Business Analytics
Logistic and Poisson Regressions

Prof. Dr. Martin Bichler

Decision Sciences & Systems

Department of Informatics

Technische Universität München

Course Content

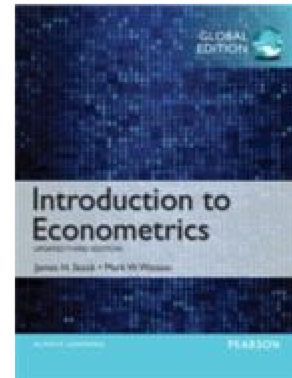
- Introduction
- Regression Analysis
- Regression Diagnostics
- **Logistic and Poisson Regression**
- Naive Bayes and Bayesian Networks
- Decision Tree Classifiers
- Data Preparation and Causal Inference
- Model Selection and Learning Theory
- Ensemble Methods and Clustering
- High-Dimensional Problems
- Association Rules and Recommenders
- Neural Networks

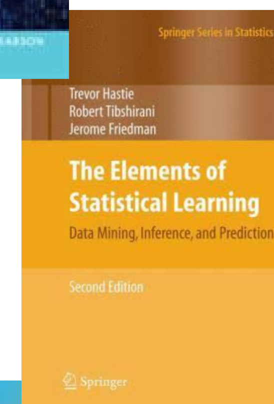
Recommended Literature

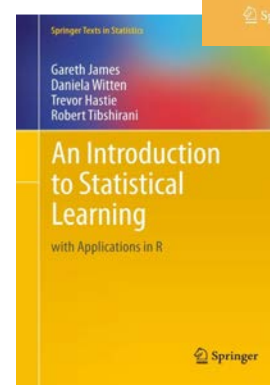
- **Introduction to Econometrics**
  - Stock, James H., and Mark W. Watson
  - Chapter 11

- **The Elements of Statistical Learning**
  - Trevor Hastie, Robert Tibshirani, Jerome Friedman
  - http://web.stanford.edu/~hastie/Papers/ESLII.pdf
  - Section 4.4: Logistic Regression

- **An Introduction to Statistical Learning: With Applications in R**
  - Gareth James, Trevor Hastie, Robert Tibshirani
  - Section 4.1-4.3: Logistic Regression







Logistic Regression

- Models of discrete choice have been a topic in (Micro-) Econometrics and are nowadays widely used in Marketing
- Logit, and probit models extend the principles of general linear models (regression) to better treat the case of dichotomous and categorical target variables
- They focus on categorical dependent variables, looking at all levels of possible interaction effects
  - Predicting a categorical dependent variable is also known as classification
- McFadden got the 2000 Nobel prize in Economics for fundamental contributions in discrete choice modeling

Application

- Why do commuters choose to fly or not to fly to a destination when there are alternatives
- Available modes = Air, Train, Bus, Car
- Observed:
  - Choice
  - Attributes: Cost, terminal time, other
  - Characteristics of commuters: Household income
- Choose to fly iff $U_{fly} \geq 0$
  - $U_{fly} = \beta_0 + \beta_1\, Cost + \beta_2\, Time + \beta_3\, Income + \varepsilon$

Data for the Estimation

Choose Air	Gen.Cost	Term Time	Income
1.0000	86.000	25.000	70.000
.00000	67.000	69.000	60.000
.00000	77.000	64.000	20.000
.00000	69.000	69.000	15.000
.00000	77.000	64.000	30.000
.00000	71.000	64.000	26.000
.00000	58.000	64.000	35.000
.00000	71.000	69.000	12.000
.00000	100.00	64.000	70.000
1.0000	158.00	30.000	50.000
1.0000	136.00	45.000	40.000
1.0000	103.00	30.000	70.000
.00000	77.000	69.000	10.000
1.0000	197.00	45.000	26.000
.00000	129.00	64.000	50.000
.00000	123.00	64.000	70.000

The Linear Probability Model

- In the OLS regression:
  $Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon$; where $Y = \{0,1\}$
- The predicted probabilities of the linear model can be greater than 1 or less than 0
- ε is not normally distributed because Y takes on only two values
- The error terms are heteroscedastic

The Logistic Regression Model

- The "logit" model solves the problems of the linear model:

$$\ln\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$

- $p(X)$ is the probability that the event Y occurs given X, $\Pr[Y = 1|X]$
- $\frac{p(X)}{1-p(X)}$ describes the "odds"
  - The 20% probability of winning describes odds of .20/.80 = .25
  - A 50% probability of winning leads to odds of 1
- $\ln\left(\frac{p(X)}{1-p(X)}\right)$ is the *log odds*, or "logit"
  - p = 0.50, then logit = 0
  - p = 0.70, then logit = 0.84
  - p = 0.30, then logit = -0.84

