LOGISTIC REGRESSION

Applied Analytics: Frameworks and Methods 1

Outline

- About Logistic Regression
- Comparison to Linear Regression
- Logit model and estimation
- Visualization of logit function on binary data
- Interpreting Results
- Case of iPad sales on eBay

Logistic Regression

- Is a predictive modeling technique
- Similar to linear regression in many ways
- Differs in the nature of the outcome variable. In logistic regression, outcome can only take two values
  - *Therefore, this is commonly used for classification problems*
- Simplest of a family of models called Classification Models. Other classification models include discriminant analysis, trees and other tree-based models.

Research Questions Answered

- What are the variables that distinguish loyal customers from switchers?
- Who is likely to respond to an email marketing offer?
- Who is going to buy and who isn't going to buy?
- Which variables influence whether a person responds to a credit card offer or not? Are men under 24 with cats as pets good prospects for credit card offers!
- Which is the most important variable in determining a persons responsiveness to a credit card offer? Is it income, age, pet ownership, etc.?

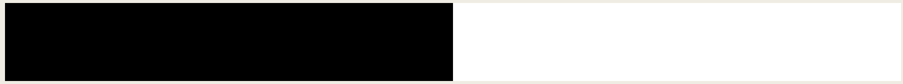
Linear Regression vs. Logistic Regression

Linear Regression

- How much is John Doe going to spend?
- Shades of grey

Logistic Regression

- Will John Doe buy?
- Black or White



Logit Model

- In practice, models the probability of a binary outcome being 1.
- Resembles a linear regression model
  - $p(X) = \beta_0 + \beta_1 X$
- Modeling probability of an outcome using a conventional linear regression model above generates meaningless predictions such as probabilities less than 0 or more than 1.
- Using the logistic function ensures $0 < p(Y=1|X) < 1$

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

Logit Model

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

- This can be rewritten as
  - *p(X)/1-p(X)* = $e^{\beta 0\ +\ \beta 1X}$
- ln(p(X)/1-p(X)) = $\beta_0 + \beta_1 X$
- The right side of the above equation resembles linear regression
- The left side is called the log-odds because the term in parentheses is the ratio of likelihood of winning (p(X)) to likelihood of losing (1 – p(X)).

Logit Model

- Logit Model is estimated by a technique called Maximum Likelihood.
- This technique maximizes the following likelihood function to derive estimates of the coefficients

