B.4 Model diagnostics, repeated events

1. (a) For simplicity, we assume no ties. If $t_1 < t_2 < \cdots < t_l$ are the event times, the partial likelihood is

$$L_P(\beta) = \prod_{j=1}^{l} \frac{\exp\left\{\beta^T \mathbf{x}_{i_j}\right\}}{\sum_{i\in\mathcal{R}_j} \exp\left\{\beta^T \mathbf{x}_i\right\}},$$

where i_j is the individual who had an event at time t_j, and $\mathcal{R}_j$ is the set of those at risk at time t_j. The log partial likelihood is

$$\ell_P(\beta) = \beta^T \sum_{j=l}^{l} \mathbf{x}_{i_j} - \sum_{j=1}^{l} \log\left(\sum_{i\in\mathcal{R}_j} \mathrm{e}^{\beta^T \mathbf{x}_i}\right).$$

(The first term could be simplified to $\beta^T \sum_{i=1}^{n} \delta_i \mathbf{x}_i$, but it is perhaps better understood in the form stated here.) The score function has k-th component

$$\frac{\partial \ell_P}{\partial \beta_k} = \sum_{j=1}^{l}\left(x_{i_j k} - \sum_{i\in\mathcal{R}_j} x_{ik} \frac{\mathrm{e}^{\beta^T \mathbf{x}_i}}{\sum_{i\in\mathcal{R}_j} \mathrm{e}^{\beta^T \mathbf{x}_i}}\right).$$

That is, it is the total difference between k-th covariate of the individual with event at time t_j and the average k-th covariate of those at risk, weighted according to the relative risk with parameter β.

(b) The observed partial information has (k, m) coordinate given by

$$-\frac{\partial^2 \ell_P}{\partial \beta_k \partial \beta_m} = \sum_{j=1}^{l}\left(\sum_{i\in\mathcal{R}_j} x_{ik} x_{im} \frac{\mathrm{e}^{\beta^T \mathbf{x}_i}}{\sum_{i\in\mathcal{R}_j} \mathrm{e}^{\beta^T \mathbf{x}_i}} - \left(\sum_{i\in\mathcal{R}_j} x_{ik} \frac{\mathrm{e}^{\beta^T \mathbf{x}_i}}{\sum_{i\in\mathcal{R}_j} \mathrm{e}^{\beta^T \mathbf{x}_i}}\right)\left(\sum_{i\in\mathcal{R}_j} x_{im} \frac{\mathrm{e}^{\beta^T \mathbf{x}_i}}{\sum_{i\in\mathcal{R}_j} \mathrm{e}^{\beta^T \mathbf{x}_m}}\right)\right)$$

That is, it is the sum of covariances between the k-th and m-th components of individuals at risk at time t_j, where an individual is selected in proportion to relative risk.

2. (a) That would treat the categorical variable as though it were quantitative. That would force the relative risks into particular proportions that have no empirical basis. There may be good reason to expect the relative risk to increase with stage, but not to expect particular proportions.

(b) We could fit the model without any covariates — so just find the Nelson–Aalen estimator— and use that as a basis for adding in the stage as a covariate and checking the martingale residuals. Here we will use age as an additional covariate. So we will fit the model $\alpha_i(t) = \alpha_0(t)e^{\beta\cdot\mathrm{age}}$, and check for the behaviour of `stage` as an additional covariate. We show a box plot in figure B.3, showing the distributions of martingale residuals for the 4 different stages. What we see is that the residuals have essentially the same mean for stages 1 and 2, rise substantially for stage 3, and somewhat less for stage 4.

(c) We would compute the scaled Schoenfeld residuals, and plot them as a function of event time. If the proportional-hazards assumption holds — that is, if the proportionality parameter associated with age is effectively constant — this should stay close to 0, with no apparent patterns or trends.

