Capital Bikeshare's Busy-ness

(Student: Vinh Luong - 442069)

Data Pre-Processing

We first read the Bikeshare data into a *biketab* data table and perform the following pre-processing steps:

1. Rename the levels of the ***season*** factor variable to "*(01) Spring*", "*(02) Summer*", "*(03) Fall*" and "*(04) Winter*"
2. Rename the levels of the ***yr*** factor variable to *2011* and *2012*
3. Rename the levels of the ***mnth*** factor variable to "*(01) Jan*", "*(02) Feb*", "*(03) Mar*", "*(04) Apr*", "*(05) May*", "*(06) Jun*", "*(07) Jul*", "*(08) Aug*", "*(09) Sep*", "*(10) Oct*", "*(11) Nov*" and "*(12) Dec*"
4. Rename the levels of the ***weekday*** factor variable to "*(01) Sun*", "*(02) Mon*", "*(03) Tue*", "*(04) Wed*", "*(05) Thu*", "*(06) Fri*" and "*(07) Sat*"
5. Rename the levels of the ***weathersit*** factor variables to "*(01) Good*", "*(02) Cloudy*", "*(03) Bad*" and "*(04) Very Bad*"

QUESTION 1: Models, Outliers and False Discovery

QUESTION 1.1:

We first consider a simple linear regression of the daily rental bike totals (*total*) on an interaction between *yr* and *mnth*:

```
daylm <- glm(total ~ yr*mnth, data=daytots)
```

This regression has an in-sample SSE (deviance) of **7.263e+08**, compared with a SST (null deviance) of 2.74e+09. Its R^2 statistic is hence **0.734876**.

QUESTION 1.2:

The mathematical formula for this simple regression is:

$$\begin{aligned} &total_t = \beta_0 + \beta_1 \cdot yr_t + \beta_2 \cdot mnth_t + \beta_3 \cdot yr_t \cdot mnth_t + \epsilon_t, \\ &\text{where: } \epsilon_t \sim \mathcal{N}(0, \sigma^2) \end{aligned}$$

where *yr* and *mnth* are factor variables indicating the year (2011 or 2012) and the month (Jan - Dec), and t is the indicator of the sample date. The β_0 coefficient represents the average rental bike total for a typical day in Jan 2011. The β_1, β_2 and β_3 coefficients measure how the average daily rental bike total varies as the month and year differs from Jan and 2011.

In estimating this model, we are maximizing the log likelihood:

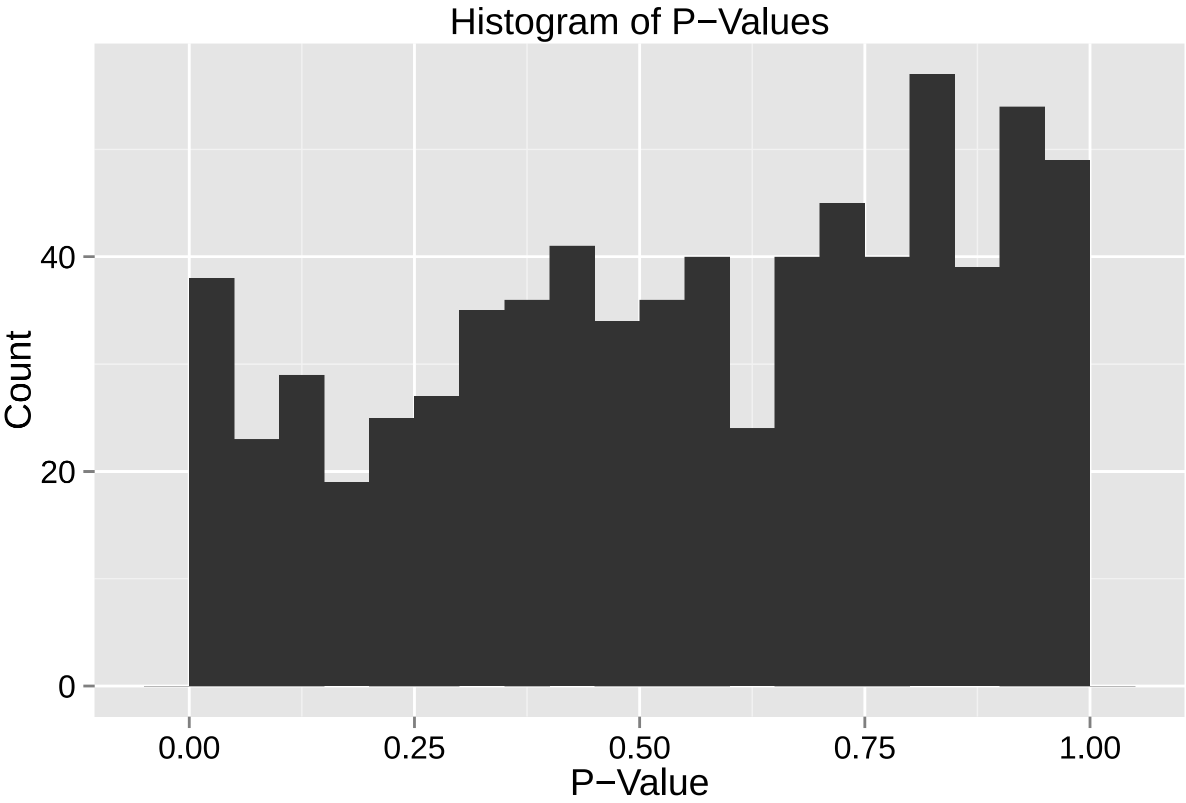
$$l = \log \prod_{t=1}^{T} \mathrm{p}(y_t|\mathbf{x}_t) = \sum_{t=1}^{T} \log \mathrm{p}_{\mathcal{N}(\mathbf{E}[y_t|\mathbf{x}_t],\sigma^2)}(y_t) = \text{constant} - \sum_{t=1}^{T} (y_t - \hat{y}_t)^2$$

