PMAP 8131 Applied Research Methods

Statistical Review

Matteo Zullo

Georgia State University & Georgia Institute of Technology

June 6, 2022

Outline

Table of Contents

Probability

Probability recap

- Sampling space: Ω
    - Collection of events
- Probability: $P(Y = 1)$
    - Frequency of event Y in Ω
- Conditional probability: $P(Y = 1|X)$
    - Frequency of event Y in subset of Ω identified by X

Probability

Example: F1 2021 World Championship

GP	Pole	Winner	GP	Pole	Winner
Bahrain	Verstappen	Hamilton	Belgian	Verstappen	Verstappen
Emilia Romagna	Hamilton	Verstappen	Dutch	Verstappen	Verstappen
Portuguese	Bottas	Hamilton	Italian	Verstappen	Ricciardo
Spanish	Hamilton	Hamilton	Russian	Norris	Hamilton
Monaco	Leclerc	Verstappen	Turkish	Bottas	Bottas
Azerbaijan	Leclerc	Pérez	United States	Verstappen	Verstappen
French	Verstappen	Verstappen	Mexico City	Bottas	Verstappen
Styrian	Verstappen	Verstappen	São Paulo	Bottas	Hamilton
Austrian	Verstappen	Verstappen	Qatar	Hamilton	Hamilton
British	Verstappen	Hamilton	Saudi Arabian	Hamilton	Hamilton
Hungarian	Hamilton	Ocon	Abu Dhabi	Verstappen	Verstappen

Probability

- Probability of Hamilton winning a race:

- Probability of Verstappen winning a race:

- Probability of other driver winning a race:

Probability

- Probability of Hamilton winning a race:

$$P(HAM) = \frac{\text{HAM wins}}{\#\ \text{races}} = \frac{8}{22} \approx 0.36$$

- Probability of Verstappen winning a race:

$$P(VER) = \frac{\text{VER wins}}{\#\ \text{races}} = \frac{10}{22} \approx 0.45$$

- Probability of other driver winning a race:

$$P(\neg VER \vee HAM) = \frac{\text{other driver wins}}{\#\ \text{races}} = \frac{4}{22} \approx 0.18$$

Probability

- Probability of Verstappen winning from pole:

Probability

- Probability of Verstappen winning from pole:

GP	Pole	Winner	GP	Pole	Winner
Bahrain	Verstappen	Hamilton	**Belgian**	**Verstappen**	**Verstappen**
Emilia Romagna	Hamilton	Verstappen	**Dutch**	**Verstappen**	**Verstappen**
Portuguese	Bottas	Hamilton	Italian	Verstappen	Ricciardo
Spanish	Hamilton	Hamilton	Russian	Norris	Hamilton
Monaco	Leclerc	Verstappen	Turkish	Bottas	Bottas
Azerbaijan	Leclerc	Pérez	**United States**	**Verstappen**	**Verstappen**
French	**Verstappen**	**Verstappen**	Mexico City	Bottas	Verstappen
Styrian	**Verstappen**	**Verstappen**	São Paulo	Bottas	Hamilton
Austrian	**Verstappen**	**Verstappen**	Qatar	Hamilton	Hamilton
British	Verstappen	Hamilton	Saudi Arabian	Hamilton	Hamilton
Hungarian	Hamilton	Ocon	**Abu Dhabi**	**Verstappen**	**Verstappen**

Probability

Bayes Rule

- Conditional probabilities $P(Y|X)$ and $P(X|Y)$:

$$P(Y|X) = \frac{P(X \cap Y)}{P(X)} \tag{1}$$

$$P(X|Y) = \frac{P(X \cap Y)}{P(Y)} \tag{2}$$

Probability

Bayes Rule

- From (1) and (2) follows:

$$P(X|Y) = \frac{P(Y|X)P(X)}{\mathbf{P}(\mathbf{Y})} = \frac{P(Y|X)P(X)}{\mathbf{P}(\mathbf{Y}|\mathbf{X})P(X) + \mathbf{P}(\mathbf{Y}|\neg\mathbf{X})P(\neg X)}$$

