Evaluating Regression Assumptions and Residual Diagnostics

Dr Austin R Brown

Kennesaw State University

Introduction

Some of the materials in today's lecture were adapted from those created by Dr. Taasoobshirazi as well as my former professor, Dr. Khalil Shafie (Thank you Drs. S & T!).

Introduction

- In the past few classes, we've learned how to fit simple and multiple linear regression models to various datasets.

- We also briefly discussed the assumptions necessary in order to perform linear regression analysis:
  1. Approximate linear relationship between the response and explanatory variables
  2. The errors have a mean of 0 (i.e., $E[\varepsilon_i] = 0$).
  3. The errors all have a constant variance (i.e., $Var[\varepsilon] = \sigma^2$)
  4. The errors are independent of each other.
  5. Basically, 2, 3, and 4 can be summed up as: $\varepsilon_i \sim N(0, \sigma^2)$, $COV[\varepsilon_i, \varepsilon_j] = 0 i \neq j$.

Introduction

- Okay, well what happens if those assumptions aren't met and how can we evaluate whether they're met or not?

- If the assumptions are only slightly violated, this usually doesn't substantially impact the results. But if there are severe violations, then we can't reasonably rely upon the results insofar as statistical inference is concerned.

- Because R, SAS, SPSS, etc., will still all run the regression analyses regardless if the assumptions are violated or not, we can't use the standard output to evaluate model assumptions.

- There are several tools and methods we have available to us that are quite useful.

Residual Analysis

- Since most of our assumptions have to do with the residuals, it is of value to examine the estimated residuals. What is an estimated residual?

$$\hat{\varepsilon}_i = y_i - \hat{y}_i$$

- Some texts denote the estimated residual as e_i. The estimated variance of the residuals is MSE:

$$\frac{\sum_{i=1}^{n}(e-\bar{e})^2}{n-p} = \frac{\sum_{i=1}^{n} e^2}{n-p} = \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-p} = \frac{SSE}{n-p} = MSE$$

Residual Analysis

- One very common way of assessing the residuals is by standardizing them. The standardized residuals are simply the raw residuals scaled by the square root of MSE (the estimated model standard deviation!).

$$d_i = \frac{e_i}{\sqrt{MSE}}$$

- If $|d_i| > 3$, this could be evidence that value is an outlier.

Residual Analysis

- The one drawback of using the standardized residuals is that using MSE as their variance is actually just an approximation, not their exact variance.

- Before we get into that, recall that the matrix method of calculating our vector of estimated $\hat{\beta}$'s is:

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

- Then, if we multiply the resulting matrix on the left by the design matrix, X, we get our vector of fitted values, $\hat{Y}$:

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

Residual Analysis

- The matrix, $X(X^TX)^{-1}X^T$ is referred to as the "hat matrix" and often denoted as H (my professor used P_x since it is a projection matrix, but we don't need to go too deep down that rabbit hole).

- If we go through all of the linear algebra, it can be shown that the true variance of the e's is:

$$Var[e_i] = \sigma^2(1 - h_{ii})$$

- Where h_{ii} is the ith diagonal element (along the main diagonal) of H and is $0 \leq h_{ii} \leq 1$.

Residual Analysis

- Using MSE to estimate σ^2, the *studentized* residuals are calculated by:

$$r_i = \frac{e_i}{\sqrt{MSE(1 - h_{ii})}}$$

Residual Analysis

- Another way we can use the residuals to help us potentially identify outliers is by taking the difference between an observed value, y_i, and a predicted value obtained from a model containing all the data points except y_i, denoted $\hat{y}_{(i)}$.

- Why it makes sense to examine this value is because if y_i is really that unusual compared to the other values, the model will be overly influenced by that point and thus, $y_i - \hat{y}_{(i)}$ will be large.

- These residuals are called the "prediction error" residuals (or PRESS residuals).

Residual Analysis

- The PRESS residuals can be calculated using:

$$e_{(i)} = \frac{e_i}{1 - h_{ii}}$$

- Why the above is true is a complicated but very cool proof.

- The standardized PRESS residual works out to be the studentized residual.

Residual Analysis

- The point of all of this is for us to be able to identify points which may detrimental to the fit of our model.
- We can categorize outliers into two types: pure leverage and influential.
- A pure leverage point is an extreme value, but it follows the pattern given by the rest of the data.
- An influential point is also an extreme value, but doesn't follow the pattern given by the rest of the data.

