Diagnostics and Model-Building

Stat 203 Lecture 9

Dr. Janssen

Assumptions for linear regression

Typical assumptions

Recall that the typical assumptions for a linear regression model are:

- Lack of outliers (the same model is appropriate for all observations).
- Linearity: the linear predictor captures the true relationship between μ_i and the explanatory variables.
- Constant variance: The responses y_i have *constant* variance, apart from known weights w_i.
- Independence: the responses y_i are statistically *independent* of each other.
- Distribution: The responses y_i are normally distributed around μ_i.

The process of examining and identifying possible violations of model assumptions is called *diagnostic analysis*.

In what follows, we assume that the important explanatory variables are available for our linear predictor. We'll also explore ways of improving linearity by changing the scale of the covariate or response, or by building new covariates from existing ones (e.g., in an interaction).

Exploring the assumptions

We mention a few of the assumptions here; a fuller treatment is available on pp. 94–97.

Independence

> Question
>
> One of the goals of experimental design/data collection is that the responses y_i are statistically independent. However, dependence can arise because responses share a common source or because data are collected in a hierarchical manner. Can you think of some examples of how this can happen?

Normality

The assumption of normality justifies the use of F- and t-tests. When the number of observations is large and there are no serious outliers, the tests tend to behave well even when the residuals are not normally distributed.

Residuals for Normal Linear Regression Models

Recall that the *raw residuals* are

$$r_i = y_i - \hat{\mu}_i,$$

and that $\text{RSS} = \sum w_i r_i^2$.

It turns out that the variance of r_i is given by

$$\text{var}[r_i] = \sigma^2(1 - h_i)/w_i, \tag{1}$$

where h_i is the *leverage* which y_i has in estimating its own fitted value $\hat{\mu}_i$ (more on this in a minute).

A consequence of Equation 1 is that the residuals do not have constant variance. A modified residual that *does* have constant variances is defined by

$$r_i^* = \frac{\sqrt{w_i}(y_i - \hat{\mu}_i)}{\sqrt{1 - h_i}}, \tag{2}$$

with $\text{var}[r_i^*] = \sigma^2$. After estimating σ^2 with s^2, we define the *standardized residuals* by

$$r_i' = \frac{r_i^*}{s} = \frac{\sqrt{w_i}(y_i - \hat{\mu}_i)}{s\sqrt{1-h_i}}. \tag{3}$$

The standardized residuals defined in Equation 3:

- estimate the standardized distance between the data y_i about the fitted values $\hat{\mu}_i$.
- are approximately standard normal in distribution (more precisely: r_i' follows a t-distribution on $n - p'$ df)

Given a linear model `model`, we can calculate the residuals in `R` using `resid(model)` and the standardized residuals by `rstandard(fit)`.

> Exploration
>
> Calculate some residuals for one of the models we've explored.

Example 0.1. Here's some residuals for the `lungcap` dataset.

```
library(GLMsData); data(lungcap);
lungcap$Smoke <- factor(lungcap$Smoke,
                        levels=c(0, 1),
                        labels=c("Non-smoker","Smoker"))
LC.lm <- lm( FEV ~ Ht + Gender + Smoke, data=lungcap)

resid.raw <- resid( LC.lm )      # The raw residuals
resid.std <- rstandard( LC.lm )  # The standardized residuals
c( Raw=var(resid.raw), Standardized=var(resid.std) )
```

```
      Raw Standardized
0.1812849    1.0027232
```

Leverages

Our next goal is to explore the notion of *leverage*, which, roughly, is the measure of the location of an observation relative to the average location of an observation. This enables us to detect unusual combinations of the explanatory variables, as well as influential observations.

Put another way, it's the distance between the observation and its fitted value.

To define the leverages, we first need to standardize the responses so they have constant variance. Write the **standardized responses** as $z_i = \sqrt{w_i}y_i$. Then $E[z_i] = \nu_i = \sqrt{w_i}\mu_i$ and

$\text{var}[z_i] = \sigma^2$. Then the fitted values $\hat{\nu}_i = \sqrt{w_i}\hat{\mu}_i$ can be considered to be a linear function of the responses z_i. The *hat-values* are defined as the values h_{ij} that relate the responses z_i to the fitted values ν_i, satisfying

$$\hat{\nu}_i = \sum_{j=1}^{n} h_{ij} z_j$$

The hat-value h_{ij} is the coefficient applied to the standardized observation z_j to obtain the standardized fitted value $\hat{\nu}_i$. When $w_i = 1$ for all i,

$$\hat{\nu}_i = \hat{\mu}_i = h_{i1}y_1 + h_{i2}y_2 + \cdots + h_{in}y_n = \sum_{j=1}^{n} h_{ij}y_j.$$

The **leverages** are the diagonal hat-values $h_{ii} =: h_i$. These measure the weight that response y_i (or z_i) receives in computing its own fitted value: $h_i = \sum_{j=1}^{n} h_{ij}^2$. The leverages h_i depend on the values of the explanatory variable and weights, not on the values of the responses. The n leverages satisfy $1/n \leq h_i \leq 1$ and have sum equal to p'.

In the case of simple linear regression without weights,

$$h_i = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\text{SS}_x}$$

.

In general, a small leverage for Observation i indicates that many observations are contributing to the estimation of the fitted value.

Example 0.1. We can calculate leverages in `R` using `hatvalues()`:

```
h <- hatvalues( LC.lm )             # Produce the leverages
sort(h,decreasing=TRUE) [1:2]      # The largest two leverages
```

```
      629       631
0.02207842 0.02034224
```

Why hat-values?

Let's consider the matrix formulation for (unweighted) regression. You may recall that

$$\hat{} = (X^T X)^{-1} X^T \mathbf{y},$$

where X is known as the *model matrix*, and contains columns which are $n \times 1$ vectors of values for x_j.

The fitted values are given by $\hat{} = X \ = H\mathbf{y}$, where

$$H = X(X^T X)^{-1} X^T.$$

We call H the *hat matrix* because it puts the "hat'' on $\mathbf{y}$. The leverages are the diagonal elements of H.