MID_RC

Haohong Shang

2023-06-26

$$Y = X\beta + \epsilon, \epsilon \sim N(0, \sigma^2)$$

0. Assumptions

- About the model form(*L2P8*)
    - Linear in parameters, not in predictors.
- About the errors(*L2P6, L12P5*)
    - ϵ is independent and is normally distributed with mean 0 and variance σ^2.
- About the predictors(*L12P6*)
    - Non-random fixed values
    - Measured without error
    - Linearly independent
- About the observations(*L12P7*)
    - All observations are equally reliable and have an approximately equal role in determining the regression results and in influencing conclusions.

1. What is this model?

- SLR(*L2P5*)
- MLR(*L2P12*)

2. Why is this model? Gauss-Markov Model

- A Simple Idea(*L3P5*): To make all residuals on average close to zero, consider minimizing the sum of squares**(OLS)**.
- Gauss-Markov(*Wikipedia*): the OLS estimator has the lowest sampling variance within the class of linear unbiased estimators, if the errors in the linear regression model are uncorrelated, have equal variances and expectation value of zero.
- ϵ(*L4P10*): $E(\epsilon) = 0$, $cov(\epsilon) = \sigma^2 I_n$.
- BLUE(*L4P16*): the OLS estimator is the best linear unbiased estimator is the sense that $cov(\hat{\beta}) \preceq cov(\tilde{\beta})$ for any estimator $\tilde{\beta}$ satisfying …

3. How to get unknown values? What's their properties?

- $\hat{\beta}$
    - Value(*L3P6~P17*): $\hat{\beta} = (X^T X)^{-1} X^T Y$. $X^T X$ **is non-degenerate.**
    - Property(*L4P11*): $E(\hat{\beta}) = \beta$, $cov(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$.
    - **SLR case(*L3P8, L6P23*)**
    - Distribution(*L6P7*): $\frac{\hat{\beta}_j - \beta_j}{\sqrt{\widehat{Var}(\hat{\beta}_j)}} \sim t_{n-p-1}$, as σ^2 is unknown. Usually, $\beta_j = 0$(in lm()).
    - $100(1-\alpha)\%$ CI for β_j(*L6P18*): $\hat{\beta}_j \pm t_{n-p-1,\alpha/2} s.e.(\hat{\beta}_j)$.
- $\hat{e}$
    - Value(*L4P12*): $\hat{\epsilon} = (I_n - H)Y$.
    - Property(*L4P12*): $E(\hat{\epsilon}) = 0$, $cov(\hat{\epsilon}) = \sigma^2 (I_n - H)$.
- $\hat{\sigma}^2$
    - Value(*L4P14*): $\hat{\sigma}^2 = RSS/(n-p-1)$.
    - Property(*L4P14*): $E(\hat{\sigma}^2) = \sigma^2$.
- $\hat{Y}$ and $\hat{\epsilon}$ are uncorrelated(*L4P12*).
- Joint distribution of $(\hat{\beta}, \hat{\sigma}^2)$(*L5P5*).
- Joint distribution of $(\hat{Y}, \hat{\epsilon})$(*L5P7*).
- Sum of squares(*L7P21*).

4. How to use this model if assumptions are met?

- Confidence Interval and Prediction Interval
    - MLR(*L7P18*)
    - SLR(*L6P25*)
- Comparison
    - R^2(*L7P23*) and Adjusted R^2(*L7P24*)

