Regression Analysis (II) Assignment.
Fall, 2020

1. Suppose that tow different sets of treatments are of interest. Let y_{ijk} be the kth observation level i of the first treatment type and level j of the second treatment type. The two way analysis-of-variance model is

$$y_{ijk} = \mu + \tau_i + \gamma_j + (\tau\gamma)_{ij} + \epsilon_{ijk}, i = 1, \ldots, a, \quad j = 1, \ldots, b, \quad k = 1, \ldots, n$$

where τ_i is the effect of level i of the first treatment type, γ_j is the effect of level j of the second treatment type, $(\tau\gamma)_{ij}$ is an interaction effect between the two treatment types, and ϵ_{ijk} is an $i.i.d.N(0, \sigma^2)$. random errors.

(a)For the case $a = b = n = 2$, write down a regression model that corresponds to the two-way analysis of variance.

(b) What are the response vector y and design matrix X for the regression model?

(c) Discuss how the regression model could be used to test the hypotheses (i) $H_0 : \tau_1 = \tau_2 = 0$(treatment type 1 means are equal), (ii) $H_0 : \gamma_1 = \gamma_2 = 0$ (treatment tye 2 means are equal), and (iii) $H_0 : (\tau\gamma)_{11} = (\tau\gamma)_{12} = (\tau\gamma)_{21} = (\tau\gamma)_{22} = 0$ (no interaction between treatment types).

2. Consider the multiple linear regression model

$$y = X\beta + \epsilon,$$

where X is $n \times p$ design matrix, β is $p \times 1$ coefficient vector and $\epsilon \sim MVN(0, \sigma^2 I)$. We assume that all regressors x_j's ,$j = 1, \ldots, p$ and the response variable y are centered and scaled to unit length. (You may consider intercept model.) The variance inflation factor (VIF_j) of the regressor x_j is defined to be $Var(\hat{\beta}_j)/\sigma^2$. Show that

$$VIF_j = \frac{1}{1 - R_j^2},$$

where R_j^2 is the coefficient of determination from the regression of x_j on the remaining $p - 1$ regressor variables.

4. Let $\hat{\beta}$ be the least squares estimator of β and $\hat{\beta}_R = (X'X + \lambda I)^{-1}X'y$ be the ridge estimator of β.

(a) Show that there exist λ such that for any constant vector a,

$$MSE(a'\hat{\beta}_R) < MSE(a'\hat{\beta}).$$

(b) Show that ridge estimator $\hat{\beta}_R$ minimizes

$$(y - Xb)'(y - Xb),$$

subject to the constraint $b'b \leq d^2$, for some d.

4. Assume that there are k regressors and

$$y = X\beta + \epsilon$$

is the true regression model, where X is $n \times k + 1$ design matrix. Let $X = (X_p, X_r)$, where X_p is $n \times p + 1$ matrix and X_r is $n \times r$ matrix and $r = k - p$. Note that X_r consists of the last r columns of X. The the true regression model can be expressed as $y = X_p\beta_p + X_r\beta_r + \epsilon$. Let $\hat{\beta}^*$ be the LSE of β from fitting the true regression model and $\hat{\beta}_p$ be LSE of β_p from fitting $y = X_p\beta_p + \epsilon$. If $\hat{\beta}_p^*$ is the LSE of β_p from $\hat{\beta}^*$, then show that $MSE(\hat{\beta}_p^*) - MSE(\hat{\beta}_p)$ is positive semi-definite under some condition. Specify the required condition.

5. Consider the nonlinear regression model

$$y_i = f(x_i, \theta) + \epsilon_i, i = 1, \ldots, n,$$

where $x_i = (x_{i1}, \ldots, x_{ik})'$ and $\theta = (\theta_1, \ldots, \theta_p)'$. Discuss the Gauss-Newton iteration method for computing the least squares estimate $\hat{\theta}$ of θ.

6. In fitting nonlinear regression model using the above Gauss-Newton iteration method, the choice of good starting values is important.

(a) For the Michaelis-Menten model

$$E(y) = f(x, \theta_1, \theta_2) = \frac{\theta_1 x}{x + \theta_2}$$

discuss how to find a reasonable starting values of θ_1, θ_2.
(b)Consider the nonlinear regression model

$$y = \theta_1 - \theta_2 exp(-\theta_3 x) + \epsilon.$$

This is called the Mitcherlich equation. Discuss how you would obtain reasonalbe starting values of the parameters θ_1, θ_2, and θ_3.

