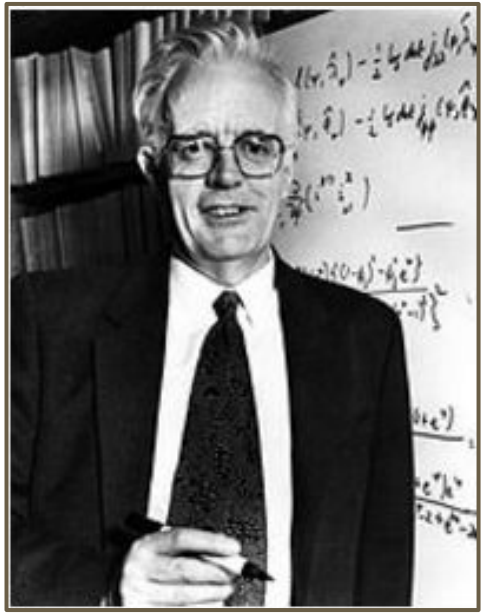
Logistic Regression

Í þessum fyrirlestri:

- Saga Logistic Regression
- Kynning á sýnidæmi
- Logistic Regression?
- Afhverju ekki einhver önnur aðferð?
- Líkur, gagnlíkindi og odds ratio
- Logit
- Estimated regression equation
- Probit vs. logit
- Fastarnir í Estimated regression equation
- MLE
- Multiple logistic regression
- Multinomial logistic regression
- Ordered logistic regression
- Demo í Jupyter

Sagan

- Logistic regression var þróað af tölfræðingnum David Cox árið 1958
- Notað í vélrænu gagnanámi (e. Machine Learning), Læknisfræði, Félagsvísindum.



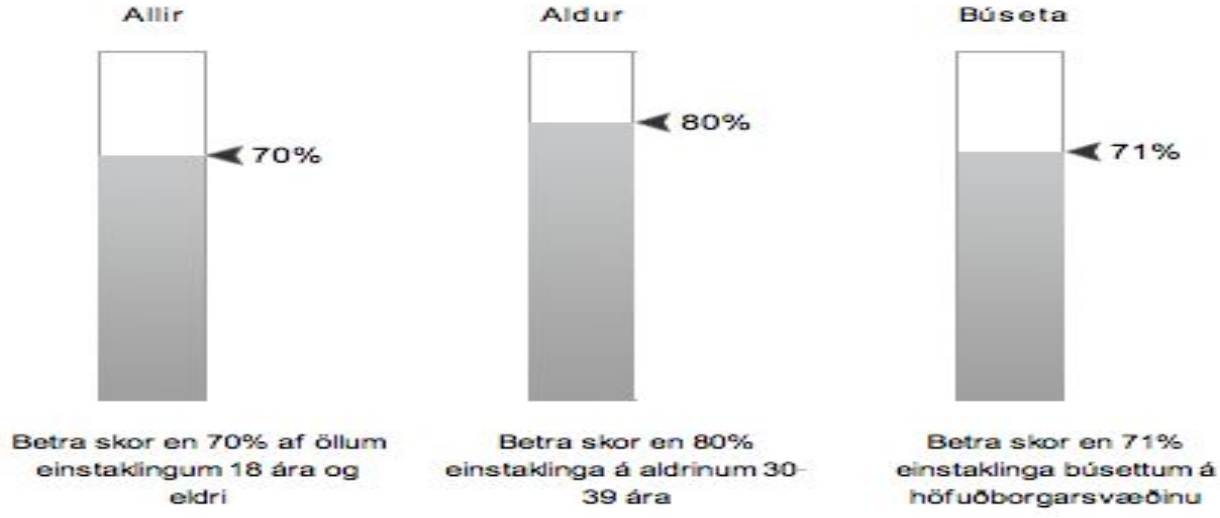
Dæmi - Get ég fengið lán?

- Þegar taka skal lán í fyrsta skipti þarf að huga að því hvernig fjárhagsleg staða einstaklings er.
- Við viljum reikna út lánshæfieinkunn sem ákvarðar hvort að umsókn um lán verði samþykkt eða ekki samþykkt.

Stig	Áhættuflokkur	Líkur á skráningu á vanskilaskrá
351	B1	0,53% innan 12 mánaða

Staðsetning einstaklings á áhættuskala

Samanburður við einstaklinga 18 ára og eldri með lögheimili á Íslandi.