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'})).$$

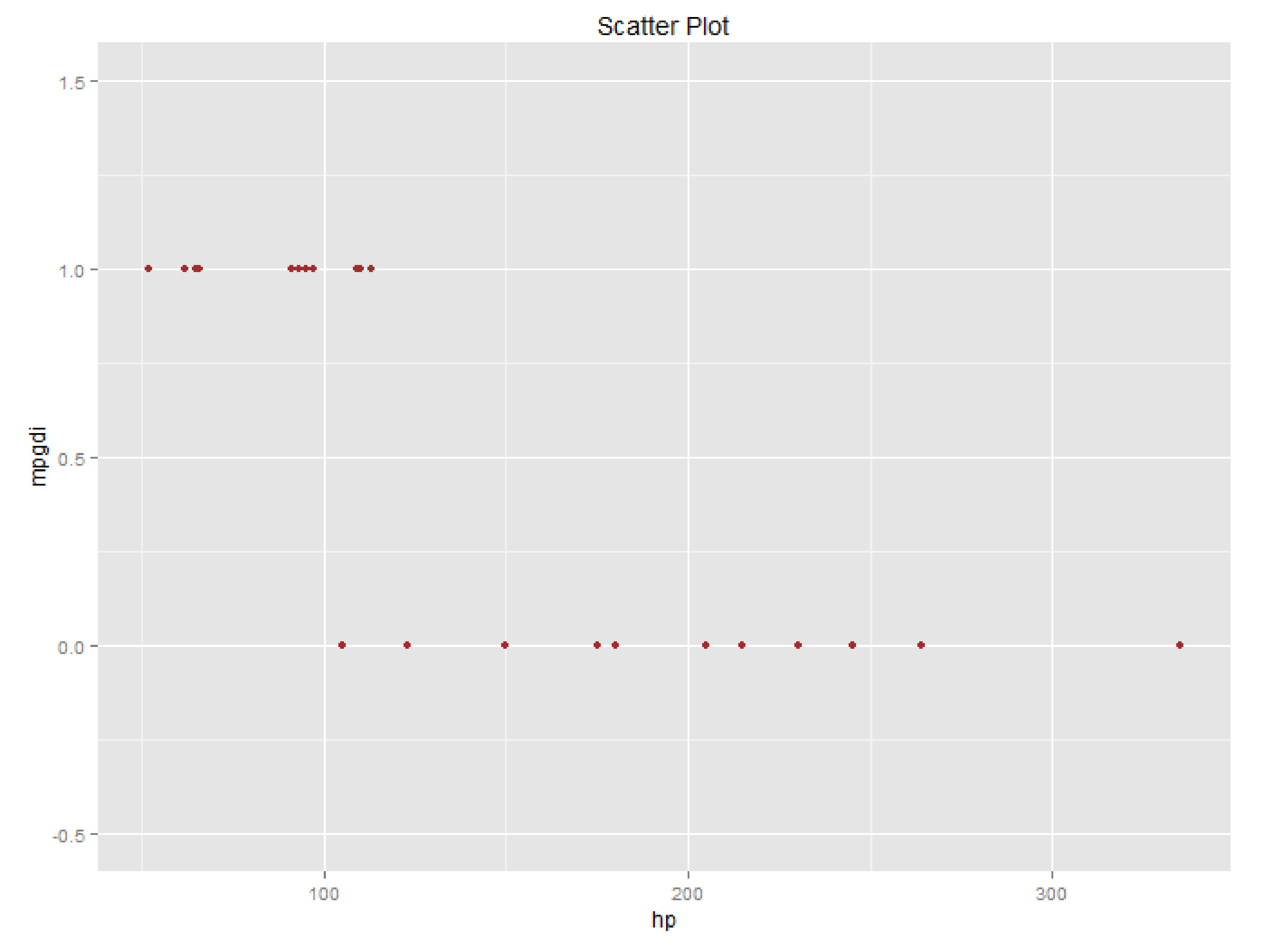
Logit Model Example

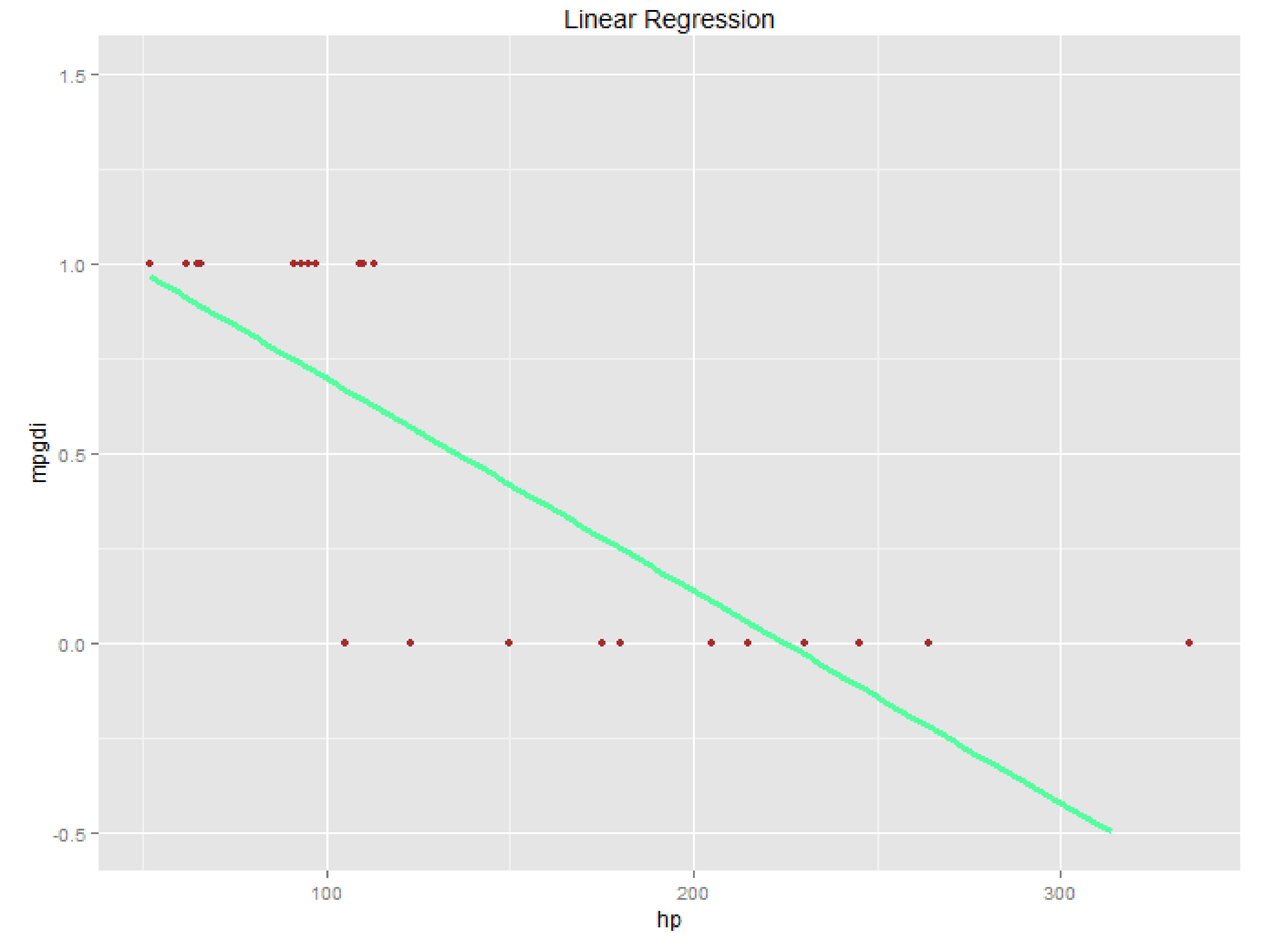
- Imagine a realtor is interested in identifying factors that influence the sale of a house
- Outcome is the Sale of a house in two weeks and p is the probability of sale in two weeks

