Linear Regression 2

MAE301 Applied Experimental Statistics

Max Yi Ren

Department of Mechanical Engineering
Arizona State University

November 3, 2015

1. analysis of variance
2. inference of linear model
3. summary

Recap of OLS

Consider training data $(\mathbf{X}, \mathbf{y})$, with each row of $\mathbf{X}$ being a data point and that of $\mathbf{y}$ the corresponding response. A linear model assumes:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{\varepsilon}$ are random errors following i.i.d. normal, i.e., $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. The goal of OLS is to estimate the model parameters $\boldsymbol{\beta}$ so that the estimations $\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta}$ are close to the observed $\mathbf{y}$. This can be formulated as follows:

$$\boldsymbol{\beta}^* = arg \min_{\boldsymbol{\beta}} \; ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2, \tag{2}$$

which has an analytical solution:

$$\boldsymbol{\beta}^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \tag{3}$$

unbiased estimator

Note that the estimator

$$\boldsymbol{\beta}^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{4}$$

is random since the observations $\mathbf{y}$ involves random error. The expectation of $\boldsymbol{\beta}^*$ is

$$\begin{aligned} E(\boldsymbol{\beta}^*) &= E\left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\right) \\ &= E\left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{\varepsilon}\right) \\ &= E(\boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{\varepsilon}) \\ &= \boldsymbol{\beta}, \end{aligned} \tag{5}$$

i.e., $\boldsymbol{\beta}^*$ is unbiased.

variance of β^*

The variance of β^* is

$$\begin{aligned} Var(\beta^*) &= Var\left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\right) \\ &= Var\left(\beta + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\varepsilon\right) \\ &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}. \end{aligned} \tag{6}$$

estimation of σ^2

We use the sample variance to estimate the variance of the error:

$$s^2 = \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*)^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*)}{n - p}. \tag{7}$$

Here $\mathbf{e} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*$ is the residual (realizations of the random error), $n - p$ is the degree of freedom of the residual. The covariance of the estimator $\boldsymbol{\beta}^*$ can be approximated by $s^2(\mathbf{X}^T\mathbf{X})^{-1}$.

degree of freedom

Why is the degree of freedom $n - p$? Note that during the derivation of $\boldsymbol{\beta}^*$, we set the gradient of the squared error to zero to have:

$$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) = \mathbf{0}, \tag{8}$$

or $\mathbf{X}^T\mathbf{e} = \mathbf{0}$. These are the p equations $\mathbf{e}$ needs to follow, i.e., while there are n residual terms, only $n - p$ of them can freely be chosen, and the rest p are determined by $\mathbf{X}^T\mathbf{e} = \mathbf{0}$.

analysis of variance (ANOVA)

In the context of linear regression, ANOVA tests the null hypothesis: $\beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$, i.e., the linear model does not explain the variance in the responses at all. Note that the intercept is not included in the analysis. The alternative hypothesis says that some linear coefficients are none-zero, i.e., are useful at explaining the variance. The test uses the following F-statistic (a ratio of two variances):

$$F = \frac{MSM}{MSE}, \tag{9}$$

where

$$MSM = \frac{||\mathbf{X}\beta^* - \bar{y}\mathbf{1}||^2}{p-1}, \tag{10}$$

$$MSE = \frac{||\mathbf{y} - \mathbf{X}\beta^*||^2}{n-p}. \tag{11}$$

The null hypothesis is rejected when F is larger than a critical value.

one-way ANOVA

We can use ANOVA to test whether multiple data sets have the same mean. Consider a linear model with one covariate:

$$Y = \beta_1 + \beta_0 X + \varepsilon. \tag{12}$$

If X is binary, then the model represents two sample sets: When $X = 0$, Y represents a random variable with mean β_1; when $X = 1$, Y represents another random variable with mean $\beta_0 + \beta_1$. Since ANOVA tests the hypothesis $\beta_0 = 0$, it essentially tests whether the two random variables have the same mean.
For more than two sample sets, we can introduce more binary (dummy) variables and their coefficients to represent the differences between means of those sample sets and that of the baseline set.

