APPLIED STATISTICS, MATH10096

Workshop 3: solutions

1. (a) The required design matrix and parameter vector are as follows

$$\begin{pmatrix} 1 & 1 & 0 & 1 & 0 & x_1 \\ 1 & 1 & 0 & 1 & 0 & x_2 \\ 1 & 1 & 0 & 0 & 1 & x_3 \\ 1 & 1 & 0 & 0 & 1 & x_4 \\ 1 & 0 & 1 & 1 & 0 & x_5 \\ 1 & 0 & 1 & 1 & 0 & x_6 \\ 1 & 0 & 1 & 0 & 1 & x_7 \\ 1 & 0 & 1 & 0 & 1 & x_8 \end{pmatrix} \begin{pmatrix} \alpha \\ \nu_1 \\ \nu_2 \\ \gamma_1 \\ \gamma_2 \\ \beta \end{pmatrix}.$$

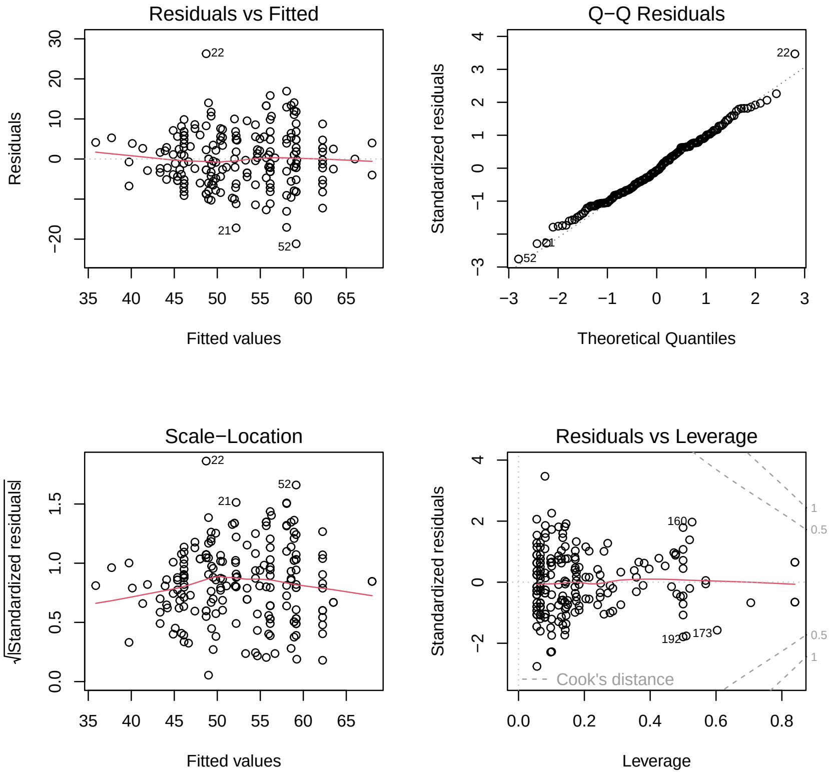
Obviously this model is not identifiable, as the sum of the second and third columns gives the first column and, similarly, the sum of the fourth and fifth columns gives also the first column.

(b) Setting $\nu_1 = \gamma_1 = 0$ is one possibility for achieving identifiability, in which case,

$$\begin{pmatrix} 1 & 0 & 0 & x_1 \\ 1 & 0 & 0 & x_2 \\ 1 & 0 & 1 & x_3 \\ 1 & 0 & 1 & x_4 \\ 1 & 1 & 0 & x_5 \\ 1 & 1 & 0 & x_6 \\ 1 & 1 & 1 & x_7 \\ 1 & 1 & 1 & x_8 \end{pmatrix} \begin{pmatrix} \alpha \\ \nu_2 \\ \gamma_2 \\ \beta \end{pmatrix}.$$

2. (a) We implement the required model as follows:

```r
require(faraway)
res_inter <- lm(math ~ (gender + race + ses + schtyp + prog)^2,
                data = hsb)
par(mfrow = c(2, 2))
plot(res_inter)
```



