STATS305A - Lecture 5

John Duchi
Scribed by Michael Howes

05/10/21

Contents

1 Announcements **1**

2 Inference **1**
2.1 Testing linear functionals 2
2.2 T-test of significance 2
2.3 Using R 4
2.4 F-tests 4
2.5 Testing submodels 5

1 Announcements

- John's office hours will be after class 11:15-12. Tuesdays and Thursdays.
- Homework 1 is due on Friday. There is an extra coding question which was added this weekend.
- Etude coming on Friday.
- Data for coding questions is available on website unless otherwise stated.
- Course notes from last week to be uploaded soon.

2 Inference

We want to know if various parameters in our model "matter". We will design tests that ask this question. Recall that we have

Theorem 1. *If $Y = X\beta + \varepsilon$ and $\varepsilon \sim \mathsf{N}(0, \sigma^2 I)$, then (key independenc results)*

(a) $\widehat{\beta} = (X^T X)^{-1} X^Y \sim \mathsf{N}(\beta, (X^T X)^{-1})$.

(b) $\widehat{Y} = X\widehat{\beta} = HY \sim \mathsf{N}(X\beta, \sigma^2 H)$.

(c) $\widehat{\varepsilon} = (I - H)Y = (I - H)\varepsilon \sim \mathsf{N}(0, \sigma^2(I - H))$.

(d) $\widehat{\varepsilon} \perp\!\!\!\perp (\widehat{\beta}, \widehat{Y})$.

where $H = X(X^T X)^{-1} X$ is the projection onto the range of X.

2.1 Testing linear functionals

Say we are curious about the value of $c^T\beta$ where $c \in \mathbb{R}^d$. A common example is when $c = e_j$ (j^{th} basis vector), then $c^T\beta = e_j^T\beta = \beta_j$.

By our key independence result we have

$$c^T(\widehat{\beta} - \beta) \sim \mathsf{N}(0, \sigma^2 c^T(X^TX)^{-1}c),$$

and

$$\begin{aligned} S^2 &:= \frac{1}{n-d}\|\widehat{\varepsilon}\|_2^2 \\ &= \frac{1}{n-d}\sum_{i=1}^{n}(x_i^T\beta - Y_i)^2 \\ &= \frac{1}{n-d}\|(I-H)\varepsilon\|_2^2 \\ &\sim \frac{\sigma^2}{n-d}\chi^2_{n-d}. \end{aligned}$$

Since $I - H$ has rank $n - d$. We alos have $S^2 \perp\!\!\!\perp (\widehat{\beta} - \beta)$ and thus

$$\frac{c^T(\widehat{\beta} - \beta)}{s\sqrt{c^T(X^TX)^{-1}}} \sim T_{n-d}.$$

This gives us the t-test.

2.2 T-test of significance

Assume $c^T\beta = 0$ (this is our null hypothesis, H_0). Reject the null if

$$t := \frac{c^T\widehat{\beta}}{S\sqrt{c^T(X^TX)^{-1}C}},$$

is far from 0.

In the <u>two sided T-test</u> our p-value is

$$\mathbb{P}(|T_{n-d}| \geq |t|),$$

and we reject the null if this value is small.

In the <u>one sided T-test</u> we only care if $c^T\beta \geq 0$. In which case our p-value if

$$\mathbb{P}(T_{n-d} \geq t),$$

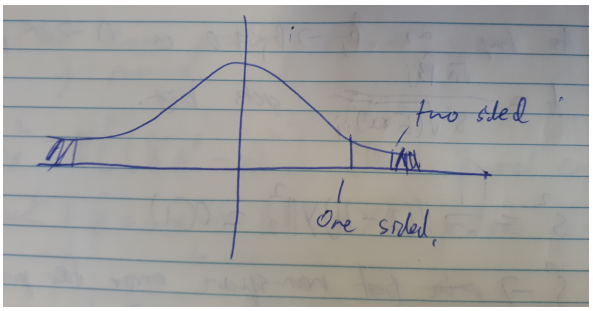
and we again reject the null if this value is small.

