FALL 2021 MATH 484/564 HOMEWORK #4

Due: November 6, 11:59PM, submit in blackboard.

Homework solution is not required to be typed, but must be legible.

All plots must be computer-generated. Hand-sketched plots are not acceptable.

Problem 1 Exercise 4.8 (b) from the TEXT.

Problem 2 Given data in Table 6.2 (attached), the response variable is n_t, representing the number of surviving bacteria (in hundreds) after being exposed to X-ray for t intervals. The predictor variable is t.

1) First regress n_t on time t, plot residuals against the fitted values $\hat{n}_t$. Conclude if the relationship between the mean response and the predictor is linear.
2) Use data transformation on the response variable, i.e., regress $\log(n_t)$ on t.
   - What is the regression line equation?
   - Plot residuals against the fitted values, and conclude if the violation of the "L" assumption still exists.

Problem 3 Given data in Table 6.6 (attached), the response variable Y is the number of injury incidents, and the predictor variable N is the proportion of flights.

1) First regress Y on N, plot residuals against the fitted values $\hat{Y}$. Conclude if error is heteroscedastic, i.e., the "E" assumption is violated.
2) Use data transformation on the response variable, i.e., regress $\sqrt{Y}$ on N. The rationale behind this transformation is that the occurrence of accidents, Y, tends to follow the Poisson probability distribution, and the variance of $\sqrt{Y}$ is approximately equal to 0.25, see Table 6.5.
   - What is the regression line equation?
   - Plot residuals against the fitted values, and conclude if there is still evidence of heteroscedasticity.

Problem 4 Given the data in Table 6.9 (attached), the response variable Y is the number of supervisors, and the predictor variable X is the number of supervised workers. Based on empirical observation, it is hypothesized that the standard deviation of the error term ϵ_i is proportional to x_i:

$$\sigma_i^2 = k^2 x_i^2, \ k > 0$$

- Use the weighted least squares (WLS) method to fit the model. Provide the regression equation.
- Use data transformation method to transform Y to $Y' = Y/X$, and transform X to $X' = 1/X$ (see equations 6.11 and 6.12), and then use the ordinary least squares (OLS) method to regress Y' on X'. Provide the regression equation.
- Compare the results from the above two methods and conclude if the two methods are equivalent. You can compare the residual vs fitted value plot side by side and conclude if they have the same effect in terms of removing heteroscedasticity.

Problem 5 For the data in Problem 4, use OLS without data transformation to fit the model, i.e., directly regress Y on X, and compare the variances of the coefficients Var($\hat{\beta}_0$) and Var($\hat{\beta}_1$) with their counterparts obtained by using WLS, conclude which method yields smaller variances.

Jane Downer
Math 484

Homework 4

Problem 1

4.8 Consider again the Examination Data used in Exercise 3.3 and given in Table 3.10:

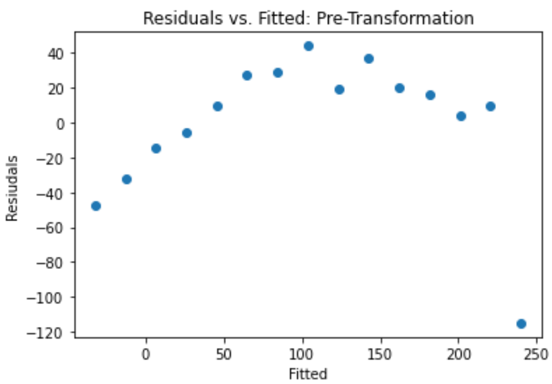
(b) What model would you use to predict the final score F?

I would use the 3rd model, which uses both P_1 and P_2 as predictors. In Homework 2 we found that Model 3 yields the lowest SSE, and in Homework 4 we found that Model 3's residuals have more evenly distributed than the other two models. This makes it a better model than Model 1 and Model 2.

Problem 2 Given data in Table 6.2 (attached), the response variable is n_t, representing the number of surviving bacteria (in hundreds) after being exposed to X-ray for t intervals. The predictor variable is t.

1) First regress n_t on time t, plot residuals against the fitted values $\hat{n}_t$. Conclude if the relationship between the mean response and the predictor is linear.
2) Use data transformation on the response variable, i.e., regress $\log(n_t)$ on t.
   - What is the regression line equation?
   - Plot residuals against the fitted values, and conclude if the violation of the "L" assumption still exists.

