tutorial_1_submission

luiyusen97

1/20/2021

Initial reading of raw data. Packages used are tidyverse, foreign, ggplot2, sandwich and lmtest.

```
dat <- read.dta(file = fil)
dat <- mutate(dat, npvis_squared = npvis**2)

model <- lm(bwght ~ npvis + npvis_squared + cigs + male, dat)
```

(a)

```
mean_med <- summary(dat[, "bwght"])
mean_bwght <- mean_med[4]
print(mean_bwght)
```

```
##     Mean
## 3401.122
```

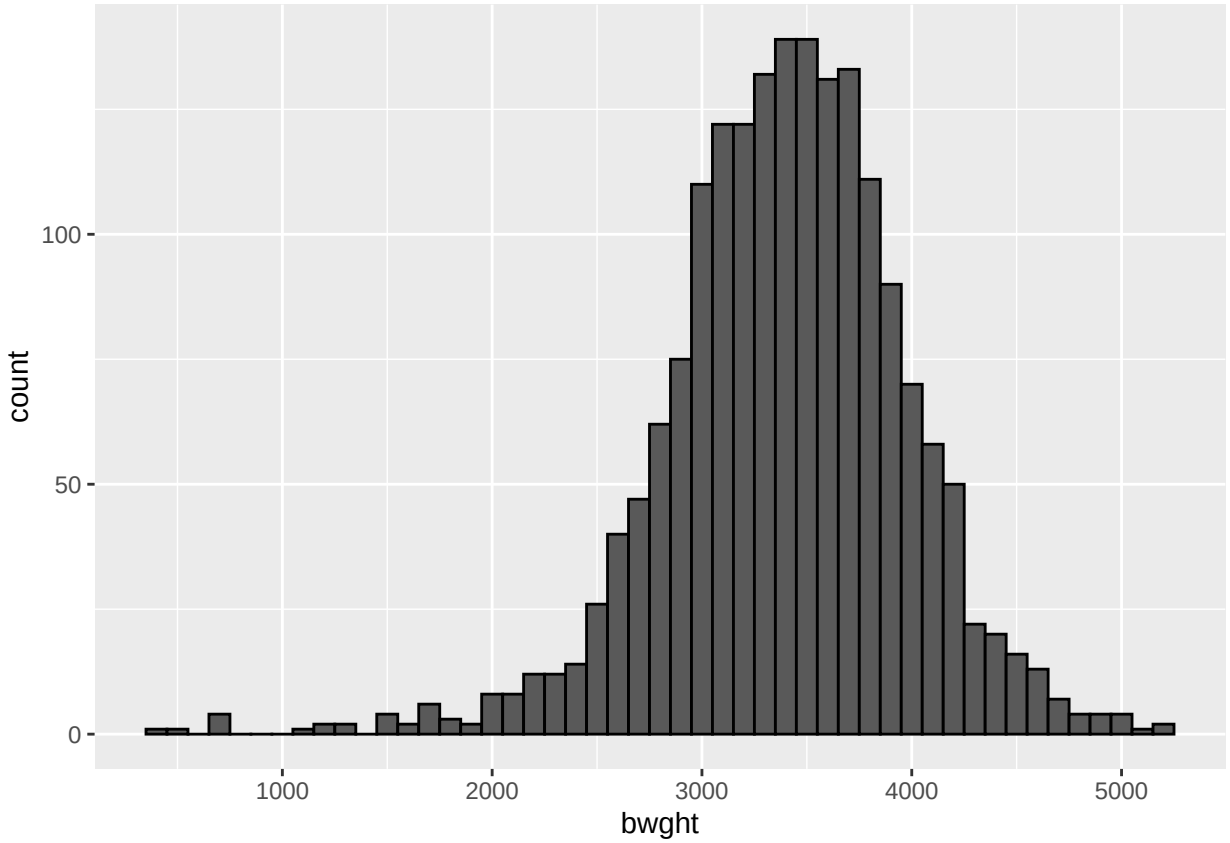
(b)

```
med_bwght <- mean_med[3]
print(med_bwght)
```

```
## Median
##   3425
```

(c)

```
plot_bwght <- ggplot(dat, aes(x = bwght)) +
    geom_histogram(binwidth = 100, color = "black")
print(plot_bwght)
```



(d)

```
beta_3 <- model$coefficients["cigs"]
print(beta_3)
```

```
##      cigs
## -9.901333
```

(e)