one-way ANOVA and t-test

One shall note that one-way ANOVA with two sample sets is equivalent to a two-sample t-test. Both are under the following assumption:

- The two sample sets are both independently and normally distributed
- The population variances of the two sets are the same

However, ANOVA can be applied when there are more than two sample sets and test whether all of the sets have the same mean.

inference for individual coefficients

We can also apply t-test to individual coefficients in a linear model. Recall that the estimator $\boldsymbol{\beta}^* \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$. We can define the variance of an individual estimator $s^2(\beta_j^*) = s^2(\mathbf{X}^T\mathbf{X})_{j,j}^{-1}$, i.e., the jth diagonal element of $s^2(\mathbf{X}^T\mathbf{X})^{-1}$.
The confidence interval of β_j is $\boldsymbol{\beta}^* \pm t_{\alpha/2}s^2(\beta_j^*)$. Here $t_{\alpha/2}$ has $n-p$ degrees of freedom.
Hypothesis test: Null hypothesis is $\beta_j = 0$. Test statistic:
$t = \frac{\beta_j^*}{\sqrt{s^2(\beta_j^*)}}$.
Note that this tests the significance of a variable given that the other variables are already in the model.

summary of the class

- ANOVA can be used for linear regression model to test if coefficients (except the intercept) are zero. Note that rejecting this null hypothesis suggests that some coefficients shall be non-zero but does not tell which one(s).
- ANOVA can also be used to test whether the mean of multiple groups of data are the same or not. Note that rejecting this null hypothesis suggests some groups have different means but does not tell which specific groups.
- For two groups of data, ANOVA provides the same conclusion as a t-test. Note: ANOVA assumes equal variances for the multiple groups.
- Variance in estimators of a linear regression model is related to the covariance matrix of inputs.

Python code for demos in the class

```python
import numpy as np
import matplotlib
import matplotlib.pyplot as plt

### ANOVA for regression
import numpy as np
import statsmodels.api as sm

# Generate artificial data (2 regressors + constant)
n_samples = 30
true_fun = lambda X: np.cos(1.5 * np.pi * X) + np.sin(0.5 * np.pi * X)
X = np.sort(np.random.rand(n_samples))
X1 = np.vstack((X**0,X**1,X**2,X**3)).T
y = true_fun(X) + np.random.randn(n_samples) * 0.1
plt.scatter(X,y)
plt.xlabel("x")
plt.ylabel("y")

# Inspect linear regression results
results = sm.OLS(y, X1).fit()
print results.summary()

# calculate F statistic
p = X1.shape[1] # model dof
beta = np.linalg.solve(np.dot(X1.T,X1),np.dot(X1.T,y)) # coefficients
ybar = y.mean() # sample mean
yhat = np.dot(X1,beta)
MSM = np.linalg.norm(yhat-ybar)**2/(p-1)
MSE = np.linalg.norm(yhat-y)**2/(n_samples-p)
F = MSM/MSE

# check std on each coefficient
sigma = np.sum((y-np.dot(X1,beta.T))**2)/(y.size-p)
np.sqrt(np.matrix.diagonal(np.linalg.inv(np.dot(X1.T,X1))*sigma))
```

Python code for demos in the class

```python
### Compare two sample t-test and one way ANOVA
import numpy as np
nobs = 1000
b2 = 0
b1 = 0.0
b0 = 0.2
X1 = b2 + np.random.random(nobs)
X2 = b2 + b1 + np.random.random(nobs)
X3 = b2 + b1 + b0 + np.random.random(nobs)
plt.hist( [X1, X2, X3] , stacked=False)

# two sample t-test
from scipy.stats import ttest_ind
ttest_ind(X1,X2)

# one way ANOVA
from scipy.stats import f_oneway
f_oneway(X1,X2)

# one way ANOVA on three groups
f_oneway(X1,X2,X3)
```