HW11 Group 1, Austin Halvorsen

Pedram Jahangiry

Dec 1 2020

# Problems

## Question 1

### (i)

1. All of the above are consequences

## Question 2

### (i)

The transformed equation would be found by dividing the original, , by , which would give us:

## Question 3

### (i)

It would be **False** because WLS estimates could have more or less bias based on the degree of correlation between the error term in explanatory variables when an important

## Question 4

### (i)

Since our *df* would be the number that values can vary, we now that in our numerator, the degrees of freedom is K+1. For our denominator, the degrees of freedom is n-k-2.

### (ii)

The will always be at least as large from the BP regression and White test because in the BP test, the added variable is , which will always be no less than the original. For the White test, the fitted values are linear functions of the regressors. So we can conclude that the original value is greater or equal to the White test .

### (iii)

If we look at our F-test formula, we can see that as increases, so does our test value. So we can conclude that our test doesn’t depend on , rather, degrees of freedom is more important. Since all three tests have different *df*, it is hard to say which test gives a smaller p-value.

### (iv)

I think that it would be a bad idea. Since our equation already includes , if we add , our model would suffer from perfect collinearity.

# Computer Exercises

## Question 5

### (i)

Given that our variance should only rely on gender, our model would then look like: where women are represented by and men by

### (ii)

df4 <- sleep75  
wls4 <- lm(sleep~totwrk+educ+age+I(age^2)+yngkid+male, df4)  
pander(summary(wls4))

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| **(Intercept)** | 3841 | 239.4 | 16.04 | 1.51e-49 |
| **totwrk** | -0.1634 | 0.01816 | -8.997 | 2.153e-18 |
| **educ** | -11.71 | 5.872 | -1.995 | 0.04645 |
| **age** | -8.697 | 11.33 | -0.7677 | 0.4429 |
| **I(age^2)** | 0.1284 | 0.1347 | 0.9538 | 0.3405 |
| **yngkid** | -0.0228 | 50.28 | -0.0004535 | 0.9996 |
| **male** | 87.75 | 34.67 | 2.531 | 0.01158 |

Fitting linear model: sleep ~ totwrk + educ + age + I(age^2) + yngkid + male

|  |  |  |  |
| --- | --- | --- | --- |
| Observations | Residual Std. Error |  | Adjusted |
| 706 | 418 | 0.1228 | 0.1152 |

wls4\_resid <- resid(wls4)  
summary(lm(wls4\_resid^2~male, df4))

Call:  
lm(formula = wls4\_resid^2 ~ male, data = df4)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-189359 -155559 -111662 26265 5465531   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 189359 20546 9.216 <2e-16 \*\*\*  
male -28850 27296 -1.057 0.291   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 359400 on 704 degrees of freedom  
Multiple R-squared: 0.001584, Adjusted R-squared: 0.000166   
F-statistic: 1.117 on 1 and 704 DF, p-value: 0.2909

Since the coefficient of male is negative, our variance error term u is higher for females than males.

### (iii)

With a p-value of 0.291, which is larger than our alpha 0.05, it is not statistically significant between male and female.

## Question 6

### (i)

df6 <- hprice1   
wls6a <- lm(price~lotsize+sqrft+bdrms, df6)  
wls6b <- lm(price~lotsize+sqrft+bdrms, df6)  
pander(summary(wls6a))

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| **(Intercept)** | -21.77 | 29.48 | -0.7386 | 0.4622 |
| **lotsize** | 0.002068 | 0.0006421 | 3.22 | 0.001823 |
| **sqrft** | 0.1228 | 0.01324 | 9.275 | 1.658e-14 |
| **bdrms** | 13.85 | 9.01 | 1.537 | 0.1279 |

Fitting linear model: price ~ lotsize + sqrft + bdrms

|  |  |  |  |
| --- | --- | --- | --- |
| Observations | Residual Std. Error |  | Adjusted |
| 88 | 59.83 | 0.6724 | 0.6607 |

coeftest(wls6a, vcov=hccm(wls6a,type="hc0"))

t test of coefficients:  
  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) -21.7703081 36.2843444 -0.6000 0.55013   
lotsize 0.0020677 0.0012227 1.6912 0.09451 .   
sqrft 0.1227782 0.0173178 7.0897 3.883e-10 \*\*\*  
bdrms 13.8525217 8.2836880 1.6723 0.09819 .   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Comparing our two tests, our errors only differ slightly with the most significant different between LOTSIZE

|  |  |  |
| --- | --- | --- |
| Variable | OLS Std | Het-robust std |
| Constant | 29.48 | 36.2843 |
| LOTSIZE | 0.0006421 | 0.0012227 |
| SQRFT | 0.01324 | 0.0173178 |
| BDRMS | 9.01 | 8.2836880 |

