Regression Diagnostics

Nikola Baci

2023-04-05

# Introduction

Linear regression is a commonly used statistical technique for modeling the relationship between a dependent variable and one or more independent variables. It has a wide range of applications in various fields, including economics, finance, social sciences, and healthcare. However, the accuracy and reliability of the linear regression model depends on a variety of factors, including the assumptions made about the data and the diagnostic steps taken to assess the model’s performance.

In this paper, we will explore the important diagnostic steps that should be undertaken to ensure that the linear regression model is as good as possible, including checking for linearity, homoscedasticity, normality, and independence of errors, as well as identifying outliers and influential points. By understanding these diagnostic steps, analysts can improve the accuracy and reliability of their linear regression models and make better-informed decisions based on the relationships between variables.

The Organization for Economic Co-operation and Development is an international organization that strives to create policies that improve quality of life. Mainly composed of developed countries, this organization fosters international co-operation and as such it has access to international data.

Using the 2010 Family Database that they have publish, we will perform a linear regression and then diagnose the regression for potential problems and improvements.

# Linear Regression Model

For this model we decided to take a look at if and how

* the proportion percentage of children age 0-2 enrolled in childcare/pre-school
* public expenditures in services for families as a percentage of GDP
* child poverty rate

effect the total fertility rate of the population.

We want to see if social and economical characteristics of a country affect the number of births in society. Below is the result of the linear model.

##   
## Call:  
## lm(formula = tot.fert ~ enr.babies + ben.ink + cpoverty, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.20605 -0.07803 -0.03120 0.02142 0.38328   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.616211 0.168220 9.608 2.86e-07 \*\*\*  
## enr.babies 0.002295 0.002481 0.925 0.3719   
## ben.ink 0.255672 0.087850 2.910 0.0122 \*   
## cpoverty -0.022125 0.010236 -2.161 0.0499 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1517 on 13 degrees of freedom  
## (17 observations deleted due to missingness)  
## Multiple R-squared: 0.727, Adjusted R-squared: 0.664   
## F-statistic: 11.54 on 3 and 13 DF, p-value: 0.0005743

As we can see (from the stars beside) public expenditure and child poverty rate have a somewhat significant affect the total fertility rate of the population, while enrollment in formal education is not significant.

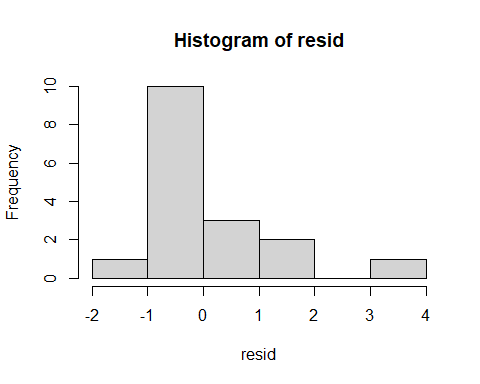
Note that the poverty rate p-value is very close to 5% which is “dangerous” to assume its significance with such slim margins.

This models suggests that: \* the increase of public expenditures by 1% is followed by 0.25 increase on the average number of children per woman \* the increase of child poverty rate by 1% is followed by 0.022 decease on the average number of children per woman \* the adjusted R-square = 0.66 informs us that the model accounts for 66% of all the data points we fed

Ta-dah! We created the model. But is this a good model? Can we improve it? Below are five ways to identify potential issues and fix them.

# 1. Distribution of errors

Here we observe distribution of residuals. We want to see a normal distribution mean zero or very close to zero. This suggests that our predictions are not far from the actual values. This is the distribution:



## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -1.53303 -0.57250 -0.21644 0.04525 0.14766 3.69320

We see from the chart and the table that we are close to what we desire but by no means is this perfect.

We can also perform the Shapiro-Wilk test to see if the above distribution is different to the normal distribution. Null hypothesis is: there is no difference.

##   
## Shapiro-Wilk normality test  
##   
## data: resid  
## W = 0.79652, p-value = 0.001818

The p-value = 0.0018 < 0.05 means that in fact this distribution is not normal, so we reject the null hypothesis. Long story short, this is a problem in our data and we should fix it.

