Regression for count data

Poisson regression and other models for counts

N. Torelli, G. Di Credico, V. Gioia

Fall 2020

University of Trieste

Introduction

Count variables

Let us now consider the case of a response variable Y that is a count.

The variable Y can take on values of 0, 1, 2 and so on.

Relevant examples of count dependent variables are:

- the number of car accidents
- the number of phone calls over a day for an assistance service
- the number of items bought on a sales portal
- the number of cases of a disease in a given territory
- the number of those accessing a web site

Usually the counts refer to all events occurred within a specified time interval (or a space interval)

Counts as response variable in a statistical model

- Also in this case we want to build a statistical model and the goal is to predict the number of events.
- We try to explain/predict the counts y_i by using observed characteristics of the i–th unit (such as age, sex, education, income, etc..) .
- We expect that the distribution of the counts varies as the other covariates vary and possibly we can try to summarize how the distribution of the counts varies by describing how the mean number of the counts vary.
- It is important to remind that there are more models that can be used for the distribution of a count variable.
- The simplest model, and possibly the best known, for this specific case is the Poisson one.

Poisson regression

The Poisson distribution

- A count variable Y is a variable that takes on 0 or any positive integer number, i.e., $0, 1, 2, \ldots$
- Its probability function is the sequence of probabilities $Pr(Y = 0), Pr(Y = 1), Pr(Y = 2), \ldots$
- It could be represented in a table like this

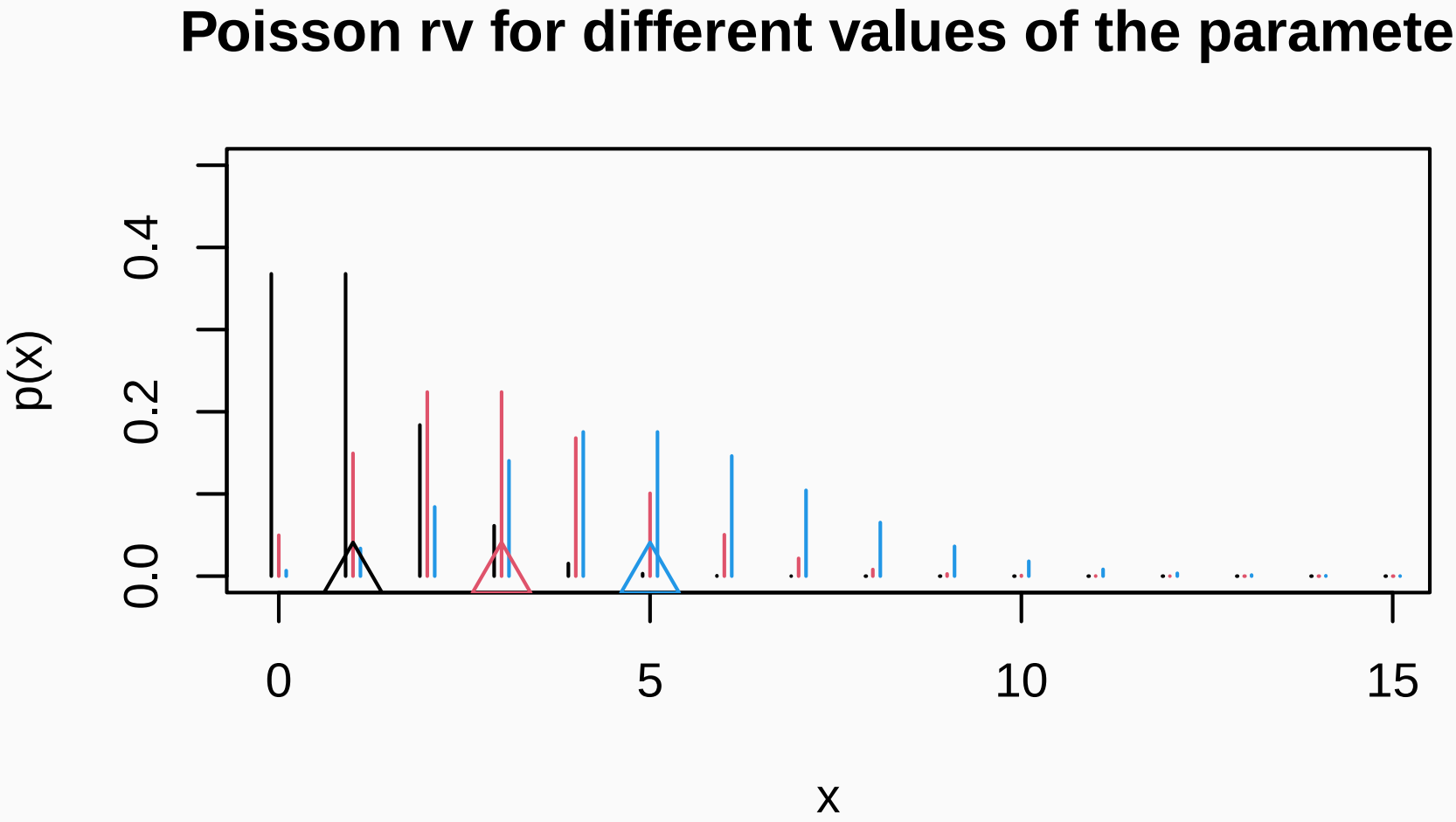
y	$Pr(Y = y)$
0	$e^{-\mu}$
1	$\mu e^{-\mu}$
2	$\frac{\mu^2 e^{-\mu}}{2!}$
3	$\frac{\mu^3 e^{-\mu}}{3!}$
$\ldots$	$\ldots$

