STA 371H Final Exam, Spring 2017

Instructions

1) Do not turn this page and begin the exam until instructed to do so.

2) The time limit for this exam is 180 minutes. Be concise in your answers, and you should have no trouble with the time limit.

3) Write in ink in a blue book. I will not grade anything written on the exam sheet itself, so you may use this as scratch paper.

4) Turn in this exam paper along with your written exam.

5) This is a closed-book exam. You are allowed pens and paper on your desk, and nothing more.

6) Switch off all cell phones, mobile communication devices, iPods, and so forth. Do not merely turn them to silent or vibrate mode. If I catch you using your phone, I will collect your exam and give you a zero.

7) The exam has three parts, with point values labeled. You may take these point values as roughly proportional to appropriate length. These point values add up to 170.

Good luck!

Here are some formulas that you may (or may not) find helpful.

Multiplication rule: The joint probability that A and B will both happen is $P(A, B) = P(A) \cdot P(B \mid A)$, where $P(B \mid A)$ is the conditional probability that B will happen, given that A happens.

Bayes' rule: The posterior probability of A, given B, is

$$P(A \mid B) = \frac{P(A) \cdot P(B \mid A)}{P(B)} \, .$$

Rule of total probability: Suppose that events $B_1, B_2, \ldots, B_N$ constitute an exhaustive partition of all possibilities in some situation. That is, the events themselves are mutually exclusive, but one of them must happen. Now consider any event A. The rule of total probability says that

$$P(A) = \sum_{i=1}^{N} P(A, B_i) = \sum_{i=1}^{N} P(B_i) \cdot P(A \mid B_i) \, .$$

Expected value.

If a random variable X has N possible outcomes $\{x_1, \ldots, x_N\}$ having corresponding probabilities $\{p_1, \ldots, p_N\}$, then the expected value is

$$E(X) = \sum_{i=1}^{N} p_i x_i \, .$$

Bivariate normal.

If X_1 and X_2 follow a bivariate normal distribution, i.e.

$$(X_1, X_2) \sim N(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) \, ,$$

then the conditional probability distribution $P(X_2 \mid X_1 = x_1)$ is an ordinary normal distribution, with mean and variance

$$E(X_2 \mid X_1 = x_1) = \mu_2 + \rho \cdot \frac{\sigma_2}{\sigma_1} \cdot (x_1 - \mu_1) \qquad (1)$$

$$\mathrm{var}(X_2 \mid X_1 = x_1) = \sigma_2^2 \cdot (1 - \rho^2) \, . \qquad (2)$$

Binomial distribution. The Binomial(N, p) distribution has expected value $N \times p$ and probability mass function

$$P(X = k) = \binom{N}{k} p^k (1 - p)^{N-k} \, .$$

Logistic regression. In a logit model, the predicted probability that $y_i = 1$ is

$$P(y_i = 1 \mid x_i) = \frac{e^{\psi_i}}{1 + e^{\psi_i}} \, , \quad \psi_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \, .$$

Question 1: Short answers (70 points, 10 points each)

A) Define the term "p-value."

B) Let A be the event "rain tomorrow," and let B be the information "the weather app on your phone says it will rain tomorrow." You know that only 15% of all days in Central Texas are rainy. Moreover, you know the track record of your weather app: when it rains, your app gives the correct forecast (and correctly says it will say tomorrow) 90% of the time. When it doesn't rain, your app raises a false alarm (and incorrectly says it will rain tomorrow) 5% of the time. Show how you would use this information to calculate the probability of rain tomorrow (A), given that your phone says it will rain tomorrow (B). You don't need to calculate a final number, but you must give an expression that *could* be calculated in terms of the information provided.

C) Concisely explain the phenomenon of "regression to the mean," and give an example. What parametric probability model can be used to provide a mathematical description of regression to the mean?

D) Under what circumstances would it be appropriate to use a binomial distribution as a parametric probability model for some real-world system?

E) The following table shows a joint frequency distribution for 1500 people undergoing a medical test. Two variables are shown: whether a patient has the disease, and whether the patient tests positive for the disease. Use the numbers in this frequency distribution to set up the calculations for three probabilities: P(tests positive | has disease); P(has disease | tests positive); and P(doesn't have disease). As with Part B, you don't need to calculate a final number, but you must give an expression that *could* be calculated in terms of the information provided.