Example 1. Suppose we are doing drug development and we have

$$X_i := \begin{cases} 1 & \text{if person } i \text{ received treatment,} \\ 0 & \text{otherwise.} \end{cases}$$

and $Y =$ survival time. We will then have the model $Y = \beta_0 + \beta_1 X + \varepsilon$. We are interested in whether or $\beta_1 \geq 0$. Thus we will use the one-side test.

When have we actually made a discovery? We reject the null H_0 if our p-value is less than α where α is the *level* of our test and is equal to the acceptable level of significance. "$\alpha = 0.05$" corresponds to a case when $\frac{1}{20}$ discoveries are false. The one sided threshold is different ot the two sided threshold. See below picture.



We can ask what happens to our t-statistic as $n \to \infty$ (we get more and more data observations). Suppose that $c = e_j$ and so $c^T\beta = \beta_j$. Then our t-statistic is

$$\frac{\widehat{\beta}_j - \beta_j}{S^2\sqrt{(X^TX)^{-1}_{jj}}}.$$

How it changes with n shows the difference between pratical and statistical significance. Note

$$X^TX = \sum_{i=1}^{n} x_i x_i^T \approx n\mathbb{E}[xx^T],$$

and thus

$$(X^TX)^{-1} = \frac{1}{n}\left(\frac{1}{n}X^TX\right)^{-1} \approx \frac{1}{n}(\mathbb{E}[xx^T])^{-1}.$$

Thus $(X^TX)^{-1}$ gets smaller as n grows. Thus our t-statistic is approximately

$$t = \frac{\widehat{\beta}_j - \beta_j}{S^2\sqrt{\frac{1}{n}(\mathbb{E}xx^T)^{-1}_{jj}}} = \frac{\sqrt{n}(\widehat{\beta}_j - \beta_j)}{S^2\sqrt{(\mathbb{E}xx^T)^{-1}_{jj}}}.$$

As long as $\widehat{\beta}_j \to \beta_j \neq 0$ as $n \to \infty$, then

$$t = \frac{\sqrt{n}\widehat{\beta}_j}{S^2\sqrt{(\mathbb{E}xx^T)^{-1}_{jj}}},$$

gets big. Note that S^2 stays roughly constant with n since

$$S^2 = \frac{1}{n-d}\|(I-H)Y\|_2^2 = O(1).$$

Also S^2 converges to the best mean-square error for the problem. Note that

$$\widehat{\beta}_j \to \text{argmin}_b\mathbb{E}[(x^Tb - y)^2],$$

and thus argmin is not 0 in any real world problems.

Thus with enough data we can reject any null hypothesis based on the t-test. This is because if our null if $\beta_j = 0$, then the statistic we compute is

$$t = \frac{\widehat{\beta}_j}{S^2\sqrt{(X^TX)_{jj}}} \approx \frac{\sqrt{n}\widehat{\beta}_j}{S^2\sqrt{(\mathbb{E}xx^T)^{-1}_{jj}}}.$$

One implication of this is that models should become more sophisticated as our amount of data grows. This may mean getting new features or having less specific assumptions on the errors.

Let's do the same calculations for $n \to \infty$ with a bit more rigor and detail. Define

$$\beta^* = \text{argmin}_b \mathbb{E}[(x^T b - y)^2],$$

so β^* is the best linear predictor under square error of y from x. Also define

$$S_*^2 = \mathbb{E}[(x^T\beta^* - y)^2],$$

the best possible square error. We can make the following assertions if $\widehat{\beta} = \text{argmin}_b \|Xb - Y\|_2^2$, then $\widehat{\beta} \to \beta^*$ and if $S^2 = \frac{1}{n-d}\left\|X\widehat{\beta} - Y\right\|_2^2$, then $S^2 \to S_*^2$ and $\frac{1}{n}\sum_{i=1}^n x_i x_i^T \to C := \mathbb{E}[xx^T] \succ 0$. Then the test statistic under the assumed null $H_0 : \beta_j = 0$, satisfies

$$\begin{aligned} t &:= \frac{\widehat{\beta}_j}{S^2\sqrt{(X^TX)^{-1}_{jj}}} \\ &= \frac{\sqrt{n}\beta_j^*}{S_*^2\sqrt{(C^{-1})_{jj}}}(1 + o(1)) \\ &= \text{sgn}(\beta_j^*) \cdot \infty. \end{aligned}$$

Thus even if β_j^* is very very close to 0, we will reject the null if we have enough data.

