Statistical Computation
Math 475

Jimin Ding

Department of Mathematics
Washington University in St. Louis
www.math.wustl.edu/ jmding/math475/index.html

October 10, 2013

Part IV

Regression

Outline I

1. Simple Regression
   - Model Setup
   - Estimation
   - Inference
   - Prediction
   - Model Diagnostic

2. Multiple Regression
   - Model Setup and Estimation
   - Model Selection
   - Collinearity and Ridge Regression

Examples

Classical regression models only deal with continuous response variable. Let Y denote response (dependent) variable and X denote explanatory (independent) variable (predictor).

- Grade point average (GPA). X: entrance test scores; Y: GPA by the end of freshman year.
- X: height; Y: weight.
- X: education level; Y: income.

The covariate X is usually continuous in regression and categorical covariates are commonly investigated in ANOVA (analysis of variance). But we can also use regression model to study the relationship between a continuous response and a categorical covariate by creating some dummy variables. This is equivalent to ANOVA/ANCOVA/MANOVA to some extend.

General Model Setup

Simple linear regression for one covariate:

$$Y_i = \beta_0 + \beta_1 X_i + e_i, \quad i = 1, \cdots, n,$$

where Y_i and X_i are the ith observed response and covariate variables, e_i is the random error, β_0 and β_1 are the unknown parameters.

Assumptions:

1. $E(e_i) = 0$ for all i's.
2. $Var(e_i) = \sigma^2$. (Homogeneity)
3. e_i and e_j are independent for any $i \neq j$. (Independence)
4. $e_i \overset{iid}{\sim} N(0, \sigma^2)$. (Normality)

Goal of Study:

- Describe the relationship between explanatory and response variables.
  $\Leftrightarrow$ Estimate $\beta_0, \beta_1, \sigma^2$.
- Predict/Forecast the response variable for a new given predictor value.
  $\Leftrightarrow$ Predict $E(Y_{new}|X_{new}) = \beta_0 + \beta_1 X_{new}$.
- Inference: testing whether the relationship is statistically significant.
  $\Leftrightarrow$ Find CI for β_1 to see whether it includes 0.
  $\Leftrightarrow$ Test: $H_0 : \beta_1 = 0$.
- Prediction interval of response variable.

- Estimator: a function of data that provides a guess about the value or parameters of interest.
  Example: $\bar{X}$ can be used to estimate μ.
- Criteria: how to choose a “good" estimator?
  The smaller the following quantities are, the better the estimator is:
  Bias: $E(\hat{\beta}) - \beta$,
  Variance: $Var(\hat{\beta})$,
  MSE: $E\{(\hat{\beta} - \beta)^2\}$.
- Question: the distribution of $\hat{\beta}$ is usually unknown since the distribution of Y is unknown!
  Approximate Bias, Variance and MSE by their sample version.

Least Square Estimator

To predict Y well in a simple linear regression, it is natural to obtain the estimators by minimizing:

$$Q = \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 X_i)]^2,$$

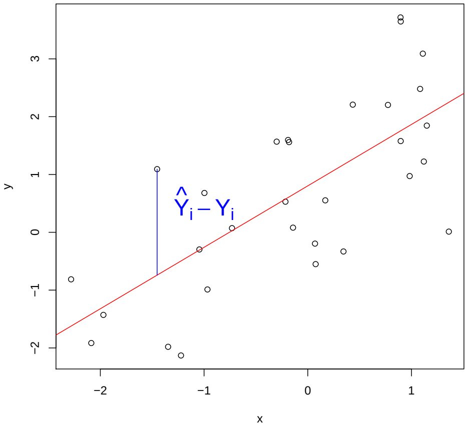
which is the so called ”Least Square Criterion“.

The obtained estimators

$$\begin{aligned}\hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X} \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \rho_{X,Y}\frac{S_Y}{S_X},\end{aligned}$$

are the so called ”Least Square Estimators“ (LSE).

