First Model

Will Jones

December 8, 2015

Introduction

As our first steps in modeling ride rating, we will start to model without route data. Instead we will focus on other question in the modeling as a start for our model:

- How much variation is there between riders in how they tend to rate rides?
- What relationship does weather, like rain or wind speed, have with ride rating?
- How does ride rating fluxuate with time of day (which we use as a proxy for traffic)?

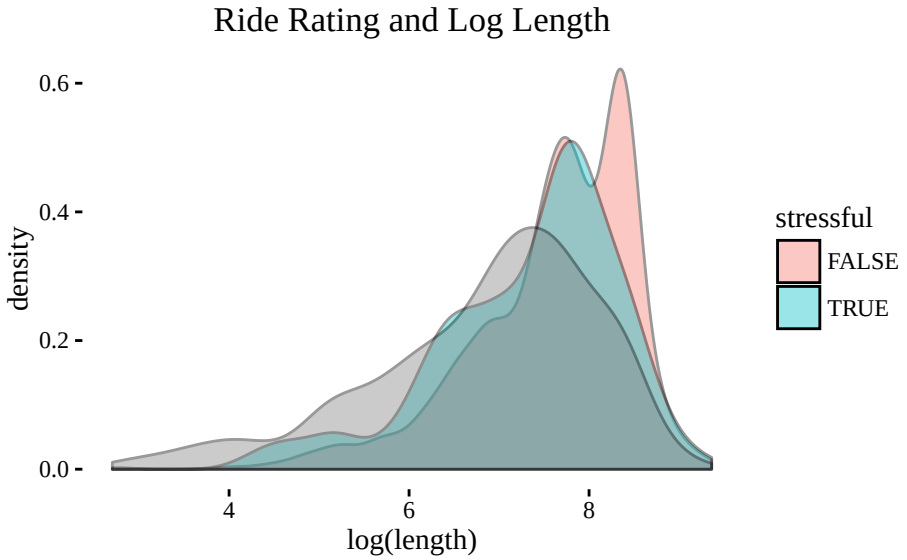
We actually expect a fair amount of the variance in ride rating to be explained by these variables, based on tests of a smaller sample.

Some Numbers about the Data

There are 1515 rides in the data set, with 238 (15.709571%) rides with no rating.

What variables will we include?

