Methods of Data Analysis – STA302 Course Notes

Yuchen Wang

January 22, 2020

Contents

1 May 7th - Introduction, p-values and statistical significance

Definition 1.1 - Statistical Analysis Data Analysis that relies on Probability theory to account for the variability of the data.

Permutation Test 1.2 Insert random premise, observe two samples Group A and Group B.
If the groups have no effect, all of the permutations are equally likely.
We can plot the Permutation Distribution with respect to difference between sample means.

Characteristics of Permutation Test 1.3

1. Involves simple probability theory
2. distribution-free
3. listing all the permutation for large dataset is almost impossible

Definition 1.4 - Statistical Significance We say a difference is **statistically significant** if it's less probable than our pre-determined significance level. (when p-value $p <$ significance level α)

Definition 1.5 - Significant Effect We say the groups have a **significant effect** if it causes the variable of interest to be significantly different.

2 May 9th - Hypothesis testing, t-test and ANOVA

2.1 The basics

Fact 2.1.1 If H_0 is true, the p-value $\sim U(0,1)$

remarks: This is saying that if p-value is greater than significance level, then it does not say anything about our confidence, it's just a value. Proof can be found online.

Tradeoff Between Type I and Type II Error It's common to fix α (significance level or type-I error) and minimize type-II error.

2.2 t-test

Under the model $(y_i|X = xi) \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$, an unbiased estimator of σ^2 is

$$S^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$$

Then

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{S^2}{SXX}}} \sim t_{(n-2)}$$

2.3 One-way Analysis of variance (ANOVA)

Suppose the response Y is quantitative and the predictor X is categorical, taking t values or levels denoted $1, \ldots, t$. With the regression model, we assume that the only aspect of the conditional distribution of Y, given $X = x$, that changes as x changes, is the mean.
Suppose we are interested in assessing whether or not there is a relationship between the response and the predictor. There is no relationship if and only if all the conditional distributions are the same. This is true under our assumptions if and only if all the means are equal. In our case, one-way ANOVA is an extension of the t-test to 3 or more samples focus analysis on group differences.
H_0: All groups are the same. If groups are different, we expect there is a bigger difference between groups (the group effect) than within groups (natural variability of the data).

Basic Definitions Suppose we have T groups and n_t observations for the t-th group, and we denote each observation as y.

1. SST: This is the sum of the squared deviations between each observation and the overall mean:

$$SST = \sum_{t=1}^{T}\sum_{i=1}^{n_t}(y_{i,t} - \bar{y})^2$$

2. SSE: This is the sum of the squared deviations between each observation and the mean of the group to which it belongs:

$$SSE = \sum_{t=1}^{T}\sum_{i=1}^{n_t}(y_{i,t} - \bar{y_t})^2$$

3. SSG: This is the sum of the squared deviations between each group mean and the overall mean:

$$SSG = \sum_{t=1}^{T} \sum_{i=1}^{n_t} (\bar{y}_t - \bar{y})^2$$

Sum of Squares Decomposition Total sum of squares = Within group sum of squares + Between group sum of squares

$$\sum_{t=1}^{T} \sum_{i=1}^{n_t} (y_{i,t} - \bar{y})^2 = \sum_{t=1}^{T} \sum_{i=1}^{n_t} (y_{i,t} - \bar{y}_t)^2 + \sum_{t=1}^{T} \sum_{i=1}^{n_t} (\bar{y}_t - \bar{y})^2$$

In shorthand:

$$SST = SSE + SSG$$

proof:
add $-\bar{y}_t + \bar{y}_t$ inside the squared error term and everything is just like a short proof in STA261, nothing interesting. ■

ANOVA We want to assess how large is SSG relative to SSE, but it would be hard to establish a distribution for SSG/SSE. Knowing a sum of squares divided by its degrees of freedom has a chi-square distribution, we can conclude that

$$SSG/(T-1) \sim \chi^2_{T-1}, \quad SSE/(n-T) \sim \chi^2_{n-T}$$

Theorem 2.2.1 If between-groups and within-groups variance are equal ($\sigma_T = \sigma_\varepsilon$), then $\frac{SSG/(T-1)}{SSE/(n-T)} \sim F_{T-1,n-T}$

proof:
In STA261, we've proven that if $\sigma_x^2 = \sigma_y^2$, then $\frac{\hat{\sigma}_x^2}{\hat{\sigma}_y^2} \sim F_{n-1,n-1}$.
Since SSG/(T-1) is an estimation for the variation between groups (σ_T) and $SSE/(n-T)$ is an estimation for the variation within groups (σ_ε), then the result follows. ■

Remarks 2.2.2 Thus a small p-value indicates theses variances are different, which is evidence for the existence of some group effect.

Theorem 2.2.3 One-way ANOVA Table if p-value $< \alpha$, we reject H_0: groups have no effect.

Source	Sum of Squares	df	Mean Squares	Test Statistic
Between	SSG	$T-1$	$MSG = \frac{SSG}{T-1}$	$F = \frac{MSG}{MSE}$
Within	SSE	$n-T$	$MSE = \frac{SSE}{n-T}$	
Total	SST	$n-1$		

3 May 14th - Linear Regression: Least Square Error Formulation

3.1 Matrix Notation

$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ is a random variable.

