STATS305A - Lecture 4

John Duchi
Scribed by Michael Howes

09/30/21

Contents

1 Anouncements

HW1 has been posted on the course webpage. An additional question will be added. The extra question will relate to our first Etude.

2 Distributions

2.1 Recap

We write $X \sim \mathsf{N}(\mu, \Sigma)$ is $X \in \mathbb{R}^d$ has a density

$$f(x) = \frac{1}{(2\pi)^{d/2} \det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right).$$

We will say X has a normal or Gaussian distribution with mean μ and covariance Σ. We have the following magical properties

- A normal distribution is determined by its first and second moments.
- If $Z \sim \mathsf{N}(\mu, \Sigma)$, then $Az + b$ is also Gaussian.
- If
$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right),$$
then $X \perp\!\!\!\perp Y$ (X is independent of Y) if and only if $\Sigma_{12} = 0$. (Note that $\Sigma_{21} = \Sigma_{21}^T$ so we have $\Sigma_{21} = 0 \iff \Sigma_{12} = 0$).

2.2 Chi-squared distributions

If $Z_i \overset{\text{iid}}{\sim} \mathsf{N}(0,1)$, then

$$S^2 = \sum_{i=1}^{n} Z_i^2,$$

has a χ^2_n-distribution. That is S^2 has a chi-squared distribution with n degrees of freedom (d.o.f).

Example 1 (Quadratic forms of Gaussians)**.** If $X \sim \mathsf{N}(\mu, \Sigma)$ with $\Sigma \succ 0$ and $X \in \mathbb{R}^n$, then we have $S^2 = (X-\mu)^T\Sigma^{-1}(X-\mu) \sim \chi^2_n$.

Proof. Note that $X - \mu \sim N(0, \Sigma)$ and $Z = \Sigma^{-1/2}(X-\mu) \sim N(0, I)$. Thus

$$\begin{aligned}(X-\mu)^T\Sigma^{-1}(X-\mu) &= (\Sigma^{-1/2}(X-\mu))^T(\Sigma^{-1/2}(X-\mu)) \\ &= Z^TZ \\ &= \|Z\|_2^2 \\ &\sim \chi^2_n.\end{aligned}$$

□

2.3 Projections

A matrix $\Pi \in \mathbb{R}^{n\times n}$ is a *projection matrix* if $\Pi^2 = \Pi$ (that is Π is idempotent). Recall that if we wanted to project onto $\text{span}\{a_i\}_{i=1}^k$ we used $\Pi_A = A(A^TA)^{-1}A^T$ where $A = [a_1, \ldots, a_k]$.

Suppose we know $\Pi^2 = \Pi$, what are the eigenvalues of Π? They have to be 0 or 1. Since if $\Pi = UDU^T$ where $D = \text{diag}(d_1, d_2, \ldots, d_n)$ and $U \in \mathbb{R}^{n\times n}$ satisfies $U^TU = I_n$, then

$$\begin{aligned}\Pi^2 = \Pi &\iff UD^2U^T = UDU^T \\ &\iff D^2 = D.\end{aligned}$$

Thus $d_i^2 = d_i$ for $i = 1, \ldots, n$ and thus $d_i = 0, 1$. We can thus write $\pi = \sum_{i=1}^k u_iu_i^T$ where k is the number of non-zero eigenvalues of Π which is also the rank of Π, ie $k = \dim(\text{range}(\Pi))$. The dimension of the space Π projects onto also has dimension k.

Suppose Π is a projection matrix that projects onto

$$S = \text{span}\{u_1, \ldots, u_k\} \subseteq \mathbb{R}^n.$$

How do we projection onto $S^\perp = \{v \in \mathbb{R}^n : v^Tu = 0 \text{ for all } u \in S\}$. We define $\Pi_\perp = I - \Pi$. Note that

$$\begin{aligned}\Pi_\perp^2 &= (I-\Pi)^2 \\ &= 1 - 2\Pi + \Pi^2 \\ &= I - 2\Pi + \Pi \\ &= I - \Pi.\end{aligned}$$

Also $(I-\Pi)\Pi = \Pi - \Pi^2 = 0$. Thus $I - \Pi$ is a projection and $I - \Pi$ is orthogonal to Π. We will now relate this to normally distributed random variables.

