Advanced Statistics
DS2003 (BDS-4A)
Lecture 26

Instructor: Dr. Syed Mohammad Irteza

Assistant Professor, Department of Computer Science, FAST

26 May, 2022

Previous Lecture

- Checking Model Conditions Using Graphs
- Transformations (model improvement)
- Logistic Regression
- Donner Party
- Concept of Odds
- Logit Function

Example - Donner Party - Interpretation

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.8185	0.9994	1.82	0.0688
Age	-0.0665	0.0322	-2.06	0.0391

Simple interpretation is only possible in terms of *log odds* and *log odds ratios* for intercept and slope terms.

Intercept: The *log odds* of survival for a party member with an age of 0. From this we can calculate the odds or probability, but additional calculations are necessary.

Slope: For a unit increase in age (being 1 year older) how much will the *log odds ratio* change, not particularly intuitive. More often than not we care only about sign and relative magnitude.

Hypothesis tests for a coefficient

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.6331	1.1102	1.47	0.1413
Age	-0.0782	0.0373	-2.10	0.0359
SexFemale	1.5973	0.7555	2.11	0.0345

We are however still able to perform inference on individual coefficients, the basic setup is exactly the same as what we've seen before except we use a Z-test.

Note: The only tricky bit, which is way beyond the scope of this course, is how the standard error is calculated.

Testing for the slope of Age

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.6331	1.1102	1.47	0.1413
Age	-0.0782	0.0373	-2.10	0.0359
SexFemale	1.5973	0.7555	2.11	0.0345

$$H_0 : \beta_{age} = 0$$

$$H_A : \beta_{age} \neq 0$$

$$Z = \frac{\hat{\beta_{age}} - \beta_{age}}{SE_{age}} = \frac{-0.0782 - 0}{0.0373} = -2.10$$

$$\text{p-value} = P(|Z| > 2.10) = P(Z > 2.10) + P(Z < -2.10)$$

$$= 2 \times 0.0178 = 0.0359$$

Confidence interval for age slope coefficient

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.6331	1.1102	1.47	0.1413
Age	-0.0782	0.0373	-2.10	0.0359
SexFemale	1.5973	0.7555	2.11	0.0345

Remember, the interpretation for a slope is the change in log odds ratio per unit change in the predictor.

Log odds ratio:

$$CI = PE \pm CV \times SE = -0.0782 \pm 1.96 \times 0.0373 = (-0.1513, -0.0051)$$

Odds ratio: [note: typo below, (0.85960, 0.9949) – exp is Euler's number, 2.71828^x]

$$\exp(CI) = (\exp -0.1513, \exp -0.0051) = (0.85960.9949)$$

Example - Birdkeeping and Lung Cancer

A 1972 - 1981 health survey in The Hague, Netherlands, discovered an association between keeping pet birds and increased risk of lung cancer. To investigate birdkeeping as a risk factor, researchers conducted a case-control study of patients in 1985 at four hospitals in The Hague (population 450,000). They identified 49 cases of lung cancer among the patients who were registered with a general practice, who were age 65 or younger and who had resided in the city since 1965. They also selected 98 controls from a population of residents having the same general age structure.