In addition, $A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix}$, $\mathbf{c} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix}$.

Then

$$E[\mathbf{x}] = \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$

$$Var[\mathbf{x}] = \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12}^2 \\ \sigma_{12}^2 & \sigma_2^2 \end{bmatrix}$$

where $\sigma_{ij}^2 = cov(x_i, x_j)$.

Theorem 3.1.1 Let $\mathbf{z} = A\mathbf{x} + \mathbf{c}$. Then

$$E[\mathbf{z}] = A\mu + \mathbf{c}$$

$$Var[\mathbf{z}] = A\Sigma A^T$$

3.2 Linear Regression

We have a vector of n predictors $\mathbf{x} = [x_1, \ldots, x_n]$, as well as n associated response variables $\mathbf{y} = [y_1, \ldots, y_n]$. We want to estimate the parameters β_0 and β_1 that best fit the model $y = \beta_0 + \beta_1 x$. (In matrix notation: $\mathbf{y} = X\beta$ where $\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$).

3.2.1 Least Square Estimation

Minimize sum of squared errors:

$$\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i - y_i)^2$$

We take derivative of $\sum_{i=1}^{n}(\mathbf{y} - X\hat{\beta})^2$ wrt $\hat{\beta}$, set this to 0 and get

Theorem 3.2.1.1

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

Remark 3.2.1.2 $X(X^T X)^{-1} X^T$ is called the hat matrix since it puts the hat on $\mathbf{y}$. This matrix (H) has the following properties:

1. $H^T = H$
2. $HH = H$
3. $HX = X$

3.2.2 ANOVA

Estimate how good a linear regression model is.

Basic Definitions $\bar{y}$ is called base estimation.

1. SST: This is the sum of the squared deviations between each observation and the mean:
$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$$
2. SSE: This is the sum of the squared deviations between each observation and the corresponding prediction
$$SSE = \sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$
unexplained variation: How much our explanation is away from the true observation?

3. SSG: This is the sum of the squared deviations between each prediction and the mean.
$$SSG = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$
explained variation: How much our explanation takes us away from the base prediction?

Coefficient of Determination

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

$(0 \leq R^2 \leq 1)$
The closer R^2 is from 1, the better the fit is.

4 May 16th - Linear Regression: Maximum Likelihood Formulation

4.1 the Linear Regression Model

Definition 4.1.1 The best linear unbiased estimator (BLUE) is the unbiased estimator with the lowest variance.

Gauss-Markov Assumptions 4.1.2 If $E[e_i] = 0, Cov(e_i, e_j) = 0 \, \forall i \neq j$ and $Var(e_i) = \sigma^2 < \infty \, \forall i$, then the best linear unbiased estimator for β's are given by minimizing the MSE

the Linear Regression Model

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

where e_i is a random variable that represents the residual.

Assumptions

1. $e_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$ which follows the Gauss-Markov assumptions.
2. $(\mathbf{y}|\mathbf{x}) \sim N(X\beta, I\sigma^2)$ or $(Y|X = x) \sim N(\beta_0 + \beta_1 x, \sigma^2)$

4.2 Maximum Likelihood Estimation

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

Maximizing this term wrt β is equivalent to minimizing $(\mathbf{y}-X\beta)^T(\mathbf{y}-X\beta)$, which gives

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

Corollary 4.2.1 Minimizing MSE and the likelihood function leads to the same estimate $\hat{\beta}$.

A Biased Estimator of σ^2

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

Maximizing the likelihood function wrt to σ^2:

$$\hat{\sigma}^2 = \frac{(\mathbf{y} - X\hat{\beta})^T(\mathbf{y} - X\hat{\beta})}{n} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}$$

This is MLE, a biased estimator of σ^2.
The unbiased estimator of σ^2 is

$$\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2}$$

4.3 Inference

The action of extracting information about parameters given a dataset.

Mean and Variance of y Since $\mathbf{y} \sim N(X\beta, I\sigma^2$, then $E[\mathbf{y}] = X\beta$ and $Var[\mathbf{y}] = I\sigma^2$.

Mean and Variance of $\hat{\beta}$

$$\begin{aligned} E[\hat{\beta}] &= E[(X^TX)^{-1}X^T\mathbf{y}] \\ &= (X^TX)^{-1}X^TE[\mathbf{y}] \\ &= (X^TX)^{-1}X^TX\beta \\ &= \beta \end{aligned}$$

$\implies$ $\hat{\beta}$ is an unbiased estimator of β.

$$\begin{aligned}
Var[\hat{\beta}] &= Var[(X^TX)^{-1}X^T\mathbf{y}] \\
&= (X^TX)^{-1}X^T\, Var[\mathbf{y}|X]((X^TX)^{-1}X^T)^T \\
&= (X^TX)^{-1}X^T\, I\sigma^2((X^TX)^{-1}X^T)^T \\
&= \sigma^2(X^TX)^{-1}X^T((X^TX)^{-1}X^T)^T \\
&= \sigma^2(X^TX)^{-1}X^T(X((X^TX)^{-1})^T) \\
&= \sigma^2(X^TX)^{-1}X^T(X((X^TX)^T)^{-1}) \\
&= \sigma^2(X^TX)^{-1}X^TX(X^TX)^{-1} \\
&= \sigma^2(X^TX)^{-1}
\end{aligned}$$

