Regression analysis Module 22: Residual deviance intuition and practice problems

(The attached PDF file has better formatting.)

Intuition: Residual deviance

[This dialogue should help you understand the rationale for the residual deviance. The final exam problems on residual deviance use the formula at the end of the dialogue.]

An actuary builds a generalized linear model to relate personal auto claim frequency to explanatory variables, using maximum likelihood to derive the rate relativities (the beta coefficients for each explanatory variable). The explanatory variables are sex of the driver (male vs female) and territory of the car (urban vs rural).

Each driver has an observed claim frequency. Each driver is one exposure, so the observed claim frequencies are integers: 0, 1, 2, 3, …. Most drivers have zero claims; a few have one claim; very few have more claims.

The sample has many points, such as 10,000 male-urban drivers, 15,000 female-urban drivers, 8,000 male-rural drivers, and 5,000 female-rural drivers.

Each driver has an expected claim frequencies. For this exercise, we assume the classes are homogeneous: each driver in the class has the same expected claim frequency. The generalized linear model solves for the expected claim frequencies by class.

Jacob: Why do we use generalized linear models? The best estimate of the expected value is the mean of the observed values. The estimated claim frequency for male-urban drivers is the average observed value of the 10,000 drivers.

Rachel: In a 2 × 2 class system (two dimensions with two levels in each dimension), using average observed values is reasonable, though not necessarily optimal. To see why we need generalized linear models, suppose the class system has ten dimensions with an average of five levels in each dimension. For example, territory might have 20 levels, age might have five levels, type of car might have ten levels, and so forth. The number of possible classes is 5^{10} = 9,765,625. If the sample has 100,000 cars, most classes have no data points, and the rest have one or two data points each. The observed averages by class are either missing (if the class has no data points) or claim frequencies of 0%, 50%, 100%, if the class has only one or two data points. The GLM determines class relativities, such as male vs female, which each have 50,000 observations, so the estimated relativities are credible. Even for territory with 20 levels, each level has about 100,000 / 20 = 5,000 observations, so the estimated relativities are credible.

Jacob: Actuaries have always estimated class relativities, which they combine to get class rates, and they use credibility procedures when data are sparse. What do generalized linear models add?

Rachel: Traditional actuarial class ratemaking has two faults which GLMs solve. First, the traditional actuarial methods examine each class separately: all male vs all female; all youthful vs all adult vs all retired; all single vs all married and divorced; and so forth. GLMs look at all class dimensions simultaneously. Second, the traditional actuarial methods use *ad hoc* credibility adjustments. The GLM maximizes the likelihood, so it gives the optimal class relativities.

Jacob: Why is the simultaneous perspective important?

Rachel: Fox distinguishes marginal from partial effects. Suppose a state has one city with a large university and exciting night life surrounded by suburbs and rural areas. Most young people live in the city; most adults live in the suburbs and rural areas. If class relativities are determined for each class dimension separately, urban drivers may have twice the average claim frequency as suburban and rural drivers, and youthful drives may have twice the average claim frequency as adult drivers. These are marginal effects, and they overlap. A youthful urban driver does not have 2 × 2 = 4 times the average claim frequency as an adult rural driver.

Simultaneous class relativity models, like multiple regression, give partial effects: the class relativity for urban vs rural drivers holding age constant and the class relativity for age holding territory constant. GLMs give the partial effects, which might be 1.8 for urban/rural and 1.2 for youthful/adult. Fox explains this in the chapter on multiple regression. The β for territory alone might be 2.0 for urban/rural, and the β for age along might be 2.0 for youthful/adult. These are marginal β's; we want the partial effects of each explanatory variable holding other explanatory variables constant.

Jacob: You explained why we use simultaneous class relativity procedures. But why do you say the actuarial credibility procedures are *ad hoc*?

Rachel: Classical credibility has little theoretical support. GLMs adjust the class relativities so the observed values are maximized. They determine regression coefficients so that the likelihood of obtaining the observed values is maximized.

Jacob: How do we maximize the likelihood?

Rachel: The likelihood depends on the probability density function of the claim frequencies in each class. To simplify the explanation, assume each class has one driver with an observed claim frequency from 50 years of driving. We want to estimate the expected claim frequencies by class. After explaining this scenario, we assume each class has many drivers with one year of experience each.