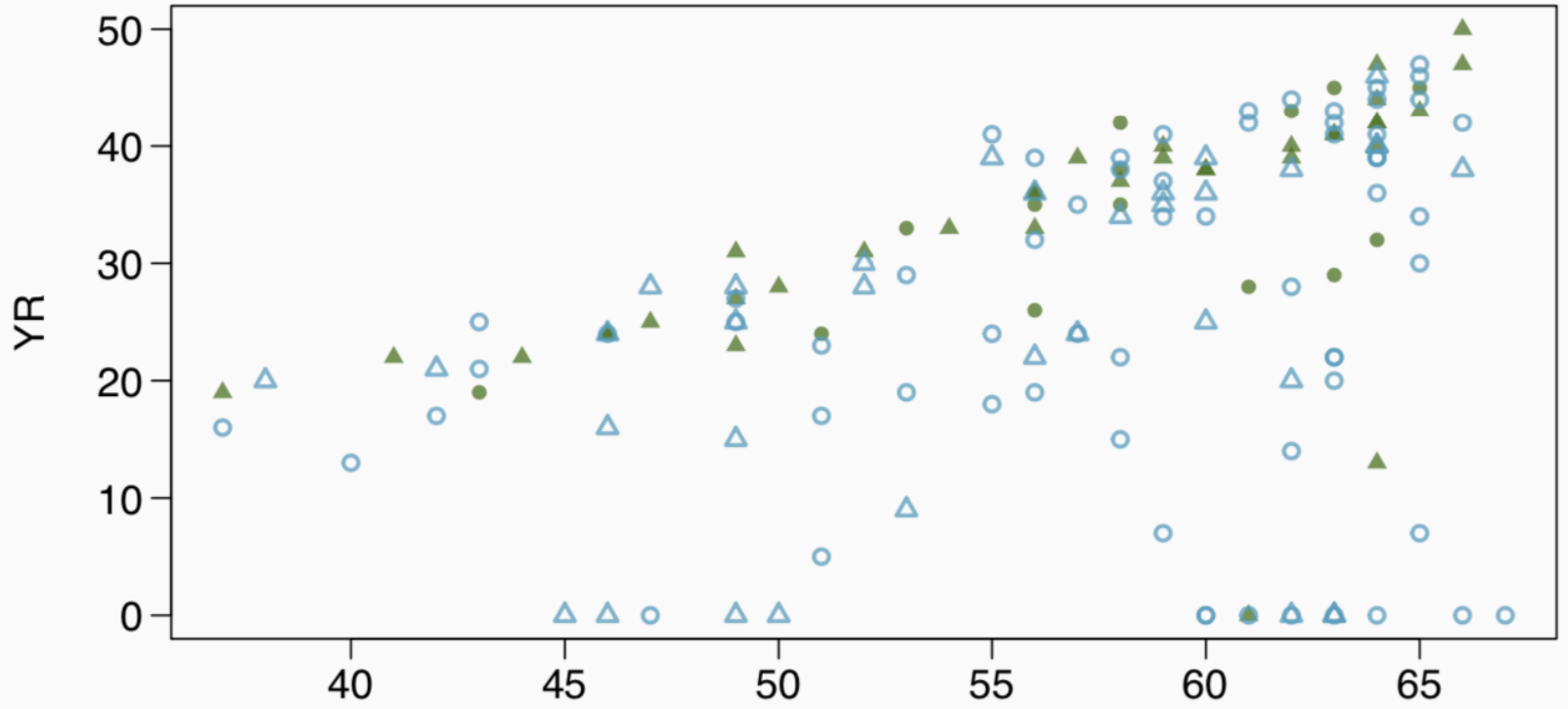
From *Ramsey, F.L. and Schafer, D.W. (2002). The Statistical Sleuth: A Course in Methods of Data Analysis (2nd ed)*

Example - Birdkeeping and Lung Cancer - Data

	LC	FM	SS	BK	AG	YR	CD
1	LungCancer	Male	Low	Bird	37.00	19.00	12.00
2	LungCancer	Male	Low	Bird	41.00	22.00	15.00
3	LungCancer	Male	High	NoBird	43.00	19.00	15.00
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
147	NoCancer	Female	Low	NoBird	65.00	7.00	2.00

- `LC` Whether subject has lung cancer
- `FM` Sex of subject
- `SS` Socioeconomic status
- `BK` Indicator for birdkeeping
- `AG` Age of subject (years)
- `YR` Years of smoking prior to diagnosis or examination
- `CD` Average rate of smoking (cigarettes per day)

Example - Birdkeeping and Lung Cancer - EDA



	Bird	No Bird
Lung Cancer	▲	●
No Lung Cancer	△	○

Example - Birdkeeping and Lung Cancer - Model

```
summary(glm(LC ~ FM + SS + BK + AG + YR + CD, data=bird, family=binomial))
## Call:
## glm(formula = LC ~FM + SS + BK + AG + YR + CD, family = binomial,
##     data = bird)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.93736  1.80425    -1.074 0.282924
## FMFemale     0.56127  0.53116     1.057 0.290653
## SSHigh       0.10545  0.46885     0.225 0.822050
## BKBird       1.36259  0.41128     3.313 0.000923 ***
## AG          -0.03976  0.03548    -1.120 0.262503
## YR           0.07287  0.02649     2.751 0.005940 **
## CD           0.02602  0.02552     1.019 0.308055
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 187.14  on 146  degrees of freedom
## Residual deviance: 154.20  on 140  degrees of freedom
## AIC: 168.2
```

Example - Birdkeeping and Lung Cancer - Interp

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.9374	1.8043	-1.07	0.2829
FMFemale	0.5613	0.5312	1.06	0.2907
SSHigh	0.1054	0.4688	0.22	0.8221
BKBird	1.3626	0.4113	3.31	0.0009
AG	-0.0398	0.0355	-1.12	0.2625
YR	0.0729	0.0265	2.75	0.0059
CD	0.0260	0.0255	1.02	0.3081

Keeping all other predictors constant then,

- The odds ratio of getting lung cancer for bird keepers vs non-bird keepers is exp(1.3626) = 3.91.
- The odds ratio of getting lung cancer for an additional year of smoking is exp(0.0729) = 1.08.