# Heteroscedasticity

OLS models work under the assumption that the variance of errors (residuals) is constant across all predicted values. If the variance is not constant then the model suffers from heteroscedasticity. Below we investigate this graphically and statistically (Breusch-Pegan or BP Test).

Chart

Description automatically generated with low confidence

##   
## studentized Breusch-Pagan test  
##   
## data: model  
## BP = 0.4575, df = 3, p-value = 0.9281

The BP test is a chi-square test with null hypothesis being that there is no heteroscedasticity. In our case the p-value is 0.92 so we do not reject the null hypothesis. Our model has constant error variation.

# No Collinearity

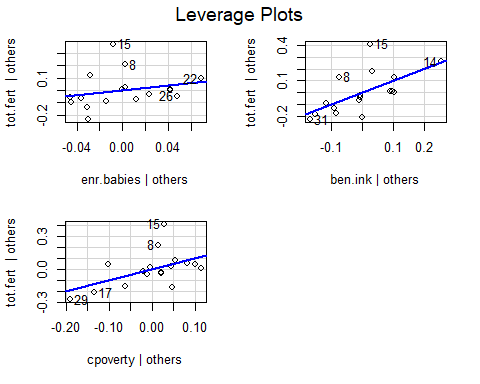
Multicollinearity reduces the precision of the estimated coefficients, which weakens the statistical power of your regression model. Here we will make sure that predictive variables are not co-linear. We will use the Variance Inflation Factor (vif()) from the car package.

## enr.babies ben.ink cpoverty   
## 1.354618 1.652352 1.271690

The vif scores under 4 indicate that the variables are not collinear. If there are above 4 then we should be careful, but if there are above 10 then we should definitely remove one of them

# Outliers

Outliers are data points that are abnormal and effect the models predictive power by pulling it one way or the other. There are two ways to check for outliers, graphically and statistically using Nenferroni OUtlier Test.

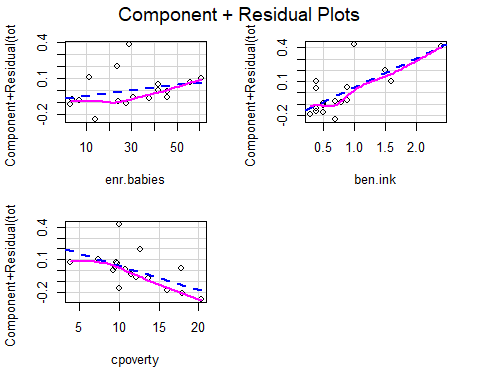


## No Studentized residuals with Bonferroni p < 0.05  
## Largest |rstudent|:  
## rstudent unadjusted p-value Bonferroni p  
## 15 3.693205 0.0030737 0.052253

Both methods suggest that data point 15 is an outlier in our 34 row dataset. We might have to remove this point wich corresponds to Ireland.

