STA302 Methods of Data Analysis

Notes of Lecture Slides

Heather T

May-June, 2019

Contents

1 Lecture 1

1.1 Permutation Test

We randomly select elements from the same distribution. If the groups have no effect, all of them are equally likely.

1.2 P-value

Under the assumption that groups have no effect,the probability of sampling a data set with a difference between groups as extreme as what we observed.
We say a difference is statistically significant if it's less probable than our pre-determinded significance level.
We say the groups have a significant effect if it causes the variable of interest to be significantly different.

2 Lecture 2

2.1 P-value

The p-value is the probability of a result as improbable as we've observed given the groups are the same.

2.1.1 Hypothesis testing

The statistical hypothesis testing is the evaluation of the compatibility of H_0 with the observe data.
By definition, the p-value is the probability of the data given H_0 is true. If the probability is samll, we reject the null. If H_0 is true, the p-value$\sim U(0,1)$.
With significance level α we expect to have to reject the null even though it is true with probability α.

Table of error types

Table of error Types	H_0	H_A
Fails to reject	Correct Inference	Type II Error
Reject	Type I Error	Correct Inference

2.2 The two-sample t-test for Mean

When two samples are independent random samples from a normal distribution with different mean μ_A and μ_B but same variance. To compare the two groups we will compare their empirical means $\bar{y}_A$ and $\bar{y}_B$

2.2.1 Known Variance

$$\bar{y}_A - \bar{y}_B \sim N(\mu_A - \mu_B, \sigma^2(\frac{1}{n_A} + \frac{1}{n_B}))$$

$$\frac{(\hat{y}_A - \hat{y}_B) - (\mu_A - \mu_B)}{\sigma\sqrt{(\frac{1}{n_A} + \frac{1}{n_B})}} \sim N(0,1)$$

2.2.2 Unknown Variance

$$s^2 = \frac{\sum_{i=1}^{n_A}(y_{iA} - \bar{y}_A)^2 + \sum_{i=1}^{n_B}(y_{iB} - \bar{y}_A)^2}{n_A + n_B - 2}$$

$$\frac{(\bar{y}_A - \bar{y}_B) - (\mu_A - \mu_B)}{s\sqrt{1/n_A + 1/n_B}} \sim t_{n_A+n_B-2}$$

2.2.3 Hypothesis testing

$H_0 : \mu_A = \mu_B$

Test stats:

$$\frac{(\bar{y}_A - \bar{y}_B)}{s\sqrt{1/n_A + 1/n_B}} \sim t_{n_A+n_B-2}$$

2.3 Analysis Of Variance (ANOVA)

If groups are different we expect there is a bigger difference between groups(reflecting the group effect) than within groups(natural variability of data)

There are k independent random samples available from k normal distributions $N(\mu_j, \sigma^2)$, for $j = 1, \cdots, k$

Set the **hypothesis test**:

H_0: $\mu_1 = \cdots = \mu_k$ vs. H_1: not all the μ_js are the same

2.3.1 Statistics associated with one-way ANOVA

- **The jth sample mean** is:

$$\bar{X}_{\cdot j} = \frac{1}{n_j}\sum_{i=1}^{n_j} X_{ij\cdot}$$

- **The overall sample mean** is:

$$\bar{X} = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j} X_{ij} = \frac{1}{n}\sum_{j=1}^{k} n_j \bar{X}_{\cdot j}$$

  Where $n = \sum_{j=1}^{k} n_j$ is the total number of the observation across all k groups.

- **The total variation**The total variation is:

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij} - \bar{X})^2$$

  with $n - 1$ degrees of freedom.

- **The between-groups variation SSG** is:

$$B = \sum_{j=1}^{k} n_j(\bar{X}_{\cdot j} - \bar{X})^2$$

  with $k - 1$ degrees of freedom

- **The within-groups variation SSE** is:

$$W = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_{\cdot j})^2$$

with $n - k = \sum_{j=1}^{k} (n_j - 1)$ degrees of freedom.

- **The ANOVA decomposition** is:

$$\sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_{ij} - \bar{X})^2 = \sum_{j=1}^{k} n_j (\bar{X}_{\cdot j} - \bar{X})^2 + \sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_{\cdot j})^2$$

The total variation is a measure of the overall variability in the data from all k groups about the overall sample mean.

B and W are also called, respectively **between-treatments variation** and **within-treatment variation**. In fact, W is effectively a *SSR(residual sum of squares)*, representing the variation which cannot be explained by the treatment or group factor.

There are some key formulae for manual computations:

- $n = \sum_{j=1}^{k} n_j$
- $\bar{X}_{\cdot j} = \sum_{i=1}^{n_j} \frac{X)_{ij}}{n_j}$ and $\bar{X} = \frac{\sum_{j=1}^{k} n_j \bar{X}_{\cdot j}}{n}$
- Total variaion = Total sum of squares $= B + W = \sum_{j=1}^{k} \sum_{i=1}^{n_j} X_{ij}^2 - n\bar{X}$
- $B = \sum_{j=1}^{k} n_j \bar{X}_{\cdot j}^2 - n\bar{X}$
- SSR = W $= \sum_{j=1}^{k} \sum_{i=1}^{n_j} X_{ij}^2 - \sum_{j=1}^{k} n_j \bar{X}_{\cdot j}^2 = \sum_{j=1}^{k} (n_j - 1) S_j^2$, where S_j^2 is the jth sample variance

Based on the given calculation result, we form the following result:

1. $B \perp W$
2. $\frac{W}{\sigma^2} = \frac{\sum_{j=1}^{k} \sum i{=}1^{n_j} (X_{ij} - \bar{X}_{\cdot j})^2}{\sigma^2} \sim \chi_{n-k}^2$
3. Under $H_0 : \mu_1 =, \ldots, = \mu_k$, then $\frac{B}{\sigma^2} = \sum_{j=1}^{k} n_j (\bar{X}_{\cdot j} - \bar{X})^2 / \sigma^2 \sim \chi_{k-1}^2$ (Then the numerator would be the form of sample minus the same sample mean).

