CHE384, From Data to Decisions: Measurement, Uncertainty, Analysis, and Modeling

Lecture 24
Heteroscedasticity: When Variance Varies

Chris A. Mack

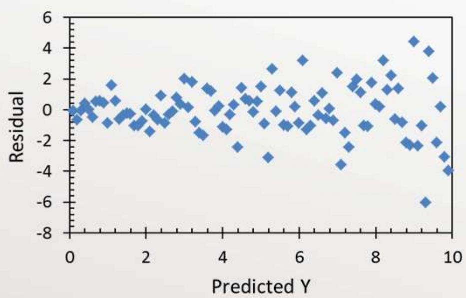
Adjunct Associate Professor

http://www.lithoguru.com/scientist/statistics/

Assumptions in OLS Regression

1. ε is a random variable that does not depend on *x* (i.e., the model is perfect, it properly accounts for the role of x in predicting y)
2. $E[\varepsilon_i] = 0$ (the population mean of the true residual is zero); this will always be true for a model with an intercept
3. All ε_i are independent of each other (uncorrelated for the population, but not for a sample)
4. All ε_i have the same probability density function (pdf), and thus the same variance (called homoscedasticity)
5. $\varepsilon \sim N(0,\sigma_\varepsilon)$ (the residuals, and thus the y_i, are normally distributed)
6. The values of each x_i are known exactly

Variance can Vary



Plotting Residuals (visual inspection)

- Plotting residuals is the first step in detecting variation of variance
  - Plot esr versus each predictor variable, and versus the predicted response variable
  - Very hard to see heteroscedasticity unless there are many points
  - A plot of the absolute value of the residual is sometimes more revealing (sign doesn't matter when considering variance)

Consequences of Heteroscedasticity

- Note that heteroscedasticity is often a by-product of other violations of assumptions
  - Wrong model, existence of outliers, non-normal errors
  - We'll assume here that only heteroscedasticity is present in our data
- Result of heteroscedasticity will be an unbiased estimator that is inefficient
  - The standard errors of the estimates are biased
  - Only fairly large heteroscedasticity matters much

Common Ways Variance Varies

- If the experimental y-value is a mean, but the sample size is different for each calculated mean
  - $SE(\bar{y}) = \sigma/\sqrt{n}$
  - Ex: Average income vs. years of college
- Variance or standard error is a constant percentage of the y-value
- Variance has been experimentally determined for each y-value
- Some distributions naturally have variance that is a function of the mean (Poisson), or mean and variance both a function of parameters (Gamma)

Checking the Variance

- Constant variance (variance is independent of the value of the predictor variable) is called homoscedasticity
- Non-constant variance (variance is not independent of the value of the predictor variable) is called heteroscedasticity
- Two ways to check for heteroscedasticity:
  - Independent knowledge of the variance of the measured y-values
  - Statistical tests for homoscedasticity

Statistical Tests for Homoscedasticity

- Divide the residuals (esr for fits) into two or more sub-groups (sort by magnitude of $\hat{y}$)
  - Test to see if the sub-groups share the same variance (Null hypothesis: all groups have the same variance)
  - The Bartlett test compares variances; it assumes a normal distribution and is sensitivity to deviations from normality
  - The Brown-Forsythe test (modified Levene test) compares deviations from the median; it is insensitive to departures from normality, but has somewhat less power

Bartlett Test

- The Barlett statistic is χ^2 distributed with k-1 degrees of freedom

$$T = \frac{(N-k)\ln s_{pool}^2 - \sum_{j=1}^{k}(n_j-1)\ln s_j^2}{1+\left(1/\left(3(k-1)\right)\right)\left(\left(\sum_{j=1}^{k} 1/(n_j-1)\right) - 1/(N-k)\right)}$$

N = total number of data points
k = number of sub-groups
n_j = sample size of the j^th sub-group
s_j^2 = variance of the j^th sub-group

$$s_{pool}^2 = \sum_{j=1}^{k} \frac{(n_j-1)s_j^2}{N-k}$$

Bartlett Test

- For two equal-sized subgroups (e.g., after rank-ordering by $\hat{y}$ and dividing in half),

$$T = \frac{(N-2)^2}{N-1}\ln\frac{s_{pool}^2}{s_1 s_2} \qquad s_{pool}^2 = \frac{s_1^2+s_2^2}{2}$$

- The null hypothesis (that the two sub-groups have equal variance) can be rejected if T is greater that the critical $\chi^2(1)$
  - For α = 0.05, the critical value is 3.84
  - For α = 0.01, the critical value is 6.63

Brown-Forsythe Test

- Divide the data into two subgroups (of size n_1 and n_2), calculate the median of each group (m_1 and m_2), then the absolute deviation from the median for each data point

$$d_{i1} = |x_{i1} - m_1| \qquad d_{i2} = |x_{i2} - m_2|$$

- Calculate the mean absolute deviation for each group ($\bar{d}_1$ and $\bar{d}_2$) and the variance of the absolute deviations for each group (s_{d1}^2 and s_{d2}^2)
- The pooled variance is

$$s_{pool}^2 = \frac{(n_1-1)s_{d1}^2 + (n_2-1)s_{d2}^2}{n_1+n_2-2}$$

Morton B. Brown and Alan B. Forsythe, "Robust Tests for the Equality of Variances", *Journal of the American Statistical Association*, **69**(346), pp. 364-367 (Jun., 1974).

Brown-Forsythe Test

- The studentized difference between the mean absolute deviations for each group ($\bar{d}_1 - \bar{d}_2$) is about t-distributed with $n-2$ degrees of freedom

$$t = \frac{|\bar{d}_1 - \bar{d}_2|}{s_{pool}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

  - Assumes a not too small value of n ($n_1, n_2 > 25$)
  - Because we use deviations from the median, the statistic is insensitive to the distribution of x
  - Two tailed test, null hypothesis: constant variance

Other Tests for Homoscedasticity

- White test: perform linear regression of ε_i^2 with x and test nR^2 as χ^2(p-1)
- Breucsh-Pagan test: a variation of the White test where x is replaced with any variable(s) of interest
- Park test: perform linear regression of ln(ε_i^2) with ln(x) and test the significance of the slope (is it significantly different from 0)

Lecture 24: What have we learned?

- Define homoscedasticity and heteroscedasticity
- What are the consequences of heteroscedasticity to your regression?
- What are some of the causes of heteroscedasticity?
- What are the advantages of either the Bartlett test or the Brown-Forsythe test?