(1)

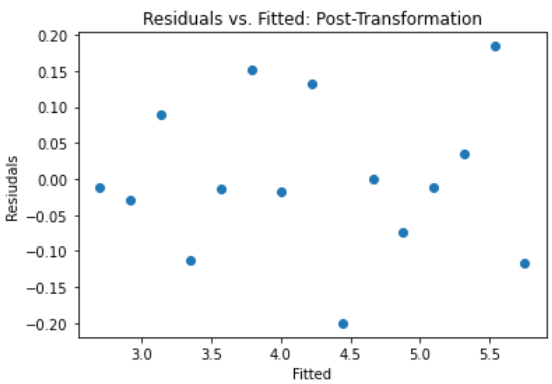


Calculated in Python

$$\boxed{\hat{n}_t = 259.58 - 19.464 * t}$$

Not linear

(2)



$n_{t_i}' = \ln(n_{t_i})$

$t' = t$

Calculated in Python

$$\boxed{\hat{n}_t' = 5.97 - 0.218 * t'}$$

"L" assumption not violated

$$\hat{\beta}_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

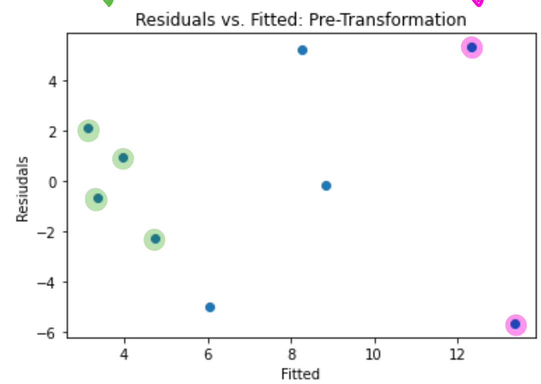
Problem 3 Given data in Table 6.6 (attached), the response variable Y is the number of injury incidents, and the predictor variable N is the proportion of flights.

1) First regress Y on N, plot residuals against the fitted values $\hat{Y}$. Conclude if error is heteroscedastic, i.e., the "E" assumption is violated.
2) Use data transformation on the response variable, i.e., regress $\sqrt{Y}$ on N. The rationale behind this transformation is that the occurrence of accidents, Y, tends to follow the Poisson probability distribution, and the variance of $\sqrt{Y}$ is approximately equal to 0.25, see Table 6.5.
   - What is the regression line equation?
   - Plot residuals against the fitted values, and conclude if there is still evidence of heteroscedasticity.

(1)

less variance

more variance



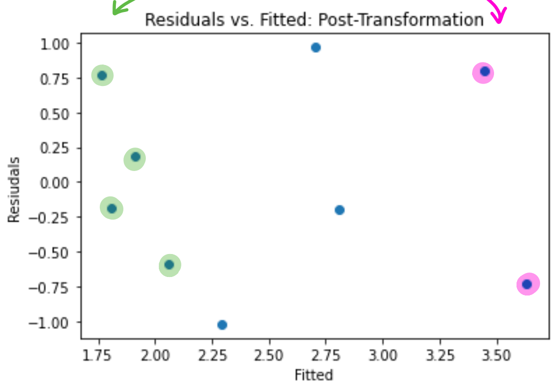
Calculated in Python

$$\boxed{\hat{y} = -0.14 + 64.975 * N}$$

Visible heteroscedasticity

(2)

roughly equal variance



$$y_i' = \sqrt{y_i}$$
$$N' = N$$

Calculated in Python

$$\boxed{\hat{y}' = 1.169 + 11.856 * N'}$$

No obvious heteroscedasticity

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Problem 4 Given the data in Table 6.9 (attached), the response variable Y is the number of supervisors, and the predictor variable X is the number of supervised workers. Based on empirical observation, it is hypothesized that the standard deviation of the error term ϵ_i is proportional to x_i:

$$\sigma_i^2 = k^2 x_i^2,\ k > 0$$

- Use the weighted least squares (WLS) method to fit the model. Provide the regression equation.
- Use data transformation method to transform Y to $Y' = Y/X$, and transform X to $X' = 1/X$ (see equations 6.11 and 6.12), and then use the ordinary least squares (OLS) method to regress Y' on X'. Provide the regression equation.
- Compare the results from the above two methods and conclude if the two methods are equivalent. You can compare the residual vs fitted value plot side by side and conclude if they have the same effect in terms of removing heteroscedasticity.

