MASM22/FMSN30: Linear and Logistic Regression, 7.5 hp
FMSN40: ... with Data Gathering, 9 hp

Lecture 7, spring 2023
Logistic regression:
probabilities, odds and odds ratios
Maximum-likelihood estimates, Wald test

Mathematical Statistics / Centre for Mathematical Sciences
Lund University

24/4-23

Introduction to Logistic regression

- In this part of the course we consider a *nonlinear model* (nonlinear in the parameters).
- However, it will be a monotonous transformation of a linear relationship making it a **Generalized Linear Model** (GLM)
- This time our response variable Y will be a **discrete**, **binary variable** (success/failure, yes/no, etc).
- The nature of the response will make the Bernoulli (a special case of the Binomial) distribution a natural choice.
- The resulting regression model is called **logistic regression**, because we will use a logistic transformation.
- Our expected response will be the probability of success.

Why is this relevant?

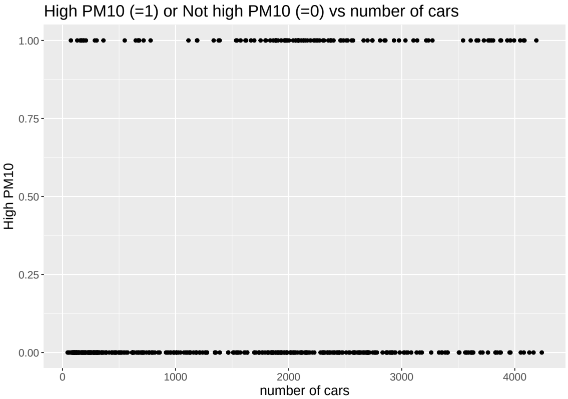
Examples:

- political election: response is win/lose. What factors (covariates) affect the probability to win? (e.g. money spent on campaign; age of the candidate etc.)
- result of some medical test (positive/negative): estimate the probability to have a "positive" result, depending on several physiological covariates.
- crash test dummies. Probability of "survival" of a dummy, depending on several test conditions.
- ...

We consider logistic regression with binary response. But extension to multicategory (or polytomous) response are possible, assuming a multinomial distributed response.

Example: particles in Oslo

A random subsample of 500 observations from the Norwegian Public Roads Administration measuring whether the concentration of atmospheric particles with a diameter between 2.5 and 10 μm, PM$_{10}$, exceeds the limit 50 μg/m^3.



Model???

Binomial distribution (a reminder)

Let Y be the number of successes in n independent trials, each with the same probability of success, p. Then $Y \sim Bin(n, p)$ with

$$Pr(Y = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n$$

$$E(Y) = np, \qquad V(Y) = np(1-p).$$

For the estimate $\hat{p} = Y/n$ we have

$$\hat{p} \approx N(p, \frac{p(1-p)}{n}) \qquad I_p \approx (\hat{p} \pm \lambda_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}})$$

when n is large enough, typically when $np(1-p) > 10$.
Warning: If n is too small the interval can go outside $[0, 1]$.
We will have $n = 1$. Not even close to "large enough".

Before (linear regression)

Y_i was a continuous variable with

$$Y_i = \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i \text{ where } \epsilon_i \sim N(0,\, \sigma^2) \Leftrightarrow Y_i \sim N(\mu_i,\, \sigma^2)$$
$$E(Y_i) = \mu_i = \mathbf{x}_i\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

Now (logistic regression)

Y_i is discrete with two possible outcomes: success (1) or failure (0) with probabilities $Pr(Y_i = 1) = p_i$ and $Pr(Y_i = 0) = 1 - p_i$

$$Y_i \sim Bin(1, p_i) \text{ with } Pr(Y_i = k) = p_i^k(1-p_i)^{1-k},\ k = 0, 1$$
$$E(Y_i) = \mu_i = p_i = \text{ some function of } \mathbf{x}_i$$
$$V(Y_i) = p_i(1-p_i) \text{ also depends on } \mathbf{x}_i$$

Choosing $\mu_i = p_i = \mathbf{x}_i\boldsymbol{\beta}$ is *not* good since $0 \leq p_i \leq 1$.