(d) We compute the Breslow estimator $\widehat{H}_0(t)$ for the baseline hazard. The Cox–Snell residual is $r_i = \widehat{H}_0(T_i)\mathrm{e}^{\beta x_i}$, where T_i is the time for individual i. We then compute a Nelson–Aalen estimator for the right-censored times (r_i, δ_i). If the Cox model is a good fit, the estimated cumulative hazard should look approximately like an upward sloping line through the origin with slope 1.

3. (a) The `R` computation below shows that the coefficient for stage 2 is clearly not statistically significant; the coefficient for stage 3 is borderline ($p = 0.071$); and the coefficient for stage 4 is highly significant ($p = 0.000053$).

```
lar.cph=coxph(Surv(time,delta)~factor(stage)+age,data=larynx)
                  coef exp(coef) se(coef)     z       p
factor(stage)2 0.140       1.15    0.4625 0.303    0.762
factor(stage)3 0.642       1.90    0.3561 1.804    0.071
factor(stage)4 1.706       5.51    0.4219 4.043 5.3e-05
age            0.019       1.02    0.0143 1.335    0.182

Likelihood ratio test=18.3  on 4 df, p=0.00107  n= 90, number of
    events= 50

stage2=(stage==2)
stage3=(stage==3)
stage4=(stage==4)
lar2.cph=coxph(Surv(time,delta)~stage2+stage3+stage4+age,data=larynx)
               coef exp(coef) se(coef)    z       p
stage2TRUE  0.1400    1.1503   0.4625 0.30    0.762
stage3TRUE  0.6424    1.9010   0.3561 1.80    0.071
stage4TRUE  1.7060    5.5068   0.4219 4.04 5.3e-05
age         0.0190    1.0192   0.0143 1.33    0.182

Likelihood ratio test=18.3  on 4 df, p=0.00107
n= 90, number of events= 50
```

(b)

```
require(survival)
require(KMsurv)

data(larynx)
lar.cph=coxph(Surv(time,delta)~age,data=larynx)

coef exp(coef) se(coef)    z    p
age 0.0233      1.02   0.0145 1.61 0.11

Likelihood ratio test=2.63  on 1 df, p=0.105  n= 90, number of events=
     50

lar.fit=survfit(lar.cph)

# The coxph object has a list of times
#   We want to find the index of the time corresponding to individual
    i.
whichtime=sapply(larynx$time,function(t) which(lar.fit$time==t))

cumhaz=-log(lar.fit$surv[whichtime])

beta=lar.cph$coefficients
relrisk=exp(beta*(larynx$age-mean(larynx$age)))
# Baseline hazard is for mean value of covariate

resids=larynx$delta-cumhaz*relrisk
#Note: We could get the same numbers out as lar.cph$residuals
resids.bystage=lapply(1:4,function(i) resids[larynx$stage==i])
boxplot(resids.bystage,xlab='Stage',ylab='Martingale residual')
```