What do the numbers not mean ...

The most common mistake made when interpreting logistic regression is to treat an odds ratio as a ratio of probabilities.

Bird keepers are not 4x more likely to develop lung cancer than non-bird keepers.

This is the difference between relative risk and an odds ratio.

$$RR = \frac{P(\text{disease}|\text{exposed})}{P(\text{disease}|\text{unexposed})}$$

$$OR = \frac{P(\text{disease}|\text{exposed})/[1 - P(\text{disease}|\text{exposed})]}{P(\text{disease}|\text{unexposed})/[1 - P(\text{disease}|\text{unexposed})]}$$

Back to the birds

What is probability of lung cancer in a bird keeper if we knew that P(lung cancer|no birds) = 0.05?

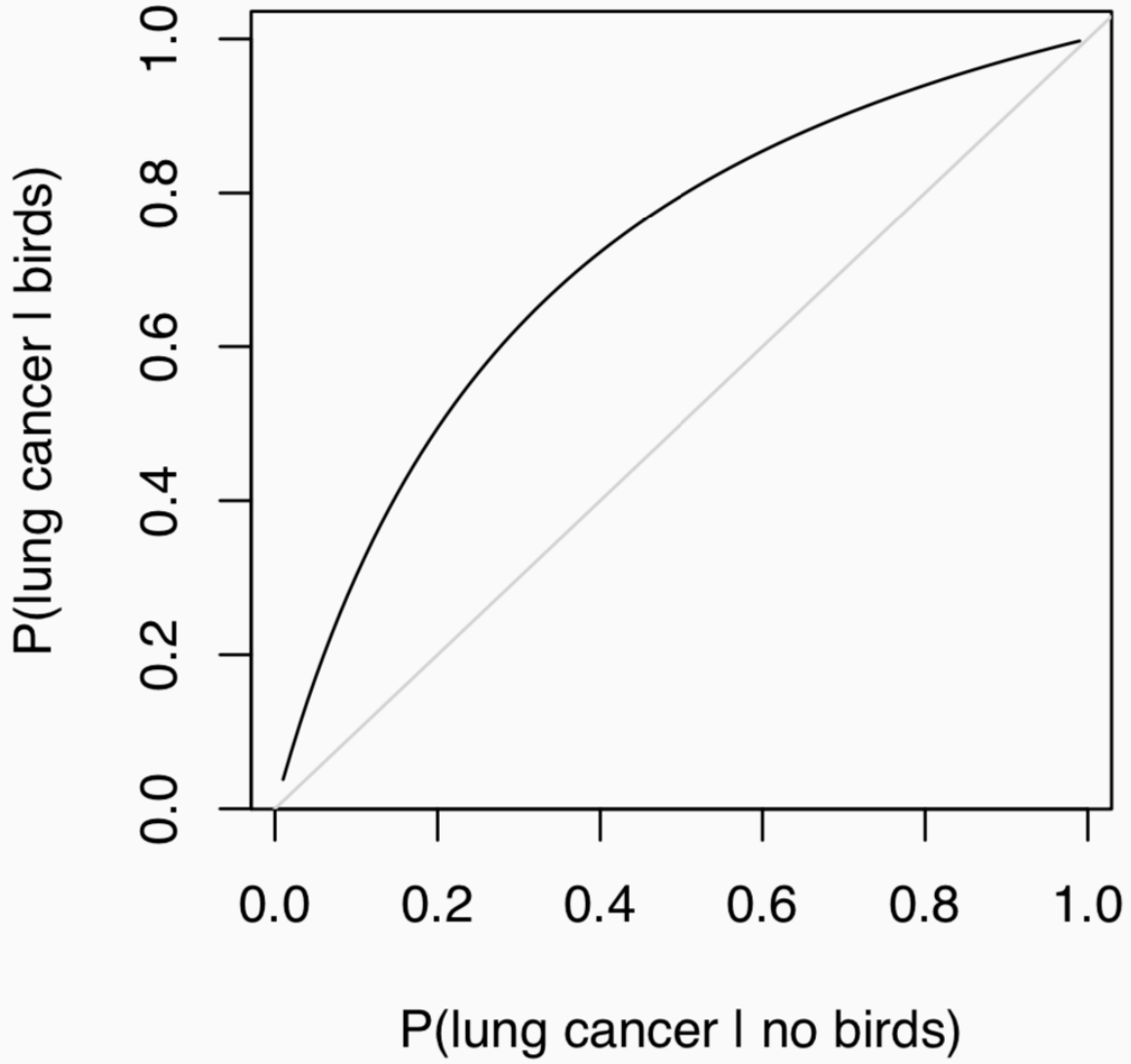
$$OR = \frac{P(\text{lung cancer}|\text{birds})/[1 - P(\text{lung cancer}|\text{birds})]}{P(\text{lung cancer}|\text{no birds})/[1 - P(\text{lung cancer}|\text{no birds})]}$$

