Residuals

BIOS 6611

CU Anschutz

Week 9

1. Linear Regression Assumptions Revisited

2. Residuals

3. Examining Residuals

Linear Regression Assumptions Revisited

Linear Regression Assumptions

Existence: For any fixed value of the variable X, Y is a random variable with a certain probability distribution having finite mean and variance.

Independence: The Y-values are statistically independent of one another.

Linearity: The mean value of Y, $\mu_{Y|X}$, is (approximately) a straight-line function of X.

Homoscedasticity: The variance of Y is the same for any X. That is,

$$\sigma^2_{Y|X} = \sigma^2_{Y|X=1} = \sigma^2_{Y|X=2} = ... = \sigma^2_{Y|X=x}$$

Normal Distribution: For any fixed value of X, Y has a normal distribution. Note this assumption does *not* state that Y is normally distributed, but that $Y|X$ is normally distributed.

The Normality Assumption

Recall the normality assumption:

- Not required to obtain estimate of the regression coefficients (β's)
- Needed to perform statistical tests (t- or F-tests depend on normality assumption)
- Obtain confidence intervals (rely on t- or F-distributions, which assume normality)
- Estimates are asymptotically normal (i.e. assume normal when sample size is large)

Regression Diagnostics

1. Regression diagnostics are tools that can be used to assess the linearity, homoscedasticity, and normality assumptions of linear regression. (Used more often)
2. Regression diagnostics are also used to help identify outliers and influential points in a regression model. (Used less often)

Regression Diagnostics

With linear models, the assumptions of linearity, homoscedasticity, and normality are so intertwined that they often are met or violated as a set.

On the other hand, actions taken to correct violations of one assumption may result in violations of another assumption (e.g., transformations to stabilize the variance can lead to non-linearity).

Prior to fitting a regression model, simple descriptive statistics should be evaluated to look for data errors, potential outliers, and other potential violations of assumptions:

- Univariate descriptive statistics (mean, SD, min, max; frequency tables; histograms).
- Bivariate descriptive statistics (correlations/scatterplots).

Residuals

Errors

In regression analysis we assume that the *unobserved* error terms (ϵ_i):

- are independent (uncorrelated)
- have a mean of zero
- have a common variance of $\sigma^2_{Y|X}$
- follow a normal distribution (required for performing parametric tests of significance and for calculating confidence intervals)

Residuals

Recall that the *observed* residuals ($\hat{e}_i = Y_i - \hat{Y}_i$) are estimates of the *unobserved* error terms.

The $\hat{e}_i$ are not independent random variables (since they must sum to zero). However, if the number of residuals (n) is large relative to the number of independent variables (p), the dependency effect can, for all practical purposes, be ignored in any analysis of the residuals.

Examining the observed residuals (or functions of observed residuals) can be used to evaluate our OLS assumptions. We will introduce five types:

1. Observed
2. Standardized
3. Studentized
4. Press
5. Jackknife

Observed Residual

The **observed residual** is the difference between the observed and predicted values:

$$\hat{e}_i = Y_i - \hat{Y}_i$$

The observed residuals have a

- mean of 0
- variance of $S_e^2 = \frac{1}{n-p-1} \sum_{i=1}^{n} \hat{e}_i^2$ (i.e., the residual mean square error: MSE)

The magnitude of the observed residuals depends on the scaling of Y. This makes defining general rules challenging, so different methods of standardizing the residuals have been developed.

Standardized Residual

The **standardized residual** (also known as the *semi-studentized residual*) is the observed residual divided by $\sqrt{MSE}$:

$$z_i = \frac{\hat{e}_i}{\hat{\sigma}_{Y|X}} = \frac{\hat{e}_i}{\sqrt{MSE}}$$

Standardized residuals have a

- mean of 0
- variance of 1

In other words, the standard residuals follow a standard normal distribution.

Studentized Residual

The **studentized residual** is the observed residual divided by the standard deviation of the *i*th residual [i.e., $Var(\hat{e}_i) = MSE \times (1 - h_i)$]:

$$r_i = \frac{\hat{e}_i}{\sqrt{MSE \times (1 - h_i)}} = \frac{z_i}{\sqrt{(1 - h_i)}} = \frac{(\text{standardized residual})_i}{\sqrt{(1 - h_i)}}$$

h_i is the **leverage**, which is a measure of the importance of the *i*th observation in determining model fit and is also the *i*th diagonal element of the hat matrix ($\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, more later).

Studentized residuals have a

- mean near 0 (but not exactly 0)
- variance of $\frac{1}{n-p-1}\sum_{i=1}^{n} r_i^2$, that is slightly larger than 1

The studentized residuals follow *approximately* a *t*-distribution with $n - p - 1$ degrees of freedom (assuming the assumptions of the errors are satisfied).

Deleted Residual

The **deleted residual** (also known as the *press residual*) is the standardized residual with the current observation deleted in the estimation of the β's (and thus from the calculation of the MSE):

$$d_i = Y_i - \hat{Y}_{i(-i)}$$

where $\hat{Y}_{i(-i)}$ is the predicted value for the *i*th observation from a model fit *without* it (i.e., deleted from that model for estimating the β's and predicted using the resulting estimates).