- For K partitions of the sampling space:

$$P(X|Y) = \frac{P(Y|X)P(X)}{\sum_{x_k=1}^{K} P(Y|X = x_k)P(X = x_k)}$$

Probability

- Probability of Verstappen starting on pole if winning:

$$P(pole|VER) = \frac{P(VER|pole)P(pole)}{P(VER|pole)P(pole) + P(VER|\neg pole)P(\neg pole)}$$

Probability

- Probability of Verstappen starting on pole if winning:

$$
\begin{aligned}
P(pole|VER) &= \frac{P(VER|pole)P(pole)}{P(VER|pole)P(pole) + P(VER|\neg pole)P(\neg pole)} \\
&= \frac{\overbrace{P(VER|pole)}^{\frac{7}{10}}\,\overbrace{P(pole)}^{\frac{10}{22}}}{\underbrace{(VER|pole)}_{\frac{7}{10}}\,\underbrace{P(pole)}_{\frac{10}{22}} + \underbrace{P(VER|\neg pole)}_{\frac{3}{12}}\,\underbrace{P(\neg pole)}_{\frac{12}{22}}} \\
&= \frac{\frac{7}{10} \times \frac{10}{22}}{\left(\frac{7}{10} \times \frac{10}{22}\right) + \left(\frac{3}{12} \times \frac{12}{22}\right)} \\
&= \frac{7}{10}
\end{aligned}
$$

Probability

Probability distributions

- Gaussian
- Standard Gaussian
- Lognormal
- Exponential
- Chi-square
- Bernoulli
- Binomial
- Geometric
- Poisson
- etc.

Probability

Probability Distributions: Functions

- Probability Distribution Function (PDF)

$$PDF_Y = f_Y(y) = P(Y = y)$$

- Cumulative Distribution Function (CDF)

$$CDF_Y = F_Y(y) = P(Y \leq y) = \int_Y f_Y(y)dy$$

Distributions

Probability Distributions: Moments

- Mean

$$\mu_Y = E[Y] = \int_Y \underbrace{y}_{\text{value of function}} \cdot \underbrace{f_Y(y)}_{\text{PDF evaluated at value}} dy$$

- Variance

$$\sigma_Y^2 = E[(Y - \mu_Y)^2] = \int_Y \underbrace{(Y - \mu_Y)^2}_{\text{delta from mean}} \cdot \underbrace{f_Y(y)}_{\text{PDF evaluated at value}} dy$$

Distributions

- Example: $PDF_{Income} \sim \mathcal{N}(\mu = \$50,000, \sigma = \$20,000)$

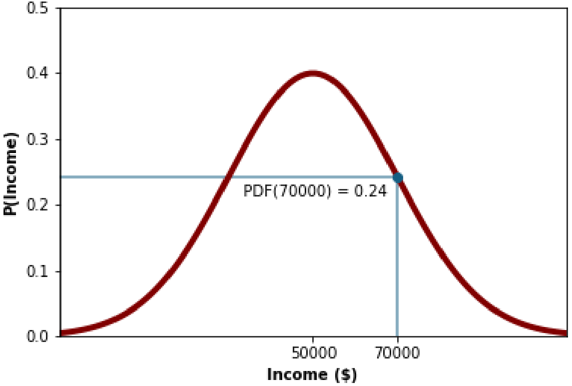

Figure: Probability distribution function of income

Distributions

- Example: $PDF_{Income} \sim \mathcal{N}(\mu = \$50,000, \sigma = \$20,000)$

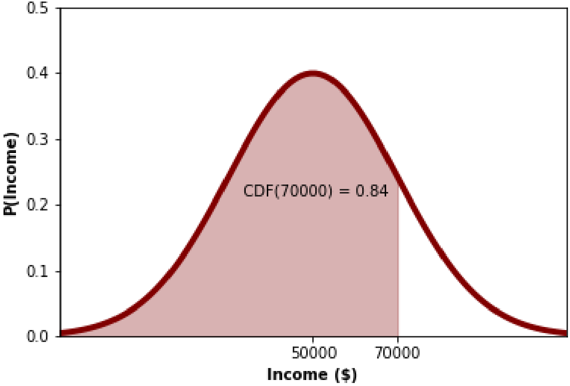

Figure: Cumulative distribution function of income

Probability

Policy example

- A policy analyst run the numbers on two high school coursework policies. What would the preferred policy be?