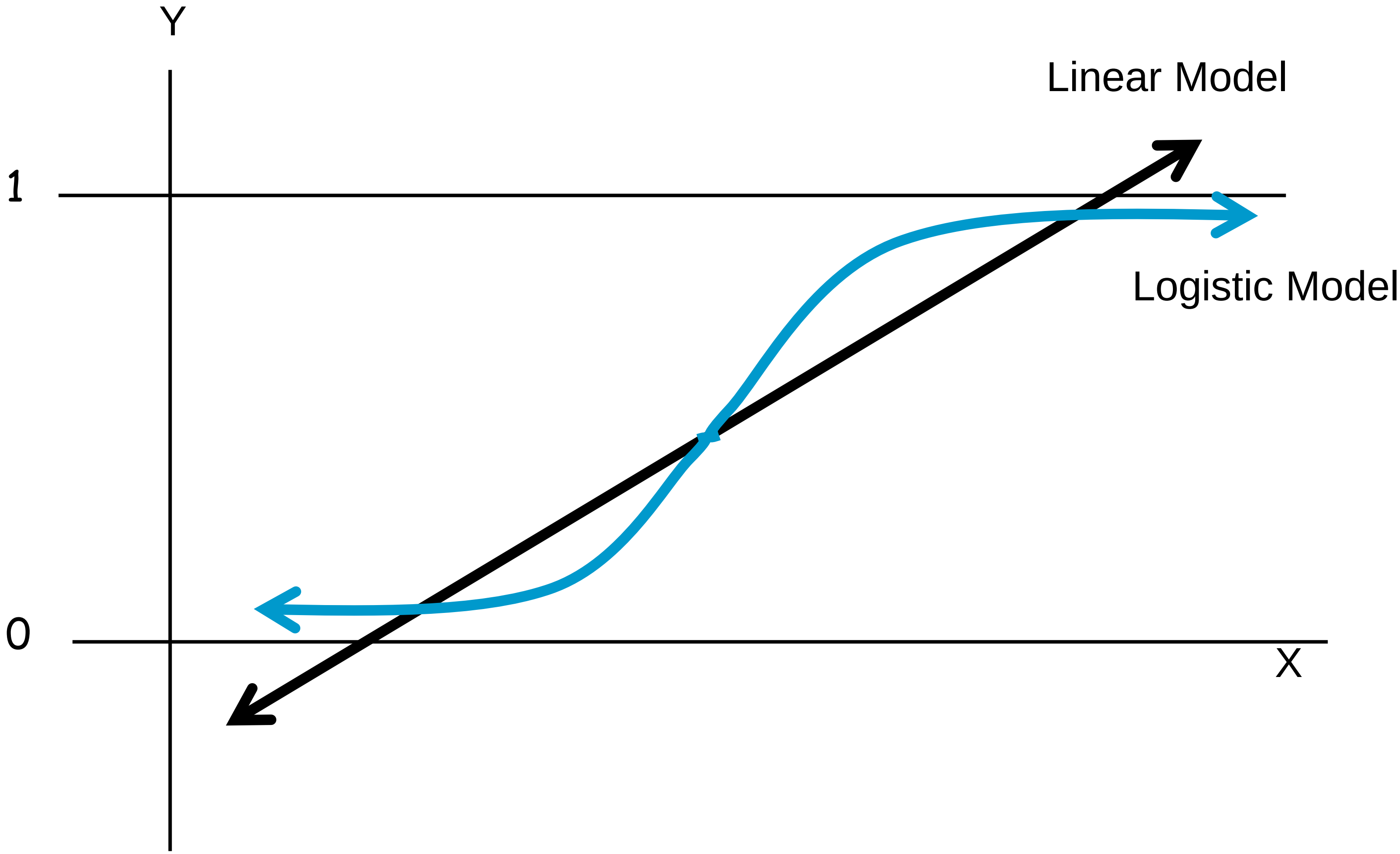
Logistic Function

- The logistic function $\Pr[Y|X]$ constraints the estimated probabilities to lie between 0 and 1 ($0 \leq \Pr[Y|X] \leq 1$)

$$\Pr[Y|X] = p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

- $\Pr[Y|X]$ is the estimated probability that the i^{th} case is in a category and $\beta_0 + \beta_1 X$ is the regular linear regression equation
- This means that the probability of a success ($Y = 1$) given the predictor variable (X) is a non-linear function, specifically a logistic function
  - if you let $\beta_0 + \beta_1 X = 0$, then $p(X)$ = .50
  - as $\beta_0 + \beta_1 X$ gets really big, $p(X)$ approaches 1
  - as $\beta_0 + \beta_1 X$ gets really small, $p(X)$ approaches 0

Linear and Logistic Models



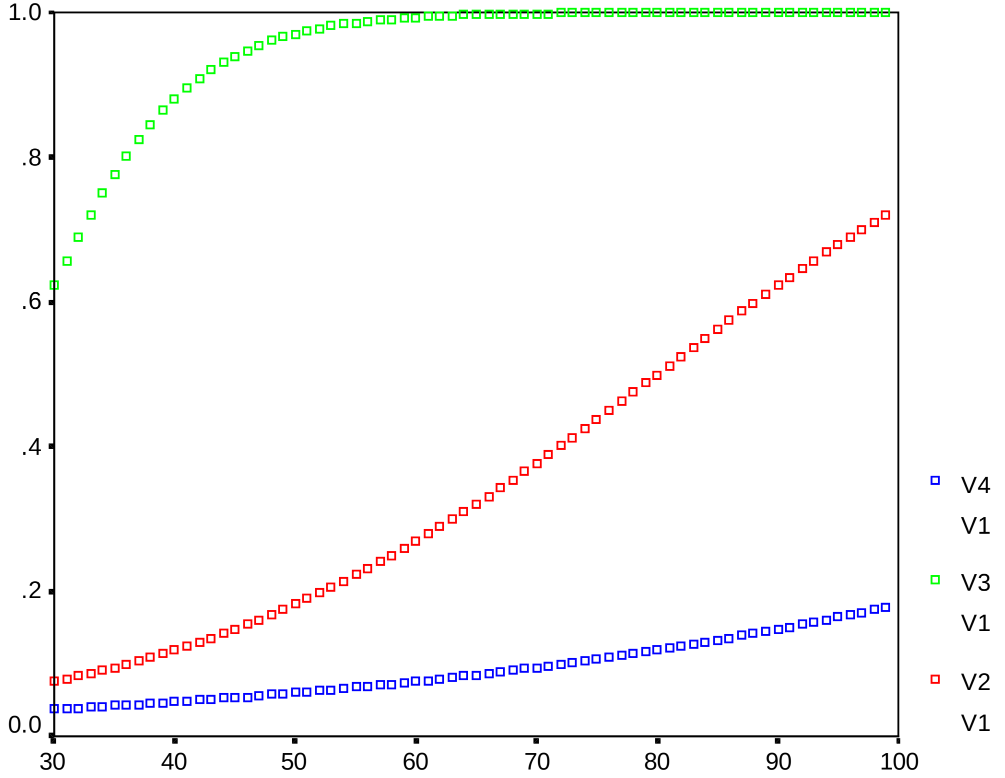
The Logistic Function

- The values in the regression equation β_1 and β_0 take on slightly different meanings
  - β_0, The regression constant (moves curve left and right)
  - β_1, The regression slope (steepness of curve)
  - $-\beta_0/\beta_1$, The threshold, where probability of success = .50

Logistic Function

Fixed regression constant, different slopes

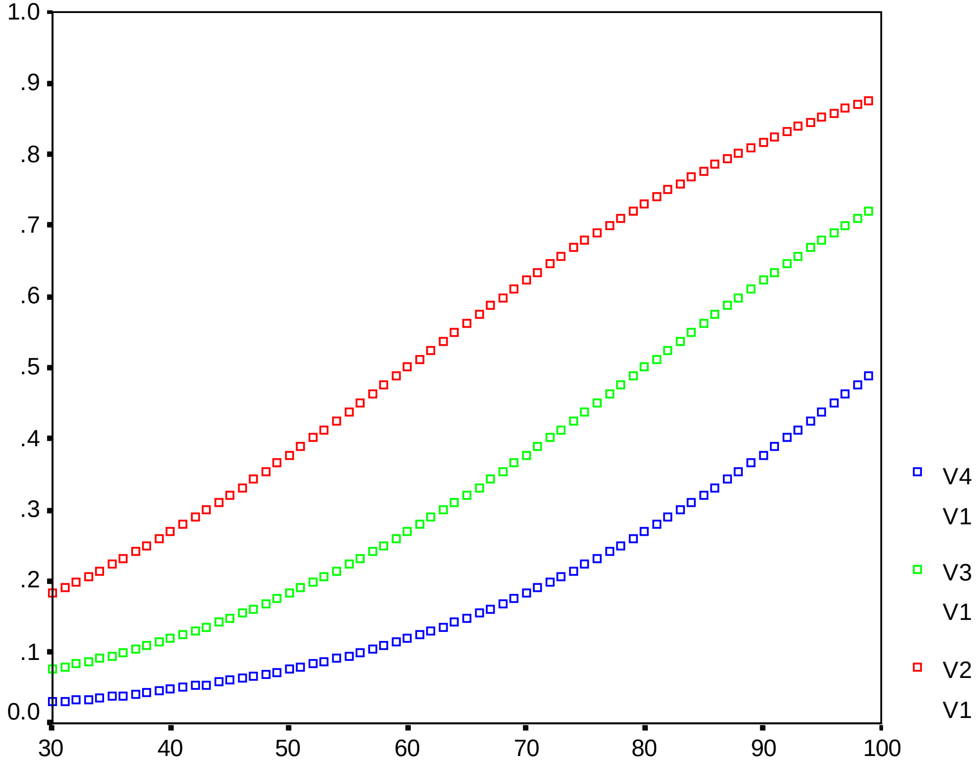
- v3: β_0 = -4.00
  β_1 = 0.15 (top)
- v2: β_0 = -4.00
  β_1 = 0.05 (middle)
- v4: β_0 = -4.00
  β_1 = 0.025 (bottom)