	Tests positive	Tests negative
Has disease	85	15
Doesn't have disease	110	1290

F) Suppose that you want to build a regression model to understand how a vehicle's weight affects its gas mileage. However, you suspect that vehicle class (a categorical variable with many levels: sedan, truck, SUV, etc.) may modulate the relationship between weight and gas mileage. So you decide to test for the presence of an interaction between class and weight. Describe (in no more than a few sentences) how you might do this.

G) One way to estimate a causal relationship between a predictor X and a response Y is by running an experiment. Briefly describe two other ways, and list one potential shortcoming of each way.

Question 2: Essay (50 points)

We have encountered the idea of bootstrapping, or resampling from a sample, in two separate contexts in this course:
1) in quantifying uncertainty about parameters in statistical models.
2) in assessing the risk/return properties of financial portfolios.

Explain the use of bootstrapping in each context. How is it done, and why is it done that way? For the "how" part, make sure to provide pseudo-code: that is, the logical steps of a computer program that would allow you to execute the algorithm. Offer some comments on what is similar, and what is different, about the use of bootstrapping in each context.

Question 3: Interpreting data analysis (50 points)

This problem concerns a data set on 673 businesses in downtown New Orleans, Louisiana. The goal is to model how quickly these businesses re-opened following the catastrophic floods caused by Hurricane Katrina, which hit New Orleans on August 29, 2005. The locations of these stores in the data set can be seen in the map below; as you can see, they cluster along three major streets. The fundamental social-scientific question here is to understand the factors that influenced a business's decision to re-open in the presence of post-disaster uncertainty. As an outcome variable, we will look at a simple binary indicator: whether the business re-opened within the first 12 months after the flood.