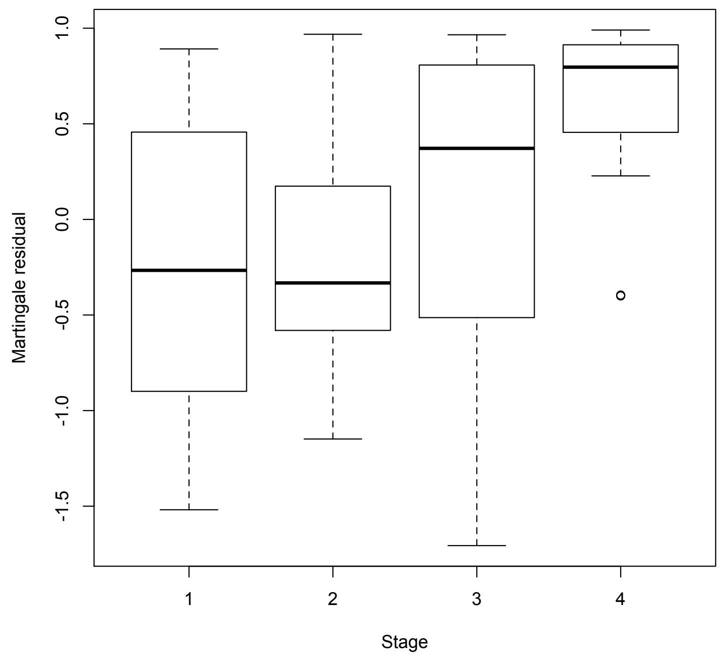

Figure B.3: Box plot of martingale residuals for `larynx` data, stratified by `stage`.

(c) The plot is shown in Figure B.4. We see that there seems to be no effect of the `age` variable until age 70, after which it seems to increase linearly.

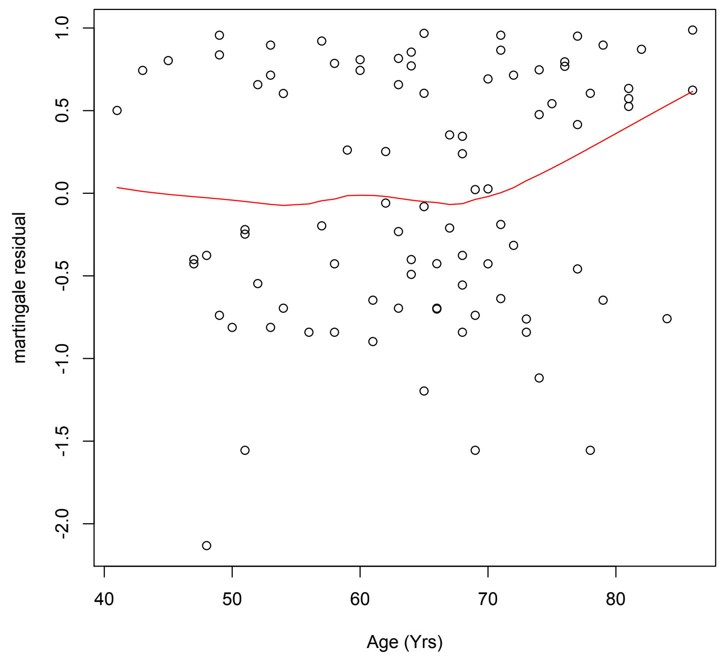

Figure B.4: Plot of martingale residuals against `age` for `larynx` data.

```
########## Residual plot to test age
aord=order(age)
resids=lar.cph2$residuals[aord]
plot(age[aord],resids,xlab='Age (Yrs)',ylab='martingale residual')
lines(lowess(resids~age[aord]),col=2)

########## New model with age starting from 70
newage=pmax(age[aord]-70,0)
lar.cph=coxph(Surv(time,delta)~factor(stage)+newage,data=larynx)
```

(d) The scaled Schoenfeld residuals are plotted with `plot(cox.zph(lar.cph))`.

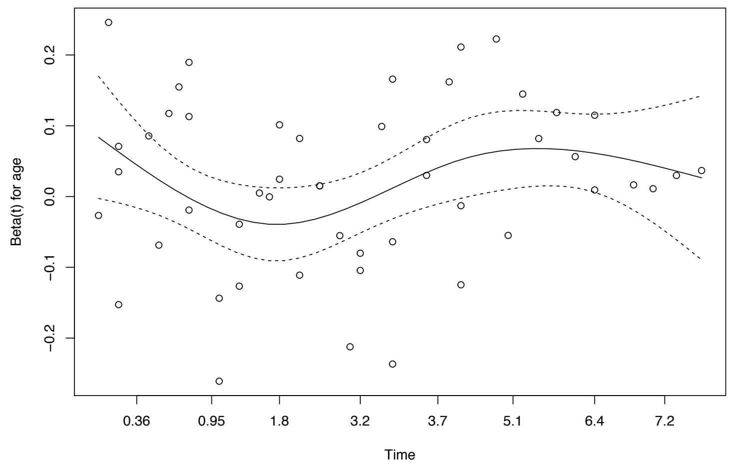

Figure B.5: Plot of scaled Schoenfeld residuals to test whether age parameter is constant.