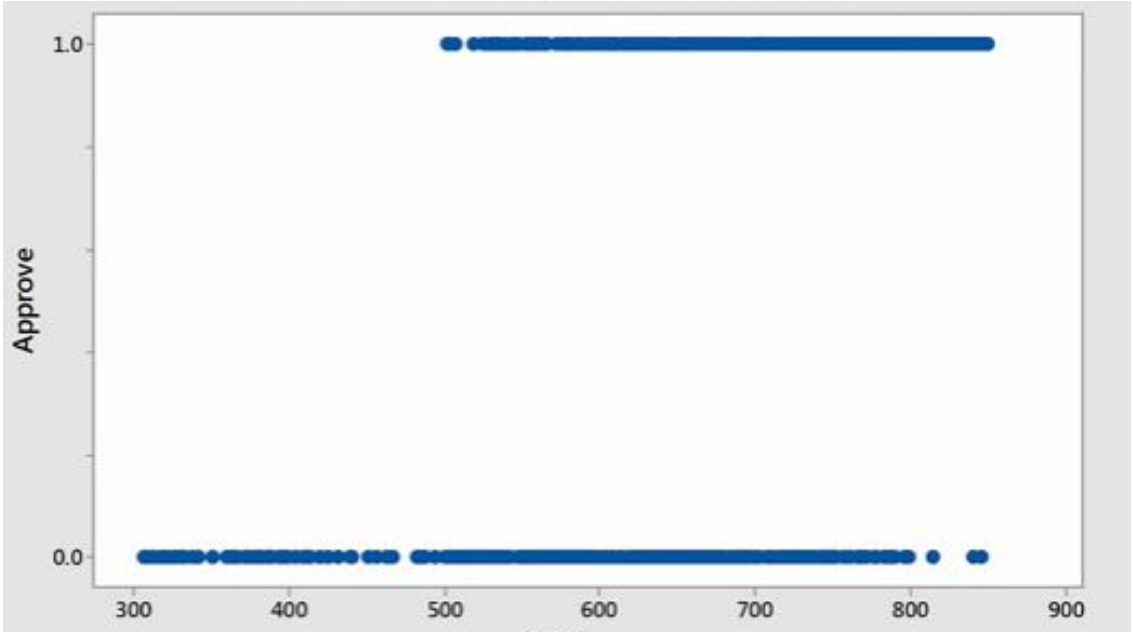


Dæmi - Get ég fengið lán?

- Skoðum lánshæfiseinkunnir á skalanum 300-850.
- Gerum ráð fyrir að okkar einkunn sé 720.
- Þessi einkunn er mikilvæg í ákvörðuninni um það hvort að umsóknin okkar um lán verði samþykkt eða ekki.
- Skoðum graf sem sýnir gögn 1000 umsækjenda þar sem kemur fram einkunn og hvort lánabeiðni var samþykkt eða ekki samþykkt.

Dæmi - Get ég fengið lán?

- Hvernig ætlum við að finna bestu línuna?



Dæmi - Get ég fengið lán?

- Með því að nota þessi gögn viljum við:
  - Búa til model sem segir okkur til um líkur (e. probability) og gagnlíkindi (e. odds) um það að vera samþykkt fyrir allar einkunnir.
  - Finna um það bil þá einkunn þegar líkurnar eru jafnar eða 50% til þess að fá lán samþykkt.
  - Setja okkar einkunn inn í model-ið og ákvarða probability og odds um að fá samþykkt lán.

Hvað er Logistic Regression

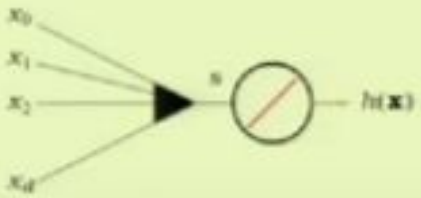
- Logistic Regression gerir líkan (e. model) um líkurnar (e. probability) á að einhver atburður eigi sér stað út frá óháðum gildum (e. independent variables).
- Áætla líkurnar á því að einhver atburður gerist út frá handahófskenndum athugunum á móti líkunum á að atburðurinn gerist ekki.
- Athuganir eru flokkaðar með því að meta hvort tiltekið gildi sé í flokknum “samþykkt” eða “ekki samþykkt” (já eða nei flokkar)

Afhverju ekki önnur Regression aðferð?

