STK3100
Introduction to Generalized Linear Models

OBLIG 2

Egil Furnes
Student: 693784

Problem 1

a)

The response `ceb` is a non–negative count (number of children). A Poisson GLM with log link,

$$Y_i \mid x_i \sim \text{Poisson}(\mu_i), \qquad \log(\mu_i) = x_i^\top \beta,$$

is a natural starting point: it ensures $\mu_i > 0$, has an interpretable multiplicative effect of covariates on the mean, and treats births as conditionally independent events within covariate strata. The residuals–versus–fitted plot (Fig. 1) suggests increasing variability with the mean (fan–shape), hinting that the Poisson assumption $\text{Var}(Y_i) = \mu_i$ may be tight; this motivates the overdispersion check in Problem 2.

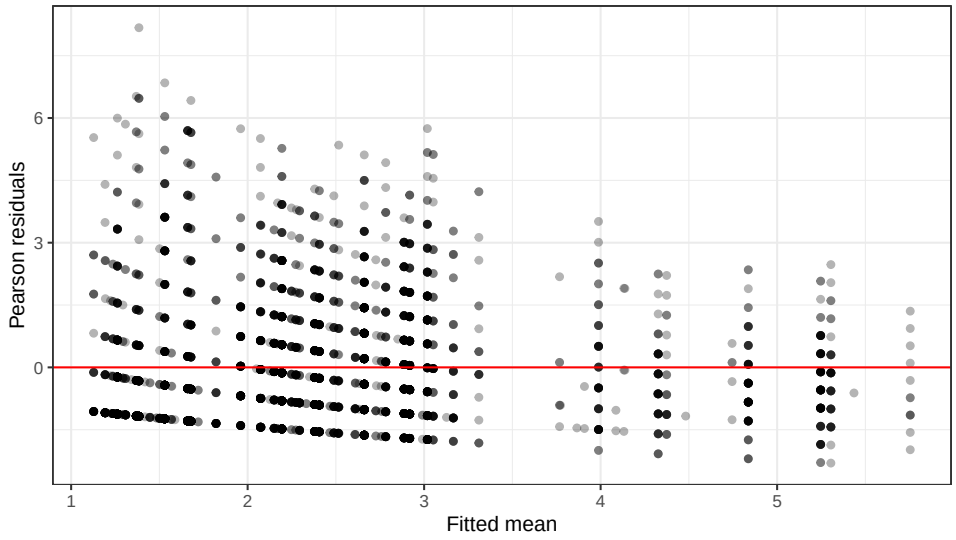

Figure 1: Residuals versus fitted

b)

We fit a Poisson GLM including all seven covariates. The summary output shows strong evidence that `educ0`, `usemeth`, `urban`, `radio`, and `bicycle` affect fertility. The coefficients for `electric` and `tv` are small and not statistically significant at conventional levels. Exponentiating the coefficients gives rate ratios: education and contraceptive use are associated with higher rates, urban residence and radio ownership with lower rates, and bicycle ownership with slightly higher rates. (See printed output for exact estimates and CIs.)

```
> summary(fit_full)

Call:
glm(formula = ceb ~ educ0 + usemeth + urban + electric + radio +
    tv + bicycle, family = poisson(link = "log"), data = fertility)
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.50704    0.02659  19.071  < 2e-16 ***
educ01       0.59763    0.02312  25.843  < 2e-16 ***
usemeth1     0.55292    0.02202  25.105  < 2e-16 ***
urban1      -0.19288    0.02153  -8.959  < 2e-16 ***
electric1   -0.05697    0.03734  -1.526 0.127079
radio1      -0.08135    0.02277  -3.573 0.000353 ***
tv1         -0.05561    0.04509  -1.233 0.217453
bicycle1     0.09236    0.02284   4.044 5.25e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 10138.6  on 4169  degrees of freedom
Residual deviance:  8760.1  on 4162  degrees of freedom
AIC: 17647

Number of Fisher Scoring iterations: 5
```

```
> tidy(fit_full, exponentiate = TRUE, conf.int = TRUE)
# A tibble: 8 × 7
  term        estimate std.error statistic   p.value conf.low conf.high
  <chr>          <dbl>     <dbl>     <dbl>     <dbl>    <dbl>     <dbl>
1 (Intercept)    1.66     0.0266     19.1  4.43e- 81    1.58      1.75
2 educ01         1.82     0.0231     25.8  2.89e-147    1.74      1.90
3 usemeth1       1.74     0.0220     25.1  4.41e-139    1.67      1.82
4 urban1         0.825    0.0215     -8.96 3.27e- 19    0.790     0.860
5 electric1      0.945    0.0373     -1.53 1.27e-  1    0.878     1.02
6 radio1         0.922    0.0228     -3.57 3.53e-  4    0.882     0.964
7 tv1            0.946    0.0451     -1.23 2.17e-  1    0.865     1.03
8 bicycle1       1.10     0.0228      4.04 5.25e-  5    1.05      1.15
```