- The Poisson random variable then defines the probabilities of any value $Y = 0, 1, 2, 3, \ldots$ as follows

$$Pr(Y = y) = \frac{\mu^y e^{-\mu}}{y!}$$

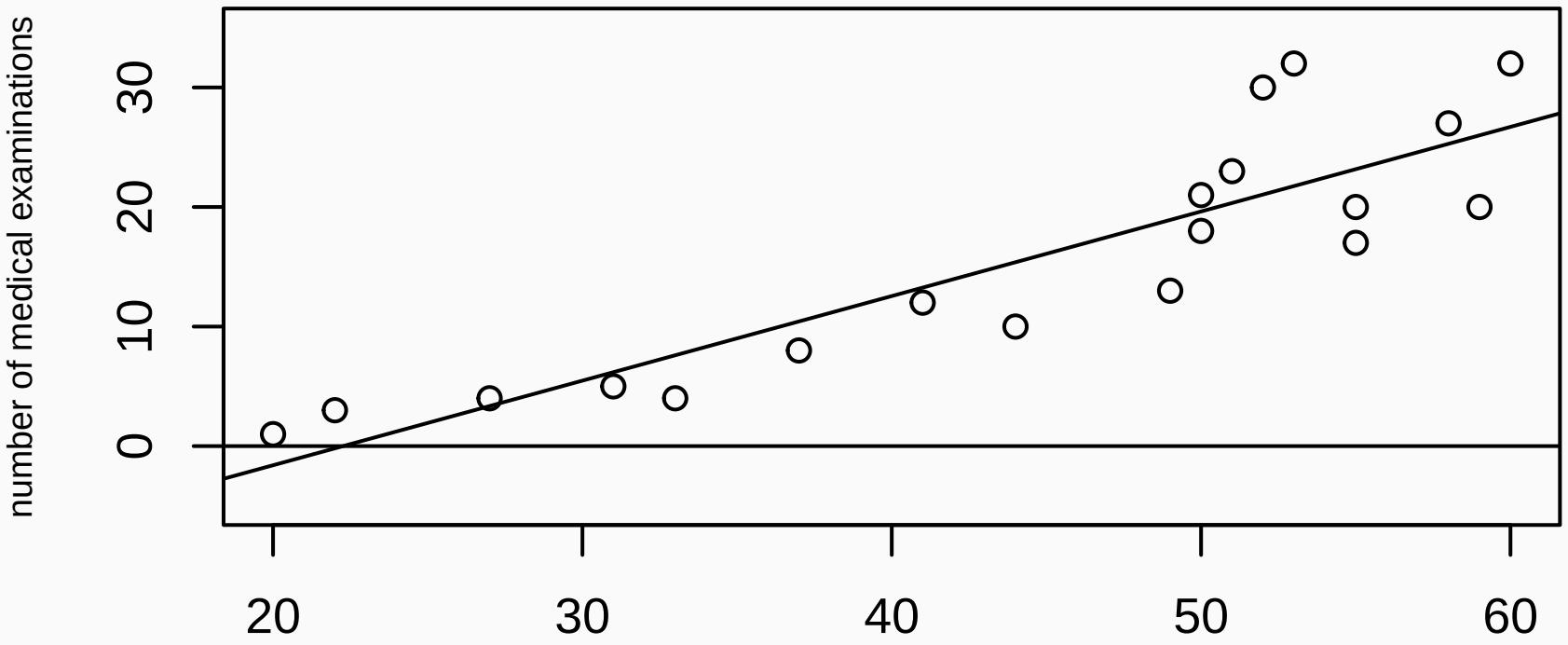
- the parameter μ is the expected value of the counts Y distributed according to a Poisson random variable. μ is greater than 0

The Poisson distribution: some examples



The graph shows the Poisson distribution for different values of the parameter, i.e., the mean ($\mu = 1$ black, $\mu = 3$ red, $\mu = 5$ blue)

Count data: an example with medical examinations



- Y_i (number of examinations) can be assumed Poisson: $Y_i \sim \text{Poisson}(\mu_i)$.
- We can assume that $\mu_i = h(x)$, i.e., it is a function of the covariate x.
- A linear specification is clearly inappropriate (also because it will predict negative values). We should choose among functions that $h(\cdot) \to [0, \infty)$.