Odds: number of successes for each failure

The odds of "success" is defined as

$$\text{odds} = \frac{Pr(\text{success})}{Pr(\text{failure})} = \frac{p}{1-p} \quad \Leftrightarrow p = \frac{\text{odds}}{1+\text{odds}}$$

$$\text{log-odds} = \ln \text{odds} = \ln \frac{p}{1-p} = \text{logit}(p)$$

$$\text{odds}_{\text{failure}} = \frac{1}{\text{odds}_{\text{success}}} \qquad \ln \text{odds}_{\text{failure}} = -\ln \text{odds}_{\text{success}}$$

	min	middle	max	
p	0	1/2	1	
odds	0	1	∞	
$\ln$ odds	$-\infty$	0	∞	no bounds!

Logistic regression model

We assume that

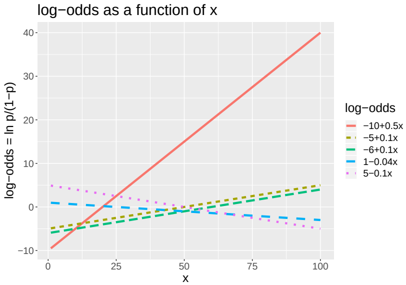
$$Y_i = \text{“success”}\ (=1) \text{ or } \text{“failure”}\ (=0)$$
$$Pr(Y_i = 1) = 1 - Pr(Y_i = 0) = p_i$$
$$Y_i \sim Bin(1, p_i), \quad i = 1, \ldots, n, \text{ (independent)}$$
$$\text{logit}(p_i) = \ln \frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} = \mathbf{x}_i \boldsymbol{\beta}$$
$$p_i = \frac{\mathrm{e}^{\mathbf{x}_i \boldsymbol{\beta}}}{1 + \mathrm{e}^{\mathbf{x}_i \boldsymbol{\beta}}}$$

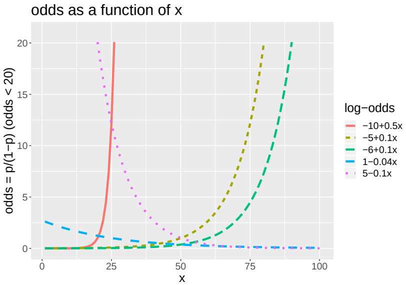
Parameters

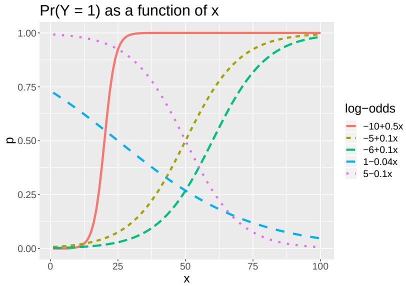
$\beta_0 =$ log-odds and $e^{\beta_0} =$ odds when all x_{ij} are 0,

$\beta_j =$ additive change in log-odds and...

$\mathrm{e}^{\beta_j} =$ relative change in odds when x_{ij} is increased by 1, $j = 1, \ldots, p$







$$Y_i \sim Bin(1,\, p_i)$$

The log-odds is linear:

$$\ln \text{odds}_i = \beta_0 + \beta_1 x_i$$

The odds is exponential:

$$\text{odds}_i = \mathrm{e}^{\beta_0+\beta_1 x_i} = \mathrm{e}^{\beta_0} \cdot (\mathrm{e}^{\beta_1})^{x_i}$$

The probability is S-shaped:

$$p_i = \frac{\mathrm{e}^{\beta_0+\beta_1 x_i}}{1+\mathrm{e}^{\beta_0+\beta_1 x_i}}$$

Interpretation of β_1: log odds ratio

- What happens when we increase x by 1?

$$\text{odds ratio} = \mathrm{OR} = \frac{e^{\beta_0+\beta_1(x+1)}}{e^{\beta_0+\beta_1 x}} = e^{\beta_1}$$

If $\beta_1 = 0.04$ then $e^{\beta_1} = 1.04$ and the odds increases by 4 %.
If $\beta_1 = -0.04$ then $e^{\beta_1} = 0.96$ and the odds decreases by 4 %.