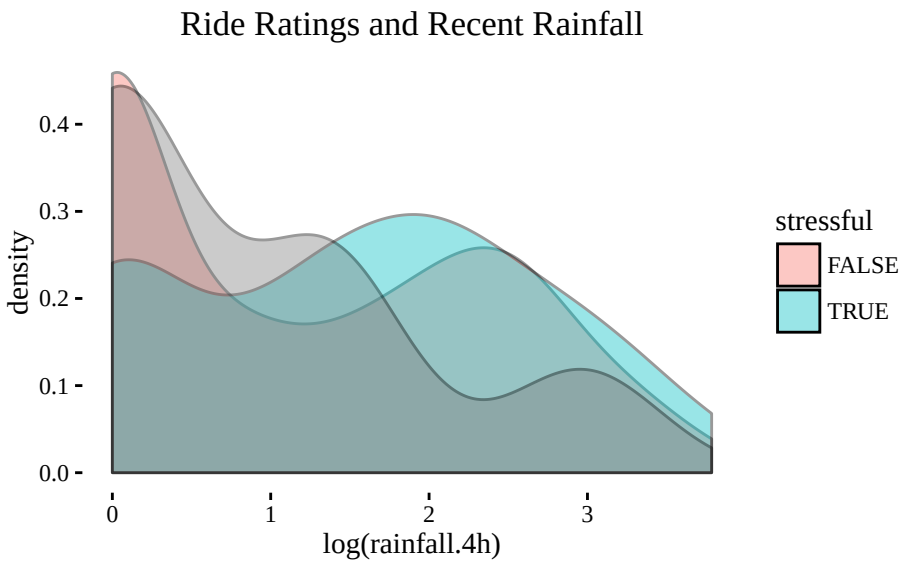
Length



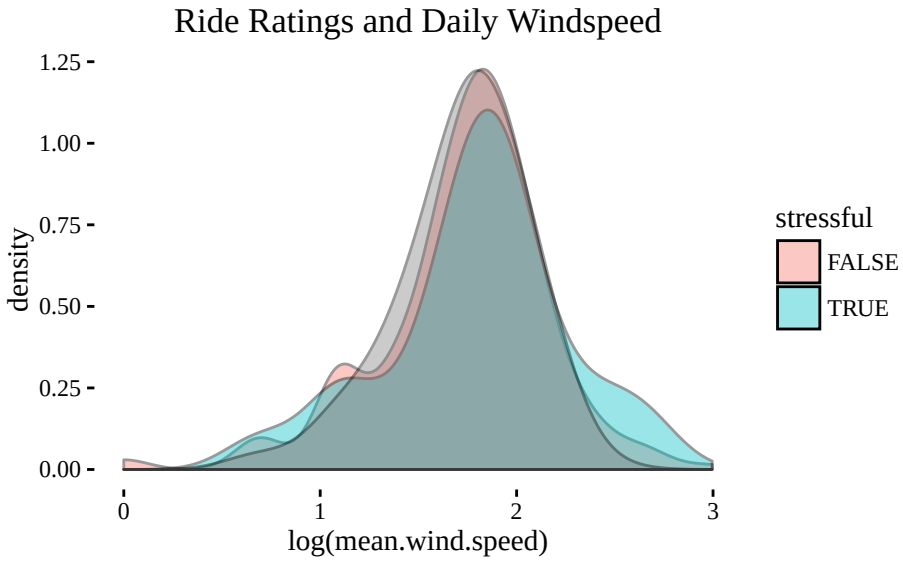
Weather

We also want to consider patterns with weather. We have data on daily weather, including wind speed, temperature highs and lows, and rain data. But we also have hourly rain data from a local fire station.

```
## Warning: Removed 2533 rows containing non-finite values (stat_density).
```

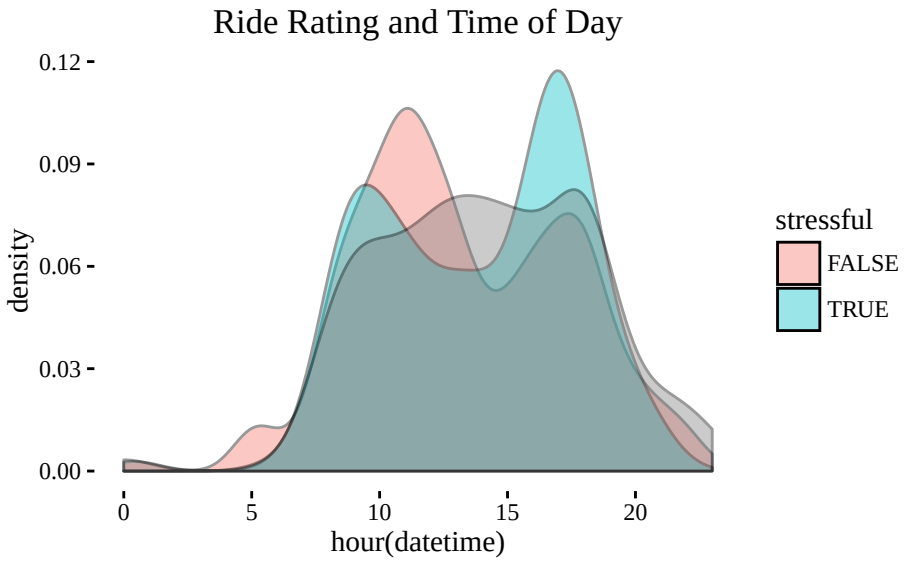



```
## Warning: Removed 176 rows containing non-finite values (stat_density).
```



Traffic / Daily Trends

We would like to incorporate traffic, but to simplify our model, we may simple use time of day as a proxy.



The Models

Intercept Baseline Model

First, we might simply try to model ride rating by modeling ride rating Y as

$$Y \sim \text{Bernoulli}(p),$$

where p is just the probability that a ride is rated "stressful". Essentially, what is p?

Classical Model

We also want to consider how a classical logistic regression model compares to a model with a random intercept for riders. So we will model:

$$Y = \text{logit}^{-1}\left(\alpha + \beta_1 \cdot \text{log.length} + \beta_2 \cdot \text{log.windspeed} + \beta_3 \cdot \text{log.rainfall.4h}\right).$$

Just Rider Random Effects

Now we want to explore how we can capture variance with and between riders. So we will use the basic model

$$Y \sim \text{Bernoulli}(\text{logit}^{-1}(\alpha_{j[i]})), \quad \alpha_{j[i]} \sim \text{Normal}(\mu_\alpha, \sigma_\alpha^2).$$

Add Time of Day Effects

Now we want to add effects based on time of day. We will try using polynomial regression to do this first, by adding to our regression the terms,

$$\beta_1 \cdot \text{hour} + \beta_2 \cdot \text{hour}^2 + \beta_3 \cdot \text{hour}^3 + \beta_4 \cdot \text{hour}^4.$$

All Effects

Our last model will take the rider intercepts and day effects and add the terms we had in our first regression with variables.