Poisson regression: Basic framework

- We assume, like in other regression models, that for any value of the covariate X the curve represents the mean μ of the dependent variable Y.
- In the example, we assume that for a given age x_i the number of medical examinations has a Poisson distribution whose parameter (mean) is μ_i
- The mean μ_i of the variable Y_i is assumed to lie on a curve (a function) like the one depicted in the previous slide.
- More specifically, we assume:

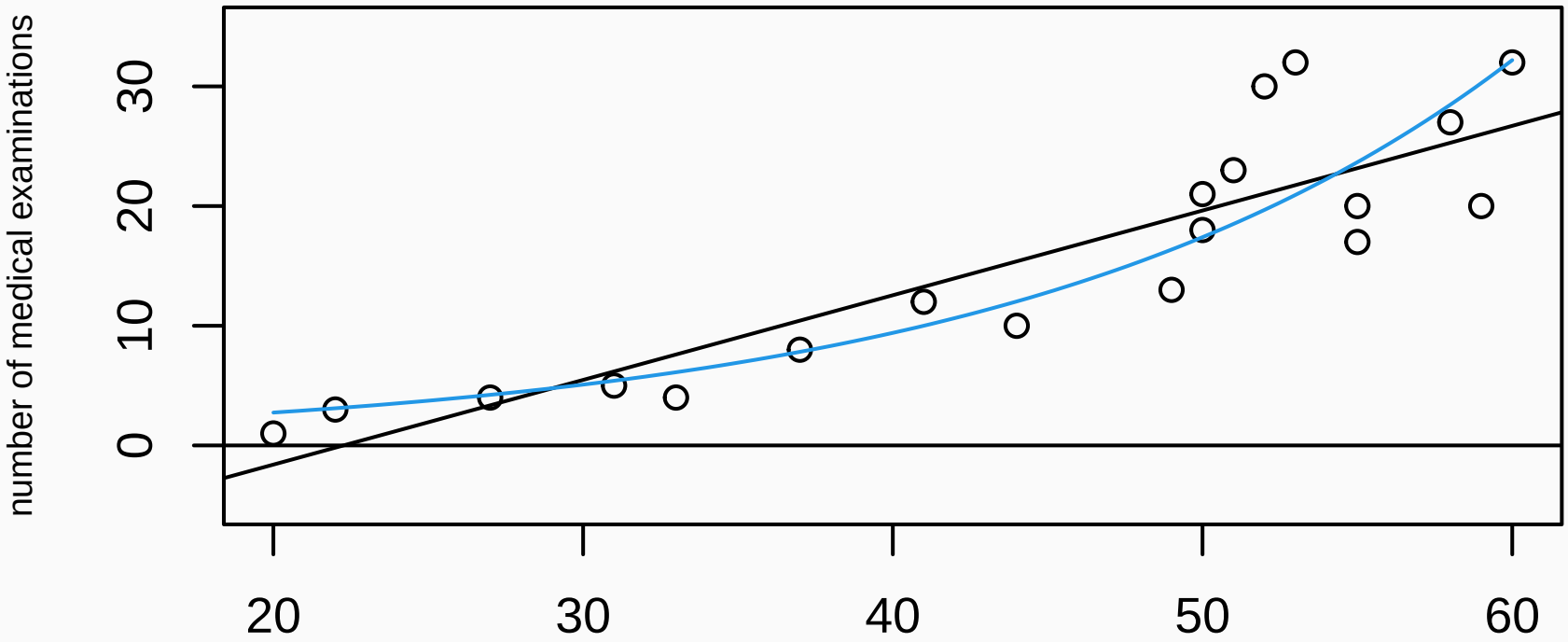
$$\mu_i = E(y_i) = e^{\beta_0 + \beta_1 x_i}$$

- The linear component $\beta_0 + \beta_1 x_i$ is transformed by the exponential function and as a result it takes on only positive values.
- Remind that the mean of a count variable can be only positive
- This will define the Poisson Regression

Poisson regression: Interpretation of the parameters

- The linear specification within the exponential function ease interpretation.
- For this simple model the most interesting parameter is β_1 that is associated to the covariate X.
- $exp(\beta_1)$ can be interpreted as the proportional change in the mean correponding to a unit change in X. We can multiply it by 100 and subtract it form 100 to interpret it as the percentage variation in Y.
- In the example presented above we obtained $\beta_0 = -0.220$ and $\beta_1 = 0.062$. Then exp(0.062)=1.063, it means that the model predicts that if we add one year to age the mean number of medical examinations raises of about 6.3 percent.