        - If I have 2 models: M_1=lm(y~x1), M_2=lm(y~x1+x2), R_2^2 will always larger than R_1^2, but adj-R^2 is not always this case, so the model with the larger adj-R^2 is preferred.
    - F-Statistic and ANOVA(*L8, L10, Appendix*)
        - Suppose I have a full model M_1 and a reduced submodel M_2, and I want to use F-test to compare these 2 models. Set M_2 to be H_0, M_1 to be H_1. If H_0 is true, then F-stats should follow $F_{n,n-p-1}$ distribution under H_0, and the corresponding p-value should be larger(usually >0.05). If p-value is smaller than 0.05, then H_0 is rejected at the 0.05 level of significance, so the assumption that "M_2 is true" is possibly wrong. As an alternative, H_1 is preferred.(**SUGGEST REVIEWING VE401!**)
        - ANOVA table helps you to deal with the calculation.
        - As SSE_{reduce} is always larger than SSE_{full}, if I have a reduced submodel with similar SSE but larger df, then $f-stats$ will be small, and thus the reduced model will be more favoured.
    - Other Methods(*L17P20*): MSE, AIC, BIC
    - Dummy Variables and Interaction(*L9, L10*)
- Feature Selection(*L17*)
    - What Happens if We Miss Necessary Predictors(*L17P8~P13*)?
    - What Happens if We Include Unnecessary Predictors(*L17P14*)?
    - Criteria(*L17P15~P18*): Description, Prediction, Control
    - Procedure/Algorithm(*L17P22~P27*): Forward selection, Backward elimination, Stepwise selection.

5. What if assumptions are violated?

- Linearity: non-linear(*L11*)
    - Residual Plot(*L12P21~P23, L15P7*)
    - Polynomial Models(*L11P11*)
        - Fitting the polynomial model doesn't mean we believe it is correct. It is just a decent approximation to the true underlying nonlinear model. One can try higher-order polynomials if lower-order ones don't capture the nonlinear pattern well(*L11P11, HW1Q3d*).
        - **Extrapolating the model beyond the range of data is dangerous**(*L11P12*).
        - Interpretation of Coefficients in a Polynomial Model(*L11P15*): it makes no sense to interpret a single coefficient for a polynomial.
    - Ordinal Categorical Predictors(*L11P20*)
        - Incorporate the ordinal info of a ordinal predictor by assigning a score to each its category(*L11P20~P22*).
        - Pros and Cons(*L11P26*).
    - Check Interactions of 2 Numerical Variables(*L13P23*)
    - Transformation
        - Transform Linearizable Models(*L15P5*)
        - QQ Plot(*L12P28~P37*): reduce skewness(*L15P16~P20*). Histogram of the residuals should be bell-shaped if normal.
        - Residual Plus Component Plot(*L13P41*): detect non-linearity.
        - Box-Cox Method(*L15P26*): an automatic procedure to select the "best" power λ that make the residuals of the model $Y = X\beta + \epsilon$ closest to normal and constant variability.
- Errors: unequal variance (mainly)
    - Residual Plot(*L12P19*)
    - Weighted Least Squares(*L16*)
        - Idea is simple and calculation is very similar to OLS.
    - Bootstrap Method(*L17P4*): sample with replacement.
- Observation: unreliable data
    - A good blog to tell the difference among high leverage points, influential points, and outliers: [link][https://www.cnblogs.com/HuZihu/p/12017890.html (https://www.cnblogs.com/HuZihu/p/12017890.html)].
    - Leverage h_{ii}(*L12P9~P11*): using hat matrix H, h_{ii} is the leverage of the i-th observation.
        - If it's close to 1, then this i-th observation is a high leverage point.
        - SLR case.
    - Standardized Residuals r_i(*L12P13*)
        - Observation with large $|r_i|$ (over 2 or 3 or 4) are potential outliers.
        - (Externally) Studentized residuals r_i^*(*L12P15, L14P21*): If an observation is not an outlier, $r_i^* \approx r_i$.
        - Comparisons of 3 Types of Residuals(*L12P16*).
    - Influential Points(*L14P6*)
        - An influential point has an unduly large effect on the model. Observations whose removal will cause major changes in the regression analysis are called influential points.
        - Check using Cook's Distance(*L14P16*).
        - **Influential points are not necessarily outliers. A point can be influential, an outlier, or both.**

6. Review Suggestions

1. ANOVA table and related calculations.
2. Difference between t-test and f-test. What's hypothesis and confidence level?
3. Sample questions and quiz.

Best Wishes! Best Wishes! Best Wishes!