Regression Study Guide

Produced by David Diez

Definitions

n - number of trials (ie, number of points we are working with)
p - number of parameters to be fit, including the intercept (in general, $p \geq 2$)
$\mathbf{X}$ - the *model matrix*, which is n-by-p and generally has its first column as 1's
    -the first column of 1's is for fitting the intercept of the model (mean)
    -row i of $\mathbf{X}$ are the values of the independent variable in trial i
β - vector of parameters, which is p-by-1
Y, y - vector of results of the trials (ie, response variable), which one row for each response
    -dimensions of Y are n-by-1
    -when we write y and not Y, we are talking about actual results, not a random variable
ϵ_i - error in the measurement i, the collection of which we tend to assume is $\epsilon \sim N(0, \sigma^2 I)$
    -that is, $\epsilon_i \sim N(0, \sigma^2)$ and the ϵ_i are independent of each other
$Y_i = X_i\beta + \epsilon_i$ - linear model for a single trial
$Y = \mathbf{X}\beta + \epsilon$ - linear model in matrix form

Computing Estimates of β and ϵ

Finding an estimate of β based on minimizing the sum of squares of the residuals/errors (ie, minimizing $\epsilon^T\epsilon = (y - \mathbf{X}\beta)^T(y - \mathbf{X}\beta)$ over β), we obtain

$$\begin{aligned} \hat{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y \\ \hat{\epsilon} &= y - \mathbf{X}\hat{\beta} = y - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y = (1 - H)y \\ RSS &= (y - \mathbf{X}\hat{\beta})^T(y - \mathbf{X}\hat{\beta}) \end{aligned}$$

Here we have $H = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. Note that $\hat{\beta} \sim N(\beta, \sigma(\mathbf{X}^T\mathbf{X})^{-1})$, where $\sigma(\mathbf{X}^T\mathbf{X})^{-1}$ is a covariance matrix when there is more than one predictor variable.

Remark: In designed experiments where $\mathbf{X}$ is chosen under the design, it is advisable to choose $\mathbf{X}$ such that $\mathbf{X}^T\mathbf{X}$ is diagonal, ensuring the inverse exists theoretically and is computationally feasible. It also allows $cov(\hat{\beta})$ to be diagonal, making the estimates independent.

Example: Comparing models Ω and ω with p parameters and $p-1$ parameters, respectively; suppose β_j is the parameter in Ω but not in ω. Then we can look at

$$t = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{jj}}}$$

where t is from a t distribution with $n-p$ degrees of freedom under H_0.

Comparing Models

It is often difficult to determine the number of parameters that should be fit. However, we can often compare models using each model's RSS. Since we typically favor smaller models over larger ones, our null hypothesis is typically favoring the small model and rejecting the small model in favor of the larger one if there is a substantial amount of information gained (ie, the RSS of the larger model is much lower).

Suppose we test a model (called ω) against a larger model (Ω) where the parameters in ω are all in Ω (and $\omega \neq \Omega$), then we compare RSS_ω to RSS_Ω (ie, we compute estimators of both models and the resulting RSS of both models). We know that models with more parameters have a smaller RSS than the RSS of a subset of those parameters – $RSS_\Omega \leq RSS_\omega$. The maximum likelihood ratio test statistic is given by

$$F = \frac{(RSS_\omega - RSS_\Omega)/(df_\omega - df_\Omega)}{RSS_\Omega/df_\Omega} = \frac{(RSS_\omega - RSS_\Omega)/(p-q)}{RSS_\Omega/(n-p)}$$

where the number of parameters in model ω is q and in Ω is p. df_ω and df_Ω are $n-q$ and $n-p$, respectively. We reject the null hypothesis if F is large enough. More specifically, if $F > F_{p-q,n-p}$ (we compare F to an F-distribution). Let us make a few definitions for convenience:

$SS_E = RSS_\Omega$ - the sum of squares (SS) of the error under Ω
$SS_T = \sum (y_i - \bar{y})^2$ - the total SS
$SS_R = SS_T - SS_E$ - the regression SS (the SS avoided by the larger model)
$MS_i = MS_i/(df$ attributable to $i)$ - mean SS of i, where $i \in \{E, T, R\}$

Remark: $R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$ measures how much of the variation is explained by the model. However, the more parameters we add, the closer the fit (R^2 increases with the number of parameters), so we can also consider the *adjusted* R^2, $R^2_{adj} = 1 - \frac{SS_E/(n-p)}{SS_T/(n-1)}$.