$$= \frac{P(\text{lung cancer}|\text{birds})/[1 - P(\text{lung cancer}|\text{birds})]}{0.05/[1 - 0.05]} = 3.91$$

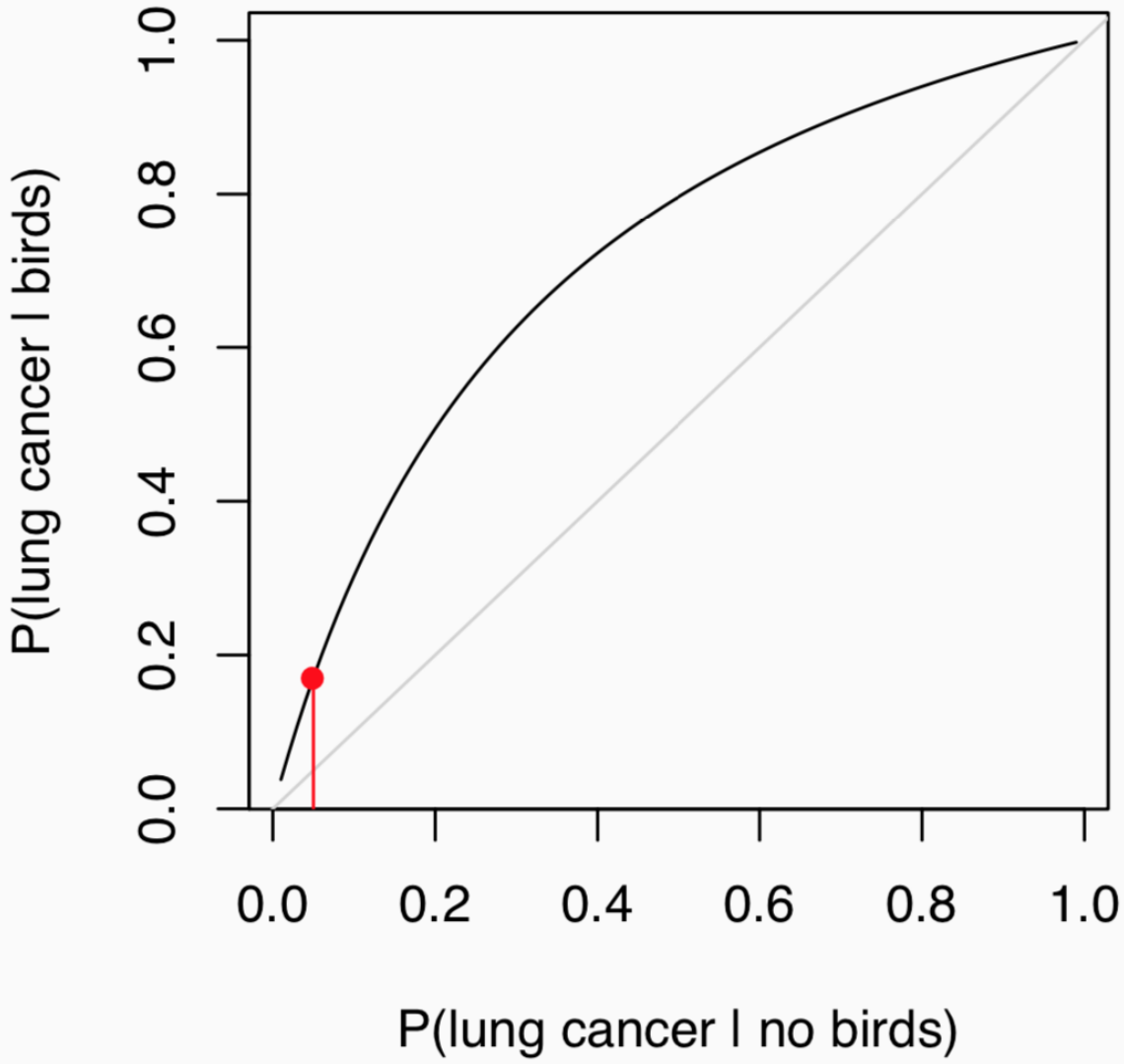
$$P(\text{lung cancer}|\text{birds}) = \frac{3.91 \times \frac{0.05}{0.95}}{1 + 3.91 \times \frac{0.05}{0.95}} = 0.171$$

