Regression Analysis Tutoring2

Seung Bong Jung

Seoul National University

November 1, 2021

Model Adequacy Checking

- In the usual regression, we assume the linearity of the model. But what if the linear model is not adequate for the given data? There are some useful methods for diagnosing the model adequacy based on the residual.
- Under the model assumption: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathsf{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$, it follows that the residual vector $\hat{\mathbf{e}} \equiv \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$, where $\mathbf{H} \equiv \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t = (h_{ij})$ is the sol-called hat matrix.
- Recalling the decomposition of the projection operator $\Pi_{\mathbf{X}} = \Pi_{\mathbf{1}} + \Pi_{\mathbf{X}_{1,\perp}}$ and the formula $\mathbf{X}_{1,\perp} = (\mathbf{x}_1 - \bar{x}_1\mathbf{1}, \ldots, \mathbf{x}_p - \bar{x}_p\mathbf{1})$, we get

$$h_{ii} = n^{-1} + \sum_{j,k=1}^{p} S^{jk}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$$

where S^{jk} is the (j,k)-entry of the inverse matrix, $(\mathbf{X}_{1,\perp}^t\mathbf{X}_{1,\perp})^{-1} = \mathbf{S}^{-1} = (S_{jk})^{-1}$.

Residuals and Leverages

- Some simple algebra shows

$$0 \leq h_{ii} \leq 1, \quad \sum_{i=1}^{p} h_{ii} = p + 1$$

- From this one can deduce that $\frac{1}{n} \leq h_{ii} \leq 1$. Note that $\frac{1}{n} \leq h_{ii} + \frac{\hat{\mathbf{e}}_i^2}{\hat{\mathbf{e}}^t \hat{\mathbf{e}}} \leq 1$.(Assignment2)
- The value h_{ii} determines the strength of the leverage of the ith predictor data point $(x_{i1}, \ldots, x_{ip})^t$.
- Since $\text{var}(\hat{\mathbf{e}}_i) = (1 - h_{ii})\sigma^2$, a data point that has a larger leverage has a smaller variance for its residual.

Regression Diagnostics

- A point $(x_{i1}, \ldots, x_{ip}, Y_i)^t \in \mathbb{R}^{p+1}$ with high leverage tends to drag the fitted plane toward itself, in case it is far distant from the fitted plane without it.
- Leverage point: A data point with a high leverage h_{ii}. A common practice is to use the criterion $h_{ii} > 2(p+1)/n$.
- Influential point: A data point such that two regression fits, one with it and the other without it, give quite different results.
- Note that leverage points are not necessarily influential points.

Regression Diagnostics

- Leverage points are determined solely by the design points, i.e., $\mathbf{X}$, while influential points are identified only after the regression model being fitted with the observed Y_i at hands.
- One popular measure of influence is Cook's distance defined by

$$D_i = \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})^t \mathbf{X}^t \mathbf{X} (\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})}{(p+1)\hat{\sigma}^2}, \quad 1 \le i \le n,$$

  where $\hat{\boldsymbol{\beta}}_{(i)}$ denotes the estimator with the ith observation being deleted.
- Some algebra may prove $D_i = \frac{\hat{\mathbf{e}}_i^2}{(p+1)\mathsf{MSE}} \frac{h_{ii}}{(1-h_{ii})^2}$.
- Typically, those points with $D_i > 1$ are considered to be influential.

Scaled Residuals

- PRESS residuals: $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$, where $\hat{Y}_i$ denotes the prediction of Y_i based on the fitted regression equation with the ith observation being deleted.
- It turns out that $Y_i - \hat{Y}_i = \frac{\hat{\mathbf{e}}_i}{1-h_{ii}}$
- Studentized residuals: $(Y_i - \hat{Y}_i)/\hat{\text{Var}}(Y_i - \hat{Y}_i) = \hat{\mathbf{e}}_i/\sqrt{(1-h_{ii})\text{MSE}}$

Checking Model Assumptions

The model assumption: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathrm{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$, may be checked by examining the residuals. This is done in the hope that the behavior e_i mimics that of ϵ_i if the regression is correctly specified.