Note that at a fixed significance level, the thresholds don't change very much with n since the T distribution converges to a standard normal as n goes to infinity.

The upshot is that pratical significance does not equal statistical significance.

2.3 Using R

In R if you run "`L <- lm()`" and then run "`summary(L)`", then you will get T-statistics for each regressor. These are the T-statistics we have been talking about.

These are the p-values for β_j controlling for $\{\beta_0, \ldots, \beta_d\} \setminus \{\beta_j\}$.

2.4 F-tests

Recall the F-distribution. If $U \sim \chi_d^2$ and $V \sim \chi_n^2$ and $U \perp\!\!\!\perp V$, then

$$\frac{\frac{1}{d}U}{\frac{1}{n}V} \sim F_{d,n},$$

we say that $\frac{\frac{1}{d}U}{\frac{1}{n}V}$ has an F-distribution with n degrees of freedom on top and d degrees of freedom on the bottom.

Tests based on F-statistics are more directly tied to how well our model fits the data.

Theorem 2. *Assume we have a linear model $Y = X\beta + \varepsilon$, $\varepsilon \sim \mathsf{N}(0, \sigma^2 I)$. Let $C \in \mathbb{R}^{d\times r}$ with $r \le d$ (C is tall) and suppose C has rank r. Then*

$$\frac{\frac{1}{r}\left(\left(C^T(\widehat{\beta} - \beta)\right)^T \left(C^T(X^TX)^{-1}C\right)^{-1}\left(C^T(\widehat{\beta} - \beta)\right)\right)}{S^2} \sim F_{r,n-d}, \tag{1}$$

where, as before, $S^2 = \frac{1}{n-d}\sum_{r=1}^n (y_i - x_i^T\widehat{\beta})^2$.

Example 2. If $c = e_j$, then $r = 1$ and

$$\frac{1}{S^2}\frac{(\widehat{\beta}_j - \beta_j)^2}{[(X^TX)^{-1}]_{jj}} \sim F_1, n-d.$$

This tests if $\beta_j \neq 0$ or if feature j should be included in our model. We would reject the null $\beta_j = 0$ if the above statistic if large.

Proof. [Of equation (1)] By our main independence result

$$C^T(\widehat{\beta} - \beta) \sim \mathsf{N}(0, \sigma^2 C^T(X^TX)^{-1}C),$$

and $S^2 = \frac{1}{n-d}\|(I-H)Y\|_2^2 = \frac{1}{n-d}\|(I-H)\varepsilon\|_2^2 \sim \frac{\sigma^2}{n-d}\chi^2_{n-d}$ since $I-H$ has rank $n-d$. Also $c^T(\widehat{\beta}-\beta)$ is independent of S^2.

Recall that we have seen if $w \sim \mathsf{N}(0, B)$, then $B^{-1/2}w \sim \mathsf{N}(0, I)$ and thus $wB^{-1}w \sim \chi^2_{\text{rank}(B)}$. Thus

$$\begin{aligned}
&\left(\left(C^T(\widehat{\beta} - \beta)\right)^T \left(C^T(X^TX)^{-1}C\right)^{-1} \left(C^T(\widehat{\beta} - \beta)\right)\right) \\
&= \sigma^2 \sum_{i=1}^{r} z_i^2 \quad \text{where } z_i \sim \mathsf{N}(0,1) \\
&\sim \sigma^2 \chi^2_r.
\end{aligned}$$

Thus the ratio in equation (1) equals

$$\frac{\frac{\sigma^2}{r}\chi^2_r}{\frac{\sigma^2}{n-d}\chi^2_{n-d}} = \frac{\frac{1}{r}\chi^2_r}{\frac{1}{n-d}\chi^2_{n-d}} \sim F_{n,d}.$$

□

2.5 Testing submodels

Question: is there a submodel in $Y = X\beta + \varepsilon$ that is just as good as the full model.

Suppose we have an X_0 such that range$(X_0) \subseteq$ range(X) and we assume that $Y = X_0\gamma + \varepsilon$. For example we may have $X = [1, x^{(1)}, \ldots, x^{(d)}]$ and $X_0 = [1, x^{(1)}, \ldots, x^{(d-k)}]$.