# Linearity

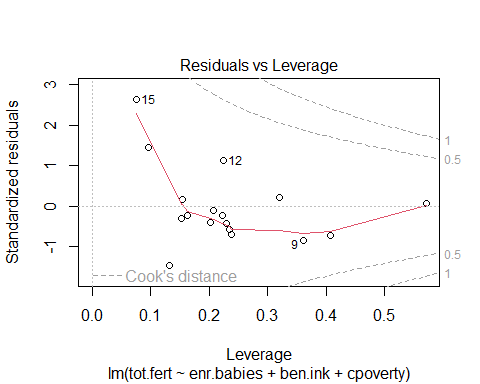
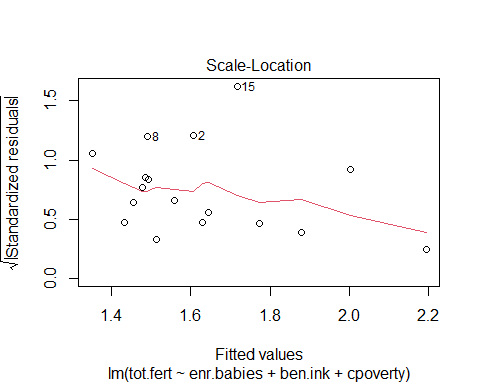
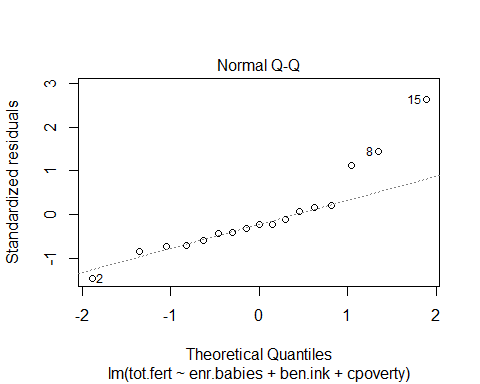
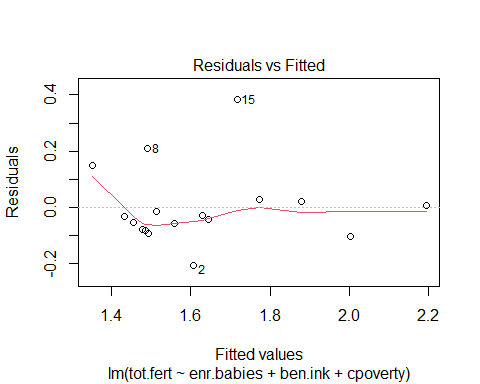
We need to watch out and determine if the independent and dependent variables are linear or not (maybe the relationship is quadratic or other). We use Component Residual Plot to graphically determine this representation.



The blue dotted line shows the expected residuals if the relationship between the predictor and response variable was linear. The pink line shows the actual residuals relationship. The two lines follow the same pattern so we can safely assume that the variables are linear relationship.

# One command for all

A quick way to do what we did in all the previous steps would be to plot the model which gives us similar results to what we got.



# Can we improve the model?

From the diagnostics it seem that the outliers might be throwing off (pulling away) out model, especially outlier number 15 which corresponds to Ireland. Lets remove this outlier, run the model again and see if that improves the predictive power of the model.

##   
## Call:  
## lm(formula = tot.fert ~ enr.babies + ben.ink + cpoverty, data = df\_1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.17259 -0.05240 -0.01757 0.04754 0.22761   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.563255 0.120637 12.958 2.05e-08 \*\*\*  
## enr.babies 0.002686 0.001770 1.518 0.15495   
## ben.ink 0.242054 0.062663 3.863 0.00226 \*\*   
## cpoverty -0.019771 0.007317 -2.702 0.01923 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.108 on 12 degrees of freedom  
## (17 observations deleted due to missingness)  
## Multiple R-squared: 0.8373, Adjusted R-squared: 0.7966   
## F-statistic: 20.58 on 3 and 12 DF, p-value: 5.052e-05

Awesome! We managed to improve our model quite a bit.

First the public expenditures is more significant (2 stars) than before (1 star - see above) when it comes to predicting fertility. While child poverty rate still has the same level of significance, we can see that the p-value = 0.019 is much more acceptable compared to 0.0499 we got previously.

This models suggests that: \* the increase of public expenditures by 1% is followed by 0.24 increase on the average number of children per woman \* the increase of child poverty rate by 1% is followed by 0.019 decease on the average number of children per woman \* the adjusted R-square = 0.79 informs us that the model accounts for 79% of all the data points we fed

# Conclusion

In conclusion, linear regression is a powerful tool for modeling relationships between variables, but it is not without its limitations. To ensure that the model is as good as possible, diagnostic steps must be taken to identify any potential issues with the model’s assumptions or performance. These steps include checking for linearity, homoscedasticity, normality, and independence of errors, as well as examining outliers and influential points. By following these diagnostic steps, analysts can ensure that their linear regression models are as accurate and reliable as possible, providing valuable insights into the relationships between variables and informing important decisions.