Policy	**Scenario**	**Probability**	**Earnings ($)**
A	Worst-case	0.1	40,000
	Most-common	0.7	50,000
	Best-case	0.2	100,00
B	Worst-case	0.2	30,000
	Most-common	0.5	60,000
	Best-case	0.3	80,000

Table: Expected student outcomes for coursework policies I and II

Probability

- Calculate expected outcomes:

$$\mu_A = E[A] = \$59,000$$
$$\mu_B = E[B] = \$60,000$$

- Calculate variances:

$$\sigma_A^2 = E[(A - \mu_A)^2] = \$429,000$$
$$\sigma_B^2 = E[(B - \mu_B)^2] = \$300,000$$

- Conclusion: Choose policy B because its has higher expected payoff and lower variance than policy A

Probability

- PDF question: What is the probability of a graduate from Policy B making $\$60,000$?

$$P(B = \$60,000) = F_B(\$60,000) = 0.5$$

- CDF question: What is the probability of a graduate from Policy B making $\$60,000$ or less?

$$P(B \leq \$60,000) = F_B(\$60,000) = 0.5 + 0.2 = 0.7$$

Table of Contents

Ordinary Least Squares

Linear regression problem

1. Express Y as a weighted combination of covariates X:

$$Y = \beta X + \epsilon$$

2. Minimize sum of squared errors SSE:

$$\min_{\beta} SSE = \epsilon^2 = (Y - \hat{Y})^2 = \sum (Y - \beta X)^2$$

Ordinary Least Squares

3. Obtain closed-form expression for parameters:
   - Bivariate regression

$$\beta_X = \frac{(Y - \overline{Y})(X - \overline{X})}{\sum(X - \overline{X})^2} = \frac{S_{XY}}{S_{XX}}$$

   - Multivariate regression

$$\beta_{X|Z} = \frac{S_{ZZ}S_{XY} - S_{XZ}S_{ZY}}{S_{XX}S_{ZZ} - S_{XZ}}$$

     - where $S_{ZZ} = (Z - \overline{Z})^2, S_{XZ} = (X - \overline{X})(Z - \overline{Z})$, and $S_{ZY} = (Z - \overline{Z})(Y - \overline{Y})$

Ordinary Least Squares

- When $\rho(X, Z) = 0$, then $S_{XZ} = 0$:

$$\beta_{X|Z} = \frac{S_{XY}S_{ZZ} - 0 \cdot S_{ZY}}{S_{XX}S_{ZZ} - 0} = \frac{S_{XY}}{S_{XX}}$$

  - Multivariate coefficient reduces to bivariate coefficient!

- In general, multivariate coefficients capture:
  - **Partial correlation** (i.e., variation left between X and Y after removing variation shared with other predictors)

Ordinary Least Squares

- Standardized regression coefficients
  - Bivariate regression

$$\beta_X = \beta_X \left(\frac{\sigma_X}{\sigma_Y} \right) = \rho(X, Y)$$

  - Multivariate regression

$$\beta_{X|Z} = \beta_{X|Z} \left(\frac{\sigma_X}{\sigma_Y} \right) \neq \rho(X, Y)$$

Ordinary Least Squares

- Error terms

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-p} \sim \chi_{n-p}$$

  - where p is the # of parameters (<u>must include intercept!</u>)

- Standard errors of estimate
  - Bivariate regression: $\hat{\sigma}^2_{\beta_X} = \frac{\hat{\sigma}^2}{S_{XX}}$
  - Multivariate regression: $\hat{\sigma}^2_{\beta_{X|Z}} = \frac{\hat{\sigma}^2}{S_{XX}(1-\hat{\rho}^2_{XZ})}$

Ordinary Least Squares

- Sum of squares

Model	**Total**	**Regression**	**Error**
df	df_T n	df_R p	df_E $n-p$
Total	SST $\sum(Y-\overline{Y})$	SSR $\sum(\hat{Y}-\overline{Y})$	SSE $\sum(\hat{Y}-Y)$
Mean	MST SST/df_T	MSR SSR/df_R	MSE SSE/df_E

