Part II: Predicting Surgery Outcomes

A) (6 points) Explain why an ordinary least squares method would not be a good model choice. In particular, explain which assumptions of OLS are violated and why.

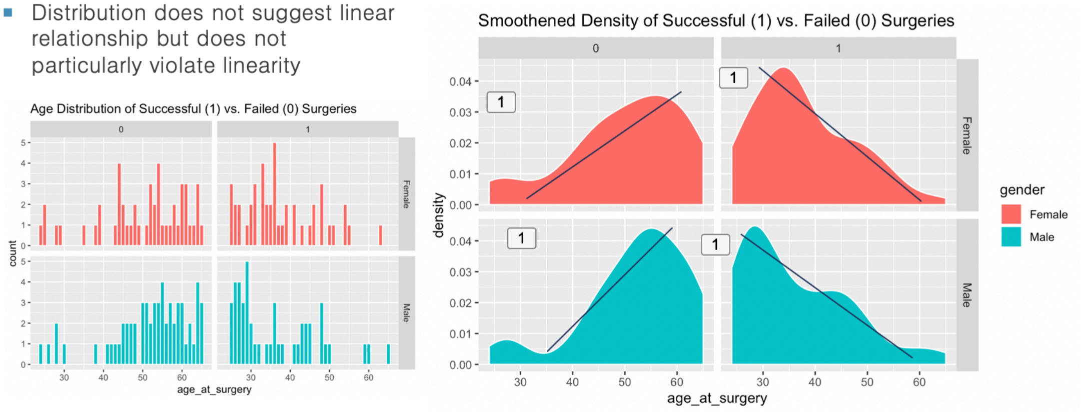
OLS regression assumes the underlying data follow the 'standard statistical mode' given by

$$Y = \beta_0 + \sum_{i=1}^{n}[\beta_i x_i + e_i], \text{ where } \mathbb{E}(e_i) = 0 \text{ and } \mathrm{Var}(e_i) = \sigma^2$$

and thus involves a number of key assumptions:

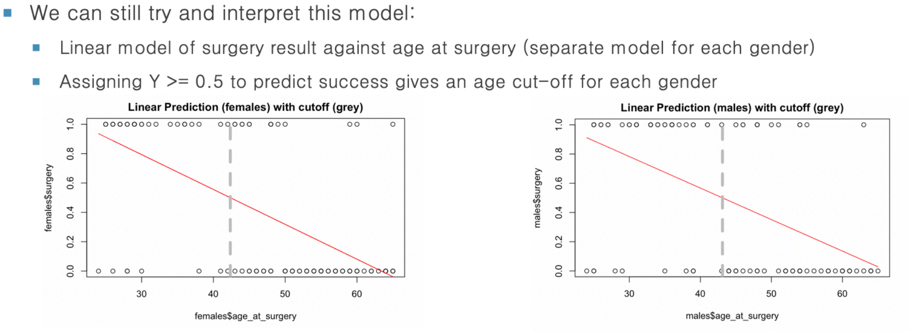
- Linearity: the relationship between predictor and response variables is linear
- Independence: observations are independent from each other (SRS)
- Homoscedasticity: variance of residual e_i is the same for any fixed predictor value
- Normality: the response variable follows a gaussian distribution for a fixed predictor value

The assumptions of linearity and normality are not met, but this does not directly mean we should not at all consider a least-squares linear model. Looking at the observed data, we see:



- Distribution does not suggest linear relationship but does not particularly violate linearity

Trying the OLS regression, we get:



- We can still try and interpret this model:
  - Linear model of surgery result against age at surgery (separate model for each gender)
  - Assigning Y >= 0.5 to predict success gives an age cut-off for each gender