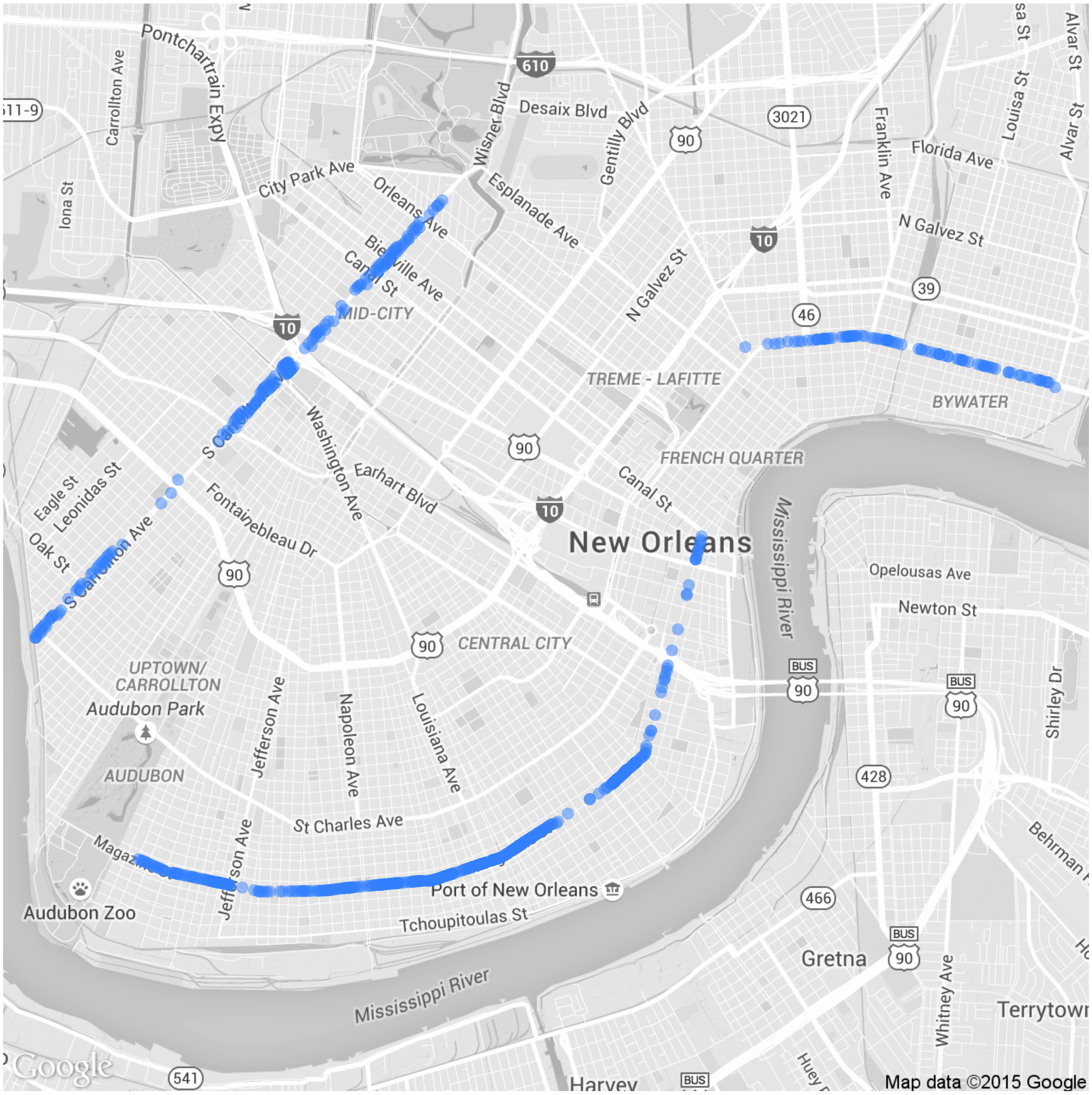


Figure 1: Map of New Orleans. Each dot is a store in the data set.

The data set

The data set has the following variables on 673 businesses in downtown New Orleans:

- flooddepth: maximum flood depth during and after Hurricane Katrina (measured in feet)
- owntype: is the store owned by a sole proprietor, a local chain, or a national chain?
- size: is the store small, medium, or large?
- street: the street where the store is located (Magazine, South Carrollton, or St. Claude).
- reopen12: a binary indicator for whether the store reopened within 12 months after Katrina (1) or not (0).

Over the next several pages, the results of several statistical plots and analyses are shown. Use these results to decide whether the following statements are true, false, or undecidable/ambiguous in light of the evidence provided. If true, cite supporting evidence. If false, propose a correction and cite supporting evidence. If undecidable/ambiguous, explain what evidence you'd like to see in order to decide the question to your satisfaction. (Note: all quoted numbers are rounded off a bit; I'm not trying to trick you here by making subtle rounding errors that invalidate an otherwise true statement.)

The statements: true, false, or undecidable?

A) Higher flood depths were significantly associated with a lower probability that a store would open in the 12 months after Katrina.

B) Stores owned by a national chain had a significantly lower predicted probability of reopening within 12 months after Katrina, compared to stores owned by a sole proprietor, once we account for the severity of the flood at those stores.

C) Higher median incomes were associated with a lower probability of reopening, holding other relevant factors equal.

D) The flood affected stores on different streets in different ways. In particular, the probability of reopening declined more steeply as a function of flood depth for stores on Magazine Street than it did for stores on other streets.

E) The flood affected stores of different sizes in different ways. In particular, the probability of reopening declined more steeply as a function of flood depth for large stores than it did for small stores.

Exploratory tables and plots

The data were read in and stored in a data frame called `nola`.

Reopening status

First, three two-way contingency tables were constructed showing `reopen12` versus `street`, `owntype`, and `size`:

```
xtabs(~reopen12 + street, data=nola)
```

```
##         street
## reopen12 Magazine SCarrollton StClaude
##        0       42         100       53
##        1      359          79       40
```

```
xtabs(~reopen12 + owntype, data=nola)
```

```
##         owntype
## reopen12 local_chain national_chain sole_proprietor
##        0          48               7             140
##        1          61              13             404
```

