Problem Set 2: Heteroskedasticity

EC 421: Introduction to Econometrics

Solutions

DUE Your solutions to this problem set are due *before* midnight on Saturday, 08 February 2020. Your files must be uploaded to Canvas.

IMPORTANT You must submit **two files**:

1. your typed responses/answers to the question (in a Word file or something similar)
2. the R script you used to generate your answers. Each student must turn in her/his own answers.

OBJECTIVE This problem set has three purposes: (1) reinforce the econometrics topics we reviewed in class; (2) build your R toolset; (3) start building your intuition about causality and time series within econometrics.

1. Suppose we are interested in estimating a model on house prices

$$\text{Price}_i = \beta_0 + \beta_1 \text{Area}_i + u_i \tag{1}$$

where Area_i is a continuous variable that takes positive values.

Suppose that we know $\text{Var}(u_i|\text{Area}_i) = \sigma^2 + \delta\text{Area}$.

1a. If $\delta = 0$, do we have heteroskedasticity? Is OLS unbiased? Explain your answer.

Answer

No, we do not have heteroskedasticity. If $\delta = 0$, then $\text{Var}(u_i) = \sigma^2$, which means we have homoskedasticity. Either way, OLS's unbiasedness is unaffected by heteroskedasticity.

1b. If $\delta = 1$, do we have heteroskedasticity? Is OLS still unbiased? Explain your answer.

Answer

Yes, we do have heteroskedasticity. If $\delta = 1$, then $\text{Var}(u_i) = \sigma^2 + \text{Area}_i$, which means the variance of u_i changes with Area_i. Thus, we have heteroskedasticity. However, OLS's unbiasedness is unaffected by heteroskedasticity.

1c. *Goldfeld-Quandt* Suppose we decide to run a Goldfeld-Quandt test to detect heteroskedasticity.

If the ratio of the two SSE's (*i.e.*, $\text{SSE}_1/\text{SSE}_2$) is exactly 1, do we have evidence of heteroskedasticity? Explain.

Answer

No. If the ratio of SSE's is exactly 1, then we will fail to reject the null hypothesis of homoskedasticity.

1d. Define the term "standard error" and explain why it is important for econometrics.

Answer

The standard error gives the standard deviation of an estimator's distribution—it describes how much the estimator varies around its mean. Standard errors are important because they quantify the uncertainty underlying our estimates—they tell us how *confident* we should be in the unknown parameter being close to the point estimate.

2. Now for some real data. The data in the this problem come from housing sales in Ames, Iowa. We are going to think about modeling the house price at the time of a sale.

2a. Setup: Open up Rstudio and a fresh R script. Load whichever packages you want.

Now load the data in `002-data.csv`.

Answer

```
# Load 'pacman'
library(pacman)
# Load additional packages
p_load(tidyverse, broom, magrittr, ggplot2)
# Load the dataset
housing_df = here("002-data.csv") %>% read_csv()
# Check the dataset
head(housing_df)
```

```
## # A tibble: 6 x 8
##    price year_sold   age  area n_rooms n_bedrooms n_bathrooms quality
##    <dbl>     <dbl> <dbl> <dbl>   <dbl>      <dbl>       <dbl>   <dbl>
## 1 215000      2010    50  1656       7          3         1         6
## 2 105000      2010    49   896       5          2         1         5
## 3 172000      2010    52  1329       6          3         1.5       6
## 4 244000      2010    42  2110       8          3         2.5       7
## 5 189900      2010    13  1629       6          3         2.5       5
## 6 195500      2010    12  1604       7          3         2.5       6
```

2b Describe the distribution of our main variable of interest (`price` in US dollars). You can provide statistical or graphical descriptions of this variable—try `summary(dataset$variable)` and `hist(dataset$variable)`, among others.