Residual Analysis

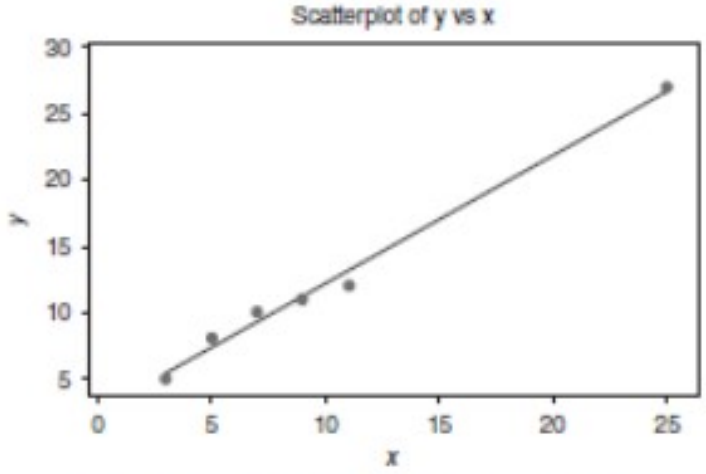


Figure 4.1 Example of a pure leverage point.

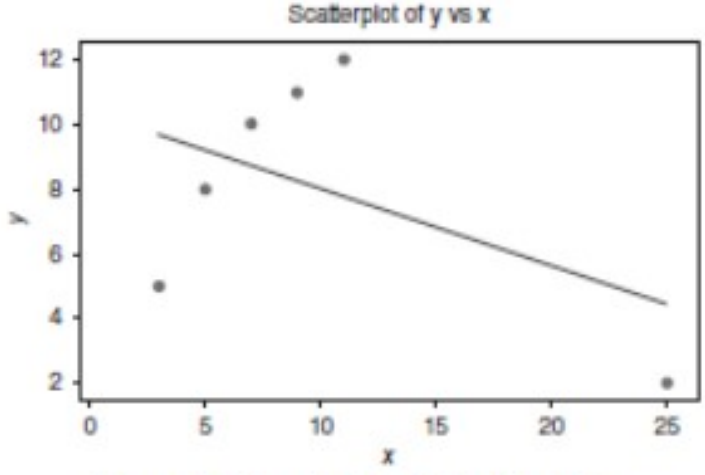


Figure 4.2 Example of an influential point.

Residual Analysis: Leverage Points

- While leverage points may not necessarily seem like they'd cause too many problems, they still can wreak havoc on our parameter estimates, their standard errors, as well as fitted values.

- How do we identify leverage points? Recall that for a single ε_i, its variance is:

$$Var[\varepsilon_i] = \sigma^2(1 - h_{ii})$$

- where h_{ii} is the ith element along the main diagonal of the hat matrix, $H = X(X^TX)^{-1}X^T$.

Residual Analysis: Leverage Points

- We focus on h_{ii} because it can be thought of as a measure of standardized distance.
- More specifically, if a particular value (or vector) of X is far away from its mean (or centroid), h_{ii} is going to be large. A large value of h_{ii} will clearly affect the variance of ε_i, and creates a sort of ripple effect.
- Traditionally, any value of $h_{ii} > 2p/n$ is considered a leverage point.

Residual Analysis: Influential Points

- We can see from the scatterplot example that an influential point, one which has an unusual X and Y value, can pull our fitted line toward it, thusly resulting in poor fit.
- We saw how we can assess the effect of a leverage point, to see if it is influential, through model comparison.
- We have a variety of measures we can use to make this determination on all points, not just those which are clearly leverage.

Residual Analysis: Influential Points - Cook's D

- For the ith observation, Cook's D is defined as:

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})^{\mathbf{T}} \mathbf{X}^{\mathbf{T}} \mathbf{X} (\hat{\beta}_{(i)} - \hat{\beta})}{pMSE}$$

- D_i is a standardized distance our vector $\hat{\beta}$ is from a vector of estimated coefficients generated by omitting point i, defined here as $\hat{\beta}_{(i)}$. If $D_i > 1$, this means the point is likely influential.

Residual Analysis: Influential Points - DFBETAS & DFFITS

- We also rely on two other commonly used measures of influence, $DFBETAS$ and $DFFITS$. Both of these are quite similar, conceptually, to Cook's D.

