Logistic Regression:

- So far y_i was a continuous variable. What if y_i is categorical?
- Assume we have **2 classes**.
- Even if we can make these classes numerical (i.e. translate labels such as "yes"/"no" into 1 / 0), these numbers don't have a mathematical meaning in the context of linear models and what we learn will be as arbitrary as the numerical labels we assigned (i.e. using "yes" =2/"no"=7 instead of "yes"=1/"no"=0 might "fit" a better model…).
- Maybe we can use the probability of belonging to a given class as a proxy for how confidently we can classify a given point? Maybe we can fit a linear model to the probability of being in a given class!
- So the output of our regression model could be a probability. But how can we enforce that $X\beta_{LS}$ from our model is always constrained to [0,1]? i.e. how can we learn a β_{LS} such that $0 \leq X\beta_{LS} \leq 1$ even for unseen X?
- Instead define the odds = p / 1 - p where p = P(Y = class 1 | X)
- Now the range of $X\beta_{LS}$ is $[0, \infty)$
- But again how can we enforce that the $X\beta_{LS}$ are constrained to $[0, \infty)$? We need $(-\infty, \infty)$ - but how?
- Let's take the log! This is also convenient numerically because in the previous odds format, tiny variations in p have large effects on the odds!
- Our goal is to fit a linear model to the log-odds of being in one of our classes (in the 2-class case) i.e.
  - $$\log\left(\frac{P(Y=1|X)}{1-P(Y=1|X)}\right) = \alpha + \beta X$$
- Suppose we have such a model. How do we recover the P(Y=1|X)?
  - $$\begin{aligned}\log\left(\frac{P(Y=1|X)}{1-P(Y=1|X)}\right) &= \alpha + \beta X \\ \frac{P(Y=1|X)}{1-P(Y=1|X)} &= e^{\alpha+\beta X} \\ \frac{P(Y=1|X)}{1-P(Y=1|X)} + 1 &= e^{\alpha+\beta X} + 1 \\ \frac{P(Y=1|X)}{1-P(Y=1|X)} &= e^{\alpha+\beta X} + 1 \\ P(Y=1|X) &= \frac{e^{\alpha+\beta X}}{1+e^{\alpha+\beta X}}\end{aligned}$$

The function we apply to our probability to obtain the log odds is called the **logit** function. The function used to retrieve our probability from the log odds is called **logit^{-1}**

-
- How do we learn our model? I.e. the α and β parameters.
- We know:

- $$P(y_i = 1|x_i) = \begin{cases} logit^{-1}(\alpha + \beta x_i) \text{ if } y_i = 1 \\ 1 - logit^{-1}(\alpha + \beta x_i) \text{ if } y_i = 0 \end{cases}$$
  $$= (logit^{-1}(\alpha + \beta x_i))^{y_i}(1 - logit^{-1}(\alpha + \beta x_i))^{1-y_i}$$

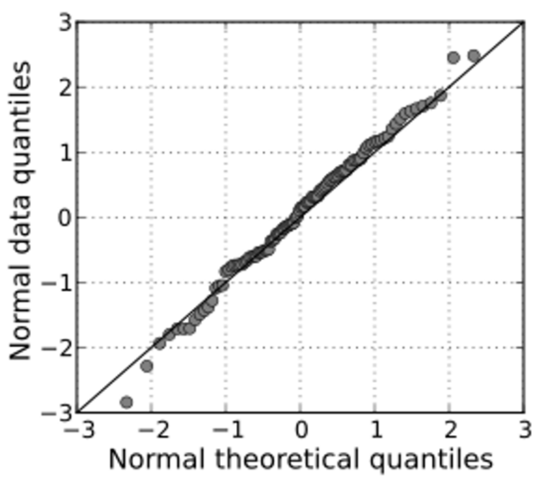
- $$L(\alpha, \beta) = \prod_i (logit^{-1}(\alpha + \beta x_i))^{y_i}(1 - logit^{-1}(\alpha + \beta x_i))^{1-y_i}$$
    - And try to maximize this quantity!
    - Unfortunately, there is no closed form solution here and we need to use numerical approximation methods to solve this optimization problem
- Evaluating our Regression Model:
  - $$TSS = \sum_i^n (y_i - \bar{y})^2$$
    $$RSS = \sum_i^n (y_i - \hat{y}_i)^2$$
    $$ESS = \sum_i^n (\hat{y}_i - \bar{y})^2$$
  - $$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$
- R^2 measures the fraction of variance that is explained by ŷ
- Each parameter of an independent variable **x** has an associated confidence interval
- If the parameter / coefficient is not significantly distinguishable from 0 then we cannot assume that there is a significant linear relationship between that independent variable and the observations **y** (i.e. if the interval includes 0)
- Confidence interval:
    - How do we build a confidence interval?

