APTS Statistical Modelling

Lecture Discussion 2

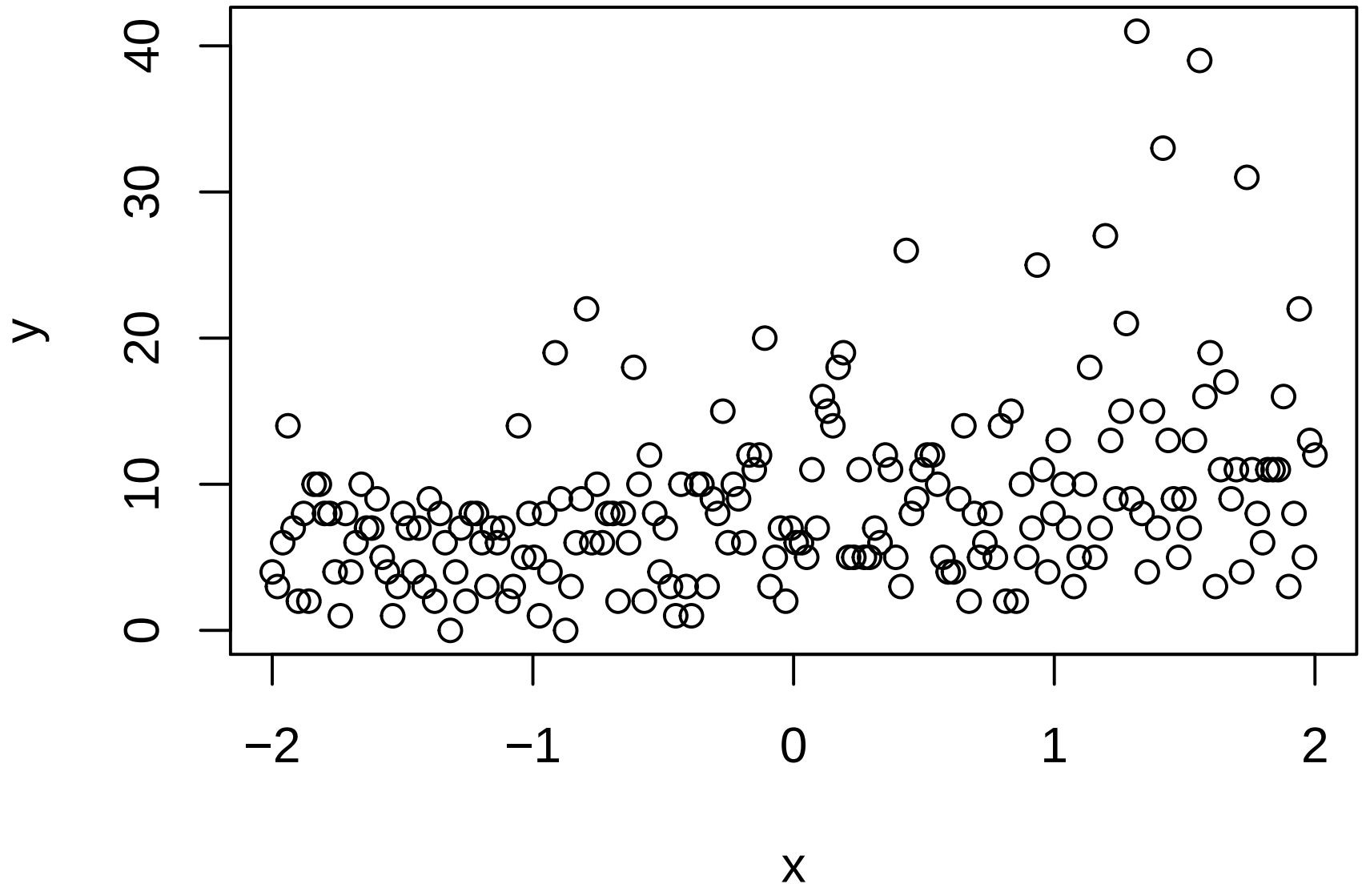
Helen Ogden

April 2021

Lecture discussion 2

- Please mute your microphone and turn off video, unless you are asking a question.
- Feel free to ask questions about any of the lecture content so far.
- Use the chat to ask questions directly, or to state that you have a question.