In order to test $H_0 : \mu_1 = \ldots = \mu_k$, we define the following test statistic:

$$F = \frac{B/(k-1)}{W/(n-k)} = \frac{\sum_{j=1}^{k} n_j (\bar{X}_{\cdot j} - \bar{X})^2/(k-1)}{\sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_{\cdot j})^2/(n-k)}$$

While MSG = SSG/(k-1) $= B/(k-1) \sim \chi_{k-1}^2$, MSE = SSE/(n-k) $= W/(n-k) \sim \chi_{n-k}^2$
One way analysis use B and W from the set of random variables whose means are assumed to be equal, in order to get the F distribution test statistics. Which contain the information for only one variable .

2.3.2 Distribution Theory

We first get the unbiased estimator for σ_x^2

$$\hat{\sigma}_x^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

A ratio of estimator such like $\frac{\hat{\sigma}_x^2}{\hat{\sigma}_x^2}$ has a F distribution if and only if $\sigma_x^2 = \sigma_y^2$:

$$\frac{\hat{\sigma}_x^2}{\hat{\sigma}_x^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2/n-1}{\sum_{i=1}^n (y_i - \bar{y})^2/n-1} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2/\sigma_x^2 n-1}{\sum_{i=1}^n (y_i - \bar{y})^2/\sigma_y^2 n-1} \sim F_{n-1,n-1}$$

3 Lecture 3

3.1 Linear Relationship

$y = \beta_0 + \beta_1 x$ is the type of relationship we are interested in. How can we fit the best linear model to explain the relationship between y and x observed in data?

3.2 matrix notation

Respect the Dimension!

For a random variable:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix}, \mathbf{c} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix}, \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$E[x] = \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, cov[x] = \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12}^2 \\ \sigma_{12}^2 & \sigma_2^2 \end{bmatrix}$$

3.2.1 Expectation

For univariate case $\mathbf{E}[ax + b] = a\mathbf{E}[x] + c = a\mu + c \implies$ for matrix, $\mathbf{E[Ax+c]} = \mathbf{A}\mu + c$

3.2.2 Variance

For univariate case $\mathbf{V}[ax + b] = a^2\mathbf{V}[x] = a^2\sigma^2 \implies$ for matrix, $\mathbf{V[Ax+c]} = \mathbf{A}\Sigma\mathbf{A}^T$

3.3 Linear Regression matrix set-up

$\mathbf{x}$=$[x_1, \cdots, x_n]$

$\mathbf{y}$=$[y_1, \cdots, y_n]$

3.4 Least Squares

parameter: β, estimator: $\hat{\beta}$

predicted response given x: $\hat{\beta}_0 + \hat{\beta}_1 x = \hat{y}$

fitted value: $\hat{\beta}_0 + \hat{\beta}_1 x_i = \hat{y}_i \forall i$

3.4.1 The least squares line of best fit

Select $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the sum of squared errors $\sum_{i=1}^{n} \hat{e}_i^2$

$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{y}_1 \\ \cdots \\ \cdots \\ \hat{y}_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ \cdots & \\ \cdots & \\ 1 & x_n \end{bmatrix}, \hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}$$

- $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$
- $\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\beta}$

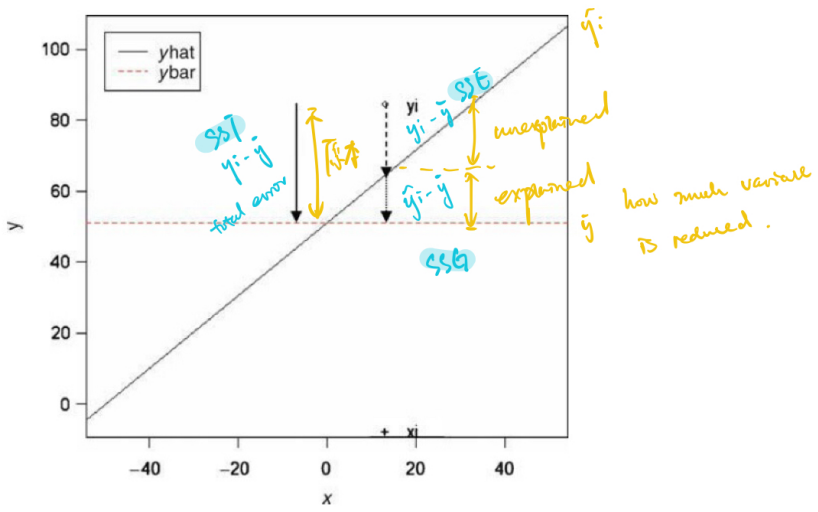
We wanna minimized $\sum_{i=1}^{n} e_i^2 = \mathbf{e}^T\mathbf{e}$ with respect to β, so we take partial deriavative.
Recall $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$

$$\begin{aligned} \frac{d}{d\hat{\beta}}(\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta}) &= 0 \\ \frac{d}{d\hat{\beta}}(y^Ty - y^TX\hat{\beta} - (X\hat{\beta})^Ty + (X\hat{\beta})^TX\hat{\beta}) &= 0 \\ \frac{d}{d\hat{\beta}}(y^Ty - 2\hat{\beta}^TX^Ty + \hat{\beta}^TX^TX\hat{\beta}) &= 0 \\ (0 - 2X^Ty + 2X^TX\hat{\beta}) &= 0 \\ 2X^Ty &= 2X^TX\hat{\beta} \\ X^Ty &= X^TX\hat{\beta} \\ \hat{\beta} &= (X^TX)^{-1}X^Ty \end{aligned}$$

- $\hat{y} = X\hat{\beta} = X(X^TX)^{-1}X^Ty$
- $X(X^TX)^{-1}X^T$ is often referred as the hat matrix H.
- $H^T = H, H^TH = H, HX = X, for M = H - I, M^T = M$

3.5 Sum of Squares revisited

If the predictor is **usefull**, we would like the predictor to explained parts of the variability and use it to produce better predictions.