- Linear regression
  - Binary gögn hafa ekki normal dreifingu sem er skilyrði fyrir flestar aðrar tegundir af regression
  - Gildi breytunnar á y-ás getur verið meiri en 0 og 1 sem brýtur reglu um skilgreiningu á líkum þar sem þær eru aðeins á skalanum 0-1
- Simple linear regression
  - Magnbundinn breyta sem ákvarðar næstu breytu - þannig er það ekki með binary gildi. Þau eru aðeins 0 eða 1
- Multiple regression
  - Er eins og simple linear regression nema með fleiri óháðum breytum
- Nonlinear regression
  - Tvær breytur en gögnin eru flokkast ekki í tvo flokka

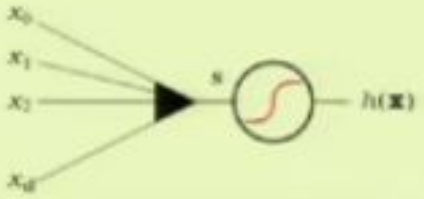
linear regression

$$h(\mathbf{x}) = s$$



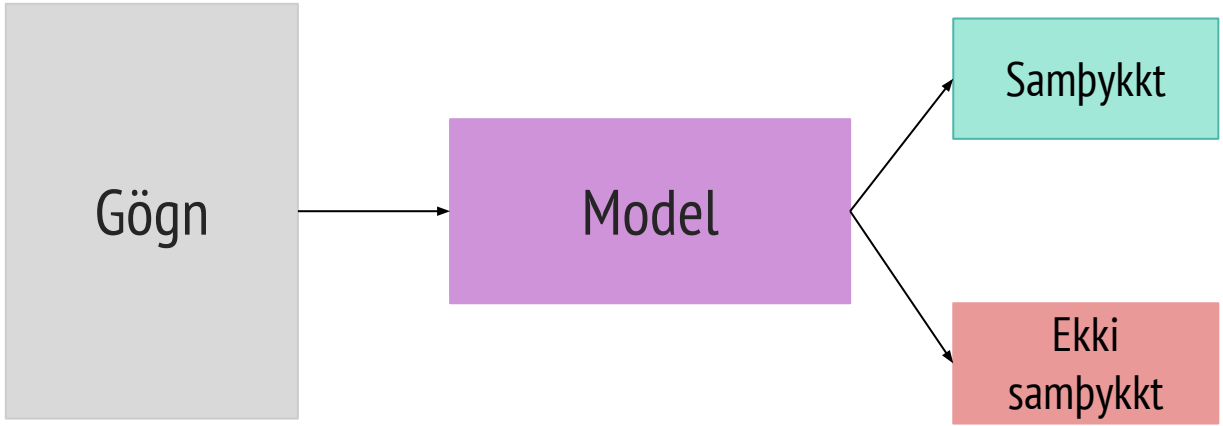
logistic regression

$$h(\mathbf{x}) = \theta(s)$$



Dæmi - Get ég fengið lán?

- Við viljum setja einkunnina í líkanið og fá líkurnar á því hvort við fáum “Já” eða “Nei”
- Viljum skilja þetta líkan: hvað verður til þess að eitthvað sé samþykkt?



Líkur (e. probability) og gagnlíkindi (e. odds)

- Táknum líkurnar með P og gagnlíkindi með O

P = Fjöldi mögulegra möguleika / allar mögulegar útkomur

O = líkurnar á að eitthvað gerist / líkurnar á að eitthvað gerist ekki

$$O = \frac{P}{1-P}$$

Ördæmi um líkur og gagnlíkindi

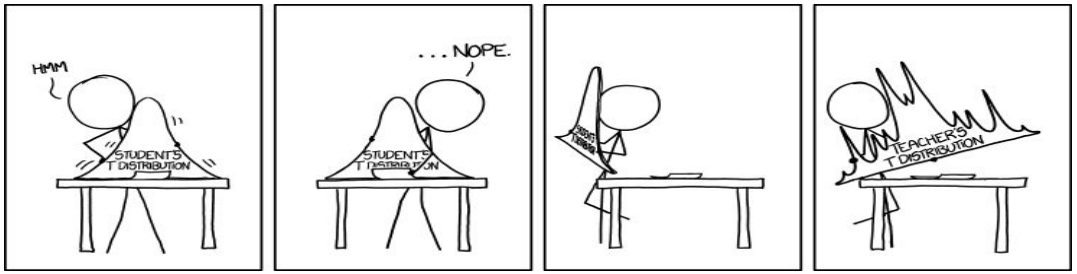
- Líkurnar á því að við fáum skjaldamerki
  - P(skjaldamerki) = ½ = 0.5 eða 50% líkur
- Gagnlíkindi
  - O = 0.5 / 1 - 0.5 = 1 (Odds are even / 1 á móti 1)
- Líkurnar á að fá 1 eða 2 á tening
  - P(1 or 2) = 2/6 = ⅓
- Gagnlíkindi
  - O = 0.333 / 0.666 = 0.5