Answer

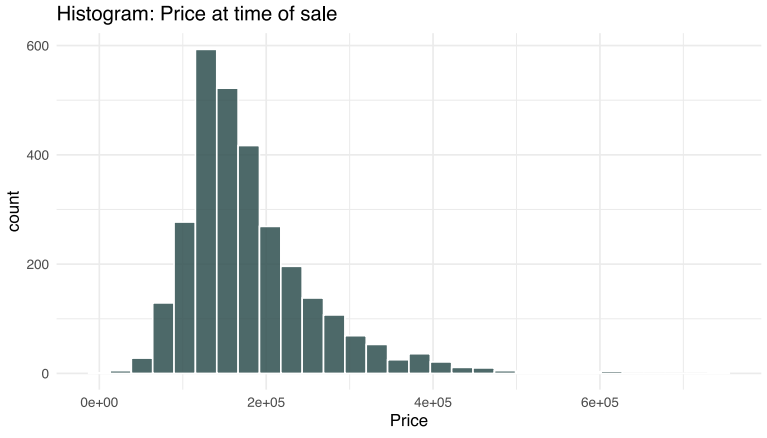
There is substantial variation in the price of a house at the time of the sale. The median is approximately $160,000, and the mean is approximately $180,796.

```
# Summarize variable
summary(housing_df$price)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12789  129500  160000  180796  213500  755000
```

```
# A histogram using 'hist'
hist(
  housing_df$price,
  breaks = 25,
  col = "grey85",
  xlab = "Price",
  main = "Histogram: Price at time of sale"
)
```

```
# A histogram using 'ggplot'
ggplot(data = housing_df, aes(x = price)) +
  geom_histogram(fill = "darkslategrey", color = "white", alpha = 0.85) +
  xlab("Price") +
  ggtitle("Histogram: Price at time of sale") +
  theme_minimal()
```



2c. Regress the sales price (`price`) on an intercept, the area (square feet) of the house (`area`), and the quality of the house (`quality` on a 1–10 scale). Report your findings—the coefficients, brief interpretations of the coefficients, and whether the coefficients are statistically significant.

Answer

```
# Estimate the model
reg_2c = lm(price ~ area + quality, data = housing_df)
# Report the results
reg_2c %>% tidy()
```

```
## # A tibble: 3 x 5
##   term          estimate std.error statistic   p.value
##   <chr>            <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept) -109919.    3421.       -32.1 2.74e-194
## 2 area             58.8      1.84      31.9 4.97e-192
## 3 quality       33241.     660.        50.4 0.
```

We estimate that the coefficient on the area of the home is approximately 58.754. This coefficient says that each additional square foot is associated with an increase in sales price of approximately $58. This coefficient is statistically significant (different from zero) at the 5% level.

For the coefficient on `quality`, we estimate that increasing one "level" of quality corresponds to in an increase in price by approximately $33,241.40. This coefficient is also statistically significant (different from zero) at the 5% level.

2d. Does it make sense to interpret the intercept in this case? Explain.

Answer

It probably does not make sense to interpret the intercept here. The intercept is our estimated price for a house with zero square feet (zero area) and zero quality. The smallest house in our dataset has 334 square feet, and quality ranges from 1 to 10, so it really doesn't make sense to make an estimate where both variables are equal to zero. *(Note: It may also not make sense to treat quality like a continuous variable, but we will ignore that for now.)*

2e. Plot the residuals from your regression in (2c) on the y axis and `quality` on the x axis. Do you see evidence of heteroskedasticity? Explain.

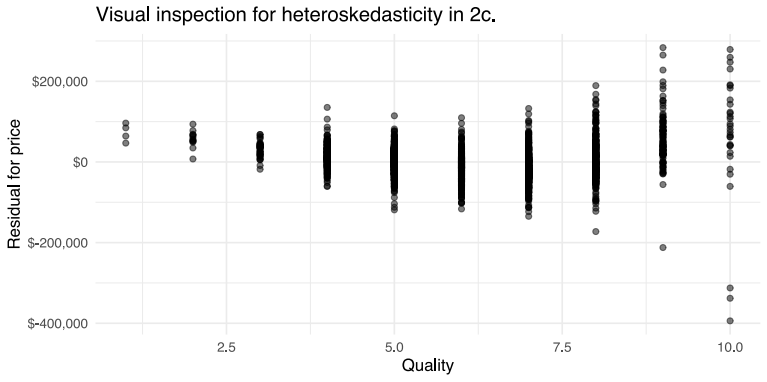
Hint$_1$: You can grab the residuals from a saved `lm` object by (1) using the `residuals()` function or (2) adding the suffix `$residuals` to the end of the `lm` object, *e.g.*, `my_reg$residuals` grabs the residuals from the `lm` object `my_reg`.

Hint$_2$: `plot(x = dataset$variable1, y = dataset$variable2)` makes quick and simple plots. You can also try `qplot()` from the package `ggplot2`, *i.e.*, `qplot(x = variable1, y = variable2, data = dataset)`.

Answer

Based upon the funnel-like figure below, heteroskedasticity seems likely.

```
# Add residuals to the dataset
housing_df %<>% mutate(e_2c = residuals(reg_2c))
# Plot with ggplot
ggplot(data = housing_df, aes(x = quality, y = e_2c)) +
  geom_point(alpha = 0.5) +
  ggtitle("Visual inspection for heteroskedasticity in 2c.") +
  scale_x_continuous("Quality") +
  scale_y_continuous("Residual for price", labels = scales::dollar) +
  theme_minimal()
```



2f. Conduct a Breusch-Pagan test for heteroskedasticity in the regression model in (2c). Describe your hypotheses, the test statistic, the *p*-value, and your conclusion.