We think of $\bar{y}$ as the base prediction.

For Anova, given the group/treatment t our prediction for an observation in group t is $\bar{y}_t$ R^2 is the coefficient of determination, is one diagonstic tools to check how good our model is:

$$R^2 = \frac{SSG}{SST} = 1 - \frac{SSE}{SST}, 0 \leq R^2 \leq 1$$

The closer R^2 is from 1, the better the fit is, the closer it is to 0 the worse the fit is.

In the case $y_i = \hat{y}_i$, SSG/SST=1

4 Lecture 4

4.1 The linear regression model

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

- Gauss-Markov Assumptions(conditions)
  - $E(e_i) = 0$
  - $Cor(e_i, e_j) = 0, i \neq j$
  - $Var(e_i) = \sigma^2 \leq \inf \forall i$
- The best linear unbiased estimator(BLUE) for β's are given by minimizing the mean square error
- Distribution implication:
  - $e_i \sim N(0, \sigma^2)$
  - $y_i | x_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$
  - Matrix notation $\mathbf{y}|\mathbf{x} \sim N(\mathbf{X}\beta, I\sigma^2)$

4.2 Maximum Likelihood

$$\hat{\theta}_{\text{MLE}} = \text{argmax } p_\theta(x_1, \cdots, x_n)$$

In our regression model, we have three parameters: $\beta_0, \beta_1, \text{and} \sigma$

Since $y_i|x_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$, then:

$$p_\theta(y_i|x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(y_i - (\beta_0 + \beta_1 x_i))^2}{2\sigma^2})$$

With matrix notation:

$$l(\theta|x_i) = -\frac{n}{2} log(2\pi) - nlog(\sigma) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta})$$

Maximizing this term with respect to β is equivalent to minimizing $(\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta})$:

$$\frac{d}{d\hat{\beta}}(\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta}) = 0$$

$$(X^T X)^{-1} X^T y = \hat{\beta}$$

Maximizing this ter with respect to σ:

$$\begin{aligned} 0 &= -\frac{n}{\sigma} log(2\pi) + \frac{1}{\sigma^3}(\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta}) \\ \implies \hat{\sigma}^2 &= \frac{(\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta})}{n} \\ &= \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n} \end{aligned}$$

which is the mean squared error and a biased estimator of $\hat{\sigma}^2$

4.3 Inference of Estimators

4.3.1 Inference of $\hat{\beta}$

Inference $y \sim N(X\beta, I\sigma^2) \implies \hat{\beta} \sim N(\beta, (X^TX)^{-1}\sigma^2)$.
No matter by minimize MSE or MLE,$\hat{\beta} = (X^TX)^{-1}X^Ty$:

$$\begin{aligned}
E[\hat{\beta}] &= E[(X^TX)^{-1}X^Ty] \\
&= (X^TX)^{-1}X^TE[y] \\
&= (X^TX)^{-1}X^TX\beta = \beta \\
Var[\hat{\beta}] &= Var[(X^TX)^{-1}X^Ty] \\
&= (X^TX)^{-1}X^TVar[y|X]((X^TX)^{-1}X^T)^T \\
&= (X^TX)^{-1}X^TI\sigma^2((X^TX)^{-1}X^T)^T \\
&= \sigma^2(X^TX)^{-1}X^T((X^TX)^{-1}X^T)^T \\
(AB)^T = B^TA^T &\implies ((X^TX)^{-1}X^T)^T = (X((X^TX)^{-1})^T) = (X((X^TX)^T)^{-1}) \\
Var[\hat{\beta}] &= \sigma^2(X^TX)^{-1}X^TX(X^TX)^{-1} \\
&= \sigma^2(X^TX)^{-1}
\end{aligned}$$

4.3.2 Inference of $\hat{\beta}_0$ and $\hat{\beta}_1$

Inference $\hat{\beta} \sim N(\beta, (X^TX)^{-1}\sigma^2) \implies \hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{SSX}) \vee \hat{\beta}_0 \sim N(\beta_0, \frac{\sigma^2}{SSX}\frac{1}{n}\sum_{i=1}^n x_i^2)$

$$\begin{aligned}
X^TX &= \begin{bmatrix} 1 & \cdots & \cdots & 1 \\ x_1 & \cdots & \cdots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ \cdots & \cdots \\ \cdots & \cdots \\ 1 & x_n \end{bmatrix} \\
&= \begin{bmatrix} \sum_{i=1}^n 1^2 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix} = n \begin{bmatrix} 1 & \bar{x} \\ \bar{x} & \frac{1}{n}\sum_{i=1}^n x_i^2 \end{bmatrix} \\
(X^TX)^{-1} &= \frac{1}{det} \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} \\
&= \frac{1}{n(1/n\sum_{i=1}^n (x_i^2 - (\bar{x})^2))} \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} \\
&= \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}
\end{aligned}$$

4.3.3 Hypo Test for $\hat{\beta}_1$

Since $\hat{\beta}_1 \sim N(\beta_1, \sigma^2 \frac{1}{SSX})$:

$$\frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{SSX}} \sim N(0,1) \text{ or } \frac{\hat{\beta}_1 - \beta_1}{s/\sqrt{SSX}} \sim t_{n-2}$$

$H_0 : \beta_1 = 0$ The parameter has no effect i.e x affect y

When σ is unknown, we estimate with $s^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)}{(n-1)}$

CI: $\hat{\beta}_1 \pm Z_{\alpha/2}\sigma/\sqrt{SSX}$ or $\hat{\beta}_1 \pm t_{(n-2,\alpha/2)}s/\sqrt{SSX}$

5 Lecture 5

5.1 Predictive inference

When there is a unobserved predictor x*, a simple prediction for the response could be $\hat{y}^* = \hat{\beta}_o + \hat{\beta}_1 x^*$.