Assuming that the linear model $Y = X_0\gamma + \varepsilon$ is "good enough", define

$$\begin{aligned}
H_0 &= X_0(X_0^TX_0)^{-1}X_0^T, \\
H &= X(X^TX)^{-1}X^T
\end{aligned}$$

That is, H_0 projects onto range(X_0) and H projects onto range(X). Note that H_0 is no longer notation for the null hypothesis but rather for a matrix.

Note that $H \succeq H_0$ i.e. $H - H_0 \succeq 0$. This is because $H_0 = \sum_{i=1}^r u_i u_i^T$ and $H = \sum_{i=1}^r u_i u_i^T + \sum_{i=r+1}^d u_i u_i^T$ where $\{u_i\}$ are orthonormal vectors. Then $H - H_0 = \sum_{i=r+1}^d u_i u_i^T$. We can also think geometrically. Since H_0 projects onto a subspace of H, we can conclude that $H - H_0$ projects onto the subspace of vectors in range(X) that are orthogonal to range(X_0). Since $H - H_0$ is a projection, we have $H - H_0 \succeq 0$. Note $HH_0 = H_0 = H_0H$ and $H_0 \perp H - H_0$. This last point is because $H - H_0$ is the part of the full model that is orthogonal to the submodel.

Proposition 1. *Assume $Y = X_0\gamma + \varepsilon$, $\varepsilon \sim \mathsf{N}(0, \sigma^2 I)$ and rank$(X_0) = d - r < d =$ rank(X). Then*

$$\frac{\frac{1}{r}\left\|(H - H_0)Y_2^2\right\|}{\frac{1}{n-d}\|(I-H)Y\|_2^2} \sim F_{r,n-d} \tag{2}$$

We have the following intepretation. Let $\widehat{Y} = HY$ = predictions using the full model and $\widehat{Y}^{sub} = H_0 Y$ = predictions using the submodel. Then $SS_{full} = \sum_{i=1}^n (y_i - \widehat{y}_i)^2 = \|(I-H)Y\|_2^2$ and $SS_{sub} = \sum_{i=1}^n (y_i - \widehat{y}_i^{sub})^2 = \|(I-H_0)Y\|_2^2$. Thus

$$\begin{aligned} SS_{extra} &= SS_{sub} - SS_{extra} \\ &= \|(I-H_0)Y\|_2^2 - \|(I-H_0)Y\|_2^2 \\ &= -Y^T H_0 Y + Y^T H Y \\ &= Y^T (H - H_0) Y \\ &= \|(H-H_0)Y\|_2^2 . \end{aligned}$$

Thus if the ratio in equation (2) is very large we should use the full model. If it is very small we will use the submodel. We will now prove (2).

Proof.

$$\begin{aligned} (H-H_0)Y &= (H-H_0)(X_0\gamma + \varepsilon) \\ &= (H-H_0)\varepsilon \sim N(0, \sigma^2(H-H_0)) \end{aligned}$$

and

$$\begin{aligned} (I-H)Y &= (I-H)(X_0\gamma + \varepsilon) \\ &= (I-H)\varepsilon \sim N(0, \sigma^2(I-H)) \end{aligned}$$

We also have $(H-H_0)Y \perp\!\!\!\perp (I-H)Y$ since

$$\begin{aligned} (H-H_0)(I-H) &= H - H_0 - H^2 + H_0 H \\ &= H - H_0 - H + H_0 \\ &= 0. \end{aligned}$$

Since $\text{rank}(I-H) = n-d$ and $\text{rank}(H-H_0) = d-(d-r) = r$, the result follows. □

The intuition is that $(I-H)Y$ is outside the full model and $(H-H_0)Y$ is the part of the full model orthogonal to the sub model. Thus $(I-H)Y$ are $(H-H_0)Y$ are independent.

One more comment. In "science" when people report F-tests they are testing $X_0 = [1]$ against $X = [1, x^{(1)}, \ldots, x^{(d)}]$. What does rejecting the null mean in this case? Implicit in our derivation of the test is the assumption that the full model is the truth. The rejecting the null means we are not in the world where the full model and the sub model are both true. We may in the world were the full model is true and the sub model is false or we may be in the world were both are false.