Logistic Function

Constant slopes with different regression constants

- v2: β_0 = -3.00
  β_1 = 0.05 (top)
- v3: β_0 = -4.00
  β_1 = 0.05 (middle)
- v4: β_0 = -5.00
  β_1 = 0.05 (bottom)



Odds and Logit

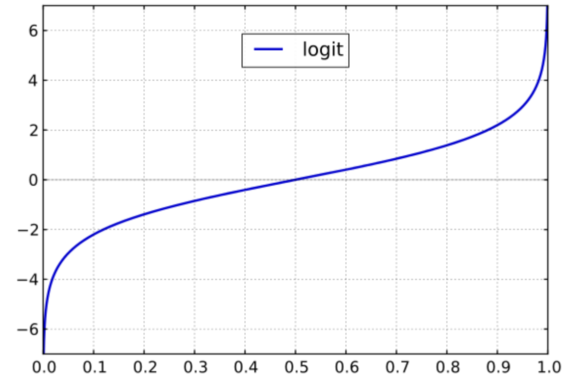
- By algebraic manipulation, the logistic regression equation can be written in terms of an odds of success:

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

- Odds range from 0 to positive infinity
- If $\frac{p(X)}{1-p(X)}$ is
  - less than 1, then less than .50 probability
  - greater than 1, then greater than .50 probability

$p(X)$	$\frac{p(X)}{1-p(X)}$
0,1	0,1111
0,2	0,2500
0,3	0,4286
0,4	0,6667
0,5	1,0000
0,6	1,5000
0,7	2,3333
0,8	4,0000
0,9	9,0000
1	INF

The Logit



- Finally, taking the natural log of both sides,
  we can write the equation in terms of logits (log-odds):

$$\ln\left(\frac{\Pr[Y=1|X]}{1-\Pr[Y=1|X]}\right)=\ln\left(\frac{p(X)}{\left(1-p(X)\right)}\right)=\beta_0+\beta_1 X$$

- Probability is constrained between 0 and 1
- Log-odds are a linear function of the predictors
- Logit is now between - ∞ and + ∞ (as the dependent variable of a linear regression)
- The regression coefficients go back to their old interpretation
- The amount the logit (log-odds) changes, with a one unit change in X

$p(X)$	$\frac{p(X)}{(1-p(X))}$	Logit
0	0	-∞
0,1	0,11	-2,20
0,2	0,25	-1,39
0,3	0,43	-0,85
0,4	0,67	-0,41
0,5	1,00	0,00
0,6	1,50	0,41
0,7	2,33	0,85
0,8	4,00	1,39
0,9	9,00	2,20
1	∞	∞

Can you try to derive the logistic function for probability from the logit model?

Estimating Coefficients of a Logistic Regression

- Maximum Likelihood Estimation (MLE) is a statistical method for estimating the coefficients β of a model, so that the observed data is most probable.
- The probability of one data point y_i can be modeled as a Bernoulli trial
  - $p^{y_i}(1-p)^{1-y_i}$
- The likelihood function (L) equals the joint probability of observing the particular set of dependent variable values that occur in the random (i.i.d.) sample
  - $L = \prod_{i=1} p^{y_i}(1-p)^{1-y_i}$

Likelihood Function for the Logit Model

The logistic function is an example of a sigmoid function often used in feed-forward neural networks as activation function.

$$\Pr[Y_i = 1|X] = p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}} = \sigma(\beta_0 \ + \beta_1 X_{1i})$$

- $\Pr(Y_i \ = \ 1) \rightarrow$ 0 as $\beta_0 \ + \beta_1 X_{1i} \rightarrow -\infty$
- $\Pr(Y_i \ = \ 1) \rightarrow$ 1 as $\beta_0 \ + \beta_1 X_{1i} \rightarrow \ \infty$

Likelihood function models a sequence of Bernoulli trials

$$L = \prod_{i=1} p^{y_i}(1-p)^{1-y_i} = \prod_{i=1} \sigma(\beta_0 \ + \beta_1 X_{1i})^{y_i} * [1 - \ \sigma(\beta_0 \ + \beta_1 X_{1i})]^{1-y_i}$$

The Likelihood Function for the Logit Model

Use $L = \prod_{i=1} p^{y_i}(1-p)^{1-y_i}$ and take the log

$$LL = \ln(L) = \sum_{i=1} y_i \ln p_i + (1-y_i)\ln(1-p_i)$$

We look for the vector β that maximizes LL

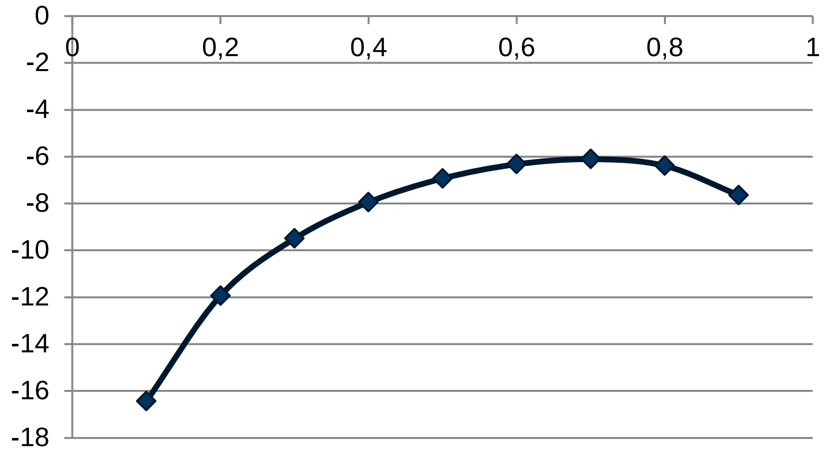
$$\beta = \text{argmax}_\beta LL(\beta) = \text{argmax}_\beta \left[\sum_{i=1} y_i \ln \sigma(\beta_0 + \beta_1 X_{1i}) + (1-y_i)\ln(1-\sigma(\beta_0 + \beta_1 X_{1i})) \right]$$

The LL function is twice differentiable and concave!

