11. Significance Tests

Significance Tests

this is the same thing that we did before.

8:37 / 8:37 1.50x

Video

Download video file

Transcripts

Download SubRip (.srt) file

Download Text (.txt) file

Setup:

A geneticist at the Broad Institute wishes to study the relationship between a collection of five genes and obesity. In particular, he suspects that the number of mutations in these five genes $\mathbf{X} = (X_1, \ldots, X_5)$ is correlated to the blood sugar level Y, when all other factors such as diet are kept identical.

A dataset consisting of measurements obtained from $n = 125$ patients is obtained from a nearby hospital. As statisticians, we attempt to perform linear regression with the assumption that the relationship of Y given $\mathbf{X}$ is linear.

All problems on this page refers to this setup.

Building a hypothesis test

2/2 points (graded)

Let's say we suspect that the number of mutations in gene 1 has some (non-zero) correlation with blood sugar level. To test this, we beign by defining the null hypothesis $H_0 : \beta_1 = 0$, and the alternative hypothesis $H_1 : \beta_1 \neq 0$.

Using the setup given above, what is an appropriate choice for the unit column vector $\mathbf{u} \in \mathbb{R}^5$? That is, what $\mathbf{u}$ gives $\mathbf{u}^T \beta = \beta_1$?

(For convenience, enter your answers to all answer boxes in this problem as a row vector to represent $\mathbf{u}^T$. For instance, if your answer is $\mathbf{u} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$, type “[1,2]". *Do not round; enter exact fractional values if applicable.*)

$\mathbf{u}^T =$ [1,0,0,0,0] ✔ **Answer:** [1,0,0,0,0]

Alternatively, we could also test whether gene 2 has a more positive correlation than gene 3. In this scenario, we setup the null hypothesis $H_0 : \beta_2 \leq \beta_3$ and $H_1 : \beta_2 > \beta_3$. Alternatively, we could write this as $H_0 : \beta_2 - \beta_3 \leq 0$ and $H_1 : \beta_2 - \beta_3 > 0$.

What choice of unit vector $\mathbf{u}$ satisfies $\mathbf{u}^T \beta \leq 0 \iff \beta_2 - \beta_3 \leq 0$?

$\mathbf{u}^T =$ [0,1/sqrt(2),-1/sqrt(2),0,0 ✔ **Answer:** [0,1/sqrt(2),-1/sqrt(2),0,0]

Solution:

For the first setup, $\mathbf{u} = (1, 0, 0, 0, 0)$ is the right choice, since we just want the first coordinate β_1. In the second setup, we want the second coordinate minus the third. Therefore, we ought to normalize the vector $(0, 1, -1, 0, 0)$. Therefore, $\mathbf{u} = (0, \frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}, 0, 0)$ is the correct choice.

Submit

You have used 2 of 3 attempts

Answers are displayed within the problem

Statistics for the LSE

1/1 point (graded)

Again, use the setup as in the previous problem.

We assume that the model is homoscedastic; i.e. $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_{125})$, so that $\mathbf{Y} = \mathbb{X}\beta^* + \varepsilon$.

In the linear regression model, we derived $\hat{\beta} = \beta^* + (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\epsilon$, so $\hat{\beta}$ is a p-dimensional Gaussian. We saw previously that $\hat{\sigma}^2 = \frac{1}{n-p}\|\mathbf{Y} - \mathbb{X}\hat{\beta}\|_2^2$ is an unbiased estimator of σ^2.

Let $\mathbf{u}$ be a unit vector in $\mathbb{R}^5$. What distribution does the quantity $S = \frac{\mathbf{u}^T\hat{\beta} - \mathbf{u}^T\beta}{\hat{\sigma}\sqrt{\mathbf{u}^T(\mathbb{X}^T\mathbb{X})^{-1}\mathbf{u}}}$ obey?

- $\mathcal{N}(0, 1)$, the standard normal distribution.
- t_{120}, a t-distribution with $n - p = 120$ degrees of freedom.
- χ^2_{120}, a chi-squared distribution with 120 degrees of freedom.