The output of the function is

```
> cox.zph(lar.cph)
                    rho  chisq      p
factor(stage)2  -0.0158 0.0133 0.9083
factor(stage)3  -0.2599 3.2313 0.0722
factor(stage)4  -0.1105 0.5431 0.4611
age              0.1138 0.8647 0.3524
GLOBAL               NA 4.6755 0.3222
```

This estimates the slope of the change in the various Cox-model parameters over time. None of the slopes is significantly different from 0.

(e) There seems to be a marked curvature of the residual plot, suggesting that the model is underestimating the cumulative hazard later on.

```
lar.cph=coxph(Surv(time,delta)~factor(stage),data=larynx)
lar.fit=survfit(lar.cph)

whichtime=sapply(larynx$time,function(t) which(lar.fit$time==t))

cumhaz=-log(lar.fit$surv[whichtime])

beta=lar.cph$coefficients
relrisk=exp(matrix(beta,1,3) %*% rbind(st2-mean(st2),st3-mean(st3),st4
    -mean(st4)))
coxsnell=c(relrisk*cumhaz)

CS.surv=Surv(coxsnell,delta[aord])
CS.fit=survfit(CS.surv~1)

plot(CS.fit$time,-log(CS.fit$surv),xlab='Time',
ylab='Fitted cumulative hazard for Cox--Snell residuals')
abline(0,1,col=2)
```

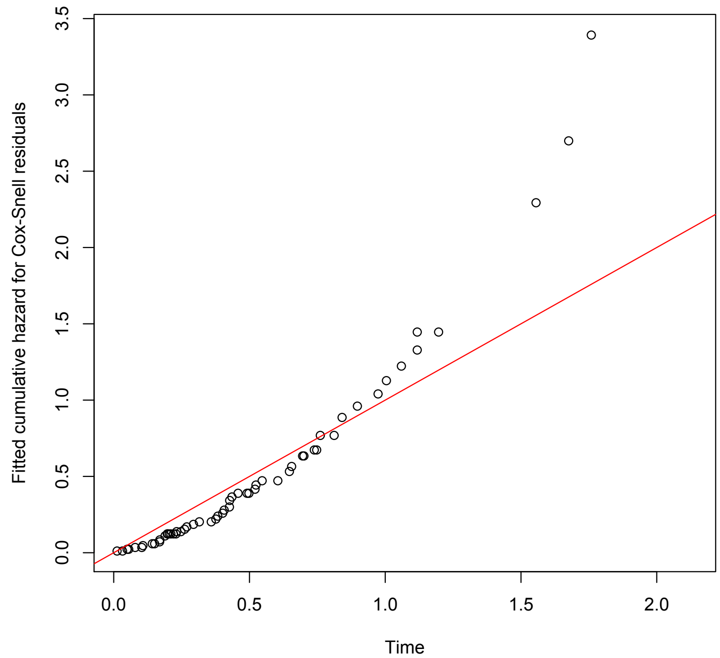

Figure B.6: Cox–Snell residual plot for `larynx` data.

