Stat 102

Introduction to Business Statistics

Class 20

Richard P. Waterman

Wharton

Table of contents I

Today's module

Topics to be covered in this module:

- Last time
- Train and test paradigm
- Introduction to logistic regression
- Summary
- Next time

Last time

- The need for tools like stepwise – the space of all models is typically too big to exhaustively explore.
- The mechanics of stepwise; stopping rules, variable selection criterion.
- The big issue with stepwise: over-fitting.
- The Information criteria that penalize complexity.
- How JMP treats categoricals in stepwise.

Crossvalidation

- Two questions that crossvalidation can help with:

1. How do I think my model will perform when it sees new data – estimate the test error.
2. Of these competing models, which one should I prefer – the one with the lowest test error.

Test and training errors

- The **training error** is the average error computed on the same data that was used to fit the model.
- The **test error** is the average error when the model is applied to new data.
- We saw last class with the Apple dataset that these can be radically different and in general TRAINING ERROR $<$ TEST ERROR.
- Unfortunately we don't always have a test dataset to calculate the test error.
- The AIC and BIC methods from last class can be understood as calculations to estimate the TEST error directly, even when there is no test dataset. Choose the model with the lowest test error, i.e. lowest AIC or lowest BIC.
- Crossvalidation is a way of using the available dataset to get an estimate of the test error, but doing so directly, rather than through the AIC/BIC calculation approach.

The validation approach

- Randomly split the dataset into two parts.
- The part that is used to fit the model is the training dataset.
- The part that is used to estimate the test error, is the validation dataset. (JMP uses slightly different terminology.)
- Often these are called *in-sample* and *out-of-sample*.
- When it is a regression model, Mean Squared Error (MSE) is the typical measure of error.
- MSE $= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2/n$, but this is calculated over the out-of-sample data component.

Training and test sample

Training sample rows {1,3,5,7,10}. Test sample rows {2,4,6,8,9}. Implementation: fit the model on the training sample alone. Estimate the error on the test sample alone.

i	x	Training	Test	y_i	$\hat{y}_i$	$(y_i - \hat{y}_i)^2$
1	0.17	0.17	0.17	355	372.21	
2	0.16	0.16	0.16	328	339.47	131.58
3	0.17	0.17	0.17	350	372.21	
4	0.18	0.18	0.18	325	404.95	6391.63
5	0.25	0.25	0.25	642	634.12	
6	0.16	0.16	0.16	342	339.47	6.40
7	0.15	0.15	0.15	322	306.73	
8	0.19	0.19	0.19	485	437.69	2238.61
9	0.21	0.21	0.21	483	503.16	406.54
10	0.15	0.15	0.15	323	306.73	

The blue cells are excluded from the calculations, the white cells are included: first for the training data and then for the test data.

Example using the car dataset and train/test approach

- Erase the excluded rows from the dataset.
- Split the remaining data in two halves randomly with the Col Shuffle() formula command.
- Temporarily Exclude the test sample.
- Use stepwise with AICc as the model selection criteria and minimize it.
- Bring back all the data with the Clear row states command.
- Create a new column with the squared forecast error.
- Calculate the training RMSE and the test RMSE using the by command from the Distribution menu.
- JMP-Pro has the command Column→New column→Initialize data→Random→Random Indicator to simplify the construction of training and validation datasets.
- We can also do this using the Column→Modeling Utilities→Make Validation Column and then within the Fit Model platform, drop in the Validation column to the Validation field.

Crossvalidation output

Crossvalidation

Source	RSquare	RASE	Freq
Training Set	0.7773	4.1174	377
Validation Set	0.7517	4.0704	122

- The acronym RASE is the same as RMSE.
- Ideally, there is not much difference between the training and the test/validation summaries.
- If the out-of-sample R^2 is much smaller than the in-sample and the out-of-sample RMSE is much larger than the in-sample, then the model has been overfit.

Introduction to Logistic Regression

- There is an outcome variable Y that takes on two values YES, NO.
- There is a set of feature variables that can be used to predict which value Y takes on.
- For a new observation, for which I only know the X's, what do I think Y is going to be, YES or NO?