Solution:

The correct answer is **"t_{120}, a t-distribution with $n - p = 120$ degrees of freedom."**

The formula provided gives $\mathbf{u}^T\hat{\beta} - \mathbf{u}^T\beta^* = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\epsilon$, which obeys the Gaussian distribution $\mathcal{N}\left(0, \sigma^2\mathbf{u}^T(\mathbb{X}^T\mathbb{X})^{-1}\mathbf{u}\right)$.

To see why, note that the covariance must be $\left(\mathbf{u}^T(\mathbb{X}^T\mathbb{X})^{-1}\right)(\sigma^2 I)\left(\mathbf{u}^T(\mathbb{X}^T\mathbb{X})^{-1}\right)^T = \sigma^2\mathbf{u}^T(\mathbb{X}^T\mathbb{X})^{-1}\mathbf{u}$.

From the definition of the t-distribution, we conclude that S obeys the law t_{120}, since S uses the unbiased estimate $\hat{\sigma}$ in place of σ.

Submit You have used 2 of 2 attempts

Answers are displayed within the problem

Designing the test

1/1 point (graded)
Let us work with the first scenario from the previous problem. We have the two-tailed hypotheses test $H_0 : \beta_1 = 0, H_1 : \beta_1 \neq 0$. Consider the test statistic

$$T := \frac{\mathbf{u}^T\hat{\beta}}{\hat{\sigma}\sqrt{\mathbf{u}^T(\mathbb{X}^T\mathbb{X})^{-1}\mathbf{u}}}$$

where $\mathbf{u}$ is the appropriate **unit vector** (a vector of length 1) such that $\mathbf{u}^T\beta = \beta_1$.

Keep in mind the following intuition: **we ought to reject H_0 if $\hat{\beta}_1$ is far away from zero, the presumed value of β_1 under the null hypothesis**. How far is "far"? We studied this previously in the Hypothesis Testing unit, and we now apply that knowledge to this setting.

We design the two-sided test with level α

$$\psi := \mathbf{1}\left(|T| \geq q_{\alpha/2}\right).$$

where q_α is the $(1-\alpha)$ quantile of the distribution of T, which has a certain distribution under H_0 (refer to the solution to the previous problem, which asks for the distribution of a certain random variable S). If we decide to test at the level $\alpha = 0.001$, what is the numerical value of $q_{\alpha/2}$? Round to the nearest 10^{-3}.

$q_{\alpha/2} =$ 3.373454 ✔ **Answer:** 3.374

Solution:

We saw previously that the statistic T, under the null hypothesis $\beta_1 = 0$, obeys the t-distribution with $n - p = 125 - 5 = 120$ degrees of freedom. Since we are doing a two-tailed test at significance level $\alpha = 0.001$, we wish to compute $q_{\alpha/2}$ such that $\Pr\left(|T| > q_{\alpha/2}\right) = 0.001$. Plugging this into a calculator (or looking the values up in a t-distribution table) gives $q_{\alpha/2} \approx 3.373$. (Note that this is very different from the quantile function q_α for a normal distribution!)

Submit

You have used 1 of 3 attempts

Answers are displayed within the problem

Discussion

Hide Discussion

Topic: Unit 6 Linear Regression:Lecture 20: Linear Regression 2 / 11. Significance Tests

Add a Post

All Posts

[STAFF] Solution for "Statistics for LSE" is missing some terms.

discussion posted 4 days ago by DriftingWoods

u^t on RHS of equation on line 2.

Extra X^T on LHS two times in each parentheses in equation on line 3.

This post is visible to everyone.

Add a Response

1 response

ya_mukhin (Staff)
3 days ago

Thank you, @DriftingWoods, this has now been fixed.

Add a comment

sandipan_dey
5 minutes ago

Since we have 5 features (genes), should not we add a 1^T vector (a column of 1s for the intercept) to have $p = 5 + 1 = 6$?

Add a comment

Preview

Submit

Showing all responses

Add a response:

Preview

Submit