L8. Modeling Problems - Other Diagnostics

Modeling Assumption Violations:

(1) Heteroscedasticity

(2) Non-Normality Residuals

(3) False assumption of linearity between response variable Y and any predictor.

1. Heteroscedasticity.

Recall Model Assumption: $\varepsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2),\ i=1,\cdots,n$

or $\vec{\varepsilon} \sim MVN(\vec{0}, \sigma^2 I_n)$

The OLSE method requires all ε_is have a constant variance σ^2, s.t. the parameter estimate

$\vec{b} = (\vec{X}^T\vec{X})^{-1}\vec{X}^T\vec{y}$ is the "best" estimator

in "BLUE".

• Hesteroscedasticity occurs when this assumption is violated, which means the variance of ~~residuals~~ errors are non-constant, more commonly, it referes to the spead of residuals changes symastematically with predictors.

(a) Problem.

(i) The OLSE estimate $\vec{b}$ is still linear and unbiased, but not the "best" anymore: we can't say $\vec{b}$ has the smallest variance among all unbiased linear estimators. There is another estimator with a smallest variance.

(ii) $se(\hat{\beta}_K)$ for all OLS output are incorrect estimate of the standard deviation of $\hat{\beta}_K$. It will result in misleading t test and C.I.

(iii) Predictions of y are still unbiased, but the prediction intervals are incorrect.

Recall: $Var(\vec{b}) = E(\vec{b}-\vec{\beta})(\vec{b}-\vec{\beta})^T$

$$= E[(\vec{X}^T\vec{X})^{-1}\vec{X}^T \vec{\varepsilon}\vec{\varepsilon}^T \vec{X}(\vec{X}^T\vec{X})^{-1}]$$ lecture 4

$$= (\vec{X}^T\vec{X})^{-1}\vec{X}^T \cdot E(\vec{\varepsilon}\vec{\varepsilon})^T \cdot \vec{X}(\vec{X}^T\vec{X})^{-1}$$

When applying OLS method:

assume $E(\vec{\varepsilon}\vec{\varepsilon})^T = \sigma^2 I_n = \begin{pmatrix} \sigma^2 & & 0 \\ & \ddots & \\ 0 & & \sigma^2 \end{pmatrix}$

so $Var(\vec{b})_{OLS} = \sigma^2(\vec{X}^T\vec{X})^{-1}$

and is estimated by $MSE(\vec{X}^T\vec{X})^{-1}$

which is an unbiased estimator of $Var(\vec{b})$ when the assumption of constant variance holds.

When Heteroscedasticity exist:

the real $Var(\vec{b}) = (\vec{X}^T\vec{X})^{-1}\vec{X}^T \begin{pmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_n^2 \end{pmatrix} \vec{X}(\vec{X}^T\vec{X})^{-1}$

so $MSE(\vec{X}^T\vec{X})^{-1}$ is now biased estimate for $Var(\vec{b})$

In this case, if we still apply OLS method:

(i) $Var(\vec{b})_{OLS} \neq Var(\vec{b})$.

Not "best" anymore

(ii) $\widehat{Var}(\vec{b})_{OLS} = MSE \cdot (\vec{X}^T \vec{X})^{-1}$ is

not unbiased anymore, so

$se(\hat{\beta}_K)_{OLS}$ is incorrect.

(iii) Let's give some result about the

prediction first

Textbook section 6.7

For a new observation with $\vec{X}_h = \begin{pmatrix} 1 \\ X_{h1} \\ \vdots \\ X_{h,p-1} \end{pmatrix}$

Note: this is a matching notation as in the textbook, might look confusing since they transposed the old way of writng X matrix.

The fitted value $\hat{Y}_h = \vec{X}_h^T \vec{b}$

- unbiased : $E(\hat{Y}_h) = \vec{X}_h^T \vec{\beta} = E(Y_h)$

- $Var(\hat{Y}_h) = Var(\vec{X}_h^T (\vec{X}^T \vec{X})^{-1} \vec{X}^T \vec{y})$

$\overset{\text{when assumption holds}}{=} \sigma^2 \vec{X}_h^T (\vec{X}^T \vec{X})^{-1} \vec{X}_h$

$$Var(\hat{Y}_h) = Var(\vec{X}_h^T \vec{b})$$

$$= Var(\vec{X}_h^T (\vec{X}^T\vec{X})^{-1} \vec{X}^T \vec{y})$$

$$= [\vec{X}_h^T (\vec{X}^T\vec{X})^{-1} \vec{X}^T] \underbrace{Var(\vec{y})}_{\sigma^2 I_n} [\vec{X}_h^T (\vec{X}^T\vec{X})^{-1} \vec{X}^T]^T$$