Theorem 4.3.0.1

$$\hat{\beta} \sim N(\beta, (X^TX)^{-1}\sigma^2)$$

proof:
$\hat{\beta} = (X^TX)^{-1}X^T\mathbf{y}$, so $\hat{\beta}$ is a linear combination of normal r.v.'s(y_i's), therefore $\hat{\beta}$ follows normal distribution with mean and variance we have calculated. ■

4.3.1 Inference for β_1

$$\hat{\beta_1} \sim N(\beta_1, \frac{\sigma^2}{SSX})$$

where $SSX = \sum_{i=1}^{n}(x_i - \bar{x})^2$ Then

$$\frac{\hat{\beta_1} - \beta_1}{\sigma/\sqrt{SSX}} \sim N(0,1)$$

Theorem 4.3.1.1

$$\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{(n-2)}$$

where $S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$

Theorem 4.3.1.2

$$\frac{\hat{\beta_1} - \beta_1}{S/\sqrt{SSX}} \sim t_{n-2}$$

where $SSX = \sum_{i=1}^{n}(x_i - \bar{x})^2$

Model Checking H_0: $\beta_1 = 0$ Then under H_0, Theorem 4.3.1.2 applies.

1. If the p-value is small, then $\mathbf{y}$ and $\mathbf{x}$ are statistically significant.
2. 0.95 confidence level for β_1:

$$(\hat{\beta}_1 - t_{(n-2)(1-\frac{\alpha}{2})}\frac{S}{\sqrt{SSX}}, \hat{\beta}_1 + t_{(n-2)(1-\frac{\alpha}{2})}\frac{S}{\sqrt{SSX}})$$

4.3.2 Inference for β_0

$$\hat{\beta}_0 \sim N(\beta_0, \sigma^2\frac{\sum x_i^2}{n\,SSX})$$

5 May 23th - Diagnostic for the linear regression model

Review of the model

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

where $e_i \sim N(0, \sigma^2)$

$$\mathbf{y} = X\beta + \mathbf{e} \implies \mathbf{y}|X \sim N(X\beta, I\sigma^2)$$

where

$$\hat{\beta}_{MLE} = (X^TX)^{-1}X^T\mathbf{y}$$
$$\hat{\beta} \sim N(\beta, \sigma^2(X^TX)^{-1})$$

Notes In all cases, we have $\sum_{i=1}^n e_i = 0$.

5.1 Predictive Inference

Since $\hat{\mathbf{y}} = X\hat{\beta}$, then

$$\hat{\mathbf{y}} \sim N(X\beta, \sigma^2X(X^TX)^{-1}X^T)$$

Prediction For a new, unobserved predictor x^*, a simple prediction for the response could be

$$y^* = \beta_0 + \beta_1 x^*$$

Predictive Distribution We have $\hat{\beta} = [\hat{\beta}_0, \hat{\beta}_1]^T$, then

$$\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$$

Matrix notation:

$$\hat{\mathbf{y}}^* = X^* \hat{\beta}$$

where X^* is a vector of new observation $\mathbf{x}^*$ added a column of 1s.
then

$$\hat{\mathbf{y}}^* \sim (X^*\beta, \sigma^2 X^*(X^T X)^{-1} X^{*T})$$

Confidence Interval for $X^*\beta$

$$\frac{\hat{\mathbf{y}}^* - X^*\beta}{\sigma\sqrt{X^*(X^T X)^{-1} X^{*T}}} \sim N(0, I)$$

Then

$$0.95CI = \hat{y}_i^* \pm 1.96 * \sigma\sqrt{[X^*(X^T X)^{-1} X^{*T}]_{ii}}$$

Remarks The confidence interval reflects our uncertainty about the population regression line

the Prediction Error

$$y^* - \hat{y}^* = \beta_0 + \beta_1 x^* + \varepsilon^* - (\hat{\beta}_0 + \hat{\beta}_1 x^*)$$

Matrix notation:

$$\mathbf{y}^* - \hat{\mathbf{y}}^* = X^*\beta + \varepsilon - X^*\hat{\beta}$$

Distribution:

$$\mathbf{y}^* - \hat{\mathbf{y}}^* \sim N(\mu, \Sigma)$$

where

$$\begin{aligned}
\mu &= E[X^*\beta + \varepsilon - X^*\hat{\beta}] \\
&= X^*\beta + E[\varepsilon] - E[X^*\hat{\beta}] \\
&= X^*\beta + 0 - X^*\beta \\
&= 0 \\
\Sigma &= Var[X^*\beta + \varepsilon - X^*\hat{\beta}] \\
&= Var[\varepsilon - X^*\hat{\beta}] \\
&= Var[\varepsilon] + Var[X^*\hat{\beta}] + 2Cov[\varepsilon, X^*\hat{\beta}] \\
&= \sigma^2 I + \sigma^2 X^*(X^T X)^{-1} X^{*T} + 0 \\
&= \sigma^2[I + X^*(X^T X)^{-1} X^{*T}]
\end{aligned}$$

Then we can construct a CI for $\mathbf{y}^*$ using t-distribution ($df = n - 2$)

Remarks Prediction intervals reflect uncertainty from both $\hat{\beta}$ and ε (i.e. the irreducible error). The irreducible error is estimated using training sample.