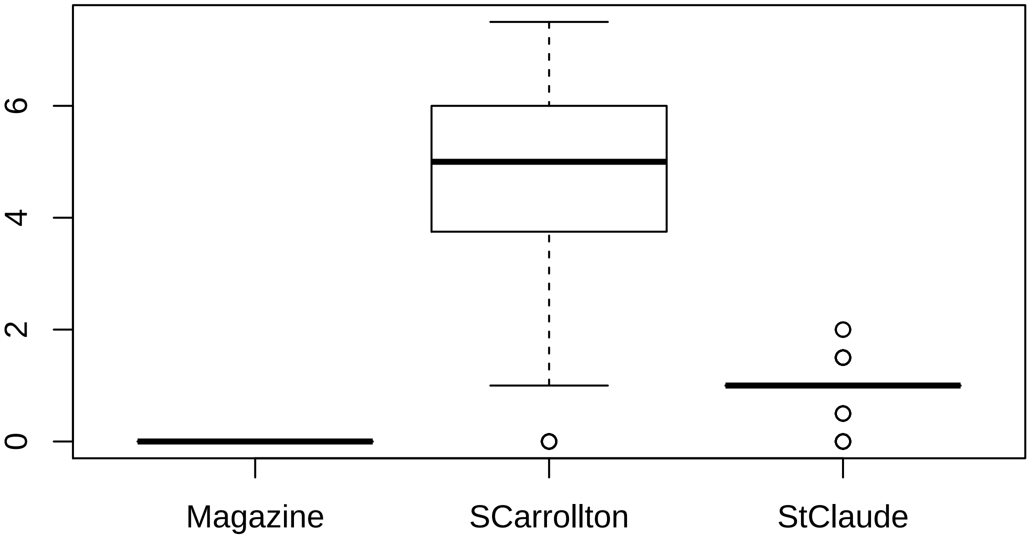
```
xtabs(~reopen12 + size, data=nola)
```

```
##         size
## reopen12 large medium small
##        0    14     62   119
##        1    17    147   314
```

Flood depth

A boxplot of flood depth versus street was created:

```
boxplot(flood_depth ~ street, data=nola)
```



Next, the mean flood depth was calculated under several different ways of stratifying the stores into groups: by reopening status, by ownership type, and by size.

```
mean(flood_depth ~ reopen12, data=nola)
```

```
##     0     1
## 3.085 0.571
```

```
mean(flood_depth ~ owntype, data=nola)
```

```
##     local_chain  national_chain sole_proprietor
##           2.771           1.500           0.997
```

```
mean(flood_depth ~ size, data=nola)
```

```
##  large medium  small
##   2.37   1.75   1.01
```

Models

Several logistic regression models were fit to the data, using `reopen12` as the response variable.

Model 1

The first model used `flood_depth` as a predictor of re-opening status. This model was bootstrapped to get 95% confidence intervals for model parameters.

```
glm1 = glm(reopen12 ~ flood_depth, data=nola, family=binomial)
boot1 = do(1000)*{
  glm(reopen12 ~ flood_depth, data=resample(nola), family=binomial)
}
confint(boot1)
```

```
##          name  lower  upper level     method estimate
## 1   Intercept  1.458  1.914  0.95 percentile    1.666
## 2 flood_depth -0.582 -0.417  0.95 percentile   -0.493
```

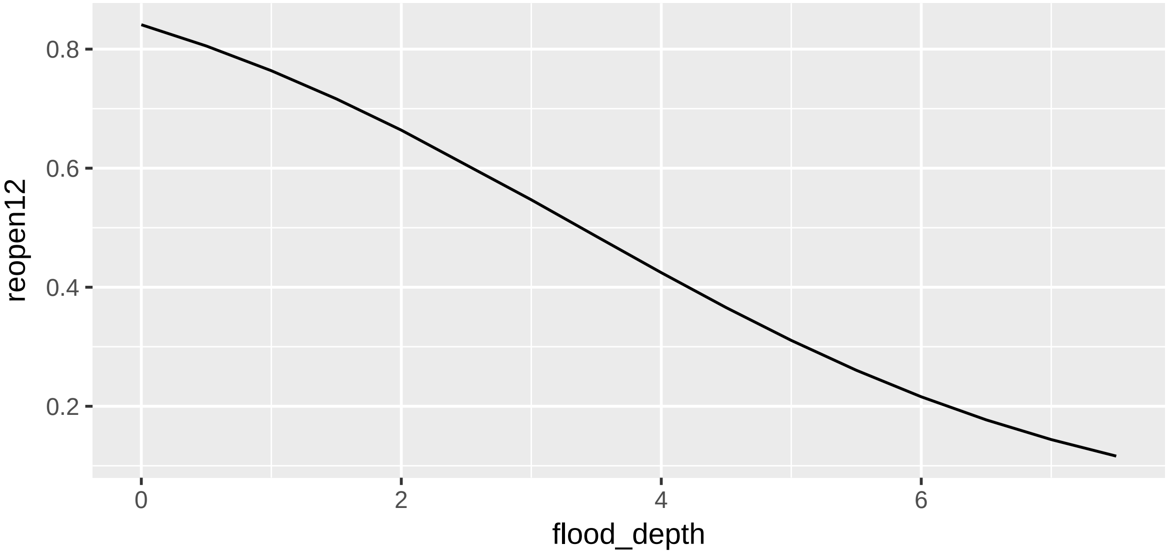
The model was used to calculate predicted probabilities that `reopen12 = 1` for a set of a set of representative data points. This was done using a function called `evaluate_model`, which is a convenience function that forms the linear predictor, and passes the linear predictor through the link function, for a representative set of data points.

```
evaluate_model(glm1, type='response')
```

```
##   flood_depth model_output
## 1           0        0.841
## 2           5        0.311
```