$$= \sigma^2 \vec{X}_h^T \cancel{(\vec{X}^T\vec{X})^{-1} \vec{X}^T \vec{X}} (\vec{X}^T\vec{X})^{-1} \vec{X}_h^T$$

$$= \sigma^2 \vec{X}_h^T (\vec{X}^T\vec{X})^{-1} \vec{X}_h$$

Hint: $Var(\underset{\substack{\uparrow \\ \text{constant matrix}}}{C} \cdot \vec{X})$ ← r.v. vector

(Quiz 2 Q7)

$$= C \, Var(\vec{X}) \, C^T$$

$Var(\hat{Y}_h)$ can also be found as $Var(\vec{X}_h^T \vec{b})$

$= \vec{X}_h^T \, Var(\vec{b}) \, \vec{X}_h = \vec{X}_h^T \sigma^2 (\vec{X}^T\vec{X})^{-1} \vec{X}_h$ (same as above)

If $s^2(\vec{b})$ is given by $MSE(\vec{X}^T\vec{X})^{-1}$ (sample estimation)

$s^2(\hat{Y}_h) = \vec{X}_h^T s^2(\vec{b}) \vec{X}_h$ (HMK 3 6.25)

- Confidence Interval for the Mean Response. μ_h

$$\hat{Y}_h \pm t_{\alpha/2, df=n-p}\ se(\hat{Y}_h)$$

where $se(\hat{Y}_h) = \sqrt{MSE\ \vec{X}_h^T (\vec{X}^T\vec{X})^{-1} \vec{X}_h^T}$

- Prediction Interval for the new observation Y_h

$$\hat{Y}_h \pm t_{\alpha/2, df=n-p} \sqrt{se(\hat{Y}_h)^2 + MSE}$$

when Heteroscedasticity exists.

$$Var(\hat{Y}_h) = Var(\vec{X}_h^T (\vec{X}^T\vec{X})^{-1} \vec{X}^T \vec{y})$$

$$= [\vec{X}_h^T (\vec{X}^T\vec{X})^{-1} \vec{X}^T]\ \underbrace{Var(\vec{y})}_{\neq \sigma^2 I_n}\ [\vec{X}_h^T (\vec{X}^T\vec{X})^{-1} \vec{X}^T]^T$$

So $MSE\ \vec{X}_h^T (\vec{X}^T\vec{X})^{-1} \vec{X}^T$ is not unbiased estimate anymore.

Therefore prediction/confidence interval calculated this way would be incorrect.

(b) Detection.

(i) Residual v.s. Fitted Value Plot:

You would observe obvious change of bandwidth in the plot

(ii) Brensch - Pagan Test

Basic Idea is the variance of error should not change given difference prediction values.

If the assumption of constant variance is violated, then the variance would be changed with predictor values.

1) Fit your model $y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{ip-1} + \varepsilon_i$.

2) Obtain the residuals $e_1, e_2, \cdots, e_n$

3) Build an auxiliary regression model:

$$e_i^2 = \gamma_0 + \gamma_1 X_{i1} + \cdots + \gamma_{p-1} X_{i,p-1} + \xi_i \quad (E)$$

4) $H_0: \gamma_1 = \gamma_2 = \cdots = \gamma_{p-1} = 0$ v.s H_a: at least one $\gamma_i \neq 0$

5> Test stat : $\chi_s^2 = nR^2$

where R^2 is the coef. of determination of model (E)

where Ho is true: $\chi_s^2 \sim \chi^2(p)$

6> Test Result:

when Fail to reject Ho: $\chi_s^2 < \chi^2_{\alpha, df=p}$ or p-value > ~~0.05~~ α

we conclude there's not significant heteroscedasticity.

when reject Ho: $\chi_s^2 \geq \chi^2_{\alpha, df=p}$ or p-value $\leq \alpha$.

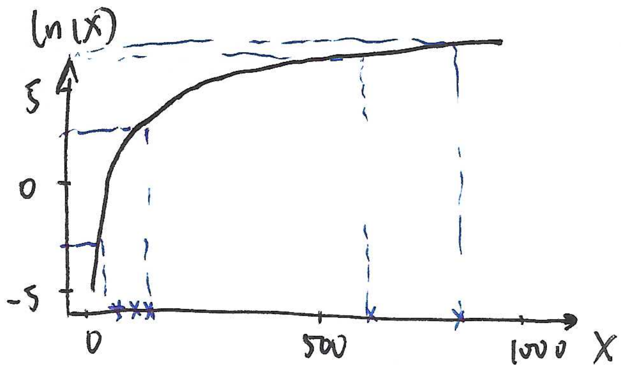
we conclude there's significant heteroscedasticity problem.

