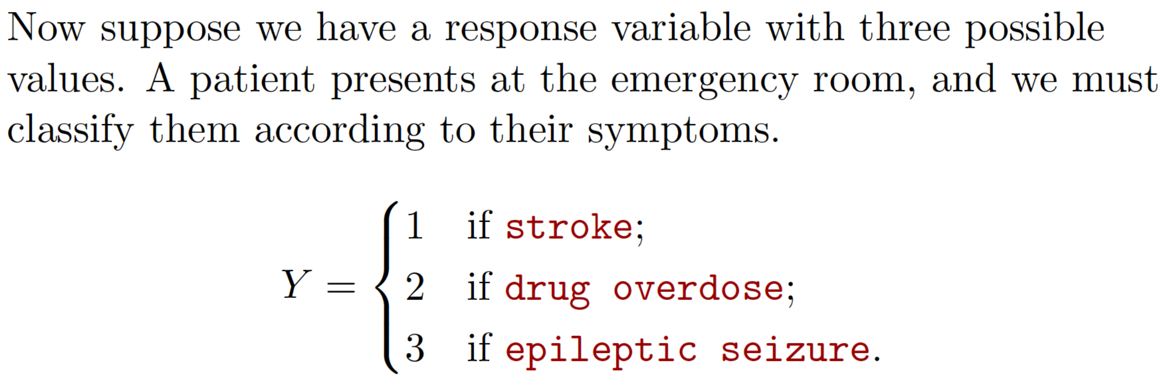
INFO 6105 – Data Sci Eng Methods
Practice Exam Solutions

Student Name: ____
Professor: Nik Bear Brown

Rules:
1. NO COMPUTER, NO PHONE, NO DISCUSSION or SHARING.
2. Ask if you don't understand a question.
3. You may have five 8½"×11" sheets of notes (you may use both sides, written or printed as small as you like).
4. Time allowed. 1 hour 30 minutes.
5. Bring pen/pencil. The midterm will be written on paper.

The exam will cover *"An Introduction to Statistical Learning with Applications in R"* chapters 1 to 5 as well as the additional regression and classification metrics that were covered in class.

Q1 (5 Points)

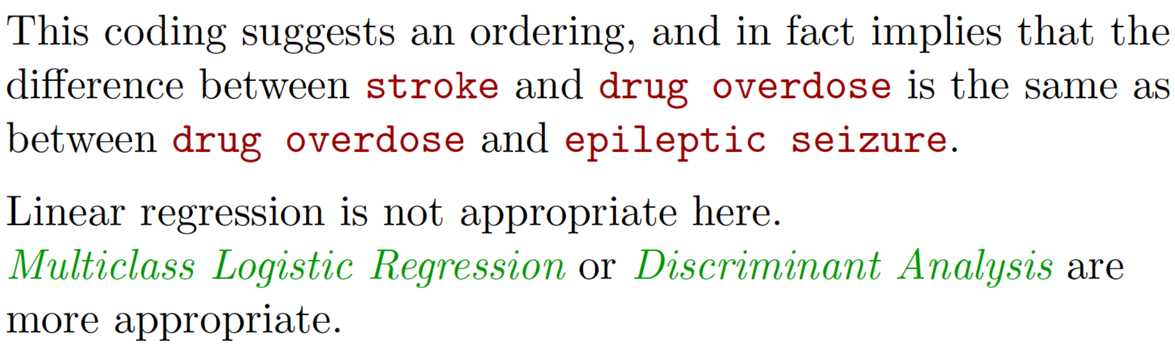


Now suppose we have a response variable with three possible values. A patient presents at the emergency room, and we must classify them according to their symptoms.

$$Y = \begin{cases} 1 & \text{if } \texttt{stroke}; \\ 2 & \text{if } \texttt{drug overdose}; \\ 3 & \text{if } \texttt{epileptic seizure}. \end{cases}$$

Is linear regression an appropriate analysis?

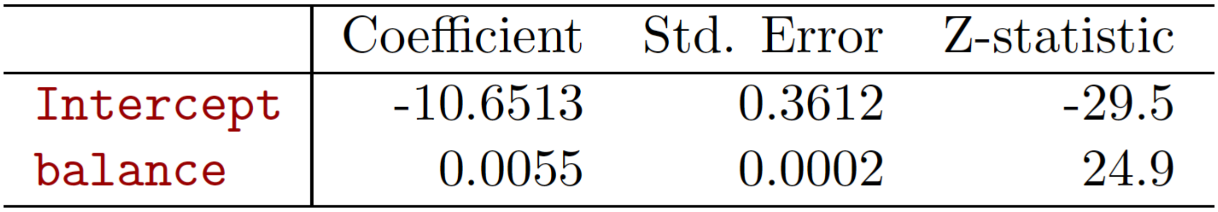
Solution: No.



