Jaywalking Probability Estimation at the Crosswalk between USC Campus and Exposition Park/USC Metro Station

EE517 Project Report

Nikolaos Flemotomos
Karan Singla

September 25, 2018

Abstract

We create a model to estimate the probability that a pedestrian will not wait for the green light to cross Exposition Blvd in order to go from USC to the Exposition Park/USC Metro Station or from the Exposition Park/USC Metro Station to USC. We record individuals crossing the road from Wednesday, April 19th to Tuesday, April 25th 2017 and we observe various characteristics, concerning both the individuals themselves, as well as the time and day of the recording or whether a train was coming or not. We found that using only two interaction terms, and specifically an interaction between the adjusted day type (Weekend vs Weekday) and the direction of the pedestrian, together with an interaction between the adjusted day type and whether a train is coming or not, we are able to create a model that gives a relative improvement of 8.58% when compared to the by chance model.

1 Problem Description

The goal of this project is to build a model that can estimate the probability of crosswalk violation between USC campus and Exposition Park/USC Metro Station. For our case, we define crosswalk violation (or jaywalking) as the act of starting crossing the crosswalk while the traffic light for the pedestrians is red. We provide a map describing the problem in Figure 1.

2 Data Collection and Feature Extraction

We set up a digital camera inside the USC campus campus facing towards the Metro Station in such an angle that allows us to observe and keep track of the various individuals' characteristics. We do all the recordings from Wednesday, April 19th to Tuesday, April 25th 2017 from 11:00am to 12:00pm in the morning and from 16:00pm to 17:00pm in the afternoon. From all the recordings available, we randomly chose to analyze 20 minutes from each day, which gave us a total of 140 minutes of recordings with 970 samples. Out of them, we observe 423 to violate the crosswalk

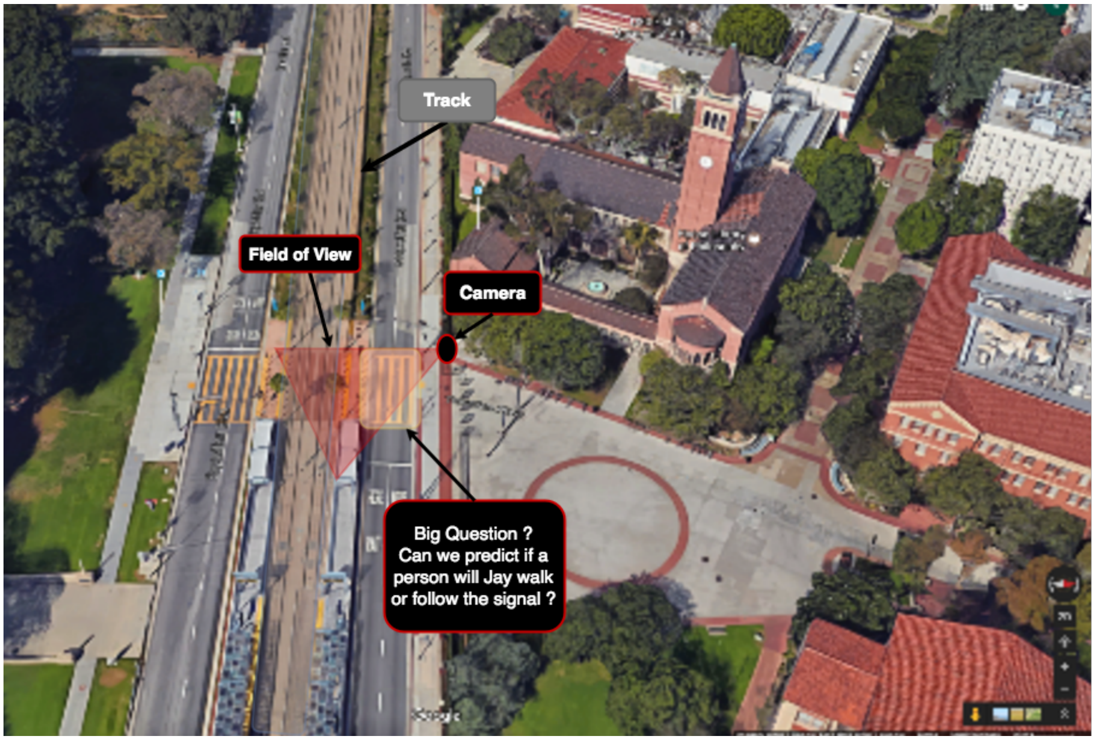
Track
Field of View
Camera
Big Question ?
Can we predict if a person will Jay walk or follow the signal ?