- What happens when we increase x by 10?

$$\mathrm{OR} = \frac{e^{\beta_0+\beta_1(x+10)}}{e^{\beta_0+\beta_1 x}} = e^{10\beta_1} = (e^{\beta_1})^{10}$$

If $\beta_1 = 0.04$ then $(e^{\beta_1})^{10} = 1.04^{10} = 1.49$ and the odds increases by 49 %.
If $\beta_1 = -0.04$ then $(e^{\beta_1})^{10} = 0.96^{10} = 0.67$ and the odds decreases by 33 %.

Size of the change

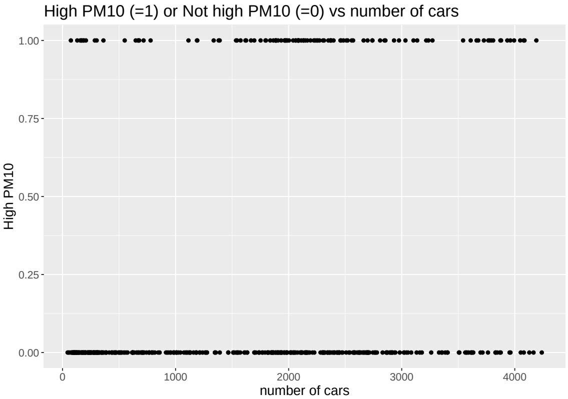
Marginal change = derivative ($\mathbf{x}\boldsymbol{\beta} = \beta_0 + \beta_1 x$):

$$\begin{aligned}
\frac{d\text{logodds}}{dx} &= \frac{d}{dx}\mathbf{x}\boldsymbol{\beta} = \beta_1 && \text{constant,} \\
\frac{d\text{odds}}{dx} &= \frac{d}{dx}\mathrm{e}^{\mathbf{x}\boldsymbol{\beta}} = \beta_1 \mathrm{e}^{\mathbf{x}\boldsymbol{\beta}} = \beta_1 \cdot \text{odds} && \text{prop. to the odds,} \\
\frac{dp}{dx} &= \frac{d}{dx}\frac{\mathrm{e}^{\mathbf{x}\boldsymbol{\beta}}}{1+\mathrm{e}^{\mathbf{x}\boldsymbol{\beta}}} = \\
&= \beta_1 \cdot \frac{\mathrm{e}^{\mathbf{x}\boldsymbol{\beta}}}{1+\mathrm{e}^{\mathbf{x}\boldsymbol{\beta}}}\left(1 - \frac{\mathrm{e}^{\mathbf{x}\boldsymbol{\beta}}}{1+\mathrm{e}^{\mathbf{x}\boldsymbol{\beta}}}\right) = \\
&= \beta_1 \cdot p(1-p) && \text{prop. to } V(Y|x)
\end{aligned}$$

The size of the change in p is largest around $p = 0.5$ and gets smaller as $p \to 0$ or $\to 1$.

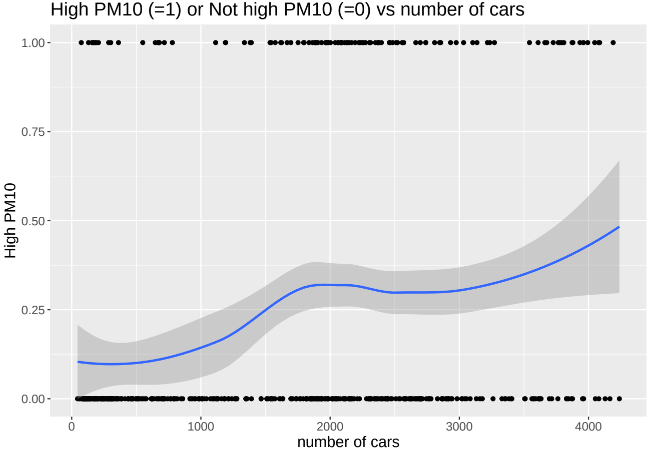
Example: particles in Oslo

A random subsample of 500 observations from the Norwegian Public Roads Administration measuring whether the concentration of atmospheric particles with a diameter between 2.5 and 10 μm, PM_{10}, exceeds the limit 50 $\mu\text{g}/\text{m}^3$.