Gagnlíkindahlutfall (e. odds ratio)

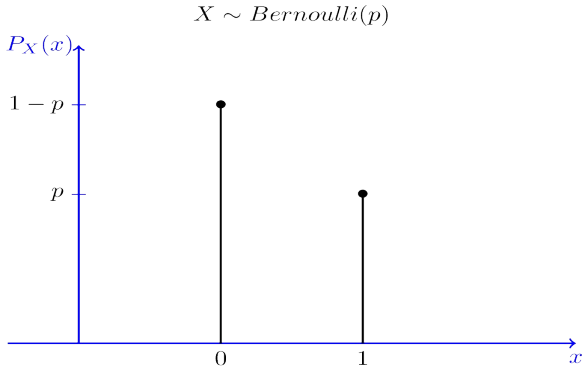
- Gagnlíkindahlutfall í Logistic Regression fyrir breytu sýnir hvernig gagnlíkindin breytast ef óháð gildi (x-gildið) er hækkuð um 1

$$Odds\ ratio = \frac{P_1}{P_0} = \frac{\frac{P_1}{1-P_1}}{\frac{P_0}{1-P_0}}$$

Bernoulli dreifing

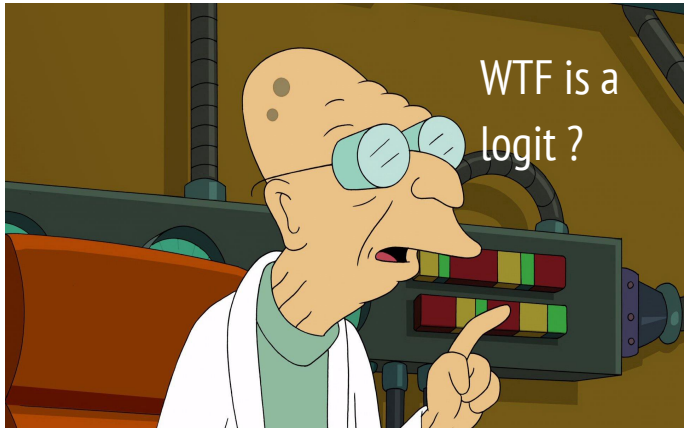


- Háð gildi (e. depentent variables) í Logistic Regression fylgja Bernoulli dreifingunni með óþekktar líkur p.
- Bernoulli dreifingin er sérstakt tilvik af **Binomial dreifingu** þar sem n = 1 (bara ein tilraun)
- Í LR erum við að meta óþekkt gildi P fyrir línlega samsetningu á óháðu gildunum.
- Þess vegna viljum við tengja saman óháðu breytunar til þess að líkja eftir Bernoulli distribution
- Þessi tenging er kölluð **logit**