Figure 1: Problem Description

and the rest 547 not to. That means that the classification of a by-chance model (always selecting the majority class) is 56.4%.

For each sample we keep track of the characteristics listed in Table 1. Whenever we are not confident about the value of a specific variable, we discard the particular sample. We assign the value of the variable `cellphone` to be `Yes` whenever the individual is using a cellphone for no matter what reason. We choose to adjust the days of the week to include only two types; namely weekday or weekend. This is because during the weekend of our recordings the Festival of Books took place in USC, which led to significant differences between the individuals recorded during Saturday and Sunday and those recorded during the other days, who were primarily USC students. Finally, it is to our knowledge that we miss some potentially important features. For example, our camera could not capture whether or not a vehicle was approaching the crosswalk, since this was out of Field Of View. Typical examples of the samples we recorded and annotated are given in Figure 2.

3 Logistic Regression

Logistic regression is often used in order to analyze the impact of multiple independent variables presented simultaneously to predict membership of one or other of the two categories of a dichotomous dependent variable [BB08]. Formally, we consider the dependent variable Y to be

Feature Name	**Possible Values**
direction	Towards_Campus, Away_from_Campus
train_coming	Yes, No
day	Weekday, Weekend
backpack	Yes, No
having_kids	Yes, No
age*	Below_30, Above_30
gender	Male, Female
time_of_the_day	Morning, Afternoon
race*	Asian, Black, Hispanic/Latino, White
earphones	Yes, No
dressing	Business, Casual
cellphone	Yes, No
hat	Yes, No
sunglasses	Yes, No

* Even though we initially include those features, we decide not to use them at all for our final analysis, because we are not confident about a lot of samples. At any case, they prove to be non-significant, and thus do not affect the results.

Table 1: Features we use for the data we have collected. We list the features in descending order of significance. We order the variables based on the step they were removed by a backward selection modeling (Wald) or on their p-value after the backward selection (if not removed) (see Section 4)

Bernoulli distributed, given the independent variables:

$$Y|\mathbf{X} \sim Bernoulli(p) \Rightarrow E\{Y|\mathbf{X}\} = p \tag{1}$$

What we want is to model the unknown parameter p. The approach in logistic regression is to assume a relationship between the dependent and independent variables of the form

$$logit(p) = \ln \frac{p}{1-p} = \mathbf{X}^T\beta \Leftrightarrow p = \frac{1}{1+e^{-\mathbf{X}^T\beta}} \tag{2}$$

The β coefficients can then be estimated using Maximum Likelihood Estimation (MLE) [Sko15]:

$$\hat{\beta}^{ML} = \arg\max L = \arg\max \sum_{i=1}^{n} \{y_i \ln p_i + (1-y_i)\ln(1-p_i)\} \tag{3}$$

where n the total number of samples. However, there is no closed form solution and we have to use some numerical software for the iterative computation. In our case, we use SPSS.

Of course, as with any regression model, we have to check about the significance of the model we are constructing and its goodness-of-fit. Additionally, we have to worry about possible multicollinearity issues and the potential existence of outliers. We deal with those issues through a variety of statistical tests and widely-used procedures which we will gradually explain in the following section.

