- **Tests of Normality**

I. **Graphical Summary** :

- use a QQ plot (quantile-quantile) plot to plot the residuals of a fitted model against the expected quantiles of a $N(0,1)$ random variable.
- If our model assumption is correct, then we expect that we will see a straight line on this plot. Sharp deviations from this suggest evidence against normality.

II. **Shapiro-Wilk Test of Normality**

- The QQ-plot is inherently testing whether or not the residuals follow a straight line when plotted against the quantiles of a $N(0,1)$ random variable.

- This is ~~challeg~~ challenging to test directly, but what one can do is test whether the correlation between the standardized residuals $e_1^*, \ldots, e_n^*$ and normal quantiles $z_1^\dagger, \ldots, z_n^\dagger$ is close to 1. If yes, we have good evidence of normality.

In particular, define

$$z_i^+ := z_{(i-0.375)/(n+0.25)}, \quad e_i^* = \frac{e_i - \bar{e}}{SD(e_1, \dots, e_n)}$$

Then the test statistic is

$$r^* := \frac{\sum z_i^+ e_i^*}{\sqrt{\sum z_i^{+2} \sum e_i^{*2}}}$$

The formal test is (the <u>Shapiro Wilk test</u>)

$$H_0: \rho^* = 1 \quad \text{vs.} \quad H_1: \rho^* < 1$$

* Generally, one calculates r^* through simulation and calculates an approximate p-value to make a decision on the above test.

III. <u>Kolmogorov - Smirnoff Test</u>

* The K-S test is one of a suite of empirical distribution tests. The <u>empirical distribution function</u> (EDF) of n observations $x_1, \dots, x_n$ is:

$$F_n(y) := \frac{1}{n} \# \{\text{observations } x_i \le y\}, \quad y \in \mathbb{R}$$

To test whether $e_1^*, \cdots, e_n^*$ ~~is~~ are normally distributed, we can calculate the EDF of its values and compare directly with the distribution of a $N(0,1)$ random variable.

* The K-S statistic is defined as

$$D = \max_{y \in \mathbb{R}} | F_n(y) - F_0(y) |$$

where F_0 is the cdf of the RV under the null hypothesis. D measures the maximum distance between the cdfs F_n & F_0.

~~In particular~~,

For testing normality of $e_1^*, \cdots, e_n^*$, let $F_n(y)$ = distribution (EDF) of $e_1^*, \cdots, e_n^*$ and let $F_0(y)$ = cdf value of a $N(0,1)$ RV at y. Then, the K-S test is

$$H_0: F_n(y) = F_0(y) \quad \text{vs.} \quad H_1: F_n(y) \neq F_0(y)$$

Rejecting H_0 gives evidence against our Normal assumptions.

Testing and Correcting for Heteroscedasticity

- Weighted Least Squares Model

$$y_i = \sum_{j=1}^{p} x_{ij} \beta_j + \beta_0 + \varepsilon_i$$

where $\{\varepsilon_i\}$ are uncorrelated with mean 0, but we allow differing variances:

$$Var(\varepsilon_i) = \sigma^2 v_i$$

with v_i <u>known</u>.

- This model, in vector form, can be generalized as

$$Y = X\beta + \varepsilon \quad \text{where}$$

(1)

i) $E[\varepsilon] = \underset{\sim}{0}$

ii) $Cov(\varepsilon) = \sigma^2 V$

and $V = diag(v_i)$.

This is known as the generalized linear regression model.

- Minimizing $SSE(X\beta)$ wrt β gives the <u>generalized least squares estimators</u>:

$$\hat{\beta}_{GLS} = (X^T V^{-1} X)^{-1} X^T V^{-1} Y$$

<u>Properties of $\hat{\beta}_{GLS}$</u>:

1) $E[\hat{\beta}_{GLS}] = \beta$ under model (1)

~~2) $Var(\hat{\beta}_{GLS}) = \sigma^2 (X^T X)^{-1} (X^T V X)(X^T X)^{-1}$~~

3) $\hat{\beta}_{GLS}$ is BLUE for ~~(1)~~ β under model (1)

2) $Var(\hat{\beta}_{GLS}) = (X^T V^{-1} X)^{-1} \sigma^2$

- <u>Note</u>: (3) implies that $\hat{\beta}_{GLS}$ has smaller variance than $\hat{\beta}_{OLS}$ if model (1) is true.

- <u>Testing for Constant Variance</u>

- There are asymptotic tests that have been developed for this; however, authors have pointed out that they are a bit challenging to understand and implement. The accepted procedure to test for constant variance is the following.

- <u>Test</u>: H_0: σ^2 fixed for obs. $i = 1, 2, \ldots n$

  vs. H_1: there exists an obs. j for which $\sigma_j^2 \neq \sigma_i^2$

Procedure:

1) Fit squared residuals as a regression on pairs $x_{ij} x_{ik}$ for all $\frac{p(p+1)}{2}$ possible pairs:

$$e_i^2 = \alpha_0 + \alpha_1 x_{i1} x_{i2} + \cdots + \alpha_{\frac{p(p+1)}{2}} x_{ip-1} x_{ip}$$

to get $\hat{\alpha}_0, \hat{\alpha}_1, \ldots, \hat{\alpha}_{\frac{p(p+1)}{2}}$

2) Calculate the R^2 value of the fit (the squared multiple correlation coefficient)

3) Reject H_0 if

$$nR^2 > \chi^2_{\frac{p(p+1)}{2}, 1-\alpha}$$

* Main strategy for testing for constant variance

1. Fit the regression model
2. Look at $\underset{\sim}{e}$ against $\underset{\sim}{y}$. You should hope for constant variability
3. If not constant, perform the above test
4. If you identify non-constant variance, either consider transformations on y or use generalized least squares.

* Do Ch. 4 in Faraway