Regression Line



- Regression lines: $y = \hat{\beta}_0 + \hat{\beta}_1 x$.
- Fitted values: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$.
- Residuals: $\hat{e}_i = \hat{Y}_i - Y_i$.

For LSE, we have $\sum_{i=1}^n \hat{e}_i = 0$ and $\sum_{i=1}^n \hat{e}_i^2$ is minimized.

Sum Squares in ANOVA Table

- SSE: sum of square errors $\sum_{i=1}^n \hat{e}_i^2$.
- SSTO: $\sum_{i=1}^n (Y_i - \bar{Y})^2$.
- SSR: $\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$.
- DF: the degree of freedom.
  DF of the residuals
  = the number of observations - the number of parameters in the model.
- MSE: SSE/DF of the error term.
  $\frac{1}{n-2} \sum_{i=1}^n \hat{e}_i^2 = \hat{\sigma}^2$ (estimator for σ^2).

Best Linear Unbiased Estimation (BLUE)

Note that

- LSE are unbiased estimators.
  $E(MSE) = E(\hat{\sigma}^2) = \sigma^2$; $E(\hat{\beta}_0) = \beta_0$; $E(\hat{\beta}_1) = \beta_1$.
- Both $\hat{\beta}_0$ and $\hat{\beta}_1$ are linear functions of observations $Y_1, \cdots, Y_n$.
- Among all unbiased estimators, LSE has smallest variance. In another words, it is more precise than any other unbiased linear predictors.

Remark: BLUE property still hold without the normality assumption in (4).

Other Type of Estimators

- Least absolute deviation estimator (LAD): a robust estimator.
- Weighted least square estimator (WLSE): an estimator to adjust for heterogeneity.
- Maximum Likelihood Estimator (MLE): MLE is based on the likelihood function and hence is only valid under the assumption (4).

MLE

Recall Likelihood:

$$L = \prod_{i=1}^{n} f(Y_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(Y_i - \beta_0 - \beta_1 X_i)^2\}.$$

The parameters $(\beta_0, \beta_1, \sigma^2)$ which maximize above likelihood function, $L(\beta_0, \beta_1, \sigma^2)$, are defined as MLE of $(\beta_0, \beta_1, \sigma^2)$, and denoted by $(\hat{\beta}_{0,ML}, \hat{\beta}_{1,ML}, \hat{\sigma}^2_{ML})$.

Properties of MLE

In the simple linear regression with the normality assumption on errors, LSE for β's are same as MLE for β's. (They are different for σ^2.) So MLE for β's are also BLUE. Usually in most of models, MLE are

- Consistent: $\hat{\beta} \to \beta$ in probability or a.s.
- Sufficient: $f(Y_1, \cdots, Y_n|\hat{\theta}_{ML})$ does not depend on θ.
- MVUE: minimum variance unbiased estimator.
- Asymptotic efficient: $Var(\hat{\beta})$ reach the Cramér-Rao lower bound. (minimum variance)

Confidence Interval

Note: $\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2$ are functions of data and so are random variables. As point estimators, they only provide a guess about true parameters and will change if the data are changed. We also want to get a range guess for $\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2$, which guarantee that with certain probability the true parameter will be in the range. For example, if we repeat the experiments 100 times and collect 100 set of data, 95 out the 100 guessed range will contain the true parameters. This range is called confidence interval.

CI for β_0 : $\hat{\beta}_0 \pm t_{1-\alpha/2,n-2} se(\hat{\beta}_0)$;
CI for β_1 : $\hat{\beta}_1 \pm t_{1-\alpha/2,n-2} se(\hat{\beta}_1)$.

Standard Errors

$$\begin{aligned}
Var(\hat{\beta}_0) &= Var(\bar{Y} - \hat{\beta}_1 \bar{X}) \\
&= \\
&= \sigma^2(\frac{1}{n} + \frac{\bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2}) \\
Var(\hat{\beta}_1) &= Var(\frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}) \\
&= \\
&= \sigma^2(\frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2})
\end{aligned}$$

But σ^2 is still unknown, so we use the estimator $\hat{\sigma}^2 = MSE$ to replace σ^2 in estimating standard errors.