This coding suggests an ordering, and in fact implies that the difference between `stroke` and `drug overdose` is the same as between `drug overdose` and `epileptic seizure`.

Linear regression is not appropriate here.
Multiclass Logistic Regression or *Discriminant Analysis* are more appropriate.

Q2 (5 Points) Assume logistic regression is being used to predict whether someone will default on a loan and an independent variable called balance is the only independent variable in the model. It returns the stats below. Is balance a significant predictor of default?



	Coefficient	Std. Error	Z-statistic
Intercept	-10.6513	0.3612	-29.5
balance	0.0055	0.0002	24.9

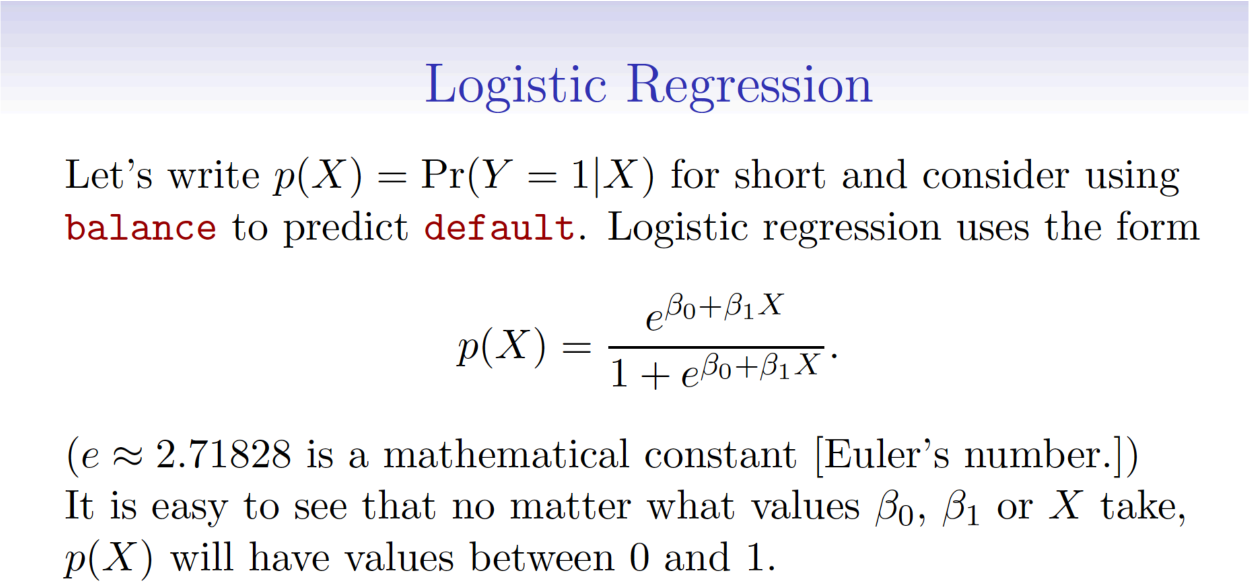
Solution:

Yes. The z-score of 24.9 for balance is very significant, about 25 standard deviations from the mean.

Q3 (5 Points) Calculate the probability of defaulting given a balance of 1000 from the stats in Q2.

Solution:

Just plug the intercept beta zero and slope beta one in to the equation below; along with an X of 1000.

