when explanatory variables equal to zero, which is unrealistic.

2(e)

```
model2 <- lm(cooking_temp ~ 1 + riskaverse + log(yrs_ed) + rand_age +
               I(rand_age^2) + cheat, data = steakeaters)

stargazer(model1, model2, header = F, single.row = T,
          title = "Comparison between Model (1) and Model (2)",
          keep.stat = c("n", "rsq", "ser"))
```

Table 3: Comparison between Model (1) and Model (2)

	Dependent variable:	
	cooking_temp	
	(1)	(2)
riskaverse	−0.069 (0.646)	−0.124 (0.661)
log(yrs_ed)		−5.366** (2.670)
rand_age		0.022 (0.093)
I(rand_age^2)		−0.0004 (0.001)
cheat		0.380 (0.883)
Constant	134.390*** (0.465)	149.056*** (7.432)
Observations	426	403
R^2	0.00003	0.018
Residual Std. Error	6.658 (df = 424)	6.628 (df = 397)
Note:		*p<0.1; **p<0.05; ***p<0.01

```
beta_3 <- as.numeric(coef(model2)["riskaverse"])
beta_4 <- as.numeric(coef(model2)["rand_age"])
beta_5 <- as.numeric(coef(model2)["I(rand_age^2)"])

# marginal effect of 1 additional year of education on cooking temperature
m_yrs_ed <- as.numeric(mean(steakeaters$yrs_ed, na.rm = T))
me1 <- beta_3 * (log(m_yrs_ed + 1) - log(m_yrs_ed))

# marginal effect of having cheated on a spouse on cooking temperature
me2 <- coef(model2)["cheat"]

# marginal effect of 10 additional years of age on cooking temperature
m_rand_age <- as.numeric(mean(steakeaters$rand_age, na.rm = T))
me3 <- beta_4 * 10 + beta_5 * ((m_rand_age + 10)^2 - m_rand_age^2)

cat(sprintf("Marginal effect of 1 addtional year of education is %.3f
            \nMarginal effect of having cheated on a spouse is %.3f
            \nMarginal effect of 10 additional years of age is %.3f",
            me1, me2, me3))
```

```
## Marginal effect of 1 addtional year of education is -0.008
##
## Marginal effect of having cheated on a spouse is 0.380
##
## Marginal effect of 10 additional years of age is -0.224
```

2(f)

```r
# create a dataframe to store mean value for each variable
mean_data <- data.frame(
  riskaverse = mean(steakeaters$riskaverse, na.rm = T),
  yrs_ed = mean(steakeaters$yrs_ed, na.rm = T),
  rand_age = mean(steakeaters$rand_age, na.rm = T),
  cheat = mean(steakeaters$cheat, na.rm = T)
)

# predict cooking temperature when all explanatory variables are set to their mean
predict(model2, newdata = mean_data)
```

```
##        1
## 134.4202
```

It is not an informative number to look at. For the dummy variables (*riskaverse* and *cheat*), it does not make sense to set them at their mean value.

2(g)

It is not a good idea to include both the estimated age and the categorical age variable.

- Imperfect Multicollinearity.
- It does not make sense to square a categorical variable.