How should we model this data?



Fitting a Poisson GLM

We can fit the Poisson log-linear model

$$Y_i \sim \text{Poisson}(\mu_i), \quad \log \mu_i = \beta_0 + \beta_1 x_i.$$

```
mod <- glm(y ~ x, family = "poisson")
mod
```

```
##
## Call:  glm(formula = y ~ x, family = "poisson")
##
## Coefficients:
## (Intercept)            x
##      2.1425       0.2291
##
## Degrees of Freedom: 199 Total (i.e. Null);  198 Residual
## Null Deviance:       780.9
## Residual Deviance: 658.7     AIC: 1421
```

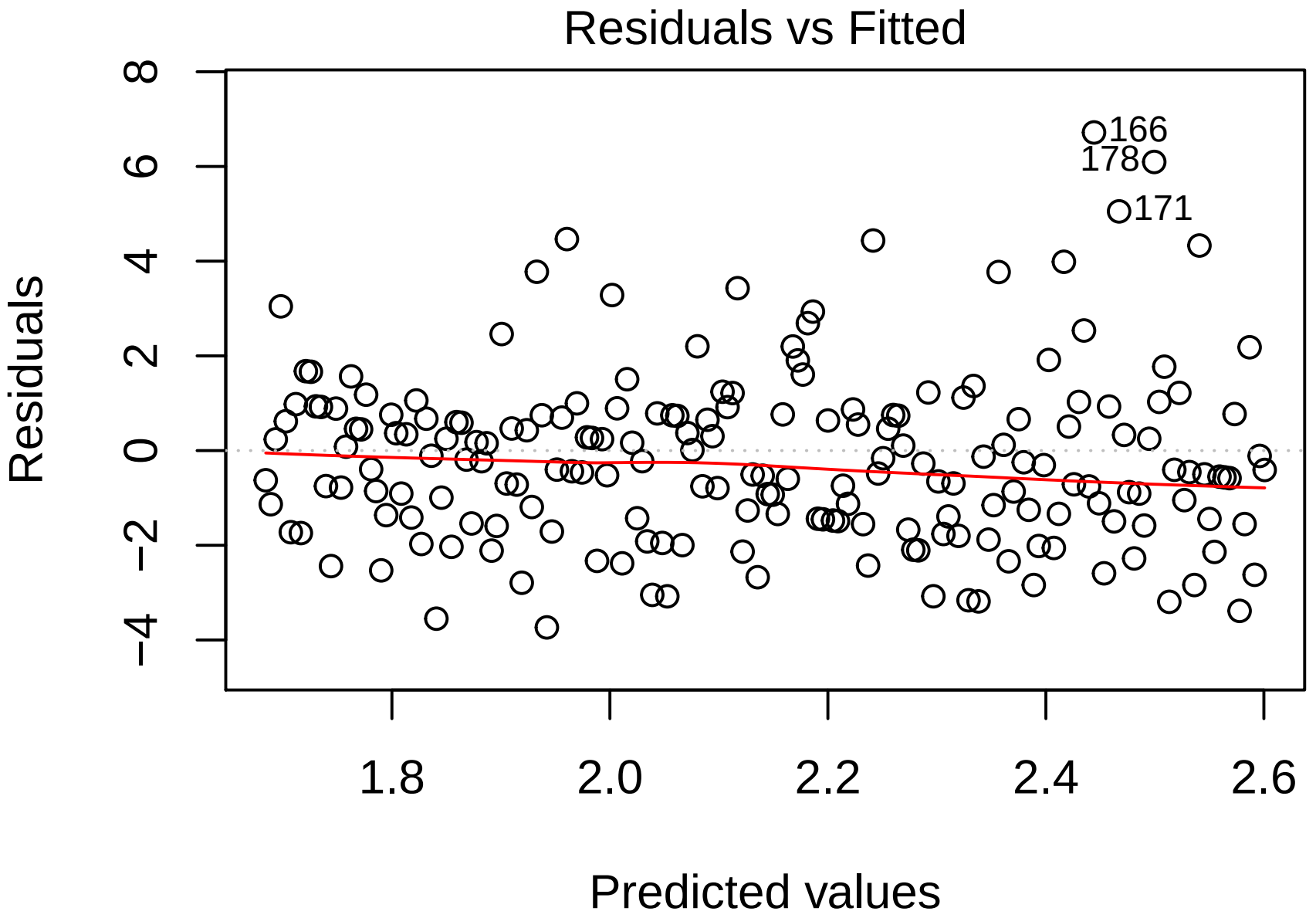
Fitting a Poisson GLM

```r
summary(mod)
```

```
##
## Call:
## glm(formula = y ~ x, family = "poisson")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.7346  -1.4462  -0.4038   0.7526   6.7182
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.14248    0.02464   86.95   <2e-16 ***
## x            0.22913    0.02095   10.94   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 780.89  on 199  degrees of freedom
## Residual deviance: 658.68  on 198  degrees of freedom
## AIC: 1421.3
##
## Number of Fisher Scoring iterations: 5
```