4. (a) The estimator $\mathbf{B}$ is defined in (5.14) as

$$\hat{\mathbf{B}}(t) = \sum_{t_j \le t} \mathbf{X}^{-}(t_j)\, \mathrm{d}\mathbf{N}(t_j).$$

In a small interval of time $[t, t + \mathrm{d}t)$ in which the cumulative hazard covariates $\mathbf{B}$ are incremented by $\mathrm{d}\mathbf{B}(t) = \beta(t)\,\mathrm{d}t$, the expected number of events is incremented by

$$\mathbb{E}\left[\mathrm{d}\mathbf{N}(t)_i\right] = \big(\mathbf{X}(t)\,\mathrm{d}\mathbf{B}(t)\big)_i\,,$$

independent of any past events. If we now condition on an event happening at time $t = t_j$ we can say that

$$\mathbb{E}\left[\mathrm{d}\mathbf{N}(t)_i\right] = \frac{\big(\mathbf{X}(t_j)\beta\big)_i}{\sum_k \big(\mathbf{X}(t_j)\beta\big)_k}.$$

Thus, conditioned on any past information, the expected increment to the martingale residual at $t = t_j$ is

$$\begin{aligned}
\mathbb{E}\left[\mathbf{M}_{res}(t) - \mathbf{M}_{res}(t-)\right] &= \mathbb{E}\left[\mathrm{d}\mathbf{N}(t)\right] - \mathbf{X}(t_j)\mathbf{X}^{-}(t_j)\mathbb{E}\left[\mathrm{d}\mathbf{N}(t)\right] \\
&= \frac{1}{\sum_k \beta^T \mathbf{X}(t_j)\mathbf{1}} \left(I - \mathbf{X}(t)\left(\mathbf{X}(t)^T\mathbf{X}(t)\right)^{-1}\mathbf{X}(t)^T\right)\mathbf{X}(t)\beta \\
&= \frac{1}{\sum_k \beta^T \mathbf{X}(t_j)\mathbf{1}} \left(\mathbf{X}(t) - \mathbf{X}(t)\left(\mathbf{X}(t)^T\mathbf{X}(t)\right)^{-1}\mathbf{X}(t)^T\mathbf{X}(t)\right)\beta \\
&= 0.
\end{aligned}$$

Of course, the expected increment is 0 conditioned on no event at time t. Thus, the expectation of $\mathbf{M}_{res}(t)$ is constant, and since it starts at 0, it is identically 0.

(b) We write $\mathbf{Y}(s)$ for the $n \times n$ diagonal matrix with the at-risk indicators on the diagonal. We note that for right-censored data $\mathbf{Y}(s')\mathbf{Y}(s) = \mathbf{Y}(s)$ for $s' \le s$, and $\mathbf{Y}(s)\,\mathrm{d}\mathbf{N}(s) = \mathrm{d}\mathbf{N}(s)$ because $Y_i(s) = 0$ implies that $\mathrm{d}\mathbf{N}_i(s) = 0$. Thus

$$\mathbf{Y}(s)\,\mathrm{d}\mathbf{M}_{res}(s) = \mathbf{Y}(s)\,\mathrm{d}\mathbf{N}(s) - \mathbf{Y}(s)\mathbf{X}(s)\,\mathrm{d}\mathbf{B}(s) = \mathrm{d}\mathbf{M}_{res}(s).$$

Also, $\mathbf{X}(t) = \mathbf{Y}(t)\mathbf{X}$, where we write $\mathbf{X}$ for $\mathbf{X}(0)$. Since $\mathbf{M}_{res}$ is constant except at times $s = t_j$, we consider the increments at $s = t_j$, obtaining

$$\begin{aligned}
\mathbf{X}^T\,\mathrm{d}\mathbf{M}_{res}(t_j) &= \big(\mathbf{X}^T(t_j) - \mathbf{X}^T(t_j)\mathbf{Y}(t_j)\mathbf{X}\mathbf{X}^-(t_j)\big)\,\mathrm{d}\mathbf{M}(t_j)\\
&= \Big(\mathbf{X}^T - \mathbf{X}^T\mathbf{Y}(t_j)\mathbf{X}\big(\mathbf{X}^T\mathbf{Y}(t_j)\mathbf{X}\big)^{-1}\mathbf{X}^T\Big)\mathbf{Y}(t_j)\,\mathrm{d}\mathbf{M}(t_j)\\
&= \Big(\mathbf{X}^T\mathbf{Y}(t_j) - \big(\mathbf{X}^T\mathbf{Y}(t_j)\mathbf{X}\big)\big(\mathbf{X}^T\mathbf{Y}(t_j)\mathbf{X}\big)^{-1}\mathbf{X}^T\mathbf{Y}(t_j)\Big)\,\mathrm{d}\mathbf{M}(t_j)\\
&= 0.
\end{aligned}$$