Suppose the expected claim frequency for the driver is μ and the observed claim frequency is y. If the driver has a Poisson distribution for claim frequency, the likelihood of observing a value y when the mean is μ is $f(\mu|y) = \mu^y e^{-\mu} / y!$

To maximize the likelihood, set the partial derivative with respect to μ equal to zero.

- The partial derivative with respect to μ is $(y \times \mu^{y-1} e^{-\mu} - \mu^y e^{-\mu}) / y!$.
- For simplicity, we drop the subscript on y. Don't forget that y is a scalar.
- Setting this to zero gives

$$(y \times \mu^{y-1} e^{-\mu} - \mu^y e^{-\mu}) / y! = 0$$
$$\Rightarrow y \times \mu^{y-1} e^{-\mu} = \mu^y e^{-\mu}$$
$$\Rightarrow y = \mu$$

To maximize the likelihood, set the expected claim frequency equal to the observed claim frequency.

We can do the same by maximizing the loglikelihood. The loglikelihood for the Poisson distribution is

$$ln(f(\mu|y) = ln(\mu^y e^{-\mu} / y!) = y\, ln(\mu) - \mu - ln(y!).$$

- The partial derivative with respect to μ is $y / \mu - 1$.
- Setting this to zero gives $y / \mu - 1 = 0 \Rightarrow y = \mu$.

If we examine each class separately, the best estimate of the expected claim frequency is the mean of the observed claim frequencies. GLMs are useful if we have a model for claim frequency.

Link functions and additive vs multiplicative models.

Jacob: Multiple regression uses linear models. If β_1 is the claim frequency relativity for males vs females and β_2 is the claim frequency relativity for urban vs rural drivers, the claim frequency relativity for male-urban drivers vs female-rural drivers is $\beta_1 + \beta_2$. GLMs also use linear models. But actuaries use multiplicative models for class relativities: the claim frequency relativity for male-urban drivers vs female-rural drivers is $\beta_1 \times \beta_2$.

Rachel: Fox discusses this issue for multiple regression and for generalized linear models.

For multiple regression, we transform the data. If $Y = \alpha \times X_1^{\beta} \times X_2^{\beta'} \times \epsilon$ then

$$ln(Y) = ln(\alpha \times X_1^{\beta} \times X_2^{\beta'} \times \epsilon) = ln(\alpha) + \beta \times ln(X_1) + \beta' \times ln(X_2) + ln(\epsilon)$$

The multiplicative model becomes a linear model by taking logarithms of the response variable.

GLMs do a similar transformation with a log-link function: the logarithm of the response variable is a linear function of the explanatory variables.

Take heed: The log-link function differs from the logarithmic transformation of the response variable two ways:

- The log-link function takes logarithms of the fitted values, not of the observed values.
- The log-link function keeps the conditional distribution of the response variable.

Jacob: How does maximum likelihood estimate the parameters in a multiplicative model?

Rachel: Let us call the parameters α for the base rate, β_1 for male/female, and β_2 for urban/rural. For example, α might be 4%, $\beta_1 = 1.5$, and $\beta_2 = 2.5$. We don't know the values of these parameters until we fit the GLM.

The fitted and observed claim frequencies are shown in the table below.

Class	*Observed*	*Fitted*
Male-urban	16%	$\alpha \times \beta_1 \times \beta_2$
Female-urban	10%	$\alpha \times \beta_2$
Male-rural	8%	$\alpha \times \beta_1$
Female-rural	4%	α

For each class, we determine the likelihood function. For female-rural drivers, the likelihood function is

$$likelihood(\alpha) = \alpha^{4\%} e^{-\alpha} / 4\%!$$

(The meaning of 4%! is not a problem. The 4% is actually a sample of drivers, such as 100 drivers, of whom two have one claim, one has two claims, and the rest have no claims. 0!, 1!, and 2! are well-defined.)

Jacob: Do we use the likelihood or the loglikelihood?

Rachel: Our goal is to maximize the likelihood. The loglikelihood = *ln*(likelihood) is a monotonic function of the likelihood, so maximizing the loglikelihood maximizes the likelihood. It is easier to work with the loglikelihood.