c)

Using single–term deletions and likelihood–ratio tests:

- Removing `tv` is not significant (p=0.216) and slightly improves AIC (17647.01 $\rightarrow$ 17646.54).
- `bicycle` is important: removing it significantly worsens fit ($p < 10^{-4}$) and increases AIC.
- `educ0`, `usemeth`, and `urban` are highly significant; `radio` is also significant. `electric` is borderline in the full model.

Preferred model:

$$\log \mu = \beta_0 + \beta_1 \texttt{educ0} + \beta_2 \texttt{usemeth} + \beta_3 \texttt{urban} + \beta_4 \texttt{electric} + \beta_5 \texttt{radio} + \beta_6 \texttt{bicycle},$$

i.e. the full model *without* `tv`. This choice matches the tests and AIC.

```
> drop1(fit_full, test = "Chisq")
Single term deletions

Model:
ceb ~ educ0 + usemeth + urban + electric + radio + tv + bicycle
         Df Deviance   AIC    LRT  Pr(>Chi)
<none>        8760.1 17647
educ0     1   9383.4 18268 623.30 < 2.2e-16 ***
usemeth   1   9424.2 18309 664.10 < 2.2e-16 ***
urban     1   8840.8 17726  80.77 < 2.2e-16 ***
electric  1   8762.4 17647   2.35 0.1252082
radio     1   8772.7 17658  12.67 0.0003709 ***
tv        1   8761.6 17647   1.53 0.2157022
bicycle   1   8776.2 17661  16.16 5.811e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> anova(fit_full, fit_no_tv, test = "Chisq")
Analysis of Deviance Table

Model 1: ceb ~ educ0 + usemeth + urban + electric + radio + tv + bicycle
Model 2: ceb ~ educ0 + usemeth + urban + electric + radio + bicycle
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      4162     8760.1
2      4163     8761.6 -1  -1.5327   0.2157
```

```
> anova(fit_no_tv, fit_no_bicycle, test = "Chisq")
Analysis of Deviance Table

Model 1: ceb ~ educ0 + usemeth + urban + electric + radio + bicycle
Model 2: ceb ~ educ0 + usemeth + urban + electric + radio
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      4163     8761.6
2      4164     8777.3 -1  -15.725 7.327e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> AIC(fit_full, fit_no_tv, fit_no_bicycle)
               df      AIC
```

```
fit_full        8 17647.01
fit_no_tv       7 17646.54
fit_no_bicycle  6 17660.27
```

```
Call:
glm(formula = ceb ~ educ0 + usemeth + urban + electric + radio +
    bicycle, family = poisson(link = "log"), data = fertility)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.50856    0.02656  19.147  < 2e-16 ***
educ01       0.59841    0.02312  25.879  < 2e-16 ***
usemeth1     0.55283    0.02203  25.098  < 2e-16 ***
urban1      -0.19546    0.02144  -9.117  < 2e-16 ***
electric1   -0.07939    0.03276  -2.423 0.015397 *
radio1      -0.08400    0.02268  -3.704 0.000213 ***
bicycle1     0.09100    0.02281   3.989 6.65e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 10138.6  on 4169  degrees of freedom
Residual deviance:  8761.6  on 4163  degrees of freedom
AIC: 17647

Number of Fisher Scoring iterations: 5
```

d)