- For $DFBETAS$, we still fit a full model and a model without the ith observation, and then compare our $\hat{\beta}$'s.
  Computationally:

$$DFBETAS_{j,i} = \frac{\hat{\beta}_j - \hat{\beta}_{j,(i)}}{S^2_{(i)} C_{jj}}$$

- Where C_{jj} is the jth diagonal element in the $(X^T X)^{-1}$ matrix. If $|DFBETAS_{j,i}| > 2/\sqrt{n}$, this indicates a likely influential observation.

Residual Analysis: Influential Points - DFBETAS & DFFITS

- $DFFITS$ is very similarly constructed. Here, similar to the rationale for the PRESS residuals, we are examining the standardized difference between a fitted value, $\hat{y}_i$, and a fitted value from a model excluding the ith observation, $\hat{y}_{(i)}$. Computationally:

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{S^2_{(i)} h_{ii}}}$$

- Traditionally, if $|DFFITS_i| > 2\sqrt{p/n}$, this indicates we may have an influential point.

A Measure of Model Performance: COVRATIO

- Cook's D, $DFFITS$, and $DFBETAS$ are all very useful tools for assessing the effect of individual observations on regression coefficient estimation and in turn, the fitted values.
- What they don't tell us, however, is overall, how precise are our estimates? Remember, precision is the inverse of variance.
- A new measure called $COVRATIO$, solves this problem.

A Measure of Model Performance: COVRATIO

- Let's assume we have a vector of $\hat{\beta}$'s. In general, the variance of this vector is the variance-covariance matrix:

$$Var[\hat{\beta}] = \sigma^2(X^T X)^{-1}$$

- Often, the determinant of a matrix (which is a scalar value) is used as a measure of precision. So we can think of the determinant of the above matrix as a generalized measure of the regression coefficients' variability.

A Measure of Model Performance: COVRATIO

- Thus, the logic behind $COVRATIO$ is an extension of those other fit statistics we've already discussed: What is the generalized variance with and without observation i?

$$COVRATIO_i = \frac{\left|(X_{(i)}^T X_{(i)})^{-1} S_{(i)}^2\right|}{\left|(X^T X)^{-1} MSE\right|}$$

- If $COVRATIO_i > 1 + 3p/n$ or $COVRATIO_i < 1 - 3p/n$, then the ith observation is considered influential.

Treatment of Influential Observations

- So we have a laundry list of ways to see if a point is influential or not. But what do we do when we find one (or more)?

- The answer is: it sorta depends. As a general rule, throwing out observations shouldn't be done unless:
  1. The data point was a typo
  2. There is justification for why/how this observation from this observational unit doesn't represent the analyzed population of interest.

- In academic research, if those two conditions aren't met, then we just proceed with the analysis (so long as the assumptions are reasonably met of course).

Treatment of Influential Observations

- In industry, where prediction is often the primary goal of model building, influential points may be quite detrimental. So what do we do then?

- Sometimes influential points are omitted with the justification that they don't represent typically occurring or observed values.
  - This is part of data preprocessing

- Alternatively, we may have to use a different method which is robust to outliers. With robust regression, we have a different distributional assumption on the residuals (Lapalcian, sometimes also called "Double Exponential"). Gradient Boosting Decision Trees (an ML method) also uses a robust algorithm.

Assumption Checking - Normality

- Now, part of the reason it is a good idea to perform residual analysis is because it gives us an idea as to whether or not our assumptions of normality and constant variance are going to be reasonably met.

- If we have a big number of outliers, this is a good indication that normality and constant variance may be called into question.

- But how do we know this for certain? Let's start with evaluating normality using both a visual method and a testing method.

Assumption Checking - Normality

- To visually assess the assumption of normality, there are several techniques we can use, but most commonly we use a Quantile-Quantile (Q-Q) plot.

- A Q-Q plot is a graphical method used to compare the theoretical quantiles of our residuals (that is, the expected values they would assume if they followed a normal distribution) to the sample quantiles (the actual values we observed).

- If the residuals follow a normal distribution, the points will plot closely to the identity line ($y = x$).

- Deviations from the identity line indicate non-normality.