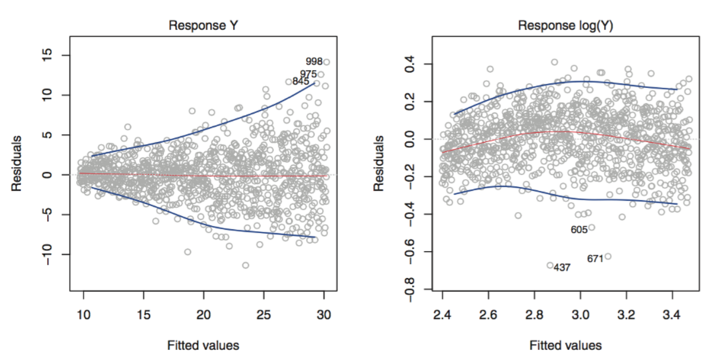
      Assume $\mathbf{Y_i \sim N(5, 25)}$, for $\mathbf{1 \le i \le 100}$ and $\mathbf{y_i = \mu + \epsilon}$ where $\mathbf{\epsilon \sim N(0, 25)}$. Then the Least Squares estimator of $\boldsymbol{\mu}$ ($\boldsymbol{\mu_{LS}}$) is

      the sample mean **ȳ**

      What is the 95% confidence interval for $\boldsymbol{\mu_{LS}}$?

      $\mathbf{CI_{.95} = [\bar{y} - 1.96 \times SE(\mu_{LS}), \bar{y} + 1.96 \times SE(\mu_{LS})]}$
      $\mathbf{= [\bar{y} - 1.96 \times .5, \bar{y} + 1.96 \times .5]}$
- Z-values:
    - These are the number of standard deviations from the mean of a N(0,1) distribution required in order to contain a specific % of values were you to sample a large number of times.
    - To find the .95 z-value (the number of standard deviations from the mean that contains 95% of values) you need to solve:
    -

    - $$\int_{-z}^{z} \frac{1}{2\pi} e^{-\frac{1}{2}x^2} dx = .95$$
- QQ Plot:
    - We need to check our assumption that our residuals / noise estimates are normally distributed.
    - How do can you check that a variable follows a specific distribution?
    - Need to check that our variable is **distributed** in the same way that a variable following our target distribution would be.
    - Plot the quantile of your target distribution against the quantiles of your data/ variable! If they match then your data probably comes from that distribution.
- Quantiles are the values for which a particular % of values are contained below it.
- For example the 50% quantile of a N(0,1) distribution is 0 since 50% of samples would be contained below 0 were you to sample a large number of times.
    - 

- Constant Variance:
    - One of our assumptions was that our noise had constant variance. How can we verify this?
    - We can plot our fitted values against our residuals (noise estimates)
    - 

- Extending our linear model:

- Changing the assumptions we made can drastically change the problem we are solving. A few ways to extend the linear model:
- 1.Non-constant variance - used in WLS (weighted least squares)
- 2.Distribution of error is not Normal - used in GLM (generalized linear models)