Suppose that Π is a projection matrix and $X \sim N(0, I)$. Then $X = \Pi X + (I-\Pi)X = Z + Y$. The variable Z and Y are both normal with mean 0 and

$$\mathbb{E}[ZY^T] = \Pi\mathbb{E}[XX^T](I-\Pi) = 0,$$

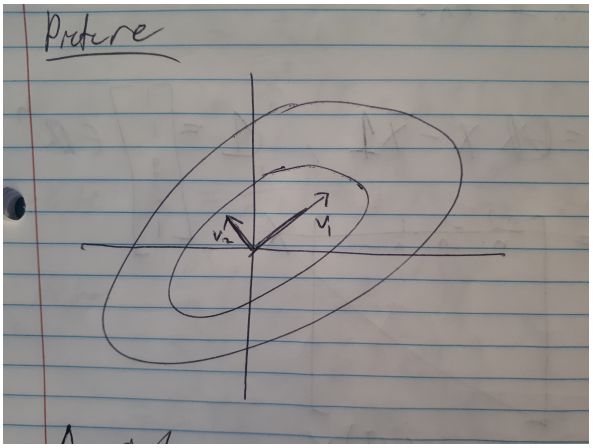
since $\mathbb{E}[XX^T] = I$. Thus $Z \perp\!\!\!\perp Y$. We can ask what is the distribution of $\|Z\|_2^2 = \|\Pi X\|_2^2$. Recall $\Pi = VV^T$ where $V = [v_1, \ldots, v_k]$ is orthogonal. Then

$$\Pi X = V \begin{bmatrix} v_1^TX \\ \vdots \\ v_k^TX \end{bmatrix} = Vw.$$

Thus

$$\|\Pi X\|_2^2 = \|Vw\|_2^2 = w^T V^T V w = \sum_{i=1}^{n} w_i^2 \sim \chi_k^2.$$

We can think about this when looing at the level sets of the normal distribution. See below picture.



2.4 Hypothesis tests and the T distribution

Suppose we have data X_i from some process and we want to test

$$\mathbb{E}[X_i] = 0 \text{ vs } \mathbb{E}[X_i] \neq 0.$$

We form a null hypothesis $H_0 : \{X_i \overset{\text{iid}}{\sim} N(0, \sigma^2)\}$ where σ^2 is unknown. If we assume that H_0 is true, then

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(0, \sigma^2/n),$$

$$S_n^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \sim \frac{\sigma^2}{n-1}\chi_{n-1}^2,$$

and $S_n^2 \perp\!\!\!\perp \bar{X}_n$.

Proof. Let $Z = X - \bar{X}_n \mathbf{1}$, where

$$\mathbf{1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \in \mathbb{R}^n.$$

We can rewrite this as $Z = X - \bar{X}_n \mathbf{1} = (I - \frac{1}{n}\mathbf{1}\mathbf{1}^T)X$. Set $\Pi = I - \frac{1}{n}\mathbf{1}\mathbf{1}^T$, then $\Pi^2 = \Pi$ so Π is a projection. In fact Π projects onto the orthogonal complement of $\mathbf{1}$. Also note that $\bar{X}_n = \frac{1}{n}\mathbf{1}^T X$. Thus

$$\begin{bmatrix} \bar{X}_n \\ Z \end{bmatrix} \sim \mathsf{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \frac{\sigma^2}{n} & - \\ - & \sigma^2\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) \end{bmatrix} \right).$$

We wish to fill in the blanks in the above covariance matrix. In particular we want to show that they are zero. Note that these blanks equal

$$\begin{aligned}
\mathbb{E}[Z\bar{X}_n] &= \frac{1}{n}\mathbb{E}\left[\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)XX^T\mathbf{1}\right] \\
&= \frac{1}{n}\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\mathbb{E}[XX^T]\mathbf{1} \\
&= \frac{1}{n}\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\sigma^2 I\mathbf{1} \\
&= \frac{\sigma^2}{n}\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\mathbf{1} \\
&= \frac{\sigma^2}{n}\left(\mathbf{1} - \frac{1}{n}\mathbf{1}\left\|\mathbf{1}\right\|_2^2\right) \\
&= \frac{\sigma^2}{n}\left(\mathbf{1} - \mathbf{1}\right) \\
&= 0.
\end{aligned}$$