(iii) White Test * is another common one.

(c) Solution

(i) log-transformation on y

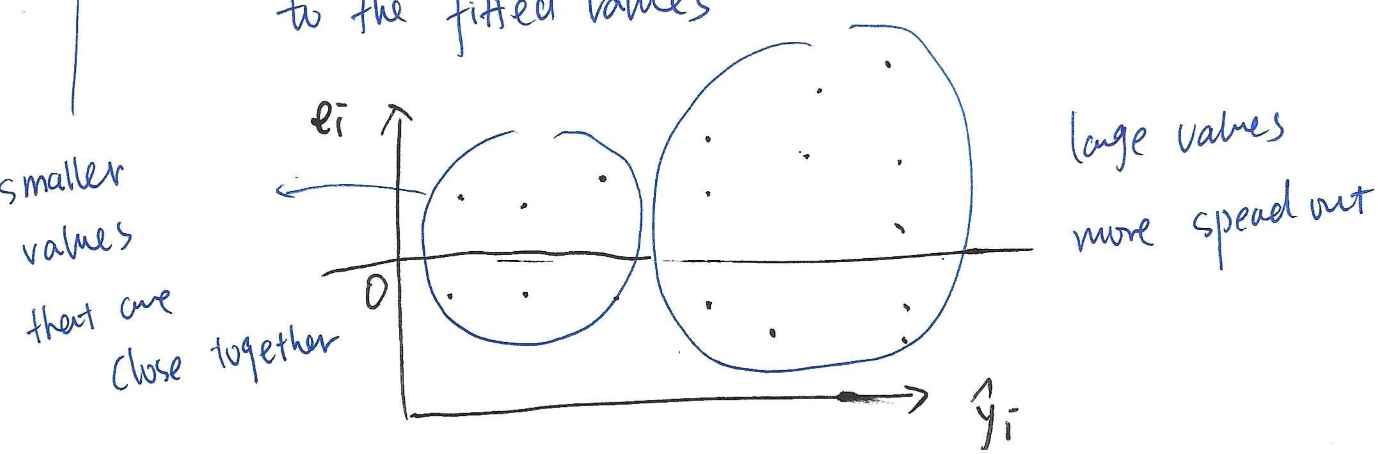
- Natural logarithm, ln, transformation is the most common transformation in MLR

- 

From the plot, the effects of taking natural log transformation are:

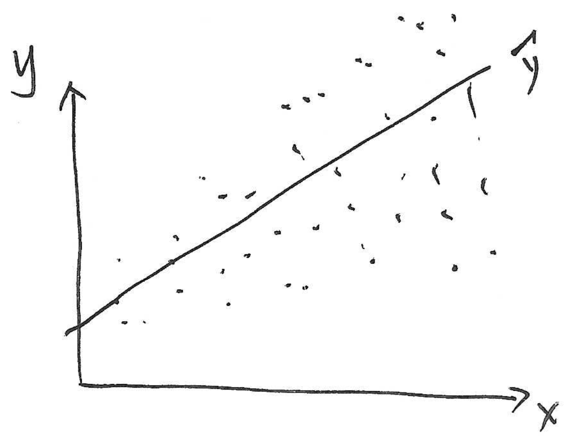
- small values that are close together are spread further out
- large values that are spread out are brought close together

Therefore, it works when the residuals changes proportionally to the fitted values



since $e_i = y_i - \hat{y}_i$

the heteroscedasticity problem lands on "y" in the data.



distribution of y around the fitted line also changes from clusters of smaller values to the spead of large values.

- In this case, a natural-log transformation on y could help "even-out" the points:

$y_{-new} = \ln(y)$

smf.ols("y_new ~ X1+ ··· + Xp-1")

$\hat{y}_{new} = \ln(\hat{y}) \rightarrow \hat{y} = e^{\hat{y}_{-new}}$

Since $\ln y$ only works for the case of e_i is proportional with $\hat{y}_i$

or "Funnel" shape, other solutions can be

(ii) Weighted least-square regression. — count different variances in the estimation.