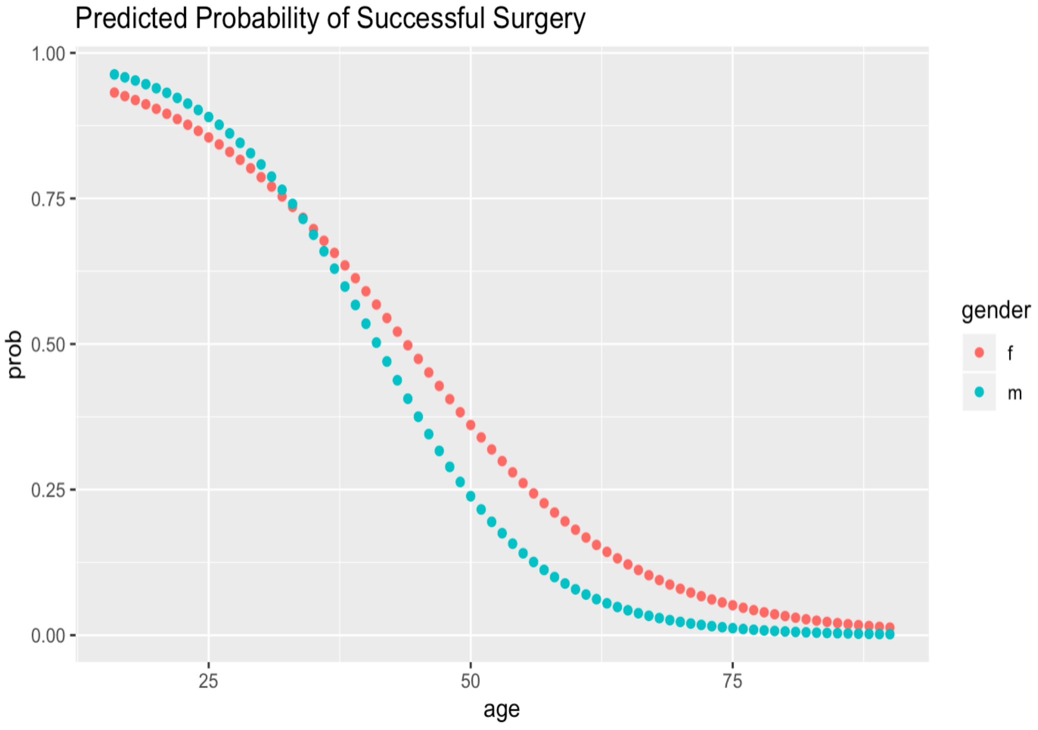
where for example, our model suggests a negative prediction for ages 70+. This is difficult to interpret. Luckily, we can do much better than OLS regression. There is a number of statistical/machine learning methods for this classification, but with the amount of data we have (200 patients), we'll use logistic regression.

B) (10 points) Fit an appropriate model; use Likelihood Ratio Tests for any model selection.

We use logistic regression with an 80/20 Train/Test split. This model is much more appropriate for the binary classification system we have, where 1 denotes successful surgery and 0 otherwise.

Observations are independent of each other and independent variables are linear. Additionally, with binary outcomes in the response variable, our sample size of 200 should be sufficient to drive the model.

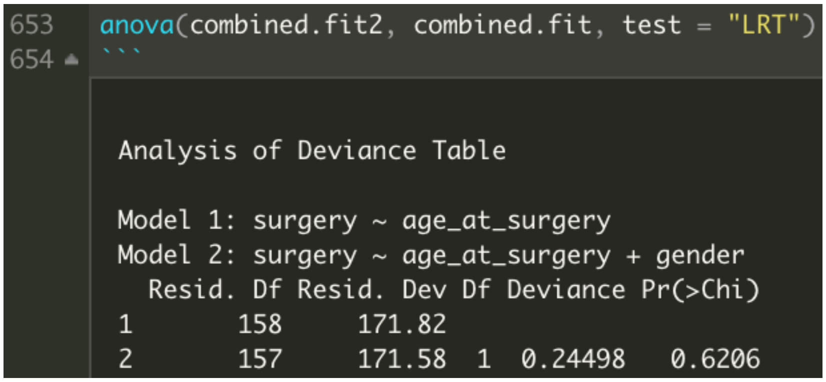
The resulting predicted probability for each age is given by the following plot:



and training / testing accuracies are:

Model & Predictor Variables	Train Accuracy	Test Accuracy
Genders: Combined Predictors: Age at Surgery, Gender	76.24%	84.99%
Genders: Combined Predictors: Age at Surgery	76.87%	82.49%
Genders: Female Predictors: Age at Surgery	73.33%	89.47%
Genders: Male Predictors: Age at Surgery	77.38%	86.36%

Now a Likelihood Ratio Test (which in the context of analysis of variance is equivalent to the chi-square goodness of fit test) gives:



```
anova(combined.fit2, combined.fit, test = "LRT")
```
```

```
Analysis of Deviance Table

Model 1: surgery ~ age_at_surgery
Model 2: surgery ~ age_at_surgery + gender
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       158     171.82
2       157     171.58  1  0.24498   0.6206
```

which suggests that gender is not a significant predictor of surgery outcome at the 0.05 significance level.
```