where y is the dependent variable *total*, $\mathbf{x}$ captures the independent variables *yr* and *mnth*, and $\hat{y}$ indicates the predicted values $\mathbf{E}[y_t|x_t]$. This log likelihood maximization is equivalent to minimizing the sum of squared errors (SSE) $\sum_{t=1}^{T}(y_t - \hat{y}_t)^2$, which is the model's deviance in this case.

This model is likely to be too simplistic to be a good predictor of bike rental demand, as its granularity is only at the monthly level and it excludes day-to-day variables that affect demand, such as weather and holidays.

QUESTION 1.3:

We now consider the standardized residuals $r_t = (y_t - \hat{y}_t)/\hat{\sigma}$ from this model, and the corresponding outlier p-values:



The null hypothesis here for each day is **that day's actual bike rental total is indeed generated by the normal distribution $\mathcal{N}(\mathbf{E}[y_t|\mathbf{x}_t], \sigma^2)$ (which is being approximated by $\mathcal{N}(\hat{y}_t, \hat{\sigma}^2)$)**. A low p-value indicates that the probability of this being true under the null hypothesis is low, i.e. the actual bike rental total is unlikely to have been generated by the concerned normal distribution.

QUESTION 1.4:

The p-value rejection cut-off associated with a 5% False Discovery Rate here is **1.647e-06**. The days that are in this rejection region are:

```
##        dteday  weekday holiday      season predicted_total total
## 1: 2012-04-22 (01) Sun       0 (02) Summer            5807  1027
## 2: 2012-10-29 (02) Mon       0 (04) Winter            6414    22
## 3: 2012-10-30 (03) Tue       0 (04) Winter            6414  1096
```

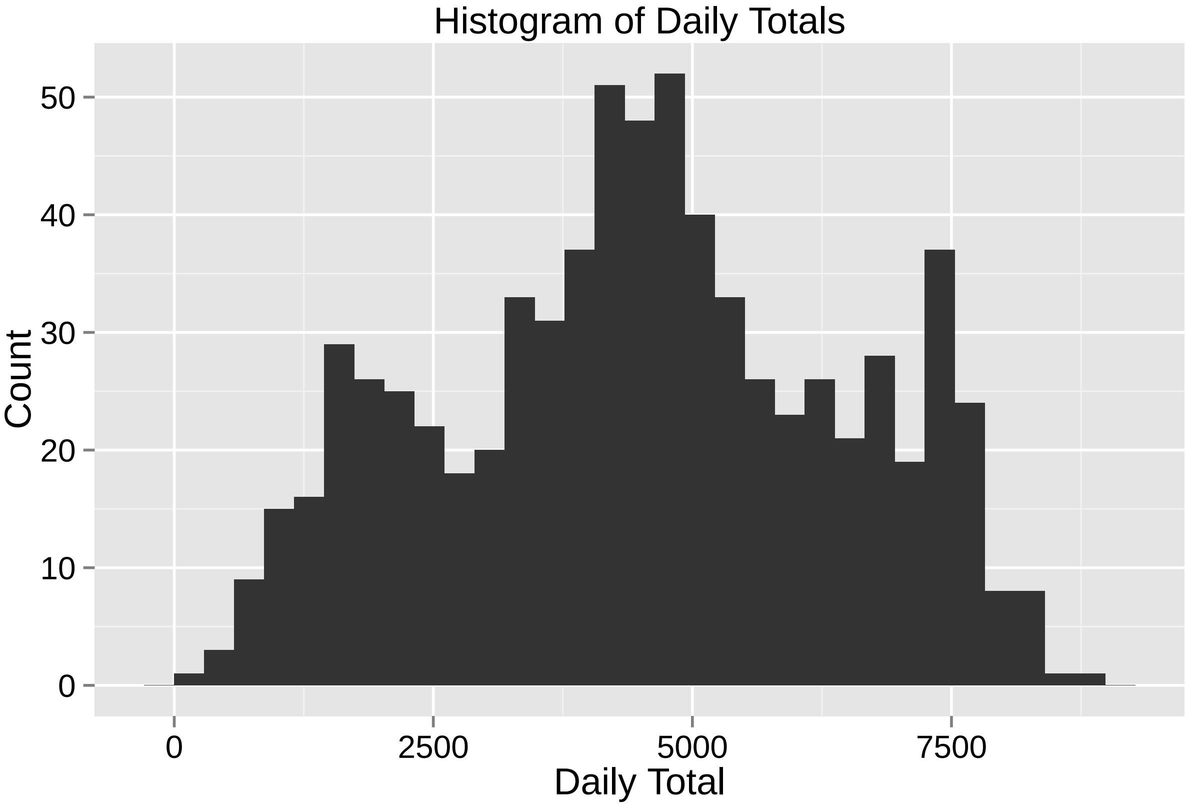
The latter two outliers correspond to the dates on which Hurricane Sandy hit the Northeastern U.S. With a hurricane capable of blowing anything in its path up to the sky, it is probably even surprising that on October 29, 2012 there were still people renting bicycles…

QUESTION 1.5:

We refer to the histogram of the outlier p-values plotted under Question 1.4.