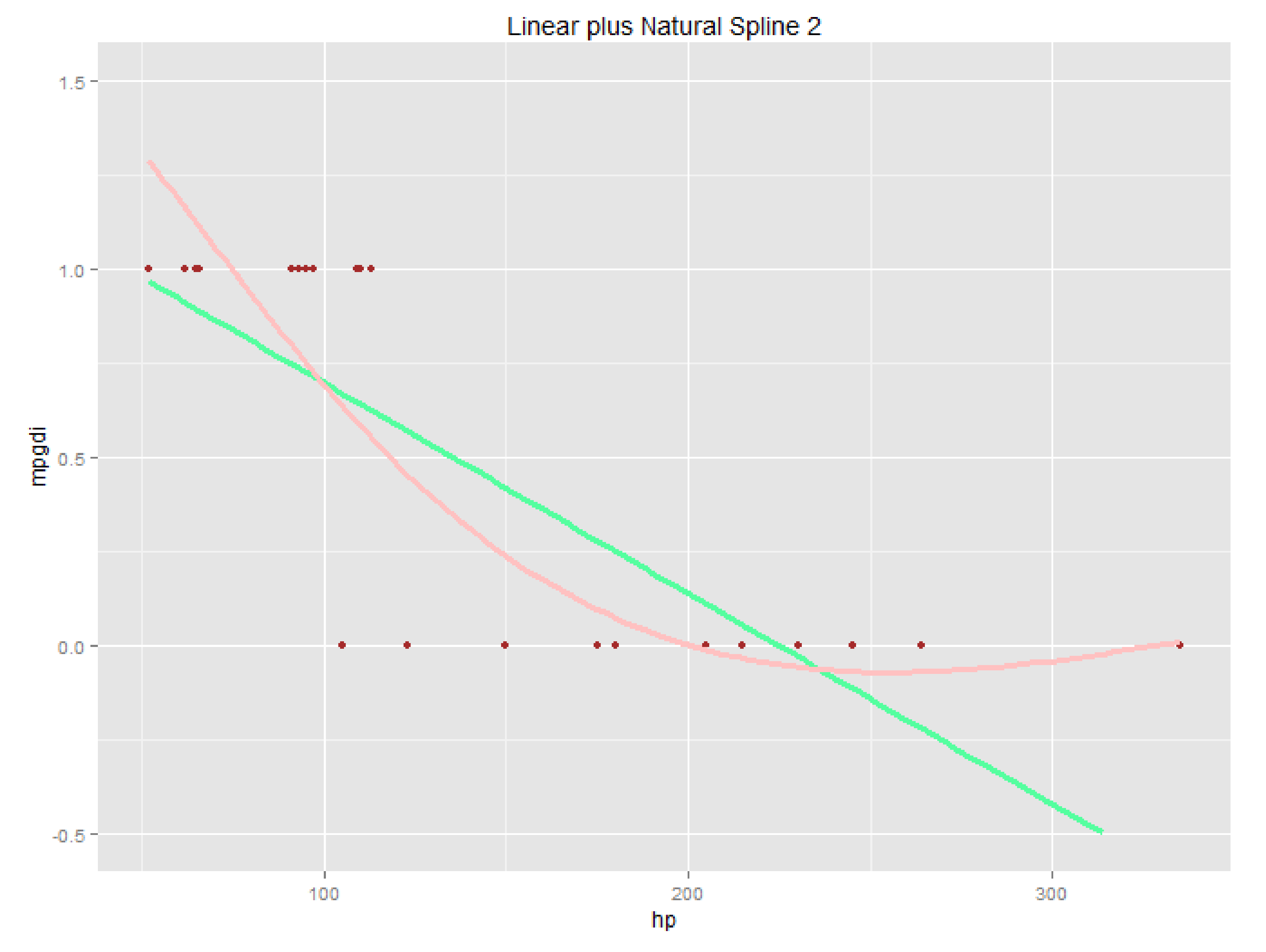
1. Outcome = f(total area, number of rooms, age)
2. ln(p/1-p) = b_0 + b_1*area + b_2*number of rooms + b_3*age
3. p = $e^{\text{b0 + b1*area + b2*number of rooms + b3*age}}$ /1+$e^{\text{b0 + b1*area + b2*number of rooms + b3*age}}$

To generate a binary prediction, probabilities are converted to a binary outcome based on a cutoff value. E.g., if p > 0.6 then 1 else 0.