We creat matrix of predictors by adding a column of 1s to get X*, and the vector of prediction $\hat{\mathbf{y}}^* = \hat{\mathbf{X}}^* \hat{\beta}$

5.1.1 Predictor distribution

For a given predictor X*

- $\hat{\beta}|X \sim N(\beta, (X^T X)^{-1}\sigma^2)$
- $y^* \sim N(\mu = \mathbf{X}^*\beta, \Sigma = \sigma^2 X^*(X^T X)^{-1} X^{*T})$
- $var(\hat{y}^*) = \sigma^2[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}]$

5.1.2 Confidence Interval

$$\text{CI} : \hat{y}_i^* \pm t_{(\frac{\alpha}{2}, df=n-2)}\hat{\sigma}\sqrt{X^*(X^T X)^{-1} X_{i,i}^{*T}},$$

which is the confidence interval for $\mathbf{X^*}\beta$.

i.e 95 percent chance that the prediction line would locate in this zone

5.1.3 Prediction Interval

$$y^* - \hat{y}^* \sim N(0, \sigma^2[I + X^*(X^T X)^{-1} X^{*T}])$$

$$var(Y^* - \hat{y}^*) = \sigma^2[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}]$$

Proof:

$$
\begin{aligned}
\mu &= E[X^*\beta + \epsilon - X^*\hat{\beta}] \\
&= X^*\beta + E[\epsilon] - E[X^*\hat{\beta}] \\
&= X^*\beta + 0 - X^*\beta = 0 \\
\Sigma &= Var[X^*\beta + \epsilon - X^*\hat{\beta}], \text{ Since } X^*\beta \text{ is constant parameter} \\
&= Var[\epsilon - X^*\hat{\beta}] \\
&= Var[\epsilon] + Var[X^*\hat{\beta}] + 2Cov[\epsilon, X^*\hat{\beta}] \\
&= \sigma^2 I + \sigma^2[X^*(X^TX)^{-1}X^{*T}] + 0 \\
&= \sigma^2[I + X^*(X^TX)^{-1}X^{*T}] \qquad \blacksquare
\end{aligned}
$$

Since the variance of prediction interval is the sum of variance of $X^*\beta$ and ϵ, so the confidence interval is always concluded in the predicted interval.

5.2 Check the Model Assumption

A good model is good at explaining the response - this is a useful model.
For a model to be a valid model, the assumption must be respected. i.e. the parameters shall be all significant(e.g the error term).

- **Error** Check whether the model follow the Gauss-Markov Assumptions and independently distributed according to $N(0, \sigma^2)$:
  - $E(e_i) = 0$
  - $Cor(e_i, e_j) = 0, i \neq j$
  - $Var(e_i) = \sigma^2 \leq \inf \forall i$
- **Model** We have assumed that the structural part of the model is $E(y_i) = \beta_0 + \beta_1 x_i$ or $E(Y|X) = X\beta$
- **i.i.d** i.i.d refers to independent and identical distributed. We assumed all data points are generated from the same distribution. There might be some unusual observations that does not fit the model and might change the choice and fit of the model.

5.2.1 Check error assumption

observed errors: $\hat{e}_i = y_i - \hat{y_i}$, in the form of matrix which is $\hat{\mathbf{e}} = (1 - H)\mathbf{y}$, with $Var(\hat{e}_i) = (I - H)_{i,i}\sigma^2$
Check Constant Variance

- **Residual against fitted** We expected there is no clear patter of the distribution of residuals around the fittest line.

- **Residual against predictor** To see whether the variance distribution remain constant as predictor varies. Also the pattern of residual plots reflect the pattern of actual observation in comparison with the fittest line

Check Normality

- **QQ-plot** Sample quantile against theoretical quantiles.Test whether the distribution of error is normal.
  - If the two distribution are the same we would expect a 45 degree straight line.
  - mild nonnormality can be safe to ignored since increase the size of sample may reduce the troublesome of nonnormality

Check Uncorrelatedness

- **standardized residuals plot** standardized residuals is calculated as $r_i = \frac{\hat{e}_i}{s\sqrt{(I-H)_{i,i}}}$ where $S = \sqrt{SSE/n-1}$ is the unbiased estimator for σ.

5.2.2 Unusual observations

An influential point is one whose removal from the data set would cause a larch change in the fit.

- **Leverage point:** Points that have a predictor value far from other points.
  - $h_i = H_{i,i} = \frac{1}{n} + \frac{(x_i-\bar{x})^2}{\sum_{j=1}^{n}(x_j-\bar{x})^2}$
  - $\sum_{i=1}^{n} h_i = 2$(number of parameters), then the average value of h is $2/n$. so leverage greater than 4/n should be looked at more closely.
- **Outlier:** point do not fit well within the model such like the residual $\hat{y}_i - y_i$ has a large value
- **Bad Leverage:** An outlier with a large leverage will definitely be an influential observation.

The cook distance

$$D_i = \frac{r_i^2}{2} * \frac{h_i}{1-h_i}$$

where r_i is the standardize residuals and h_i is the leverage.
There is a problem when

- $D_i \leq 4/n$ on large data sets
- $D_i \leq 1$ on small data sets
- D_i is separated by a large gap from the other D_js.

6 Lecture 6

6.1 Transformation

When there is a clear pattern, we need a new model.

6.1.1 Box-Cox

there is a family of possible transformation $g_\lambda(y)$

$$g_\lambda(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda} \text{ if } \lambda \neq 0 \\ log(y) \text{ if } \lambda = 0 \end{cases}$$

Log transformation reduces large values of y and tends to fix some non-constant variance issues.

6.2 Dummy variables

6.2.1 Example of model with 2 categorical variables

Group	X
A	0
B	1

This model we have $y \sim N(\beta_0 + \beta_1 x + e)$. Since x is a binary variable so this model implies that:

$$E(y) = \begin{cases} \beta_0 & \text{if x=0(Group=A)} \\ \beta_0 + \beta_1 & \text{if x=1(Group=B)} \end{cases}$$

This is a simple model consist of two different intercept and there is no slope.