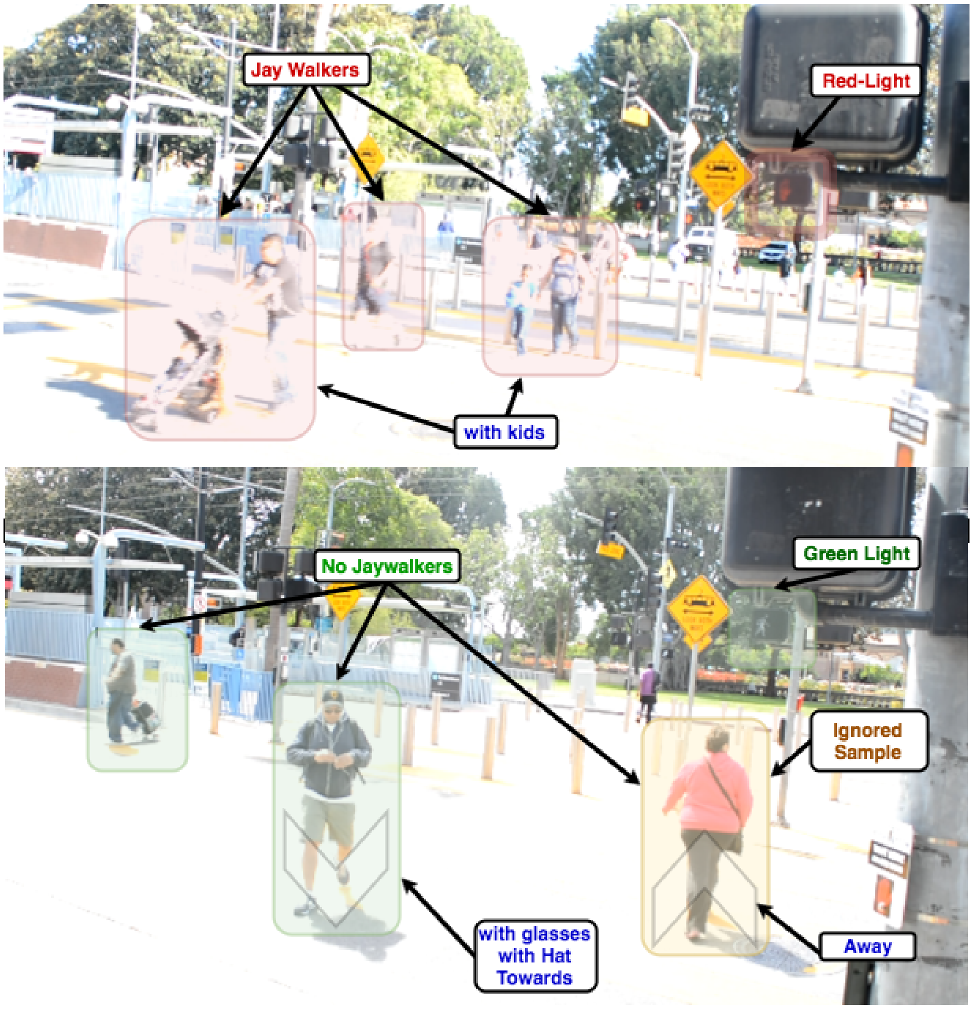

Figure 2: Typical samples from our recordings.

4 Analysis and Results

From all the available features we collect, the first step was to choose which of them are significant to be included as independent variables in our model. To that end, we used both a Forward and a Backward selection scheme, based on the Wald statistic. Those two schemes yielded the exact same final results. Specifically, the `direction`, the `train_coming` and the `day` (given in descending order of significance) were the three independent variables chosen to be included in the model, in the sense that the null hypothesis that the corresponding coefficient is 0 can be rejected at the 0.05 level.

Using those three independent variables, the model is described by the equation

$$\ln(\text{odds}) = \tilde{\beta}_0 + \tilde{\beta}_1 \text{direction} + \tilde{\beta}_2 \text{train\_coming} + \tilde{\beta}_3 \text{day} \tag{4}$$

To check the overall model adequacy, we used the Omnibus Test, the Cox & Snell and the

Nagelkerke versions of pseudo-R^2 measure, and the Hosmer & Lemeshow goodness-of-fit test, with the results listed in Table 2.

Test Name	$\mathcal{X}^2$ **Test Statistic**	p**-value**
Omnibus	61.382	0.000
Hosmer & Lemeshow	11.553	0.009

pseudo-R^2	**value**
Cox & Snell	0.061
Nagelkerke	0.082

Table 2: Statistical tests for the overall adequacy of the model described by equation (4).

The very low p-value of the Omnibus test suggests that we can reject the null hypothesis that all the coefficients of the model are 0. However, the low values of the pseudo-R^2 measures suggest that a great amount of the observed variability cannot be explained by our model, while the low p-value of the Hosmer & Lemeshow test gives evidence of poor fit. These results pushed us towards trying to find a better model for our data.