7. Assume that $y_i \sim indepBer(\pi_i), i = 1, \ldots, n$, where π_i may depend on the i-th level of predictor variable, $x_i = (1, x_{i1}, \ldots, x_{ip-1})'$. We want to fit the logistic regression model to the data.
(a) Write down the model
(b) To compute the MLE of parameters, we will use either Newton-Raphson method or Fisher scoring method. Explain these methods.
(c) When there is only one predictor variable, i.e. $p = 1$, express the suitable hypotheses for the significance test and obtain the likelihood ratio test.
(d) Explain about the goodness-of-fit test for the logistic model.

8. We have binary data with one predictor variable. At each level x_i of the predictor variable, we have n_i repeated observations.
(a) Write down the logistic regression model.
(b) After 10 iterations using Fisher scoring method, we got the following results

$$\hat{\beta}_0 = 60.717, \hat{\beta}_1 = 34.270, D = 11.23$$
$$I^{-1} = \begin{bmatrix} 26.802 & 15.061 \\ 15.061 & 8.469 \end{bmatrix}$$

What are the standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$? Do the goodness-of-fit test for the logistic regression model.

9. Suppose that we have regression data $(x_i, y_i), i = 1, \ldots, n$. The pdf $f(y_i; \theta_i, \phi)$ of y_i is assumed to belong to the exponential family and it can be written in the form

$$f(y_i; \theta_i, \phi) = exp(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + h(y_i, \phi)),$$

where θ_i is a natural parameter and ϕ is a dispersion parameter (assumed to be known).
(a) Using the above pdf, express $u_i = E(Y_i)$ and the canonical link function.
(b) If y_i has Bernoulli distribution, that is, $f(y_i, \pi_i) = \pi^{y_i}(1-\pi_i)^{1-y_i}$, express the natural parameter θ_i as function of π_i. What is the canonical link function?
(c) Let $U = \frac{\partial l}{\partial \beta}$ (score function) and $I = -E(\frac{\partial^2}{\partial\beta\partial\beta'})$ (Fisher information matrix). Show that

$$U = XW\Delta(y-\mu), \qquad I = X'WX,$$

where X is the design matrix and $diag(w_{11}, \cdots, w_{nn})$, $\Delta = diag(\partial\eta_1/\partial\mu_1, \cdots, \partial\eta_n/\partial\mu_n)$, and $w_{ii} = (a(\phi)\nu_{ii})^{-1}(\partial\mu_i/\partial\eta_i)^2$.
(d)Explain the 'Iteratively Reweighted Least Squares' algorithm for computing the MLE.

10. Consider the regression data $(x_1, y_1), \cdots, (x_n, y_n)$, where $y_i \sim indepB(n_i, \pi_i), i = 1, \cdots, n$. We want to fit the logistic regression model, $log(\pi_i/(1-\pi_i)) = \beta_0 + \beta_1 x_i$. Derive the Fisher information matrix I.

11. Consider the simple linear regression model with first-order autoregressive errors;

$$y_t = \beta_0 + \beta_1 x_t + \epsilon_t, \qquad \epsilon_t = \phi\epsilon_{t-1} + a_t,$$

where $a_t \sim i.i.d.N(0, \sigma_a^2)$ and $|\phi| < 1$.
(a) Explain Durbin-Watson test for testing $H_0 : \phi = 0,\ v.s.\ H_1 : \phi \neq 0$.
(b) Explain maximum likelihood method for estimating the parameters.
(c) Give a reasonable forecast of $y_{T+\tau}$ for the period $T+1$ at the end of the current time period T, where τ is a positive integer.

12. Consider the linear model $y = x'\beta + \epsilon$. The M-estimator of β is the minimizer of

$$\sum_{i=1}^{n} \rho((y_i - x_i'b)/s)$$

with respective to b, where ρ is a differentiable function with derivative ψ and s is a robust estimate of scale. Explain how to find the M-estimator using 'Iteratively reweighted leasst square' method.

13. Consider the linear model $y = x'\beta + \epsilon$. We want to find $100(1-\alpha)\%$ confidence interval for

$$\Delta = \frac{a'\beta + d}{c'\beta + q},$$

where a, c are constant vectors and d, q are constants.

(a) Explain Fieller's method to find the confidence interval.

(b) Suppose that $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$, where $\epsilon_i \sim iidN(0, \sigma^2)$. Let x_m be the value for which

$$E(y|x) = \beta_0 + \beta_1 x + \beta_2 x^2$$

achieves an extremum (maximum or minimum). Derive a formula for a 95% confidence region for x_m. When is this region an interval?