Test related $\mu_A = \beta_0$ and $\mu_B = \beta_0 + \beta_1$, and $H_0 : \beta_1 = 0$ or $H_0 : \mu_A = \mu_B$

Since the response is normally distributed and we assumed a fixed variance, so this is exactly the two sample t-test. We can use dummy variable linear model to replace t test and anova.

T Test we can use dummy variable to indicate whether there is a difference between two group by telling whether the second parameter exist. To see how to replace an ANOVA test, see the following section

6.3 Multiple Linear Regression

There is a new model called$y = \beta_o + \beta_1 X_1 + \cdots + \beta_p x_p + e$ in a matrix form which is $y = \mathbf{X}\beta + e$ and $e \sim MVN(0, I\sigma^2)$(A vector of independent Normal Variables with mean 0 and variance σ^2).

6.3.1 Key parameters

$y \sim N(X\beta, I\sigma^2)$

$\hat{\beta} = (X^T X)^{-1} X^T y$

6.3.2 Example of model with 3 categorical variables

For a simple categorical variable representing three(or T) groups: A, B and C ,we can use a model with two(or T-1) binary variables $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$

Group	x_1	x_2
A	0	0
B	1	0
C	0	1

with $\mu_A = \beta_0, \mu_B = \beta_0 + \beta_1$ and $\mu_C = \beta_0 + \beta_2$

Test Related

$H_0 : \beta_1 = 0$ checks if Group B is different from group A;

$H_0 : \beta_2 = 0$ checks if Group C is different from Group A;

$H_0 : \beta_1 = \beta_2 = 0$ checks if all the groups are the same.

F Test and ANOVA we use as.factor() method to create categorical variables. In the anova table we check the within-group variacne and the between-group variance.

The linear regression model checks if $\beta_1 = \beta_2 = \cdots = \beta_p = 0$ which is the same as ANOVA $\frac{SSreg/p-1}{SSE/n-p}$

SST $= \sum_{i=1}^{n}(y_i - \bar{y})^2$

SSG=SSreg $= \sum_{i=1}^{n}(\hat{y} - \bar{y})^2$ which is the explained variation.

SSE $= \sum_{i=1}^{n}(\hat{y}_i - y_i)^2$ which is the unexplained variation.

6.3.3 Example of model with one categorical and numerical predictor

The model would be $y = \beta_0 + \beta_1 x_1 + \beta_2 X_2 + e$, $x_2$would be a dummy variable, and would move the intercept. When expected value would be $E(y) = \beta_0 + \beta_1 x$ or $E(y) = \beta_0 + \eta_1 x + \beta_2 = (\beta_0 + \beta_2) + \beta_1 x$ and the categorical predictor would move the intercept.

7 Lecture 7

7.1 Interaction

Interactions only make sense when we have multiples predictors.

Interaction is the effect of one predictor on the effect of the other predictor on y. i.e. if x_1 varies, the effect of x_2 on y is different or the relationship between is different.

By plot(lm(model)) we can see whether the model is valid, but to see whether parameters are significant, we need to check the t-test.

- **Interactions between two categorical predictors** actually each combination are a new category. e.g MBTI personality test.

- **Interaction between a cate and a num predictor** Implies model with different intercept and slopes. i.e different category have different distribution, which will make the predicted result based on such variable become less accurate(not a usable predictor).

An interaction is actually a product of effect:$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{1,2} x_1 x_2 + e$
For a fixed predictor, observation of group A would be $E(y_A) = \beta_0 + \beta_1 x$, and of group B would be $E(y_B) = \beta_0 + \beta_1 x + \beta_2 + \beta_{1,2} x = (\beta_0 + \beta_2) + (\beta_1 + \beta_{1,2})x$. x_1 is the numeric predictor while x_2 is a dummy variable that indicate the categorical predictor.

7.1.1 Interaction between two categorical predictors

In the model if we just simply add two variables together, we cannot make the difference between different choice of groups significant.To make the difference between each new category formed by two of the original category more significant, we add an interaction term

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{1,2} x_1 X_2 + e$$

$E(y\|A1, A2)$	β_0
$E(y\|B1, A2)$	$\beta_0 + \beta_1$
$E(y\|A1, B2)$	$\beta_0 + \beta_2$
$E(y\|B1, B2)$	$\beta_0 + \beta_1 + \beta_2 + \beta_{1,2}$

$\beta_{1,2}$ emphasize the change from A1 to B1.

Test related in R we have method interaction.plot(x1,x2,y) to see whether the changing in x2 value will change the predicting of y based on x1

7.2 Polynomial fit

i.e As x_1 varies, the effect of x_1 on y is different.
The model of prediction would be $y = \beta_0 + \beta_1 x_1 + \beta_{1,1} x_1 x_1 = \beta_0 + \beta_1 x_1 + \beta_{1,1} x_1^2$
When we see there is a pattern in the residual vs fitted plot, we might consider about adding polynomial predictor to reduce residual standard error and increase the R^2.

7.3 Model Checking

See lecture 5 for further in formation about checking error assumption, leverage points and outliers and as well as – Collinearity.

7.4 Collinearity

Usually we consider the new predicting variable is independent with old ones, however this is not an assumption of the model.We hope the new variable added contain different information with old ones, instead

of contain same or similar information. i.e $x_2 \implies x_1$

7.4.1 Collinearity fact reflected by matrix

- If a vector is a linear combination of otehr vectors of the matrix(i.e not independent), det($\mathbf{A}$)=0
- since the inverse of a matrix can be written as $\mathbf{A}^{-1} = \frac{1}{\det(\mathbf{A})} adj(\mathbf{A})$, when the determinant of a matrix is 0.
- $\mathbf{A}$= $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$, adj($\mathbf{A}$) = $\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$
- when two variables are **perfectly correlated**, $det(X) = det(X^T) = det(X^T X) = 0$ and $X^T X$ has no inverse, i.e $\hat{\beta} = (X^T X)^{-1} X^T y$ DNE.
- whent eh correlation is close to 1 or -1, $det(X^T X)$ would become closer and closer to 0 and $Var(\hat{\beta} = \sigma^2 (X^T X)^{-1}$ would become extremely large.