Ordinary Least Squares

- Sum of squares

$$SST = SSR + SSE$$

- Overall Significance: F-statistic (F)

$$F = \frac{MSR}{MSE} \sim F_{df_R, df_E = n-p}$$

- Parameter significance: t-statistic (t) and p-value (p)

$$t = \frac{\hat{\beta} - 0}{\hat{\sigma}_{\beta}} = \frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}}$$

$$p = 2 \cdot PDF_{t_{df_E}}(t)$$

Ordinary Least Squares

Assumptions

- **Linearity**
- **Independence**
- **Homoskedasticity**
- **Normality**

Ordinary Least Squares

Assumptions

- **Linearity**
  - The outcome is linearly related to the predictors
- **Independence**
  - Observations are i.i.d.
- **Homoskedasticity**
  - Error terms are homoskedastic
- **Normality**
  - Errors terms are normally distributed

Table of Contents

Generalized Linear Models

- Nonlinear link function $h(\cdot)$:

$$Y = h(\beta X + \epsilon)$$

Model	**Link function**	**Coefficient**	**Application**
Normal	μ	Units change	Measures
Logit	$\log\left(\frac{\mu}{1-\mu}\right)$	Log-Odds change	Responses
Probit	$\phi^{-1}(\mu)$	Z-score change	Responses
Poisson	$\log(\mu)$	Incidence Rate Ratio	Counts
Negative Binomial	$\log\left(\frac{\mu}{k(1-m/k)}\right)$	Incidence Rate Ratio	Counts

Generalized Linear Models

Logistic regression

- Canonical form

$$y_i = \frac{1}{1+e^{-\beta X}} = \frac{e^{\beta X}}{1+e^{\beta X}}$$

- Linear form

$$\ln\left(\frac{p}{1-p}\right) = \beta X$$

Generalized Linear Models

Log-odds to probabilities

- STATA's *margins*
    - Average Partial Effect (APE)
        - Average derivative across the logistic curve
    - Marginal Effect at the Mean (MEM)
        - Derivative at the mean of the logistic curve

Generalized Linear Models

Maximum Likelihood Estimation (MLE)

- Likelihood function

$$\mathcal{L} = \prod_{i=1}^{N} p_i^{y_i}(1 - p_i)^{1-y_i}$$

Generalized Linear Models

Maximum Likelihood Estimation (MLE)

- Likelihood function

$$\mathcal{L} = \prod_{i=1}^{N} p_i^{y_i}(1 - p_i)^{1-y_i}$$

- Substitute probabilities and take natural log:

$$\begin{aligned}
\log(\mathcal{L}) &= \log \prod_{i=1}^{N} \left(\frac{1}{1+e^{-\beta X}}\right)^{y_i} \left(1 - \frac{1}{1+e^{-\beta X}}\right)^{1-y_i} \\
&= \sum_{i=1}^{N} y_i \log\left(\frac{1}{1+e^{-\beta X}}\right) + (1-y_i)\log\left(1 - \frac{1}{1+e^{-\beta X}}\right)
\end{aligned}$$

Generalized Linear Models

- Maximize log-Likelihood with respect to coefficients:

$$\frac{\partial \log(\mathcal{L})}{\partial \beta} = 0$$

- Closed-form solution <u>unavailable</u>

Generalized Linear Models

- Likelihood-Ratio test: Base vs new model

$$G = -2\log\left(\frac{\mathcal{L}(\beta_{new})}{\mathcal{L}(\beta_{base})}\right) = -2\Bigg(\log\mathcal{L}(\beta_{new}) - \log\mathcal{L}(\beta_{base})\Bigg)$$

Generalized Linear Models

- Overall significance: Full vs intercept-only model (G)

$$G = -2\log\left(\frac{\mathcal{L}(\beta_{full})}{\mathcal{L}(\beta_{null})}\right) = -2\Big(\log\mathcal{L}(\beta_{full}) - \log\mathcal{L}(\beta_{null})\Big)$$

- Predictor significance: z-statistic (z) and p-value (p)

$$z = \frac{\hat{\beta}_k}{\hat{\sigma}_{\beta_k}} \sim \mathcal{N}(0,1)$$
$$p = 1 - 2\phi(z)$$