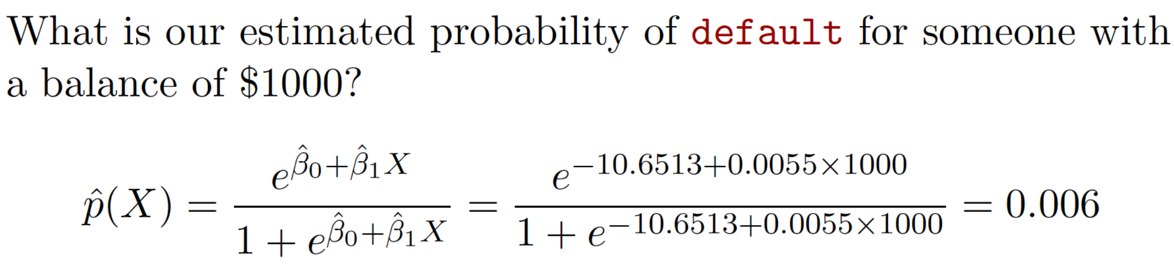


Logistic Regression

Let's write $p(X) = \Pr(Y = 1|X)$ for short and consider using `balance` to predict `default`. Logistic regression uses the form

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

($e \approx 2.71828$ is a mathematical constant [Euler's number.])
It is easy to see that no matter what values β_0, β_1 or X take, $p(X)$ will have values between 0 and 1.

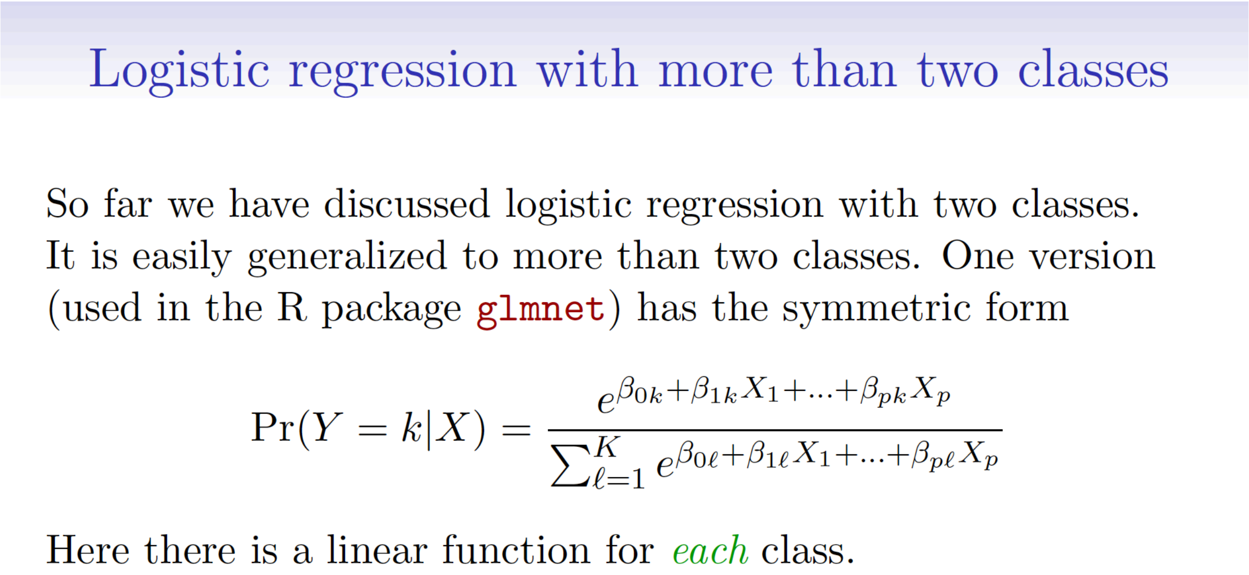


What is our estimated probability of `default` for someone with a balance of $1000?

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1000}}{1+e^{-10.6513+0.0055\times 1000}} = 0.006$$

Q4 (5 Points) What is the equation for logistic regression with more than three classes?

Solution:



Logistic regression with more than two classes

So far we have discussed logistic regression with two classes. It is easily generalized to more than two classes. One version (used in the R package `glmnet`) has the symmetric form

$$\Pr(Y=k|X) = \frac{e^{\beta_{0k}+\beta_{1k}X_1+\ldots+\beta_{pk}X_p}}{\sum_{\ell=1}^{K} e^{\beta_{0\ell}+\beta_{1\ell}X_1+\ldots+\beta_{p\ell}X_p}}$$