Does the data follow an S-shape? Well...

We can get a rough estimate of the shape using a moving average which calculates the average Y-value in an interval moving along the x-axis.



Sort of S-shaped. Obviously $\beta_1 > 0$. More cars give larger probability of exceeding the concentration limit.

How should we estimate $\boldsymbol{\beta}$

Least squares estimates?

- Minimize $Q(\boldsymbol{\beta}) = \sum_{i=1}^{n}(\ln \frac{Y_i}{1-Y_i} - \mathbf{x}_i\boldsymbol{\beta})^2$?
  No, $\ln \frac{Y_i}{1-Y_i} = \ln 0 = -\infty$ or $\ln \infty = \infty$. Useless!
- Minimize $Q(\boldsymbol{\beta}) = \sum_{i=1}^{n}(Y_i - p_i)^2 = \sum_{i=1}^{n}(Y_i - \frac{e^{\mathbf{x}_i\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i\boldsymbol{\beta}}})^2$?
  No, since $V(Y_i) = p_i(1-p_i)$ is not constant. We would need to do a weighted least squares but the weights $1/V(Y_i)$ are unknown.
- Minimize $Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{(Y_i-p_i)^2}{p_i(1-p_i)} = \sum_{i=1}^{n} \frac{(Y_i - \frac{e^{\mathbf{x}_i\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i\boldsymbol{\beta}}})^2}{\frac{e^{\mathbf{x}_i\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i\boldsymbol{\beta}}}(1-\frac{e^{\mathbf{x}_i\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i\boldsymbol{\beta}}})}$
  using iteratively re-weighted least squares?
  No, it can be done but it is a very inefficient method with a slow convergence rate.

Totally different method? Yes!

Maximum likelihood-method

Since we know what type of distribution our data come from, $Y_i \in Bin(1, p_i)$, we can find the $\boldsymbol{\beta}$-values that maximize the probability of getting exactly the observation values that we got. That means that we should maximize the likelihood function

$$L(\boldsymbol{\beta}; \mathbf{Y}) = Pr(Y_1 = y_1, \ldots, Y_n = y_n) = \prod_{i=1}^{n} Pr(Y_i = y_i)$$

$$= \prod_{i=1}^{n} p_i^{Y_i}(1-p_i)^{1-Y_i} = \prod_{i=1}^{n} (\frac{\mathrm{e}^{\mathbf{x}_i\boldsymbol{\beta}}}{1+\mathrm{e}^{\mathbf{x}_i\boldsymbol{\beta}}})^{Y_i}(1 - \frac{\mathrm{e}^{\mathbf{x}_i\boldsymbol{\beta}}}{1+\mathrm{e}^{\mathbf{x}_i\boldsymbol{\beta}}})^{1-Y_i}$$

$$= \prod_{i=1}^{n} \left(\frac{\mathrm{e}^{\mathbf{x}_i\boldsymbol{\beta}}}{1+\mathrm{e}^{\mathbf{x}_i\boldsymbol{\beta}}}\right)^{Y_i} \left(\frac{1}{1+\mathrm{e}^{\mathbf{x}_i\boldsymbol{\beta}}}\right)^{1-Y_i} = \prod_{i=1}^{n} \frac{\mathrm{e}^{Y_i\mathbf{x}_i\boldsymbol{\beta}}}{1+\mathrm{e}^{\mathbf{x}_i\boldsymbol{\beta}}}$$

It is easier to maximize the log-likelihood function instead:

$$\ln L(\boldsymbol{\beta}; \mathbf{Y}) = \sum_{i=1}^{n} \left(Y_i\mathbf{x}_i\boldsymbol{\beta} - \ln(1+\mathrm{e}^{\mathbf{x}_i\boldsymbol{\beta}})\right)$$