- For the OLS estimator, we set the FOC=0.
- For the logistic regression, this does not get us a closed-form solution.
- We use a numerical algorithm to find the maximum: gradient ascent!

Illustrative Example of MLE

- Suppose 10 individuals make travel choices between auto and public transit.
- All travelers are assumed to possess identical attributes (unrealistic), and so the probabilities are not functions of β_i but simply a function of p, the probability p of choosing auto.
  - $L = p^x(1-p)^{n-x} = p^7(1-p)^3$
  - $LL = \ln(L) = 7\ln(p) + 3\ln(1-p)$, maximized at 0.7



Reminder: Gradient Ascent

One numerical technique to find the maximium of the LL function is gradient ascent. Gradient ascent is typically part of calculus classes. We'll revisit the essentials here, but will spend more time in the context of neural networks.

Input: Concave, continuously differentiable function $f\colon \mathbb{R}^n \to \mathbb{R}$,
feasible start point $x^{(1)} \in \mathbb{R}^n$, and parameter $\varepsilon > 0$

$k = 1$

While($\left\|\nabla f(x^{(k)})\right\| \geq \varepsilon$){

- Choose step size $\alpha > 0$
  e.g. compute $\alpha^* > 0$, to maximize $f\left(x^{(k)} + \alpha \nabla f\left(x^{(k)}\right)\right)$ //"line search"
- Set $x^{(k+1)} = x^{(k)} + \alpha^* \nabla f\left(x^{(k)}\right)$
- $k++$

}

Gradient Ascent Example

$k = 1$
While($\|\nabla f(x^{(k)})\| \geq \varepsilon$){
- Determine $\alpha^* > 0$, for maximizes $f\left(x^{(k)} + \eta \nabla f\left(x^{(k)}\right)\right)$.
- Set $x^{(k+1)} = x^{(k)} + \alpha^* \nabla f\left(x^{(k)}\right)$
- $k++$

}

$$f(x_1, x_2) = 2(x_1 + x_2) - x_1^2 - 2x_2^2$$

$$\nabla f(x) = \begin{pmatrix} 2(1 - x_1) \\ 2(1 - 2x_2) \end{pmatrix}$$

Start at $x^{(1)} = (0,0), f\left(x^{(1)}\right) = 0, \nabla f\left(x^{(1)}\right) = \binom{2}{2}$

Maximize $f(0 + 2\alpha, 0 + 2\alpha) = 8 - 12\alpha^2$

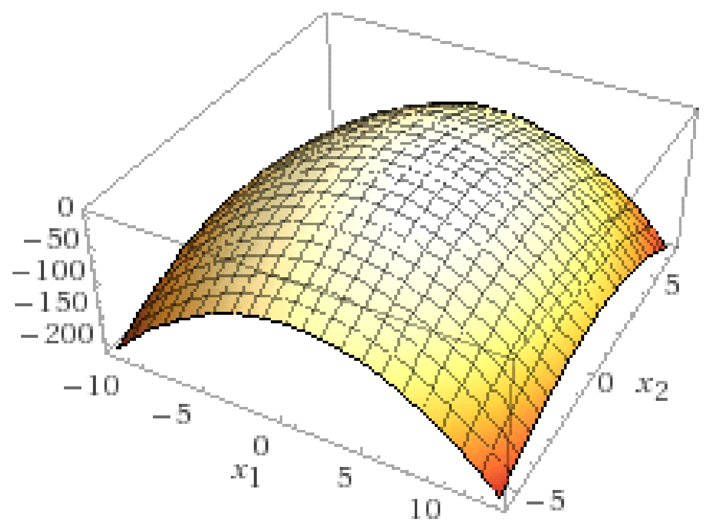
→ Set the derivative to null: $8 - 24\alpha = 0$

→ Maximized at $\alpha = \frac{1}{3}$

→ $x^{(2)} = \left(\frac{2}{3}, \frac{2}{3}\right), f\left(x^{(2)}\right) = \frac{4}{3}, \nabla f\left(x^{(2)}\right) = \begin{pmatrix} \frac{2}{3} \\ -\frac{2}{3} \end{pmatrix}$

this is the direction of steepest ascent

...



Gradient for the LL Function

$$\beta = \text{argmax}_{\beta} LL(\beta) = \text{argmax}_{\beta} \left[\sum_{i=1} y_i \ln \sigma(\mathbf{X}\beta) + (1 - y_i) \ln(1 - \sigma(\mathbf{X}\beta)) \right]$$

We can use the chain rule:

$$LL(\beta) = y \ln p + (1 - y) \ln(1 - p)$$

$$\frac{\partial LL(\beta)}{\partial p} = \frac{y}{p} - \frac{1 - y}{1 - p}$$

$$p = \sigma(z), z = \mathbf{X}\beta$$

$$\frac{\partial p}{\partial z} = \sigma(z)[1 - \sigma(z)]$$

$$z = \mathbf{X}\beta, \frac{\partial z}{\partial \beta_j} = x_j$$

$$\frac{\partial LL(\beta)}{\partial \beta_j} = \frac{\partial LL(\beta)}{\partial p} * \frac{\partial p}{\partial z} * \frac{\partial z}{\partial \beta_j} =$$

$$\left[\frac{y}{p} - \frac{1 - y}{1 - p}\right] \sigma(z)[1 - \sigma(z)] x_j =$$