7.4.2 Collinearity checking in R

correlation matrix cor(data)
Eigen value of $X^T X$ eigen(t(X)%*%X) -¿v$values
R_i^2 summary(lm(X[,1] X[-1]))$r.squared

7.4.3 Variance Inflating factors(VIFs)

For the model

$$\mathbf{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p x_p + e$$

, we have:

$$Var(\hat{\beta}_j) = \frac{1}{1 - R_j^2} \frac{\sigma^2}{(n-1) SSX_j^2} \implies$$

correlation between predictors increase and variance between estimate of single parameter increase.
$\frac{1}{1-R_j^2}$ are defined as VIFs.
R code vif(X)

8 Lecture 8

8.1 BIG DATA

Key Challanges

- Data Storage
- Data Analysis
- Data Visualization
- Information Privacy

Big data usually includes data sets with sizes beyond the ability of commonly used software tools. We can define multiple axis for which a data set can considered large:

- Volume/Tall data: Large number of observations(large n)
- Wide data: Large number of predictors(large p)
- Variety: Multiple styles of data from texts to images to audio and video files
- Velocity: The speed at which the data is generated.

Big data usually includes data sets with sizes beyond the ability of commonly used software tools.
5V features: Volumn, Velocity, variaty, value and Veracity.
To be considered large, the data must have features including :Large volumn, Wide data(large number of predictors), variety,velocity.

8.1.1 Large number of predictors(Large p)

Problems

- **Reduce interpretation**
  - Occam's Razor – simplest is the best
  - in order to get the big picture, we are willing ot sacrifice small details
- **Increase the variance of the estimates** reflect by reducing the denominator of calculating sample variance $s^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i^2)/(n-(p+1))$
- **trend to be overfitting** Fitting the training set too much might not be able to provide a better view for the entire population.
- **trend to cause collinearity issues** Collinearity is caused by having too many variable s providing similar information.

8.2 Overfitting

We say a model overfits when it offers poor generalization abilities. Even with when the degree of polynomial is high(M=9) and all points are captured by the graph, but the relationship would be quiet different with the true distribution. To make a better model, we might sacrifice some details.

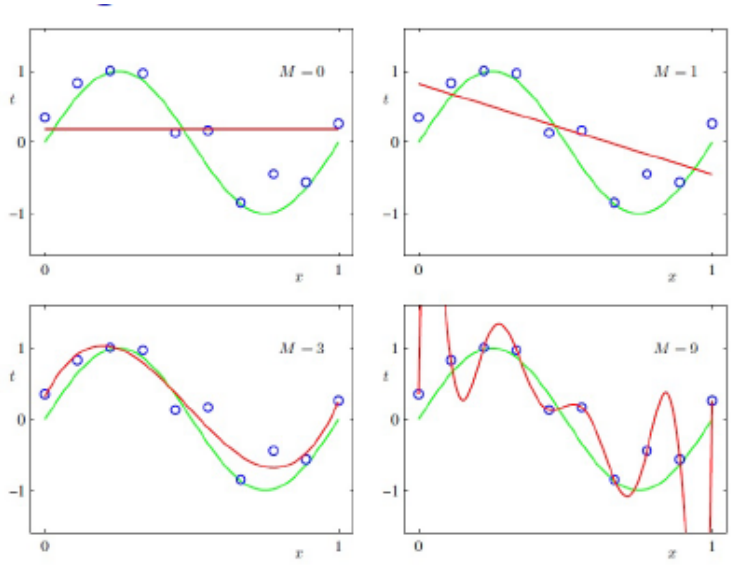


Figure 1: Allowing polynomial terms of high order can cause overfitting

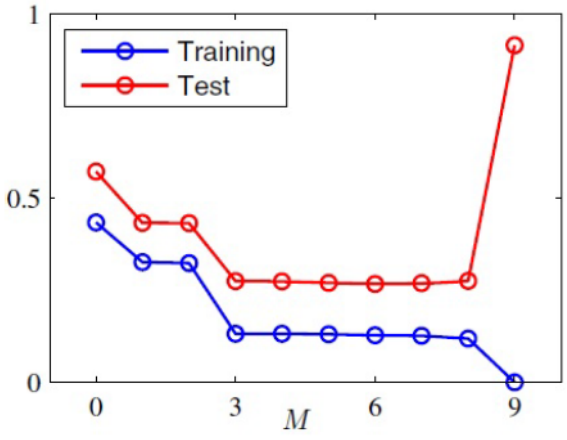
Figure 1: Greenline: population relation;Redline: modelled relation

8.2.1 Training set and test set

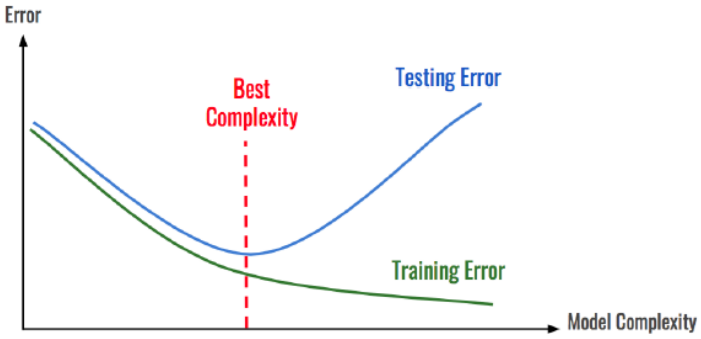
the training set is used for training to get our estimated parameters $\hat{\beta}$ by predictors $\mathbf{X}$ and response $\mathbf{y}$

We say a model overfits if it has been fitted to have too good performances on the training set to the detriment of test set performance.