Assumption Checking - Normality

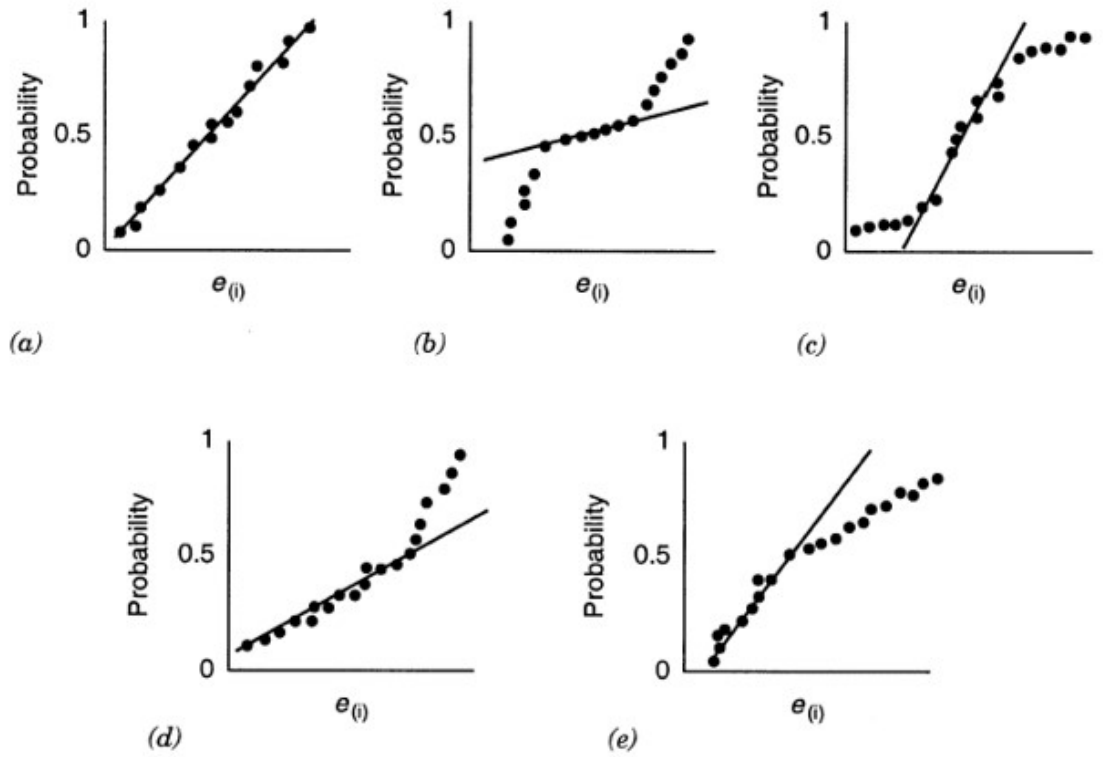


Figure 4.1 Normal probability plots: (a) ideal; (b) heavy-tailed distribution; (c) light-tailed distribution; (d) positive skew; (e) negative skew.

Assumption Checking - Normality

- For testing normality using a hypothesis test, there are lots of options available to us, including:
  1. Kolmogorov-Smirnov
  2. Shapiro-Wilk
  3. Anderson-Darling
  4. Cramer-von Mises
  5. Lilliefors

- Regardless of test, each has the same null and alternative hypothesis:

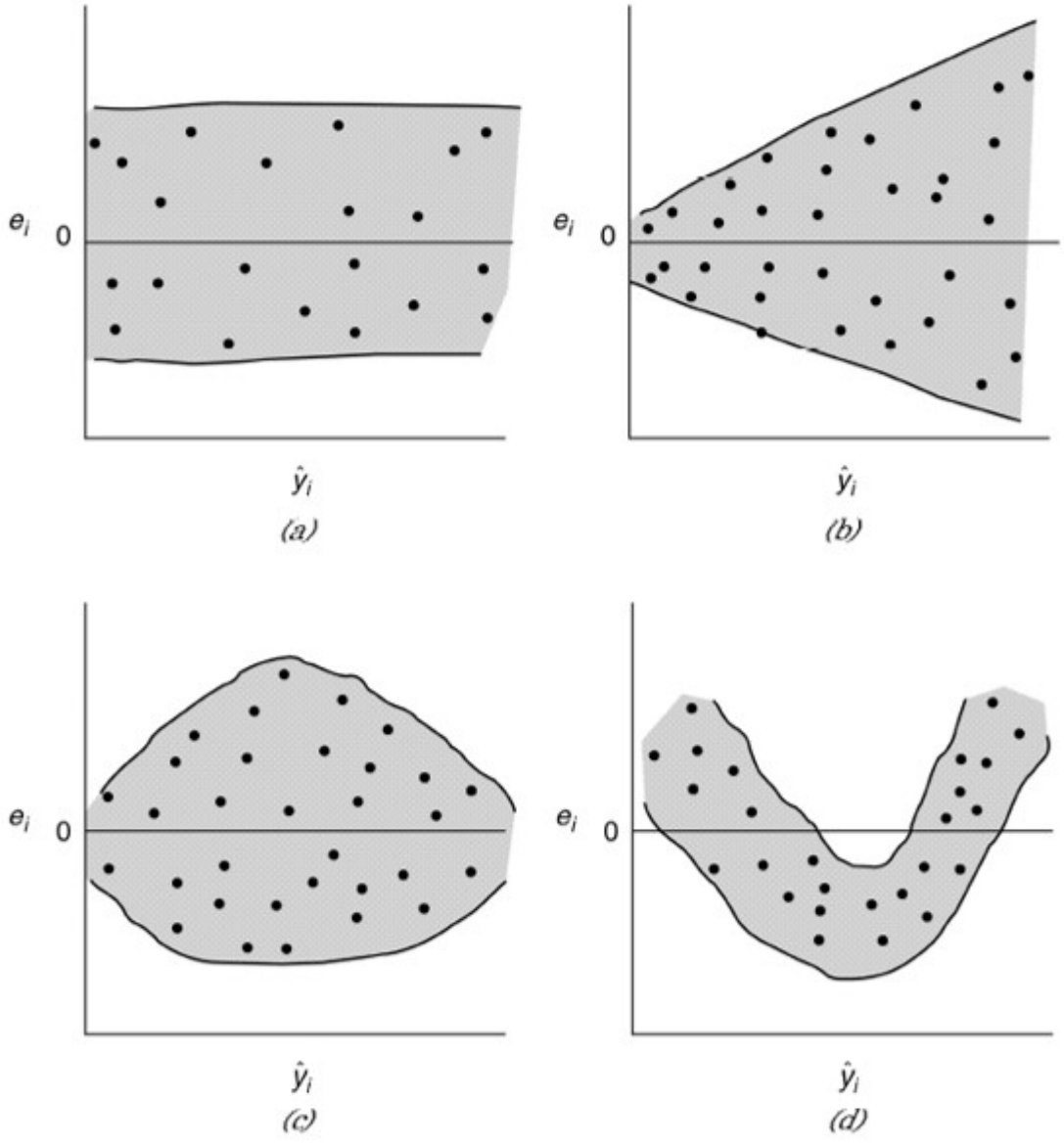
$$H_0 : \text{The data follow a normal distribution}$$

$$H_1 : \text{The data do not follow a normal distribution}$$

Assumption Checking - Constant Variance

- What the "constant variance" assumption is telling us is that as the value of y_i (and consequently, $\hat{y}_i$) changes, its variance stays the same.
- As we have learned so far, model variance is a function of the residuals, $e = y_i - \hat{y}_i$.
- So if we plot the residuals versus the y's or $\hat{y}$'s, we should expect to see no pattern as a pattern would indicate a relationship between the value of y and the variance.

Assumption Checking - Constant Variance



Assumption Checking - Constant Variance

- To test the assumption of constant variance, we use the Breusch-Pagan test. Its hypotheses are:

$$H_0 : \text{The model has constant variance}$$

$$H_1 : \text{The model does not have constant variance}$$

Conclusion

- In sum, we care about residual diagnostics because the residuals:
  1. Affect the accuracy of the estimation of our β's
  2. Affect the accuracy of the prediction of our $\hat{y}$'s

- Further, we care about the assumption of normality and constant variance because if these aren't reasonably met:
  1. Our F and t test statistics, for the omnibus test and for the individual coefficient tests respectively, may not follow the F and t distributions.
     - So we may be making mistakes in our conclusions since we don't know for certain the distributions of the test statistics!!
  2. Statistical power (the probability of rejecting H_0 when we should reject H_0) goes down.