Reducible and irreducible errors

1. Reducible error is the error arising from the mismatch between $\hat{f}$ and f. f is the true relationship between X and Y, but we can't see f directly - we can only estimate it. We can reduce the gap between our estimate and the true function by applying improved methods.
2. Irreducible error arises from the fact that X doesn't completely determine Y. That is, there are variables outside of X - and independent of X - that still have some small effect on Y. The only way to improve prediction error related to irreducible error is to identify these outside influences and incorporate them as predictors.

5.2 Checking the Model Assumption

We will divide model checking into 3 pieces:

1. Error assumption ($e_i \overset{i.i.d}{\sim} N(0, \sigma^2)$)
2. Identical distribution (checking for unexpected observations)
3. Model assumption (linearity)

5.2.1 Checking Error Assumption

We only have access to the residuals (observed errors)

$$\hat{e}_i = y_i - \hat{y}_i$$

$$\hat{\mathbf{e}} = (I - H)\mathbf{y}$$

Constant variance Check $Var(e_i) = \sigma^2 \, \forall i$
Plot the residuals against the fitted values($\hat{y}_i$)

Normality of residuals Quantile to Quantile plot (QQplot)

Uncorrelatedness / Independence

1. Scatter plot (y against x)
2. Residual plot against predictors (see if clustered around zero and looks random)
3. Residual Sequence plot

Remarks We rarely check for this assumption

5.2.2 Unusual Observations

Definition 5.2.2.1 - Leverage points A leverage point is a point whose x-value is distant from the other x-values

leverages In the linear regression model, the leverage for the i-th observation is defines as:

$$h_i = H_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$$

where $H = X(X^T X)^{-1} X^T$
Property: $\sum_{i=1}^{n} h_i = 2$ (number of parameters)

Remarks The average value for h is $2/n$. Usually leverages larger than $4/n$ should be looked at more closely.

Definition 5.2.2.2 - Outliers An outlier is a data point whose y-value differs significantly from other observations.

Remarks Usually large residual $\hat{y}_i - y_i$ might indicate outliers

Definition 5.2.2.3 - Influential observations An influential point is one whose removal from the dataset would cause a large change in the fit. They could be leverage points, outliers but usually the both.

Remarks An outlier with a large leverage will definitely be an influential observation. It is sometimes called a bad leverage

Definition 5.2.2.4 - Cook's Distance For observation i, the Cook's distance is

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

where r_i is the standardized residual and h_i is the leverage.

Remarks r_i measures the extent of outlying, h_i measures the leverage. Thus, a large value of D_i indicates influential observations.

Simple rules of thumb There is a problem when

1. $D_i > 4/n$ on large datasets
2. $D_i > 1$ on small datasets
3. D_i is separated by a large gap from the other D_js

6 May 28th - Dummy variables and introduction to multiple linear regression

6.1 Transformation

We can use transformations to fix 2 problems:

1. Non-constant variance
2. Non-linearity

When the variance is exploding, typically we raise y to a power between 0 and 1 or apply a logarithmic transformation.

Logarithmic Transformation

$$\log(y_i) = \beta_0 + \beta_1 x_i + e_i$$

$$y_i = \exp(\beta_0) \exp(\beta_1 x_i) \exp(e_i)$$

Box-Cox Transformation Let's consider a family of possible transformations $g_\lambda(y)$

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

λ is selected by the model achieving the highest log-likelihood:

$$g_\lambda(y_i) \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

$$l(\lambda|\mathbf{y}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(g_\lambda(y_i) - \beta_0 - \beta_1 x_i)^2$$

Conclusions

1. Transforming the response (or the predictors) might help with violated assumptions
2. It makes the model less interpretable.

6.2 Multiple Linear Regression

Suppose we have p predictors.

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + e$$

where $e \sim N(0, \sigma^2)$

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1p} \\ \vdots & \vdots & \ldots & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

Matrix notation:

$$\mathbf{y} = X\beta + \mathbf{e}$$

where $\mathbf{e} \sim N(0, I\sigma^2)$ (a vector of independent normal variables) Therefore,

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

Inference

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

which leads to

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

6.3 Dummy variables

A set of binary variables to represent a categorical variable. Allows to fit a parameter for every possible categories of a categorical variable.

Model - 2 groups Fit a linear regression:

$$y = \beta_0 + \beta_1 x + e$$

where $e \sim N(0, \sigma^2)$

$$y \sim N(\beta_0 + \beta_1 x, \sigma^2)$$

This is a model that consists only two different intercepts, no slope.

Inference

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

$$H_0 : \beta_1 = 0$$

Applying t-test, this is exactly same to lecture 2.

Model - 3 groups

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$$

where $e \sim N(0, \sigma^2)$

Inference

$$E(y) = \begin{cases} \beta_0 & \text{for Group A} \\ \beta_0 + \beta_1 & \text{for Group B} \\ \beta_0 + \beta_2 & \text{for Group C} \end{cases}$$

$H_0 : \beta_1 = \beta_2 = 0$ for testing if all groups are the same; $H_0 : \beta_1 = 0$ for testing if Group B is same as Group A.