Under the null hypothesis that the actual rental bike totals are generated by the normal distribution specified earlier, the p-values ought to be uniformly distributed over the (0, 1) interval.

The plotted histogram looks somewhat uniform, although certainly not perfect, suggesting that modelling the *total* variable on a linear scale as a normally-distributed variable is probably reasonable. This is justified by the following histogram of the *total* variable, which looks sufficiently symmetric:

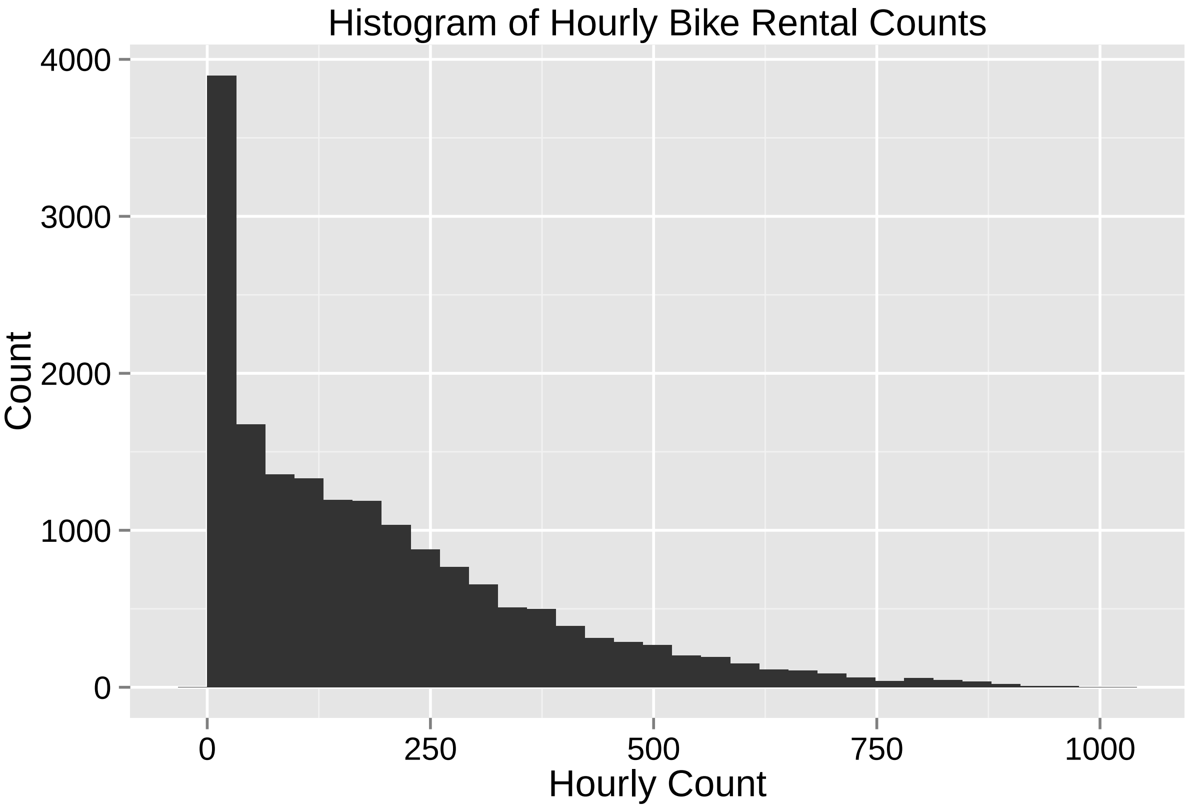


QUESTION 2: LASSO Linear Regression and Model Selection

Question 2.1:

We now consider a cross-validated LASSO regression of the log of the *cnt* variable on other variables plus interactions between *yr* and *mth* and between *hr* and *notbizday*.