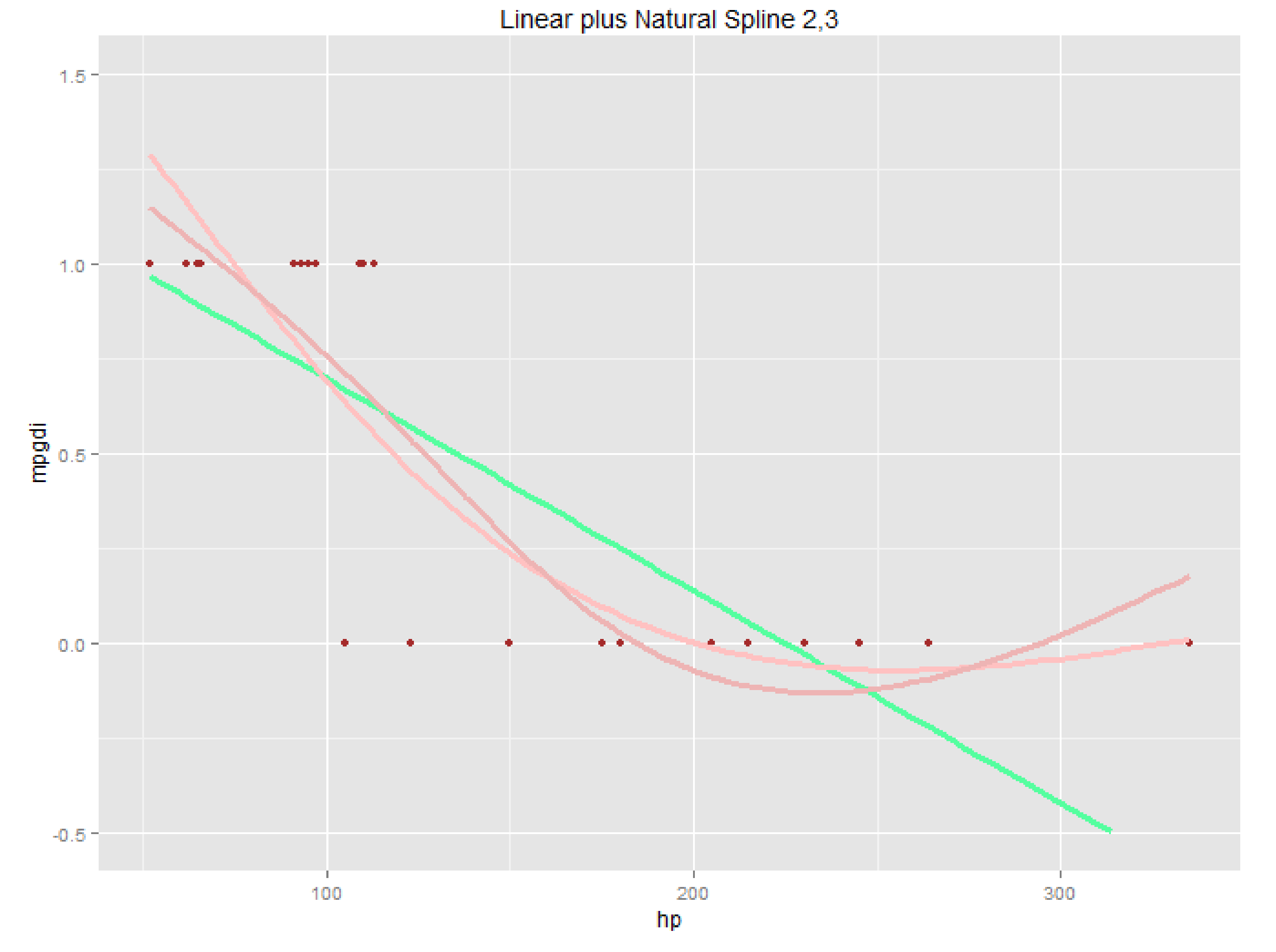
VISUALIZATION OF LOGIT FUNCTION ON BINARY DATA: LOGIT FITS BETTER


Scatter Plot
1.5
1.0
0.5
0.0
-0.5
mpgdi
100
200
300
hp

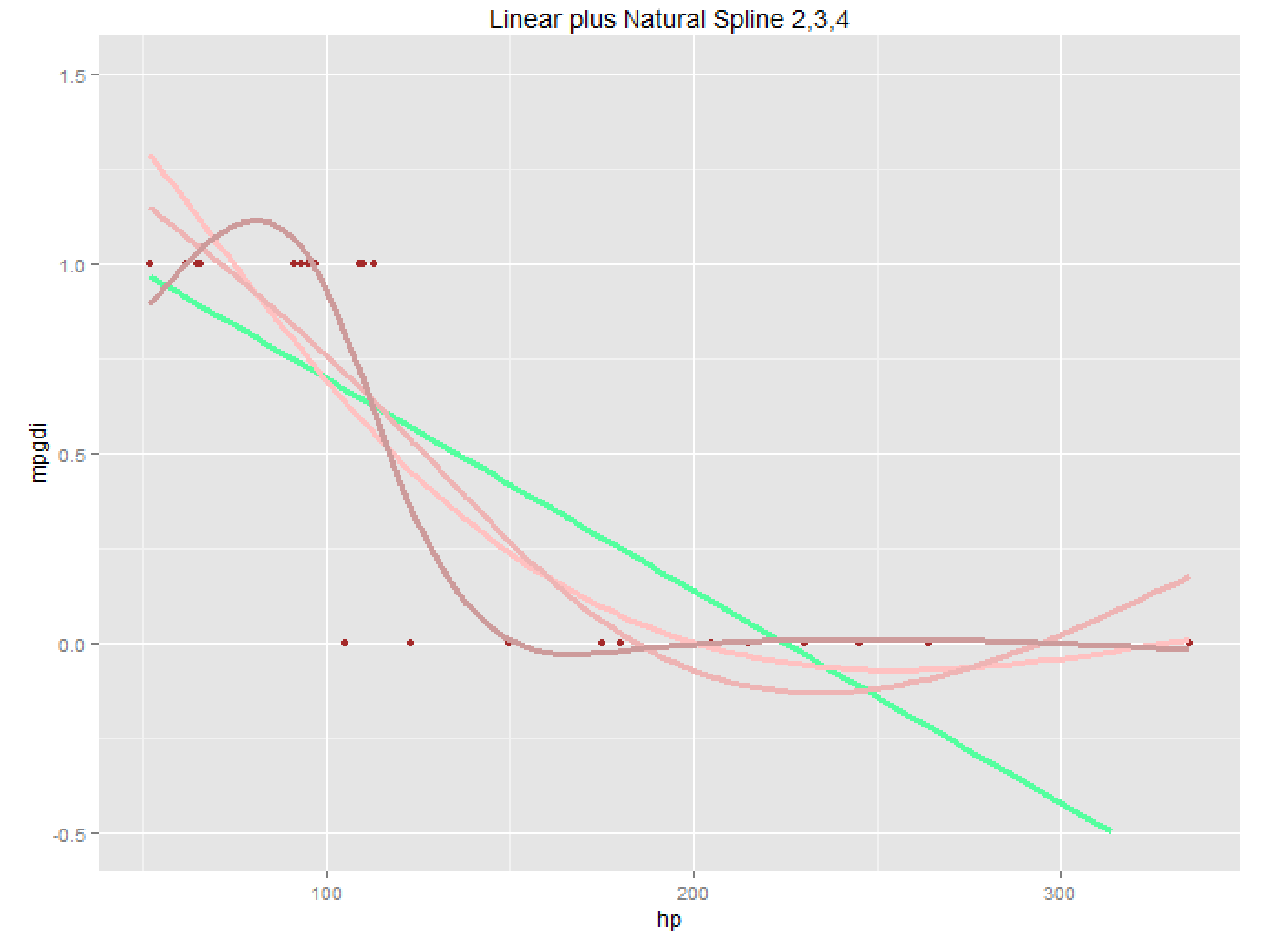

Linear Regression
1.5
1.0
0.5
0.0
-0.5
mpgdi
100
200
300
hp

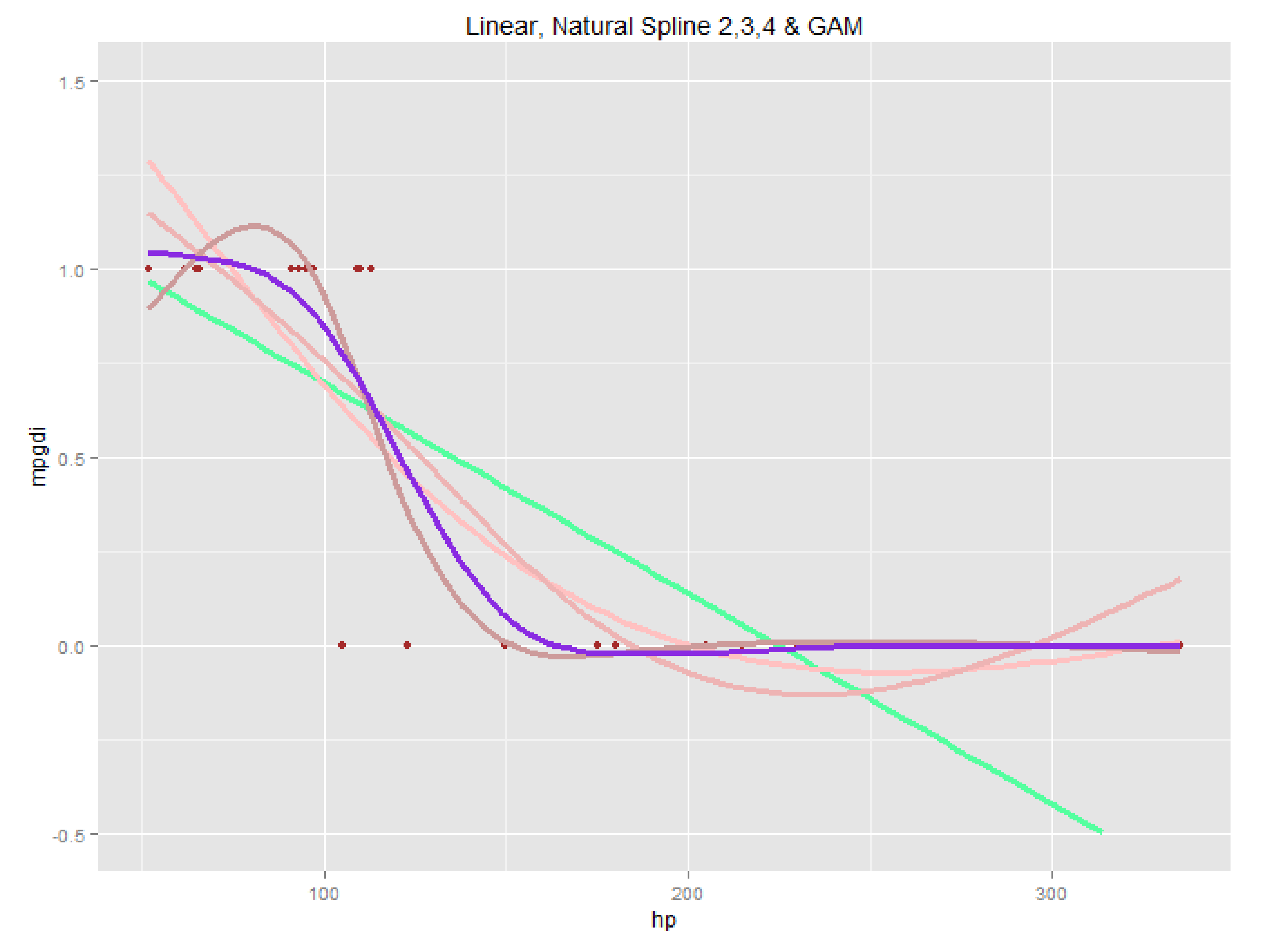


Linear plus Natural Spline 2
1.5
1.0
0.5
0.0
-0.5
mpgdi
100
200
300
hp