Thus $\bar{X}_n \perp\!\!\!\perp Z$ and so $\bar{X}_n \perp\!\!\!\perp S_n^2$ and also $S_n^2 = \frac{1}{n-1}\left\|Z\right\|_2^2 \sim \frac{\sigma^2}{n-1}\chi^2_{n-1}$. □

Definition 1. Let $X \sim N(0,1)$ and $S^2 \sim \chi^2_d$ with $X \perp\!\!\!\perp S^2$, then the ratio

$$\frac{X}{\sqrt{\frac{1}{d}S^2}},$$

has a T-distribution (also called a Student's t-distribution) with d d.o.f.

Recall our hypothesis $H_0 : \{X_i \sim N(0, \sigma^2)\}$. Under this hypothesis the statistic

$$T_n := \frac{\bar{X}_n}{\sqrt{\frac{1}{n-1}S_n^2}} \sim T_{n-1}.$$

That is our statistic has a T distribution with $(n-1)$ degrees of freedom regardless of what σ^2 is. We next calculate our p value which is the probability under the null of observing data as "weird" as what we observed (where "weird" is something subjective that can vary). We then check if our p value is below some predetermined threshold α.

In the one-sided T-test we reject if

$$\mathbb{P}(T_{n-1} \geq t_n) \leq \alpha,$$

where α is the level of our test, T_{n-1} has a T distribution with $(n-1)$ d.o.f. and t_n is our statistic calculated from the data. In this case observed data is "weird" if it is very big and positive.

In the two-side T-test we reject if

$$\mathbb{P}(|T_{n-1}| \geq |t_n|) \leq \alpha,$$

where α, T_{n-1} and t_n are as above. In this case our data is "weird" if it is very big and either positive or negative. We could also have some other meaning of "weird". Note the following important points:

- p-values say *nothing* about the truth.
- They only say something about what *isn't* true.
- The t-test is fairly robust to non-normal data.
- The t-test is not robust to the case when the variables are dependent.

2.5 F-distributions

let $Z_1 \sim \chi_d$ and $Z_2 \sim \chi_n$ with $Z_1 \perp\!\!\!\perp Z_2$, then the ratio

$$\frac{\frac{1}{d}Z_1}{\frac{1}{n}Z_2},$$

has Fisher's F-distribution with (d, n) d.o.f. We will sometimes say d d.o.f on the top and n d.o.f on the bottom. We expect

$$\frac{\frac{1}{d}Z_1}{\frac{1}{n}Z_2},$$

to be close to 1.

3 Least squares and linear models

Starting assumptions

$$Y = X\beta + \varepsilon,$$

where $X \in \mathbb{R}^{n\times d}$, $\beta \in \mathbb{R}^d$ and $\varepsilon \in \mathbb{R}^n$. Our assumptions on ε will vary. We will always assume (a) but sometimes we will make stronger assumptions like (b) and (c)

(a) $\mathbb{E}[\varepsilon] = 0$ and $\text{cov}(\varepsilon) = \sigma^2 I_n$,

(b) $\varepsilon_i \overset{\text{iid}}{\sim} (0, \sigma^2)$,

(c) $\varepsilon_i \overset{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2)$.

We will represent X in the following way

$$X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix},$$

and usually we will have $X_{i,1} = 1$ for all i which means our model has an intercept.

We saw previously that our estimator is given by

$$\widehat{\beta} = \underset{b}{\text{argmin}} \frac{1}{2} \|Xb - Y\|_2^2 .$$

Recall that this is equivalent to $\widehat{\beta}$ satisfying the normal equations. That is we want

$$X^T X\widehat{\beta} = X^T Y.$$

(Typically X will have rank d (full rank) and then $\widehat{\beta} = (X^T X)^{-1} X^T Y$.)

3.1 Geometric picture

$\mathcal{M} = \{Xb : b \in \mathbb{R}^d\}$ is our model space. $H = X(X^T X)^{-1} X^T$ is the matrix that projects onto $\mathcal{M}$ and we call it the hat matrix.