Count data: an example with medical examinations



This new curve is more appropriate for the mean μ of the number of examinations as a function of age.

Poisson regression: Estimation of the parameters

- The non linear specification adopted makes a bit more difficult to detect the curve that approximate the data points. There are many possible criteria to find it.
- Since we have postulated a data generating mechanism based on the Poisson distribution we can again use the maximum likelihood method.
- If we observe a random sample of data (y_i, x_i) than we can evaluate the probability of obtaining those data.
- More specifically, it is assumed that each data point is drawn from a Poisson distribution with mean $\mu_i = e^{\beta_0 + \beta_1 x_i}$.

Then using independence assumption (implied by random sampling) we can evaluate the probability $L(\beta_0, \beta_1)$ of observing that specific set of data points for each possible couple (β_0, β_1)

- The maximum likelihood estimate is the couple $(\hat{\beta}_0, \hat{\beta}_1)$ that corresponds to the highest value of $L(\beta_0, \beta_1)$
- locating the maximum likelihood estimates $(\hat{\beta}_0, \hat{\beta}_1)$ is not strightforward and requires the use of an iterative procedure.

Multiple Poisson regression

- The Poisson regression model defined in the example is very simple. It can be easily extended to include more covariates (quantitative variables or qualitative factors).
- A more general specification for the model is then + Y_i has a Poisson distribution with mean μ_i + $\mu_i = e^{\beta_0+\beta_{i1}x_{i1}+\beta_{i2}x_{i2}+\cdots+\beta_{ip}x_{ip}}$

- This is a log–linear model since a linear regression model is assumed for the logarithm of μ_i
  $log(\mu_i) = \beta_0 + \beta_{i1}x_{i1} + \beta_{i2}x_{i2} + \cdots + \beta_{ip}x_{ip} = \boldsymbol{x}_i^T\boldsymbol{\beta}$

Inference

Poisson regression: parameters estimates

- The Poisson assumption for Y_i allows us to use maximum likelihhod for parameters estimation.
- log-likelihood, assuming random sampling, is:

$$log(L(\beta)) = \ell(\beta) = \sum_{i=1}^{n}(y_i log(\mu_i) - \mu_i)$$

the constant $-nlog(y_i!)$ is omitted since it does not depends on β

- using the link function $log(\mu_i) = \eta_i$ we obtain

$$\ell(\beta) = \sum_{i=1}^{n} \ell_i(\beta) = \sum_{i=1}^{n} y_i \boldsymbol{x}_i^T \beta - exp(\boldsymbol{x}_i^T \beta) = \sum_{i=1}^{n}(y_i \eta_i - exp(\eta_i))$$

we can evaluate the score function to obtain the likelihood equations

$$s(\beta) = \sum_{i=1}^{n} \frac{\partial l_i(\beta)}{\partial \beta} = \sum_{i=1}^{n} \boldsymbol{x}_i(y_i - exp(\eta_i)) = \sum_{i=1}^{n} \boldsymbol{x}_i(y_i - \mu_i) \quad \text{and equate it to 0} \quad s(\beta) =$$

expected Fisher information $i(\beta)$ can be also obtained
$i(\beta) = E(s(\beta)s(\beta)^T) = \sum_{1=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T \mu_i$

- solving likelihhod equations is not straightforward and numeriacal solutions are necessary (e.g. Newton-Raphson)

Inference in Poisson regression: testing significance of single β_s

- Once the parameters of the model are estimated we are interested in deciding if a given covariate is relevant or not to predict the response variable.
- Also for this model, for large samples, maximum likelihood method provides also good estimates of the standard errors of the βs.
- We can then evaluate if a given β_j associated to the covariate X_j is large enough by looking at the ratio $\frac{\hat{\beta}_j}{s.e.(\hat{\beta}_j)}$.
- If the absolute value of this ratio is large there is evidence that the variable X_j affects the mean of Y. For large samples, the ratios above are well approximated by a standard Gaussian distribution when the parameter is actually 0. This allow us to judge if the ratio is large enough.
- The rule of thumb is that large means greater than 2. But it is always a good idea to look at p-values associated to the estimated parameters.