Interpreting the preferred model's exponentiated coefficients (rate ratios) with 95% CIs:

- `educ0`=1 $\Rightarrow$ RR $= 1.82$ [1.74, 1.90]: women with education have $\sim$82% higher expected `ceb`.
- `usemeth`=1 $\Rightarrow$ RR $= 1.74$ [1.66, 1.82]: contraceptive users have $\sim$74% higher expected `ceb`.
- `urban`=1 $\Rightarrow$ RR $= 0.822$ [0.789, 0.858]: urban residence is associated with $\sim$18% lower `ceb`.
- `electric`=1 $\Rightarrow$ RR $= 0.924$ [0.866, 0.985]: electricity is associated with a modest reduction in `ceb`.
- `radio`=1 $\Rightarrow$ RR $= 0.919$ [0.880, 0.961]: radio ownership is associated with lower `ceb`.
- `bicycle`=1 $\Rightarrow$ RR $= 1.10$ [1.05, 1.15]: bicycle ownership is associated with a small increase in `ceb`.

(Intercept is the baseline mean rate on the multiplicative scale.)

```
> rate_ratios
# A tibble: 7 × 5
  term        estimate conf.low conf.high   p.value
  <chr>          <dbl>    <dbl>     <dbl>     <dbl>
1 (Intercept)    1.66     1.58      1.75  1.02e- 81
2 educ01         1.82     1.74      1.90  1.16e-147
3 usemeth1       1.74     1.66      1.82  5.20e-139
4 urban1         0.822    0.789     0.858 7.74e- 20
5 electric1      0.924    0.866     0.985 1.54e-  2
6 radio1         0.919    0.880     0.961 2.13e-  4
7 bicycle1       1.10     1.05      1.15  6.65e-  5
```

e)

For a woman with `educ0`=0, `usemeth`=0, `urban`=0, `electric`=1, `radio`=1, `bicycle`=1 (and no TV in the chosen model), the predicted rate is

$$\widehat{\mu} = 1.547 \quad \text{with 95\% CI } [1.429,\ 1.674].$$

This corresponds to the expected number of living children under the preferred Poisson model.

```
> cbind(newdata, rate_est, rate_low, rate_high)
  educ0 usemeth urban electric radio bicycle rate_est rate_low rate_high
1     0       0     0        1     1       1 1.546757 1.428844    1.6744
```