The mean squared error or maximum likelihood are examples of reasongable matrix.



Cases in the above model



Genral case

8.3 Variable Selection

We should tempted to select the set of variables that maximizes the likelihood or the R^2 coefficient but this can only lead to overfits.

The motivation behind the commonly used matrices is to use the R^2 coefficient or log-likelihood but to penalizes for high number of parameters.

8.3.1 Adjusted R^2

$$R^2 = \frac{SSreg}{SST} = 1 - \frac{SSE}{SST}$$

R^2 adjusted includes a penalty per parameters

$$\text{Adjusted}R^2 = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$$

A large Adjusted R^2 indicates a good improvement over $\hat{y}$.

8.3.2 AIC and BIC to select the amount of parameters

The Akaike Information Criterion and Bayesian Information Criterion are likelihood-based metrics with a penalty for the number of parameters and the smaller the better.

$$AIC = 2p - 2l$$

$$BIC = log(n)p - 2l$$

Where p is the number of parameter and l the log-likelihood of the current model.

BIC has a stronger penalty for the number of parameter($n > 7 \implies log(n) > 2$).

Low AIC a model is considered to be closer to the truth.

Low BIC a model is considered to be more likely to be the true model.
Sometimes it lead to using BIC select the underdeveloped model and AIC select the overfitted model.i.e.

$$p \in [p_{\text{minBIC}}, p_{\text{minAIC}}]$$

Select the amount of predictors that has the highest adjusted R^2 of lowest AIC/BIC.

8.3.3 How to establish the list of model to compare

Traditionally, we try all the models.
When there is p predictors, it would lead to 2^p different models. Then fit all of the 2^p models and select the one with the highest adjusted R^2 of lowest AIC/BIC.
There is some logic using during the comparison to determind the order of comparison:

- **Hierarchical Models** e.g:Polynomial fits: We can increase the order of the model as long as
  - adjusted R^2 increasing
  - AIC or BIC decreasing
  - adding term is significant
- **Stepwise subsets**
  - **Forward Selection** starts with the simplest model and sequentially add predictors to the model.
    - add the one with smallest p-value from not-selected
    - stop when the R^2 decrease or AIC/BIC increase when all parameters selected are added.
  - **Backward Elimination** starts with all of the possible predictors
    - take out the one with biggest p-value
    - stop when the R^2 decrease or AIC/BIC increase when all parameters selected are added.
  - **Pros** easy to use, intuitive and easy computed
  - **Cons** p-value not used properly, since we test sequentially, we might stop before finding the best model and the selection procedure disturbs inference and prediction.

8.3.4 Post-selection inference

The selection process changes the properties of the estimators as wel as the standard inferential procedures such as tests and confidence intervals. The regression coefficients obtained after variable selection are biased.i.e.Before we actually doing the analysis, if we have a assumption on the distribution would cause highly misleading of our analysis.
Cause The sampling properties of post-model-selection estimators are typically significantly different from the nominal distributions that arise if a fixed model is supposed.
Solution computing the conditional distribution of the parameters.

9 Lecture 9

9.1 Principal Component Analysis(PCA)

9.1.1 Motivation

Principal components are a reparametrization of the current system in order to create uncorrelated predictors. We project more than one predictors on a lower dimensional place that preserves as much variability as possible.

Features

- solve the collinearity
- rich in information
- reduce the number of variable

9.1.2 A matrix algebra problem

We would like to do a projection that keeps observations as distinguishable as possible(i.e for x_1 and x_2 we make all $x_2 = 0$). So we project the preictors onto the axis that maximize the variance of new vectors(make the most effect of reduce $x_2 = 0$).

Defining $\mathbf{S}$ as the observed covariance matrix:

$$\mathbf{S} = \begin{bmatrix} \sum_{i=1}^{n}(x_{i,1} - \bar{x}_1)^2 & \cdots & \sum_{i=1}^{n}(x_{i,1} - \bar{x}_1)(x_{i,p} - \bar{X}_p) \\ \cdots & \cdots & \cdots \\ \sum_{i=1}^{n}(x_{i,p} - \bar{x}_p)(x_{i,1} - \bar{x}_1) & \cdots & \sum_{i=1}^{n}(x_{i,p} - \bar{x}_p)^2 \end{bmatrix}$$

- $\mathbf{z}$=$\mathbf{Xu}$, $\mathbf{Z} \in \mathbb{R}^m, m < p$ and $\mathbf{U}_{p\times 1}$ is the projection vector.
- $Var(\mathbf{z}) = \mathbf{u}^T\mathbf{Su}$
- we want $\mathbf{u}$ to be a direction in the original predictor space(in stead of creating new variable while reducing dimension), so $\mathbf{u}$ is a vector of norm 1 and $\mathbf{u}^T\mathbf{u} = 1$
- we wanna maximize the variance subject to given norm, then we apply the Lagrange multiplier and get

$$\mathbf{u}^T\mathbf{Su} + \lambda(1 - \mathbf{u}^T\mathbf{u})$$

- $\mathbf{Su}$=$\lambda\mathbf{u}$
- λ is an eigenvalue of $\mathbf{S}$ and $\mathbf{u}$ an eigenvector of $\mathbf{S}$
- $\mathbf{u}^T\mathbf{Su} = \lambda$ and thus λ is the variance of the projected data.
- to maximize the variance, we select $\mathbf{u}$ as the eigenvector associated with the largest eigenvalue.

Generalization We can project the predictor matrix $\mathbf{X}$ on a lower dimension orthogonal space $\mathbf{Z}$ of size $m < p$ using a projection matrix $\mathbf{U}_{p\times m}$.