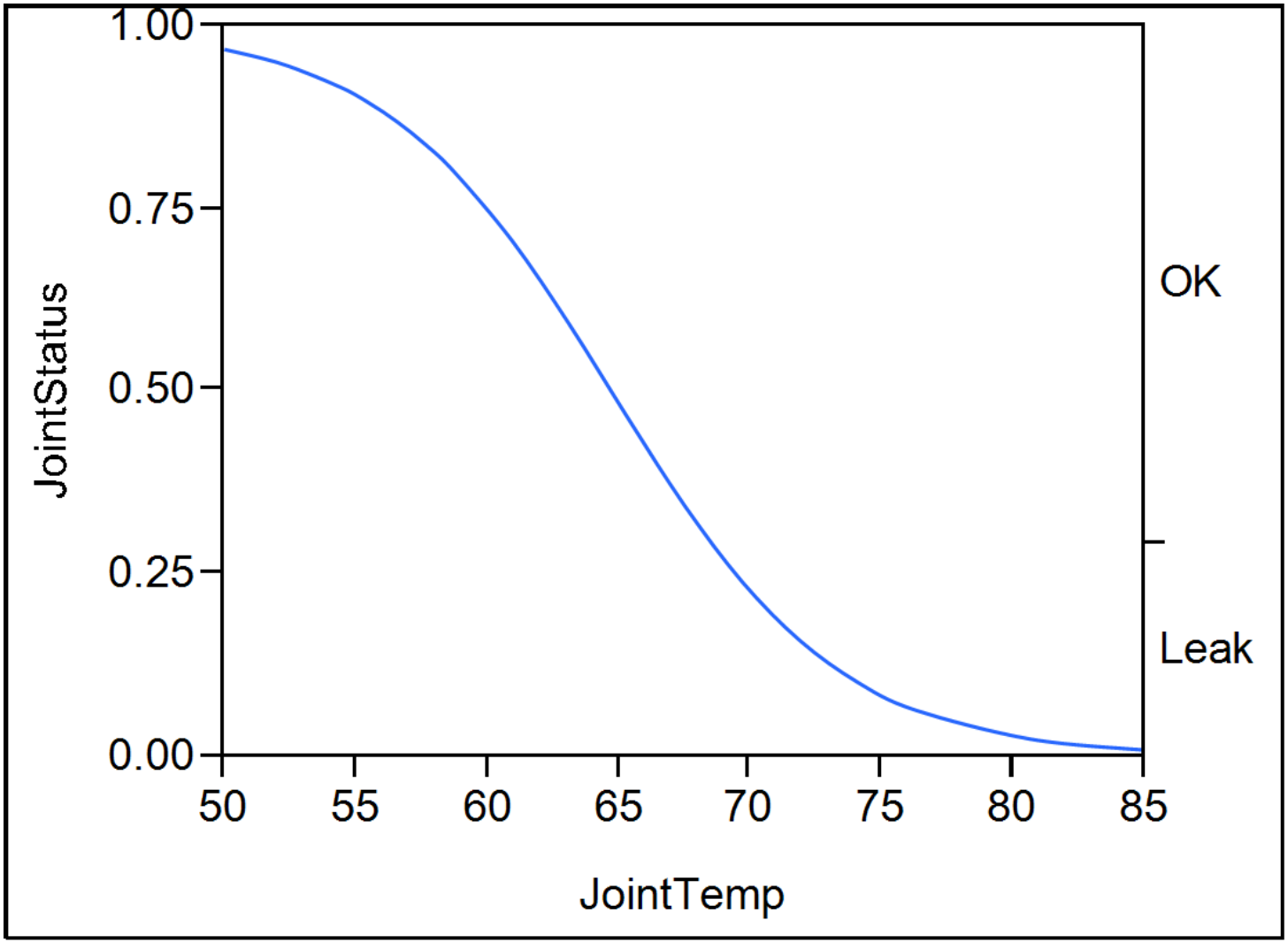
Logistic regression

- Very popular in practice, and with a good grounding in regular regression it is not a huge step.
- It models the probabilities of group membership directly $P(Y = 1|X) = p(x)$.
- We still want a linear model but a straight line will not respect the range of probabilities (which must lie between 0 and 1).
- Solution: transform the probabilities before modeling them.
- A very special transformation: the logit.

$$logit(p(x)) = \log\left(\frac{p(x)}{1 - p(x)}\right).$$

- If the probability is modeled as linear in X, i.e. $logit(p(x)) = \beta_0 + \beta_1 X$, then on the probability scale we see ...