R-code

```r
library(tidyverse)
library(broom)
library(MASS)

set.seed(3100)
theme_set(theme_bw())

# problem 2
  ------------------------------------------------------------------

# read data (same as P1)
fertility <- read.csv("data/fertility_data.csv", header = TRUE) %>%
  as_tibble() %>%
  mutate(
    educ0     = as.factor(educ0),
```

```r
    usemeth  = as.factor(usemeth),
    urban    = as.factor(urban),
    electric = as.factor(electric),
    radio    = as.factor(radio),
    tv       = as.factor(tv),
    bicycle  = as.factor(bicycle)
  )

# a)-e) (analytical): handled in LaTeX write-up later.

# f) Fit NB-GLM for the preferred model from Problem 1 (no TV)
# Poisson reference (from P1)
fit_pois <- glm(
  ceb ~ educ0 + usemeth + urban + electric + radio + bicycle,
  data = fertility,
  family = poisson(link = "log")
)

# Negative binomial (MASS::glm.nb uses log link by default)
fit_nb <- glm.nb(
  ceb ~ educ0 + usemeth + urban + electric + radio + bicycle,
  data = fertility, link = log
)

summary(fit_nb)
tidy(fit_nb, exponentiate = TRUE, conf.int = TRUE)

# Theta (k) estimate and SE
theta_hat <- fit_nb$theta
theta_se  <- fit_nb$SE.theta
c(theta_hat = theta_hat, theta_se = theta_se)

# Compare Poisson vs NB
comp_aic <- AIC(fit_pois, fit_nb)
comp_ll  <- tibble(
  model = c("Poisson", "NegBin"),
  logLik = c(as.numeric(logLik(fit_pois)), as.numeric(logLik(fit_nb)
    )),
  df = c(attr(logLik(fit_pois), "df"), attr(logLik(fit_nb), "df"))
)

comp_aic
comp_ll

# LR-style comparison (treat NB as Poisson with extra parameter k)
lr_stat <- 2 * (as.numeric(logLik(fit_nb)) - as.numeric(logLik(fit_
   pois)))
lr_pval <- pchisq(lr_stat, df = 1, lower.tail = FALSE)
c(lr_stat = lr_stat, df = 1, p_value = lr_pval)

# Pearson dispersion check
```

```r
disp_pois <- sum(residuals(fit_pois, type = "pearson")^2) / df.residual(fit_pois)
disp_nb   <- sum(residuals(fit_nb,   type = "pearson")^2) / df.residual(fit_nb)
c(pearson_overdisp_poisson = disp_pois, pearson_overdisp_negbin = disp_nb)

# Residuals vs fitted for NB (diagnostic)
resid_nb_plot <- fertility %>%
  mutate(
    fitted = fitted(fit_nb),
    resid  = residuals(fit_nb, type = "pearson")
  ) %>%
  ggplot(aes(x = fitted, y = resid)) +
  geom_point(alpha = 0.3) +
  geom_hline(yintercept = 0, color = "red") +
  labs(x = "Fitted values (NB-GLM)", y = "Pearson residuals") +
  theme_bw()

ggsave("plots/problem2_nb_residuals_vs_fitted.pdf", resid_nb_plot, width = 7, height = 4)

# Rate ratios from NB model (for interpretation in the report)
nb_rate_ratios <- tidy(fit_nb, exponentiate = TRUE, conf.int = TRUE) %>%
  dplyr::select(term, estimate, conf.low, conf.high, p.value)

nb_rate_ratios

# Prediction for the same covariate combo as in P1(e)
newdata <- tibble(
  educ0   = factor(0, levels = c(0,1)),
  usemeth = factor(0, levels = c(0,1)),
  urban   = factor(0, levels = c(0,1)),
  electric = factor(1, levels = c(0,1)),
  radio    = factor(1, levels = c(0,1)),
  bicycle  = factor(1, levels = c(0,1))
)

pred_nb <- predict(fit_nb, newdata = newdata, type = "link", se.fit = TRUE)

nb_rate_est  <- exp(pred_nb$fit)
nb_rate_low  <- exp(pred_nb$fit - 1.96 * pred_nb$se.fit)
nb_rate_high <- exp(pred_nb$fit + 1.96 * pred_nb$se.fit)

cbind(newdata, nb_rate_est, nb_rate_low, nb_rate_high)
```

Problem 2

We use the parameterisation

$$p(y;\mu,k) = \frac{\Gamma(y+k)}{\Gamma(k)\,\Gamma(y+1)}\left(\frac{k}{\mu+k}\right)^k\left(\frac{\mu}{\mu+k}\right)^y, \qquad y = 0,1,2,\ldots,$$

so that $\mathbb{E}(Y) = \mu$ and $\text{Var}(Y) = \mu + \mu^2/k$.

a)

Let $Y^* = Y/k$. From (2),

$$p(y^*;\mu,k) = \frac{\Gamma(ky^*+k)}{\Gamma(k)\,\Gamma(ky^*+1)}\left(\frac{k}{\mu+k}\right)^k\left(\frac{\mu}{\mu+k}\right)^{ky^*}, \qquad y^* \in \{0,\tfrac{1}{k},\tfrac{2}{k},\ldots\}.$$

Write this on exponential–dispersion form

$$p(y^*;\theta,\phi) = \exp\left(\frac{y^*\theta - b(\theta)}{a(\phi)} + c(y^*;\phi)\right),$$

with $a(\phi) = 1/k$. Taking logs,

$$\log p(y^*;\mu,k) = \underbrace{\log\frac{\Gamma(ky^*+k)}{\Gamma(k)\Gamma(ky^*+1)}}_{c(y^*;k)} + k\log\frac{k}{\mu+k} + ky^*\log\frac{\mu}{\mu+k}.$$