Answer

```
# B-P regression
reg_2f = lm(e_2c^2 ~ area + quality, data = housing_df)
# B-P test statistic
lm_2f = summary(reg_2f)$r.squared * nrow(housing_df)
# B-P p-value
pchisq(q = lm_2f, df = 2, lower.tail = F)
```

```
## [1] 1.151965e-99
```

Hypotheses Our Breusch-Pagan test here tests the hypotheses H_0 $\alpha_1 = \alpha_2 0$ vs. H_a $\alpha_1 \neq 0$ and $\alpha_2 \neq 0$ for $e_i^2 = \alpha_0 + \alpha_1 \text{Area}_i + \alpha_2 \text{Quality} + v_i$ (where we are using e_i^2 to estimate u_i^2, which gives us an estimate for σ_i^2.) If we reject H_0, then we have evidence of heteroskedasticity.

Test statistic We calculate a B-P test statistic of approximately 455.63.

***p*-value** Under the distribution of a χ_1^2, the implied *p*-value for our LM statistic (the probability of seeing this test statistic or greater) is approximately 0.

Conclusions Because our *p*-value is less than our standard significance of 0.05, we reject the null hypothesis that $\alpha_1 = 0$ and/or $\alpha_2 = 0$. Thus, there is statistically significant evidence at the 5% level that $\alpha_1 \neq 0$ and/or $\alpha_2 \neq 0$, meaning there is statistically significant evidence of a relationship between e_i^2 and our explanatory variables (area and quality). Therefore, we have statistically significant evidence of heteroskedasticity.

2g. Conduct a White test for heteroskedasticity in the regression model in (2c). Describe your hypotheses, the test statistic, the *p*-value, and your conclusion.

Hint: To square the variable `x` in `lm()`, we write `lm(y ~ x + I(x^2), data = dataset)`. To take the interaction between two variables `x1` and `x2` in `lm()`, we use the colon (`:`), *i.e.*, write `lm(y ~ x1 + x2 + x1:x2, data = dataset)`.

Answer

```
# White regression
reg_2g = lm(e_2c^2 ~ area + I(area^2) + quality + I(quality) + area:quality, data = housing_df)
# White test statistic
lm_2g = summary(reg_2g)$r.squared * nrow(housing_df)
# White p-value
pchisq(q = lm_2g, df = 2, lower.tail = F)
```

```
## [1] 5.132206e-319
```

Hypotheses Our White test in this question tests the hypotheses H_0 $\alpha_1 = \cdots = \alpha_5 = 0$ *vs.* H_a $\alpha_i \neq 0$ for some i, where $e_i^2 = \alpha_0 + \alpha_1 \text{Area}_i + \alpha_2 \text{Area}_i^2 + \alpha_3 \text{Quality}_i + \alpha_4 \text{Quality}_i^2 + \alpha_5 \text{Area}_i \times \text{Quality}_i + v_i$ (where, again, we are using e_i^2 to estimate u_i^2, which gives us an estimate for σ_i^2.) If we reject H_0, then we have evidence of heteroskedasticity.

Test statistic We calculate a White test statistic of approximately 1,466.

***p*-value** Under the distribution of a χ_2^2, the implied *p*-value for our LM statistic (the probability of seeing this test statistic or greater) is approximately 0.

Conclusions Because our *p*-value is less than our standard significance of 0.05, we reject the null hypothesis $\left(\alpha_j = 0\right)$—there is statistically significant evidence at the 5% level that at least one of the $\alpha_j \neq 0$. Therefore we find statistically significant evidence of a relationship between e_i^2 and our explanatory variables (area and quality). We have statistically significant evidence of heteroskedasticity.

2h. Let's imagine that we think heteroskedasticity is present. Estimate heteroskedasticity-robust standard errors. Do your standard errors change? What about the coefficients? Why is this the case?

Hint: To do this, use the `felm()` function in the `lfe` package. `felm()` takes a regression formula just like `lm()`. Then use `summary(the_estimate, robust = T)` to show the heteroskedasticity-robust standard errors.