Output Interpretation

- Regression equation

$$y_i = \beta_1 x_i + \beta_2 x_i^2 + \beta_3 z_i + \beta_4 (x_i \times z_i) + \epsilon_i$$

  - where x_i continuous, z_i binary (1 = group A, 0 = B)

Output Interpretation

- Regression equation

$$y_i = \beta_1 x_i + \beta_2 x_i^2 + \beta_3 z_i + \beta_4 (x_i \times z_i) + \epsilon_i$$

  - where x_i continuous, z_i binary ($1 =$ group A, $0 =$ B)

- Effects
  - β_1, β_3: Main effects
  - β_2: Quadratic effect
  - β_4: Interaction effect

Output Interpretation

- Quadratic effect

$$\frac{\partial \hat{y}_i}{\partial x_i} = \beta_1 + 2x_i\beta_2$$

- Interpretation: **Marginal rate**
  - $\beta_2 > 0$: Increasing rate
  - $\beta_2 < 0$: Decreasing rate
  - $-\frac{1}{2}\frac{\beta_1}{\beta_2}$: Max/Min

Output Interpretation

- Interaction effect
  - Group A: $\frac{\partial \hat{y}_i}{\partial x_i} = \beta_2$
  - Group B: $\frac{\partial \hat{y}_i}{\partial x_i} = \beta_2 + \beta_4$
- Interpretation: **Differential rate**
  - $\beta_4 > 0$: Group A has higher rate than group B
  - $\beta_4 < 0$: Group A has lower rate than group B

- OLS: Athletics and SAT (Bremmer & Kesselring, 1993)

VARIABLES	COEFFICIENTS	t-scores
Constant	966.734[a]	11.588
Sports	11.003	0.659
Tuition	0.010[a]	2.705
Volumes	–0.001	–0.215
Salary	0.353[a]	3.096
Age	–0.007	–0.047
Students/faculty	1.782	1.095
Enrollment	–0.001	–0.275
Endowment/Students	5.036[c]	1.862
Ph.D.'s/Students	–0.191	–0.453
State SAT	0.016	0.340
Accept Percent	–310.133[a]	–7.530
Football	0.802	0.280
Basketball	0.030	0.013
F	41.443	
R^2	0.837	
Adjusted R^2	0.817	
Chow test F	6.305	
n	119	

Figure: Predictors of freshman cohort average SAT

- OLS: Returns to education (Heckman et al., 2006)

		Whites		Blacks	
		Coefficient	Std. Error	Coefficient	Std. Error
1940	Intercept	4.4771	0.0096	4.6711	0.0298
	Education	0.1250	0.0007	0.0871	0.0022
	Experience	0.0904	0.0005	0.0646	0.0018
	Experience-squared	−0.0013	0.0000	−0.0009	0.0000
1950	Intercept	5.3120	0.0132	5.0716	0.0409
	Education	0.1058	0.0009	0.0998	0.0030
	Experience	0.1074	0.0006	0.0933	0.0023
	Experience-squared	−0.0017	0.0000	−0.0014	0.0000
1960	Intercept	5.6478	0.0066	5.4107	0.0220
	Education	0.1152	0.0005	0.1034	0.0016
	Experience	0.1156	0.0003	0.1035	0.0011
	Experience-squared	−0.0018	0.0000	−0.0016	0.0000
1970	Intercept	5.9113	0.0045	5.8938	0.0155
	Education	0.1179	0.0003	0.1100	0.0012
	Experience	0.1323	0.0002	0.1074	0.0007
	Experience-squared	−0.0022	0.0000	−0.0016	0.0000
1980	Intercept	6.8913	0.0030	6.4448	0.0120
	Education	0.1023	0.0002	0.1176	0.0009
	Experience	0.1255	0.0001	0.1075	0.0005
	Experience-squared	−0.0022	0.0000	−0.0016	0.0000
1990	Intercept	6.8912	0.0034	6.3474	0.0144
	Education	0.1292	0.0002	0.1524	0.0011
	Experience	0.1301	0.0001	0.1109	0.0006
	Experience-squared	−0.0023	0.0000	−0.0017	0.0000

