AAEC 4804/5804G, STAT 4804: Fundamentals of Econometrics

Your Name Here

Spring 2025 – Homework #5

Instructions

This homework is intended to help you review the material covered in Lecture 7. This week we have a greater emphasis on real world application.

When requested, please ensure that you specify the null and alternative hypotheses that you are testing. Just as important, you should also provide a brief discussion of your conclusion and interpretation of the results– "Reject the Null" is not sufficient.

You are strongly encouraged to work with your classmates, but you must submit your own Solutions.

Question 1: Consequences of Heteroskedasticity

TRUE/FALSE/MAYBE: In the presence of heteroskedasticity:

(a) Our OLS estimators are inconsistent.
(b) Our usual F statistics are no longer has an F distribution.
(c) Our OLS estimators are no longer BLUE.

Briefly explain your answers.

Question 2: Transformations and Heteroskedasticity

Consider the following linear model to explain the consumption of beer among college students:

$$beer = \beta_0 + \beta_1 income + \beta_2 price + \beta_3 age + \beta_4 female + u_t$$

Assume that: $E(u|X) = 0$ and $Var(u|X) = \sigma^2 inc^2$.)

(a) Write the transformed equation that has a homoskedastic error term.

(b) Show with quick algebra that the error term of the transformed model is indeed homoskedastic.

Question 3: Heteroskedasticity in Time Series Data

Although we did not cover time series data in class this semester, the concept and method are similar to those we have used in class. The only difference is that the data is ordered by time.

For this question, we will try detecting and correcting for heteroskedasticity in US corn yield (measured in bushels/acre). For your convenience, I have pulled data from the USDA's NASS QuickStat. The data file of interest is `corn-yield.csv` in the `Data` folder on `GitHub`.

Importing Data

(i) Using your class codes, import the data directly from GitHub. You will need to do some data cleaning to ensure that you are left with only the `Year` and `Value` columns.

Tasks:

- Under the `Data.Item` column, filter for `CORN, GRAIN - YIELD, MEASURED IN BU / ACRE`.
- Under `Period`, you are interested only in the `YEAR` line items.
- Within the `Geo.Level` column, you are interested in the `NATIONAL` line items.
- Finally, keep only the two (2) columns of interest.

Data Exploration

(ii) Using the `ggplot()` function, plot the data as points and **briefly** discuss your observations. **Please center your discussion around the dynamics and variability you observe before and after the introduction and improvement of the hybrid corn seed in the 1930 and mid-1950's, respectively.** This article can help to bolster the quality of your discussion.

Shifting Mean Regression

(iii) Notice from the dynamics above and the referenced-article that we might have different "regimes" and potential structural breaks in the data. We could potentially model these dynamics using a so-called **Shifting Means Regression model** of the form:

$$yield_t = \beta_0 + \beta_1 t + \sum_{k=1}^{3} \gamma_k sin\left(\frac{2 \cdot \pi \cdot k \cdot t}{T}\right) + \sum_{k=1}^{3} \delta_k cos\left(\frac{2 \cdot \pi \cdot k \cdot t}{T}\right) + u_t,\ t \in \{1, \ldots, T\}$$

where t is a time trend, T is the number of periods or observations ($T = 159$) in the data, and k is the number of harmonics terms that should be included.

Tasks:

- Estimate the model via the `lm()` function and store the model fit in a variable called `seg.fit`.
- Using the `patchwork` package produce a gridded plot of:
  - The actual (as points) and fitted values (as a line) over time.
  - The model residuals against the fitted values. **Add the zero line (as a red dashed line) to the residuals plot.**
- **Briefly** discuss your observations in both graphs. In particular, do you think the shifting means model does a good job of capturing the overall dynamics of the corn yield over time? Why or why not? Do the residuals appear to be homoskedastic? Why or why not?

Robust Errors

(iv) Using the `stargazer` package, report all relevant regression results along with the (a) normal standard errors and (b) White-Corrected Standard Errors (`HC3`) for the model stored in `seg.fit`. **Ensure that your models are well labeled.**

Discuss/compare the usual standard errors and t-ratios with their robust counterparts. How would inference had change if we were to rely on the OLS standard errors?

Hint: You can report the coefficient, standard errors, and t-ratios at once by including `report = c("vc*st")` as an additional argument in the `stargazer` function.