### (ii)

wls6c <- lm(I(log(price))~I(log(lotsize))+I(log(sqrft))+bdrms, df6)  
pander(summary(wls6c))

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| **(Intercept)** | -1.297 | 0.6513 | -1.992 | 0.04967 |
| **I(log(lotsize))** | 0.168 | 0.03828 | 4.388 | 3.307e-05 |
| **I(log(sqrft))** | 0.7002 | 0.09287 | 7.54 | 5.006e-11 |
| **bdrms** | 0.03696 | 0.02753 | 1.342 | 0.1831 |

Fitting linear model: I(log(price)) ~ I(log(lotsize)) + I(log(sqrft)) + bdrms

|  |  |  |  |
| --- | --- | --- | --- |
| Observations | Residual Std. Error |  | Adjusted |
| 88 | 0.1846 | 0.643 | 0.6302 |

coeftest(wls6c, vcov=hccm(wls6c,type="hc0"))

t test of coefficients:  
  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) -1.297042 0.763351 -1.6991 0.09299 .   
I(log(lotsize)) 0.167967 0.040520 4.1453 8.067e-05 \*\*\*  
I(log(sqrft)) 0.700232 0.101442 6.9028 9.014e-10 \*\*\*  
bdrms 0.036958 0.029898 1.2362 0.21984   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Again, we see that many of our standard errors are close to each other using both the OLS and the heteroskedasticity-robust standard error

|  |  |  |
| --- | --- | --- |
| Variable | OLS Std | Het-robust std |
| Constant | 0.6513 | 0.763351 |
| LLOTSIZE | 0.03828 | 0.040520 |
| LSQRFT | 0.09287 | 0.0173178 |
| BDRMS | 0.02753 | 0.101442 |

### (iii)

After comparing our standard errors, we can see that using log transformations has helped us drastically reduce heteroskedasticity within our model. However, using the OLS standard error versus the Heteroskedasticity-robust standard error over different coefficients only yields a marginal difference in in value.

## Question 7

### (i)

df7 <- vote1  
ols7 <- lm(voteA~prtystrA+democA+lexpendA+lexpendB, df7)   
pander(summary(ols7))

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| **(Intercept)** | 37.66 | 4.736 | 7.952 | 2.564e-13 |
| **prtystrA** | 0.2519 | 0.07129 | 3.534 | 0.0005296 |
| **democA** | 3.793 | 1.407 | 2.697 | 0.007717 |
| **lexpendA** | 5.779 | 0.3918 | 14.75 | 4.03e-32 |
| **lexpendB** | -6.238 | 0.3975 | -15.69 | 9.343e-35 |

Fitting linear model: voteA ~ prtystrA + democA + lexpendA + lexpendB

|  |  |  |  |
| --- | --- | --- | --- |
| Observations | Residual Std. Error |  | Adjusted |
| 173 | 7.573 | 0.8012 | 0.7964 |

ols7\_resid <- resid(ols7)  
ols7a <- lm(ols7\_resid~prtystrA+democA+lexpendA+lexpendB, df7)   
pander(summary(ols7a))

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| **(Intercept)** | 3.752e-15 | 4.736 | 7.923e-16 | 1 |
| **prtystrA** | -2.306e-17 | 0.07129 | -3.234e-16 | 1 |
| **democA** | -8.381e-16 | 1.407 | -5.959e-16 | 1 |
| **lexpendA** | 3.309e-17 | 0.3918 | 8.446e-17 | 1 |
| **lexpendB** | -1.865e-16 | 0.3975 | -4.691e-16 | 1 |

Fitting linear model: ols7\_resid ~ prtystrA + democA + lexpendA + lexpendB

|  |  |  |  |
| --- | --- | --- | --- |
| Observations | Residual Std. Error |  | Adjusted |
| 173 | 7.573 | 6.145e-32 | -0.02381 |

When we regress on the error term, we get an . This is because the dependent variable, *voteA*, gave us the estimated coefficients in a way that the error term was uncorrelated with the independent variables. So we are seeing the errors terms that are uncorrelated with our independent variables.