Standard Errors

Hence

$$\begin{aligned}
se(\hat{\beta}_0) &= \sqrt{MSE\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)}, \\
se(\hat{\beta}_1) &= \sqrt{MSE\left(\frac{1}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)}.
\end{aligned}$$

Distribution of Estimators

Under the normality assumption,

$$\frac{\hat{\beta}_p - \beta_p}{se(\hat{\beta}_p)} \sim t_{n-2},$$

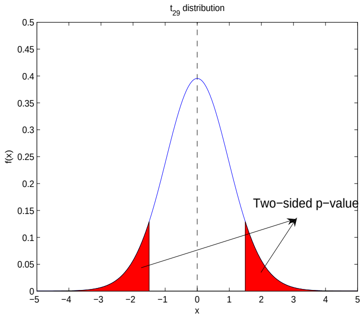
and without the normality assumption,

$$\frac{\hat{\beta}_p - \beta_p}{se(\hat{\beta}_p)} \sim t_{n-2} \quad \text{approximately},$$

for $p = 1, 2$.

Hypothesis Test

- $H_0 : \beta_1 = 0 \quad \text{v.s.} \quad \beta_1 \neq 0$.
  (Whether Y_i depends on X_i or not.)
- Test statistic: $t = \frac{\hat{\beta}_1 - 0}{se(\hat{\beta}_1)} \sim t_{n-2}$.
- p-value$=P(T_{n-2} > |t|)$. (two sided)



Confidence Interval for the Mean of an Observation

Let $\hat{Y}_{new} = \hat{\beta}_0 + \hat{\beta}_1 X_{new}$.

$$\begin{aligned}
Var(\hat{Y}_{new}) &= Var(\hat{\beta}_0 + \hat{\beta}_1 X_{new}) \\
&= \\
&= \sigma^2\left(\frac{1}{n} + \frac{(\bar{X} - X_{new})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right).
\end{aligned}$$

Hence a $(1-\alpha)\%$ CI for the mean value of the observation $E(Y_{new}) = \beta_0 + \beta_1 X_{new}$ is

$$\hat{Y}_{new} \pm t_{1-\alpha/2, n-2} \sqrt{MSE\left(\frac{1}{n} + \frac{(\bar{X} - X_{new})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)},$$

which is denoted as *CLM* in SAS output.

Prediction Interval (PI)

A $(1-\alpha)\%$ prediction interval is an interval I such that $P(Y_{new} \in I) = 1-\alpha$. Note that $Y_{new} = \beta_0 + \beta_1 X_{new} + e_{new}$ is random in the sense that $Var(Y_{new}) = \sigma^2 \neq 0$. We can derive

$$Y_{new} - \hat{Y}_{new} \sim N(0, \sigma^2(1 + \frac{1}{n} + \frac{(\bar{X} - X_{new})^2}{\sum_{i=1}^n (X_i - \bar{X})^2})),$$

$$1 - \alpha = P(|\frac{Y_{new} - \hat{Y}_{new}}{\sqrt{MSE(1 + \frac{1}{n} + \frac{(\bar{X} - X_{new})^2}{\sum_{i=1}^n (X_i - \bar{X})^2})}}| \leq t_{1-\alpha/2,n-2}).$$

Hence $(1-\alpha)\%$ PI for Y_{new}, denoted as *RLCLI* in SAS, is:

$$\hat{Y}_{new} \pm t_{1-\alpha/2,n-2}\sqrt{MSE(1 + \frac{1}{n} + \frac{(\bar{X} - X_{new})^2}{\sum_{i=1}^n (X_i - \bar{X})^2})}.$$