ML-estimate for the Null model, $\ln \frac{p_i}{1-p_i} = \beta_0$

For the simplest model, having only an intercept, we have

$$p_i = \frac{\mathrm{e}^{\beta_0}}{1+\mathrm{e}^{\beta_0}}$$

and the ML-estimate can easily be derived as

$$\ln L(\beta_0) = \sum_{i=1}^{n} \Big(Y_i \beta_0 - \ln(1+\mathrm{e}^{\beta_0}) \Big) = \beta_0 \sum_{i=1}^{n} Y_i - n \ln(1+\mathrm{e}^{\beta_0})$$

$$\frac{d \ln L(\beta_0)}{d\beta_0} = \sum_{i=1}^{n} Y_i - \frac{n\mathrm{e}^{\beta_0}}{1+\mathrm{e}^{\beta_0}} = 0 \Rightarrow$$

$$\hat{\beta}_0 = \ln \frac{\bar{Y}}{1-\bar{Y}} \Rightarrow \hat{p}_i = \bar{Y} = \frac{\text{number of successes}}{\text{number of observations}}$$

ML-estimate for the full model: $\ln \frac{p_i}{1-p_i} = \mathbf{x}_i\boldsymbol{\beta}$

Find the $\boldsymbol{\beta}$ that maximizes the log-likelihood. This means setting all the partial derivatives equal to 0:

$$\frac{\partial \ln L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \left(x_{ij} \cdot Y_i - x_{ij} \cdot \frac{\mathrm{e}^{\mathbf{x}_i\boldsymbol{\beta}}}{1+\mathrm{e}^{\mathbf{x}_i\boldsymbol{\beta}}} \right) = 0, \quad j = 0, \ldots, p$$

where $x_{i0} = 1$ for $i = 1, \ldots, n$. This gives us the following relationships:

$$\sum_{i=1}^{n} x_{ij} Y_i = \sum_{i=1}^{n} x_{ij} p_i, \quad j = 0, \ldots, p$$

Matrix formulation: $\mathbf{X}'\mathbf{p} = \mathbf{X}'\mathbf{Y}$ "Normal equations"
where $\mathbf{p}$ is a $n \times 1$ vector with elements p_i, $i = 1, \ldots, n$.
Nonlinear in $\boldsymbol{\beta}$ so no closed form solutions. We need an iterative method, e.g. Newton-Raphson algorithm.

(*) Estimates via Newton-Raphson (a.k.a. Fisher-scoring)

- Start from an arbitrary guess $\hat{\boldsymbol{\beta}}^{(0)}$, then iterate until $\| \hat{\boldsymbol{\beta}}^{(k+1)} - \hat{\boldsymbol{\beta}}^{(k)} \|$ is small enough.
- A generic iteration k of Newton-Raphson/Fisher-scoring is: $\hat{\boldsymbol{\beta}}^{(k+1)} = \hat{\boldsymbol{\beta}}^{(k)} + (\mathbf{X}'\mathbf{W}^{(k)}\mathbf{X})^{-1}\mathbf{X}'(\mathbf{Y} - \hat{\mathbf{p}}^{(k)}), \quad k = 0, 1, \ldots$
- Here $\hat{\mathbf{p}}^{(k)}$ are estimated using the current $\hat{\boldsymbol{\beta}}^{(k)}$
- $\mathbf{W}^{(k)}$ is a diagonal matrix with elements $(w_{11}^{(k)}, \ldots, w_{nn}^{(k)})$ where $w_{ii}^{(k)} = \hat{p}_i^{(k)}(1 - \hat{p}_i^{(k)})$.
- At convergence (k large) we write $\mathbf{W}^{(k)} \equiv \mathbf{W}$ and $\hat{\mathbf{p}}^{(k)} \equiv \hat{\mathbf{p}}$.

