Assumptions in OLS Regression

1. \( \varepsilon \) 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 \( \varepsilon_i \) are independent of each other (uncorrelated for the population, but not for a sample)
4. All \( \varepsilon_i \) have the same probability density function (pdf), and thus the same variance (called homoscedasticity)
5. \( \varepsilon \sim N(0, \sigma^2) \) (the residuals, and thus the \( y_i \), are normally distributed)
6. The values of each \( x \) are known exactly

Variance can Vary

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
  - \( SE(y) = \sigma / \sqrt{n} \)
  - 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 (\(e_r\) for fits) into two or more sub-groups (sort by magnitude of \(y\))
  - 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 sensitive 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 Bartlett statistic is \(\chi^2\) distributed with \(k-1\) degrees of freedom.

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

- For two equal-sized subgroups (e.g., after rank-ordering by \(y\) and dividing in half),
  - The null hypothesis (that the two sub-groups have equal variance) can be rejected if \(T\) is greater than the critical \(\chi^2_{1}\).
  - For \(\alpha = 0.05\), the critical value is 3.84.
  - For \(\alpha = 0.01\), 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.

\[
d_{1j} = |x_{1j} - m_1| \quad d_{2j} = |x_{2j} - m_2|
\]

- Calculate the mean absolute deviation for each group \((d_{1} \text{ and } d_{2})\) and the variance of the absolute deviations for each group \((s_{1}^{2} \text{ and } s_{2}^{2})\).

\[
s_{pool}^{2} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}
\]

- The studentized difference between the mean absolute deviations for each group \((d_{1} - d_{2})\) is about t-distributed with \(n - 2\) degrees of freedom.

\[
t = \frac{|d_1 - d_2|}{s_{pool} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}
\]

- Assumes a not too small value of \(n_1, n_2 \geq 25\).
- 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 \( \epsilon_i^2 \) with \( x \) and test \( nR^2 \) as \( \chi^2(p-1) \)
- **Breusch-Pagan test**: a variation of the White test where \( x \) is replaced with any variable(s) of interest
- **Park test**: perform linear regression of \( \ln(\epsilon_i^2) \) 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?