$H_0 : \beta_3 = -10$, $H_a : \beta_3 \neq -10$, $\alpha = 0.05$, 2-tailed 1-sample t-test.

```
t_test_2tail_variable_hypo <- function(hypothesis, significance_lvl, df, sample_value, standard_error){
    confidence_interval <- hypothesis +c(-1, 1)*qt(p = 1-significance_lvl/2,
                                          df = df)*standard_error
    if (between(sample_value, confidence_interval[1], confidence_interval[2])){
        return(FALSE)
    } else {return(TRUE)}
}
test_beta_3 <- t_test_2tail_variable_hypo(-10, 0.05, 1651, beta_3, 3.3330) # TRUE means insignificantly
confidence_interval <- -10 + c(-1, 1)*qt(p = 0.975, df = 1651)*3.3330
print(confidence_interval)
```

```
## [1] -16.537352  -3.462648
```

Since $\beta_3 \in [-16.54, -3.46]$, then we have insufficient evidence to reject $H_0 = -10$.

(f)

$H_0 : \beta_4 = 100$, $H_a : \beta_4 \neq 100$, $\alpha = 0.05$, 2-tailed 2-sample Welsh t-test.

```
test_beta_4 <- t_test_2tail_variable_hypo(100, 0.05, 1651, beta_3, 27.9757) # TRUE means insignificantl
confidence_interval <- 100 + c(-1, 1)*qt(p = 0.975, df = 1651)*27.9757
print(confidence_interval)
```

```
## [1]  45.12841 154.87159
```

Since $+100 \in [45.1, 154.9]$, then we have insufficient evidence to reject $H_0 = +100$.

(g)

```
beta_1 <- model$coefficients["npvis"]
beta_2 <- model$coefficients["npvis_squared"]
```

The partial effect is $32.8 npvis + (-0.669) npvis^2$.

(h)

$0 = \frac{\partial bwght}{\partial npvis} = 32.8 + (-0.669) npvis * 2 \; npvis = \frac{32.8}{2*(-0.669)}$

```
turning_pt <- beta_1/(2*beta_2)
print(turning_pt)
```

```
##     npvis
## -24.50519
```

$npvis = -24.50519$ Diminishing marginal returns of additional pre-natal visits since there is not much more the doctor can check.

(i)

$H_0 : 0 = \beta_1 = \beta_2 = \beta_3 = \beta_4$, $H_a : otherwise$, F-test of squared residuals against independent variables.

```
hetero_test <- model$model[, 2:5]
hetero_test <- mutate(hetero_test, residuals = model$residuals)
hetero_test <- mutate(hetero_test, residuals = residuals**2)
hetero_test_model <- lm(residuals ~ npvis + npvis_squared + cigs + male, hetero_test)
hetero_test_p_value <- 4.476*(10**(-5)) # heteroscedastic
```

Since $4.476 * (10^{-5}) < 0.05$, then we have sufficient evidence to reject H_0, and we cannot have the homoscedascity assumption. To solve it, we can use robust standard errors as sample size $n = 1656 > 120$. The hypothesis tests for β_3 and β_4 would then be like so:

```
robust_t <- coeftest(model, vcov = vcovHC(model, type = "HC0"))
print(robust_t)
```

```
##
## t test of coefficients:
##
##                 Estimate Std. Error t value  Pr(>|t|)
## (Intercept)   3097.59754   99.79408 31.0399 < 2.2e-16 ***
## npvis           32.80329   12.76511  2.5698  0.010264 *
## npvis_squared   -0.66931    0.39520 -1.6936  0.090525 .
## cigs            -9.90133    3.28212 -3.0167  0.002594 **
## male            81.31364   27.82588  2.9222  0.003523 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
test_beta_3_robust <- t_test_2tail_variable_hypo(
    -10, 0.05, 1651, beta_3, 3.28212) # TRUE means insignificantly different from -10
confidence_interval <- -10 + c(-1, 1)*qt(p = 0.975, df = 1651)*3.28212
print(confidence_interval)
```

```
## [1] -16.437556  -3.562444
```

```
test_beta_4 <- t_test_2tail_variable_hypo(
    100, 0.05, 1651, beta_3, 27.82588) # TRUE means insignificantly different from +100
confidence_interval <- 100 + c(-1, 1)*qt(p = 0.975, df = 1651)*27.82588
print(confidence_interval)
```

```
## [1]  45.42227 154.57773
```

Since $-10 \in [-16.437556, -3.562444]$, then we cannot reject $H_0 : \beta_3 = -10$. Since $+100 \in [45.42227, 154.57773]$, then we cannot reject $H_0 : \beta_4 = 100$.

(j)

No, I do not think we can have the zero conditional mean, because income can be correlated with pre-natal visits since poorer people can come from locations where clinics are inaccessible.