Stats 201B: Statistical Modeling and Learning
Problem Set 1

January 7, 2016

Due January 14 by 11:55pm, submitted on course website.

Responses should be typeset in LaTeX, or Rmarkdown or similar.

Random Variables

1. (2pt) Consider continuous random variable X with probability distribution $p(X)$.[1]. How is $\mathbb{E}[X]$ defined? (Give the definition, not the estimator you'd use given a sample).

   $$\mathbb{E}[X] = \int_{-\infty}^{\infty} xp(x)dx$$

2. (2pt) How is $\text{Var}(X)$ defined?

   $$\text{Var}(X) = \mathbb{E}[X^2] - [\mathbb{E}[X]]^2 = \int_{-\infty}^{\infty} x^2 p(x)dx - \left(\int_{-\infty}^{\infty} xp(x)dx\right)^2$$

3. (2pt) Further suppose Y is a continuous variable as well, and you have joint density $p(X, Y)$. How is $\mathbb{E}[Y|X]$ defined? (Write it out in terms of an integral and density function).

   $$\mathbb{E}[Y|X = x] = \int yp(y|x)dy = \int y\frac{p(x,y)}{p(x)}dy = \int y\frac{p(x,y)}{\int p(x,y)dy}dy = \frac{\int yp(x,y)dy}{\int p(x,y)dy}$$

4. (2pt) If X and Y were independent, what is the relationship between $p(X, Y)$, $p(X)$, and $p(Y)$? What does $\mathbb{E}[X|Y]$ reduced to when X and Y are independent (show why this is,

[1] In problem sets we will maintain the notation used in class, in which $p(Z)$ is a density function for random variable Z. Note that we use p instead of f, and that we use this for both probability density functions and probability mass functions. Further, we drop the subscript and assume that the density is for the random variable referenced in the parentheses (i.e. $f_X(X)$ is simply $p(X)$)

writing out the definition of $\mathbb{E}[Y|X]$ first).

$$p(X,Y) = p(X)p(Y)$$
$$\mathbb{E}[Y|X=x] = \frac{\int yp(x,y)dy}{\int p(x,y)dy} = \frac{\int yp(x)p(y)dy}{\int p(x)p(y)dy} = \frac{p(x)\int yp(y)dy}{p(x)\int p(y)dy} = \frac{\int yp(y)dy}{\int p(y)dy} = \int yp(y)dy$$
$$= \mathbb{E}[Y]$$

For the following questions, draw random variables X_1, X_2,..., X_N, all independently from common density $p(X)$.

5. (2pt) Suppose you have scalars, a, b, c. What is $\mathbb{E}[aX_1 + bX_2 + cX_3]$ equal to (in terms of $\mathbb{E}[X]$)? What is $\mathrm{Var}[aX_1 + bX_2 + cX_3]$?

$$\begin{aligned}\mathbb{E}[aX_1 + bX_2 + cX_3] &= \mathbb{E}[aX_1] + \mathbb{E}[bX_2] + \mathbb{E}[cX_3] = a\mathbb{E}[X_1] + b\mathbb{E}[X_2] + c\mathbb{E}[X_3] \\ &= (a+b+c)\mathbb{E}[X] \\ \mathrm{Var}[aX_1 + bX_2 + cX_3] &= \mathrm{Var}[aX_1] + \mathrm{Var}[bX_2] + \mathrm{Var}[cX_3] \\ &= a^2\,\mathrm{Var}[X_1] + b^2\,\mathrm{Var}[X_2] + c^2\,\mathrm{Var}[X_3] \\ &= (a^2+b^2+c^2)\,\mathrm{Var}[X]\end{aligned}$$

6. (3pt) Let $\overline{X} = \frac{1}{N}\sum_i^N X_i$. Is $\overline{X}$ unbiased for $\mathbb{E}[X]$? Prove it. (Do not just cite a theorem!)

Yes.

$$\begin{aligned}\mathbb{E}[\overline{X}] &= \mathbb{E}[\sum_{i=1}^N \frac{1}{N}X_i] \\ &= \sum_{i=1}^N \frac{1}{N}\mathbb{E}[X] \qquad \text{according to 5.} \\ &= \mathbb{E}[X]\end{aligned}$$

7. (3pt) Derive the variance of $\overline{X}$. What happens to it as $N \to \infty$?

$$\begin{aligned}\mathrm{Var}[\overline{X}] &= \mathrm{Var}[\sum_{i=1}^N \frac{1}{N}X_i] \\ &= \sum_{i=1}^N \frac{1}{N^2}\,\mathrm{Var}[X] \\ &= \frac{1}{N}\,\mathrm{Var}[X]\end{aligned}$$

$$\lim_{N\to\infty} \text{Var}[\overline{X}] = \lim_{N\to\infty} \frac{1}{N} \text{Var}[X] = 0$$

Matrix Algebra,OLS, and R Practice

Consider random variables $Y \in \mathbb{R}$ and $X \in \mathbb{R}^{\mathbb{P}}$, drawn from joint density $p(X, Y)$. You collect a sample of draws from this distribution, $\{(Y_1, X_1), ..., (Y_N, X_N)\}$.

Let $\mathbf{X}$ be a $N \times (1 + P)$ matrix, with row i equal to $[1\ X_i^\top]$ (i.e., there is an intercept and then a column for each "covariate"). Consider an OLS model, $Y = \mathbf{X}\beta + \epsilon$, where $E[\epsilon|X] = 0$ by assumption.