Simultaneous Confidence/Prediction Bands

- Confidence band for $E(Y)$: Different from prediction interval and confidence interval for the mean, it is a simultaneous band for the entire regression line.

$$\hat{Y}_i \pm W\sqrt{MSE(\frac{1}{n} + \frac{(\bar{X} - X_i)^2}{\sum_{i=1}^n (X_i - \bar{X})^2})},$$

  where $W^2 = 2F(1-\alpha; 2, n-2)$ and $i = 1, \cdots, n$.
- Prediction band for the Y_i's in the entire region:

$$\hat{Y}_i \pm S(\text{or } B)\sqrt{MSE(\frac{1}{n} + \frac{(\bar{X} - X_i)^2}{\sum_{i=1}^n (X_i - \bar{X})^2})},$$

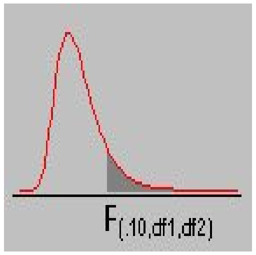
  where $S^2 = gF(1-\alpha; g, n-2)$ (Scheffé type) and $B = t(1-\alpha/2g, n-2)$ (Bonferroni type).

Coefficient of Determination

- Coefficient of Determination is defined as $R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} \in [0,1]$.
- Interpretation: the proportion reduction of total variation associated with the use of the predictor variable X. The larger R^2 is, the more the total variation of Y is explained by X.
- In simple linear regression, R^2 is same as $\hat{\rho}^2$.
- The R close to 0 does not imply that X and Y are not related, but simply means that the linear correlation between X and Y is small.

F-test for Goodness of Fit

- H_0 : Reduced model $Y_i = \beta_0 + e_i$, v.s.
  H_1 : Full model $Y_i = \beta_0 + \beta_1 X_i + e_i$.
  This is equivalent to test $H_0 : \beta_1 = 0$.
- Test statistic: $F = \frac{MSR}{MSE} \sim F(1-\alpha; 1, n-2)$, where $MSR = SSR/$the number of model parameters $-$ 1. (Note that $F_{1,df} = t_{df}^2$)
- P-value:



General F-test

If model 1 (reduced model) is nested (submodel) within model 2 (full model), the comparison between two models can be done by F-test.

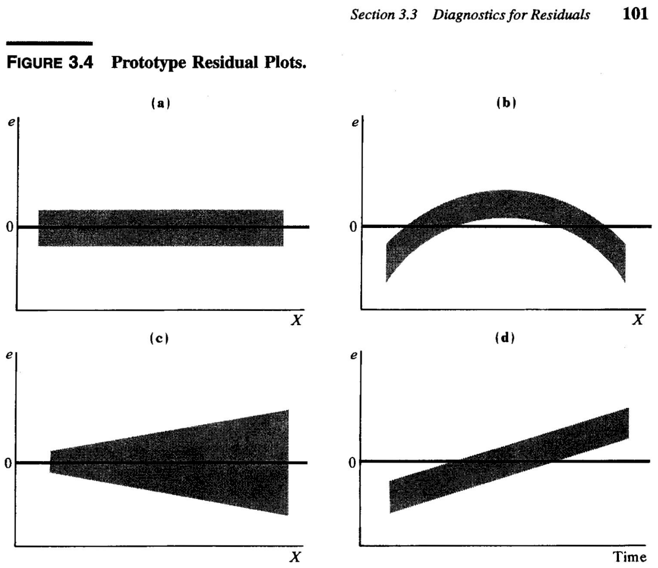
$$F^* = \frac{SSE(R) - SSE(F)}{df(R) - df(F)} \div \frac{SSE(F)}{df(F)}$$

General F-test is commonly used in model selection.

Residual Plots

- Residuals against the index of observations:
  1. symmetric around 0;
  2. constant variability;
  3. no serial correction.
- Residuals against the predicted values. (add a link to possible problematic residual plots.)
- QQ plot/normality plot: check the normality assumption. Under the normality assumption, the residuals should be close to the reference line or look linear.