$$RR = P(\text{lung cancer}|\text{birds})/P(\text{lung cancer}|\text{no birds}) = 0.171/0.05 = 3.41$$

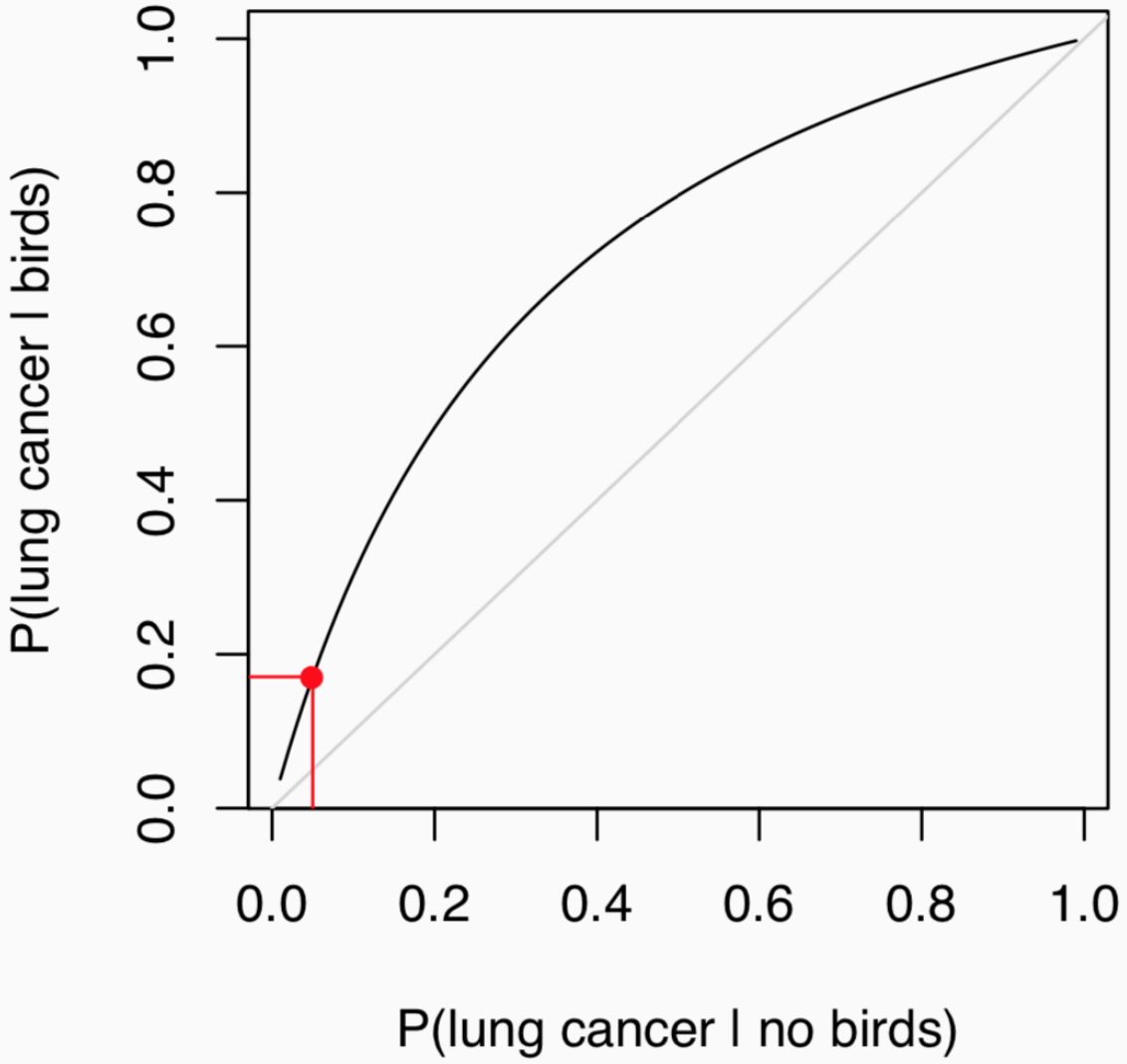
Bird OR Curve



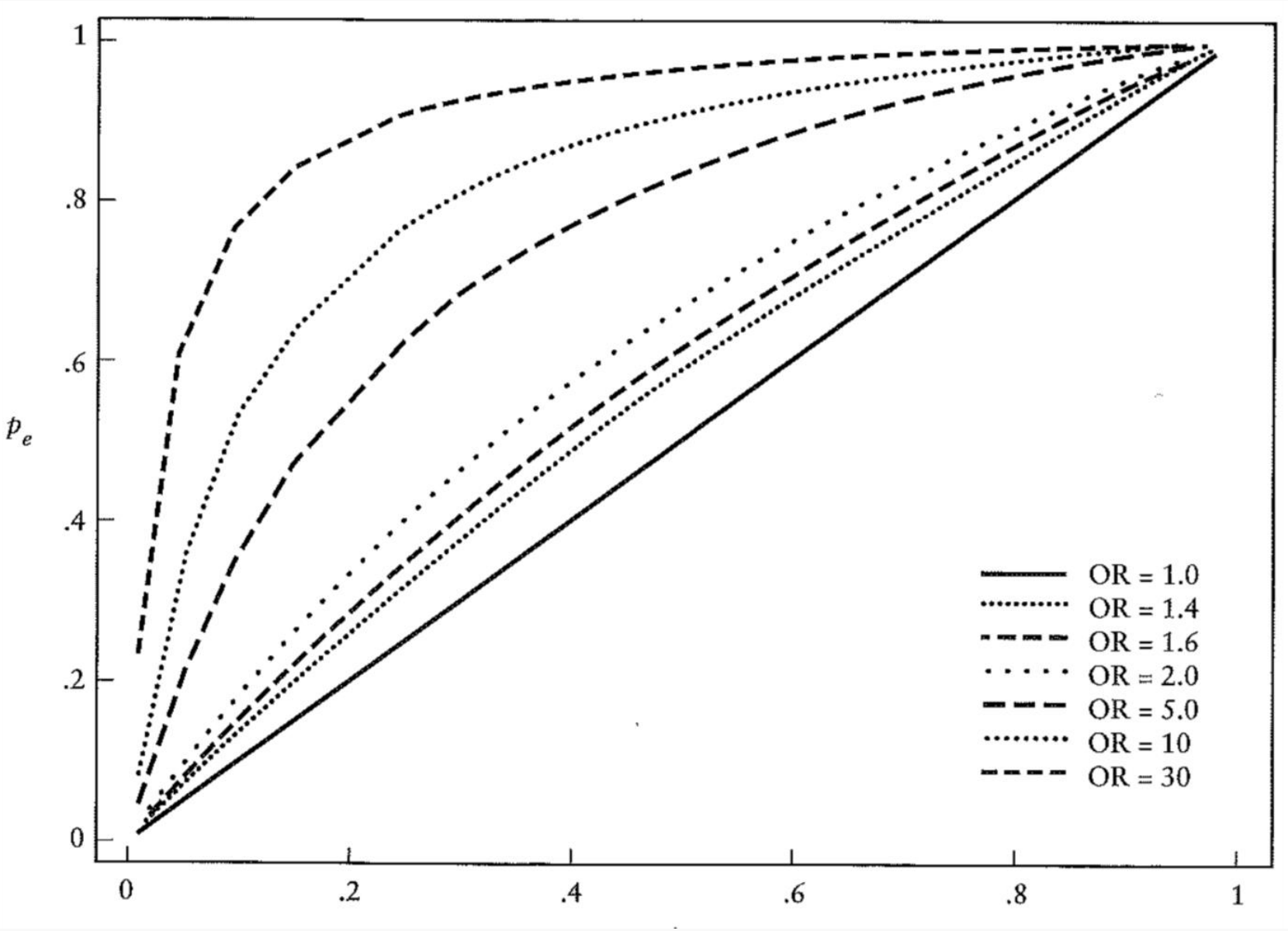
Bird OR Curve



Bird OR Curve



OR Curves



Useful Links & Resources

- **Reference**:
  - openintro.org/os (Chapter 9, Section 9.5)
- **Further Helpers:**
  - https://www.mathsisfun.com/numbers/e-eulers-number.html
  - https://www.youtube.com/watch?v=zAULhNrnuL4