- $\mathbf{Z}_{n\times m} = \mathbf{X}_{n\times p}\mathbf{U}_{p\times m}$
- $\mathbf{U}$ consist of eigenvectors associated with the m largest eigenvalues of the data correlation matrix $\mathbf{S}$.
- fit a linear model using: $\mathbf{Z} : \mathbf{y} = \mathbf{Z}\beta + \mathbf{e}$.
- we have lower number of predictors and they are all uncorrelated (i.e we reduce one of the correlated predictor by include its information in to the variance of its correlated one).
- however then it is really hard to explain what each $z_i \in \mathbf{Z}$ is

9.2 Ridge Regression

9.2.1 Motivation

Beside we control the number of predictors by PCA, we can also control the size of parameters by Ridge regression.
$\bar{y}$ does not **overfit** since $\forall i \in \{1, \cdots, p\}, \beta_i = 0 \implies \bar{y} = \beta_0$.
Intuitively, if the β's are all small they can not affect y too much and it minimize the chances of overfitting.
Techniques Shrinkage techniques and ridge regression.

9.2.2 Ridge Regression Estimators

When $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$,the observed error is $\hat{\mathbf{e}} = (\mathbf{y} - \mathbf{X}\hat{\beta})$ and the MSE is $(\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta})$.Our parameters are the solution of minimization of MSE, as well as force to find small $\hat{\beta}$. So we minimize MSE subject to minimize $\sum_{i=1}^{n} \beta_i^2 = \hat{\beta}^T\hat{\beta}$.
Thus we can establish $\hat{\beta}_{ridge}$ as the solution of minimization of:

$$(\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta}) + \lambda\hat{\beta}^T\hat{\beta}$$

where λ is an hyper-parameter controlling the penalty on $\sum_{i=1}^{p} \hat{\beta}_i^2$.
If $\lambda = 0$ we have $\hat{\beta} = (X^TX)^{-1}X^Ty$ which is the case of overfit and if $\lambda \to \infty$ then all β's goes to 0 and we have a model that is predicted only by $\bar{y}$ which is not fitted.
By Lagrance Multiplier, $(\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta}) + \lambda\hat{\beta}^T\hat{\beta}$ indicate the net work done by the two vectors, and

instead make it equals to 0, we are targeted to find the $\hat{\beta}$ that make it smallest.

$$\begin{aligned}
\frac{d}{d\hat{\beta}}(\mathbf{y}-\mathbf{X}\hat{\beta})^T(\mathbf{y}-\mathbf{X}\hat{\beta})+\lambda\hat{\beta}^T\hat{\beta}&=0\\
-2\mathbf{X}^T\mathbf{y}+2\mathbf{X}^T\mathbf{X}\hat{\beta}+2\lambda\hat{\beta}&=0\\
2\mathbf{X}^T\mathbf{y}&=2\mathbf{X}^T\mathbf{X}\hat{\beta}+2\lambda\hat{\beta}\\
\mathbf{X}^T\mathbf{y}&=(\mathbf{X}^T\mathbf{X}\hat{\beta}+\lambda I)\hat{\beta}\\
\hat{\beta}_{ridge}&=(\mathbf{X}^T\mathbf{X}+\lambda I)^{-1}\mathbf{X}^T\mathbf{y}
\end{aligned}$$

9.2.3 Effect

- penalty term is convexity $\implies$ there is a solution
- protect from overfitting by reduce the effect of parameters
- prevent collinearity since $X^TX+\lambda I$ is always invertible and λI ensure the determinant is not too small.

Conclude to a constrained optimization problem:

$$\hat{\beta}_{ridge}=\text{argmin}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\text{argmin}\sum_{i=1}^{n}(y_i-(\beta_0+(\sum_{j=1}^{p}\beta_j x_{i,j})))^2$$

$$\text{subject to}\sum_{j=1}^{p}\beta_j^2\leq t$$

Since t and λ is one-to-one, we now partition the data set into three pieces: A training set, a validation set, and a test set.

- Training set : $\hat{\beta}_{ridge}=(\mathbf{X}^T\mathbf{X}+\lambda I)^{-1}\mathbf{X}^T\mathbf{y}$
- Validation set: select λ
- Observation(test set) to test the model

9.2.4 Graph Info

- horizontal axis: $df(\lambda)$
- vertical axis: coefficient of each parameters
- vertical dotted line: $df(\lambda)$ selected by validation set
- right-most: $\hat{\beta}_{\text{LSE}}$

9.3 Lasso Regression

9.3.1 Motivation

LASSO Least Absolute Shrinkage and Selection Operator

Enforce sparsity using a penalty on **parameters** i.e. the ability to fit the model that needs a reduced amount of parameters.

9.3.2 Lasso regression Estimators

we want to minimize:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \sum_{i=1}^{n}(y_i - (\beta_0 + (\sum_{j=1}^{p}\beta_j x_{i,j})))^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$
$$= (\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta}) + \lambda\sum_{j=1}^{p}|\beta_j|$$

In Lasso we use the $\mathcal{L}_1$ norm and in Ridge we use $\mathcal{L}_2$ norm.

$\mathcal{L}_p = ||\beta||_p = (\sum_{i=1}^{n}|\beta|^p)^{\frac{1}{p}} \implies \mathcal{L}_1 = \sum_{j=1}^{p}|\beta_j|$

Expressed as a constrained optimization problem:

$$\hat{\beta}_{\text{Lasso}} = \text{argmin}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \text{argmin}\sum_{i=1}^{n}(y_i - (\beta_0 + (\sum_{j=1}^{p}\beta_j x_{i,j})))^2$$
$$\text{subject to}\sum_{j=1}^{p}|\beta_j| \leq t$$

Problem

- make the solutions nonlinear
- no closed form expression for the solution

Result making t small will actually cause some parameters to be exactly zero.

Difference with Ridge As t gets smaller and smaller all the parameter shrink in parallel while in Lasso some goes to 0 before others.

9.4 Elastic net

The constraint can be expressed as $\sum_{j=1}^{p}|\beta_j|^q$ where q=1 for Lasso and q=2 for Ridge.

If we change the value of q, we will change the contours of the constraint surface.