$$W = \begin{bmatrix} \frac{1}{\sigma_1^2} & & \cdots 0 \\ & \ddots & \vdots \\ \vdots & & \\ 0 \cdots & & \frac{1}{\sigma_n^2} \end{bmatrix} = \frac{1}{k^2} \begin{bmatrix} \frac{1}{x_1^2} & & \cdots 0 \\ & \ddots & \vdots \\ \vdots & & \\ 0 \cdots & & \frac{1}{x_n^2} \end{bmatrix}$$

$$\text{Let } W' = \begin{bmatrix} \frac{1}{x_1^2} & & \cdots 0 \\ & \ddots & \vdots \\ \vdots & & \\ 0 \cdots & & \frac{1}{x_n^2} \end{bmatrix} \qquad (\text{i.e., } W = \tfrac{1}{k^2} W')$$

$$\hat{\beta}_{WLS} = (X^T W X)^{-1} (X^T W Y)$$
$$= \left(\frac{1}{k^2} * X^T W' X\right)^{-1} \left(\frac{1}{k^2} * X^T W' Y\right)$$
$$= k^2 \cdot \frac{1}{k^2} \cdot (X^T W' X)^{-1} (X^T W' Y)$$
$$= (X^T W' X)^{-1} (X^T W' Y)$$
$$= \begin{pmatrix} 3.803 \\ 0.121 \end{pmatrix}$$

calculated in Python

$$\boxed{\hat{Y} = 3.803 + 0.121 X}$$

$$Y' = Y/X$$
$$X' = 1/X$$

$$\text{OLS: } \begin{cases} \hat{\beta}_1 = \dfrac{\Sigma(x_i' - \bar{x}')(y_i' - \bar{y}')}{\Sigma(x_i' - \bar{x}')^2} \\ \hat{\beta}_0 = \bar{y}' - \hat{\beta}_1 \bar{x}' \end{cases}$$

calculated in Python

$$\boxed{\hat{Y}' = 0.121 + 3.803 \cdot X'}$$

WLS $\longrightarrow$ $\hat{Y} = 3.803 + 0.121 X$

equivalent

$\hat{Y}/X = 3.803 \cdot 1/X + 0.121$

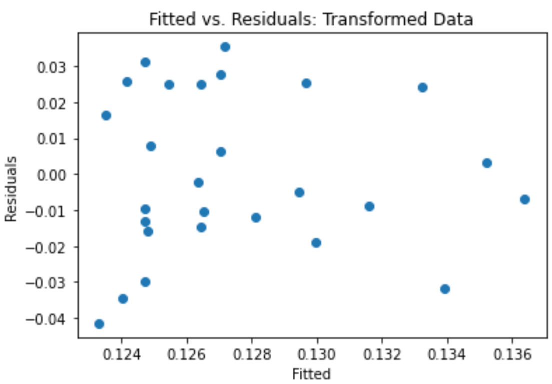
equivalent

$\hat{Y}/X = 0.121 + 3.803 \cdot 1/X$

equivalent

OLS with transformation $\longrightarrow$ $\hat{Y}' = 0.121 + 3.803 X'$

Both methods yield equivalent regression equation. Therefore, both models will yield the same changes in heteroscedasticity. We can get an idea of the new level of heteroscedasticity from the fitted vs. residuals plot of the transformed model, which reflects the results of both methods, since they are equivalent.



$\boxed{\text{The plot does not indicate any obvious heteroscedasticity. Both methods produce this result.}}$

Problem 5 For the data in Problem 4, use OLS without data transformation to fit the model, i.e., directly regress Y on X, and compare the variances of the coefficients $\text{Var}(\hat{\beta}_0)$ and $\text{Var}(\hat{\beta}_1)$ with their counterparts obtained by using WLS, conclude which method yields smaller variances.

OLS:

$$\hat{\sigma}^2 = \frac{SSE}{n-2}$$

$$= 2106.667$$

$$\widehat{var(\hat{\beta}_0)} = \hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\Sigma(x_i - \bar{x})^2}\right) = 407.946$$

$$\widehat{var(\hat{\beta}_1)} = \hat{\sigma}^2 \frac{1}{\Sigma(x_i - \bar{x})^2} = 0.000572$$

WLS:

$MSE_w = 0.00051369$

$$\widehat{var(\hat{\beta}_{WLS})} = MSE_w \, (X^T W X)^{-1}$$

$$= \left(\frac{\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2}{n-p-1}\right) \cdot (X^T W X)^{-1}$$

$$\widehat{var(\hat{\beta}_{WLS,0})} = 20.8826$$

$$\widehat{var(\hat{\beta}_{WLS,1})} = 8.098 \times 10^{-5}$$

$\boxed{\text{WLS yields smaller results.}}$