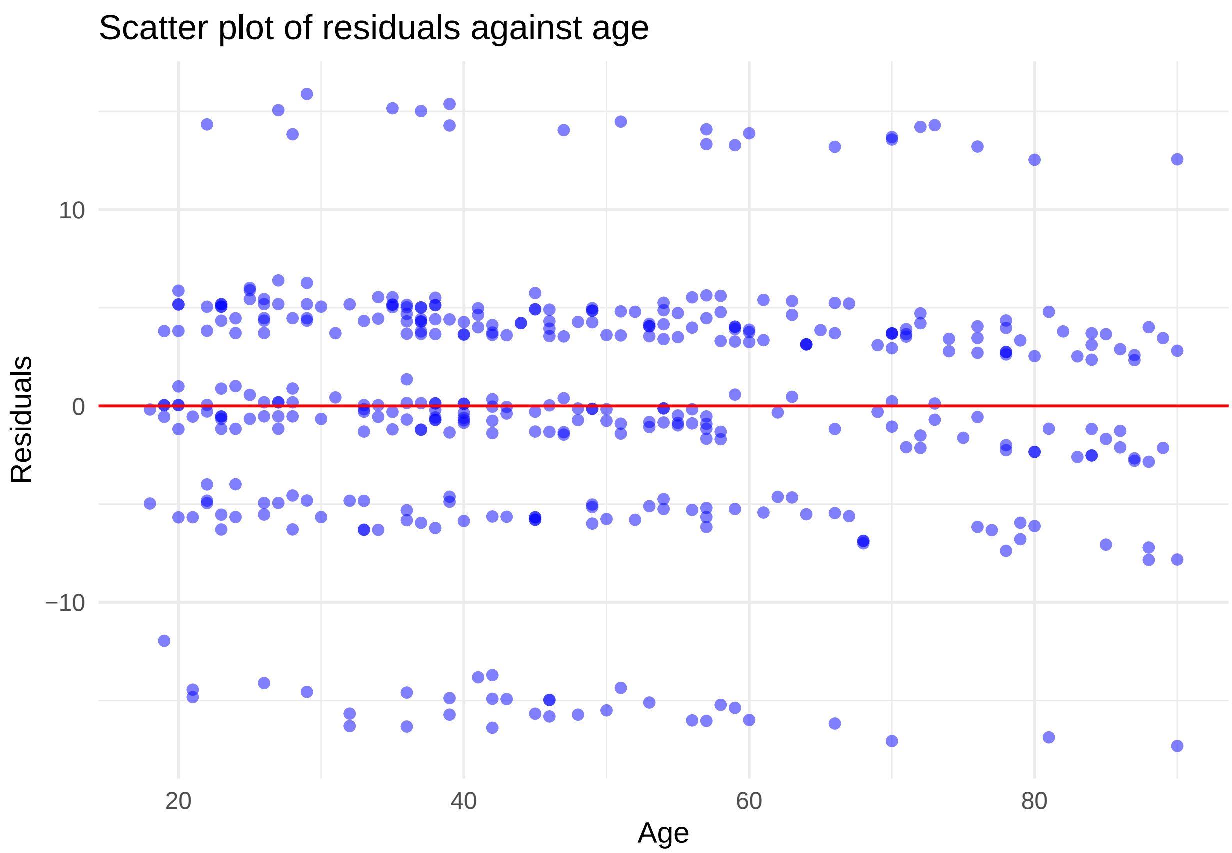
2(h)

- Data set is not large enough to capture the true effects.
- We might fit the data with a wrong model.

2(i)

```r
# drop missing values
steakeaters <- steakeaters[complete.cases(steakeaters), ]

# generate predicted residuals variable
steakeaters$pred_residuals <- NA # initialize a column
steakeaters$pred_residuals <- predict(model2, steakeaters) - steakeaters$cooking_temp
```

```r
ggplot(
  steakeaters,
  aes(x=rand_age, y=pred_residuals)
) + geom_point(color="blue", alpha=0.5) +
  geom_hline(yintercept=0, color="red") +
  xlab("Age") + ylab("Residuals") +
  theme_minimal() + ggtitle("Scatter plot of residuals against age")
```