To avoid fitting *n* different regression models, we can instead use the following formula if we know the leverage:

$$d_i = \frac{\hat{\mathrm{e}}_i}{1 - h_i}$$

This summary is most useful in the calculation of the jackknife residuals. . .

Jackknife Residual

The **jackknife residual** (also known as the *studentized deleted*, *R-student*, *studentized press*, *externally studentized* residual) is the studentized residual with the current observation deleted:

$$r_{(-i)} = r_i \sqrt{\frac{MSE}{MSE_{(-i)}}} = \frac{\hat{e}_i}{\sqrt{MSE_{(-i)}(1-h_i)}} = r_i \sqrt{\frac{(n-p-1)-1}{(n-p-1)-r_i^2}}$$

where $MSE_{(-i)}$ is the residual variance (i.e., MSE) computed with the ith observation deleted.

Jackknife residuals have a

- mean near 0
- variance of $\frac{1}{(n-p-1)-1} \sum_{i=1}^{n} r_{(-i)}^2$ that is slightly greater than 1

Jackknife residuals *exactly* follow a t-distribution with $(n-p-1)-1$ degrees of freedom.

Examining Residuals

Examining Residuals

If the standard regression assumptions are satisfied and approximately the same number of observations are made at all predictor values, then patterns in standardized, studentized, and jackknife residuals will look similar.

As potential problems arise, studentized and jackknife residuals will make suspicious values more obvious and are thus often preferred. However, jackknife residuals are more sensitive and are usually the most preferred residual for regression diagnostics.

The Standard Normal Distribution when $n > 30$

As the error degrees of freedom ($n - p - 1$ for studentized and $n - p - 2$ for jackknife) increase much above 30, the distribution of the residuals can be approximated increasingly by a standard normal distribution.

- This is useful for evaluating the size of observed residuals and for identifying outliers by appealing to properties of a standard normal distribution.
- For example, no more than 5% of the residuals would be expected to exceed 1.96 in absolute value.

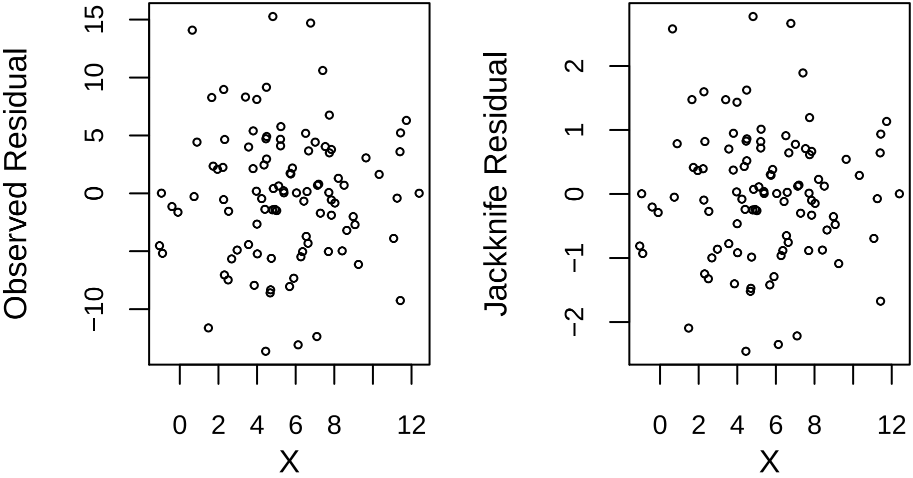
Calculation in R

Different functions exist to calculate and extract our residuals in R, so we introduce a few helpful ones here in the context of the glm function:

```
set.seed(515)
x <- rnorm(n=100, mean=5, sd=3)
y <- rnorm(n=100, mean=3+2*x, sd=5)
glm1 <- glm(y ~ x)

coef(glm1)
## (Intercept)           x
##    2.728488    2.171943

observed_res <- glm1$residuals # the observed residuals
head(observed_res)
##         1         2         3         4         5         6
##  4.426337  3.784119 -4.514892  4.036424 -7.319465 -4.961615

jackknife_res <- rstudent(glm1) # the jackknife residuals
head(jackknife_res)
##          1          2          3          4          5          6
##  0.7860515  0.6655550 -0.8121844  0.7095707 -1.2913341 -0.8755936
```

Calculation in R

```
par(mfrow=c(1,2), mar=c(5.1,4.1,0.5,1.1))
plot(x=x, y=observed_res, xlab='', ylab='Observed Residual', cex=0.5, cex.axis=0.8); mtext('X',side=1,line=2)
plot(x=x, y=jackknife_res, xlab='', ylab='Jackknife Residual', cex=0.5, cex.axis=0.8); mtext('X',side=1,line=2)
```



Quantitative Examination of Residuals (Preview)

Residuals can be used to graphically or formally evaluate:

- skewness (degree of asymmetry of a distribution)
- kurtosis (heaviness of the tails relative to the middle of the distribution)
- normality assumption
- equal variance assumption
- independence assumption (when data are collected in a time sequence)