The column called `model_output` represents the predicted probability $P(y_i = 1 \mid x_i)$ for the specified combination of x variables.

Finally, the predicted probability that `reopen12 = 1` was plotted as a function of flood depth:



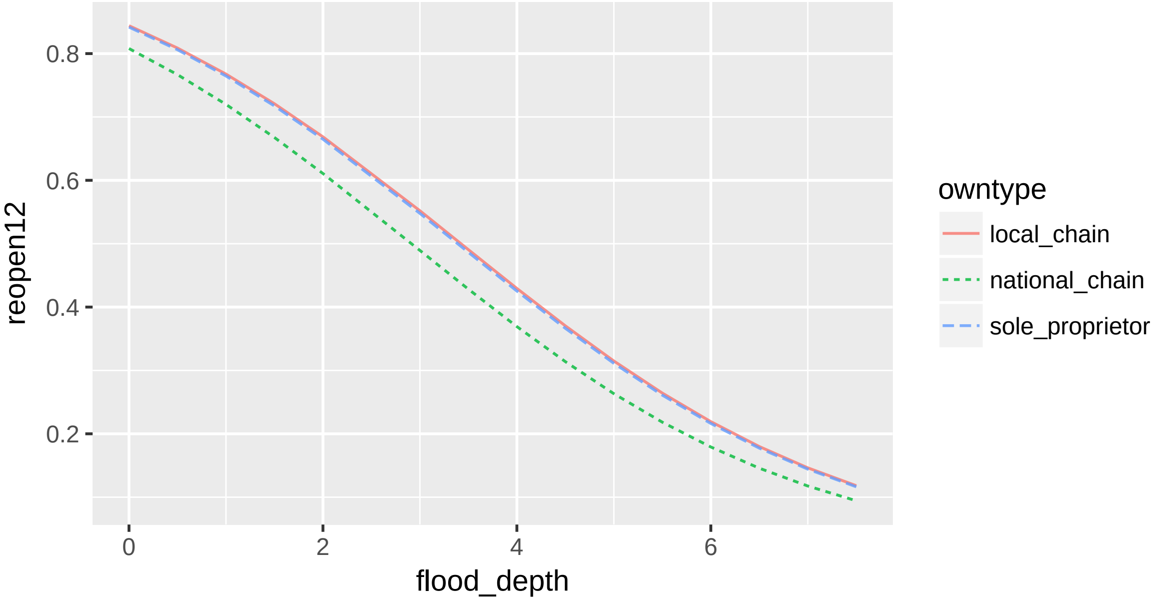
Model 2

A second model used both `flood_depth` and `owntype` as a predictor of re-opening status. This model was bootstrapped to get 95% confidence intervals for model parameters.

```
glm2 = glm(reopen12 ~ flood_depth + owntype,
           data=nola, family=binomial)
boot2 = do(1000)*{
  glm(reopen12 ~ flood_depth + owntype,
      data=resample(nola), family=binomial)
}
confint(boot2)
```

```
##                        name  lower  upper level     method estimate
## 1                 Intercept  1.164  2.240  0.95 percentile   1.6884
## 2               flood_depth -0.607 -0.420  0.95 percentile  -0.4932
## 3   owntypenational_chain -1.257  1.173  0.95 percentile  -0.2513
## 4 owntypesole_proprietor -0.559  0.545  0.95 percentile  -0.0169
```

The predicted probability that `reopen12 = 1` under Model 2 was plotted as a function of flood depth and ownership type:



Finally, Model 2 was compared against Model 1 using a permutation test. The test statistic used for the test was the sample correlation between y_i, the actual outcome, and $\hat{y}_i$, the predicted probability that $y_i = 1$ under the model. Higher values of $\text{cor}(y_i, \hat{y}_i)$, like higher values of R^2, indicate better model fit.