For each class, we determine the log-likelihood function.

- For female-rural drivers, the log-likelihood is $4\% \times ln(\alpha) - \alpha - ln(4\%!)$
- For male-urban drivers, the log-likelihood is $16\% \times ln(\alpha \times \beta_1 \times \beta_2) - (\alpha \times \beta_1 \times \beta_2) - ln(16\%!)$
- For male-rural drivers, the log-likelihood is $8\% \times ln(\alpha \times \beta_1) - (\alpha \times \beta_1) - ln(8\%!)$
- For female-urban drivers, the log-likelihood is $10\% \times ln(\alpha \times \beta_2) - (\alpha \times \beta_2) - ln(10\%!)$

Jacob: How do we solve for the parameters?

Rachel: The likelihood for the entire class system is the product of the likelihoods for each class, so the loglikelihood for the entire class system is the sum of the loglikelihoods for each class. To maximize the total loglikelihood, we set the partial derivatives with respect to each parameter equal to zero. This give three equations in three unknowns (the three parameters). They are not linear equations, so they are hard to solve

by pencil and paper. With statistical software packages, they are easily solved. Excel uses *SOLVER* to find the parameters that set the partial derivatives to zero.

Illustration: Using SOLVER for generalized linear models

In a workbook, name three cells (such as A1, B1, and C1) alpha, beta1, and beta2. On scratch paper, write out the loglikelihoods for each class (shown above) and the sum of these loglikelihoods. Take the partial derivatives of the total loglikelihood with respect to each variable. These partial derivatives are simple expressions. The partial derivative with respect to μ is $y / \mu - 1$; the terms with factorials drop out entirely. If μ is a function of α, β_1, and β_2, the partial derivative is a bit more complex.

We want to set all three partial derivatives equal to zero. Put the formulas for the three partial derivatives in cells A3, B3, and C3. Set A4 = A3^2, B4 = B3^2, and C4 = C3^2. Set D4 = A4 × B4 × C4. Use solver to minimize Cell D4 by choosing values for Cells A1, B1, and C1. This exercise is simple enough that solver will give a solution quickly.

Jacob: Fox discusses statistical inference for regression models. A regression model may give an estimated value for α or β, but unless we know the *p*-value, we don't know if the estimate is significant. For example, a regression model solving for class relativities may give male/female = 1.500 and urban/rural = 2.500. Fox shows how to use the *F*-ratio to test the significance of sex (male/female) and territory (urban/rural). Even though the urban/rural class relativity is larger in this example than the male/female relativity, the *p*-value for the male/female relativity may be low (say 5%) and significant and the *p*-value for the urban/rural relativity may be high (say 40%) and not significant. Statistical inference also give confidence intervals. The β's, *t*-values, and degrees of freedom give confidence intervals. If the 95% confidence interval for the male/female relativity is (1.400, 1.600), and the current rate relativity is 1.350, we might file for a rate level change. If the 95% confidence interval for the male/female relativity is (0.800, 2.200), and the current rate relativity is 1.350, we might wait before filing for a rate level change. How do we infer significance of generalized linear models?

Rachel: Statistical inference is related to the goodness-of-fit test. Regression uses least squares estimation: the selected parameters are those that minimize the squared error. The *F*-ratio examines the regression sum of squares (adjusted for degrees of freedom) divided by the residual sum of squares (adjusted for degrees of freedom). The regression analysis maximizes the portion of the total sum of squares explained by the regression parameters. That is, it maximizes the R^2, or the ratio of the regression sum of squares to the total sum of squares, which is the same as maximizing the ratio of the regression sum of squares to the residual sum of squares.

GLMs use maximum likelihood estimation, not ordinary least squares estimation. Statistical inference depends on the ratio of the total likelihood for the GLM compared to the likelihood for a larger model.

Jacob: What do you mean by a larger model? Do you mean a model whose likelihood is 100%?

Rachel: The likelihood is never 100% if the response variable is stochastic. Even if all the parameters are correct and the response variable has the assumed relation to the explanatory variables, random fluctuations cause the observed values to differ from the fitted values. The model with the highest likelihood is a saturated model. The saturated model has all the fitted values exactly equal to the observed values.

Jacob: If the fitted values equal the observed values, isn't the likelihood 100%?