since $p = \sigma(z)$

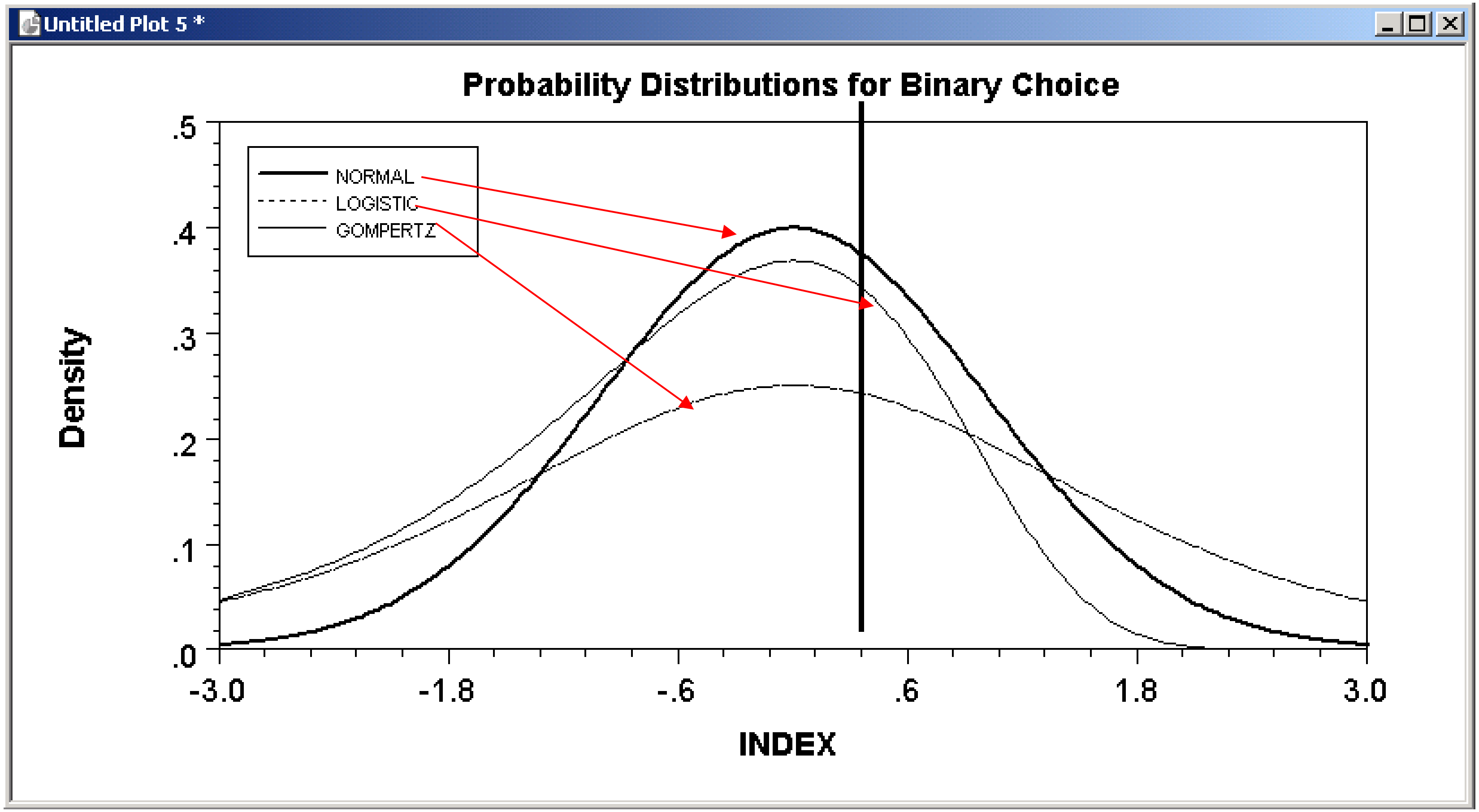
$$\left[\frac{y}{p} - \frac{1 - y}{1 - p}\right] p[1 - p] x_j =$$

$$[y(1 - p) - p(1 - y)] x_j =$$

$$[y - p] x_j =$$

$$[y - \sigma(\mathbf{X}\beta)] x_j$$ => Gradient

The Distributions



Logit models use the logistic distribution, Probit models use the Normal distribution;

Data for the Estimation

Y_i=Auto	X_{1i}=Cost	X_{2i}=Time*	X_{3i}=Income*
1.0000	86.000	25.000	70.000
.00000	67.000	69.000	60.000
.00000	77.000	64.000	20.000
.00000	69.000	69.000	15.000
.00000	77.000	64.000	30.000
.00000	71.000	64.000	26.000
.00000	58.000	64.000	35.000
.00000	71.000	69.000	12.000
.00000	100.00	64.000	70.000
1.0000	158.00	30.000	50.000
1.0000	136.00	45.000	40.000
1.0000	103.00	30.000	70.000
.00000	77.000	69.000	10.000
1.0000	197.00	45.000	26.000
.00000	129.00	64.000	50.000
.00000	123.00	64.000	70.000

Estimated Binary Choice Models

	LOGIT		PROBIT		EXTREM VALUE (GOMPERTZ)	
Variable	Estimate	t-ratio	Estimate	t-ratio	Estimate	t-ratio
Constant	1.78458	1.40591	0.438772	0.702406	1.45189	1.34775
GC	0.0214688	3.15342	0.012563	3.41314	0.0177719	3.14153
TTME	-0.098467	-5.9612	-0.0477826	-6.65089	-0.0868632	-5.91658
HINC	0.0223234	2.16781	0.0144224	2.51264	0.0176815	2.02876
Log-L	-80.9658		-84.0917		-76.5422	
Log-L(0)	-123.757		-123.757		-123.757	

Goodness of Fit

- The **null model**
  - assumes one parameter (the intercept) for all of the data points, which means you only estimate 1 parameter.
- The **fitted model**
  - assumes you can explain your data points with p parameters and an intercept term, so you have $\mathrm{p}+1$ parameters.

- **Null deviance:** $-2\ln\left(\mathrm{L}(\mathrm{null})\right)$
  - How much is explained by a model with only the intercept.
- **Residual deviance:** $-2\ln\left(\mathrm{L}(\mathrm{fitted})\right)$
  - Small values mean that the fitted model explains the data well.

Example in R

```
Call:
glm(formula = vs ~ mpg, family = binomial, data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2127  -0.5121  -0.2276   0.6402   1.6980

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -8.8331     3.1623  -2.793  0.00522 **
mpg           0.4304     0.1584   2.717  0.00659 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 43.860  on 31  degrees of freedom
Residual deviance: 25.533  on 30  degrees of freedom
AIC: 29.533
```

Wald test results are standard normal distributed: small p-values indicate that variables are needed to explain the variation in y

-2 LL

degrees of freedom = no. of observations – no. of predictors

AIC = Residual deviance + 2 (# parameters in the model)

Goodness of Fit

- **Likelihood ratio test:**
  - $\mathrm{D} = -2\ln\left(\frac{\mathrm{L(null)}}{\mathrm{L(fitted)}}\right) = -2\big(\mathrm{LL(null)} - \mathrm{LL(fitted)}\big)$
  - The logarithm of this likelihood ratio (the ratio of the null model to the fitted model) will produce a negative value, hence the need for a negative sign.
  - D follows a χ^2 distribution (the smaller, the better).
  - Non-significant χ^2 values indicate that a significant amount of the variance is unexplained.
  - The test can also be used to assess individual predictors (model with and w/o predictor).

- **A Wald test**
  - Similar purpose than the t-test for the linear regression
  - It is used to test the statistical significance of each coefficient in the model hypothesis that $\beta_i = 0$

  - Remark: t-test and Wald test assume different distributions under the null hypothesis. Therefore, the t-test is appropriate for linear regression (test statistic follows t-distribution), while the Wald test is appropriate for logistic regression (test statistic follows χ^2 distribution).

McFadden R^2 as example of a Pseudo R^2

$$R^2_{McFadden} = 1 - \frac{LL(fitted)}{LL(null)}$$

- If the fitted model does much better than just a constant, in a discrete-choice model this value will be close to 1.
- If the full model doesn't explain much at all, the value will be close to 0.
- Typically, the values are lower than those of R^2 in a linear regression and need to be interpreted with care.
  >0.2 is acceptable, >0.4 is already ok

Calculating Error Rates from a Logistic Regression

- Assume that if the estimated $p(x)$ is greater than or equal to 0.5 then the event is expected to occur and not occur otherwise.
- By assigning these probabilities 0s and 1s and comparing these to the actual 0s and 1s, the % correct Yes, % correct No, and overall % correct scores are calculated.