The actual correlation between y_i and $\hat{y}_i$ under Model 2 was calculated as follows:

```
# Actual correlation
yhat2 = fitted(glm2)
cor(nola$reopen12, yhat2)
```

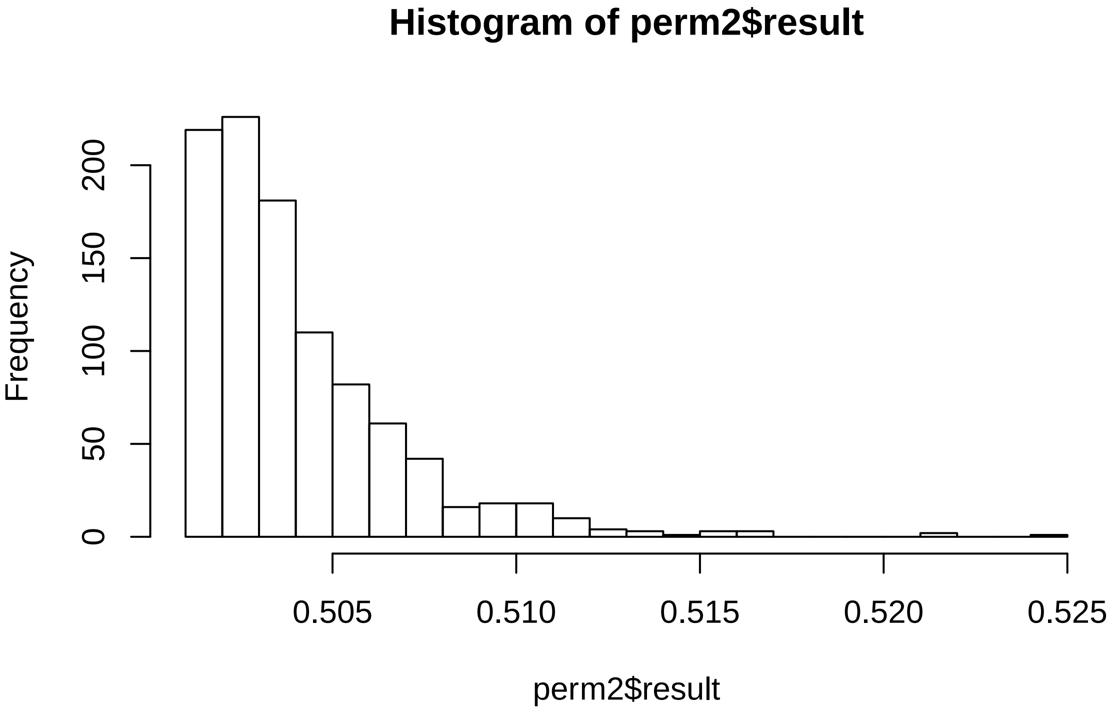
```
## [1] 0.502
```

The permutation test was run using the following code:

```
# Permutation test
perm2 = do(1000)*{
  glm_perm = glm(reopen12 ~ flood_depth + shuffle(owntype),
                 data=nola, family=binomial)
  yhat_perm = fitted(glm_perm)  # fitted values after shuffling owntype
  cor(nola$reopen12, yhat_perm)
}
```

The result of the permutation test is shown in the following histogram:

```
hist(perm2$result, breaks=20)
```