A graph of the logistic function



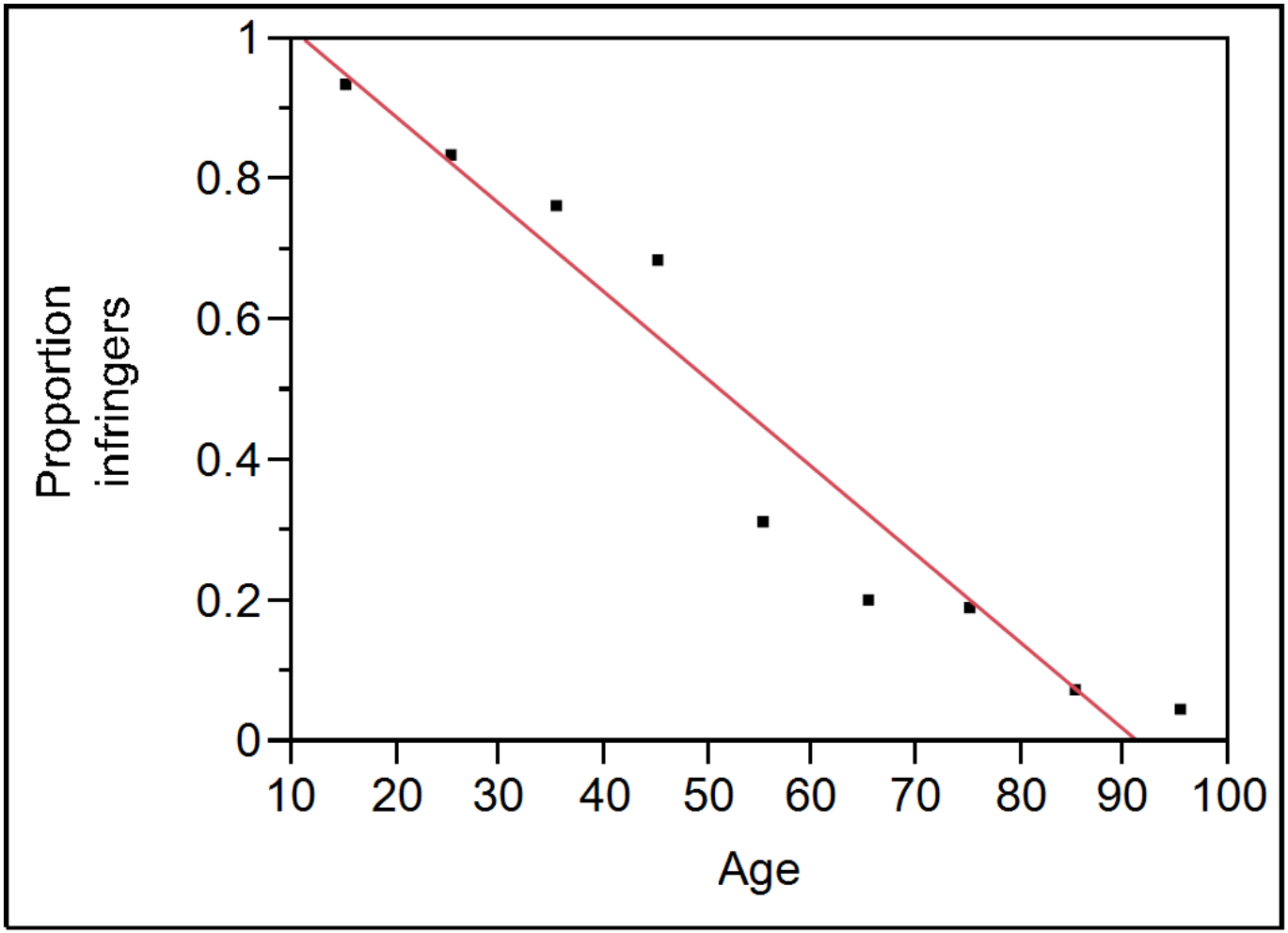
The basic idea

Response variable: Y = Infringe or don't infringe online copyrighted material.
X variable: Age, treated as continuous: [10, 20),[20, 30) ... and taken as the midpoint of the range. Based on a market research study we find

X	15	25	35	45	55	65	75	85	95
Infringe (Y = 1)	61	36	40	31	12	6	7	2	1
Not infringe (Y = 0)	4	7	12	14	26	23	29	24	20
n	65	43	52	45	38	29	36	26	21
Prop. infringe	0.94	0.84	0.77	0.69	0.32	0.21	0.19	0.08	0.05

We are interested in measuring the relationship between age and the probability of infringement

