

Example 1:

Data: IMPORTS

The gross domestic product (GDP) and imports (IMPORTS) for 25 countries are shown in Table IMPORTS Data.

1. provide the scatterplot of IMPORTS versus GDP
2. Fit a regression model to predict IMPORTS
3. Is there any problem?
4. Omit the USA and fit again the model and construct the residual plot for this new regression
5. Is there any difference from the previous model, if so how do they differ?
6. What we should do?

Example 2:

Data: nfl[1]

Data for the players' salaries were provided by the NFL Players' Association. A sample of 198 players is drawn. The salary is the dependent variable.

POSITION: position of the player.

DRAFT: is the round in which the player was drafted

YRSEXP: years of experience.

PLAYED: the number of regular season games played in previous years.

STARTED: the number of regular season games started in the previous years.

CITY POP: size of the city in which the team is domiciled.

- Study the effect of each explanatory variable to the SALARY. Interpret.

POSITION variable is a categorical variable. The meaning of the categories are:

POSITION=1, offensive back(OB)

POSITION=2, defensive back(DB)

POSITION=3, lineman(L)

POSITION=4, kicker/punter(KP)

- Check what is the category which makes an impact on SALARY controlling for the effect of all the other explanatory variable.

We will keep the INDICATOR (OB) as the important indicator variable.

- To choose the best model, use Forward Stepwise Regression and Backward Regression separately on:

- (a) DRAFT (b) YRS EXP (c) PLAYED (d) STARTED
 (e) CITY POP (f) 1/DRAFT (g) OB

- Do they give the same model? which one?
- Fit SALARY on the resulted model?
- Summarize the typical error, the coefficient of determination and so on.
- What kind of pattern the standardized residuals manifests. Why?
- We call this tendency a heteroscedastic error. One way to compensate on for this violation is to consider a transformed version of the response: $\log(y) = \text{LNSALARY}$ (using the base $e=2.718$)

Definition:

Transformations are usually applied to dependent variables.

When the residual plot shows increasing error, use $\ln(y)$. Use also \sqrt{y}

When the residual plot shows decreasing error, use y^2

- The natural log of the y , ($\ln y$) values are less variable than the original y values and may stabilize the variance.

Example:

y	1	2	5	10	50
$\ln y$	0	0.69	1.61	2.30	3.91
\sqrt{y}	1	1.41	2.23	3.16	7.071

- Use also square root of y , \sqrt{y} . The square roots of the values y are less variable than the original values and may stabilize the variance. The square roots are defined only for positive values.

Square roots are defined for count data.

- When deciding whether a transformation has been improved our results, comparing the R^2 is not always efficient. The residual plots can be used to help in this decision. If the pattern suggesting a violation in the original residual plot is no longer present, then the transformed model is preferable. If there is little or no improvement, then try another transformation

- Using now $LNSALARY$, and all the proposed explanatory variables, use the C_p criteria, what is the best model?
- It seems that the best model is the one which uses: $X_1 = 1/DRAFT$, $X_2 = YRSEXP$, $X_3 = STARTED$ and $X_4 = OB$
- Do you see any pattern in the standardized residuals?
- Plot the standardized residuals versus $YRSEXP$. What do you see? can we think of a polynomial effect of $YRSEXP$.

Now three models are proposed, they are:

1. Model 1

$$SALARY = \beta_0 + \beta_1 YRS EXP + \beta_2 STARTED + \beta_3 (1/DRAFT) + \beta_4 OB$$

2. Model 2

$$LNSALARY = \beta_0 + \beta_1 YRS EXP + \beta_2 STARTED + \beta_3 (1/DRAFT) + \beta_4 OB$$

3. Model 3

$$LNSALARY = \beta_0 + \beta_1 YRS EXP + \beta_2 STARTED + \beta_3 (1/DRAFT) + \beta_4 OB + \beta_5 EXP SQR$$

- Calculate $PRESS$ for those models and propose the best one.