Test for Heteroskedasticity

(v) Conduct a modified White test for heteroskedasticity using the squared residuals and fitted values `seg.fit`. That is, estimate a model of the form:

$$\hat{u}_t^2 = \beta_0 + \beta_1 \widehat{yield}_t + \beta_2 \widehat{yield}_t^2 + \varepsilon_t$$

- Report the regression results. Are the slope coefficients individually significant?
- Using the `linearHypothesis` function, perform an F-test for the **joint significance** of the explanatory variables. What do you conclude about the presence of heteroskedasticity? **Be sure to be explicit about the null and alternative hypotheses we are testing.**
- Based on your findings, does it seem more appropriate to rely on robust standard errors? Why?

Estimated Heteroskedasticity Function

(vi) One approach to adjusting for heteroskedasticity is to specify an explicit equation for the conditional variance, and then use the fitted values, $\hat{h}_t$, from this equation to perform weighted least squares (WLS).

From the corn yield model above consider the regression:

$$\log(\hat{u}_t^2) = \delta_0 + \delta_1 t + \sum_{k=1}^{3} \eta_k \sin\left(\frac{2\pi kt}{159}\right) + \sum_{k=1}^{3} \psi_k \cos\left(\frac{2\pi kt}{159}\right) + \xi_t$$

(a.) - Estimate and report the regression indicated above. **Be sure to also store the model fit in this step.**

- Conduct an F-test for the **joint significance** of the explanatory variables. Discuss your conclusions.

(b.)

- Obtain the fitted values from this regression in part (a)– these are your $g(x) = \widehat{\log \hat{u}_t^2}$ in your notes– and exponentiation them to get $\hat{h}(x) = \exp(g(x))$.

For clarity, $\hat{h}(x) = \exp\left(\hat{\delta}_0 + \hat{\delta}_1 t + \sum_{k=1}^{3} \hat{\eta}_k \sin\left(\frac{2\pi kt}{159}\right) + \sum_{k=1}^{3} \hat{\psi}_k \cos\left(\frac{2\pi kt}{159}\right)\right)$

- We can use these fitted values to compute the time-varying standard deviation as $\hat{\sigma}_t = \sqrt{\hat{h}_t}$. **Plot these values over time and discuss what you observe. I would like to see you use the `bquote()` function to label your y-axis as $\sqrt{\hat{h}_t}$.**

(c.) Your professor suggests that you can use estimates of $\hat{\sigma}_t$ to potentially represent approximate 90% confidence intervals for the fitted values from your **base regression** in `seg.fit`.

- Create two additional series `Upper` and `Lower` based on $\hat{yield}_t \pm 1.645 \cdot \hat{\sigma}_t$. Plot (i) the values of your upper and lower CIs (as dashed lines) over time along with (ii) the fitted values (as a solid line) from the base regression, and (iii) the actual yield values (as points with `alpha=.5`). **Ensure your each element of your graph is well labeled (you might need the `scale_color_manual()` function for this). Briefly discuss your observations.**

Feasible GLS and heteroskedasticity function

(vii) Using the fitted values, $\hat{\sigma}_t = \sqrt{\hat{h}_t}$, from the previous question, we can manually estimate a feasible GLS model in a WLS framework.

(a) Divide all original variables (including the intercept) by $\hat{\sigma}_t$ and re-estimate the base model in `seg.fit`. **Remember to exclude the intercept.** Store the model fit as `seg.wls` then report the model summary.

(b) As we mentioned in class, since the WLS model above does not have an intercept, we cannot place a lot of stock in the model's goodness of fit, R^2. As such you will need to obtain and report a correct estimate of the R^2 for this regression.

This is fairly easy. You need only regress the original dependent variable ($yield$) on the fitted values from the WLS regression divided by their respective weights ($\widehat{yield}_t/\sqrt{\hat{h}_t}$). The R^2 from this auxiliary model will serve as the corrected R^2. **Conduct this regression and report the R^2 value.**

(viii) After all your hard work, your roommate, Emmalee, pointed out that she saw where `R`'s `lm` function has a `weight` command that could have estimated the same model as you just did with the transformed variables. You are happy that she pointed this out, but you are also a bit annoyed that you did all that work while she sat on that information and just watched you.

Use the `lm()` function with the appropriate weight to confirm your results from part (vii)(a). Store the model fit as `seg.wls2` then report the model summary.

Hint: You should note that the coefficients are the same as your manual approach but the R^2 values might not be a one-to-one match.

(ix) Using the `stargazer` package, compare your results including parameter estimates, standard errors, and t-ratios in `seg.wls2` with those you reported in `seg.fit` from (iii). **Be sure to properly label your models.** Discuss your findings and any implications for inference.