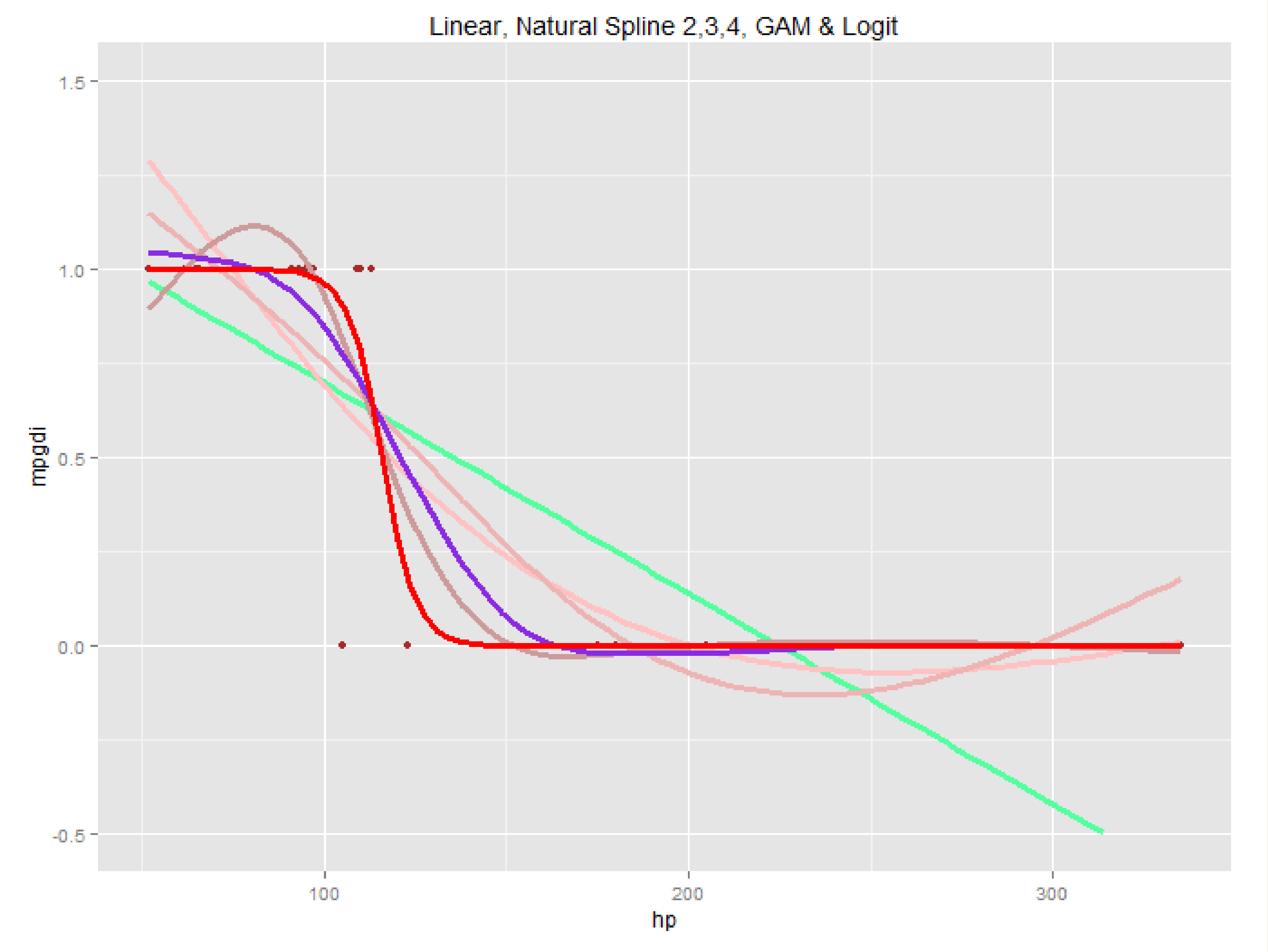


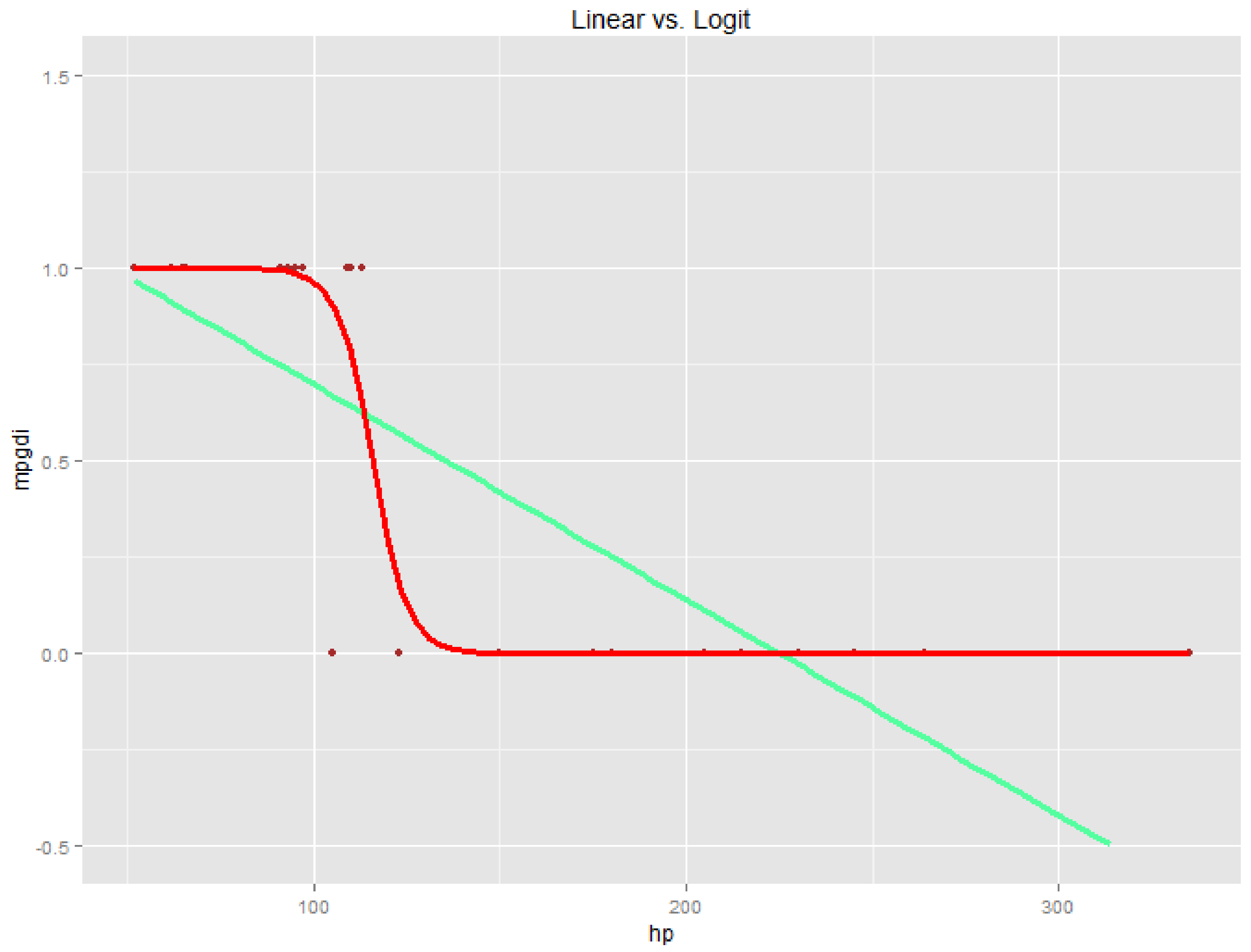
Linear plus Natural Spline 2,3
1.5
1.0
0.5
0.0
-0.5
mpgdi
100
200
300
hp


Linear plus Natural Spline 2,3,4
1.5
1.0
0.5
0.0
-0.5
mpgdi
100
200
300
hp



Linear, Natural Spline 2,3,4 & GAM
1.5
1.0
0.5
0.0
-0.5
mpgdi
100
200
300
hp


Linear, Natural Spline 2,3,4, GAM & Logit
1.5
1.0
0.5
0.0
-0.5
mpgdi
100
200
300
hp


Linear vs. Logit
1.5
1.0
0.5
0.0
-0.5
mpgdi
100
200
300
hp

INTERPRETING RESULTS

Interpreting Results

- As in the case of linear regression, the goals of modeling may be split into
- Inference
  - *Meaningfulness and interpretability of results*
- Prediction
  - *Quality of predictions*

Inference

- Which predictors influence the outcome?
  - *Statistical test to examine individual coefficients*
  - *Statistically significance of z-test indicates an effect*
- Interpretation of coefficients?
  - *If X_i is increased by one unit, the log odds will change by b_i units, when the effect of other independent variables is held constant.*
  - *The sign of b_i will determine whether the likelihood increases (if the sign is positive) or decreases (if the sign is negative) by this amount.*