Example: Error Rate of Predicting Loan Decisions

Classification Table[a]

Observed			Predicted loaned .00	Predicted loaned 1.00	Predicted Percentage Correct
Step 1	loaned	.00	6	11	35.3
		1.00	2	32	94.1
	Overall Percentage				74.5

a. The cut value is .500

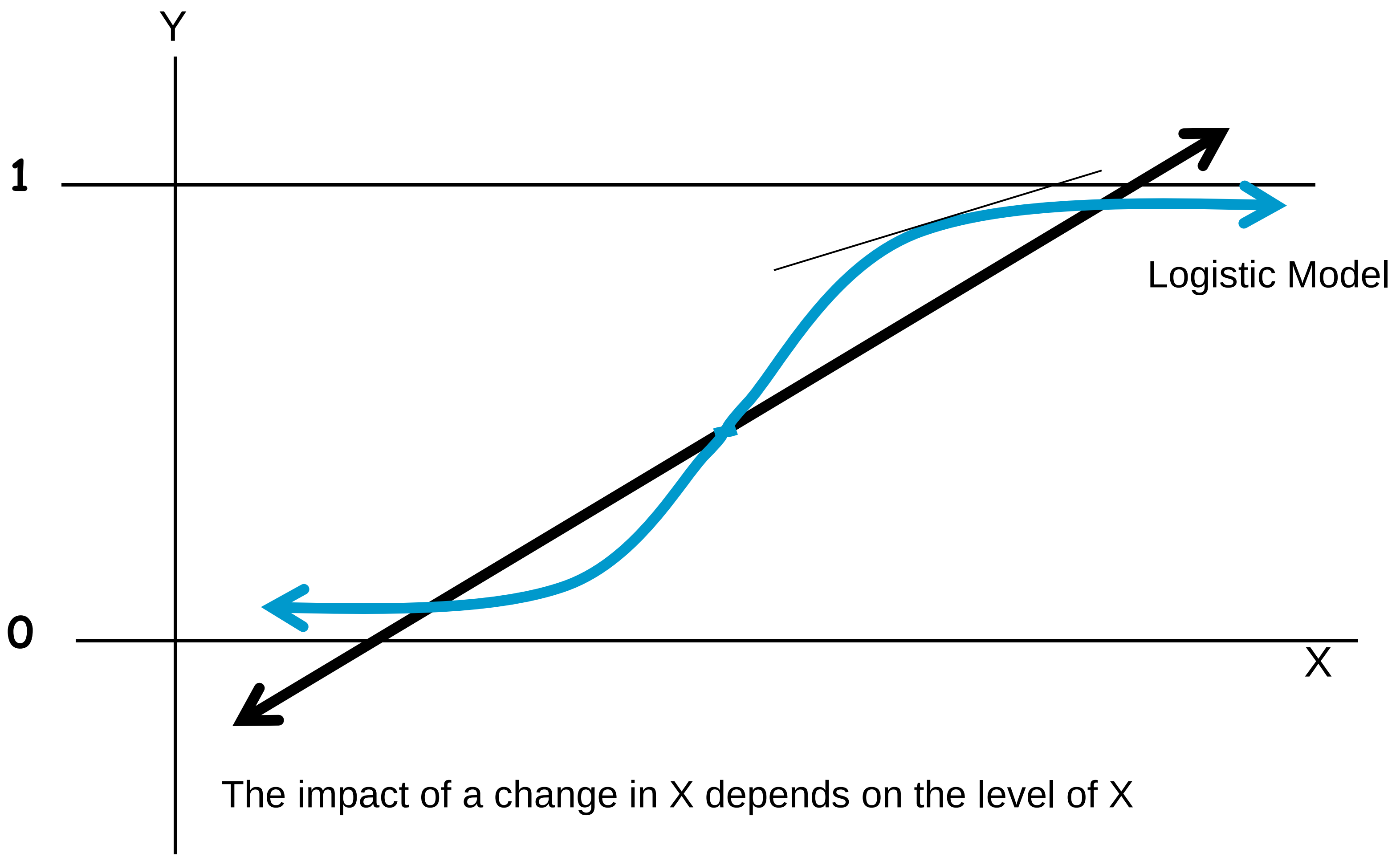
35% of loan rejected cases (0) were correctly predicted

94% of loan accepted cases (1) were correctly predicted

75% of all cases (0,1) were correctly predicted

Note: The model is much better at predicting loan acceptance than loan rejection – this may serve as a basis for thinking about additional variables to improve the model

Interpreting the Coefficients of a Logistic Regression



The impact of a change in X depends on the level of X

Simple Interpretation of the Coefficients

- If $\beta_1 < 0$ then an increase in X_1 => $(0 < e^{\beta_1} < 1)$
  - then odds go down

- If $\beta_1 > 0$ then an increase in X_1 => $(e^{\beta_1} > 1)$
  - then odds go up

- Always check for the **significance** of the coefficients

- But can we say more than this when interpreting the coefficient values?

Example Results: Campaign Response and Age

- Results from logistic regression calculation
  $-ln(p(x)/(1-p(x))) = \beta_0 + \beta_1$ Age

Variable	Estimated Coefficient	Standard Error
Age	0.135	0.036
Constant	-6.54	1.73

Example: Campaign Response

- How can we actually interpret the value .135?
  ln(odds response person #2) = $\beta_0 + \beta_1(X_1 + 1) = \beta_0 + \beta_1(X_1) \ + \beta_1$
  ln(odds response person #1) = $\beta_0 + \beta_1(X_1)$

  → The difference is β_1 (which describes the estimator here)

- So, β_1 = ln(odds “response” person #2)
  – ln(odds “response” person #1)

Example: Campaign Response

- "Reversing" a property of logs:

$$\ln\left(\frac{x}{y}\right) = \ln(x) - \ln(y)$$

$$\beta_1 = \ln\left(\frac{odds\_of\_response\_person\#2}{odds\_of\_response\_person\#1}\right)$$

- β_1 = ln(*odds ratio* for person #2
  compared to person # 1)
  = ln(*odds ratio* comparing two age groups
  who differ by one year in age)

Example: Campaign Response

- So, $\beta_1 = \ln(odds\ ratio)$, then we can get the estimated odds ratio, OR, by e^{β_1}
- So, in our example, $OR = e^{0.135} = 1.14$
- Example:
  - If we were to compare 2 people (or two groups of people) 60 years old and 59 years old respectively, the odds ratio for response of the 60-year-old to the 59-year-old is 1.14
- In fact, if we compared any two people (groups) who differed by year of age, older to younger, the odds ratio would be 1.14 . . .
  - 27 to 26 year olds
  - 54 to 53 year olds . . . etc . .

Multicollinearity and Irrelevant Variables

- The presence of *multicollinearity* will not lead to biased coefficients, but it will have an effect on the standard errors.
  - If a variable which you think should be statistically significant is not, consult the correlation coefficients or VIF.
  - If two variables are correlated at a rate greater than .6, .7, .8, etc. then try dropping the least theoretically important of the two.
- The inclusion of irrelevant variables can result in poor model fit.
  - You can consult your Wald statistics and remove irrelevant variables.