Inference for Poisson regression models: Judging the overall performance of the model

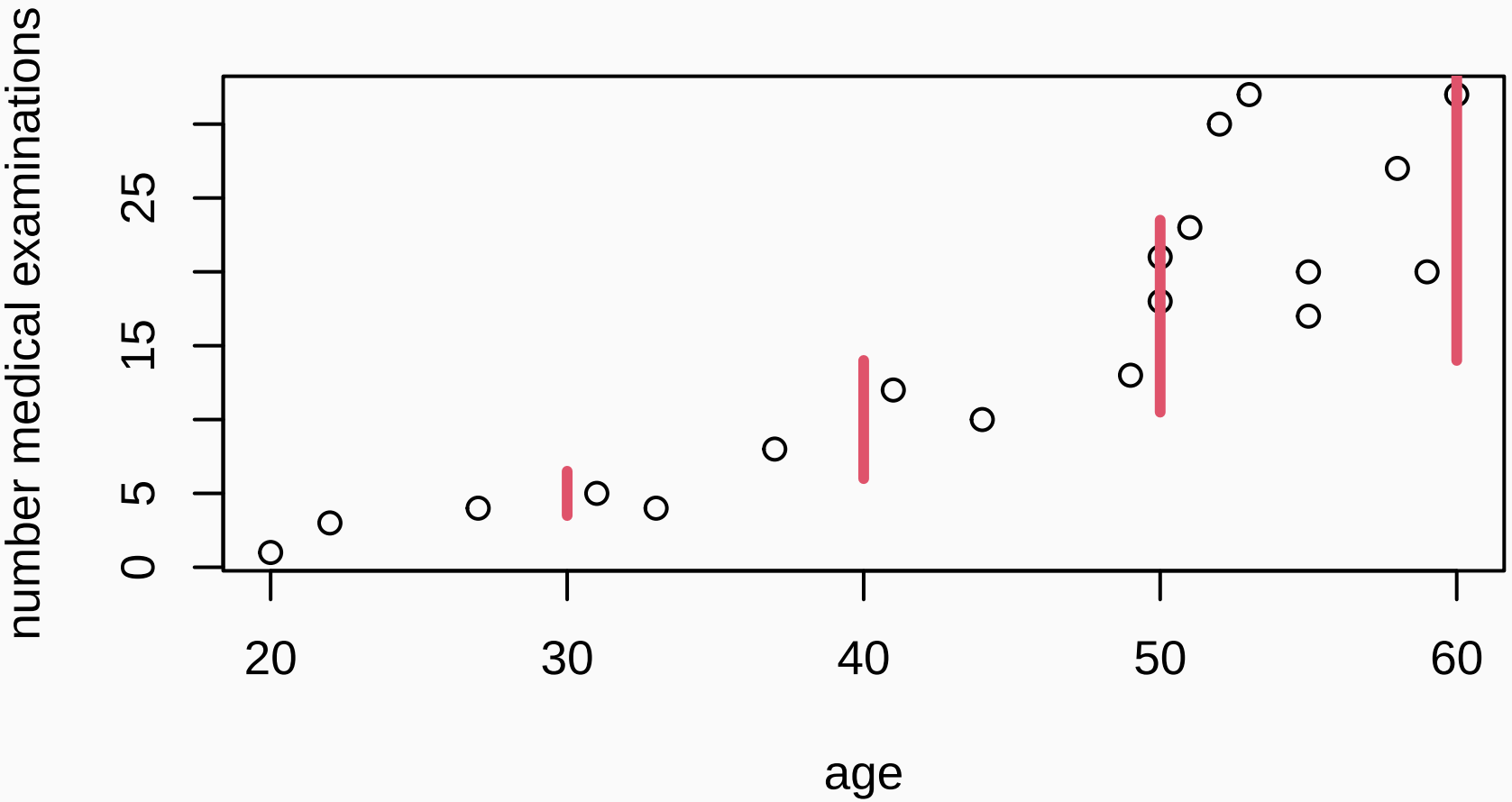
- Also in this case, just like in the logistic regression, one can measure the difference between the value of the likelihood for the estimated parameters $L_{\hat{\beta}} = L(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$ and the value of the likelihood we would obtain when considering:
  - as many parameters as the available data (that would give a perfect fit) L_{max}
  - the null model with only the intercept β_0, i.e., L_0
  - an alternative model that is equal to the one estimated but with some parameters set to 0 ($L(\hat{\beta}_R = L(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k, 0, \ldots, 0, 0))$. In the last expression the last $p - k$ parameters are set to
- Comparing those likelihoods (or their logarithm) helps to judge the quality of the model

Inference for Poisson regression models: Judging the overall performance of the model

- The difference between $logL_{\hat{\beta}}$ and $logL_0$ give a first indication: if the latter difference is small then the model is not supported by the data
- The difference between the $logL_{max}$ and $logL_{\hat{\beta}}$ should be small for good models.
- Twice this difference is called the deviance and is defined as $D_{\hat{\beta}} = 2(logL_{max} - logL_{\hat{\beta}})$.
- A way to compare two models is to compare their deviances. We could compare the deviance $D_{\hat{\beta}_R}$ where the likelihood is evaluated for a reduced model $L(\hat{\beta})_R$.
- If the difference between the two deviances is small we keep the simpler model.
- To decide when "small" is small enough we can use statistical criteria (comparing the difference with the value of a appropriate χ^2 distribution with $p - k$ degrees of freedom)

Overdispersion

Again the example with medical examinations



The vertical red strips illustrate how dispersion of the number of medical examinations increases with the mean. Can the model accommodate this?