Example:

```
# The regression
some_reg = felm(y ~ x, data = fake_data)
# Print the coefficients w/ het-robust standard errors
summary(some_reg, robust = T)
```

Answer

```
# Load the 'lfe' package
p_load(lfe)
# Same regression as in (2c)—but with 'felm'
reg_2h = felm(price ~ area + quality, data = housing_df)
# Print the coefficients w/ and w/out het-robust standard errors
reg_2h %>% summary(robust = T)
reg_2h %>% summary(robust = F)
```

```
## Coefficients:
##                Estimate Robust s.e t value Pr(>|t|)
##  (Intercept) -1.099e+05  5.788e+03  -18.99   <2e-16 ***
##  area         5.875e+01  4.432e+00   13.26   <2e-16 ***
##  quality      3.324e+04  8.464e+02   39.27   <2e-16 ***
```

```
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
##  (Intercept) -1.099e+05  3.421e+03  -32.13   <2e-16 ***
##  area         5.875e+01  1.841e+00   31.91   <2e-16 ***
##  quality      3.324e+04  6.596e+02   50.40   <2e-16 ***
```

The estimated coefficients are the same across the two sets of estimates (with and without heteroskedasticity-robust standard errors), because they both use OLS to estimate the coefficients. The standard errors change because they use different estimators for the standard errors—a heteroskedasticity-robust estimator and an estimator that assumes homoskedasticity. The heteroskedasticity-robust standard errors are larger.

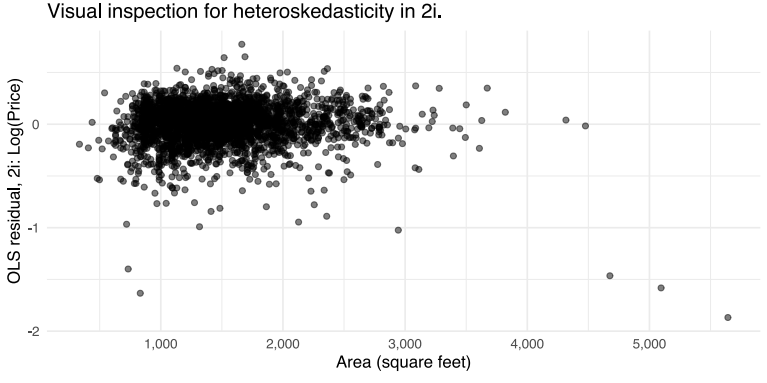
2i. As we discussed in class, we can introduce heteroskedasticity by misspecifying our regression model. Try taking the log of price. Then plot the new residuals against area. Explain whether this change in the regression specification suggests that misspecification was creating the heteroskedasticity.

Note: You do not need to formally test for heteroskedasticity.

Answer

Taking the log of price does seem to help with our heteroskedasticity a bit—at least with respect to the variable area.

```
# Regression with all variables, quadratics, and interactions
reg_2i = lm(log(price) ~ area + quality, data = housing_df)
# Add residuals to dataset
housing_df$e_2i = residuals(reg_2i)
# Plot residuals against area with ggplot
ggplot(data = housing_df, aes(x = area, y = e_2i)) +
  geom_point(alpha = 0.5) +
  scale_x_continuous("Area (square feet)", labels = scales::comma) +
  scale_y_continuous("OLS residual, 2i: Log(Price)") +
  ggtitle("Visual inspection for heteroskedasticity in 2i.") +
  theme_minimal()
```



2j. Should we interpret the regression results in (2c)—or your preferred specification in (2i)—as *causal*? Explain your answer. If we cannot interpret the regression as causal, can we still learn something interesting here? Explain.

Answer We probably should not apply a causal interpretation to our estimated coefficients in (2c). There are likely many omitted variables that are (1) correlated with *area* and *quality* and (2) affect the price of a house.

One example is the age of the house. For example, the correlation between quality and age is -0.6. If the age directly affects our outcome variable (price), then our estimate on quality will suffer from omitted-variable bias. Specifically, if we think age negatively affects our outcome variable (price)—and we already know age is negatively correlated with quality—then our coefficient on area should be an overestimate of its true effect. Let's try including *age* (`age`).

```
# The results with only share_middle
reg_2c %>% tidy()
```

```
## # A tibble: 3 x 5
##   term         estimate std.error statistic   p.value
##   <chr>           <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept) -109919.    3421.       -32.1 2.74e-194
## 2 area             58.8      1.84      31.9 4.97e-192
## 3 quality       33241.     660.        50.4 0.
```

```
# The results from adding in age
lm(price ~ area + quality + age, data = housing_df) %>% tidy()
```

```
## # A tibble: 4 x 5
##   term         estimate std.error statistic   p.value
##   <chr>           <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept) -54266.    4738.       -11.5 9.70e- 30
## 2 area            63.1      1.78      35.4 1.95e-228
## 3 quality      26009.     773.        33.7 2.90e-210
## 4 age           -495.      30.5      -16.3 5.53e- 57
```

Just as we thought: Including *age* **decreases** the estimated 'effect' of *quality*, which could suggest that the original estimate overestimating the actual effect of quality.