Checking model fit

```
plot(mod, which = 1)
```



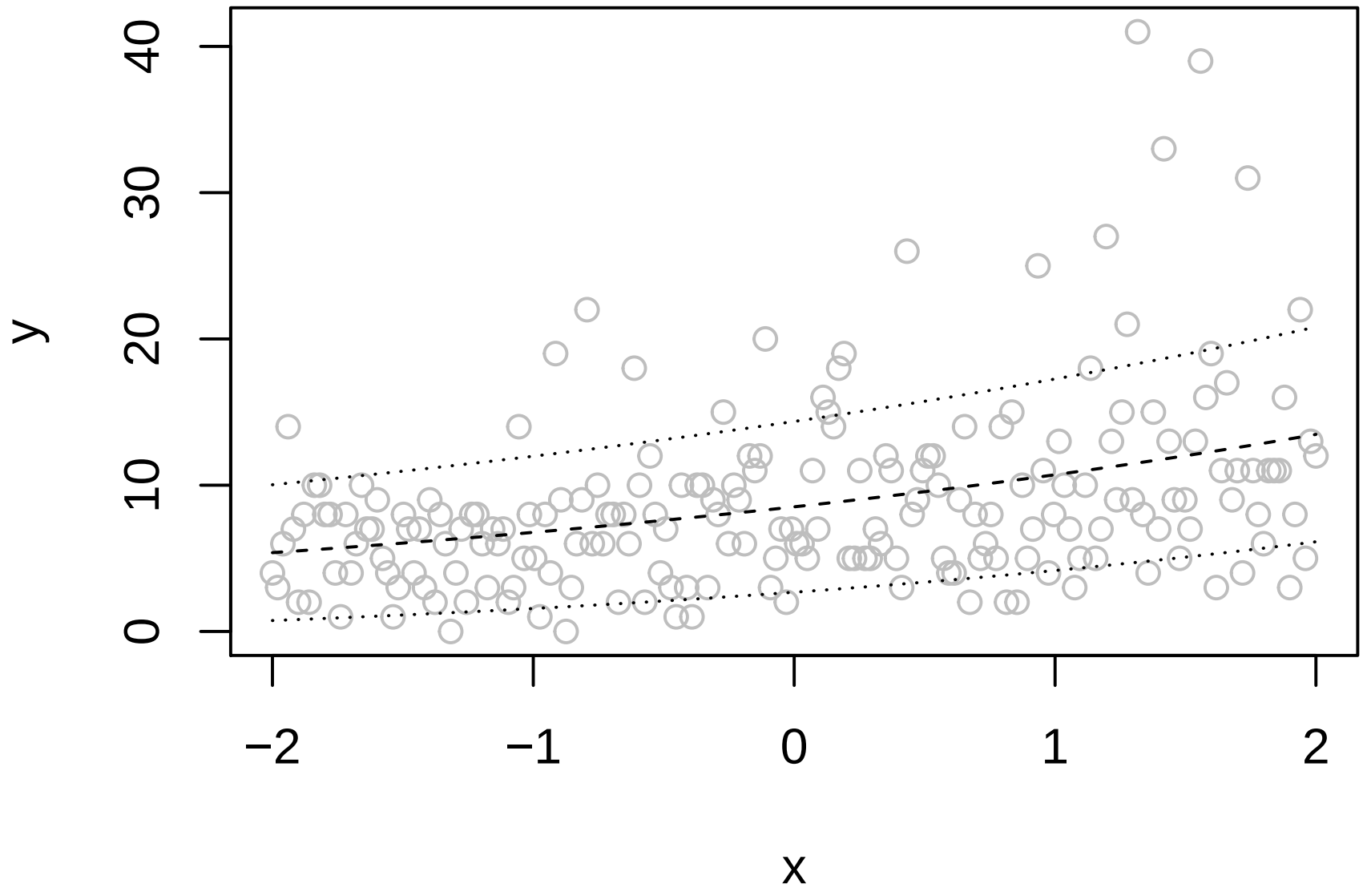
Checking model fit

Are you happy with the model fit?

- Yes, the model seems to fit well.
- No, there is some problem with the model fit.
- Not sure or need more information.

Plotting the fitted model

Adding the fitted mean plus or minus two standard deviations:



Quasilikelihood approach

Model

$$E(Y_i) = \mu_i = \exp(\beta_0 + \beta_1 x_i), \quad \text{Var}(Y_i) = \sigma^2 \mu_i$$

```
mod_quasi <- glm(y ~ x, family = "quasipoisson")
```