Thus, with $a(\phi) = 1/k$, we can choose

$$\theta = \log\frac{\mu}{\mu+k} \quad (<0), \qquad b(\theta) = -\log(1-e^\theta),$$

since $k\{y^*\theta - b(\theta)\} = ky^*\log\frac{\mu}{\mu+k} + k\log\frac{k}{\mu+k}$. Hence (2) is an EDF with canonical parameter $\theta = \log\{\mu/(\mu+k)\}$ and $c(y^*;k) = \log\Gamma(ky^*+k) - \log\Gamma(k) - \log\Gamma(ky^*+1)$.

b)

For an EDF, $\mathbb{E}(Y^*) = b'(\theta)$ and $\text{Var}(Y^*) = a(\phi)\,b''(\theta)$. With $b(\theta) = -\log(1-e^\theta)$,

$$b'(\theta) = \frac{e^\theta}{1-e^\theta} = \frac{\mu}{k}, \qquad b''(\theta) = \frac{e^\theta}{(1-e^\theta)^2} = \frac{\mu}{k}\left(1+\frac{\mu}{k}\right).$$

Since $a(\phi) = 1/k$,

$$\mathbb{E}(Y^*) = \frac{\mu}{k}, \qquad \text{Var}(Y^*) = \frac{1}{k}\frac{\mu}{k}\left(1+\frac{\mu}{k}\right) = \frac{\mu}{k^2} + \frac{\mu^2}{k^3}.$$

Therefore $Y = kY^*$ satisfies

$$\mathbb{E}(Y) = k\,\mathbb{E}(Y^*) = \mu, \qquad \text{Var}(Y) = k^2\,\text{Var}(Y^*) = \mu + \frac{\mu^2}{k}.$$

c)

Assuming $Y_1, \ldots, Y_n$ independent with pmf (1),

$$L(\mu, k; \mathbf{y}) = \sum_{i=1}^{n} \Big\{ \log\Gamma(y_i + k) - \log\Gamma(k) - \log\Gamma(y_i + 1) + k \log \frac{k}{\mu_i + k} + y_i \log \frac{\mu_i}{\mu_i + k} \Big\},$$

where $g(\mu_i) = \eta_i = \sum_j \beta_j x_{ij}$ (e.g. $g = \log$).

d)

For fixed $k > 0$, the deviance is

$$D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = 2\{L(\mathbf{y}, k; \mathbf{y}) - L(\hat{\boldsymbol{\mu}}, k; \mathbf{y})\} = 2 \sum_{i=1}^{n} \left[y_i \log \frac{y_i}{\hat{\mu}_i} + (y_i + k) \log \frac{\hat{\mu}_i + k}{y_i + k} \right],$$

with the usual convention $y_i \log(y_i/\hat{\mu}_i) = 0$ if $y_i = 0$.

e)

As $k \to \infty$,

$$(y_i + k) \log \frac{\hat{\mu}_i + k}{y_i + k} = (y_i + k) \log\left(1 + \frac{\hat{\mu}_i - y_i}{y_i + k} \right) \longrightarrow \hat{\mu}_i - y_i,$$

so

$$D(\mathbf{y}, \hat{\boldsymbol{\mu}}) \to 2 \sum_{i=1}^{n} \left[y_i \log \frac{y_i}{\hat{\mu}_i} - (y_i - \hat{\mu}_i) \right],$$

the Poisson deviance. Intuitively, $\mathrm{Var}(Y) = \mu + \mu^2/k \to \mu$, so the NB model converges to the Poisson model.

f)