We tried to do so by including the interaction terms of the independent variables. In particular, we included all the second order interaction terms between the 3 variables we already had after the forward (and backward) selection scheme. We chose not to include all the interaction terms from the very beginning, because that would lead to too many features, compared to the number of samples available. Having this new set of 6 variables (3 + 3 interaction temrs), we ran once again a forward and a backward selection scheme to choose the significant ones for our final model. This time, the two procedures yielded different results, with the corresponding correlation matrices given in Table 3.

	constant	**direction×day**	**train_coming×day**
constant	1.000	-0.549	-0.050
direction×day		1.000	-0.321
train_coming×day			1.000

(a) Forward Selection.

	constant	**train_coming**	**direction×day**
constant	1.000	-0.938	-0.441
train_coming		1.000	0.255
direction×day			1.000

(b) Backward Selection.

Table 3: Correlation matrices for the independent variables selected after a forward (a) and a backward (b) selection scheme, using as the initial set of variables the three included in equation (4) plus the interaction terms between them.

Because of obvious multicollinearity issues with the variables selected by the backward selection scheme, we chose to keep the variables selected by the forward selection scheme. We should note, however, that it is not guaranteed that the specific pair of variables is the optimal one in terms either of best classification accuracy results or least multicollinearity [Kos17].

Our final model in now described by the equation

$$\ln(\text{odds}) = \beta_0 + \beta_1 \text{direction} \times \text{day} + \beta_2 \text{train\_coming} \times \text{day} \tag{5}$$

where

$$\beta_0 = 0.147 \quad \beta_1 = -1.238 \quad \beta_2 = 0.935$$

In Table 4 we list the results of the statistical tests described earlier for the overall adequacy of our final model.

Test Name	$\mathcal{X}^2$ Test Statistic	p-value
Omnibus	76.928	0.000
Hosmer & Lemeshow	0.737	0.391

pseudo-R^2	value
Cox & Snell	0.076
Nagelkerke	0.102

Table 4: Statistical tests for the overall adequacy of the model described by equation (5).

As we can see, all the test statistics have been improved, when compared to the corresponding statistics in Table 2. Specifically, the Omnibus test statistic has been increased, which means we are even more confident we can reject the null hypothesis that all the coefficients are 0. Hosmer & Lemeshow test statistic has been significantly reduced and particularly the null hypothesis stating that there is not a difference between the values predicted by the model and the observed values cannot be rejected at any significance level above the 0.391 level. The low pseudo-R^2 values suggest that much of the observed variability cannot be explained by our model, but are still higher compared to out initial model (Table 2).

After testing the overall model adequacy, we wanted to check each independent variable's significance. As we have already said, we chose to use the Wald statistic for both the forward and the backward selection. In Table 5 we list the values of the Wald statistic, as well as the corresponding Standard Errors, the p-values, and the 95% confidence intervals for e^{β_i}. As we can see, the p-values for both the interaction terms are very small (0.000 using SPSS's numerical accuracy), but for the constant term the p-value is 0.084, which means it is not significant at the 0.05 level, but it is significant at the 0.1 level. We chose to include that constant term in our final model, because the Hosmer & Lemeshow test statistic was much smaller when we did not.

Variable	β_i	95% C.I. for e^{β_i}	St. Error	Wald Statistic	p-value
constant	0.147	–	0.084	3.044	0.081
direction×day	-1.238	[0.217, 0.388]	0.149	69.445	0.000
train_coming×day	0.935	[1.529, 4.241]	0.260	12.900	0.000

Table 5: Statistical tests for the overall adequacy of the model described by equation (5).

The interpretation of the β_i coefficients is more straightforward in terms of the exponentials e^{β_i}. If e^{β_i} is less than one, any increase in the predictor leads to a drop in the odds of the outcome occurring. If it exceeds 1, then the odds of an outcome occurring increase. For example, according to our encoding, the interaction term `direction`×`day` is 1 when an individual goes away from the campus during weekend and 0 otherwise. So, an individual heading away from the campus during weekend is $e^{-0.128} \approx 0.290$ times as likely to jaywalk compared to an individual heading towards campus or recorded during a weekday. Since the entire C.I. for e^{β_1} is below 1, we can conclude a negative association between the `direction`×`day` and the outcome at the 0.05 level [Win05]. Similarly, since the entire C.I. for e^{β_2} is above 1, we can conclude a positive association between the `train_coming`×`day` and the outcome.

The correlation matrix we provided in Table 3a suggests there might be multicollinearity issues with our model, but the Standard Errors are far below 2, which is a good sign against serious multicollinearity problems. To further analyze the probable negative effects of multicollinearity, we checked the eigenvalues of the matrix $\mathbf{X}^T\mathbf{X}$ (where $\mathbf{X}$ is the data matrix), as well as the Variance Inflation Factors (VIFs). We list those measures in Table 6.