8. (5pt) Using matrix notation at each step, derive the ordinary least squares estimator for β:

$$\beta_{OLS} = \underset{\beta\in\mathbb{R}^{P+1}}{argmin}(\mathbf{Y} - \mathbf{X}\beta)^\top(\mathbf{Y} - \mathbf{X}\beta)$$

$\text{f}(\beta) = (\mathbf{Y} - \mathbf{X}\beta)^\top(\mathbf{Y} - \mathbf{X}\beta)$

To get the minimum value of $\text{f}(\beta)$, we let $\frac{\partial f(\beta)}{\partial \beta} = 0$.

$\Rightarrow \frac{\partial f(\beta)}{\partial \beta} = -2\mathbf{X}^\top\mathbf{Y} + 2\mathbf{X}^\top\mathbf{X}\beta = 0$

$\Rightarrow \hat{\beta}_{OLS} = (\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\mathbf{Y})$

9. (4pt) Show R code that would achieve the following (there is no need to submit this code in a separate file; just include it in your problem set write-up using an environment such as verbatim):

   a. Construct a matrix X to represent $\mathbf{X}$ in the above, with $N = 100$, one column of ones, and two columns of randomly drawn numbers (from any distribution you like).

   ```
   X = cbind(rep(1, 100), matrix(rnorm(200, 0, 1), ncol = 2))
   ```

   b. Using $\beta = [1\ 2\ 3]^\top$, compute vector Y equal to $X\beta + \epsilon$, where ϵ is drawn from a standard normal distribution.

   ```
   epsilon = matrix(rnorm(100, 0, 1), ncol = 1)
   beta = matrix(1:3, ncol = 1)
   Y = X %*% beta + epsilon
   ```

   c. Compute $\hat{\beta}_{OLS} = (X^\top X)^{-1}(X^\top Y)$.

   ```
   beta_ols = solve(t(X) %*% X) %*% (t(X) %*% Y)
   > beta_ols
            [,1]
   [1,] 1.035916
   ```

```
[2,] 2.071113
[3,] 2.938582
```

d. Compare the result to the coefficients obtained using `lm` with the data you have constructed.

```
lm_ols = lm(Y~X[,1]+X[,2]+X[,3]-1)
summary(lm_ols)
> summary(lm_ols)

Call:
lm(formula = Y ~ X[, 1] + X[, 2] + X[, 3] - 1)

Residuals:
    Min      1Q  Median      3Q     Max
-2.0532 -0.6210 -0.1071  0.6383  2.5732

Coefficients:
       Estimate Std. Error t value Pr(>|t|)
X[, 1]   1.0359     0.1036   10.00   <2e-16 ***
X[, 2]   2.0711     0.0977   21.20   <2e-16 ***
X[, 3]   2.9386     0.1060   27.72   <2e-16 ***
---

Residual standard error: 1.034 on 97 degrees of freedom
Multiple R-squared:  0.9319,Adjusted R-squared:  0.9298
F-statistic: 442.8 on 3 and 97 DF,  p-value: < 2.2e-16
```

The results in (c) and (d) are the same.

10. (4pt) Show (analytically) the unbiasedness of $\hat{\beta}_{OLS}$ for β. (Hint: compute $\hat{\beta}_{OLS}$, but replacing Y with $\mathbf{X}\beta + \epsilon$).

$$\begin{aligned}\mathbb{E}[\hat{\beta}_{OLS}] &= \mathbb{E}\left[(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\mathbf{Y})\right] = (\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\mathbb{E}[\mathbf{Y}]) \\ &= (\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\mathbb{E}[\mathbf{X}\beta+\epsilon]) = (\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\mathbf{X}\beta) \qquad \text{since } \mathbb{E}[\epsilon]=0 \\ &= \beta\end{aligned}$$

11. (4pt) Compute the variance, $\mathbb{E}[(\hat{\beta}_{OLS}-\beta)(\hat{\beta}_{OLS}-\beta)^\top]$, again sticking with matrix notation. You may assume $\mathbb{E}[\epsilon\epsilon^\top|X] = \sigma^2 I_N$, where I_N is the $N \times N$ identity matrix.

$$\begin{aligned}
\mathbb{E}[(\hat{\beta}_{OLS}-\beta)(\hat{\beta}_{OLS}-\beta)^\top] &= \mathbb{E}\left[(\hat{\beta}_{OLS}-\mathbb{E}[\hat{\beta}_{OLS}])(\hat{\beta}_{OLS}-\mathbb{E}[\hat{\beta}_{OLS}])^\top\right] = \mathrm{Var}[\hat{\beta}_{OLS}] \\
&= \mathrm{Var}[(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\mathbf{Y})] = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\,\mathrm{Var}(\mathbf{Y})\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\,\mathrm{Var}(\mathbf{X}\beta+\epsilon)\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\,\mathrm{Var}(\epsilon)\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\sigma^2 I_N\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}
\end{aligned}$$

12. (2pt) What meaning would you give to the matrix $\mathbb{E}[\epsilon\epsilon^\top|X]$? Give an intuitive explanation of what the assumption that this matrix equals $\sigma^2 I$ implies.

    Covariance matrix of random errors. Random errors consist of unknown factors, uncontrolled factors, measurement errors and so on. The total effect of these factors may be either positive or negative, but in large number of trials, they have the same variation and don't affect each other.