Since this is true for all s, and since $\mathbf{M}_{res}(0) = 0$, it must be true that $\mathbf{X}^T\mathbf{M}_{res}(t) = 0$ for all $t \le \tau$.

(c) The equation $\mathbf{X}^T\mathbf{M}_{res}(\tau) = 0$ means that the n-dimensional vector $\mathbf{M}_{res}(\tau)$ is orthogonal to each of the $p+1$ distinct n-dimensional vectors of the coefficients. There is no linear trend with respect to the covariates. (In other words, in the linear regression model predicting $\mathbf{M}_{res}(\tau)$ as a function of the covariates, the coefficients are all 0.)

If the additive hazards model is true, there should be no nonlinear effect of the covariates on the martingale residuals. So one possible model test is to plot the martingale residuals against nonlinear functions of the residuals — for instance, the square of a covariate, or a product of two covariates — and look for trends. This is described briefly in section 4.2.4 of [2], and more extensively in [1].

5. (a) The correct log partial likelihood would have been

$$\ell_P(\beta) = \sum_{j=1}^{k} \beta^T\mathbf{x}_{i_j} - \sum_{j=1}^{k} \log\Big(\sum_{i\in\mathcal{R}_j} \mathrm{e}^{\beta^T\mathbf{x}_i}\Big).$$

The doubled data set will produce two identical events at time t_j, and the risk sets $\mathcal{R}'_j$ will have each individual covariate from $\mathcal{R}_j$ repeated. This produces a log partial likelihood

$$\begin{aligned}
\ell'_P(\beta) &= 2\sum_{j=1}^{k} \beta^T\mathbf{x}_{i_j} - 2\sum_{j=1}^{k} \log\Big(\sum_{i\in\mathcal{R}'_j} \mathrm{e}^{\beta^T\mathbf{x}_i}\Big)\\
&= 2\Big(\sum_{j=1}^{k} \beta^T\mathbf{x}_{i_j} - \sum_{j=1}^{k} \log\Big(2\sum_{i\in\mathcal{R}_j} \mathrm{e}^{\beta^T\mathbf{x}_i}\Big)\\
&= 2\ell_P(\beta) - 2k\log 2.
\end{aligned}$$

So ℓ'_P will be maximised at the same β that maximises ℓ_P.

Since the log likelihood is (up to an additive constant) doubled for the data set with repetition, so is the partial information. The estimated variance is therefore halved, and the SE divided by $\sqrt{2}$.

(b) We know that if the model were correctly specified, we would have $V_n(\hat\beta) \approx J_n(\hat\beta)$ for large n. (This is Fisher's identity, that the variance of the score function is equal to the expected second derivative, that is, the Fisher information.) As discussed above, the log partial likelihood for the duplicated data is twice the log partial likelihood for the correctly specified Cox model on the true data. The variance estimate is made up of squared score functions, so the estimate derived from the duplicated data will be $V'_n(\hat\beta) = 4V_n(\hat\beta)$, while $J'_n(\hat\beta) = 2J_n(\hat\beta)$. Thus we obtain the sandwich estimator from the duplicated data

$$(J'_n)^{-1}V'_n(J'_n)^{-1} = (2J_n)^{-1}(4V_n)(2J_n)^{-1} = (J_n)^{-1}V_n(J_n)^{-1} \approx J_n^{-1},$$

which is the correct variance estimate.