Eigenvalues
1.876
0.771
0.352

	VIFs
constant	–
direction×day	1.069
train_coming×day	1.069

Table 6: Multicollinearity diagnostics for our final model.

Based on those results we can claim that the multicollinearity, even if it exists, does not impose serious problems, since the value of the smallest eigenvalue is not unreasonably low, while the VIFs are well below 5 [Sko15]. We should note here that the computation of VIFs is based on the coefficients of determination of linear regression, but can still be an indicator of multicollinearity. Finally, the condition index is

$$\sqrt{\frac{\lambda_{max}}{\lambda_{min}}} = 2.307 < 10$$

The final step in our analysis is to test the behavior of the residuals to check whether our model fits well for all the data available or if there exist outliers. The casewise list that SPSS provides, produces a list of cases that did not fit the model well. By considering our outlying criterion to be individuals outside a 2-standard-deviations band, we have a first sign of non-existence of outliers in our data. However, we continue our analysis using the most widely used influence statistics tests; namely the Dfbetas, the leverage and the Cook's distance. We plot the histograms for all those tests in Figure 3.

Based on the results of all those measures, we can argue that there are not highly influential samples in our data. In particular, Cook's distance and leverage are much less than 1 [Sko15], while all the $\Delta\beta$'s are below the threshold of $2/\sqrt{n} = 2/\sqrt{970} \approx 0.064$ [Ret07].

After having done the above-mentioned analysis and being confident that we have created a reasonable model, we check the final classification accuracy. To do so, we applied a 5-fold cross-validation scheme with the results listed in Table 7. We can see that our model always outperforms the by-chance model, yielding to an average relative improvement of 8.85%.

While doing the cross-validation, we also wanted to check the robustness of our model. According to the results of the Table 8, our model is robust enough, since the values of the coefficients β_i do not change greatly with different folds.

5 Conclusions and Weaknesses

We conducted a logistic regression analysis to predict jaywalking at the crosswalk between USC Campus and Exposition Park/USC Metro Station, using a variety of observable characteristics as potential predictors. The Wald criterion with a forward selection scheme demonstrated that