Recall OLSE assumes $Var(\vec{\varepsilon}) = \begin{pmatrix} \sigma^2 & & 0 \\ & \ddots & \\ 0 & & \sigma^2 \end{pmatrix}$

while reality in heteroscedasticity is

$$Var(\vec{\varepsilon}) = \begin{pmatrix} \sigma_1^2 & & & 0 \\ & \sigma_2^2 & & \\ & & \ddots & \\ 0 & & & \sigma_n^2 \end{pmatrix} = \Sigma$$

$$Var(\vec{b}) = (\vec{X}^T\vec{X})^{-1}\vec{X}^T \cdot \underbrace{E(\vec{\varepsilon}\vec{\varepsilon})^T}_{\text{AKA } Var(\vec{\varepsilon}) = \Sigma} \cdot \vec{X}(\vec{X}^T\vec{X})^{-1}$$

We need to estimate Σ by $\hat{\Sigma} = \begin{pmatrix} e_1^2 & & & 0 \\ & e_2^2 & & \\ & & \ddots & \\ 0 & & & e_n^2 \end{pmatrix}$

then estimate $Var(\vec{b})$ by $(\vec{X}^T\vec{X})^{-1}\vec{X}^T\hat{\Sigma}\vec{X}(\vec{X}^T\vec{X})^{-1}$

$$Var(\vec{\varepsilon}) = \begin{pmatrix} \sigma_1^2 & & & \\ & \sigma_2^2 & & \\ & & \ddots & \\ & & & \sigma_n^2 \end{pmatrix} = \Sigma$$

- Weighted-least-square estimates take account of the different variances in the way that

  - An observation with <u>small</u> error variation should weigh <u>more</u> in the model.

  - An observation with <u>bigger</u> error variation should weigh <u>less</u> in the model

  - It includes the Σ in the estimation in the way that

Define the Weight Matrix

$$W = \begin{pmatrix} w_1 & & & \\ & w_2 & & 0 \\ & 0 & \ddots & \\ & & & w_n \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma_1^2} & & & 0 \\ & \frac{1}{\sigma_2^2} & & \\ 0 & & \ddots & \\ & & & \frac{1}{\sigma_n^2} \end{pmatrix}$$

· Defined the weighted sum of squares:

$$SS_{WLS}(\vec{\beta}) = \Sigma w_i e_i^2$$

and $\vec{b}_{WLS} = \text{argmin } SS_{WLS} = (X^T \hat{W} X)^{-1} X^T \hat{W} Y$

where $\hat{W} = \boxed{?}$

Practically how to decide the weights can be tricky:

If we assume $\varepsilon_1 \sim N(0, \sigma_1^2), \cdots, \varepsilon_n \sim N(0, \sigma_n^2)$

we only have one observed sample ~~for~~ e_i from each ε_i.

Robust standard error later used this appraoch.

Here we just give an simple example. It's not required

to run WLS yourself:

When e_i v.s. X_K has a Funnel shape,

you can run model = " $e_i^2 \sim X_K$ "

then use fitted values of the model as the

estimated variance of $Var(\varepsilon_i)$

then $w_i = 1/\hat{e}_i^2$

(iii) Generalized least squares: again deal with the estimate of Σ in a broader choices of cases.

logistic model is a case of GLM.

(iv) Use more Robust Regression: which is not sensitive to assumptions.

(v) Use Robust standard errors (common in Economics, see supplemental material.)

(2) Sometimes, after correcting Non-Normality, Non-linearity and model selection, heteroscedasticity can be greatly improved as well.

Summary:

- Heteroscedasticity : The assumption of constant variance of ε_i is violated.

Problem

- In OLS result

$se(\hat{\beta}_k)$ is incorrect, result in misleading t test and anova test reslults.

- $Var(\hat{Y}_h)$ is incorrect. result in misleading C.I. and prediction intervals.

*: The estimation of coefficients and fitted values are still unbiased, therefore reliable.

Detection

- Residual v.s. Fitted plot

the bandwidth changes

- Breusch-Pagan Test or White Test :

p-value ≤ 0.05 indicated serious heteroscedasticity problem.

Solution:

(i) If e_i v.s. $\hat{y}_i$ is Funnel shape, perform natural log transformation on y.

(ii) Perform Weighted- ~~square~~ least-square regression.

* Practically hard to decide a good weight

(iii) Use Robust Regression methods in ML:

Huber, RANSAC, Theil Sen

(iv) Use Robust standard errors in all formulas that need the variance-coviarance matrix of error term