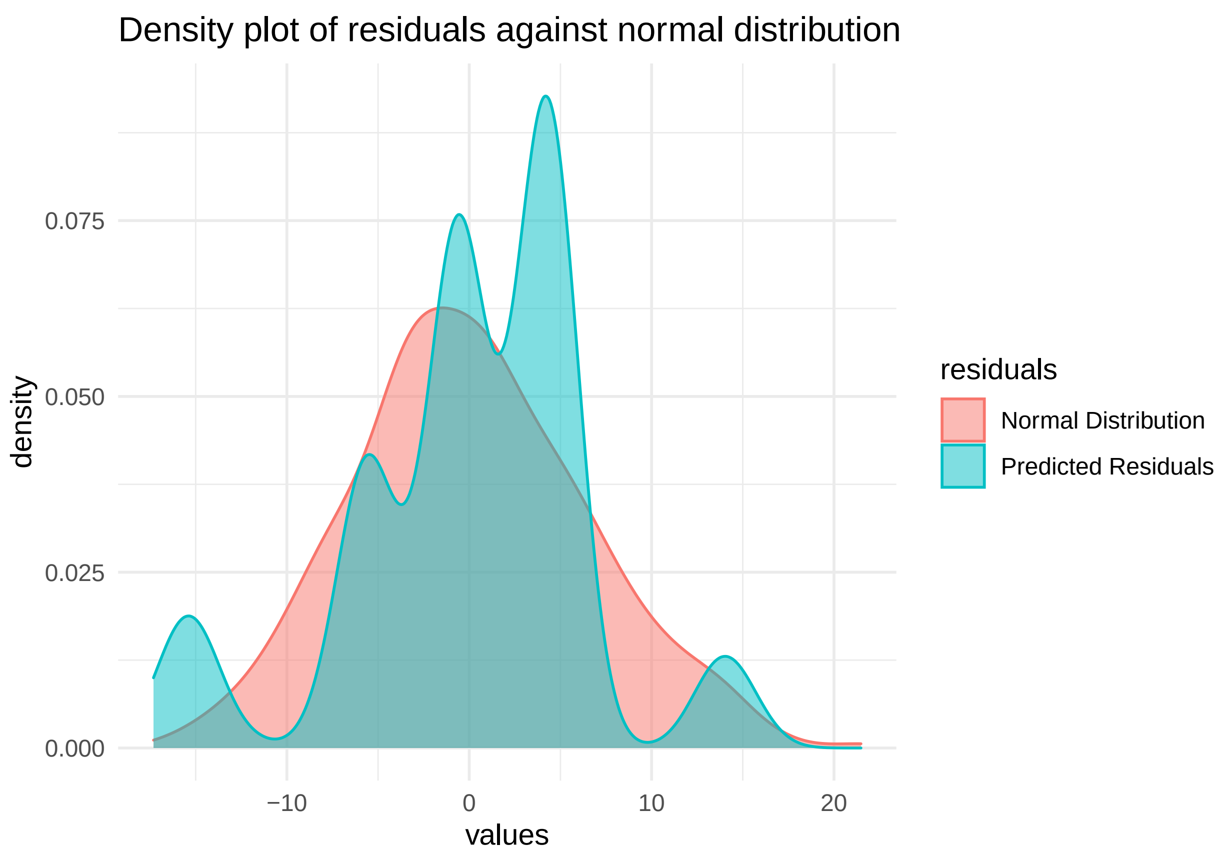


Scatter plot of residuals against age shows that residuals do not vary across different ages. Homoskedasticity assumption holds.

```r
set.seed(123)

df <- data.frame(
  steakeaters$pred_residuals,
  rnorm(nobs(model2), 0, sigma(model2))
)

colnames(df) <- c("Predicted Residuals", "Normal Distribution")

df <- reshape(
  df,
  direction = "long",
  varying = list(names(df)),
  v.names = "values",
  timevar = "residuals",
```

```
  times = c("Predicted Residuals", "Normal Distribution")
)
```

```
ggplot(df, aes(x = values)) +
  geom_density(aes(group = residuals, color = residuals, fill = residuals), alpha = 0.5) +
  theme_minimal() + ggtitle("Density plot of residuals against normal distribution")
```



Density plot of residuals against normal distribution shows that the predicted residuals are quasi-normally distributed.