- Q-Q plot(Normal probability plot): To check the normality of the error distribution. If $e_{(i)}$ denotes the ith ordered e_i: $e_{(1)} \leq \cdots e_{(n)}$, then it is a plot of

$$e_{(i)} \text{ versus } \Phi^{-1}((i - 1/2)/n), \quad i = 1, 2, \ldots, n$$

- P-P plot: To see cumulative probability.

$$(i - 1/2)/n \text{ versus } \Phi(e_{(i)}), \quad i = 1, 2, \ldots, n$$

Generalized Least Squares Regression

- Let $\text{Var}(\epsilon) = \sigma^2\mathbf{V}$, where $\mathbf{Y}$ is not the identity matrix $\mathbf{I}$, but positive definite matrix. consider the following transformation of the model:

$$\mathbf{V}^{-\frac{1}{2}}\mathbf{Y} = \mathbf{V}^{-\frac{1}{2}}\mathbf{X}\boldsymbol{\beta} + \mathbf{V}^{-\frac{1}{2}}\boldsymbol{\epsilon}$$

- Generalized least squares estimator: The least squares estimator of $\boldsymbol{\beta}$ for the above transformed model is given by

$$\hat{\boldsymbol{\beta}}_G = (\mathbf{X}^t\mathbf{V}\mathbf{X})^{-1}\mathbf{X}^t\mathbf{V}^{-1}\mathbf{Y}.$$

- E($\hat{\boldsymbol{\beta}}_G$)= $\boldsymbol{\beta}$ and $\text{Var}(\hat{\boldsymbol{\beta}}_G) = \sigma^2(\mathbf{X}^t\mathbf{V}\mathbf{X})^{-1}$
- Then, clearly

$$||\mathbf{V}^{-\frac{1}{2}}\mathbf{Y} - \mathbf{V}^{-\frac{1}{2}}\mathbf{X}\hat{\boldsymbol{\beta}}_G||^2/\sigma^2$$

$$= (\mathbf{V}^{-\frac{1}{2}}\mathbf{Y})^t[\mathbf{I}-\mathbf{V}^{-\frac{1}{2}}\mathbf{X}(\mathbf{X}^t\mathbf{V}\mathbf{X})^{-1}\mathbf{X}^t\mathbf{V}^{-\frac{1}{2}}](\mathbf{V}^{-\frac{1}{2}}\mathbf{Y})/\sigma^2 \sim \chi^2(n-p-1)$$

Weighted Least Squares Regression

- Specalization: $\mathbf{V} = \text{diag}(w_i^{-1})$ for $w_i > 0$, i.e., the error terms ϵ_i are uncorrelated but have unequal variances $\text{Var}(\epsilon_i) = \sigma^2/w_i$.
- Applying the generalized least squares method in this special case is simply doing weighted least squares that minimizes the weighted sum of squared errors

$$\sum_{i=1}^{n} w_i(Y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2,$$

with the weight to each data point being inversely proportional to the variance of corresponding response.

Lack of Fit Test

We wish to test

$$H_0 : \mathrm{E}(Y|x_1, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \text{ versus } H_1 : \text{ not } H_0$$

- The basic idea is to compare estimates of $\mathrm{E}(Y|x_1, \ldots, x_p)$ with and without the model assumption.
- Need repeated measurements: With only a single observation of Y at each design point $(x_1, \ldots, x_p)$ we cannot get a desirable estimate of $\mathrm{E}(Y|x_1, \ldots, x_p)$ without the model assumption.
- Assume we have r_i repeated measurements $Y_{i,1}, Y_{i,2}, \ldots, Y_{i,r_i}$ of Y at $(x_{i1}, \ldots, x_{ip})$, $i = 1, 2 \ldots, n$

Lack of Fit Test

- Estimation of $\mathrm{E}(Y|x_{i1},\ldots,x_{ip})$ without the model assumption:

$$\bar{Y}_i \equiv \frac{1}{r_i}(Y_{i,1}+Y_{i,2}+\cdots+Y_{i,r_i}).$$

- Estimation of $\mathrm{E}(Y|x_{i1},\ldots,x_{ip})$ with the model assumption: The regression model for the full dataset is given by $\mathbf{Y}_F = \mathbf{X}_F\boldsymbol{\beta}+\boldsymbol{\epsilon}_F$, where

$$\mathbf{Y}_F = \begin{pmatrix} \mathbf{Y}_1 \\ \vdots \\ \mathbf{Y}_n \end{pmatrix}, \quad \mathbf{X}_F = \begin{pmatrix} \mathbf{1}_{r_1} & x_{11}\mathbf{1}_{r_1} & \cdots & x_{1p}\mathbf{1}_{r_1} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{1}_{r_n} & x_{n1}\mathbf{1}_{r_n} & \cdots & x_{np}\mathbf{1}_{r_n} \end{pmatrix}.$$