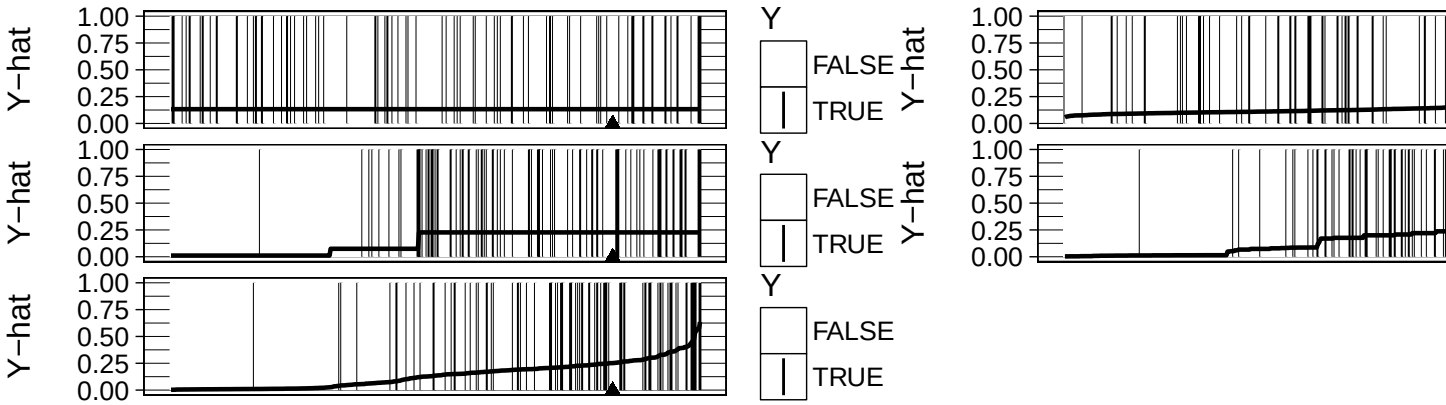
Table of coefficients

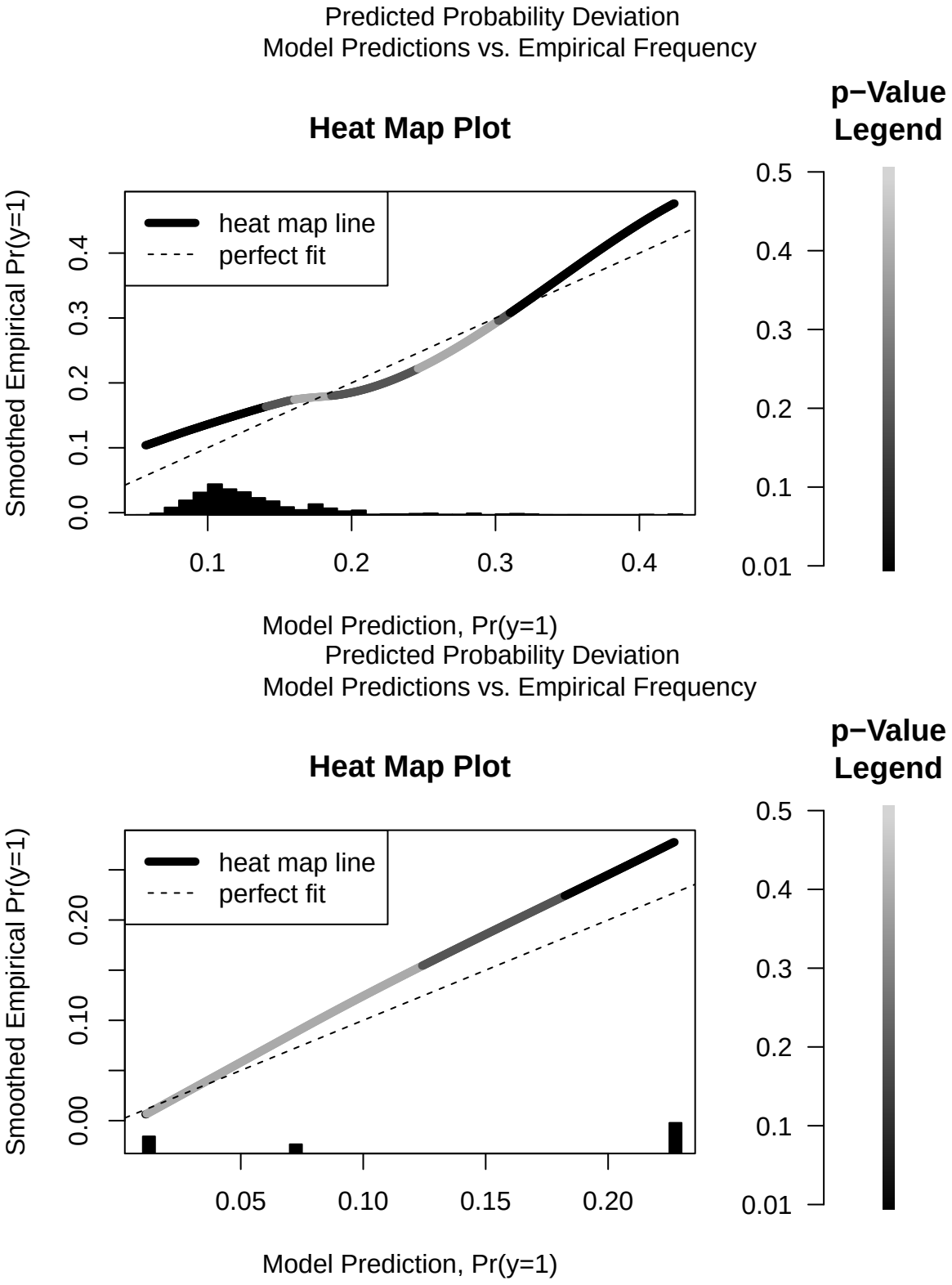
% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Jan 27, 2016 - 21:10:26 % Requires LaTeX packages: dcolumn