Naïve: put a least squares line through the proportions

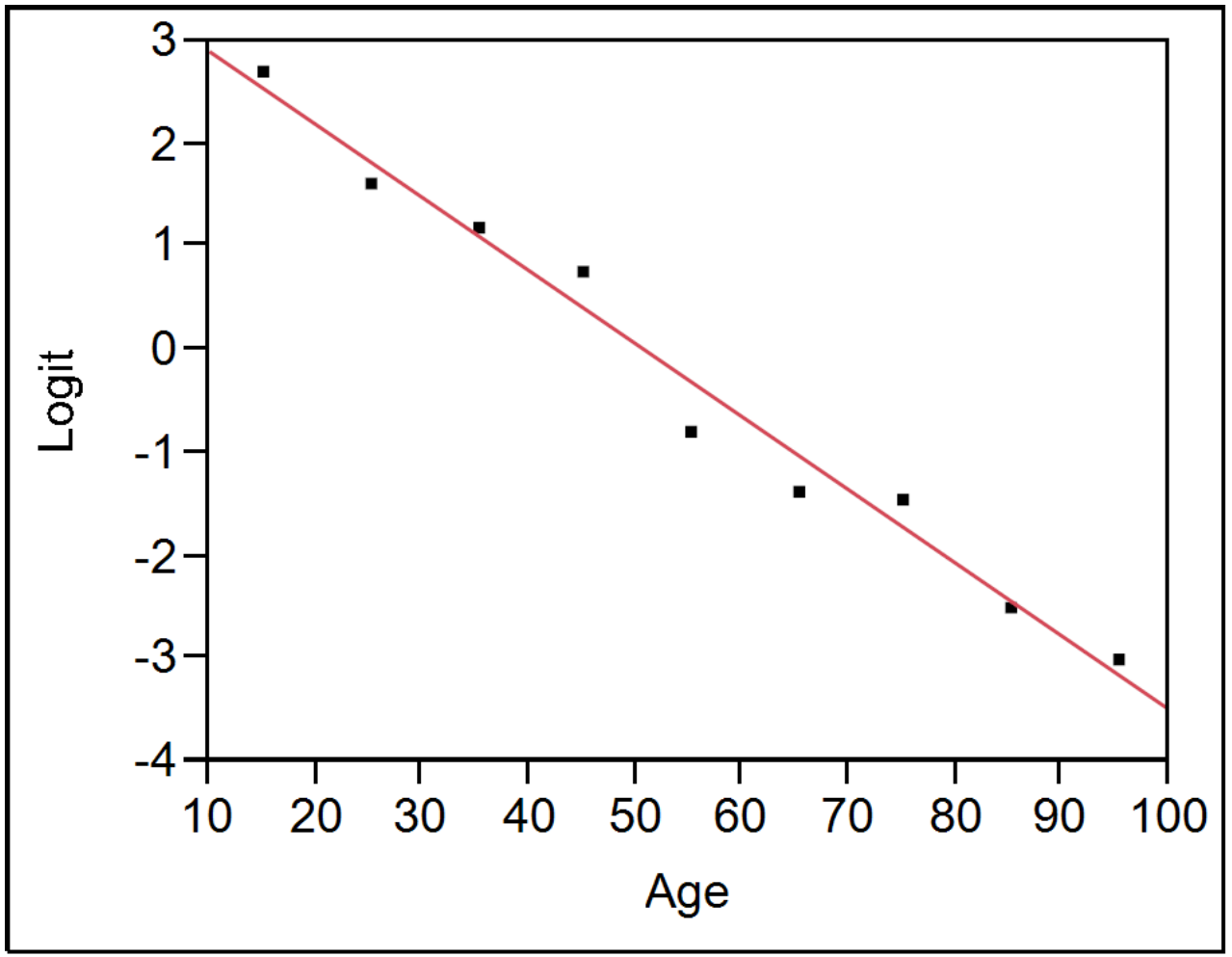


Criticism's

1. Extrapolation is crazy – the line goes outside the range [0,1] and it is meant to be estimating a proportion.
2. Doesn't take into account the frequency of responses in each age category.
3. Doesn't take into account the different variability in the proportions ($Var(\hat{p}) = \frac{p(1-p)}{n}$ so that the variance is a function of p. Different values of p have different variances i.e. non-constant variance).

A better approach

Fit the line to the logit of the proportions then the back transformed probabilities will always lie between 0 and 1 because $p = \frac{e^{logit(p)}}{1+e^{logit(p)}}$.

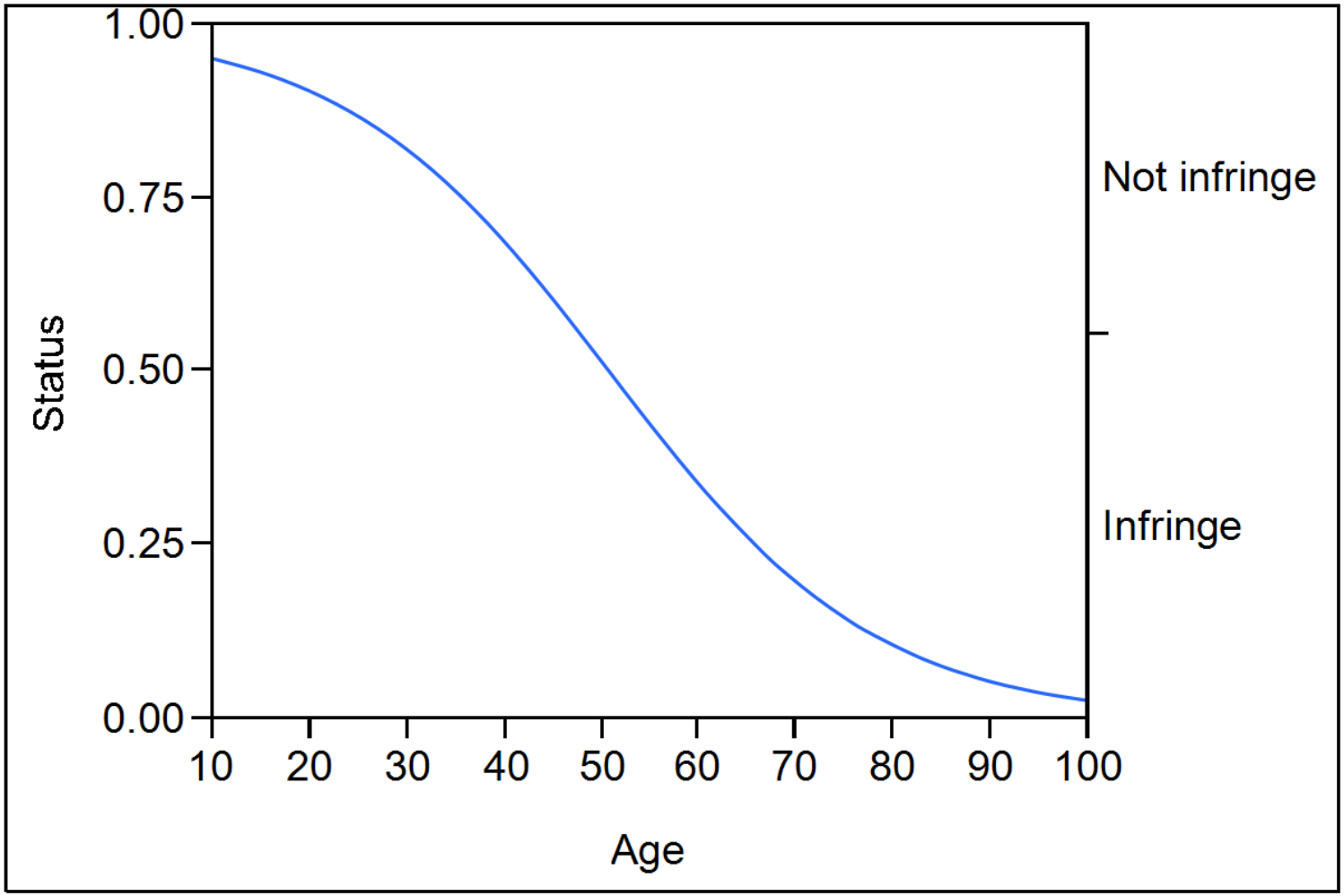


Criticisms:

- This is better but still doesn't deal with the non-constant variance issue.
- We are in real trouble if any of the proportions are 0 or 1. The logit will be $\pm\infty$ (recall that the logit equals $\log\{p/(1-p)\}$).

We need **logistic regression** to do the fitting properly.

The logistic regression fit



Above is the logistic fit, but what went on to get it?