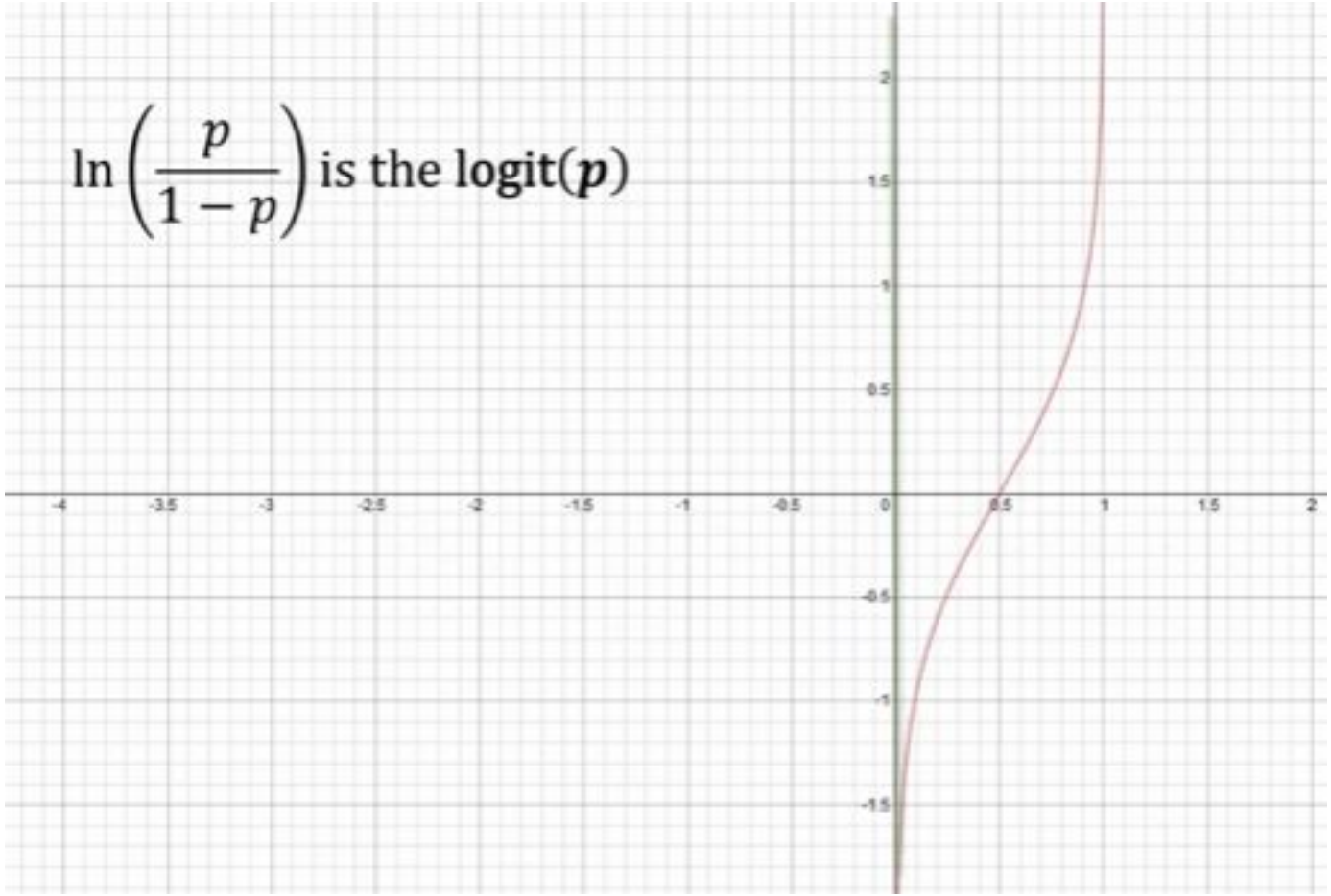


logit

- Í LR vitum við ekki líkurnar P eins og við gerum í Binomial (Bernoulli) verkefnum, þar sem við getum t.d reiknað út líkurnar á að fá 2 skjaldamerki í 5 peningaköstum.
- **Markmiðið** með LR er að meta líkurnar P með því að nota línulega samsetningu á óháðu breytunum.
- Til þess að tengja saman *línulega samsetningu á breytunum* og kjarna *Bernoulli dreifingarinnar* þurfum við fall sem tengir þess tvo hluti saman.
- Náttúrulegi logrinn af gagnlíkindahlutfallinu, **the logit**, er fallið sem tengir þessa tvo hluti saman.