Figure: Predictors of earnings (log)

- OLS: Democracy and GDP (Saha et al., 2009)

	(1)	(2)	(3)
DEMO	0.463*** (0.061)	0.435*** (0.055)	0.104* (0.061)
EF	−0.721*** (0.036)	−0.465*** (0.039)	−0.471*** (0.040)
DEMO*EF	−0.098*** (0.009)	−0.085*** (0.008)	−0.019** (0.011)
Log(RGDP)		−0.884*** (0.079)	−0.825*** (0.089)
Gini index		0.028*** (0.005)	0.045*** (0.006)
Unemployment		0.017*** (0.005)	0.019*** (0.004)
Literacy rate		0.029*** (0.003)	−0.004 (0.004)
Latin America			1.005*** (0.378)
Middle East			0.39 (0.372)
East Asia			1.924*** (0.419)
South East Asia			1.067*** (0.385)
South Asia			1.113*** (0.399)
Eastern Europe			2.076*** (0.394)
Central Asia			1.426*** (0.422)
Africa			−0.213 (0.372)
Western Europe			0.844** (0.421)
Northern Europe			−0.419 (0.443)
North America			−0.032 (0.483)
Australasia			−0.56 (0.496)
Constant	9.563*** (0.145)	11.91*** (0.659)	12.523*** (0.866)
Number of observations	981	978	978
Adj *R*-squared	0.72	0.78	0.84

Figure: Predictors of GDP growth rate

- Logit: Gambling Laws (Richard, 2010)

	β	SE	Exp(β)	Significance
Constant	−20.162**	8.030	0.000	0.012
INCOME	0.201*	0.111	1.223	0.071
FISCAL	−0.004	2.725	0.996	0.883
UNEMPL	0.683***	0.246	1.981	0.005
TOURISM	0.185*	0.095	1.203	0.051
RELIGION	−0.115***	0.042	0.891	0.006
Nagelkerke R^2		0.398		
−2 Log likelihood		37.345		Significance
χ^2 (6)		20.767***		0.002

Figure: Predictors of casino legalizations in the world

- Negative binomial: Juvenile crime rates (Osgood, 2000)

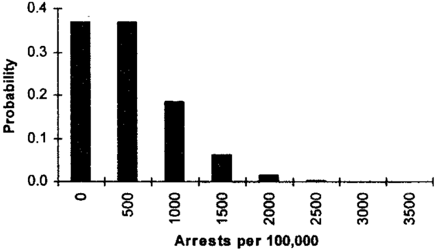

Figure: Probability distribution of juvenile arrests in non-metropolitan areas

- Negative binomial: Juvenile crime rates (Osgood, 2000)

Explanatory variable	Statistical method: OLS, rate/100,000	OLS, log(rate + 1)	OLS, log(rate + 0.2)	Basic Poisson	Negative binomial
Log population at risk					
b	11.220	0.749	1.102	1.501[a]	1.718[a]
SE	3.838	0.128	0.177	0.061	0.188
t	2.923	5.852	6.226	8.213	3.819
P	0.004	0.000	0.000	0.000	0.000
Residential instability					
b	35.573	3.017	4.366	1.567	0.162
SE	48.790	1.628	2.255	0.567	2.026
t	0.729	1.853	1.936	2.764	0.080
P	0.467	0.065	0.054	0.005	0.936
Ethnic heterogeneity					
b	63.839	2.461	3.325	2.069	2.861
SE	32.711	1.091	1.512	0.419	1.156
t	1.952	2.256	2.199	4.938	2.475
P	0.052	0.025	0.029	0.000	0.013
Female-headed households					
b	22.765	0.533	0.192	3.919	3.739
SE	71.679	2.391	3.313	1.030	2.937
t	0.318	0.223	0.058	3.805	1.273
P	0.751	0.824	0.954	0.000	0.203
Poverty rate					
b	39.474	1.405	2.181	0.499	0.021
SE	81.162	2.708	3.752	1.009	3.381
t	0.486	0.519	0.581	0.495	0.006
P	0.627	0.604	0.561	0.621	0.995

Figure: Predictors of juvenile arrests in non-metropolitan areas