Methods of Data Analysis – STA302 Course Notes

Yuchen Wang

May 28, 2019

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 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 $SSG/(T-1) = SSE/(n-T)$, 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 (MSE):

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

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

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

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 $sigma^2$:

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

This is MSE, 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

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

5.1 Predictive Inference

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

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

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

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

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

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

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 $\hat{\mathbf{y}}^*$ using t-distribution ($df = n - 2$)

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)

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

7 May 30th - Interactions and multiple linear regression assumptions

7.1 Transformation

We can use transformations to fix 2 problems:

1. Non-constant variance
2. Non-linearity

8 June 6th - Model selection and variable selection

9 June 11th - Ridge and Lasso regression

10 June 13th - Statistical analysis, data science and ethics