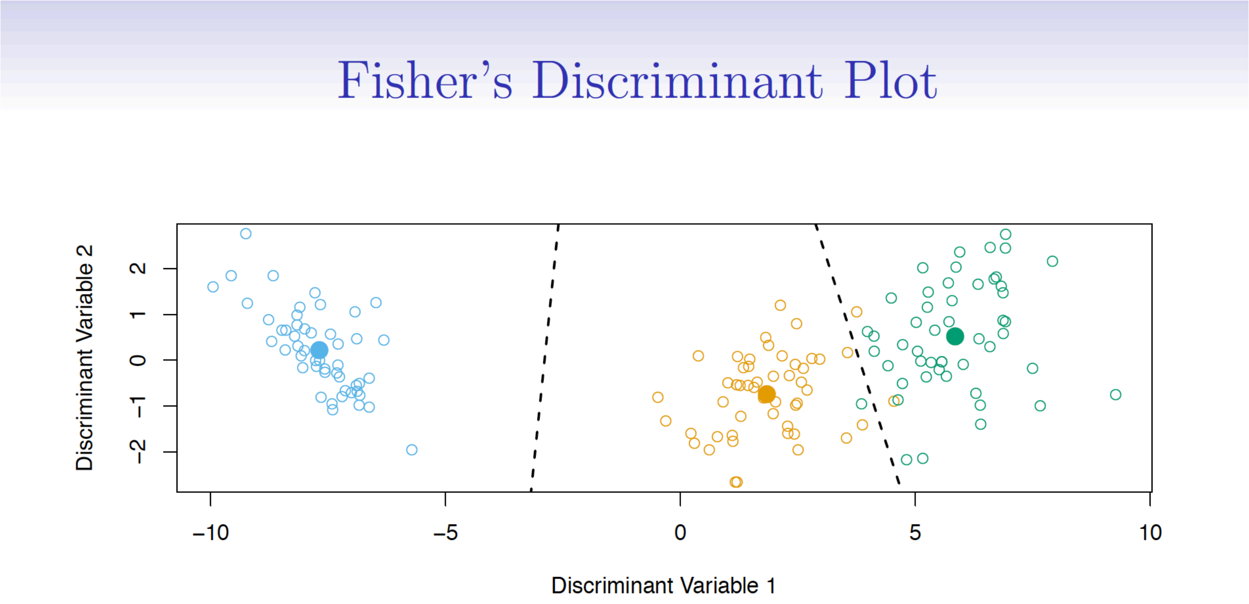
Here there is a linear function for *each* class.

Q5 (5 Points) What is a linear discriminant?

Solution:

A line that discriminates (i.e. separates in to classes).

See below



Fisher's Discriminant Plot

Q6 (5 Points) Explain whether each scenario is a classification or regression problem.
Indicate whether we are most interested in inference or prediction.

(a) We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

(b) We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

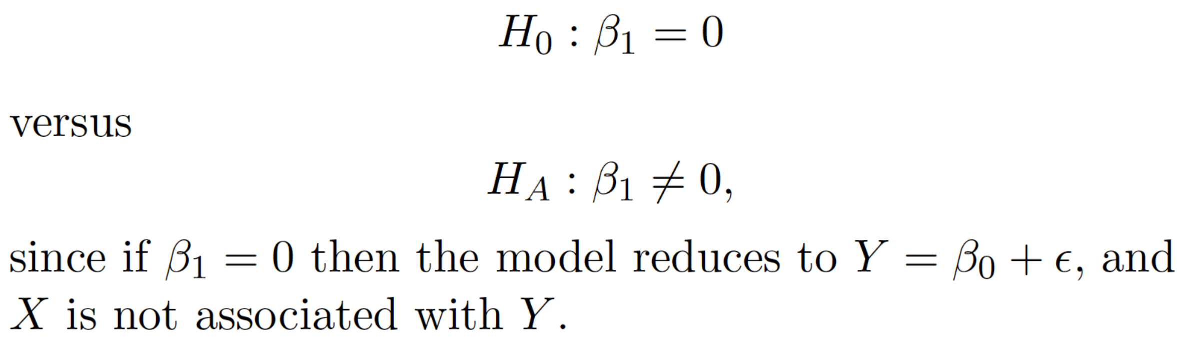
Solution:

(a) Inference because we are interested in understanding which factors affect CEO salary. Regression because our dependent variable CEO salary is numeric.

(b) Prediction because we wish to know whether it will be a success or a failure. Classification if we predict success or a failure. Logistic regression if we predict a probability of success or a failure.

Q7 (5 Points) Describe the null hypotheses to which the p-values given in linear regression correspond.