The general approach to fitting probability models

1. Write down the probability function for a single observation.
2. Assuming independence then the probability of all the data is the **product** of these probabilities.
3. Think of the data as fixed and then find the parameter that maximizes the probability of the observed data.
4. This is the method of *maximum likelihood*.
5. We usually do the maximization on the log scale because the log of a product is the sum of the logs, and sums are easier to work with than products.
6. Furthermore, sums have Central Limit Theorems going for them which is useful in understanding the properties of the maximum likelihood estimates.

A likelihood example

If I toss a coin twice and observe a head and a tail what is the maximum likelihood estimate of the probability of a head? Call the true probability θ and this is what we want to estimate.
Define a head as 1 and tail as 0 and Y_i as the outcome of the ith coin toss. Then

$$P(y;\theta) = \begin{cases} \theta & : \quad y = 1 \\ (1-\theta) & : \quad y = 0. \end{cases}$$

Or more succinctly

$$p(y_i;\theta) = \theta^{y_i}(1-\theta)^{1-y_i}$$

which is called the probability density function (pdf).

With more than one observation

$$\begin{aligned} p(Y_1 = y_1 \& Y_2 = y_2; \theta) &= \theta^{y_1}(1-\theta)^{1-y_1} \times \theta^{y_2}(1-\theta)^{1-y_2} \\ &= \theta^{y_1+y_2}(1-\theta)^{2-y_1-y_2} \end{aligned}$$

Now take the log:

$$\log\{p(Y_1 = y_1 \& Y_2 = y_2; \theta)\} = (y_1 + y_2)\log(\theta) + (2 - y_1 - y_2)\log(1-\theta)$$

and the goal is to maximize this as a function of θ, which we do by differentiating and setting the derivative equal to zero.

The maximization

Write $\log\{p(\theta; Y_1 = y_1 \& Y_2 = y_2)\} = ll(\theta; \mathbf{y})$ for convenience (ll stands for **l**og-**l**ikelihood). So

$$ll(\theta; \mathbf{y}) = (y_1 + y_2)\log(\theta) + (2 - y_1 - y_2)\log(1 - \theta)$$

and

$$\frac{d}{d\theta} ll(\theta; \mathbf{y}) = \frac{(y_1 + y_2)}{\theta} - (2 - y_1 - y_2)\frac{1}{1 - \theta}.$$

On setting the derivative equal to 0 we get

$$0 = \frac{(y_1 + y_2)}{\theta} - (2 - y_1 - y_2)\frac{1}{1 - \theta}.$$

Multiply through by $\frac{1-\theta}{\theta}$ to get

$$(1 - \theta)(y_1 + y_2) + \theta(y_1 + y_2) = 2\theta,$$

and finally $\theta = \frac{y_1+y_2}{2}$, otherwise known as the mean! So if $Y_1 = 1$ and $Y_2 = 0$ then the probability of a head is estimated at 0.5.

And more realistically ...

- There are many more observations.
- The parameter θ is modeled as a function of the predictor variables, for example
$$logit(\theta) = \beta_0 + \beta_1 X.$$
- But the rest is the same; maximize the likelihood of the parameters given the observed data – which parameter value is most concordant with the data?
- The method of maximum likelihood is general, but when you apply it to Normally distributed data it is equivalent to the method of Least Squares.
- The algorithm used to do maximum likelihood in logistic regression is called **I**teratvely **R**eweighted **L**east **S**quares (IRLS).

Properties of the maximum likelihood estimate $\hat{\theta}$

With enough data and under a few assumptions the key fact is, writing $\hat{\theta}$ for the *m.l.e.*

$$\hat{\theta} \sim Normal(\theta_0, se(\hat{\theta})),$$

where θ_0 is the true value of the parameter.
That is, the *m.l.e.* has a normal distribution so all of the ideas of confidence intervals and t-stats from Stat101/102 are still applicable.

Hypothesis testing

Say we want to test whether a bigger expanded model (M_1) provides a significantly better fit to the data than a smaller model (M_0). Then if we use the method of maximum likelihood, it is the case that

$$-2\{ll(M_0) - ll(M_1)\} \sim \chi^2_k,$$

where χ^2_k denotes the Chi-squared distribution on k degrees of freedom and k is the number of additional parameters in M_1 that are not in M_0.
Recall that *ll* stands for *log likelihood*.
This means that we can do the usual hypothesis testing using p-values, but now on the χ^2-statistic rather than t-stats or F-stats.

Prediction

Once we have fit a logistic model we can back transform from the fitted logits to the estimated probabilities.
If