Multiple Logistic Regression

- More than one independent variable

$$\ln\left(\frac{p(X)}{1-p(X)}\right)=\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots\beta_n X_n$$

- Interpretation of β_1
  - Increase in log-odds for a one unit increase in x_i with all the other $x_j, j \neq i$ constant

$$\Pr[Y=1|X]=\frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n)}}$$

Multinomial Logit Models

- The dependent variable, $Y,$ is a discrete variable that represents a choice, or category, from a set of mutually exclusive choices or categories.
  – Examples are brand selection, transportation mode selection, etc.

- Model:
  – Choice between $J > 2$ categories
  – Dependent variable $y = 1, 2, ..., J$

- If characteristics that vary over alternatives (e.g., prices, travel distances, etc.), the multinomial logit is often called “conditional logit”.
- Ordered logit models have ordinal dependent variables.

Can you explain the maximum likelihood estimator based on the logistic regression?

Generalized Linear Models (GLM)

- The logit model is an example of a *generalized linear model (GLM)*
- GLMs are a general class of linear models that are made up of three components: **Random**, **Systematic**, and **Link Function**
  - Random component: Identifies dependent variable (μ) and its probability distribution
  - Systematic component: Identifies the set of explanatory variables $(X_1, \ldots, X_k)$
  - Link function: Identifies a function of the mean of the distribution function μ as a linear function of the explanatory variables

$$g(\mu) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

Common Link Functions

- Identity link (form used in *normal* linear regression models):

  $$g(\mu) = \mu = X\beta$$

  $-\mu$ is the mean of the distribution function
- Logit link (used when μ is bounded between 0 and 1 as when data are binary):

  $$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$$
- Log link (used when μ cannot be negative as when data are *Poisson* counts):

  $$g(\mu) = \log(\mu)$$

Count Variables as Dependent Variables

- Many dependent variables are ***counts***: Non-negative integers
  - \# Crimes a person has committed in lifetime
  - \# Children living in a household
  - \# new companies founded in a year (in an industry)
  - \# of social protests per month in a city

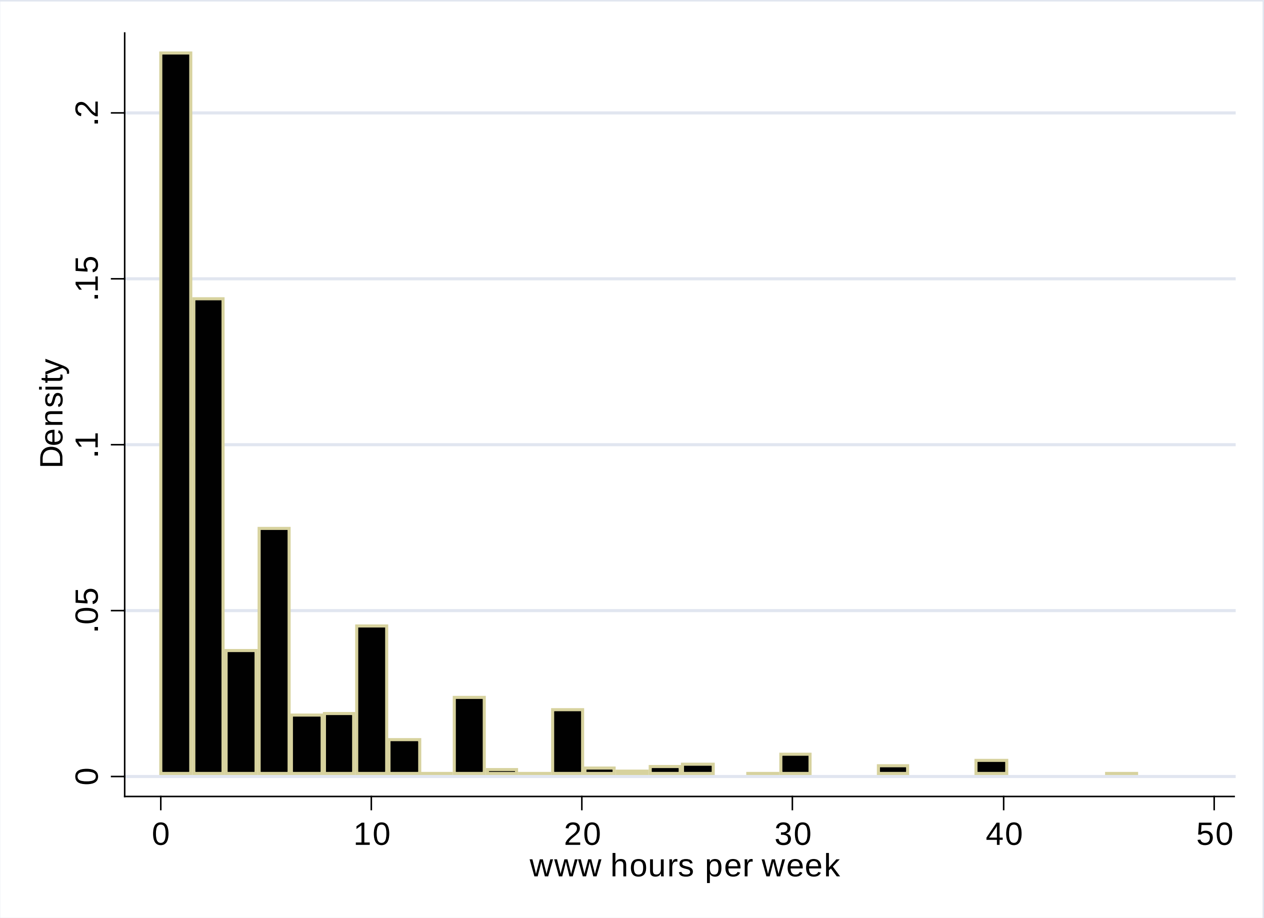
Count Variables

Count variables can be modeled with OLS regression… but:

1. Linear models can yield negative predicted values, whereas counts are never negative.
2. Count variables are often highly skewed
   - # crimes committed this year… most people are zero or very low; a few people are very high.
   - Extreme skew violates the normality assumption of OLS regression.