Overdispersed count data

- Poisson regression models in a GLM context imply that the variance function then functionally related to the mean function (it is actually the same).
- In fact, a striking characteristic of a Poisson distribution is that its mean is equal to its variance.
- A Poisson model states that the mean of our response variable varies according to the model. This implies that the variance of Y_i varies accordingly.
- The Poisson regression model implicitly introduces a form of heteroschedasticity.
- This characteristic makes the Poisson model very peculiar and in many cases reduces its flexibility and its ability to describe real situations.
- A simple way to check appropriateness of the model is to verify if data reflect the specific requirement mean(Y_i)= variance(Y_i)
- Considering a model for counts with overdispersion will be more realistic in many practical cases

Poisson regression: Residual checks

- In the Normal linear regression models residuals checks are a powerful tool for assessing model adequacy
- We can evaluate residuals also for Poisson regression
- First we can predict μ by using our model. More specifically:

$$\hat{\mu}_i = e^{\hat{\beta}_0 + \hat{\beta}_{i1} x_{i1} + \hat{\beta}_{i2} x_{i2} + \cdots + \hat{\beta}_{ip} x_{ip}}$$

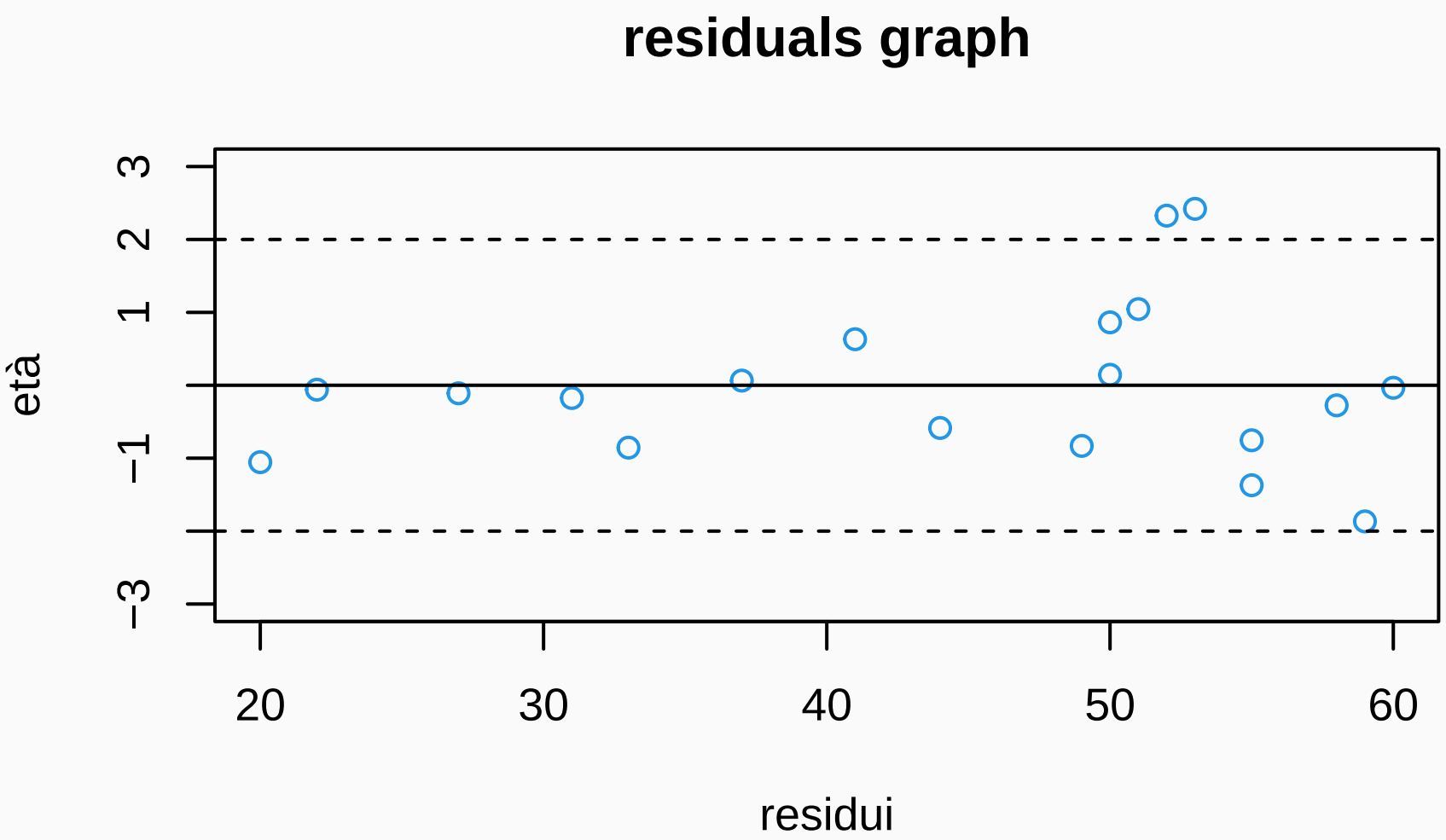
- We can now obtain residuals by comparing the predictions with the observed values and dividing them by the estimated standard deviations (remind that the model is heteroschedastic)