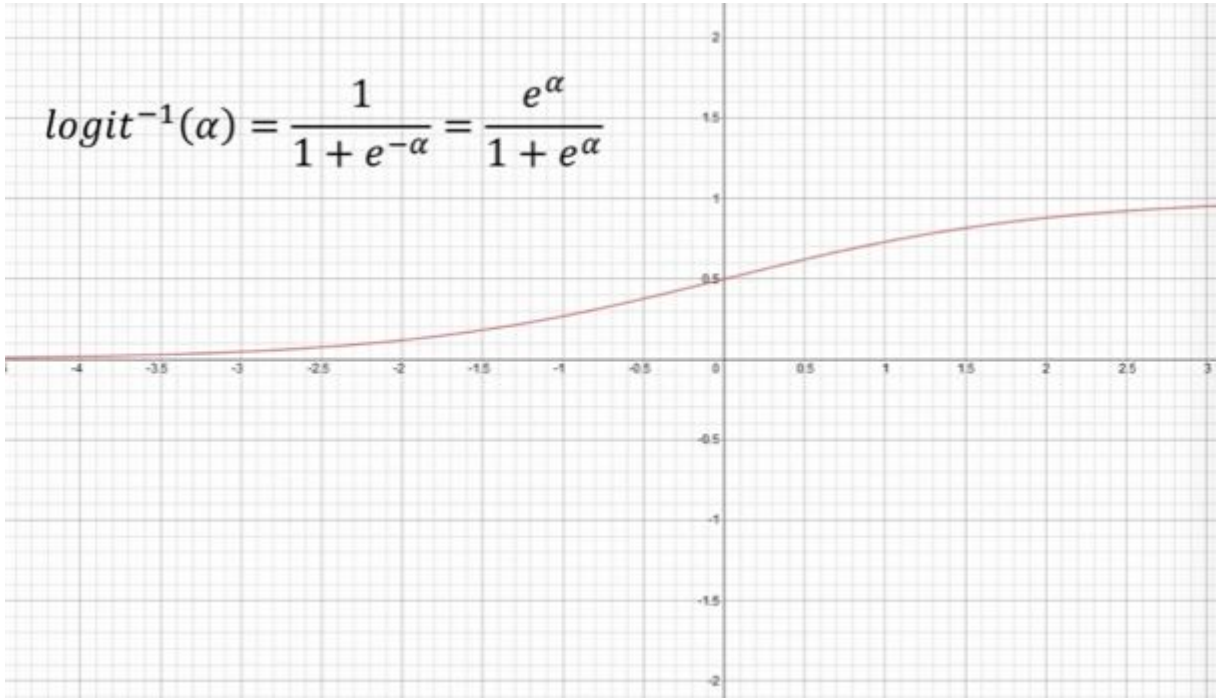


$$\ln\left(\frac{p}{1-p}\right) \text{ is the logit}(\boldsymbol{p})$$



Inverse logit

$$logit^{-1}(\alpha) = \frac{1}{1+e^{-\alpha}} = \frac{e^{\alpha}}{1+e^{\alpha}}$$



Estimated Regression Equation

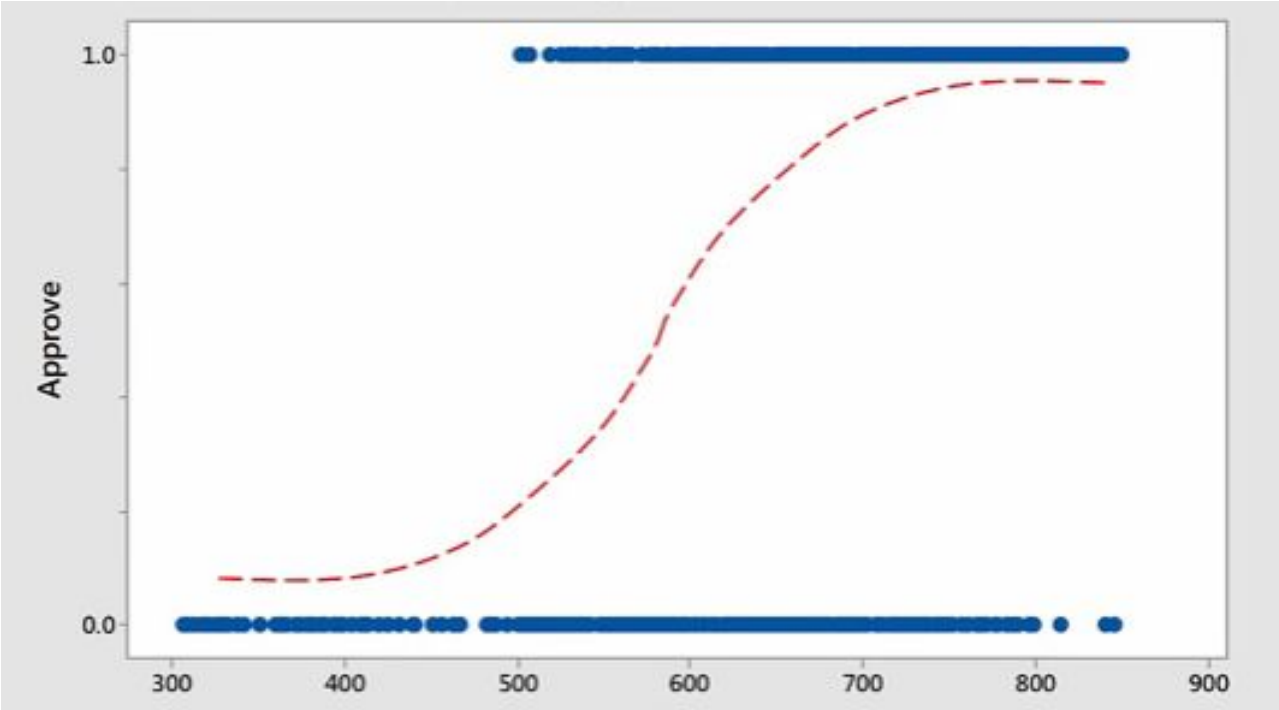
$$logit(P) = ln(\frac{P}{1-P}) = \beta_0 + \beta_1 x_1$$

Þá fæst að estimated regression equation er

$$\hat{P} = \frac{e^{\beta_0 + \beta_1 x_1}}{1 + e^{\beta_0 + \beta_1 x_1}}$$

Estimated Probability

Dæmi - Get ég fengið lán?