Quasilikelihood approach

```
summary(mod_quasi)
```

```
##
## Call:
## glm(formula = y ~ x, family = "quasipoisson")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.7346  -1.4462  -0.4038   0.7526   6.7182
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.14248    0.04692  45.667  < 2e-16 ***
## x            0.22913    0.03988   5.746 3.41e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 3.624912)
##
##     Null deviance: 780.89  on 199  degrees of freedom
## Residual deviance: 658.68  on 198  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

Checking model fit

Do the assumptions of the quasilikelihood approach seem reasonable here?

- Yes, the assumptions are now reasonable.
- No, there is some problem with the assumptions.
- Not sure or need more information.

How were the data generated?

In fact

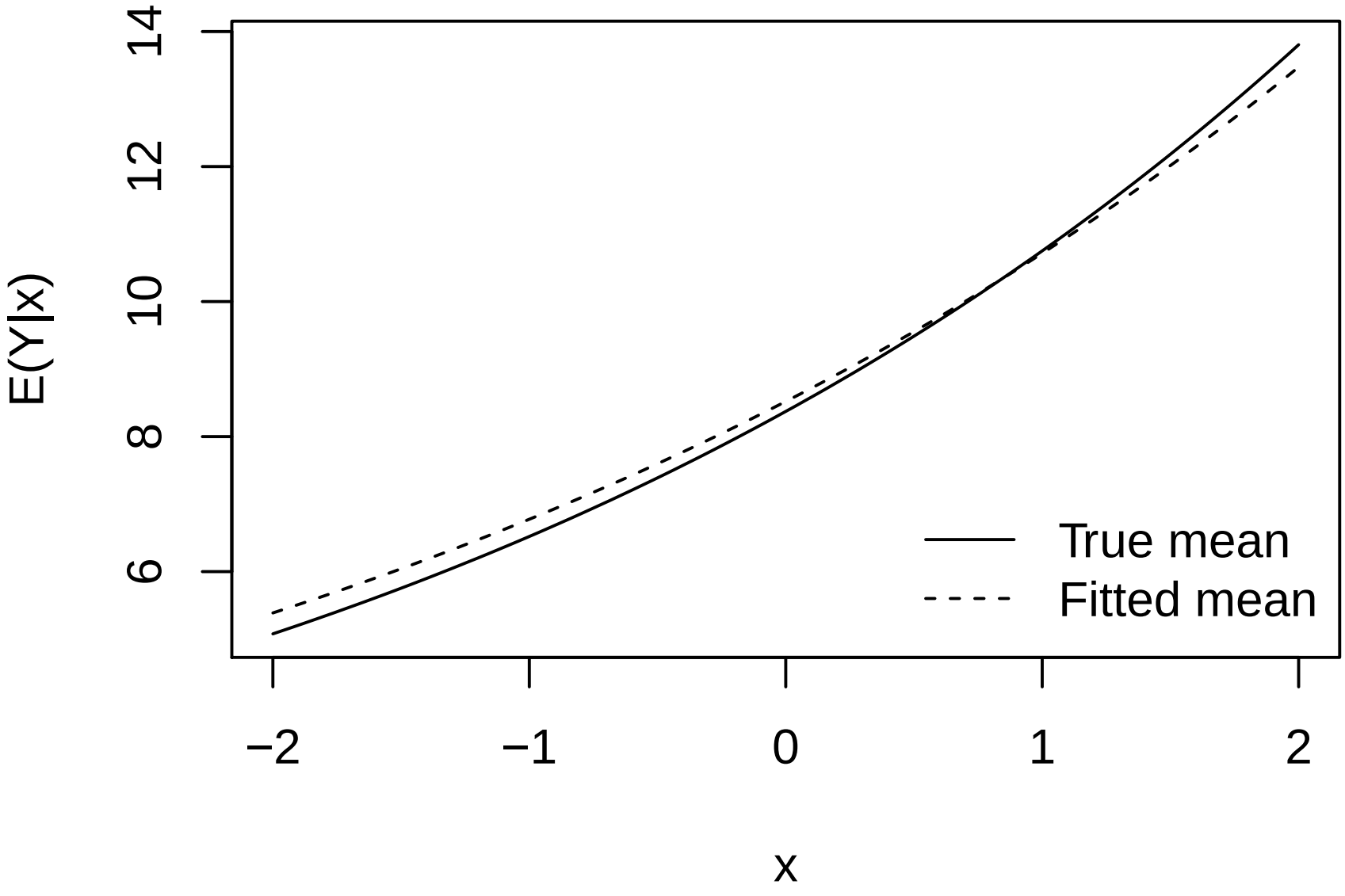
$$Y_i \mid x_i, z_i \sim \text{Poisson}(\mu_i), \quad \log \mu_i = \beta_0 + \beta_1 x_i + \beta_2 z_i,$$

