MATH 829: Introduction to Data Mining and Analysis
Linear Regression: statistical tests

Dominique Guillot

Departments of Mathematical Sciences
University of Delaware

February 17, 2016

Statistical hypothesis testing

Suppose we have a linear model for some data:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon.$$

- An important problem is to identify which variables are really useful in predicting Y.
- We want to decide if $\beta_i = 0$ or not with some level of confidence.
- Also want to test if groups of coefficients $\{\beta_{i_k} : k = 1, \dots, l\}$ are zero.

Statistical hypothesis testing (cont.)

Recall: to do a statistical test:

1. State a **null** hypothesis H_0 and an alternative hypothesis H_1.
2. Construct an appropriate **test statistics**.
3. Derive the **distribution** of the test statistic under the null hypothesis.
4. Select a **significance level** α (typically 5% or 1%).
5. Compute the test statistics and decide if the null hypothesis is rejected at the given significance level.

Example

Example: Suppose $X \sim N(\mu, 1)$ with μ unknown. We want to test:

$$H_0 : \mu = 0$$
$$H_1 : \mu \neq 0.$$

We have an iid sample $X_1, \dots, X_n \sim N(\mu, 1)$.
Recall: if $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ are independent, then $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.
Therefore, under H_0, we have

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(0, \frac{1}{n}).$$

Test statistics: $\sqrt{n} \cdot \hat{\mu} \sim N(0, 1)$.
Suppose we observed: $\sqrt{n}\hat{\mu} = k$.
We compute:

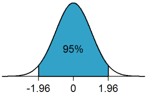
$$P(-z_\alpha \leq \sqrt{n}\hat{\mu} \leq z_\alpha) = P(-z_\alpha \leq N(0,1) \leq z_\alpha) = 1 - \alpha.$$

Reject the null hypothesis if $k \notin [-z_\alpha, z_\alpha]$.

Example (cont.)

For example, suppose: $\alpha = 0.05$, $\sqrt{n}\hat{\mu} = 2.2$.
If $Z \sim N(0,1)$, then

$$P(-1.96 \leq Z \leq 1.96) = 0.95$$



Therefore, it is very unlikely to observe $\sqrt{n}\hat{\mu} = 2.2$ if $\mu = 0$. We **reject** the null hypothesis.
In fact, $P(-2.2 \leq Z \leq 2.2) \approx 0.972$. So our **p-value** is 0.028.

- Type I error: H_0 true, but rejected $\rightarrow$ False positive. (Controlled by the level α).
- Type II error: H_0 false, but not rejected $\rightarrow$ False negative. (Power of the test).

Testing if coefficients are zero

- In practice, we often include unnecessary variables in linear models.
- These variables bias our estimator, and can lead to poor performance.
- Need ways of identifying a "good" set of predictors.

We now discuss a classical approach that uses statistical tests.

Before, we tested if the mean of a $N(\mu, 1)$ is zero:

$$H_0 : \mu = 0$$
$$H_1 : \mu \neq 0.$$

assuming $\sigma^2 = 1$ is known. What if the variance is *unknown*?

Sample variance:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2, \qquad \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Testing if coefficients are zero (cont.)

In general, suppose $X \sim N(\mu, \sigma^2)$ with σ^2 known and we want to test

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0.$$

Under the H_0 hypothesis, we have

$$\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu_0, \frac{\sigma^2}{n}\right)$$

Therefore, we use the test statistic

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1).$$

If the variance is unknown, we replace σ by its sample version s

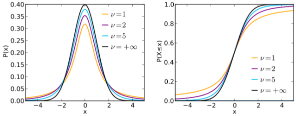
$$T = \frac{\overline{X} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}.$$

Review: the student distribution

The student t_ν distribution with ν degrees of freedom:

$$f_\nu(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

where Γ is the Gamma function.



When $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$, then

$$\frac{\overline{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}.$$

Testing if coefficients are zero (cont.)

Back to testing regression coefficients: suppose

$$y_i = x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + \epsilon_i,$$

where (x_{ij}) is a fixed matrix, and ϵ_i are iid $N(0, \sigma^2)$.

We saw that this implies

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

In particular,

$$\hat{\beta}_i \sim N(\beta_i, v_i\sigma^2),$$

where v_i is the i-th diagonal element of $(X^T X)^{-1}$.

We want to test:

$$H_0 : \beta_i = 0$$
$$H_1 : \beta_i \neq 0.$$

Note: v_i is known, but σ is **unknown**.

Testing if coefficients are zero (cont.)

Recall:

$$y_i = x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + \epsilon_i,$$

Problem: How do we estimate σ?

Let $\hat{\epsilon}_i = y_i - (x_{i,1}\hat{\beta}_1 + x_{i,2}\hat{\beta}_2 + \cdots + x_{i,p}\hat{\beta}_p)$. What about:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i^2 \quad ?$$

What is $E(\hat{\sigma}^2)$?

We have $y = X\beta + \epsilon$ and $\hat{\beta} = (X^T X)^{-1}X^T y$. Thus,

$$\begin{aligned}
\hat{\epsilon} &= y - X\hat{\beta} \\
&= y - X(X^T X)^{-1}X^T y \\
&= (I - X(X^T X)^{-1}X^T)y \\
&= (I - X(X^T X)^{-1}X^T)(X\beta + \epsilon) \\
&= (I - X(X^T X)^{-1}X^T)\epsilon \\
&= M\epsilon.
\end{aligned}$$

Testing if coefficients are zero (cont.)