Dæmi - Get ég fengið lán?

- Við erum með 1000 færslur eða 1000 einkunnir um lánshæfismat á skalanum 300-850 þar sem kemur fram hvort einkunn var samþykkt = “1” eða ekki samþykkt = “2”
- Við viljum reikna út áætlaðar líkur út frá okkar einkunn sem er 720
- Við reiknum út fastana út frá þeim gögnum sem við erum með með hjálp Maxium likelihood estimation eða MLE (útskýrt síðar)

$$\beta_0 = -9.346 \; og \; \beta_1 = 0.014634$$

$$\hat{P} = \frac{e^{\beta_0 + \beta_1 x_1}}{1 + e^{\beta_0 + \beta_1 x_1}}$$

- Setjum þessa fasta inn í jöfnuna
- x gildið er einkunnin okkar

Dæmi - Get ég fengið lán?

$$\hat{P} = \frac{e^{-9.346 + 0.014634 * 720}}{1 + e^{-9.346 + 0.014635 * 720}}$$

$$= \frac{3.289}{1 + 3.289} = \frac{3.289}{4.289} = 0.7668$$

- 76.68% líkur

$$O = \frac{0.7668}{1 - 0.7668} = 3.289$$

Dæmi - Get ég fengið lán?

- Hvaða áhrif hefur það á líkurnar ef einkunnin hækkar um **1 stig**?
- Ef við setjum 721 inn í jöfnuna fáum við að:

$$\hat{P} = \frac{e^{-9.346 + 0.014634 * 721}}{1 + e^{-9.346 + 0.014635 * 721}}$$

$$= \frac{3.337}{1 + 3.337} = 0.7694$$

og að:

$$O = \frac{0.7694}{1 - 0.7694} = 3.337$$

- Þá fæst að gagnlíkindahlutfallið (e. Odds ratio) er:

3.337 / 3.289 = **1.0146**

- Sem er breyting á gagnlíkindunum (e. Odds) um eitt stig
- Getum reiknað út breytingar fyrir mismumandi hlutfall stiga
- Ef við hækkum einkunnina okkar um 20 stig eru 81.46% líkur á að við fáum lán

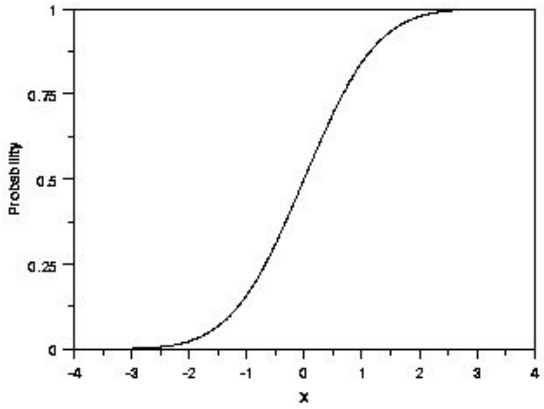
Getum notað fastana til að meta líkurnar

- Hvaða einkunn gefur helmingslíkur eða 50/50 líkur?

$$
\begin{aligned}
ln(\tfrac{0.5}{0.5}) &= -9.346 + 0.014634x_1 \\
\Leftrightarrow\ & -9.346 + 0.014634x_1 = 0 \\
\Leftrightarrow\ & x_1 \approx 638.65
\end{aligned}
$$

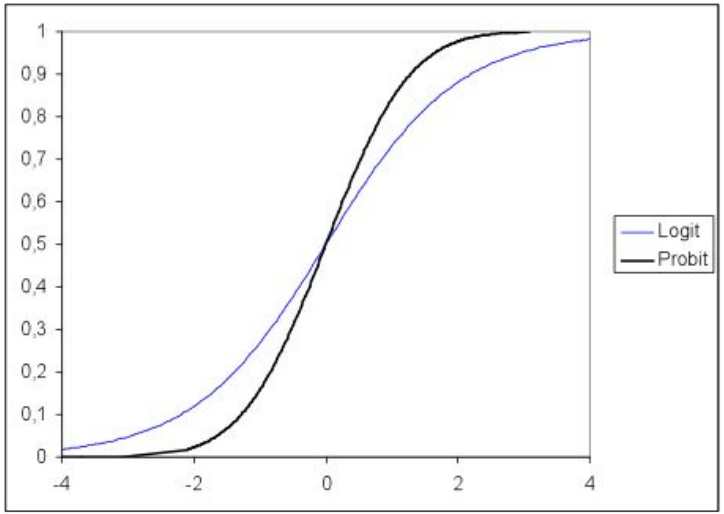
Probit - er það eins og logit?