(c)

```
> require(survival)
> require(KMsurv)
> data(bmt)
> bmt$id <- 1:dim(bmt)[1]
> bmt.double <- rbind(bmt,bmt)
> bmcox=coxph(Surv(t2,d3)~z1+z2+z1*z2,data=bmt)
> bmcox.double <- coxph(Surv(t2,d3)~z1+z2+z1*z2,data=bmt.double)
> bmcox.cluster <- coxph(Surv(t2,d3)~z1+z2+z1*z2+cluster(id),data=bmt.
    double)
> bmcox
Call:
coxph(formula = Surv(t2, d3) ~ z1 + z2 + z1 * z2, data = bmt)

coef   exp(coef)   se(coef)        z        p
z1     -0.1075108  0.8980669  0.0348843  -3.082  0.00206
z2     -0.0828731  0.9204680  0.0302421  -2.740  0.00614
z1:z2   0.0034405  1.0034464  0.0009076   3.791  0.00015

Likelihood ratio test=13.29  on 3 df, p=0.004048
n= 137, number of events= 83
> bmcox.double
Call:
coxph(formula = Surv(t2, d3) ~ z1 + z2 + z1 * z2, data = bmt.double)

coef   exp(coef)   se(coef)        z          p
z1     -0.1080799  0.8975559  0.0246902  -4.377  1.20e-05
z2     -0.0833621  0.9200180  0.0214009  -3.895  9.81e-05
z1:z2   0.0034588  1.0034648  0.0006425   5.384  7.30e-08

Likelihood ratio test=26.78  on 3 df, p=6.535e-06
n= 274, number of events= 166
> bmcox.cluster
Call:
coxph(formula = Surv(t2, d3) ~ z1 + z2 + z1 * z2 + cluster(id),
data = bmt.double)

coef   exp(coef)   se(coef)   robust se        z         p
z1     -0.1080799  0.8975559  0.0246902  0.0319525  -3.383  0.000718
z2     -0.0833621  0.9200180  0.0214009  0.0274344  -3.039  0.002377
z1:z2   0.0034588  1.0034648  0.0006425  0.0007601   4.551  5.35e-06

Likelihood ratio test=26.78  on 3 df, p=6.535e-06
n= 274, number of events= 166
```

6. (a) By the law of total probability

$$\begin{aligned}
\mathrm{P}\{\{N=n\} &= \int_0^\infty \frac{t^n}{n!}\mathrm{e}^{-t}\cdot\frac{\lambda^r}{\Gamma(r)}t^{r-1}\mathrm{e}^{-\lambda t}\,\mathrm{d}t\\
&= \frac{\lambda^r}{n!\Gamma(r)}\frac{\Gamma(n+r)}{(\lambda+1)^{n+r}}\\
&= \frac{\Gamma(n+r)}{\Gamma(r)n!}\left(\frac{1}{\lambda+1}\right)^n\left(\frac{\lambda}{\lambda+1}\right)^r\\
&= \binom{n+r-1}{n}p^n(1-p)^r
\end{aligned}$$

where $p=1/(\lambda+1)$. This is the negative binomial distribution with parameters p and r.

(b) (Note: Apologies, the notation of the statement was somewhat confusing, because λ is the true parameter of the NB distribution, but also the nominal Poisson parameter for the erroneous model that we are fitting in this section. I will call this Poisson parameter λ_P.) The MLE for the Poisson distribution is $\hat\lambda_P = n^{-1}\sum N_i$. This will converge to the expected value of the negative binomial, which is r/λ. We would expect this estimator to have variance given by the variance of the Poisson distribution, which is the same as the expected value $r/n\lambda \approx \hat\lambda_P/n$. But the variance of the negative binomial is actually $r(\lambda^{-1}+\lambda^{-2})/n$.