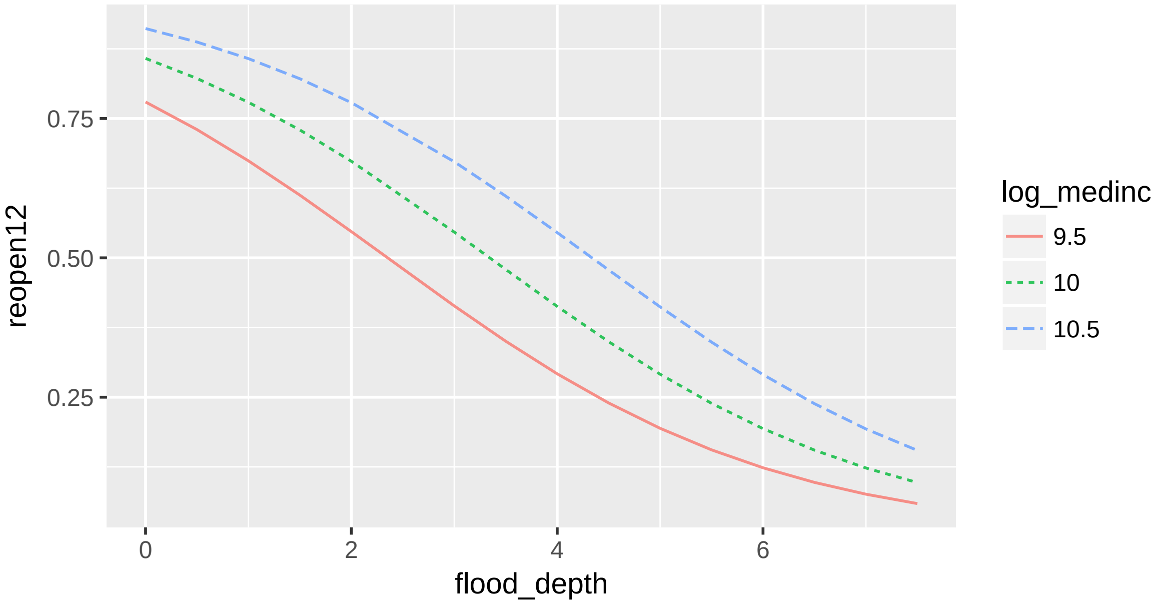
Model 3

A third model was fit, with flood depth, street, size, and log median income as predictors. 95% confidence intervals were constructed via bootstrapping.

```
glm3 = glm(reopen12 ~ flood_depth + street + size + log_medinc,
           data=nola, family=binomial)
boot3 = do(1000)*{
  glm(reopen12 ~ flood_depth + street + size + log_medinc,
      data=resample(nola), family=binomial)
}
confint(boot3)
```

```
##                name    lower  upper level     method estimate
## 1         Intercept -21.1999  1.260  0.95 percentile   -9.237
## 2       flood_depth  -0.8491 -0.351  0.95 percentile   -0.537
## 3 streetSCarrollton  -0.6478  1.947  0.95 percentile    0.370
## 4    streetStClaude  -2.1749 -0.853  0.95 percentile   -1.488
## 5        sizemedium  -0.1573  1.607  0.95 percentile    0.640
## 6         sizesmall  -0.4241  1.301  0.95 percentile    0.356
## 7        log_medinc   0.0475  2.247  0.95 percentile    1.068
```

The predicted probability of opening was plotted as a function of flood depth, for three different levels of the log median income variable:



Model 4

A fourth model was fit incorporating all the predictors from Model 3, as well as an interaction term between store size and flood depth. 95% confidence intervals were calculated by bootstrapping.

```
glm4 = glm(reopen12 ~ flood_depth + street + size + log_medinc +
  flood_depth:size, data=nola, family=binomial)
boot4 = do(1000)*{
  glm(reopen12 ~ flood_depth + street + size + log_medinc +
    flood_depth:size, data=resample(nola), family=binomial)
}
confint(boot4)
```

```
##                      name  lower  upper level     method estimate
## 1               Intercept -7.633 11.655  0.95 percentile    5.357
## 2             flood_depth -1.224 -0.197  0.95 percentile   -0.288
## 3  owntypenational_chain -1.235  1.345  0.95 percentile    1.245
## 4 owntypesole_proprietor -0.602  0.661  0.95 percentile    0.695
## 5              sizemedium -2.919  1.256  0.95 percentile   -0.568
## 6               sizesmall -3.085  1.178  0.95 percentile   -0.569
## 7              log_medinc -0.870  0.909  0.95 percentile   -0.375
## 8 flood_depth.sizemedium -0.328  0.762  0.95 percentile   -0.278
## 9  flood_depth.sizesmall -0.336  0.777  0.95 percentile   -0.196
```

Model 4 was then compared against Model 3 using a permutation test. As before, the test statistic used for the test was the sample correlation between y_i, the actual outcome, and $\hat{y}_i$, the predicted probability that $y_i = 1$ under the model.