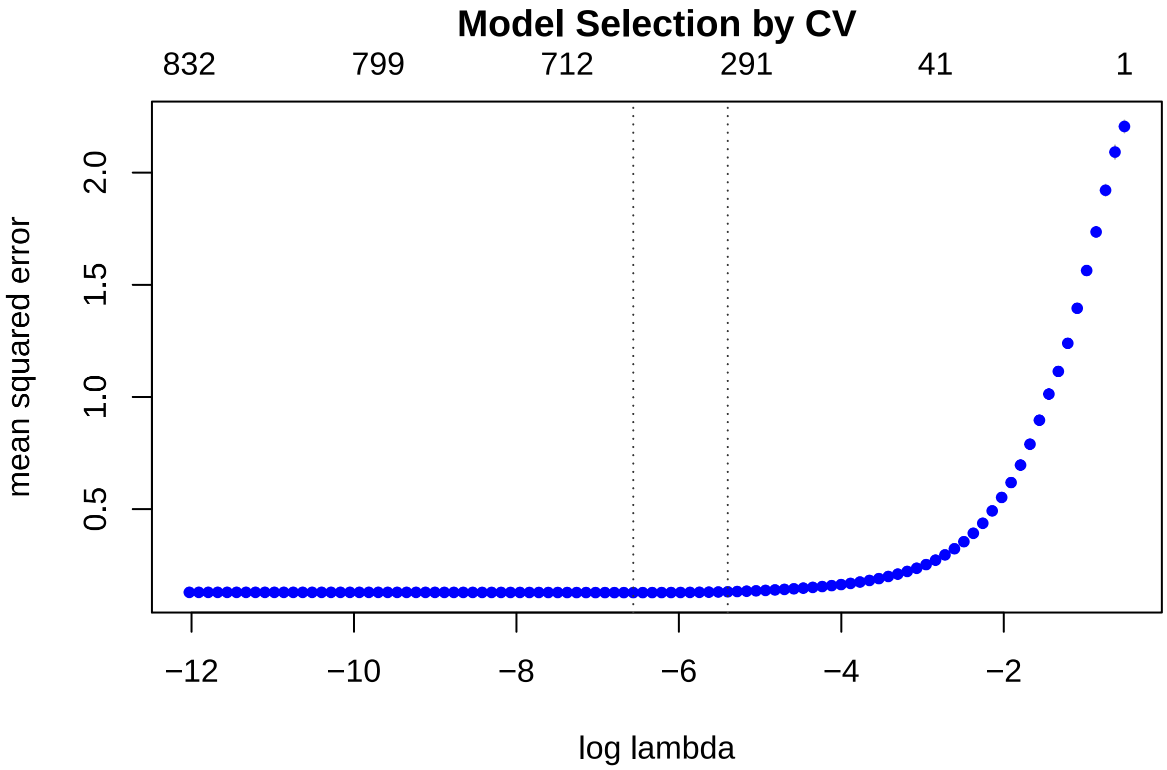
The response variable log*(cnt)* means that we are considering multiplicative changes instead of linear changes in *cnt*. This is advisable given that the *cnt* variable at the hourly granularity, unlike the daily *total* in Question 1, displays a clear exponential pattern as presented below:



The model matrix is a sparse matrix, with columns of factor variables *season*, *yr*, *mnth*, *hr*, *holiday*, *weekday*, *notbizday* and *weathersit*, plus those of the interaction terms, having values of 1 only when the concerned factor value is “on” for a certain case, and having values of 0 everywhere else.

This model addresses the outliers detected in Questions 1 by including variables indicating weather conditions and holidays, which were not considered by the previous model.

QUESTION 2.2:



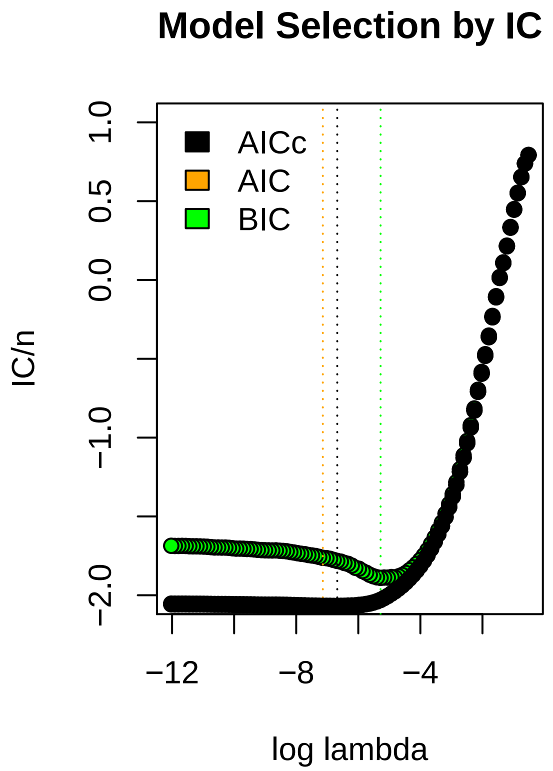
In the cross-validated model, the model selection criterion **select="min"** (corresponding to the left vertical dotted line in the above plot) chooses the LASSO regularization parameter λ that minimizes the average out-of-sample residual deviance. In this case, the estimated average out-of-sample deviance is 0.128 per case, corresponding to estimated R^2 of **0.942**.

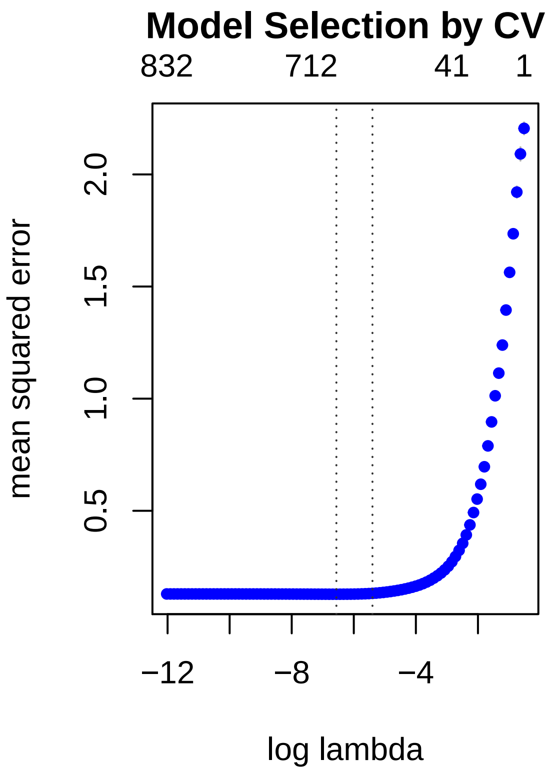
If we choose the model selection criterion **select="1se"** (corresponding to the right vertical dotted line in the above plot) instead, we'll end up with a simpler model (regularized by a larger λ) whose average out-of-sample residual deviance not more than 1 standard error away from the minimum. In this case, the estimated average out-of-sample deviance is 0.1317 per case, corresponding to estimated R^2 of **0.9404**.