Solution:



$$H_0 : \beta_1 = 0$$

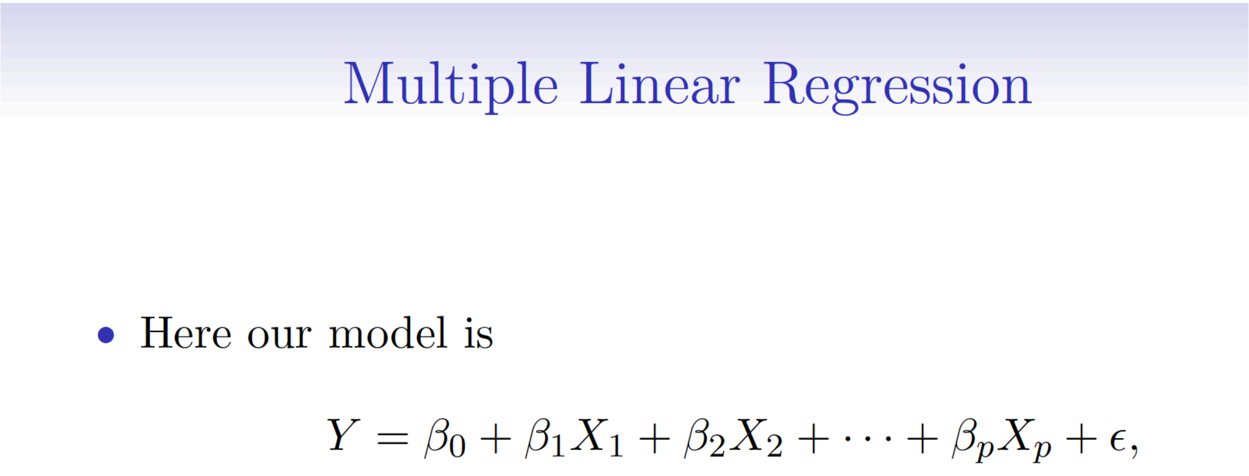
versus

$$H_A : \beta_1 \neq 0,$$

since if $\beta_1 = 0$ then the model reduces to $Y = \beta_0 + \epsilon$, and X is not associated with Y.

Q8 (5 Points) What is the equation for multiple linear regression?

Solution:



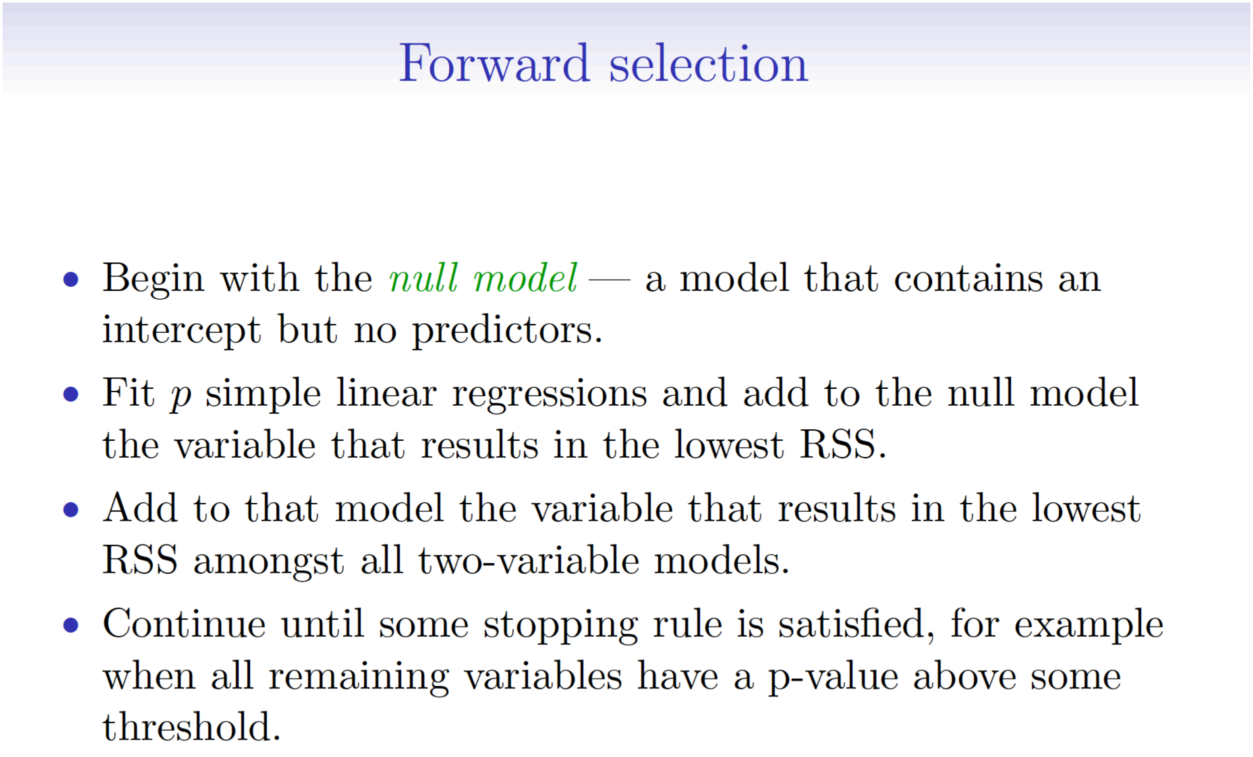
Multiple Linear Regression

- Here our model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon,$$

Q9 (5 Points) Describe forward step-wise selection.

Solution:



Forward selection

- Begin with the *null model* — a model that contains an intercept but no predictors.
- Fit p simple linear regressions and add to the null model the variable that results in the lowest RSS.
- Add to that model the variable that results in the lowest RSS amongst all two-variable models.
- Continue until some stopping rule is satisfied, for example when all remaining variables have a p-value above some threshold.

Q10 (5 Points) k-fold cross-validation.

(a) Explain how k-fold cross-validation is implemented.
(b) What are the advantages and disadvantages of k-fold cross validation relative to:

i. The validation set approach?
ii. LOOCV?

Solution:

In **k**-**fold cross**-**validation**, the original sample is randomly partitioned into **k** equal sized subsamples. Of the **k** subsamples, a single subsample is retained as the **validation** data for testing the model, and the remaining **k** − 1 subsamples are used as training data.

i. The validation set approach? his split is usually done using a 70–30 or 80-20 ratio and can beneficial because the amount of time required for training will be lower than CV. However, its estimates have higher variance than k-fold cross-validation.

ii. Leave-one-out cross-validation (LOOCV) is approximately unbiased, but it tends to have a high variance

Q11 (5 Points) Suppose that we use some statistical learning method to make a prediction for the response Y for a particular value of the predictor X.

Carefully describe how we might estimate the standard deviation of our prediction.

Solution:

Use bootsrapping. Bootstrapping is any test or metric that relies on random sampling with replacement. Bootstrapping allows assigning measures of accuracy (defined in terms of bias, variance, confidence intervals, prediction error or some other such measure) to sample estimates. This technique allows estimation of the sampling distribution of almost any statistic using random sampling methods

Q12 (5 Points) Assume one wants to use gender (male/female) as an independent variable in regression. How can we encode it?

Solution:

1 female and 0 male (or vice versa)

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male} \end{cases}$$

Resulting model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male.} \end{cases}$$

Q13 (5 Points) Why might we have high bias in our model fitting?

Solution:

The two main reasons for high bias are insufficient model capacity and underfitting because the training phase wasn't complete.

Q14 (5 Points) What is an interaction variable in regression?

Solution:

An **interaction variable** or **interaction feature** is a variable constructed from an original set of variables to try to represent either all of the interaction present or some part of it. In exploratory statistical analyses it is common to use products of original variables as the basis of testing whether interaction is present with the possibility of substituting other more realistic interaction variables at a later stage. When there are more than two explanatory variables, several interaction variables are constructed, with pairwise-products representing pairwise-interactions and higher order products representing higher order interactions.

The binary factor *A* and the quantitative variable *X* interact (are non-additive) when analyzed with respect to the outcome variable *Y*.

Thus, for a response *Y* and two variables x_1 and x_2 an *additive* model would be:

$$Y = c + ax_1 + bx_2 + \text{error}$$

In contrast to this,

$$Y = c + ax_1 + bx_2 + d(x_1 \times x_2) + \text{error}$$

Q15 (5 Points) What is the Akaike information criterion (AIC) equation?

Solution:

Let k be the number of estimated parameters in the model. Let k be the maximum value of the likelihood function for the model. Then the AIC value of the model is the following.