Prototype Residual Plots



Section 3.3 Diagnostics for Residuals **101**

FIGURE 3.4 Prototype Residual Plots.

Outliers and Influential Observations

Cook's distance is defined as

$$D_i = \frac{\sum_{j=1}^n \hat{Y}_j - \hat{Y}_{j(i)}}{p \cdot MSE} = \sum_{i=1}^n \frac{\hat{e}_i^2}{p \cdot MSE}[\frac{h_{ii}}{(1-h_{ii})^2}].$$

The value of Cook's distance for each observation represents a measure of the degree to which the predicted values change if the observation is left out of the regression.

If an observation has an unusually large value for the Cook's distance, it might be worth to further investigations. (small influence: $D < 0.1$; huge influence: $D > 0.5$)

Although above definition requires to fit regression n times, it can be simplified and only need to fit model once. Here h_{ii} is the i-th diagonal element of the hat matrix $H = X(X^TX)^{-1}X^T$.

Model Setup

- Two predictors: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + e_i$.
- More general:
  Consider $p-1$ predictors $X_{i1}, \cdots, X_{i(p-1)}$,

  $$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_{p-1} X_{i(p-1)} + e_i,$$

  for $i = 1, \cdots, n$. We may write it in the following matrix form

  $$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{e},$$

  where $\boldsymbol{Y} = (Y_1, \cdots, Y_n)^T$, $\boldsymbol{\beta} = (\beta_0, \cdots, \beta_{p-1})^T$, $\boldsymbol{e} = (e_1, \cdots, e_n)^T \sim N(0, \sigma^2 I)$ and X is the $n \times p$ design matrix.

Estimation

- The LSE for $\hat{\boldsymbol{\beta}}_{LS} = (X^TX)^{-1}X^TY$
- $Cov(\hat{\boldsymbol{\beta}}_{LS}) = \sigma^2(X^TX)^{-1}$.
- $se(\hat{\boldsymbol{\beta}}_{LS}) = [MSE(X^TX)^{-1}]^{1/2}$.
- Under normality assumption, $\hat{\boldsymbol{\beta}}_{ML} = \hat{\boldsymbol{\beta}}_{LS}$.
- The fitted values: $\hat{\boldsymbol{Y}} = X\hat{\boldsymbol{\beta}} = X(X^TX)^{-1}X^T\boldsymbol{Y}$.
- Hat matrix $H = X(X^TX)^{-1}X^T$.
- The residuals: $\hat{\mathbf{e}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = (I - H)\boldsymbol{Y}$, and $Cov(\boldsymbol{e}) = \sigma^2(I - H)$.

ANOVA Table

- SSE: $\hat{\mathbf{e}}^T\hat{\mathbf{e}} = \boldsymbol{Y}^T(I - H)\boldsymbol{Y}$.
- SSTO: $\sum_{i=1}^n (Y_i - \bar{Y})^2 = \boldsymbol{Y}^T\boldsymbol{Y} - \frac{1}{n}\boldsymbol{Y}^TJ\boldsymbol{Y}$, where J is the $n \times n$ matrix with all components equal to 1.
- SSR=SSTO-SSE.
- MSE=SSE/(n-p).
- MSTO=SSTO/(n-1).
- MSR=SSR(p-1).
- Overall F-test:
  $H_0 : \boldsymbol{\beta} = \mathbf{0} \quad (\beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0)$
  H_a : not all β's are zeros.
  $F = \frac{MSR}{MSE} \sim F_{1-\alpha;p-1,n-p}$.
- $R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$.
- Adjusted R^2: adjust for the number of predictors
  $\frac{1}{n-1}(1 - R_A^2) = \frac{1}{n-p}(1 - R^2)$.