but we **do not observe** the explanatory variable z_i (in this case $z_i \sim N(0, 1)$).

We are trying to model the marginal distribution $Y_i \mid x_i$, with particular interest in the marginal mean $\mu(x) = E(Y|x)$ and the marginal variance $\text{Var}(Y|x)$.

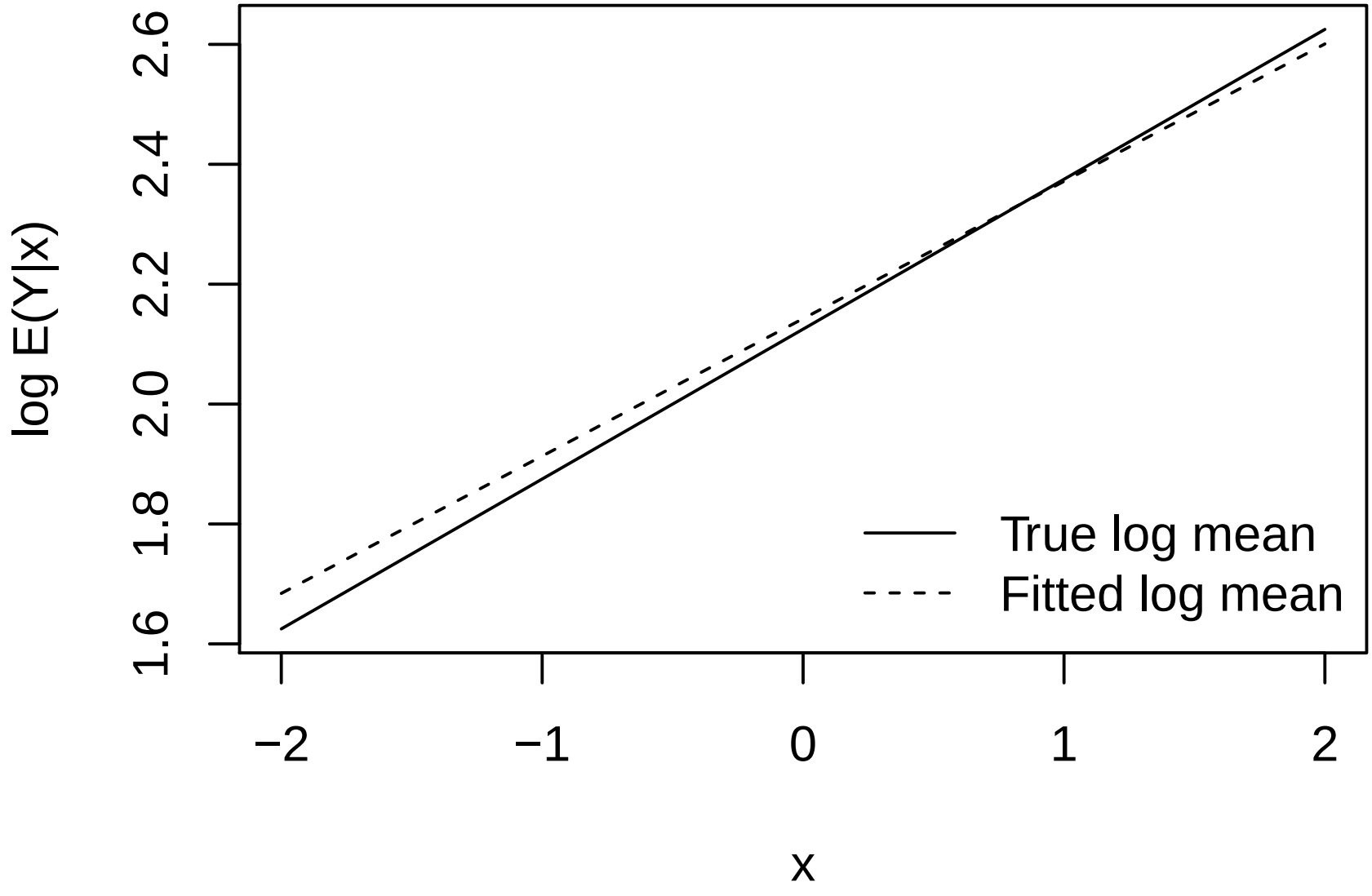
The marginal mean

Plotting the true marginal mean and the estimated marginal mean:



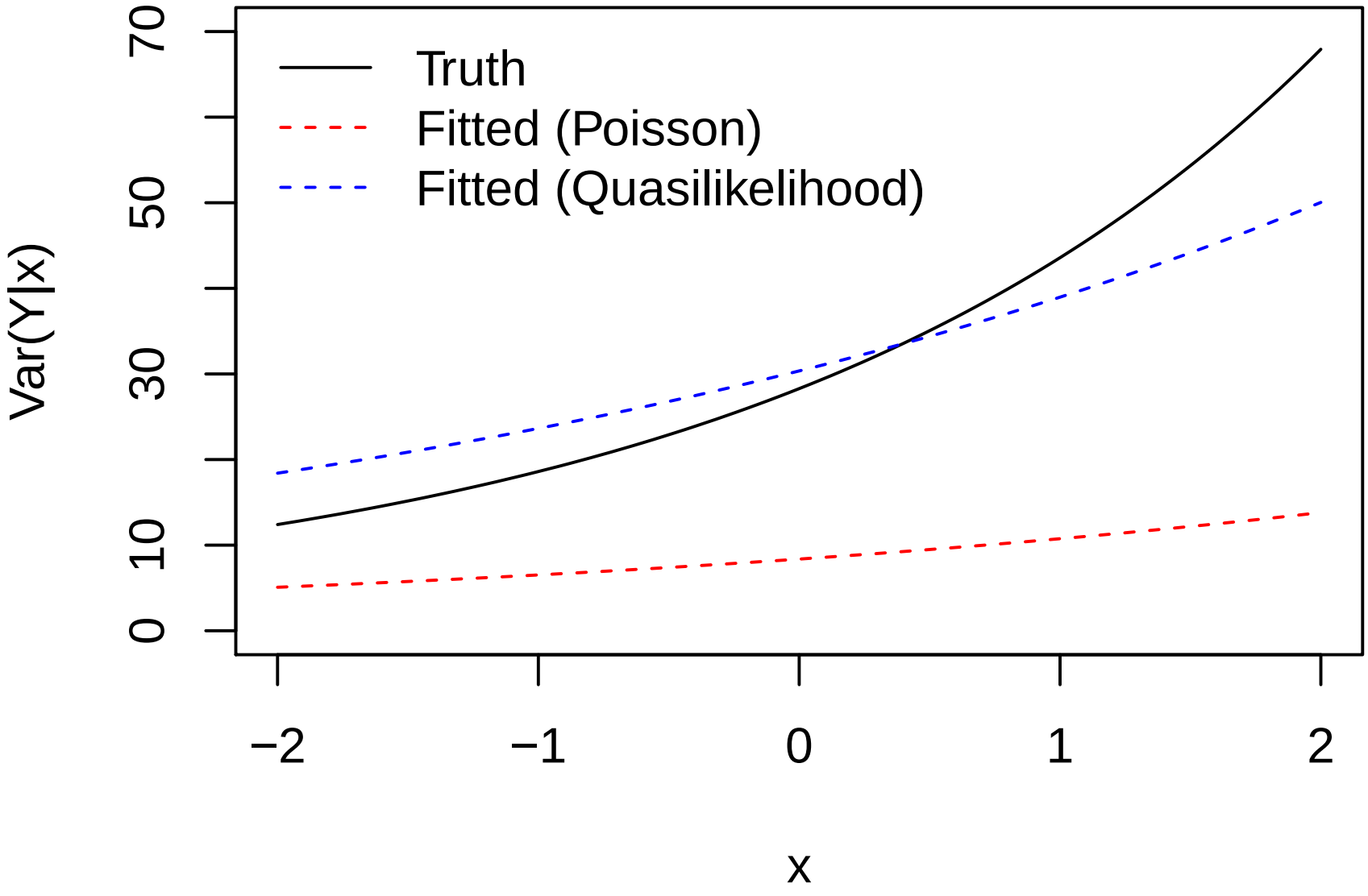
The marginal mean