Poisson Regression: Example

• Hours per week spent on web

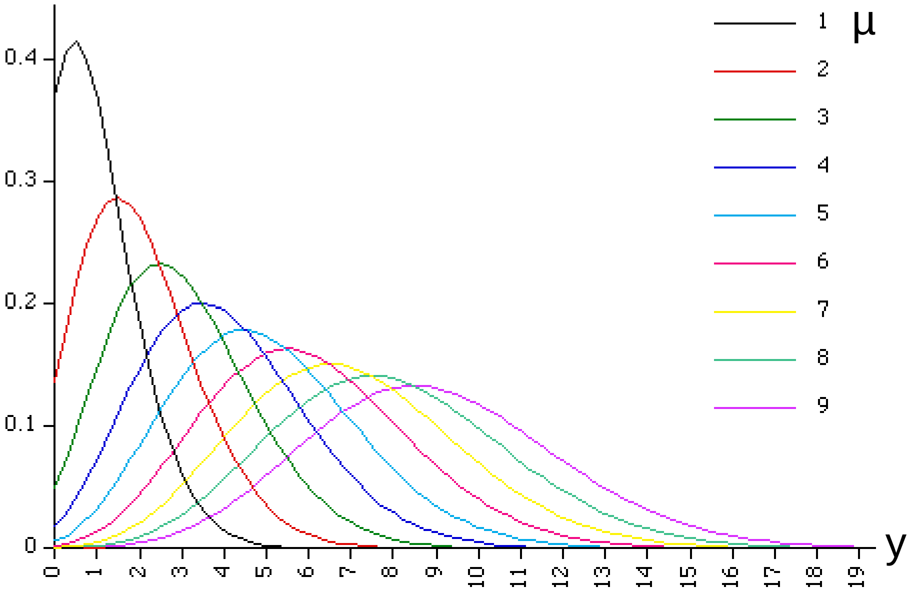


Count Models

- Two most common count models:
  - Poisson regression model (aka. log-linear model)
  - Negative binomial regression model

- Both assume the observed count is distributed according to a Poisson distribution:
  - $\mu = E(y_i|x_i) \geq 0$
    (exp. count and variance)
  - y = observed count

  $\Pr[y|\mu] = \frac{e^{-\mu}\mu^{y}}{y!}$, random component with $y = \{1,2,\ldots\}$



Specification of the Model

The most common link function of this model is the log-linear specification:

$$\ln(\mu) = \ln\left(e^{\beta' x}\right)$$

The random component, the pmf, of the model is

$$\Pr[Y|X] = p(X) = \frac{\mu^y e^{-\mu}}{y!} = \frac{e^{y\beta' x} e^{-e^{\beta' x}}}{y!}$$

$$L(\beta|X,Y) = \prod_{i=1}^{n} \frac{e^{y_i \beta' x_i} e^{-e^{\beta' x_i}}}{y_i!}$$

Now logarithmize and maximize the likelihood.

$$\log L(\beta|X,Y) = \sum_{i=1} \left(y_i \beta' x_i - e^{\beta' x_i} - \log(y_i!)\right)$$

The negative log-likelihood is a concave function and gradient ascent can again be applied to find the optimal value for β.

Interpreting Coefficients

- In Poisson Regression, y is typically conceptualized as a rate…
  - Positive coefficients indicate higher rate; negative = lower rate

- Like logit, Poisson models are non-linear
  - Coefficients don’t have a simple linear interpretation

- Like logit, model has a log form; exponentiation aids interpretation
  - Exponentiated coefficients are multiplicative
  - Analogous to odds ratios … but called “incidence rate ratios”.

Example: Purchasing Decision

How does purchasing tickets to rock concerts in a year depend on the status of a person (student/no student)

$$\ln(\mu) = -0.282 + (0.388)(X_{Student})$$

- No student
  - $\ln(\mu) = -0.282 + (0.338)(0) = -0.282$
  - $\mu = e^{-0.282} = 0.754$
  - $\mu = 0.75\ tickets\ bought$

- Student
  - $\ln(\mu) = -0.282 + (0.388)(1) = 0.106$
  - $\mu = e^{0.106} = 1.112$
  - $\mu = 1.11\ tickets\ bought$

Example: Web Use

```
. poisson wwwhr male age educ lowincome babies

Poisson regression                                Number of obs   =       1552

Log likelihood =  -8598.488                       Pseudo R2       =     0.0297

------------------------------------------------------------------------------
       wwwhr |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   .3595968   .0210578    17.08   0.000     .3183242    .4008694
         age |  -.0097401   .0007891   -12.34   0.000    -.0112867   -.0081934
        educ |   .0205217    .004046     5.07   0.000     .0125917    .0284516
   lowincome |  -.1168778   .0236503    -4.94   0.000    -.1632316   -.0705241
      babies |  -.1436266   .0224814    -6.39   0.000    -.1876892   -.0995639
       _cons |   1.806489   .0641575    28.16   0.000     1.680743    1.932236
------------------------------------------------------------------------------
```

Men spend more time on the web than women

Number of babies in household reduces web use

Interpreting Coefficients

```
. poisson wwwhr male age educ lowincome babies

Poisson regression                                Number of obs   =       1552

Log likelihood =  -8598.488                       Pseudo R2       =     0.0297

------------------------------------------------------------------------------
       wwwhr |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   .3595968   .0210578    17.08   0.000     .3183242    .4008694
         age |  -.0097401   .0007891   -12.34   0.000    -.0112867   -.0081934
        educ |   .0205217    .004046     5.07   0.000     .0125917    .0284516
   lowincome |  -.1168778   .0236503    -4.94   0.000    -.1632316   -.0705241
      babies |  -.1436266   .0224814    -6.39   0.000    -.1876892   -.0995639
       _cons |   1.806489   .0641575    28.16   0.000     1.680743    1.932236
------------------------------------------------------------------------------
```

Exponentiation of 0.359 = 1.43; Rate is 1.43 times higher for men

(1.43-1) * 100 = 43% more

Exp(-0.14) = 0.87. Each baby reduces rate by factor of 0.87

(0.87-1) * 100 = 13% less

Poisson Model Assumptions

- Poisson regression makes a big assumption:
  That $E(Y) = Var(Y) = \mu$
  - In other words, the mean and variance are the same
  - This assumption is often **not met** in real data
  - Variance is often greater than μ: **overdispersion**

- Consequence of overdispersion:
  - Standard errors will be underestimated
  - Potential for overconfidence in results; rejecting H_0 when you shouldn't!

- Negative binomial regression as alternative to Poisson regression.

Zero-Inflation

- If outcome variable has many zero values it tends to be highly skewed

- But, sometimes you have LOTS of zeros. Even Negative binomial regression isn't sufficient
  - Model under-predicts zeros, doesn't fit well

- Examples:
  - # violent crimes committed by a person in a year
  - # of wars a country fights per year
  - # of foreign subsidiaries of firms.

Zero-Inflation

- Logic of zero-inflated models: Assume two types of groups in your sample
  - Type A: Always zero – no probability of non-zero value
  - Type ~A: Non-zero chance of positive count value
    - Probability is variable, but not zero

- 1. Use logit to model group membership (A or ~A)
- 2. Use Poisson or NB regression to model counts for those in group ~A
- 3. Compute probabilities based on those results.

- Various alternative approaches to combat zero-inflation

General Remarks

- Poisson & negative binomial models suffer all the same basic issues as “normal” regression, and you should be careful about
  - Model specification / omitted variable bias
  - Multicollinearity
  - Outliers

- Also, it uses maximum likelihood
  - $n > 500$ = fine; $n < 100$ can be worrisome
    - Results aren’t necessarily wrong if $n < 100$, but less reliable
  - Plus ~10 cases per independent variable.