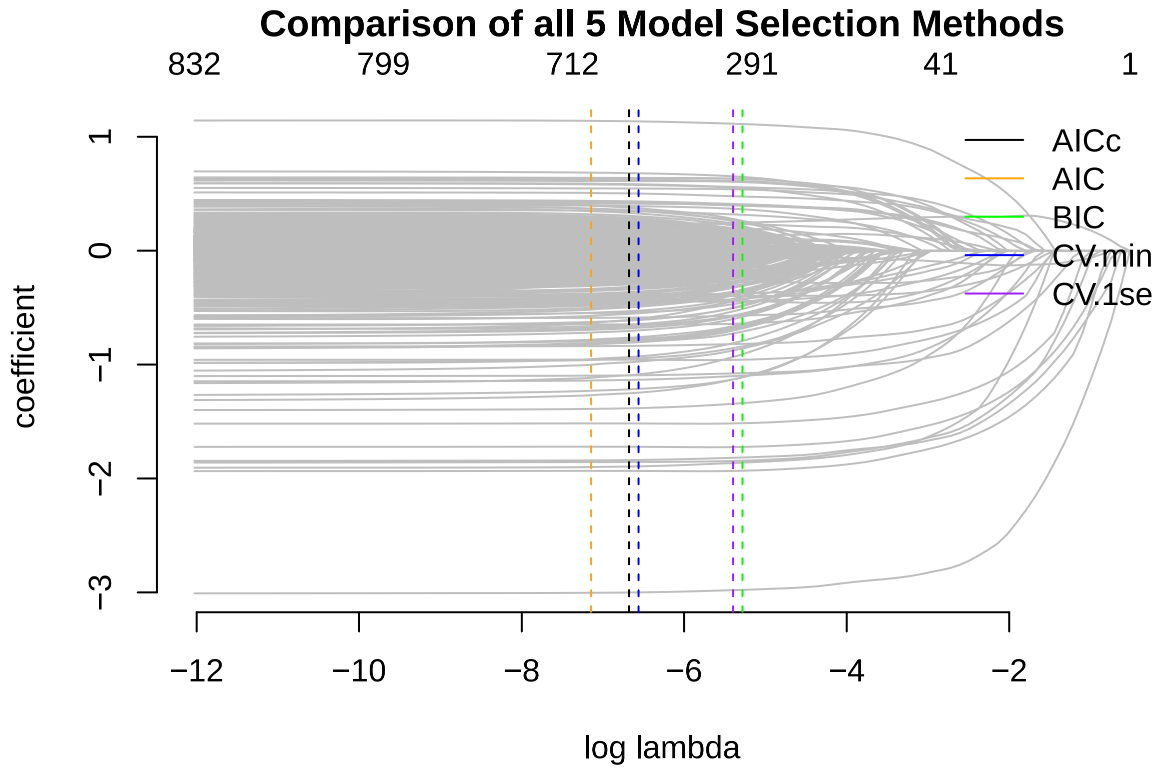
QUESTION 2.3:

The below plots and table compares model selection by information criteria AICc, AIC and BIC and cross valiation criteria "min" and "1se":

```
##                 DEVIANCE        R2      AICc       AIC       BIC
## AICc Model   0.1171119 0.9469747 -2.067532 -2.070398 -1.782283
## AIC Model    0.1165013 0.9472512 -2.067284 -2.070562 -1.762793
## BIC Model    0.1267126 0.9426277 -2.029150 -2.029813 -1.889999
## CV.min Model 0.1279921 0.9420484 -2.066847 -2.069606 -1.786851
## CV.1se Model 0.1316793 0.9403790 -2.038380 -2.039172 -1.886404
```


Model Selection by IC
AICc
AIC
BIC
IC/n
1.0
0.5
0.0
−1.0
−2.0
−12
−8
−4
log lambda


Model Selection by CV
832
712
41
1
mean squared error
2.0
1.5
1.0
0.5
−12
−8
−4
log lambda


Comparison of all 5 Model Selection Methods
832
799
712
291
41
1
AICc
AIC
BIC
CV.min
CV.1se
coefficient
1
0
−1
−2
−3
−12
−10
−8
−6
−4
−2
log lambda

QUESTION 2.4:

It seems from the table above (Question 2.3) that the five model selection criteria select models with very similar performances. I decide to go for the one selected by the AICc criterion.

Let's look at the largest three *dteday* absolute effects in the model:

```
##         date date_effect
## 1 2012-12-26   -1.252353
## 2 2011-12-25   -1.217331
## 3 2011-10-29   -1.093731
```

On this three dates, the log of the hourly bike rental totals was down by about 1.0, meaning the number of bikes rented on average per hour was only about 36.79% as many was the baseline. This is easy to explain, as two of the dates correspondeded to Christmas/Boxing Day festivities while there was a severe snow storm (the "2011 Halloween nor'easter") affecting Northeastern U.S. on Oct 29, 2011.

QUESTION 2.5:

We now bootstrap the regularization parameters λ selected by the AICc and BIC criteria.