$$r_i = \frac{\mu_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$$

Poisson regression: Residual checks

- A look at residuals plots can help detecting if our assumptions are reasonable
- Residuals move around 0, they should be reasonably small and should not show any specific pattern.
- For large samples they are approximately equivalent to draws from a standard Gaussian. This means that the large majority (about 95%) of them have a value between 2 and -2.
- A large number of residuals whose absolute value is greater than 2 could be a symptom that variance(Y_i)>Mean(Y_i).
- This situation, called overdispersion, indicates that a Poisson model could be not appropriate

Data an medical examination



Note that two residuals are outside the interval $[-2, +2]$

Overdispersed count data

- Poisson regression models in a GLM context imply that the variance function then functionally related to the mean function (it is actually the same).
- For a Poisson models the standardized residuals are

$$z_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$$

  where $\hat{\mu}_i = \exp^{x_i^T \hat{\beta}}$
- If the Poisson model holds the z_i are approximately independent and will have mean equal to 0 and variance equal to 1. Approximately $\sum_{i=1}^{n} z_i^2$ is a χ^2_{n-k} distribution if the model holds. This can be used for detecting overdispersion.
- Considering a model for counts with overdispersion will be more realistic in many practical cases

Dealing with overdispersion

- For LMs the method of LS allows to obtain estimates of the regression parameters without the specification of a probabilistic model.
- The method of LS requires only the specification of the relation between the expected value of the response variable and the linear predictor, and the specification of the variance of the error term,
- Also for the GLMs it is possible to specify only these two relations (assuming that the variance function $V(\mu_i)$ is known).
- In other words, this means that the parametric assumption $Y_i \sim EF(\cdot, \phi)$ could not even be satisfied. Only the assumption about expectations is essential:$\mu_i = E(Y_i) = g^{-1}(\eta_i)$
- the only distributional feature that must be known in order to calculate the estimating equation is the variance function $V(\mu)$.

Quasi-likelihood model

- Under suitable regularity conditions, the likelihood equations for a GLM give estimates for the coefficients β which maintain several properties, also if the parametric assumptions of Y_i are substituted with weaker assumptions:
  - $g(\mu_i) = g(E(Y_i)) = \eta_i, \qquad i = 1, \ldots, n,$
  - $var(Y_i) = \phi V(\mu_i), \qquad i = 1, \ldots, n,$
  - $cov(Y_i, Y_j) = 0$, if $i \neq j$.
- The semi-parametric statistical model specified by assumptions 1–3 is called **quasi-likelihood model**.

Quasi-likelihood model

- The assumptions 1–3 above offer an increase in flexibility with respect to the usual parametric specifications based, respectively, on the Poisson, binomial or exponential distributions.
- In general, the quasi-likelihood approach allows to deal with *overdispersion problems*: it is possible to specify $var(Y_i)$ so that there is more variability with respect to the exponential family.
- The case of *underdispersion*, i.e. $\phi < 1$, is less important in applications, but can be dealt with under the quasi-likelihood model as well.