The actual correlation between y_i and $\hat{y}_i$ under Model 4 was calculated as follows:

```
# Actual correlation
yhat4 = fitted(glm4)
cor(nola$reopen12, yhat4)
```

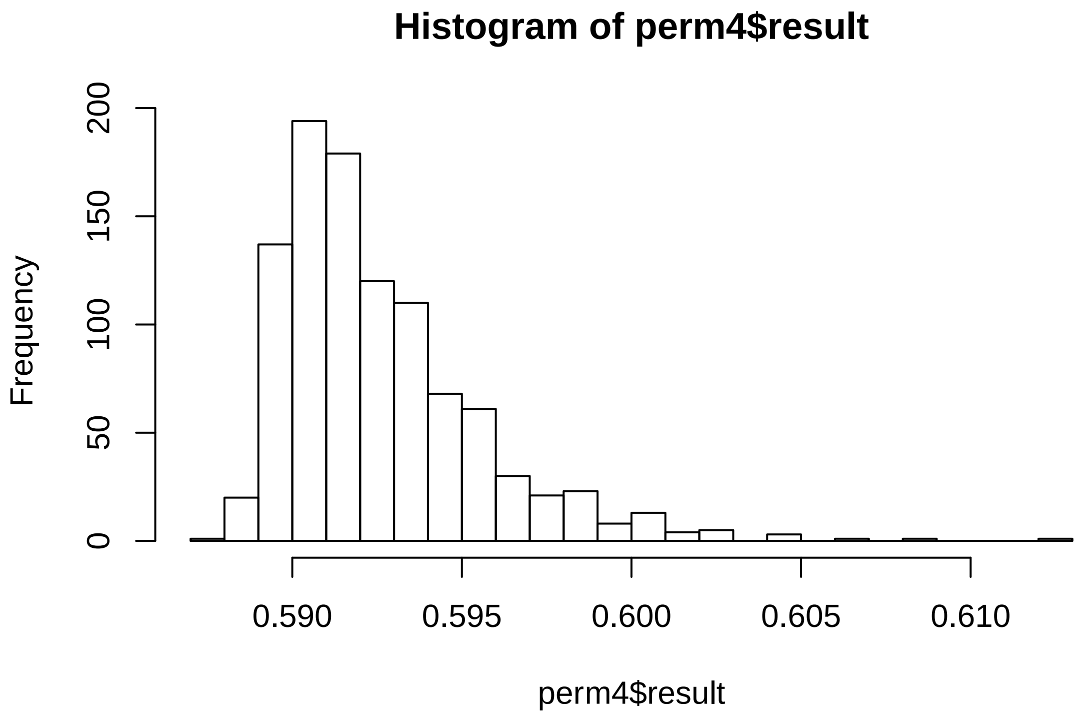
```
## [1] 0.591
```

The permutation test was run using the following code:

```
# Permutation test
perm4 = do(1000)*{
  glm_perm = glm(reopen12 ~ flood_depth + street + size + log_medinc +
            shuffle(flood_depth):shuffle(size), data=nola, family=binomial)
  yhat_perm = fitted(glm_perm)  # fitted values after shuffling owntype
  cor(nola$reopen12, yhat_perm)
}
```

The result of the permutation test is shown in the following histogram:

```
hist(perm4$result, breaks=20)
```



Finally, Model 4 was used to calculate predicted probabilities that `reopen12 = 1` for a handful of different stores with varying sizes and flood depths. This was done using the `evaluate_model` convenience function described earlier:

```
evaluate_model(glm4, type='response')
```

```
##   flood_depth      street   size log_medinc model_output
## 1           0 SCarrollton  small         10       0.8921
## 2           4 SCarrollton  small         10       0.5235
## 3           8 SCarrollton  small         10       0.1274
## 4           0 SCarrollton medium         10       0.9228
## 5           4 SCarrollton medium         10       0.5677
## 6           8 SCarrollton medium         10       0.1260
## 7           0 SCarrollton  large         10       0.9250
## 8           4 SCarrollton  large         10       0.3371
## 9           8 SCarrollton  large         10       0.0205
```