The λ's selected by AICc have a mean of 0.001117. The λ's selected by BIC have a mean of 0.006392.

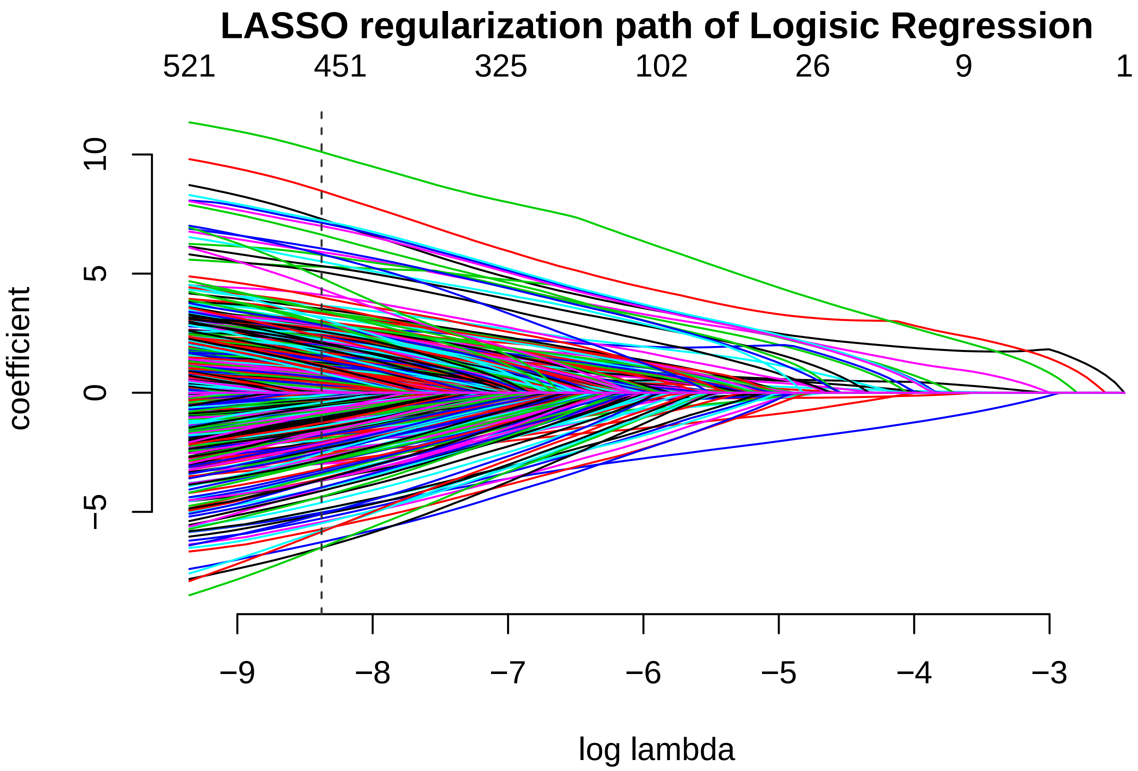
We can see clearly that BIC tends to select a larger λ because its penalization of the coefficients is stricter than that by AICc.

QUESTION 3. Logistic Regression and Classification

QUESTION 3.1:

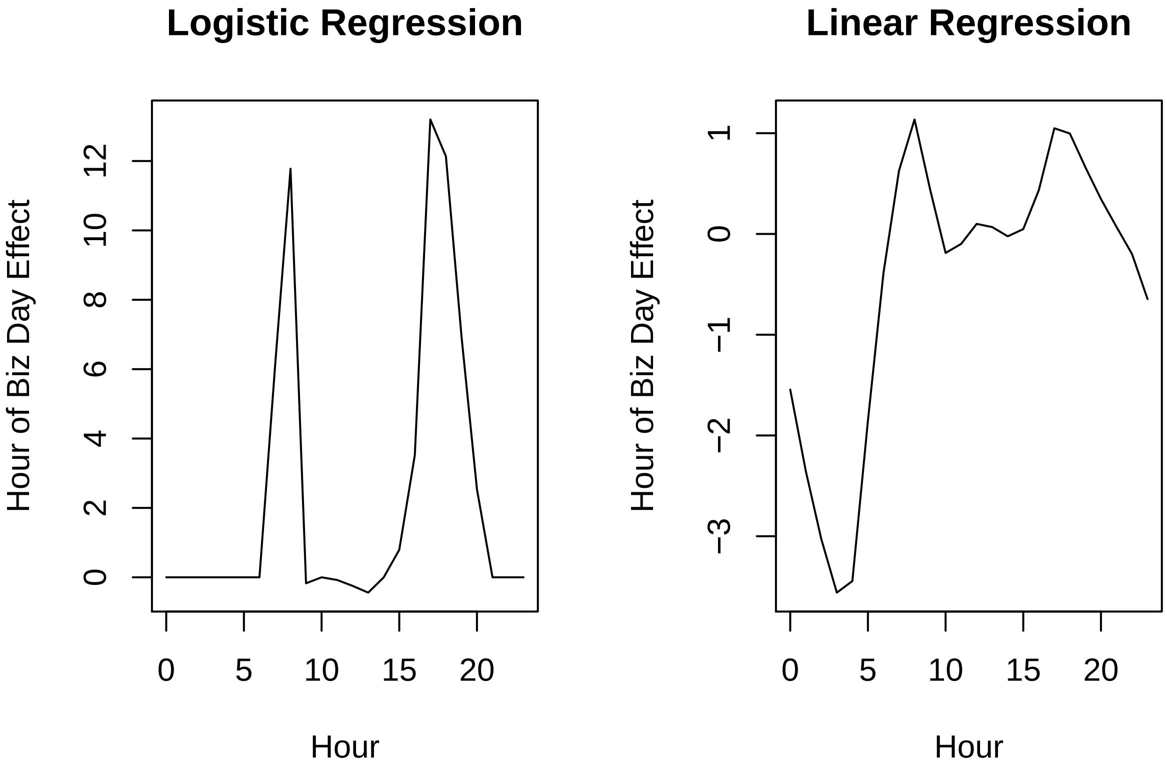
We now consider a logistic regresssion of *overload* on the same dependent variables as under Question 2.

```r
overload <- biketab$cnt > 500
fitlog <- gamlr(mmbike, overload, family = "binomial", lmr=1e-3)
plot(fitlog)
title("LASSO regularization path of Logisic Regression")
```