We showed $\hat{\epsilon} = M\epsilon$ where $M := I - X(X^T X)^{-1}X^T$. Now

$$\sum_{i=1}^n \hat{\epsilon}_i^2 = \hat{\epsilon}^T\hat{\epsilon} = \epsilon^T M^T M\epsilon.$$

Note: $M^T = M$ and

$$\begin{aligned}
M^T M = M^2 &= (I - X(X^T X)^{-1}X^T)(I - X(X^T X)^{-1}X^T) \\
&= I - X(X^T X)^{-1}X^T = M.
\end{aligned}$$

(In other words, M is idempotent.)

Therefore,

$$\sum_{i=1}^n \hat{\epsilon}_i^2 = \epsilon^T M\epsilon.$$

Now,

$$\begin{aligned}
E(\epsilon^T M\epsilon) &= E(\operatorname{tr}(M\epsilon\epsilon^T)) \\
&= \operatorname{tr} E(M\epsilon\epsilon^T) \\
&= \operatorname{tr} ME(\epsilon\epsilon^T) \\
&= \operatorname{tr} M\sigma^2 I = \sigma^2 \operatorname{tr} M.
\end{aligned}$$

Testing if coefficients are zero (cont.)

We proved:

$$E\Big(\sum_{i=1}^n \hat{\epsilon}_i^2\Big) = \sigma^2 \operatorname{tr} M,$$

where $M = I - X(X^T X)^{-1}X^T$. (Here $I = I_n$, the $n \times n$ identity matrix.)

What is $\operatorname{tr} \mathbf{M}$? Recall $\operatorname{tr}(AB) = \operatorname{tr}(BA)$. Thus,

$$\begin{aligned}
\operatorname{tr} M &= \operatorname{tr}(I - X(X^T X)^{-1}X^T) \\
&= n - \operatorname{tr}(X(X^T X)^{-1}X^T) \\
&= n - \operatorname{tr}(X^T X(X^T X)^{-1}) \\
&= n - \operatorname{tr}(I_p) \\
&= n - p.
\end{aligned}$$

Therefore,

$$\frac{1}{n-p}E\Big(\sum_{i=1}^n \hat{\epsilon}_i^2\Big) = \sigma^2.$$

Testing if coefficients are zero (cont.)

As a result of the previous calculation, our estimator of the variance σ^2 in the regression model will be

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

where $\hat{y}_i := x_{i,1}\hat{\beta}_1 + x_{i,2}\hat{\beta}_2 + \cdots + x_{i,p}\hat{\beta}_p$ is our prediction of y_i.

Our test statistic is

$$T = \frac{\hat{\beta}_i}{s\sqrt{v_i}}, \qquad v_i = ((X^T X)^{-1})_{ii}.$$

Under the null hypothesis $H_0 : \beta_i = 0$, one can show that the above T statistic has a student distribution with $n - p$ degrees of freedom:

$$T \sim t_{n-p}.$$

Thus, to test if $\beta_i = 0$, we compute the value of the T satistic, say $T = \hat{T}$ and reject the null hypothesis (at the $\alpha = 5\%$ level) if

$$P(|t_{n-p}| \geq \hat{T}) \leq 0.05.$$

Important: This procedure **cannot** be iterated to remove multiple coefficients. We will see how this is done later.

Confidence intervals for the regression coefficients

Recall that

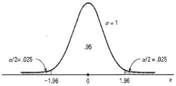
$$\hat{\beta}_i \sim N(\beta_i, v_i\sigma^2).$$

Using our esimate s^2 for σ^2, we can construct a $1 - 2\alpha$ confidence interval for β_i:

$$\left(\hat{\beta}_i - z^{(1-\alpha)}\sqrt{v_i}s, \hat{\beta}_i + z^{(1-\alpha)}\sqrt{v_i}s\right).$$

Here $z^{(1-\alpha)}$ is the $(1-\alpha)$-th percentile of the $N(0,1)$ distribution, i.e.,

$$P(Z \leq z^{1-\alpha}) = 1 - \alpha.$$



Python

- Unfortunately, `scikit-learn` doesn't compute t-statistics and confidence intervals.
- However, the module `statsmodels` provides exactly what we need.

```
import numpy as np
import statsmodels.api as sm
import statsmodels.formula.api as smf

# Load data
dat = sm.datasets.get_rdataset("Guerry", "HistData").data

# Fit regression model (using the natural log
# of one of the regressors)
results = smf.ols('Lottery ~ Literacy +
np.log(Pop1831)', data=dat).fit()

# Inspect the results
print results.summary()
```


Python (cont.)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                Lottery   R-squared:                       0.348
Model:                            OLS   Adj. R-squared:                  0.333
Method:                 Least Squares   F-statistic:                     22.20
Date:                Mon, 18 Jan 2016   Prob (F-statistic):           1.90e-08
Time:                        15:40:59   Log-Likelihood:                -379.82
No. Observations:                  86   AIC:                             765.6
Df Residuals:                      83   BIC:                             773.0
Df Model:                           2
===================================================================================
                      coef    std err          t      P>|t|      [95.0% Conf. Int.]
-----------------------------------------------------------------------------------
Intercept         246.4341     35.233      6.995      0.000       176.358   316.510
Literacy           -0.4889      0.128     -3.832      0.000        -0.743    -0.235
np.log(Pop1831)   -31.3114      5.977     -5.239      0.000       -43.199   -19.424
==============================================================================
Omnibus:                        3.713   Durbin-Watson:                   2.019
Prob(Omnibus):                  0.156   Jarque-Bera (JB):                3.394
Skew:                          -0.487   Prob(JB):                        0.183
Kurtosis:                       3.003   Cond. No.                         702.
==============================================================================
```