ML-estimates of $\boldsymbol{\beta}$

At convergence the ML-estimates of $\boldsymbol{\beta}$ become

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{Z}$$

where $\mathbf{W} = \hat{\mathrm{Var}}(\mathbf{Y})$ is a diagonal matrix with elements

$$w_{ii} = \hat{p}_i(1-\hat{p}_i), \quad i = 1, \ldots, n,$$

$\mathbf{Z}$ is a column vector with elements

$$Z_i = \mathbf{x}_i\hat{\boldsymbol{\beta}} + \frac{Y_i - \hat{p}_i}{\hat{p}_i(1-\hat{p}_i)}, \quad i = 1, \ldots, n$$

and

$$\hat{p}_i = \frac{\mathrm{e}^{\mathbf{x}_i\hat{\boldsymbol{\beta}}}}{1+\mathrm{e}^{\mathbf{x}_i\hat{\boldsymbol{\beta}}}}, \quad i = 1, \ldots, n.$$

Asymptotics from likelihood estimation

For all maximum likelihood estimates, $\hat{\boldsymbol{\theta}}$, we have

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \to N(\mathbf{0}, \mathbf{I}_{\text{Fish}}^{-1}) \qquad (n \to \infty)$$

where $\mathbf{I}_{\text{Fish}}$ is the Fisher information matrix (see any reference in inference theory and some numerical analysis).
In this case, it means that

$$\hat{\boldsymbol{\beta}} \sim N_{p+1}(\boldsymbol{\beta}, (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}) \qquad (n \to \infty)$$
$$\mathbf{x}_0\hat{\boldsymbol{\beta}} \sim N(\mathbf{x}_0\boldsymbol{\beta}, \mathbf{x}_0(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{x}_0') \qquad (n \to \infty)$$

Motivates the Wald test and confidence interval for β_j and constructing intervals for p_0 based on the log odds $\mathbf{x}_0\boldsymbol{\beta}$.
Warning: for small and medium $n \ll \infty$ the normal approximation is not good. Confidence intervals for $\mathrm{x}_0\boldsymbol{\beta}$ are usually OK. For $\boldsymbol{\beta}$, use likelihood based tests and intervals instead, see Lecture 8.

Wald test for β_j (when n is very large)

Does variable x_j have a significant effect on the probability of success, i.e., does it change the log-odds of success?

Wald test

We want to test H$_0$: $\beta_j = 0$ against H$_1$: $\beta_j \neq 0$. If H$_0$ is true then

$$Z = \frac{\hat{\beta}_j - 0}{d(\hat{\beta}_j)} \sim N(0,1) \qquad \text{if } n \text{ is large}$$

and we should reject H$_0$ at significance level α if

$$\frac{|\hat{\beta}_j - 0|}{d(\hat{\beta}_j)} > \lambda_{\alpha/2}$$

Using `summary(model)` gives Wald tests for the β-parameters.
Warning: For small and medium size data ($n \ll \infty$) you should use a likelihood ratio test instead, see Lecture 8.

Wald based confidence intervals for log odds (ratios)

If n is large, so that the normal approximation of $\hat{\boldsymbol{\beta}}$ is good, we can construct confidence intervals for β_j in the usual way (define λ_α as the α-percentile from $N(0,1)$):

$$I_{\ln \mathrm{OR}_j} = I_{\beta_j} = (\hat{\beta}_j \pm \lambda_{\alpha/2} \cdot d(\hat{\beta}_j)).$$

Warning: For small and medium size data, use a profile likelihood based confidence interval instead, see Lecture 8. This is what `confint(model)` does if the `MASS` package is installed.

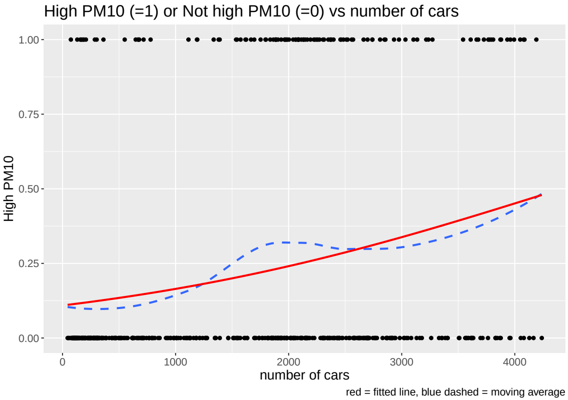
Confidence interval for odds and odds ratios