$$logit\{p(x)\} = \beta_0 + \beta_1 X,$$

then

$$p(x) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}.$$

Interpretation of the coefficients in a logistic regression

Consider X increasing by one unit. Then

$$logit\{p(x+1)\} - logit\{p(x)\} = \beta_1.$$

Exponentiating both sides gives

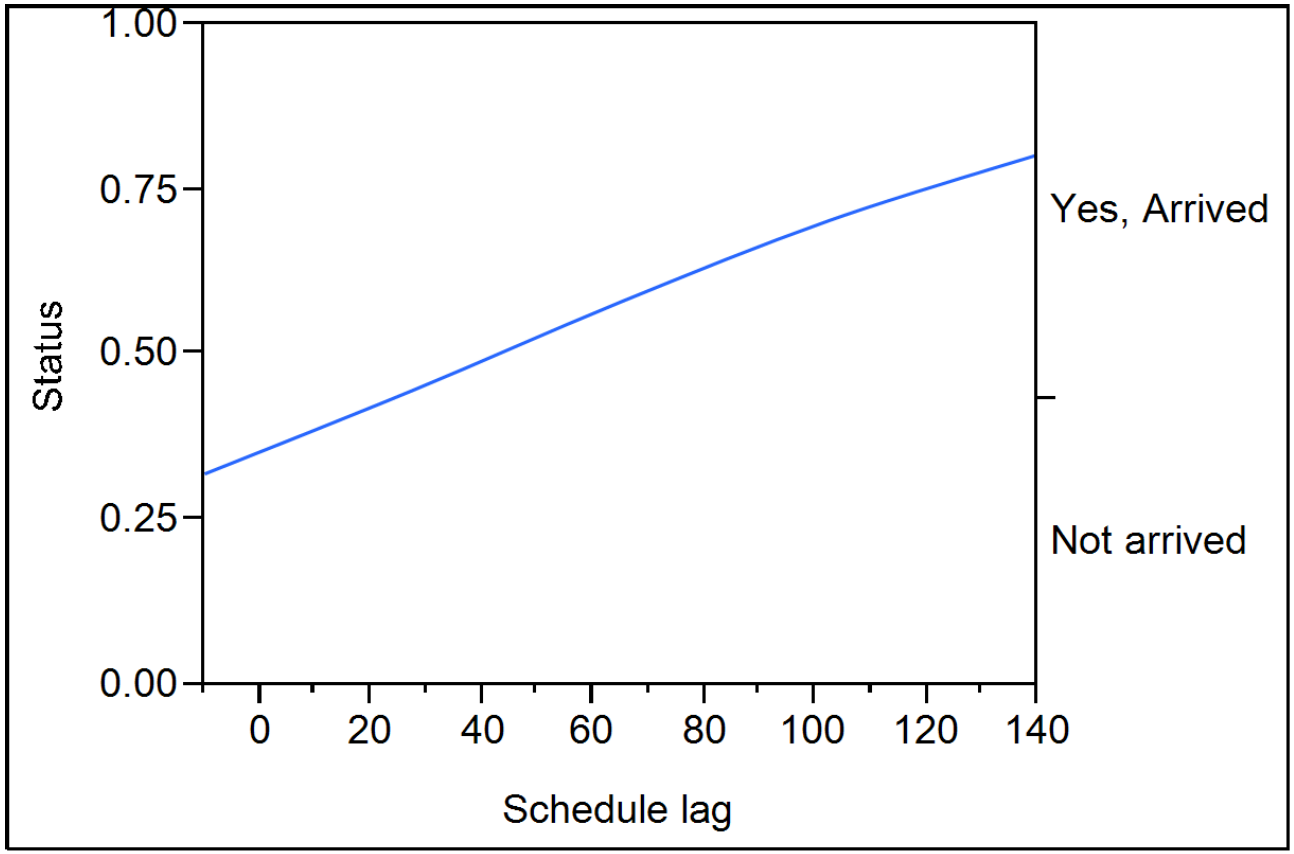
$$\frac{p(x+1)}{1-p(x+1)} \Big/ \frac{p(x)}{1-p(x)} = e^{\beta_1}.$$

This quantity is often termed the *odds* ratio.
In (approximate) English: if X goes up by one unit, then the odds of being a "Yes" goes up by a multiplicative factor of e^{β_1}.
And β_1 is termed the *log odds ratio*.

The outpatient visit data

- Y. Status. Did a patient show up. YES/NO.
- Various X's. We will focus on how far in advance the visit was scheduled. `Schedule lag`.

The logistic fit



Parsing the output

The whole model test. The equivalent of the "overall F-statistic" in regular regression:

Whole Model Test

Model	-LogLikelihood		DF	ChiSquare	Prob>ChiSq
Difference	7.34366	C	1	14.68731	0.0001*
Full	332.88779	A	D	E	F
Reduced	340.23144	B			

The whole model test

A Minus the log likelihood under the *BIG/FULL* model.

B Minus the log likelihood under the *small/reduced* model – here just the intercept.

C Minus the difference in log-likelihoods. Twice this number has the χ^2 distribution and is sometimes called the *deviance*. You may see it written as G^2.

D The degrees of freedom (df), the number of parameters between the BIG and small models.

E The value of the χ^2 statistic.

F The p-value of the χ^2 statistic.

Nested models

- Some terminology: A model with just the intercept is sometimes called the *Null* model. And a model with as many parameters as observations is called the *saturated* model. These are the two extremes.
- In this example the two models, *full* and *reduced*, are *nested*, that is, one is a special case of the other.
- If models are nested you can compare them with G^2.
- If models are **not** nested you cannot use the difference in log-likelihoods for hypothesis testing. But you can still compare the AIC values.

The individual parameter estimates

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-0.6122051	0.1309326	21.86	<.0001*
Schedule lag	0.01439497	0.003892	13.68	0.0002*
	A	B	C	D

For log odds of Not arrived/Yes, Arrived

A The estimates of β_0 and β_1.

B The estimated standard errors of β_0 and β_1.

C The ratio of the estimate to the standard error, squared. Approximately $\chi^2_{(1)}$ distributed.

D The p-value of the $\chi^2_{(1)}$ statistic.

The ratio of the estimate to the standard error is often called the *Wald* statistic and for a sufficient sample size has an approximate Normal distribution.

Summarizing the findings

Does it look like the model is useful in explaining Status?

Yes, as the overall model test is significant. And because we only have one variable you can also look at the single slope estimate p-value.

What is the sign of the association?

From looking at the sign of the slope, it is positive. As schedule lag goes up the probability of a no-show goes up.

What is the interpretation of the slope here?

Every one day increase in schedule lag is associated with an increase in the odds of a no-show by a multiplicative factor of $e^{0.0144} = 1.0145$

Is the association significant?