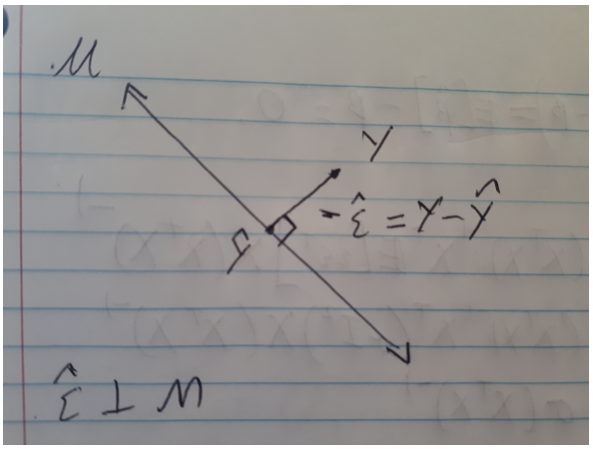
Define the predicted values

$$\widehat{Y} = X\widehat{\beta} = X(X^T X)^{-1} X^T Y = HY.$$

And the residuals/estimated error

$$\widehat{\varepsilon} = Y - \widehat{Y} = (I - H)Y.$$

We then have $\varepsilon \perp \mathcal{M}$ (ε is orthogonal to $\mathcal{M}$, see picture).



We can see this by using the normal equations which state that $X^T X\widehat{\beta} - X^T Y = 0$. Thus

$$X^T\widehat{\varepsilon} = X^T(\widehat{Y} - Y) = X^T(X\widehat{\beta} - Y) = X^T X\widehat{\beta} - X^T Y = 0.$$

3.2 Distributional results

Can we get distributional results from this picture?

Theorem 1. *Assume that X has rank $d \leq n$ and that $\mathbb{E}[\varepsilon] = 0$ and $cov(\varepsilon) = \sigma^2 I$ (this was our weakest assumption on ε), then*

(a) $\mathbb{E}[\widehat{\beta}] = \beta$ *(that is $\widehat{\beta}$ is unbiased for β).*

(b) $cov(\widehat{\beta}) = \sigma^2(X^T X)^{-1}$.

Proof. (a) $\mathbb{E}[\widehat{\beta}] = (X^T X)^{-1}X^T\mathbb{E}[Y] = X^T X)^{-1}X^T X\beta = \beta$.

(b) Note that $\mathbb{E}[\widehat{\beta} - \beta] = \mathbb{E}[\widehat{\beta}] - \beta = 0$. Thus

$$\begin{aligned}\text{cov}(\widehat{\beta}) &= (X^T X)^{-1}X^T\mathbb{E}[\varepsilon\varepsilon^T]X(X^T X)^{-1} \\ &= (X^T X)^{-1}X^T\sigma^2 I_n X(X^T X)^{-1} \\ &= \sigma^2(X^T X)^{-1}.\end{aligned}$$

□

If we use our strongest assumptions we have

Theorem 2. [Important - will be used a lot] *Assume X has rank d and $\varepsilon \sim \mathsf{N}(0, \sigma^2 I)$, then*

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

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

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

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

Proof. (a) We know $\widehat{\beta} - \beta = (X^T X)^{-1}X^T\varepsilon$ and so $\widehat{\beta} - \beta \sim \mathsf{N}(0, \sigma^2(X^T X)^{-1})$ from before.

(b) $HY = H(X\beta + \varepsilon) = X\beta + H\varepsilon$ and $H^2 = H$. Thus $HY \sim \mathsf{N}(X\beta, \sigma^2 H)$.

(c) $(I-H)Y = (I-H)(X\beta+\varepsilon) = (I-H)\varepsilon$, since $I-H$ is a projection we have $(I-H)Y \sim \mathsf{N}(0, \sigma^2(I-H))$.

For independence note that

$$\begin{aligned}\operatorname{cov}(\widehat{\varepsilon}, \widehat{Y}) &= \mathbb{E}[(I-H)Y(H(Y-\mathbb{E}Y))^T] \\ &= (I-H)\mathbb{E}[Y(Y-\mathbb{E}Y)^T]H \\ &= (I-H)\sigma^2 IH \\ &= 0.\end{aligned}$$

Thus $\widehat{\varepsilon} \perp\!\!\!\perp \widehat{Y}$ and the proof that $\widehat{\varepsilon} \perp\!\!\!\perp \widehat{\beta}$ is similar. □