QUESTION 3.2:

The hour-of-day effects on *overload* on business days are as follows:



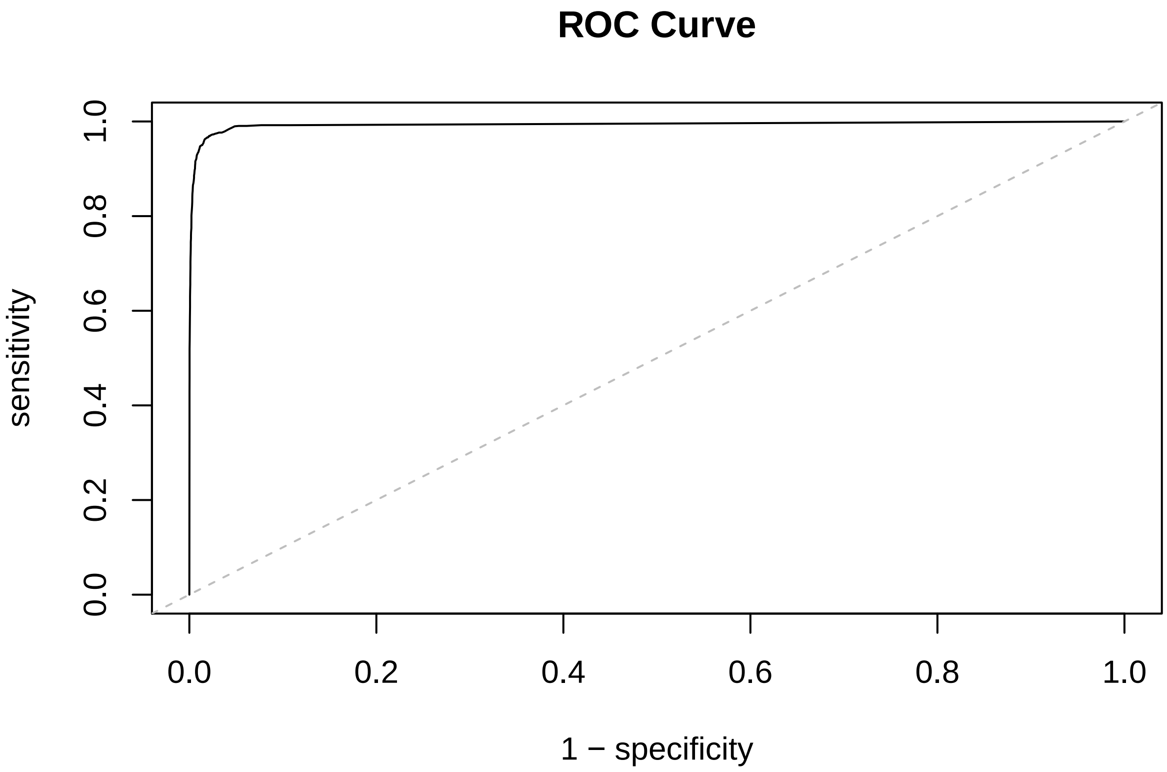
We can see two clear peaks at 8am and 5pm on business days. Take 5pm, for instance: the demand for rental bikes is about 285.4% times the baseline demand, making the odds of an overload be 5.396e+05 times such odds in the baseline.

Question 3.3:

If it costs $200/hr in overtime pay if we have an overload, and staffing an extra driver to move the bikes costs only $100/hr, then the we should staff an extra driver if the expected cost of overload exceeds the cost of staffing the driver, i.e. when probability of overload is greater than $100 / $200 = 0.5.