- logit og **probit** líkön koma til greina þegar við erum að svara spurningu eins og já/nei, sammála/ósammála.
- Probit líkanið notar það sem kallast cumulative distribution function sem tengist standard normal distribution til að skilgreina f(*)
- Bæði föllin taka inn gildi og setja það inn á grafið þannig að það fellur á milli 0 og 1.
- Hvaða fall sem er sem skilar gildi á milli 0 og 1 kemur til greina. Hinsvegar er dýpra fræðilegra líkan sem er bakvið tjöldin hjá logit og probit sem gerir kröfu á að fallið sé byggt á líkindadreifingu (e. probability distribution).
- Munurinn liggur í ólíkum föllum sem eru notuð til þess að reikna út áætlaðar líkur og geta niðurstöðurnar því verið ólíkar.



Hvort er logit eða probit betra?

- Aðferðirnar virðast mjög svipaðar (þó ekki alveg eins).
- Logit, sem þekkist einnig sem logistic regression er vinsælla í heilbrigðisvísindum eins og faraldsfræði.
- Probit model tekur tillit til misdreifni í gögnum (e. heteroskedastic) sem tengist oft inn í hagfræði eða stjórmálafræði.
- Ef hvorugt af þessum skilyrðum eru til staðar þá skiptir ekki máli hvaða aðferð er notuð.



Fastarnir í Logistic Rregression (e.coefficients)

- Fáum þá með því að nota reiknirit úr vélrænu gagnanámi.
- Fastarnir eiga að tákna breytingarnar í logit modelinu útfrá hverri breytingu í óháðu breytunum í þessu ákveðna gagnasetti.
- Notum Maximum likelihood estimation reiknirit.

Maximum Likelihood Estimation

- Vinsælt reiknirit í vélrænu gagnanámi.
- Fastarnir (Beta gildin) í lógískri aðhvarfsgreiningu þurfa að vera metnir út frá æfingar-gögnunum.
- Finnur Beta0 ássniðið (e. intercept) og Beta1 hallatöluna með „best fit“ aðferðum.

Yfirlit: MLE reikniritið

- Ítrar yfir gildin oft.
- Notar aðferðir til að finna minnstu kvaðrarót, t.d. með *Newtons method*.
- Notar lágmörkunar reiknirit til að ákvarða bestu gildin.

Afhverju LR og áræðanleiki

- Notað til að spá fyrir um hverjar séu líkurnar á að atburður gæti gerst. (eða ekki gerst)
- Leyfir manni að sjá hvort einhver áhættu faktor (e. risk factor) hafi áhrif á líkurnar um einhver ákveðin prósent.

Multiple logistic regression

- Höfum bara séð dæmi með einni tölu breytu og einni tvíundar breytu.

$$\operatorname{logit}(p(x)) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

$$p(x_1, \ldots, x_p) = \frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}}$$

Multinomial logistic regression

- Líka kallað multi-class classification
- Úttaks-breytum fjölgar.
- Notað ef háðar breytur eru 3 eða fleiri flokkar.
- Svarar spurningum:
  - Hvaða fag mun háskólanemi velja, gefið t.d. hvaða einkunnir þeir eru með.
  - Hvaða blóðflokk er persóna í gefið að við höfum læknis-upplýsingar um hann.
  - Hvaða stjórnmálaflokk mun persóna kjósa út frá upplýsingum um hann.

Ordered logistic regression

- Líka kallað Proportional odds model
- Notað ef það eru raðaðir flokkar.
- T.d. mjög óánægður, óánægður, alveg sama, ánægður, mjög ánægður.
- Svarar spurningum eins og:
    - Hversu vel gætum við séð fyrir hvaða möguleika fólk myndi velja í könnun þar sem svarmöguleikarnir væru mjög óánægður, óánægður, alveg sama, ánægður, mjög ánægður. Gefið að við höfum svör við öðrum spurningum.

Demo í jupyter

https://github.com/sigurdurb/Logistic-Regression