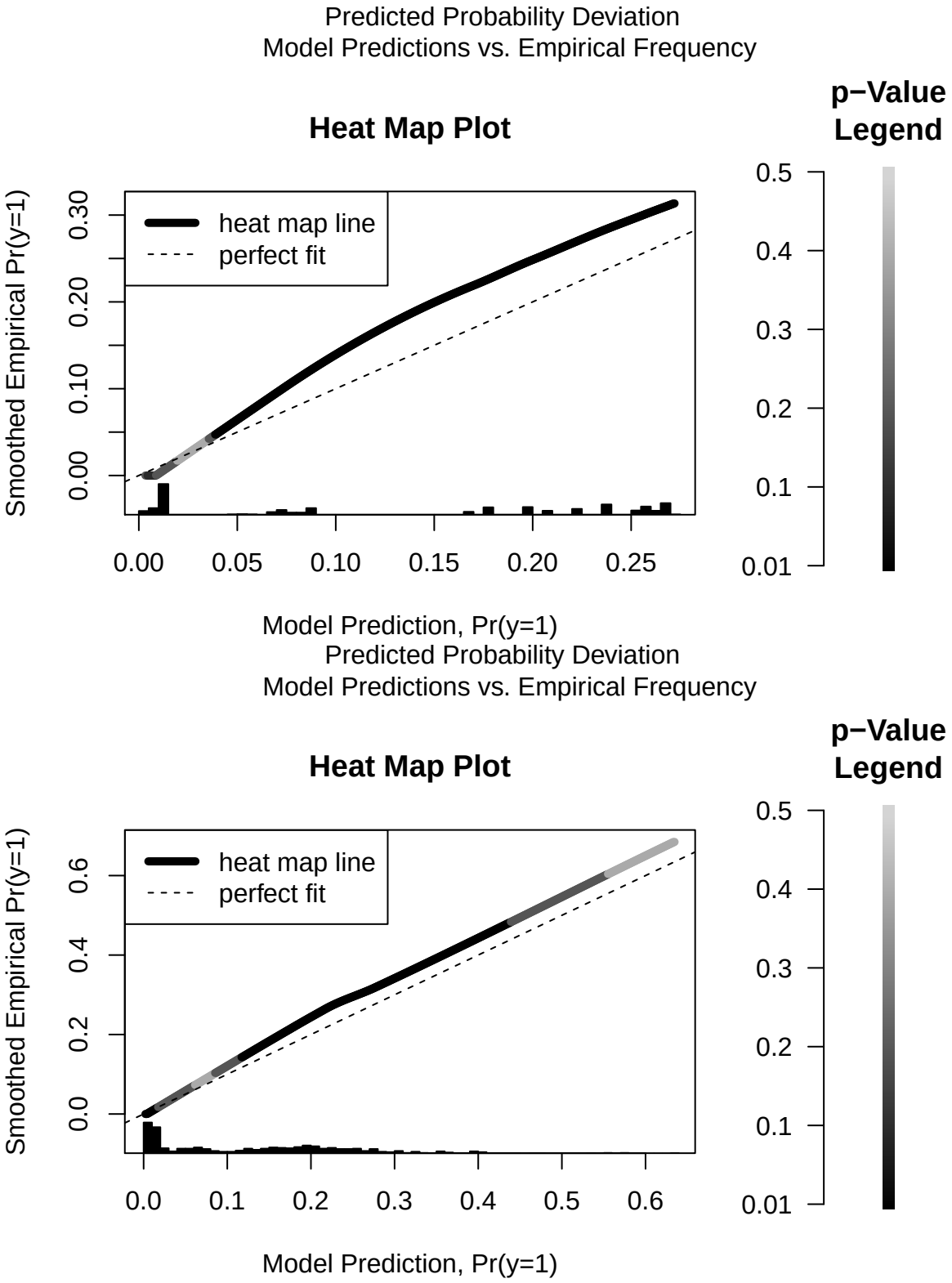
Table 1: results

	Dependent variable:				
	stressful				
	logistic			*generalized linear mixed-effects*	
	(1)	(2)	(3)	(4)	(5)
log.length		−0.343***			−0.421***
		(0.077)			(0.083)
rainfall.4h		0.002			−0.010
		(0.021)			(0.023)
mean.wind.speed		0.092***			0.070**
		(0.030)			(0.029)
hour				0.160	0.229
				(0.153)	(0.160)
I(hurˆ2)				−0.178	0.034
				(0.191)	(0.204)
I(hurˆ3)				−0.035	−0.083
				(0.089)	(0.095)
I(hurˆ4)				−0.026	−0.061
				(0.047)	(0.054)
Constant	−1.874***	−2.457***	−2.772***	−2.584***	−3.159***
	(0.074)	(0.205)	(0.819)	(0.818)	(0.826)
Observations	1,586	1,586	1,586	1,586	1,586
Log Likelihood	-621.908	-608.446	-542.192	-536.674	-522.436
Akaike Inf. Crit.	1,245.815	1,224.892	1,088.385	1,085.348	1,062.873
Bayesian Inf. Crit.			1,099.123	1,117.562	1,111.194
Note:					*p<0.1; **p<0.05; ***p<0.01

Model Accuracy and Fit




Predicted Probability Deviation
Model Predictions vs. Empirical Frequency
Heat Map Plot
p−Value
Legend
heat map line
perfect fit
Smoothed Empirical Pr(y=1)
0.0
0.1
0.2
0.3
0.4
0.1
0.2
0.3
0.4
0.5
0.4
0.3
0.2
0.1
0.01
Model Prediction, Pr(y=1)
Predicted Probability Deviation
Model Predictions vs. Empirical Frequency
Heat Map Plot
p−Value
Legend
heat map line
perfect fit
Smoothed Empirical Pr(y=1)
0.00
0.10
0.20
0.05
0.10
0.15
0.20
0.5
0.4
0.3
0.2
0.1
0.01
Model Prediction, Pr(y=1)


Predicted Probability Deviation
Model Predictions vs. Empirical Frequency
Heat Map Plot
p−Value
Legend
heat map line
perfect fit
Smoothed Empirical Pr(y=1)
0.00
0.10
0.20
0.30
0.00
0.05
0.10
0.15
0.20
0.25
0.5
0.4
0.3
0.2
0.1
0.01
Model Prediction, Pr(y=1)
Predicted Probability Deviation
Model Predictions vs. Empirical Frequency
Heat Map Plot
p−Value
Legend
heat map line
perfect fit
Smoothed Empirical Pr(y=1)
0.0
0.2
0.4
0.6
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.5
0.4
0.3
0.2
0.1
0.01
Model Prediction, Pr(y=1)