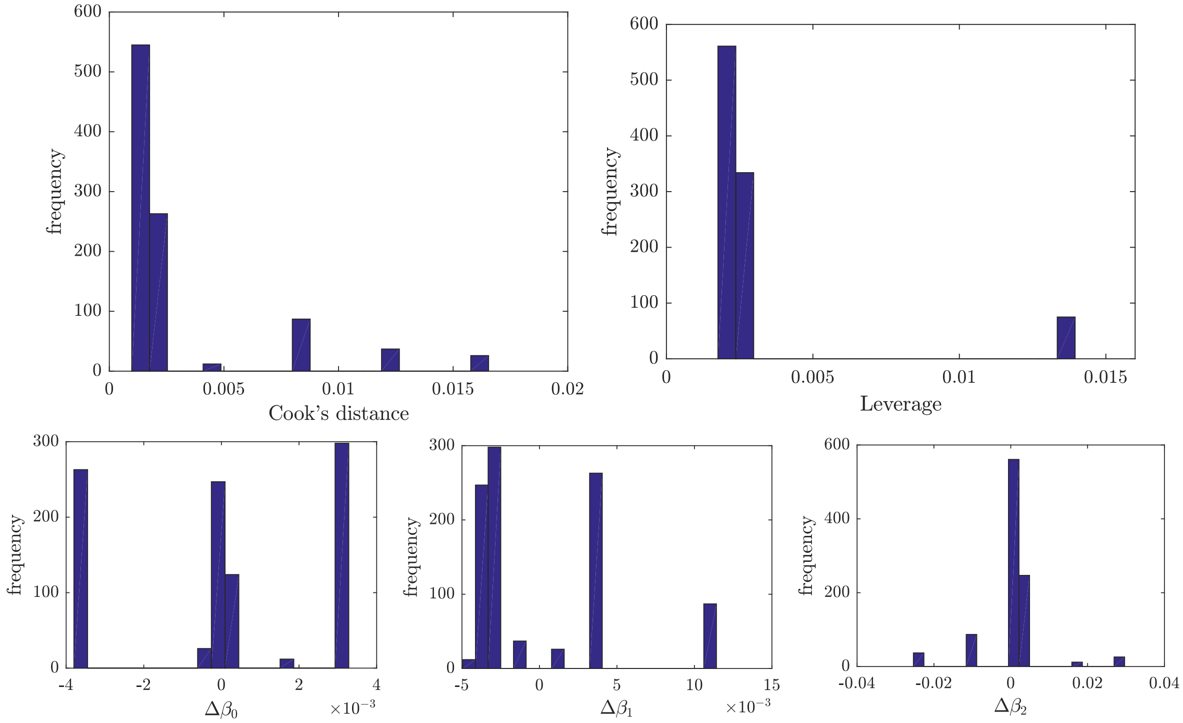

Figure 3: Influence Statistics tests.

only the interaction terms between direction and day type and between day type and whether a train was coming or not made a significant contribution to prediction at the 0.05 level. The Omnibus test indicated that our model is statistically significant against a constant-only model ($p < 0.000$), while a Hosmer & Lemeshow test indicated that our model provides a moderately good fit ($p = 0.391$). However, the low values of Cox & Snell's R^2 (0.076) and Nagelkerke's R^2 (0.102) demonstrate that our model failed to explain all the data's variability. The inspection of the correlation matrix, the eigenvalues and the VIFs indicated that the model is free of serious multicollinearity problems. An extensive residual and influential statistics analysis, including the Cook's distance, the leverage, and the Dfbetas demonstrated that our data is free of outliers, when using our model. Testing the accuracy of our model with a 5-fold cross-validation, we observed an average relative improvement of 8.58%, compared to the by-chance model (56.4% vs. 61.24%).

We are aware, however, of several weaknesses of our model. As we have already mentioned, we failed to capture all the potentially important features. For example, the level of traffic or whether a vehicle was approaching or not are parameters we did not take into consideration due to the limited field of view of the single camera we used. Additionally, we investigated only the interaction terms between the features selected by our initial forward (or backward) selection scheme. However, there is no guarantee that those are indeed the most significant or helpful in terms of classification accuracy. Higher order interaction terms could have also been proved helpful, but we did not include them to our analysis. Finally, we defined as crosswalk violation the act of beginning crossing the road while the traffic light for the pedestrians is still

Fold	**accuracy**(%)	**sensitivity**(%)	**specificity**(%)
1	58.2	0.0	100.0
2	57.2	0.0	100.0
3	55.2	0.0	100.0
4	55.2	0.0	100.0
5	56.2	0.0	100.0
average	56.4	0.0	100.0

(a) By-chance model.

Fold	**accuracy**(%)	**sensitivity**(%)	**specificity**(%)
1	60.3	69.1	54.0
2	59.3	65.1	55.0
3	63.4	77.0	52.3
4	61.9	78.2	48.6
5	61.3	76.5	49.5
average	61.24	73.18	51.88

(b) Our model.

Table 7: Accuracy, Sensitivity, and Specificity of our model, using a 5-fold cross-validation.

Fold	β_0 (**constant**)	β_1 (**direction**×**day**)	β_2 (**train_coming**×**day**)
1	0.167	-1.266	0.934
2	0.166	-1.311	0.936
3	0.118	-1.180	0.915
4	0.134	-1.209	0.943
5	0.147	-1.226	0.945

Table 8: Parameters of our final model for different folds during a 5-fold cross-validation.

red. The rest samples were assigned as not violating the crosswalk. However, we did not make a distinction between individuals that had the opportunity to jaywalk but instead chose to wait for the green light and those who did not jaywalk just because it happened for the light to be green when they reached the crosswalk.

References

[BB08] Robert Burns and Richard Burns. *Business Research Methods and Statistics using SPSS*. SAGE Publishing, 2008.

[Kos17] Bart Kosko. Lecture Notes. *EE517, USC*, 2017.

[Ret07] Karl Rethemeyer. Outliers and DFBETA. *Rockefeller University, retrieved online: http://www.albany.edu/faculty/kretheme/PAD705/SupportMat/DFBETA.pdf*, 2007.

[Sko15] Zisis Skordilis. Logistic Regression for SPSS. *TA Notes for EE517, USC*, 2015.

[Win05] Larry Winner. Chapter 15, Logistic Regression. *STA6127, University of Florida, retrieved online: www.stat.ufl.edu/ winner/sta6127/chapter15c.ppt*, 2005.