With $I_{\beta_j} = (c_1, c_2)$ we just exponentiate the bounds to get confidence intervals for the intercept odds, e^{β_0}, and the odds ratios, e^{β_j}, $j = 1, \ldots, p$:

$$I_{\mathrm{OR}_j} = I_{\mathrm{e}^{\beta_j}} = \mathrm{e}^{I_{\beta_j}} = (\mathrm{e}^{c_1}, \mathrm{e}^{c_2})$$

	param.	est.	s.e.	P-value (Wald)	95 % C.I. (profile)
Intercept	β_0	-2.10	0.22	< 0.001	$(-2.55, -1.68)$
cars/1000	β_1	0.48	0.10	< 0.001	$(0.29, 0.67)$

	param.	est.	95 % C.I.
Intercept	e^{β_0}	$e^{-2.10} = 0.12$	$(e^{-2.55}, e^{-1.68}) = (0.08,\ 0.19)$
cars/1000	e^{β_1}	$e^{0.48} = 1.61$	$(e^{0.29}, e^{0.67}) = (1.34,\ 1.95)$



Probability estimates

Since the log-odds is a linear function

$$\ln \frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} = \mathbf{x}_i \boldsymbol{\beta}$$

the corresponding probability of success becomes

$$p_i = \frac{\mathrm{e}^{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}}}{1 + \mathrm{e}^{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}}} = \frac{\mathrm{e}^{\mathbf{x}_i \boldsymbol{\beta}}}{1 + \mathrm{e}^{\mathbf{x}_i \boldsymbol{\beta}}}$$

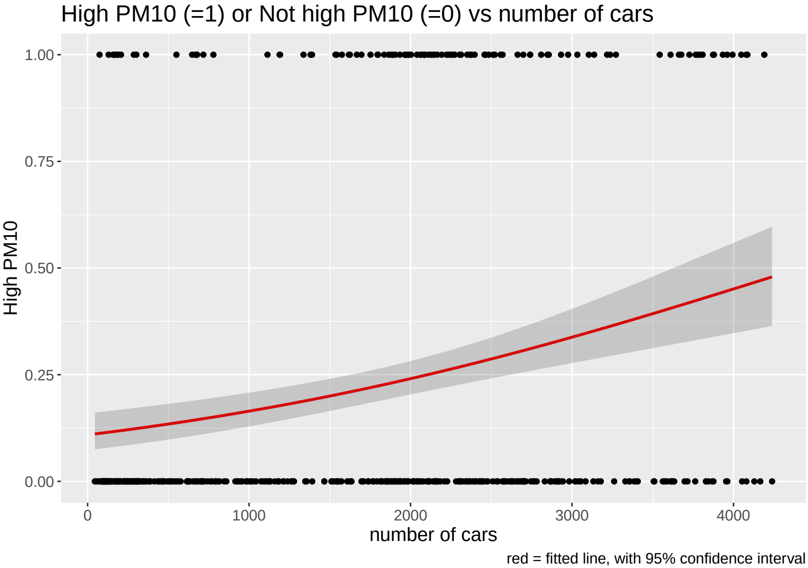
which is a non-linear function of the β-parameters.
Since $\mathbf{x}_i \hat{\boldsymbol{\beta}}$ is a linear function of (dependent, approx.) normally distributed β-estimates we can construct confidence intervals for the log odds:

$$I_{\mathbf{x}_i \boldsymbol{\beta}} = (\mathbf{x}_i \hat{\boldsymbol{\beta}} \pm \lambda_{\alpha/2} \cdot d(\mathbf{x}_i \hat{\boldsymbol{\beta}}))$$

Since $\hat{p}_i$ is a monotonous, increasing, function of $\mathbf{x}_i \hat{\boldsymbol{\beta}}$ we get

$$I_{p_i} = \frac{\mathrm{e}^{I_{\mathbf{x}_i \boldsymbol{\beta}}}}{1 + \mathrm{e}^{I_{\mathbf{x}_i \boldsymbol{\beta}}}}$$

which always lies in $[0, 1]$!



Prediction interval? The observations will always be either 0 or 1 so we will need other methods than intervals here.