Replacing the preferred Poisson model from Problem 1(c) by a negative binomial GLM (using `MASS::glm.nb`) yields:

```
> summary(fit_nb)

Call:
glm.nb(formula = ceb ~ educ0 + usemeth + urban + electric + radio +
    bicycle, data = fertility, link = log, init.theta = 2.122759)

Coefficients:
```

```
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.46801    0.03844  12.175  < 2e-16 ***
educ01       0.64862    0.03665  17.696  < 2e-16 ***
usemeth1     0.60773    0.03183  19.092  < 2e-16 ***
urban1      -0.20963    0.03169  -6.615 3.72e-11 ***
electric1   -0.07003    0.04642  -1.508  0.13143
radio1      -0.08557    0.03411  -2.509  0.01212 *
bicycle1     0.09061    0.03396   2.668  0.00764 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(2.1228) family taken to be 1)

    Null deviance: 5451.6  on 4169  degrees of freedom
Residual deviance: 4770.4  on 4163  degrees of freedom
AIC: 16410

Number of Fisher Scoring iterations: 1


              Theta:  2.123
          Std. Err.:  0.102

 2 x log-likelihood:  -16394.265
```

```
> tidy(fit_nb, exponentiate = TRUE, conf.int = TRUE)
# A tibble: 7 × 7
  term        estimate std.error statistic  p.value conf.low conf.high
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
1 (Intercept)    1.60     0.0384     12.2  4.25e-34    1.48      1.72
2 educ01         1.91     0.0367     17.7  4.51e-70    1.78      2.06
3 usemeth1       1.84     0.0318     19.1  2.92e-81    1.72      1.96
4 urban1         0.811    0.0317     -6.61 3.72e-11    0.762     0.863
5 electric1      0.932    0.0464     -1.51 1.31e- 1    0.851     1.02
6 radio1         0.918    0.0341     -2.51 1.21e- 2    0.858     0.982
7 bicycle1       1.09     0.0340      2.67 7.64e- 3    1.02      1.17
```

```
> c(theta_hat = theta_hat, theta_se = theta_se)
theta_hat  theta_se
2.1227590 0.1015833
```

```
> comp_aic
         df      AIC
fit_pois  7 17646.54
fit_nb    8 16410.26
```

```
> comp_ll
# A tibble: 2 × 3
  model   logLik    df
  <chr>    <dbl> <int>
1 Poisson -8816.     7
2 NegBin  -8197.     8
```

```
> nb_rate_ratios
# A tibble: 7 × 5
  term        estimate conf.low conf.high  p.value
  <chr>          <dbl>    <dbl>     <dbl>    <dbl>
1 (Intercept)    1.60     1.48      1.72  4.25e-34
2 educ01         1.91     1.78      2.06  4.51e-70
3 usemeth1       1.84     1.72      1.96  2.92e-81
4 urban1         0.811    0.762     0.863 3.72e-11
5 electric1      0.932    0.851     1.02  1.31e- 1
6 radio1         0.918    0.858     0.982 1.21e- 2
7 bicycle1       1.09     1.02      1.17  7.64e- 3
```

```
> cbind(newdata, nb_rate_est, nb_rate_low, nb_rate_high)
  educ0 usemeth urban electric radio bicycle nb_rate_est nb_rate_low nb_rate_high
1     0       0     0        1     1       1    1.496336    1.336035      1.67587
```

- **Model fit:** $\text{AIC}_{\text{Pois}} = 17646.5$ vs. $\text{AIC}_{\text{NB}} = 16410.3$. A likelihood–ratio comparison gives $2(\ell_{\text{NB}} - \ell_{\text{Pois}}) = 1238.3$ (df$= 1$), $p \ll 10^{-10}$, indicating a markedly better fit for NB.
- **Overdispersion:** The estimated k (reported as θ in `glm.nb`) is $\hat{k} \approx 2.12$ (SE ≈ 0.10). Since $\text{Var}(Y) = \mu + \mu^2/k$, a finite $\hat{k}$ implies extra-Poisson variation: $\mu^2/\hat{k}$ is substantial when μ is moderate. This aligns with Problem 1(a) where the residuals–vs–fitted plot suggested increasing variance with the mean.
- **Dispersion check:** Pearson dispersion drops from ≈ 2.14 (Poisson) to ≈ 1.10 (NB), consistent with the NB absorbing overdispersion.
- **Effects:** NB rate ratios are very similar in direction and magnitude to Poisson, but with more realistic standard errors due to the overdispersion parameter.

Overall, the NB model provides a significantly better fit and confirms overdispersion relative to the Poisson assumption in Problem 1(a).