### (ii)

bptest(ols7)

studentized Breusch-Pagan test  
  
data: ols7  
BP = 9.0934, df = 4, p-value = 0.05881

summary(lm(ols7\_resid^2~prtystrA+democA+lexpendA+lexpendB, df7))

Call:  
lm(formula = ols7\_resid^2 ~ prtystrA + democA + lexpendA + lexpendB,   
 data = df7)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-94.88 -46.16 -23.51 15.84 508.32   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 113.96347 50.81503 2.243 0.0262 \*  
prtystrA -0.29926 0.76493 -0.391 0.6961   
democA 15.61921 15.09117 1.035 0.3022   
lexpendA -10.30573 4.20401 -2.451 0.0153 \*  
lexpendB -0.05141 4.26452 -0.012 0.9904   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 81.25 on 168 degrees of freedom  
Multiple R-squared: 0.05256, Adjusted R-squared: 0.03   
F-statistic: 2.33 on 4 and 168 DF, p-value: 0.05806

Using the F statistic method, we get a p-value of 0.05806

### (iii)

ols\_y <- predict(ols7)  
summary(lm(ols7\_resid^2~ols\_y+I(ols\_y^2)))

Call:  
lm(formula = ols7\_resid^2 ~ ols\_y + I(ols\_y^2))  
  
Residuals:  
 Min 1Q Median 3Q Max   
-107.65 -44.80 -29.88 23.81 539.52   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 171.85840 53.14213 3.234 0.00147 \*\*  
ols\_y -4.26368 2.16653 -1.968 0.05070 .   
I(ols\_y^2) 0.03574 0.02124 1.682 0.09434 .   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 81.66 on 170 degrees of freedom  
Multiple R-squared: 0.03173, Adjusted R-squared: 0.02034   
F-statistic: 2.786 on 2 and 170 DF, p-value: 0.0645

Here, we get an F-statistic of of 0.0645, which is greater than our alpha at 0.05, which is slightly higher than our BP test p-value. This would mean that there is slightly less evidence of heteroskedasticity than our BP test showed.

## Question 8

### (i)

df8 <- k401ksubs  
ols8 <- lm(e401k~inc+incsq+age+agesq+male, df8)  
pander(summary(ols8))

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| **(Intercept)** | -0.5063 | 0.0811 | -6.243 | 4.479e-10 |
| **inc** | 0.01245 | 0.0005929 | 20.99 | 1.238e-95 |
| **incsq** | -6.165e-05 | 4.732e-06 | -13.03 | 1.837e-38 |
| **age** | 0.02651 | 0.003922 | 6.758 | 1.488e-11 |
| **agesq** | -0.0003053 | 4.501e-05 | -6.782 | 1.257e-11 |
| **male** | -0.003533 | 0.01208 | -0.2924 | 0.77 |

Fitting linear model: e401k ~ inc + incsq + age + agesq + male

|  |  |  |  |
| --- | --- | --- | --- |
| Observations | Residual Std. Error |  | Adjusted |
| 9275 | 0.4648 | 0.09428 | 0.09379 |

coeftest(ols8, vcov=hccm(ols8,type="hc0"))

t test of coefficients:  
  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) -5.0629e-01 7.8529e-02 -6.4472 1.196e-10 \*\*\*  
inc 1.2446e-02 6.0011e-04 20.7404 < 2.2e-16 \*\*\*  
incsq -6.1649e-05 5.0025e-06 -12.3235 < 2.2e-16 \*\*\*  
age 2.6506e-02 3.8222e-03 6.9347 4.342e-12 \*\*\*  
agesq -3.0527e-04 4.3739e-05 -6.9794 3.169e-12 \*\*\*  
male -3.5328e-03 1.2049e-02 -0.2932 0.7694   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Looking at our standard errors from both, the difference is small and such, the difference is not statistically important.

### (ii)

Our fitted values are all between 0 and 1.