7 May 30th - Interactions and multiple linear regression assumptions

7.1 Interactions

Definition Interaction is the effect of predictor x_1 on the effect of predictor x_2 on y. With the interaction term, the linear regression model becomes

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

where $e \sim N(0,1)$

Remarks As x_1 varies,

1. the effect of x_2 on y is different.
2. the relationship between x_2 and y is different.

Interaction between two categorical predictors $\implies$ every combination are categories with respective effects
Interaction between a categorical predictor and a numerical predictor $\implies$ different intercepts and different slopes

Example: numerical and categorical Suppose x_1 is numerical, x_2 is categorical.
Then for a fixed x_1,

$$E(y) = \begin{cases} \beta_0 + \beta_1 x_1 & \text{if } x_2 = 0 \\ (\beta_0 + \beta_2) + (\beta_1 + \beta_{1,2} x_2) & \text{if } x_2 = 1 \end{cases}$$

Now β_2 indicates difference in intercept and $\beta_{1,2}$ indicates difference in slope (interaction)

Example: categorical and categorical $\beta_{1,2}$ allow for a different effect from change of predictor 1 depending on the value of predictor 2.

Example: numerical and numerical More complicated.
The slope and intercept of x_1 is different across different values of x_2.

Remarks The model is completely wrong without the interaction term.But interaction terms make the number of parameters explode quickly. (p predictors $\implies$ $\binom{p}{2}$ interaction terms)

7.2 Polynomial fit

1. Can be understood as a special case of interactions.
2. Can also solve the issue of observable pattern in the residuals

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

7.3 Model Checking

7.3.1 Collinearity

Correlation

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

range: $[-1, 1]$.
when $\rho = 1/-1$, we call it "perfect correlation".

Theorem $\rho = \pm 1$ iff $P(Y = a + bX) = 1$ for some constants a and b.
proof:
Rice p143.

Collinearity If two variables are perfectly correlated, it implies a perfect increasing(decreasing) linear relationship ($x_i = a + bx_j$). This implies that $\det(X) = 0 = \det(X^T X) = 0$.
$\implies$ $X^T X$ not invertible
$\implies$ $\hat{\beta} = (X^T X)^{-1} X^T \mathbf{y}$ does not exist
$\implies$ perfect correlation should not happen.
We also do not want a correlation close to -1 or 1, since the variance of $\hat{\beta}$ would be extremely large.

Checking for collinearity

1. Look at the correlation matrix of the predictors. Large pairwise correlation indicates a problem.
2. Build a regression model for x_i as response on all other predictors to assess R_i^2. High R_i^2 (close to 1) indicates a problem.
3. Look at the eigenvalues of $X^T X$. Small eigenvalues indicates a problem.

Variance Inflating Factor (VIF)

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

$\frac{1}{1-R_j^2}$ are defined as the VIF.

Remarks If the variable X_j is uncorrelated then VIF = 1

8 June 6th - Model selection and variable selection

8.1 Big Data

Properties Big data usually includes datasets with sizes beyond the ability of commonly used software tools.

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

Challenges

1. Data storage
2. Data analysis
3. Data visualization
4. Information privacy

8.1.1 Large number of predictors

We have a large number of predictors when we

1. have a large number of parameters
2. want to investigate many interaction terms and interaction of high order
3. want to add multiple polynomial terms

Why is it a problem ?

1. It reduces interpretation
   "The simplest is best". it is easier to explain a simpler model. In order to get the big picture, we are willing to sacrifice small details.

2. It increases the variance of the estimates

$$s^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 / (n - (p+1))$$

3. It is more prone to overfitting.
   More parameters increase the detriment of generalization abilities.

4. It increases the chances of collinearity issues. Too many similar information.

8.2 Overfitting

Definition 8.2.1 Overfitting happens when we detect a pattern in the data set that does not exist for new observations.

Definition 8.2.2 Poor performances of the model on non-observed points.

Remarks We say that a model overfits when it offers poor generalization abilities.

Performance Metric We can observe symptoms of overfitting by selecting a performance metric and comparing its value on the training set to its value on the test set as we change the model complexity.
Example:

1. Mean Squared Error (MSE)

2. Maximum Likelihood (MLE)

3. R^2 coefficient

4. log-likelihood

8.3 Variable Selection

We could select the set of variables that maximizes the likelihood or the R^2 coefficient. But to prevent overfitting, we penalizes for high number of parameters.

Adjusted R^2 Recall that

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

R^2-adjusted includes a penalty per parameter:

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

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

Akaike information criterion (AIC) The AIC is a likelihood-based metric with a penalty for the number of parameters.

$$AIC = 2p - 2I$$

where $p =$ number of parameters and $I =$ log-likelihood of the current model.

Remarks The smaller the AIC, the better the model is.

Bayesian information criterion (BIC) The BIC is also a likelihood-based metric with a penalty for the number of parameters.

$$BIC = p \log n - 2I$$

where $p =$ number of parameters and $I =$ log-likelihood of the current model.

Remarks The smaller the BIC, the better the model is.