Model Selection: Covariates

- Forward selection: from no covariates
- Backward selection: from all covariates
- Stepwise selection: Backward+forward

Simple Regression
Model Setup
Estimation
Inference
Prediction
Model Diagnostic
Multiple Regression
Model Setup and Estimation
Model Selection
Collinearity and Ridge Regression

When Collinearity Happens

- Adding or deleting a predictor changes R^2 substantially.
- Type III SSR heavily depends on other variables in the model.
- $se(\hat{\beta}_k)$ is large.
- Predictors are not significant individually, but simple regression on each covariate is significant.

Simple Regression
Model Setup
Estimation
Inference
Prediction
Model Diagnostic
Multiple Regression
Model Setup and Estimation
Model Selection
Collinearity and Ridge Regression

Remedy for Collinearity

- Centering and standardize the predictors, which might be helpful in polynomial regression when some of the predictors are badly scaled.
- Drop the correlated variables by model selection.
  1. The predictor is not significant.
  2. The reduced model after dropping the predictor fits data nearly as well as the full model.
- Add new observations. (Economy, Business)
- Use the index of several variables (PCA)
- Ridge Regression

Simple Regression
Model Setup
Estimation
Inference
Prediction
Model Diagnostic
Multiple Regression
Model Setup and Estimation
Model Selection
Collinearity and Ridge Regression

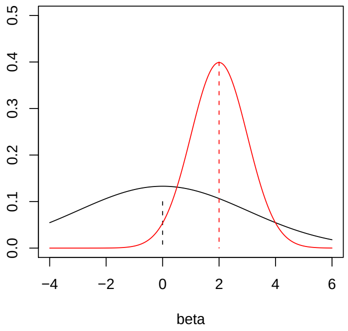
Variance Inflation Factor (VIF)

- Let R_j^2 is the coefficient of determination of X_j on all other predictors. (R_j^2 is the R^2 of regression model $X_{ij} = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{j-1} X_{i(j-1)} + \beta_{j+1} X_{i(j+1)} + \cdots + e_{ij}$.) Define $VIF_j = \frac{1}{1-R_j^2}$.
- $VIF_j \geq 1$ since $R_j^2 \leq 1$ for all j.
- If $VIF > 10$, we usually believe the variable has influential variation to cause collinearity problem.
- In standardized regression $Var\hat{\beta}_j = \sigma VIF_j$.
- In SAS, VIF table is reported in PROC REG by adding VIF option in MODEL statement.

Simple Regression
Model Setup
Estimation
Inference
Prediction
Model Diagnostic
Multiple Regression
Model Setup and Estimation
Model Selection
Collinearity and Ridge Regression

Ridge Regression

Recall that $\hat{\boldsymbol{\beta}}_{LS} = (X^T X)^{-1} X^T Y$, $E(\hat{\boldsymbol{\beta}}_{LS}) = \boldsymbol{\beta}$ and $Var(\hat{\boldsymbol{\beta}}) = (X^T X)^{-1}$. When $(X^T X)$ is nearly singular (the determinate is close to 0), LSE is unbiased but has large variance, which leads to large mean square error of the estimator.



The idea of the ridge regression is:
reduce variance at the cost of increasing bias.

Ridge Regression

The ridge regression estimator is:

$$\hat{\beta}_r(b) = (X^T X + bI)^{-1} X^T Y,$$

where b is a constant chosen by users and referred as tuning parameter.
As b increases, the bias increases but the variance decreases, $\hat{\beta}_r(b) \rightarrow 0$ (componentwise).
One may choose the tuning parameter b, such that

- ridge trace ($\hat{\beta}_r(b)$ against b) gets flat,
- VIF_j (against b) drop around 1.

SAS Program

- PROC GLM;
- PROC REG;
- PROC CATMOD;
- PROC GENMOD;
- PROC LOGISTIC;
- PROC NLINL;
- PROC PLS;
- PROC MIXED;

Reading Assignment

Textbook: Chapter 5 and Chapter 9.