(c) A simple test would be to compare the sample mean to the sample variance. Under the assumption of Poisson distribution the sample variance

$$\hat\sigma^2 := \frac{1}{n-1}\sum(N_i-\hat\lambda_P)^2$$

has expected value

$$\begin{aligned}
\mathbb{E}[\hat\sigma^2] &= \frac{n}{n-1}\mathbb{E}\left[N_i^2-2\hat\lambda_P N_i+\hat\lambda_P^2\right]\\
&= \frac{n}{n-1}\left(\lambda_P^2+\lambda_P-2(\lambda_P^2+\frac{\lambda_P}{n})+\frac{\lambda_P}{n}+\lambda_P^2\right)\\
&= \lambda_P.
\end{aligned}$$

The variance will be approximately (see below)

$$\mathrm{Var}(\hat\sigma^2) \approx \frac{\lambda_P+4\lambda_P^2}{n}+O(n^{-2}).$$

(To compute this, it makes sense to write $N_i=\tilde N+\lambda_P$ and $\hat\lambda_P=\tilde\lambda+\lambda_P$. Then higher powers of $\tilde\lambda$ have expectations $O(n^{-1})$, and products with $\tilde N$ are still of this sort.) Then we can take

$$Z := \frac{\hat\sigma^2-\hat\lambda_P}{\sqrt{(\hat\lambda_P+4\hat\lambda_P^2)/n}}$$

as an approximately standard normal test statistic.

Alternatively, we could perform a likelihood ratio test.

To calculate the variance of $\hat\sigma^2$ we rewrite it as

$$\begin{aligned}
\hat\sigma^2 &= \frac{1}{n-1}\sum_{i=1}^n\Big(\frac{1}{n}\sum_{j\neq i}(N_i-N_j)\Big)^2\\
&= \frac{1}{(n-1)n^2}\Big(\sum_{i=1}^n\sum_{j\neq i}(N_i-N_j)^2+\sum i,j,j' \text{ all distinct}(N_i-N_j)(N_i-N_{j'})\Big).
\end{aligned}$$

The variance will be the sums of all the variances of the individual terms, plus the sums of all the covariances of two different terms. We could calculate this exactly, but there would be a lot of different terms to keep track of. We note, though, that for large n we can approximate this by calculating only the first-order term in n. Each individual term contributes on the order of n^{-6} to the sum. In total the number of terms is on the order of n^3, so their contribution is on the order of n^{-3} in total. Covariances between terms involving entirely distinct N_i, N_j will be 0 (by independence), so we need consider only covariances with duplications. The first sum contributes only $O(n^3)$ of these, and pairs between the first and second sum contribute only $O(n^4)$. The only contribution on the order of n^5 is that made by pairs from the second sum. These could have the same i or the same j or j'.

In total, there are $n(n-1)(n-2)$ terms. If we are considering pairs of the form $(N_i - N_j)(N_i - N_{j'})$ and $(N_i - N_{j''})(N_i - N_{j'''})$ with j's all distinct there will be $n(n-1)(n-2)(n-3)(n-4)$ pairs (since we are taking account of order), and each one contributes

$$\mathrm{Cov}\big((N_i - N_j)(N_i - N_{j'}), (N_i - N_{j''})(N_i - N_{j'''})\big) = (3\lambda_P^2 + \lambda_P) - \lambda_P^2 = 2\lambda_P^2 + \lambda_P.$$

(since the central fourth moment of a Poisson random variable with parameter λ is $3\lambda^2 + \lambda$). Considering pairs of the form $(N_i - N_j)(N_i - N_{j'})$ and $(N_{i'} - N_j)(N_{i'} - N_{j''})$ we see that there are also $2n(n-1)(n-2)(n-3)(n-4)$ pairs (since, having chosen i, j, j' for the first term we have two possibilities of which one to repeat for the second), and each contributes

$$\mathrm{Cov}\big((N_i - N_j)(N_i - N_{j'}), (N_{i'} - N_j)(N_{i'} - N_{j''})\big) = \lambda_P^2.$$

Putting these together, we get the indicated approximation to the variance.