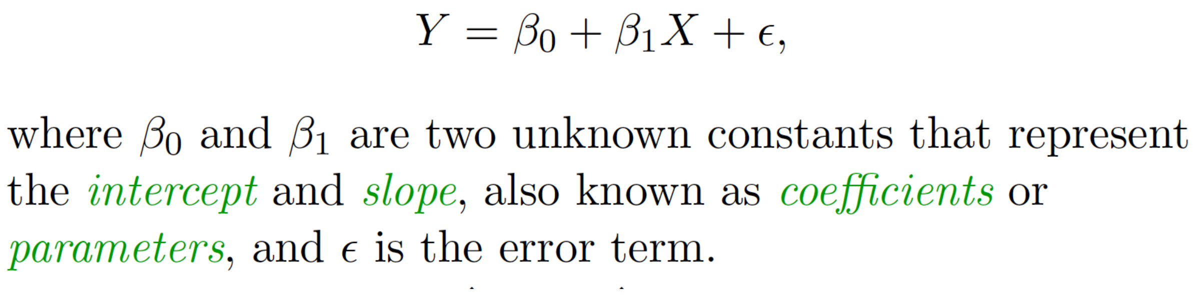
$$\mathrm{AIC} \, = \, 2k - 2\ln(\hat{L})$$

Q16 (5 Points) Describe the differences between a parametric and a non-parametric statistical learning approach. Give an example of a parametric equation and point out the parameters.

Solution:

In parametric approaches one estimates parameter values, in non-parametric approaches one doesn’t.

The equation of a line below is a parametric equation. The slope and intercept are its parameters.



$$Y = \beta_0 + \beta_1 X + \epsilon,$$

where β_0 and β_1 are two unknown constants that represent the *intercept* and *slope*, also known as *coefficients* or *parameters*, and ϵ is the error term.

Q17 (10 Points)

Suppose we have a data set with five predictors, X1 =GPA, X2 = IQ, X3 = Gender (1 for Female and 0 for Male), X4 = Interaction between GPA and IQ, and X5 = Interaction between GPA and Gender. The response is starting salary after graduation (in thousands of dollars).

Suppose we use least squares to fit the model, and get ˆβ0 = 50, ˆβ1 = 20, ˆβ2 = 0.07, ˆβ3 = 35, ˆβ4 = 0.01, ˆβ5 = −10.

(a) Is the following answer correct, and why? For a fixed value of IQ and GPA, males earn more on average than females.

(b) Predict the salary of a female with IQ of 110 and a GPA of 4.0.

(c) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

Solution:

a) Is the following answer correct, and why?
For a fixed value of IQ and GPA, males earn more on average
than females.

Males earn less.

The same values of IQ and GPA generate the same values for males and females however the gender variable ˆβ3 = 35 is positive and 1 means female so if female the prediction will be 35 more for a fixed value of IQ and GPA.

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male} \end{cases}$$

Resulting model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male.} \end{cases}$$

(b) Predict the salary of a female with IQ of 110 and a GPA of 4.0.

X1 =GPA, X2 = IQ, X3 = Gender (1 for Female and 0 for Male), X4 = Interaction between GPA and IQ, and X5 = Interaction between GPA and Gender.

ˆβ0 = 50, ˆβ1 = 20, ˆβ2 = 0.07, ˆβ3 = 35, ˆβ4 = 0.01, ˆβ5 = −10.

Use equation y = β0 + β1* GPA + β2* IQ + β3* Gender + β4*(GPA*IQ) + β5* (GPA* Gender)

(c) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

False. We care about the p-value not the coefficient.

Q18 (10 Points)

What are:

- True positive rate (TPR),

- False positive rate (FPR),

How do TPR and FPR relate to AUC or AUROC?

Solution:

AUC = Area Under the Curve.
AUROC = Area Under the Receiver Operating Characteristic curve.