Question 3.4:

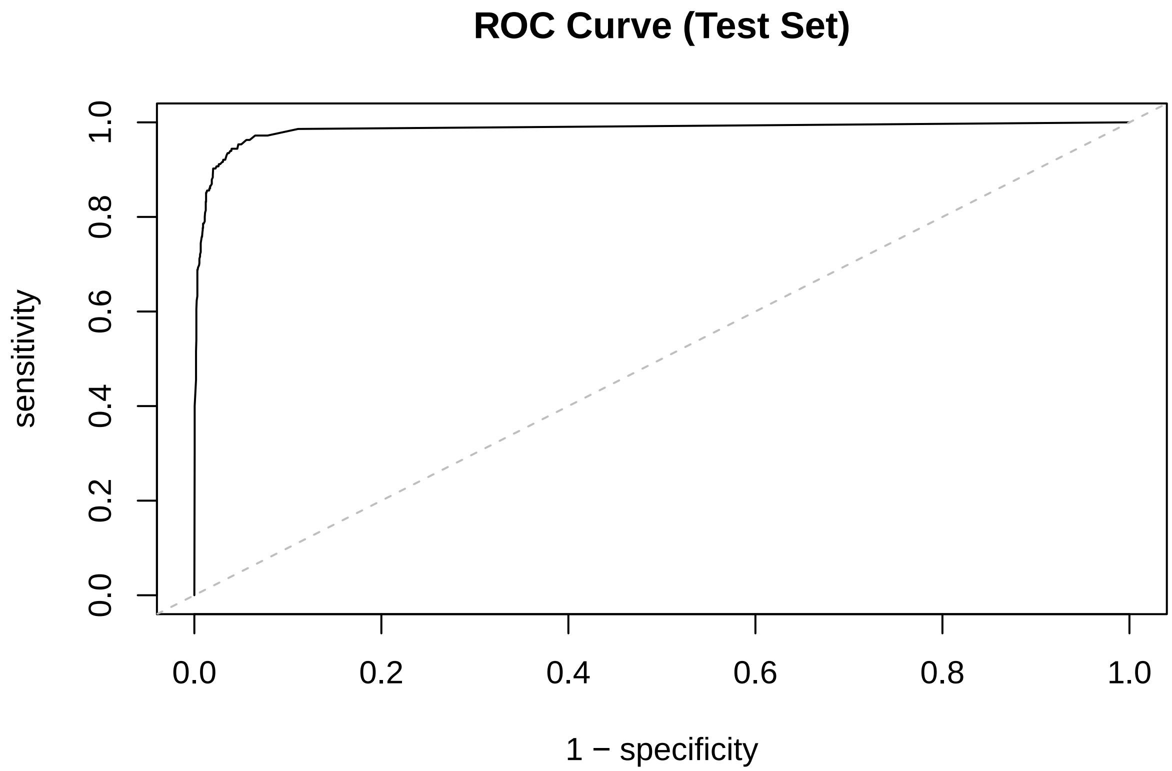
Below is the ROC curve for the AICc-selected logistic regression model:



We can see that the classifier for *overload* seems to be a very good one, with very little trade-off between sensitivity and specificity. At the decision threshold of 0.5, sensitivity is **89.91%** and specificity is **99.4%**.

Question 3.5:

We now refit the logisic regression on a Training set and evaluate its performance on a Test set, through a Test ROC curve:



The shape of the Test ROC is very similar to what we see in Question 3.4, implying that the model selected by AICc in this case has very good out-of-sample performance.

QUESTION 4. Treatment Effects Estimation

QUESTION 4.1:

Based on the “naive” model in Question 1, one standard deviation increase in hudimity decreases the log of the hourly demand by **-0.050673**, i.e. make the demand be 95.06% as much as in the baseline.

QUESTION 4.2:

We now try to estimate the independent effect of humidity. First, we fit a LASSO regression of humidity on other independent variables:

```
x <- mmbike[, -grep("hum",colnames(mmbike))]
hum <- mmbike[, "hum"] # pull humidity out as a separate vector
hum_reg <- gamlr(x, hum, lmr = 1e-4)
pred_hum <- as.vector(predict(hum_reg, newdata = x))
```

This regression has a average deviance of 2939 and R^2 of 0.8309. This suggests a lot of variation in the *hum* variable is explained by other independent variables. This is relevant because we’ll need to isolate the effect of the part of *hum* not captured by the other variables.

QUESTION 4.3:

We now fit a LASSO regression including the fitted values for *hum* from the above regression, with no penalization on the coefficient of these fitted values:

```
hum_reg_for_treatment_effect <- gamlr(cBind(pred_hum, mmbike), y,
                                      free = 1, lmr = 1e-4)
```

The effect of *hum* independent from other variables is now estimated to be -0.04819, slightly different from the coefficient in the "naive" model.

QUESTION 4.4:

We now extend the model to include the interaction between humidity and temperature:

```
mmbike <- sparse.model.matrix(
  cnt ~ . + yr*mnth + hr*notbizday + hum*temp,
  data=naref(biketab))[,-1]
fitlin <- gamlr(mmbike, y, lmr=1e-5)
```

The effect of one standard deviation increase *hum* on the log of rental bike demand is now -0.04956 + 0.04888 *temp*. This effect is a positive linear function of temperature, which means, the hotter it is, the more positive / less negative the effect of humidity on rental bike demand becomes.

QUESTION 4.5:

We now try to isolate the independent effect of humidity from the model in Question 4.4:

```
x <- mmbike[, -grep("hum",colnames(mmbike))]
hum <- mmbike[, "hum"] # pull humidity out as a separate vector
hum_reg <- gamlr(x, hum, lmr = 1e-4)
pred_hum <- as.vector(predict(hum_reg, newdata = x))
pred_hum_times_temp <- pred_hum * biketab$temp
hum_reg_for_treatment_effect <- gamlr(cBind(pred_hum, pred_hum_times_temp, mmbike), y,
                                      free = 1 : 2, lmr = 1e-4)
```

The independent effect of one standard deviation increase *hum* on the log of rental bike demand is -0.06067 + 0 *temp*, which is now not dependent on *temp*.