Using quasi-likelihood in `glm`

- When estimating a GLM by using quasi-likelihood one can use the same variance function derived from a Binomial or from a Poisson model and using the canonical link for those models. In R this leads to a specification of the `family` that is called `quasibinomial` or `quasipoisson`.
- Estimates of the β are the same since the estimating equations do not change
- But standard errors of estimates will change since a value different from 1 is estimated for ϕ. In `quasipoisson` one should take into account that variance is modelled as $Var(y_i) = \phi\mu_i$
- The parameter ϕ can be also estimated as

$$\hat{\phi} = \frac{1}{n-p}\sum \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$$

- In those cases also the Deviance of the model has to be corrected because it is computed assuming $\phi = 1$. The deviance reported has to be divided by $\hat{\phi}$
- Also the standardized residuals are different. E.g., for the Poisson:

$$z_i^{QL} = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\phi}\hat{\mu}_i}} \quad vs \quad z_i^{GLM} = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$$

Poisson regression. . . and beyond

Poisson regression: Taking into account exposure

- The basic Poisson model can be extended to take into account the fact that counts can arise under different conditions.
- Assume we want to model the average number of accidents Y_i on some roads. Obviously, this number depends on how many vehicles in a given period have been on the road. The set of those exposed to the risk of accident is different for each data unit.
- The number of vehicles is an exposure variable e_i and it should be taken into account.
- It could be sensible to model the rates instead of the counts, i.e., we could write for $\mu_i = E(y_i)$
- $log(\frac{\mu_i}{e_i}) = \beta_0 + \beta_{i1}x_{i1} + \beta_{i2}x_{i2} + \cdots + \beta_{ip}x_{ip}$
- But this is equivalent to put
  $log(\mu_i) = \beta_0 + \beta_{i1}x_{i1} + \beta_{i2}x_{i2} + \cdots + \beta_{ip}x_{ip} + log(e_i)$
- This means that we insert an additional variable $log(e_i)$ among the covariates but imposing its coefficient to be equal to 1
- The special covariate $log(e_i)$ is called offset

Other models for count data

- The basic Poisson model is the simplest one to use when modelling counts.
- Poisson random variables are not the only ones that can be used to describe the distribution of counts.
- Other slightly more complex models, for instance Negative Binomial random variables, could be more flexible to cope with situations (such as overdispersion) frequently encountered in practise.
- Poisson models could be also refined to face non standard and more complex data patterns, such as: + The case of counts that can never take on the value 0 (truncated Poisson) + The case where counts can be observed only for a portion of the sample. For the remaning portion we know that only 0 is possible (zero inflated models). This can occur in automobile claims data because insured are reluctant to report claims fearing that this will result in higher future insurance premiums.
- More complex model are also needed to accommodate situations where the theoretical conditions that give raise to Poisson counts are not met.

Negative binomial regression

- It is an alternative model that can be considered when data exhibits overdispersion. Its probability function is

$$Pr(Z = z) = \binom{z + k - 1}{z} p^k (1 - p)^z \qquad z = 0, 1, \ldots$$

where $E(Z) = k(1 - p)/p$ and $Var(Z) = k(1 - p)/p^2$.

- Interpretation: probability to observe z *failures* until the pre-specified number of *successes* k is observed.
- Compared with Poisson
  - since it has an extra parameter it proves to be more flexible
  - mean is larger than variance and then it accommodates overdispersion
  - Poisson is a limiting case of negative binomial (if $p \to 1$ and $k \to 0$ then $kp \to \lambda$)
- Recall that negative binomial emerges as a mixture of Poisson when each unit Y is Poisson with mean λ and λ are drown from a *Gamma* distribution.

Negative Binomial regression

- When building a model for Negative Binomial a different parametrization is more appropriate, by defining $Y = Z - k$ and $p = \frac{1}{1+\alpha}$
- $$Pr(Y = y) = \binom{y + k - 1}{k - 1} \frac{\alpha^y}{(1 + \alpha)^{y+k}} \qquad y = 0, 1, \ldots$$
- Then
  - $E(Y) = \mu = k\alpha$
  - $Var(Y) = k\alpha + k\alpha^2 = \mu + \mu^2/k$
- and the following link can be used $\log \frac{\alpha}{1+\alpha} = \log \frac{\mu}{k+\mu}$
- In `R` a specific function has to be used: `glm.nb(...)` included in the package MASS

Zero inflated Poisson

- Zero inflation means that we have far more zeros than what would be expected for a Poisson or BiN distribution
- Ignoring zero inflation can have two consequences:
  - the estimated parameters and standard errors may be biased
  - the excessive number of zeros can cause overdispersion
- A possible model hypotizes that the observed counts derive from a mixture of two populations:
  - for a part of the population (with probability p) Y can only be 0
  - for the remaining part (with probability $1-p$) Y is distributed as a Poisson or a BiN.
- Distribution of counts is then, in case of Poisson

$$P(y_i = 0) = p_i + (1 - p_i)e^{-\mu_i}$$

$$P(y_i = y_i | y_i > 0) = (1 - p_i)\frac{\mu_i^{y_i} e^{\mu_i}}{y_i!}$$

- Covariates can be introduced, like in GLM, for modelling p_i and μ_i