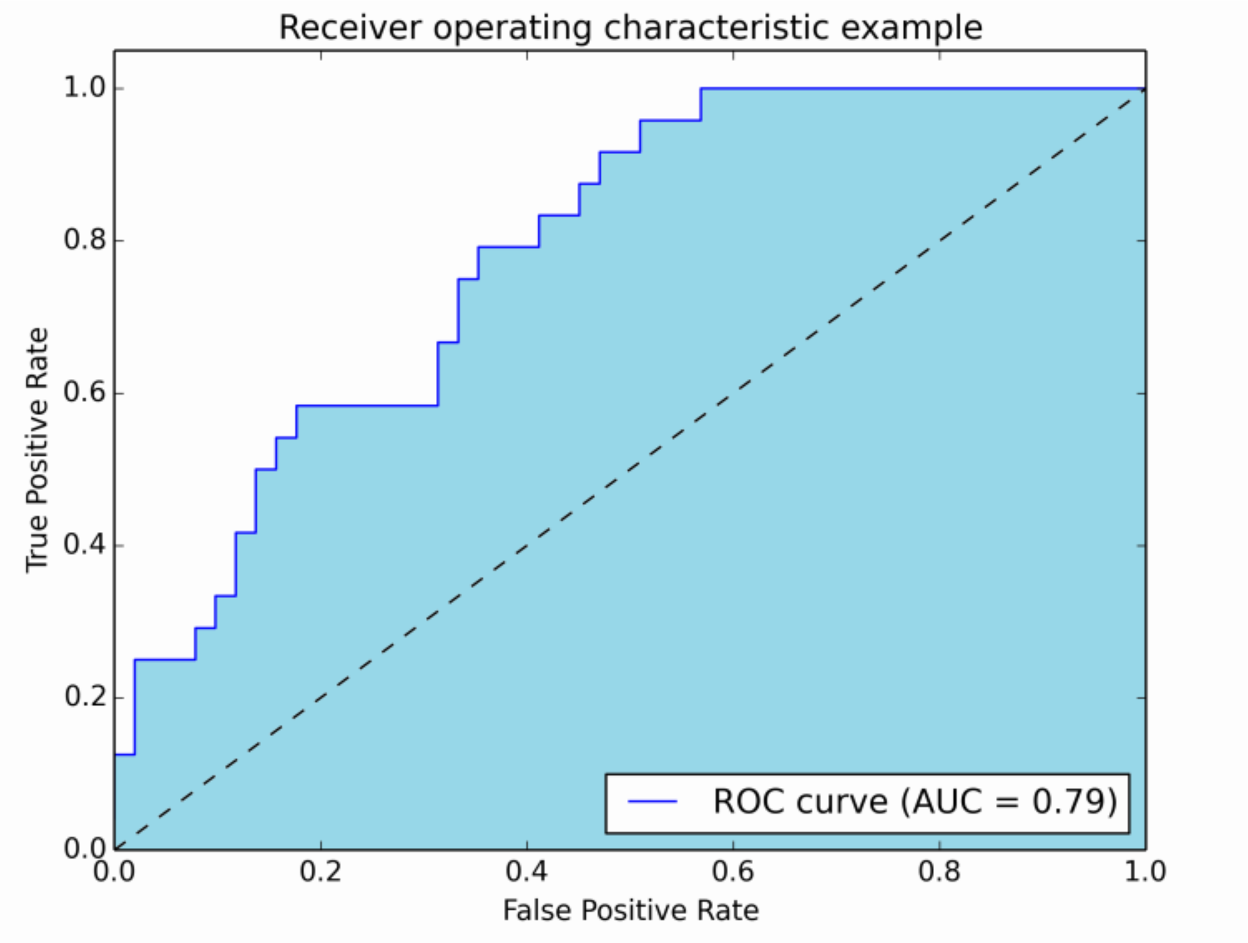
AUC is used most of the time to mean AUROC

- True positive rate (TPR), aka. sensitivity, hit rate, and recall, which is defined as TP/TP+FN. Intuitively this metric corresponds to the proportion of positive data points that are correctly considered as positive, with respect to all positive data points. In other words, the higher TPR, the fewer positive data points we will miss.

- False positive rate (FPR), aka. fall-out, which is defined as FP/FP+TN. Intuitively this metric corresponds to the proportion of negative data points that are mistakenly considered as positive, with respect to all negative data points. In other words, the higher FPR, the more negative data points will be missclassified.

To combine the FPR and the TPR into one single metric, we first compute the two former metrics with many different thresholds (for example 0.00; 0.01, 0.02,… 1.00), then plot them on a single graph, with the FPR values on the abscissa and the TPR values on the ordinate. The resulting curve is called ROC curve, and the metric we consider is the AUC of this curve, which we call AUROC.

The following figure shows the AUROC graphically:



In this figure, the blue area corresponds to the Area Under the curve of the Receiver Operating Characteristic (AUROC). The dashed line in the diagonal we present the ROC curve of a random predictor: it has an AUROC of 0.5. The random predictor is commonly used as a baseline to see whether the model is useful.