Yes, the small p-value on the slope estimate indicates that this is the case.

Predict the probability of a no show at 30 days lag.

$e^{-0.6122+30\times(0.0144)}/(1+e^{-0.6122+30\times(0.0144)}) = 0.455.$

False positives and false negatives

- When the outcome variable is a 1 define it as a *positive*.
- If we predict the outcome to be a 1, but it is in fact a 0, then we call it as a false positive (aka Type I error).
- If we predict the outcome to be a 0, but it is in fact a 1, then we call it as a false negative (aka Type II error).
- *Specificity* = 1 - false positive rate.
- *Sensitivity* = 1 - false negative rate.

Definitions ctd.

For any set of observations (in-sample or out-of-sample) we can compare the predicted values and the actual values in a table:

	Predicted negative	Predicted positive	Total
Actual negative	a	b	a + b
Actual positive	c	d	c + d
Total	a + c	b + d	a + b + c + d = n

- False positive rate = b/(a + b). False negative rate: c/(c + d).
- Overall misclassification rate = (b + c)/(a + b + c + d).

Example for the outpatient visit data

- Predict an observation to be a no-show if its predicted probability is greater than or equal to 0.5

	Predicted No-show	Predicted Arrived	Total
Actual No-show	46	170	216
Actual Arrived	35	246	281
Total	81	416	497

- Call "Arrived" the positive level.
- False positive rate: $170/216 = 0.787$. False negative rate: $35/281 = 0.125$.
- Misclassification rate: $(170 + 35)/497 = 0.4125$.

Consider probability cutoffs other than 0.5

- It costs \$2.00 to mail a high-end promotional offer.
- If a person responds, then there will be a profit of \$50 to the company.
- What is the break-even response probability, p^*, that determines if someone should be mailed to?
- Call p the probability that someone responds, then

$$E(Profit) = p \times \$50 - \$2$$

  and $p^* = 2/50 = 0.04$.
- So it makes sense to mail to anyone who has a probability of response greater than 0.04.

The ROC curve

- A graphical summary of the quality of prediction.
- ROC curves can be used to compare the predictive quality of models.
- A perfect classifier would have Sensitivity and Specificity of 1. An ROC curve plots Sensitivity against (1 - Specificity) allowing for different probability cut-offs to define the predicted positive and negatives.
- For the outpatient visit data, if we predicted a no-show whenever the predicted probability was greater than 0.4, the classification table would be:

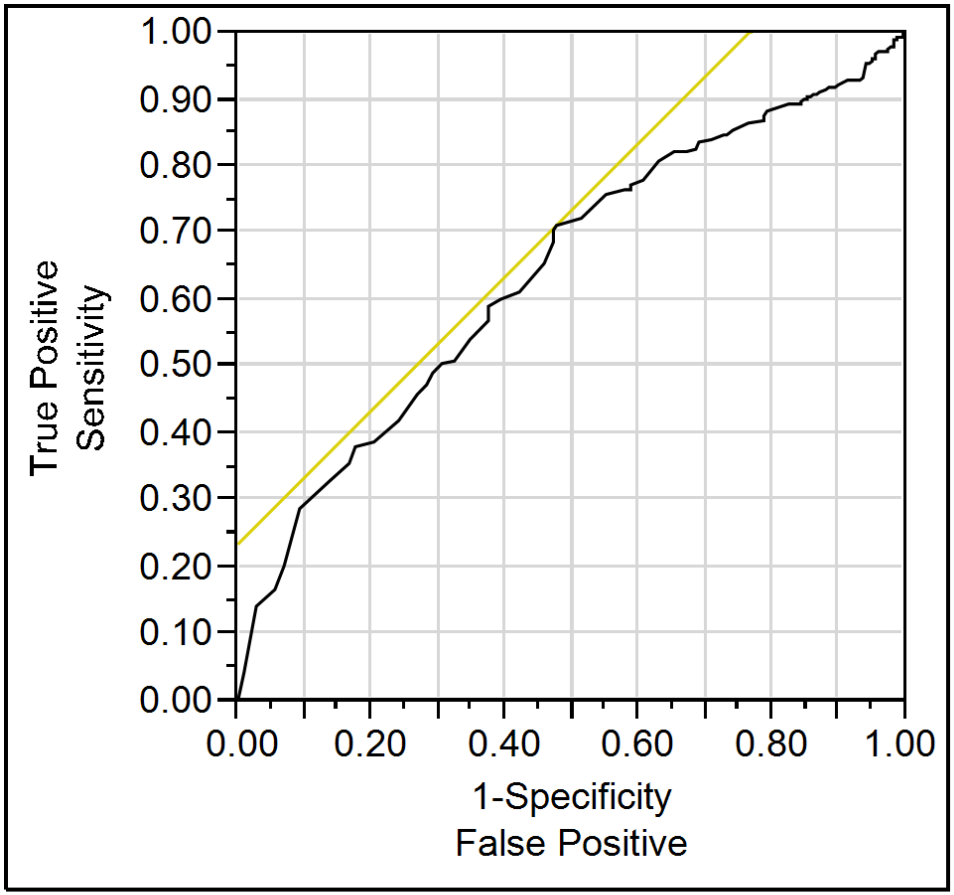
Example for the outpatient visit data

- Example: predict an observation to be a no-show if its predicted probability is greater than or equal to 0.4.