AIC vs BIC

1. BIC has a stronger penalty for the number of parameter.
2. Sometimes BIC leads to selecting an underdeveloped model but AIC leads to selecting a model that overfits.

Variable Selection If we have p predictors, it would lead to 2^p different models. It might be reasonable to fit all of the 2^p models and select the one with the highest adjusted R^2 of lowest AIC/BIC.

Hierarchical models Start with only x, then include x^2 and compare the two models with one of our metrics. We can increase the order of the model as long as the adjusted R^2 (AIC/BIC) keeps increasing (decreasing) or as long as the added terms are significant.

Forward Selection Forward Selection is a stepwise subset technique that starts with the simplest model (no predictors) and sequentially add predictors to the model. At every step, for all predictors that are not in the model, we check their p-value if they were added to the model and add the predictor that would have the smallest p-value. We stop the process when R^2 (AIC/BIC) decreases (increases) or when all parameters were added to the model.

Backward Elimination Backward elimination start with all of the possible predictors. At every step, for all predictors in the model, we take out the predictor with the largest p-value. We stop the process when R^2 (AIC/BIC) decreases (increases) or when all parameters were taken out of the model.

Advantages These two technique share similar pros:

1. They are easy to use
2. They are intuitive
3. They are computationally cheap (they are both a lot cheaper than looking through all subsets)

Weaknesses

1. They don't use p-values appropriately
2. By testing model sequentially we might stop before finding the best model
3. The selection procedure disturbs inference and predicton

Post-selection inference The selection process changes the properties of the estimators as well as the standard inferential procedures (such as tests and confidence intervals). The regression coefficients obtained after variable selection are biased.

1. The p-values are usually much smaller.
2. And the t-statistics and F-statistic can be misleading.

9 June 11th - Ridge and Lasso regression

9.1 Principal Component Analysis (PCA)

A reparametrization of the current system in order to create uncorrelated predictors.
We project the predictors onto an orthogonal space.

Remarks

1. It completely solves the collinearity problem of any predictor set
2. Reduces number of variables to keep the variance low and chances of overfitting low
3. It makes the interpretation even harder.
4. This transformation is extremely simple and fast

Steps

1. Ideally we would like to do a projection that keeps observations as distinguishable as possible. We want to maximize the variance of the new vectors.
2. Define S as the observed covariance matrix:

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

3. With $Z = X\mathbf{u}$ where Z is our 1d space and $\mathbf{u}_{p\times 1}$ is the projection vector. We want $\mathbf{u}$ to be a direction in the original predictor space, so define $\mathbf{u}$ as a vector of norm 1.

4. Since the $Var(\mathbf{z}) = \mathbf{u}^T S\mathbf{u}$, we do the maximization problem using Lagrange multipliers with one constraint:

$$\begin{cases} \text{Maximize} & \mathbf{u}^T S\mathbf{u} \\ \text{Subject to} & \mathbf{u}^T\mathbf{u} = 1 \end{cases}$$

5. This leads to

$$S\mathbf{u} = \lambda\mathbf{u}$$

which implies that λ is an eigenvalue of S and $\mathbf{u}$ is an eigenvector of S.

6. Left-multiply $S\mathbf{u}$ by $\mathbf{u}^T$ we get $\mathbf{u}^T S\mathbf{u} = \lambda$, thus λ is the variance of the projected data.

7. To maximize the variance, we select $\mathbf{u}$ as the eigenvector associated with the largest eigenvalue.

Generalization To generalize the process, suppose we have p predictors. We can project the predictor matrix X on a lower dimension orthogonal space Z with dimension $m < p$ using a projection matrix $\mathbf{u}_{p\times m}$:

$$Z_{n\times m} = X_{n\times p}\mathbf{u}_{p\times m}$$

Problem

$$\begin{cases} \text{Maximize} & \mathbf{u}^T S\mathbf{u} \\ \text{Subject to} & \mathbf{u}^T\mathbf{u} = 1 \\ \text{and} & \mathbf{u}_i\text{'s are orthogonal} \end{cases}$$

The matrix $\mathbf{u}_{p\times m}$ consists eigenvectors associated with the m largest eigenvalues of the data correlation matrix S.
Then we can fit a linear model using $Z : \mathbf{y} = Z\beta + \mathbf{e}$ that we have lower number of predictors and they are all uncorrelated.

9.2 Ridge and Lasso Regression

9.2.1 Ridge Regression

A shrinkage technique: technique that reduce the size of the parameters.

Motivation We could prevent overfitting by controlling the size of the parameters.
$\bar{y}$ does not overfit. If $\beta_i = 0$ for $i \in \{1, \ldots, p\}$, then $\bar{y} = \beta_0$. Intuitively, if the β_i's are small, they can only affect the y so that it minimizes the chances of overfitting.

Review We established our $\hat{\beta}$ by minimizing sum of square error $(\mathbf{y} - X\hat{\beta})^T(\mathbf{y} - X\hat{\beta})$

Regularization One way to force small $\hat{\beta}$ would be to minimize

$$\sum_{i=1}^{n} \beta_i^2$$

Thus we can establish $\hat{\beta}_{ridge}$ as the solution of the minimization of

$$(\mathbf{y} - X\hat{\beta})^T(\mathbf{y} - X\hat{\beta}) + \lambda \sum_{i=1}^{n} \beta_i^2$$

where λ is a hyper-parameter controlling the penalty on $\sum_{i=1}^{n} \beta_i^2$.
If $\lambda = 0$ then we have our regular $\hat{\beta} = (X^T X)^{-1} X^T \mathbf{y}$; as $\lambda \to \infty$ then all of β_i's goes to 0 and we have a model that predicts $\bar{y}$.