Rachel: Consider the example with male/female and urban/rural class relativities. Suppose the observed claim frequencies by class are as shown in the table below.

Class	*Observed*	*Fitted*
Male-urban	18%	$\alpha \times \beta_1 \times \beta_2$
Female-urban	9%	$\alpha \times \beta_2$

Male-rural	6%	$\alpha \times \beta_1$
Female-rural	3%	α

We can fit these observed values exactly with $\alpha = 3\%$, $\beta_1 = 2.00$, and $\beta_2 = 3.00$. The likelihoods by class are

- Male-urban: $0.18^{0.18} \times e^{-0.18} / 0.18! = 0.61345 / 0.18!$
- Female-urban: $0.09^{0.09} \times e^{-0.09} / 0.09! = 0.73586 / 0.09!$
- Male-rural: $0.06^{0.06} \times e^{-0.06} / 0.06! = 0.79548 / 0.06!$
- Female-rural: $0.03^{0.03} \times e^{-0.03} / 0.03! = 0.87354 / 0.03!$

Let us ignore the denominators with the factorials. The 3% for female-rural really means 300 drivers of whom three had claims. The 0.03! is really 0! × 0! × … × 0! × 1! × 1! × 1! = 1. In a few cases, there might be a 2 in the product (if one driver had two claims) or a 6 (if one driver had 3 claims). The total likelihood is

$$0.61345 \times 0.73586 \times 0.79548 \times 0.87354 = 0.31368.$$

Jacob: Is this a reasonable size for the total likelihood of a saturated model?

Rachel: Most generalized linear models for class rate relativities 100,000 (or more) exposures. If the sample has 25,000 cars in each class, the total likelihood is

$$0.61345^{25,000} \times 0.73586^{25,000} \times 0.79548^{25,000} \times 0.87354^{25,000} = 0.000000000000000.$$

In large samples, total likelihoods are extremely small figures.

Jacob: How do we work with such small figures?

Rachel: The loglikelihood of an extremely small positive figure is a large negative figure.

Illustration: If the likelihood is e^{-N}, the loglikelihood is –N. The likelihood ratio test uses –2 × the loglikelihood, and the figures are about the same size as the degrees of freedom.

Take heed: Know the section on hypothesis tests, analysis of deviance, and the residual deviance on page 408. Fox discusses several other types of residuals. The final exam tests the residual deviance, which is used for hypothesis testing, not the other residuals of GLMs.

Equation 15.20 on page 408 gives the formula for the residual deviance. Let us parse this formula.

L is the likelihood function, so $log_e L$ is the loglikelihood function. As the example above shows, the likelihood is extremely small for large sample. You might think that the residual deviance is very small for large samples, since it is the difference of two very small numbers. That is not correct; let us see why.

In the example above, the total likelihood for the saturated model with one driver in each class is about 30%. The total likelihood for the model under review is lower; it might be 25%.

The difference of the loglikelihoods is the logarithm of the ratio of the likelihoods. The ratio of these likelihoods is 0.30 / 0.25 = 1.200. If there are 1,000 cars in each class (a relatively small sample), the ratio of the likelihoods is $1.200^{1,000} = 1.51791$ E+79.

Jacob: How can this be? We compare two minuscle numbers. How do we get a large result?

Rachel: The likelihood in a large sample is an extremely small number, close to zero. As the likelihood approaches zero, the loglikelihood approaches negative infinity. The negative of the loglikelihood approaches infinity. We are dealing with large numbers in large samples, not small numbers.

Illustration: Suppose the likelihood for the saturated model is 1×10^{-20} and the likelihood for the model under review is 1×10^{-25}.

- The loglikelihoods are –46.05 and –57.56.
- Twice the difference in the loglikelihoods is 23.03.

** Exercise 22.1: Residual deviance

The likelihood for the model being tested is 8% and the likelihood for the saturated model is 10%.

1. What is the loglikelihood for the model being tested?
2. What is the loglikelihood for the saturated model?
3. What is the residual deviance for the model being tested?

Part A: $ln(0.08) = -2.52573$

Part B: $ln(0.10) = -2.30259$

Part C: $2 \times [\, ln(0.10) - ln(0.08) \,] = 0.44629$