	Predicted No show	Predicted Arrived	Total
Actual No show	146	70	216
Actual Arrived	138	143	281
Total	284	213	497

- False positive rate: $70/216 = 0.324$. False negative rate: $138/281 = 0.491$.
- Misclassification rate: $208/497 = 0.419$.
- Repeat this calculation for the complete range of probabilities and plot.

What the plot looks like



Using Status='Yes, Arrived' to be the positive level

AUC
0.64299

The extremes

- With no predictive power: the plot follows the 45° line.
- With perfect predictive power: the plot shoots straight up to the point (0,1) then moves horizontally to (1,1).
- That is, we like steep ROC curves.

AUC = **A**rea **U**nder **C**urve.
Among competing classifiers, prefer the classifier with the greatest area under the curve.

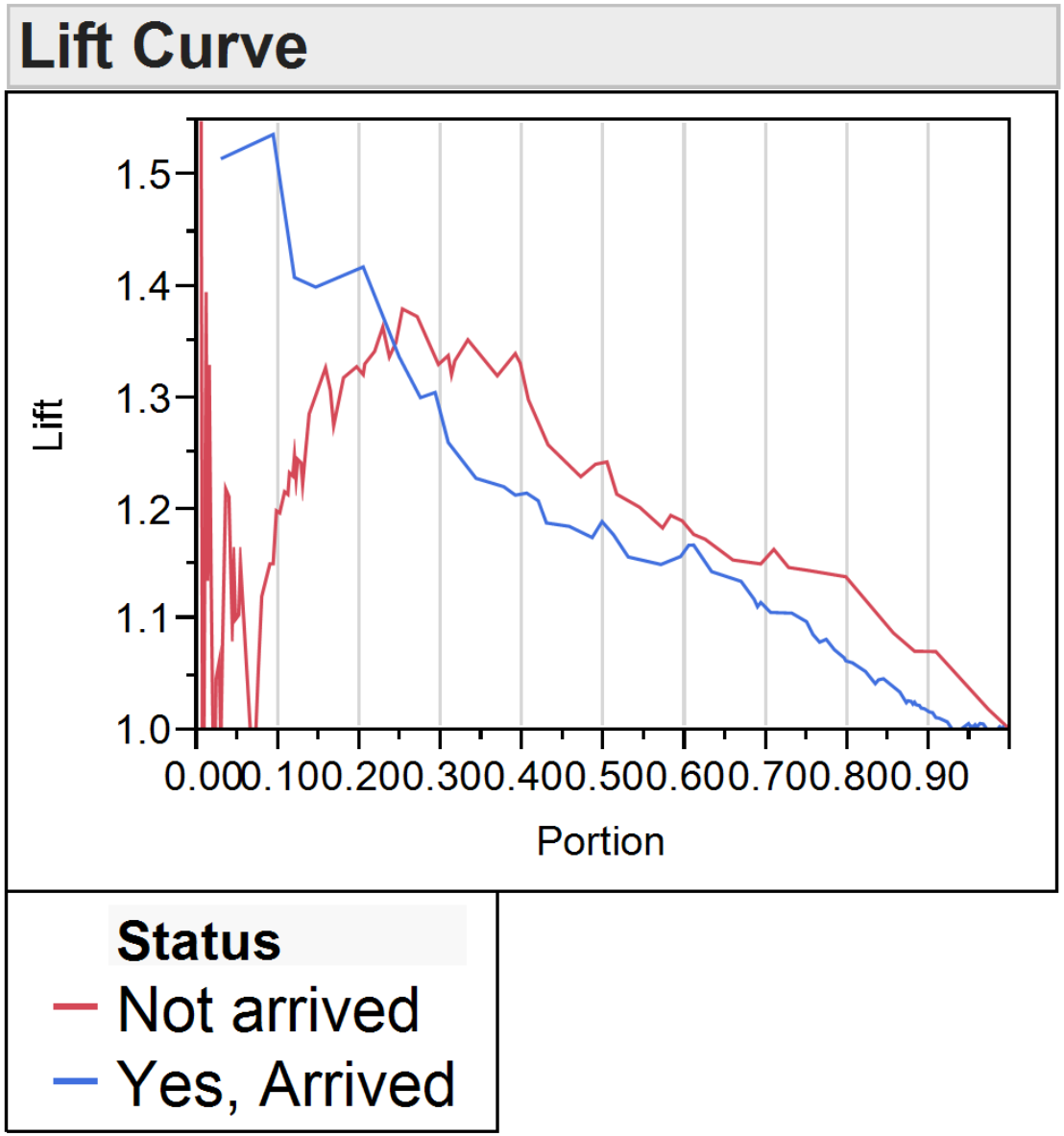
The Lift Chart

- *Lift* is a measure of the true target richness at the top of the target list.
- Say there are 1,000 people in the population of which 100 are positive for a population prevalence of 10%.
- If the top 5th percentile in the target list had 20 positives, that is 20 positives out of 50 or 40% then the **Lift** would be defined as

$$\text{Lift}(5\%) = \frac{\%\text{ in target list}}{\%\text{ in population}} = \frac{40\%}{10\%} = 4.$$

- High Lift is good.
- The Lift Chart is made from considering all the different percentiles (portions), not just the 5th.

Lift curve


Lift Curve
Lift
1.5
1.4
1.3
1.2
1.1
1.0
0.00 0.10 0.20 0.30 0.40 0.50 0.60 0.70 0.80 0.90
Portion
Status
Not arrived
Yes, Arrived

Summary

- Train/test paradigm
- Introduction to logistic regression
- Hypothesis testing in logistic regression
- False positives/false negatives
- The ROC curve
- The lift chart