A constrained optimization problem The ridge regression can be expressed as a constrained optimization problem. In fact:

$$\hat{\beta}_{ridge} = \begin{cases} \underset{\beta}{argmin} & \sum_{i=1}^{n} (y_i - (\beta_0 + (\sum_{j=1}^{p} \beta_j x_{i,j}))^2 \\ \text{Subject to} & \sum_{j=1}^{p} \beta_j^2 \leq t \end{cases}$$

Just know that there exist a one-to-one correspondence between λ and t, and that both of the formulations lead to the same solution.

Parameter tuning for λ Establish a huge list of possible λs. Then try them all and select one that best result on the validation set.

Remarks The model does NOT reduce the number of parameters, thus does not improve interpretability.

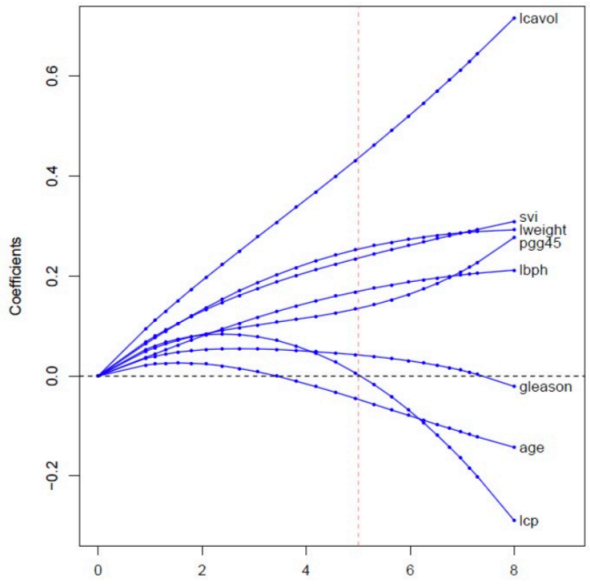

Figure 1: x-axis: $\frac{1}{\lambda}$. As the penalty decrease, the parameters converge to LSE.

9.2.2 Lasso Regression

Actually enforce sparsity: reduce some of the parameters. But this method is non-convex, so it is much more complicated to solve and we cannot compute the exact solution.

Regularization We want to minimize

$$\sum_{i=1}^{n}(y_i - (\beta_0 + \sum_{j=1}^{p}\beta_j x_{i,j}))^2 + \lambda \sum_{j=1}^{p}|\beta_j|$$

A constrained optimization problem

$$\hat{\beta}_{lasso} = \begin{cases} \underset{\beta}{argmin} & \sum_{i=1}^{n}(y_i - (\beta_0 + \sum_{j=1}^{p}\beta_j x_{i,j}))^2 \\ \text{subject to} & \sum_{j=1}^{p}|\beta_j| \leq t \end{cases}$$

There exist a one-to-one correspondence between λ and t, and that both of the formulations lead to the same solution.

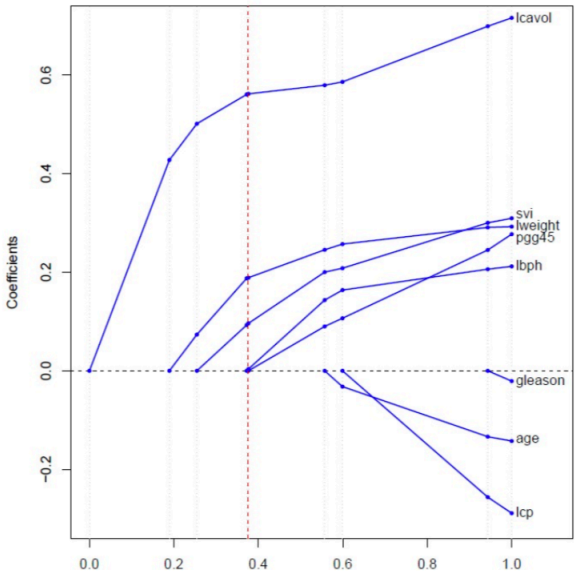

Figure 2: x-axis: $\frac{1}{\lambda}$. As the penalty decrease, the parameters converge to LSE.

Remarks Computing the lasso solution is a quadratic programming problem, but efficient algorithms are available for computing the entire path of solutions as λ varies with the same computational cost as for ridge regression.

Limitations

1. If $p > n$, the lasso selects at most n variables. (The number of selected genes is bounded by the number of samples)
2. Grouped variables: the lasso fails to do grouped selection. It tends to select one variable from a group and ignore the others.

9.2.3 Ridge v.s. Lasso

related R packages The GLMnet and Elastic net packages in R are freely available.