In this case $\log(\mu(x))$ is approximately linear in x:



The marginal variance

Plotting the true and estimated marginal variances:



Ratio of marginal variance

The quasilikelihood approach assumes

$$\text{Var}(Y|x) = \sigma^2 \mu(x),$$

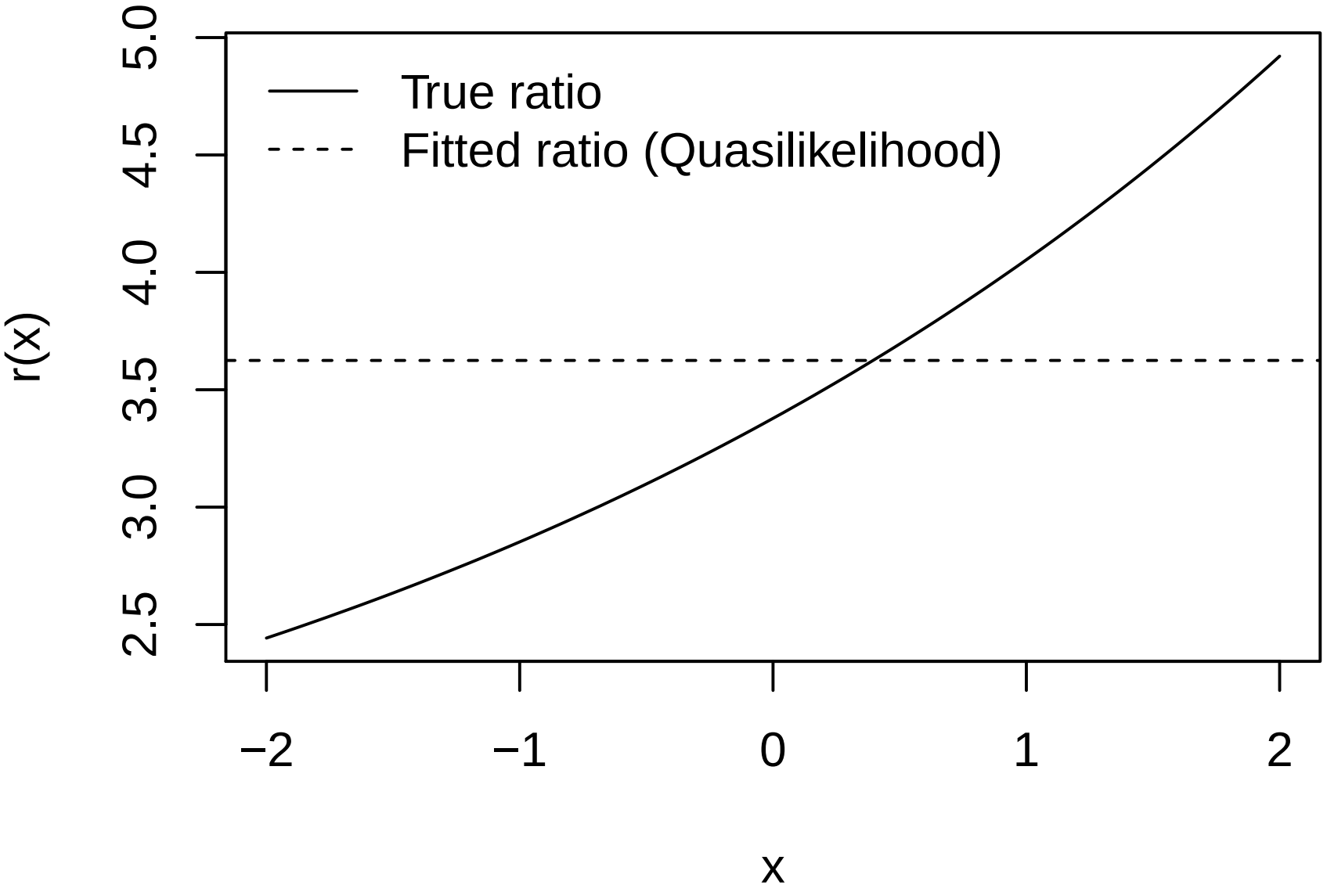
i.e. that the true marginal variance divided by the Poisson model marginal variance

$$r(x) = \frac{\text{Var}(Y|x)}{\mu(x)}$$

is constant (σ^2).

Ratio of marginal variances

Plotting the true $r(x)$ with the fitted ratio $\hat{\sigma}^2$ from quasilikelihood:



Summary: overdispersion and quasilikelihood approaches

- Overdispersion relative to Poisson model is **very** common in count data.
- Failing to observe an important explanatory variable is one cause of overdispersion.
- The assumption of quasilikelihood $\text{Var}(Y|x) = \sigma^2\mu(x)$ might not be accurate (but a big improvement over original model).

Questions on lecture content so far?