R-code

```r
library(tidyverse)
library(broom)
library(MASS)

set.seed(3100)
theme_set(theme_bw())

# problem 2
  ------------------------------------------------------------------

# read data (same as P1)
fertility <- read.csv("data/fertility_data.csv", header = TRUE) %>%
  as_tibble() %>%
  mutate(
    educ0    = as.factor(educ0),
    usemeth  = as.factor(usemeth),
    urban    = as.factor(urban),
    electric = as.factor(electric),
    radio    = as.factor(radio),
    tv       = as.factor(tv),
    bicycle  = as.factor(bicycle)
  )

# f) Fit NB-GLM for the preferred model from Problem 1 (no TV)
# Poisson reference (from P1)
fit_pois <- glm(
  ceb ~ educ0 + usemeth + urban + electric + radio + bicycle,
  data = fertility,
  family = poisson(link = "log")
)

# Negative binomial (MASS::glm.nb uses log link by default)
fit_nb <- glm.nb(
  ceb ~ educ0 + usemeth + urban + electric + radio + bicycle,
  data = fertility, link = log
)

summary(fit_nb)
tidy(fit_nb, exponentiate = TRUE, conf.int = TRUE)

# Theta (k) estimate and SE
theta_hat <- fit_nb$theta
theta_se  <- fit_nb$SE.theta
c(theta_hat = theta_hat, theta_se = theta_se)

# Compare Poisson vs NB
comp_aic <- AIC(fit_pois, fit_nb)
comp_ll  <- tibble(
  model = c("Poisson", "NegBin"),
```

```r
  logLik = c(as.numeric(logLik(fit_pois)), as.numeric(logLik(fit_nb))),
  df = c(attr(logLik(fit_pois), "df"), attr(logLik(fit_nb), "df"))
)

comp_aic
comp_ll

# LR-style comparison (treat NB as Poisson with extra parameter k)
lr_stat <- 2 * (as.numeric(logLik(fit_nb)) - as.numeric(logLik(fit_pois)))
lr_pval <- pchisq(lr_stat, df = 1, lower.tail = FALSE)
c(lr_stat = lr_stat, df = 1, p_value = lr_pval)

# Pearson dispersion check
disp_pois <- sum(residuals(fit_pois, type = "pearson")^2) / df.residual(fit_pois)
disp_nb   <- sum(residuals(fit_nb,   type = "pearson")^2) / df.residual(fit_nb)
c(pearson_overdisp_poisson = disp_pois, pearson_overdisp_negbin = disp_nb)

# Residuals vs fitted for NB (diagnostic)
resid_nb_plot <- fertility %>%
  mutate(
    fitted = fitted(fit_nb),
    resid  = residuals(fit_nb, type = "pearson")
  ) %>%
  ggplot(aes(x = fitted, y = resid)) +
  geom_point(alpha = 0.3) +
  geom_hline(yintercept = 0, color = "red") +
  labs(x = "Fitted values (NB-GLM)", y = "Pearson residuals") +
  theme_bw()

ggsave("plots/problem2_nb_residuals_vs_fitted.pdf", resid_nb_plot, width = 7, height = 4)

# Rate ratios from NB model (for interpretation in the report)
nb_rate_ratios <- tidy(fit_nb, exponentiate = TRUE, conf.int = TRUE) %>%
  dplyr::select(term, estimate, conf.low, conf.high, p.value)

nb_rate_ratios

# Prediction for the same covariate combo as in P1(e)
newdata <- tibble(
  educ0   = factor(0, levels = c(0,1)),
  usemeth = factor(0, levels = c(0,1)),
  urban   = factor(0, levels = c(0,1)),
  electric = factor(1, levels = c(0,1)),
  radio    = factor(1, levels = c(0,1)),
```

```r
  bicycle  = factor(1, levels = c(0,1))
)

pred_nb <- predict(fit_nb, newdata = newdata, type = "link", se.fit
   = TRUE)

nb_rate_est  <- exp(pred_nb$fit)
nb_rate_low  <- exp(pred_nb$fit - 1.96 * pred_nb$se.fit)
nb_rate_high <- exp(pred_nb$fit + 1.96 * pred_nb$se.fit)

cbind(newdata, nb_rate_est, nb_rate_low, nb_rate_high)
```