We obtain $\hat{\boldsymbol{\beta}} = (\mathbf{X}_F^t\mathbf{X}_F)^{-1}\mathbf{X}_F^t\mathbf{Y}_F = (\mathbf{X}^t\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^t\mathbf{W}\bar{\mathbf{Y}}$, where $\mathbf{W} = \mathrm{diag}(r_i)$ and $\bar{\mathbf{Y}} = (\bar{Y}_1,\ldots,\bar{Y}_n)^t$, and

$$\hat{Y}_{ij} \equiv \hat{Y}_i = (1, x_{i1},\ldots,x_{ip})(\mathbf{X}^t\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^t\mathbf{W}\bar{\mathbf{Y}}$$

Lack of Fit Test

- Decomposition of sum of squares: Since $\mathbf{Y}_i - \bar{Y}_i \mathbf{1}_{r_i} \perp \mathcal{C}_{\mathbf{1}_{r_i}}$,

$$
\begin{aligned}
||\mathbf{Y}_F - \hat{\mathbf{Y}}_F||^2 &= \sum_{i=1}^{n} ||\mathbf{Y}_i - \hat{Y}_i \mathbf{1}_{r_i}||^2 \\
&= \sum_{i=1}^{n} ||\mathbf{Y}_i - \bar{Y}_i \mathbf{1}_{r_i}||^2 + \sum_{i=1}^{n} ||\bar{Y}_i \mathbf{1}_{r_i} - \hat{Y}_i \mathbf{1}_{r_i}||^2 \\
&= \sum_{i=1}^{n} \sum_{j=1}^{r_i} (Y_{ij} - \bar{Y}_i)^2 + \sum_{i=1}^{n} r_i (\bar{Y}_i - \hat{Y}_i)^2, \\
\text{SSE} &= \text{SSPE} + \text{SSLF}
\end{aligned}
$$

- We have seen that $\hat{Y}_i$ is also obtained by weighted least squares regression of the response means $\bar{Y}_i$ onto $\mathbf{x}_1, \ldots, \mathbf{x}_p$, so that SSLF may be regarded as the weighted residual sum of squares

$$
(\mathbf{W}^{-\frac{1}{2}} \bar{\mathbf{Y}})^t [\mathbf{I} - \mathbf{W}^{-\frac{1}{2}} \mathbf{X} (\mathbf{X}^t \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^t \mathbf{W}^{-\frac{1}{2}}] (\mathbf{W}^{-\frac{1}{2}} \bar{\mathbf{Y}})
$$

Lack of Fit Test

- Large SSLF indicates discrepancy between $\mathsf{E}(Y|x_1, \ldots, x_p)$ and the model $\{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p : \beta_k \in \mathbb{R}\}$.
- Indeed, it folows that, under the model assumption with $\epsilon_F \sim \mathsf{N}(\mathbf{0}, \sigma^2 \mathbf{I})$,

$$\mathsf{SSLF}/\sigma^2 \sim \chi^2(n-p-1),$$

and that SSPE&SSLF are stochastically independent. Furthermore,

$$\mathsf{SSPE}/\sigma^2 \sim \chi^2(N-n), \quad N = \sum_{i=1}^{n} r_i$$

- F-statistic:

$$F \equiv \frac{\mathsf{SSLF}/(n-p-1)}{\mathsf{SSPE}/(N-n)} = \frac{\frac{\mathsf{SSLF}}{\sigma^2}/(n-p-1)}{\frac{\mathsf{SSPE}}{\sigma^2}/(N-n)} \sim F(n-p-1, N-n).$$

- Reject H_0 if $F > F_\alpha(n-p-1, N-n)$.