Remark: When testing the removal of a β_i from Ω, the test in the example and the other method of using F are actually equivalent. They produce the same p-value and $F = t^2$.

Prediction

After finding $\hat{\beta}$, we may wish to attempt to predict Y_0 (a number) for (a column vector) x_0. If so, then we start with the following:

$$\hat{y_0} = EY_0 \sim N\left(x_0^T\hat{\beta}, var\left(x_0^T\hat{\beta}\right)\right) = N\left(x_0^T\hat{\beta}, x_0^T var(\hat{\beta})x_0\right) = N\left(x_0^T\hat{\beta}, \sigma^2 x_0^T(\mathbf{X}^T\mathbf{X})^{-1}x_0\right)$$

We then let $\sigma = \hat{\sigma}$ for our estimate of the variance. Here we have obtained a confidence interval for the mean at x_0. To compute the confidence interval for prediction, we must add a variance component of ϵ onto our prediction of a future measurement, resulting in

$$Y_0 \sim N\left(x_0^T\hat{\beta}, \sigma^2\left(1 + x_0^T(\mathbf{X}^T\mathbf{X})^{-1}x_0\right)\right)$$

where we here again can put in our estimate of σ to obtain a value.

Leverage and Studentized Residuals

Computing $var(\hat{\epsilon})$, we find it is not zero, but it is instead $\sigma^2(1-H)$. Hence, the $\hat{\epsilon}_i$ do not necessarily have constant variance, even if the ϵ_i do. Letting $H = (h_{ij})$, we say h_{ii} is the *leverage* of ϵ_i and the *studentized residuals* (`rstandard` in R) are given by

$$r_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$$

If the model and assumptions hold, the r_i will have mean 0 and variance 1.

Diagnostics

If the model is correct, the residuals should be appear as normally distributed white noise. That is, they should be independent, identically distributed, and normal. Below are three typical checks for these assumptions. Due to leverage issues, it is typically better to use the studentized residuals mentioned above in place of the residuals.

Dependence between errors - Plotting residuals according to their index (the order the data was collected), check for violation of independence. We are looking for correlation (positive or negative) between successive residuals. The auto-correlation function may be useful here for those familiar with its properties.

Heteroscedasticity - We want to check that the variance is constant with respect to fitted values. For example, check that for larger y_i, there are not larger ϵ_i. This is checked by plotting the residuals against their fitted values, looking for constant variance. If there is non-constant variance, this would violate our assumption of identically distributed.

Normality - Plotting the (ordered) residuals against normal quantiles in a Q-Q plot, if the line is approximately straight, then we cannot reject the normality claim. To get a feel for how straight a plot should be, it is a good idea to generate pseudo-random normal samples of the same size and look at those Q-Q plots for comparison.

Outliers and Influential Points

We may want to know what the model would look like if we dropped a point from the analysis. For example, if we have an outlier, would dropping it make a big difference in results? Or we may ask, is this point very influential? We can make a guess at this by looking at jackknife residuals and Cook's distance. *Jackknife residuals* are defined as

$$t_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}}$$

where $\sigma_{(i)}$ is the estimated standard error of the residuals when we drop case i (called `rstudent` in R). There is a formula that allows us to consider the jackknife residuals without computing n different models, as shown below:

$$t_i = r_i\sqrt{\frac{n-p-1}{n-p-r_i^2}}$$

If case i is not an outlier, then t_i follows a t distribution with $n-p-1$ degrees of freedom. Note that if we see an outlier, and decide to check it, we should use a Bonferroni correction as if we were checking all of the residuals – we picked which of the n cases that was the most extreme, so we were essentially looking at all of the data and thinking to only check the most extreme case(s).

Cook's distance, D_i, suggests how influential a point is, which does not directly translate to if a point is an outlier.

$$D_i = \frac{r_i^2 h_{ii}}{p(1-h_{ii})}$$

For Cook's distance to be large, we can see that either the leverage (h_{ii}) must be large or the studentized residual must be large (r_i), where the latter suggests the point may be an outlier. If we plot all of the Cook's distances, we look for cases that are substantially higher than others or appear to be outliers from the others (outliers in comparing their Cook's distance, **not** the original plot).