Prediction or Model Performance

- Is there a relationship between predictors and outcome
  - *Statistical test that compares target model to null model*
  - *Indicated by significant χ^2 difference between deviance of target and null model*
- How strong is the relationship
  - *- 2Log(Likelihood), Pseudo R2 (McFadden R2, Cox and Snell R2, Nagelklerke R2), AIC (=2k-2Log(likelihood))*
- Accuracy of predictions
  - *Accuracy or hit ratio (compared to baseline)*
  - *Specificity*
  - *Sensitivity*
  - *Area Under the Curve (auc)*

Model Performance Measures

- Log-Likelihood Based
  - *Model Chi-square, - 2Log(Likelihood), Pseudo R2 (McFadden R2, Cox and Snell R2, Nagelklerke R2), AIC (=2k-2Log(likelihood))*
- Classification or Confusion matrix
  - *Accuracy or hit ratio (compared to baseline)*
  - *Specificity*
  - *Sensitivity*
- AUC

Model Performance

Log-Likelihood Based

- - 2log(likelihood): Measure of error in the model; lower is better
- Model Chi-square: Difference of error between baseline and new model; higher and statistically significant is better.
- Pseudo R2 (designed to mimic R2 from linear regression)
  - *McFadden R2,*
  - *Cox and Snell R2,*
  - *Nagelklerke R2*
- AIC: Measure of relative quality of model. Lower is better. Can only be used to compare two models. Not meaningful in an absolute sense.

Model Performance
Classification Matrix

- Logistic Regression predicts probability of the occurrence of an event
- E.g., 0.7 implies there is a 70% chance an event will occur
- To convert these probabilities into a binary outcome, we use a threshold or cutoff value
- E.g., if cutoff value is set to 0.5
  - *p = 0.72 would be predicted to have Y=1*
  - *p = 0.34 would be predicted to have Y = 0*
  - *Choice of threshold is often defined based on the cost associated with a False Positive and a False Negative; this would vary widely based on application area.*
  - *False Positive: Predicted 1 but true value is 0; Type 1 Error*
  - *False Negative: Predicted 0 but true value is 1; Type 2 Error*

Model Performance

Classification Matrix

		Predicted		
		No	Yes	
Actual	No	True Negatives (TN)	False Positive (FP)	Specificity = TN/(TN+FP) False Positive Rate = 1- Specificity)
	Yes	False Negative (FN)	True Positives (TP)	Sensitivity = TP/(TP+FN) True Positive Rate = Sensitivity
				Accuracy or Hit Ratio = (TN+TP)/(TN+TP+FN+FP)

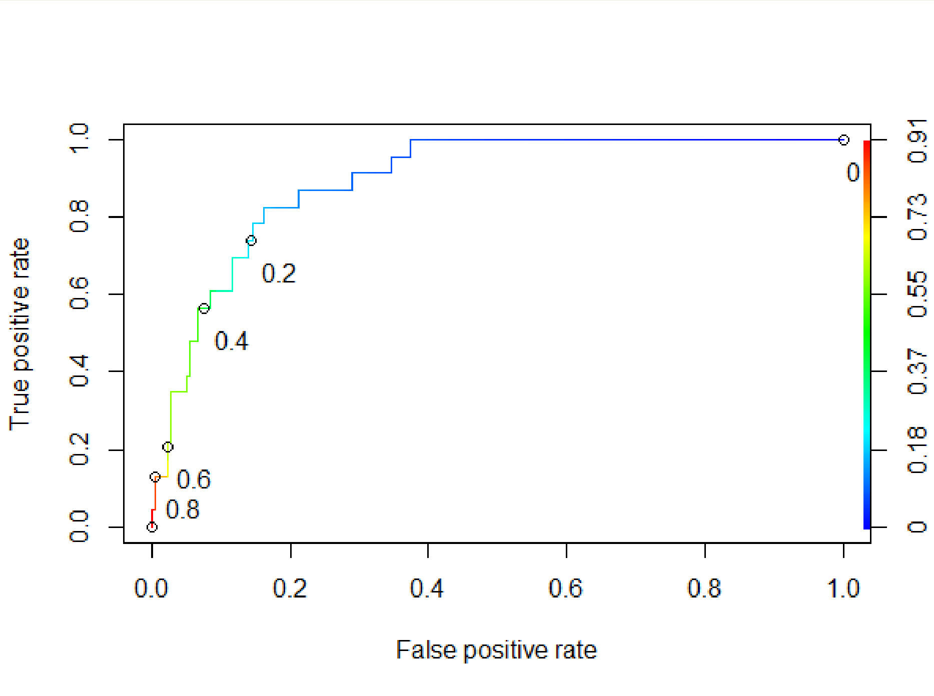
Source: Wikipedia

Model Performance
Classification Matrix

- Generally
  - *High threshold => Higher specificity and Lower sensitivity*
  - *Low threshold => Lower specificity and higher sensitivity*
- What is a good threshold?

Model Performance
AUC

- ROC curve
  - *X-axis: False positive rate (1 – specificity)*
  - *Y-axis: True positive rate (Sensitivity)*
- AUC
  - *Area under the ROC curve*
  - *0 < AUC < 1*



Summary

- In this module we
  - Learnt when logit models are useful
  - Contrasted logit with linear regression
  - Examined estimation of logit models
  - Learnt how to interpret results of logit models