Post-selection inference Post-selection inference is still a problem problem when using these techniques. In fact there is not even a distribution in

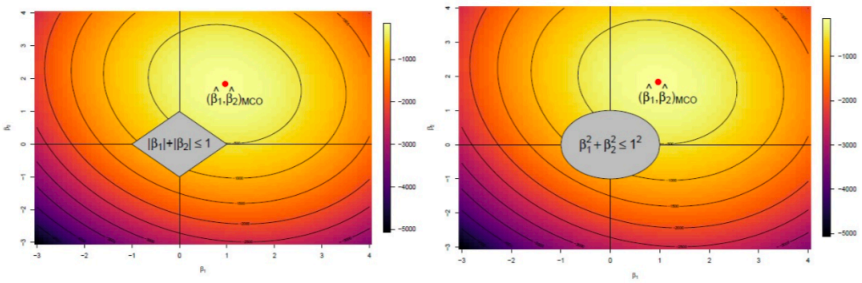

Fig. 1: `lasso` Fig. 2: `ridge`

Figure 3: Contour of the prediction errors and Regularizations. Different constraints lead to different contour shapes. Notice that lasso always has the smallest error at one of the axis, thus induces sparsity.

the model yet. SelectiveInference package in R allow to do valid inference for parameters that were selected using Elastic net.

9.2.4 Elastic Net Regularization

$$\hat{\beta} = \underset{\beta}{argmin}|\mathbf{y} - X\beta|^2 + \lambda_2|\beta|^2 + \lambda_1|\beta|$$

Properties

1. Removes the limitation on the number of selected variables
2. Encourages grouping effect
3. Stabilizes the l_1 regularization path

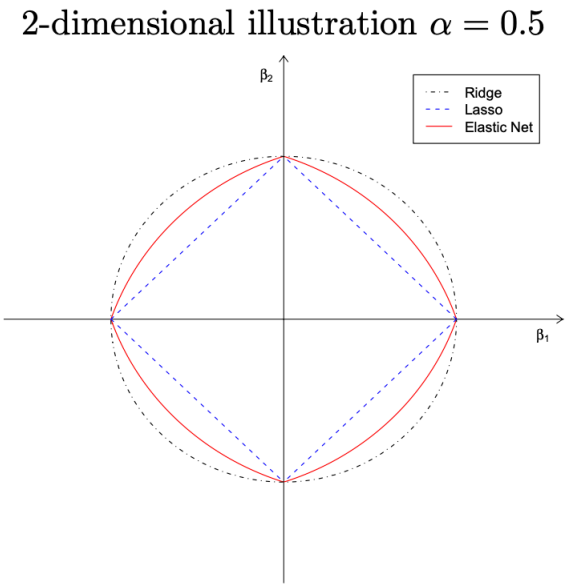

Figure 4: Singularities at the vertexes (necessary for sparsity) and strict convex edges (grouping). The strength of convexity varies with α

Lasso v.s. Elastic Net

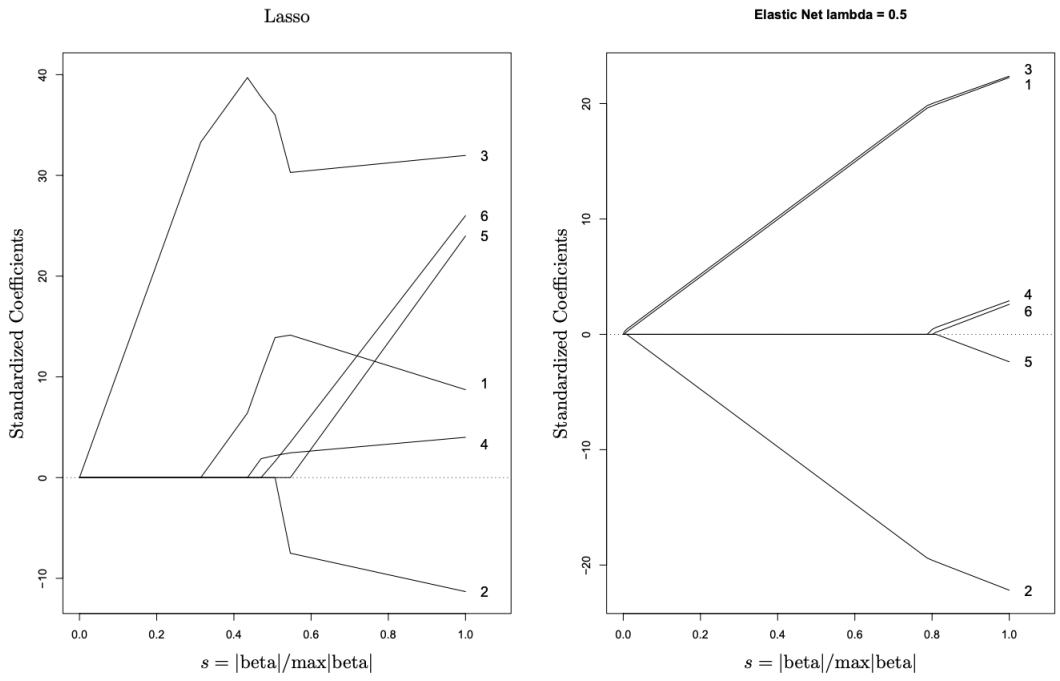

Figure 5: Elastic Net encourages group effect.