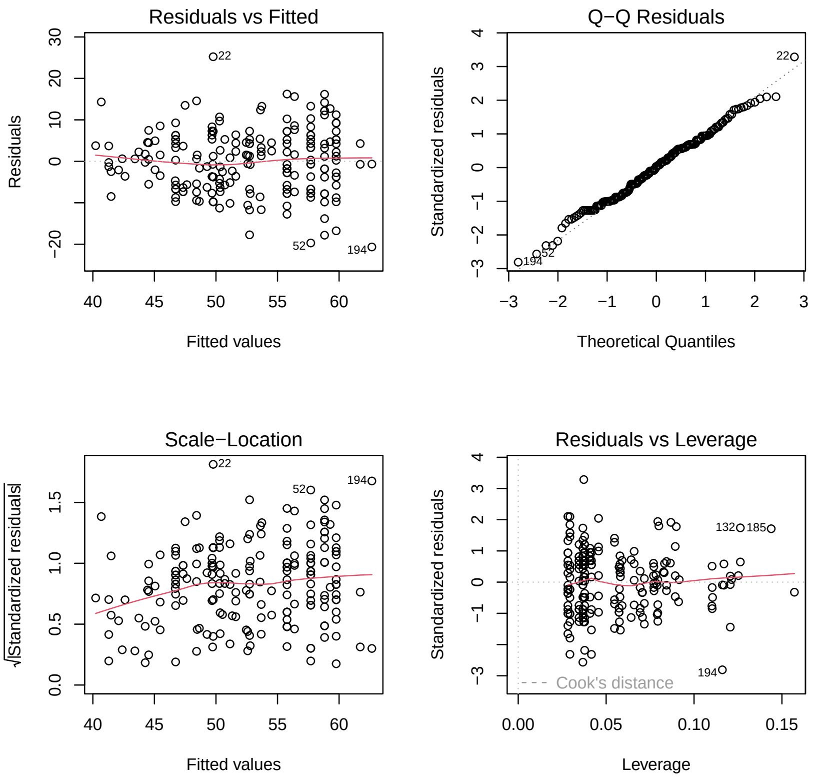
Everything appears to be fine with the residual plots. The dispersion of the residuals is smaller for both low and high fitted values. However, since there are only a few observations in these regions, it is hard to conclude that the variance of the error term is not constant.

(b) Let us implement the backwards model selection. To do this in an automatic fashion we will use the `step` function, which performs model selection based on the AIC criterion.

```
step(res_inter, direction = "backward")

## Start:  AIC=862.91
## math ~ (gender + race + ses + schtyp + prog)^2
##
##                  Df Sum of Sq     RSS    AIC
## - ses:prog        4    116.58 10042.3 857.25
## - race:schtyp     3     30.99  9956.7 857.54
## - gender:race     3     94.98 10020.7 858.82
## - gender:prog     2      5.65  9931.3 859.03
## - race:ses        6    472.36 10398.0 860.21
## - ses:schtyp      2    101.05 10026.7 860.94
## - gender:schtyp   1      3.36  9929.1 860.98
## - schtyp:prog     2    134.15 10059.8 861.60
## <none>                         9925.7 862.91
## - gender:ses      2    240.56 10166.2 863.70
## - race:prog       6    860.51 10786.2 867.54
##
## Step:  AIC=857.25
## math ~ gender + race + ses + schtyp + prog + gender:race + gender:ses +
##     gender:schtyp + gender:prog + race:ses + race:schtyp + race:prog +
##     ses:schtyp + schtyp:prog
##
##                 Df Sum of Sq   RSS    AIC
## - race:schtyp    3     33.28 10076 851.91
## - gender:prog    2      2.42 10045 853.30
## - gender:race    3    132.50 10175 853.87
## - ses:schtyp     2     62.44 10105 854.49
## - race:ses       6    480.18 10522 854.59
## - gender:schtyp  1      3.70 10046 855.32
## - schtyp:prog    2    194.69 10237 857.09
## <none>                       10042 857.25
## - gender:ses     2    253.63 10296 858.24
## - race:prog      6    779.78 10822 860.20
##
## Step:  AIC=851.91
## math ~ gender + race + ses + schtyp + prog + gender:race + gender:ses +
##     gender:schtyp + gender:prog + race:ses + race:prog + ses:schtyp +
##     schtyp:prog
##
##                 Df Sum of Sq   RSS    AIC
## - gender:prog    2      2.01 10078 847.95
## - gender:race    3    123.11 10199 848.34
## - ses:schtyp     2     61.62 10137 849.13
## - gender:schtyp  1      1.02 10077 849.93
## - race:ses       6    560.18 10636 850.73
## - schtyp:prog    2    197.49 10273 851.79
## <none>                       10076 851.91
## - gender:ses     2    291.52 10367 853.61
## - race:prog      6    821.23 10897 855.58
##
## Step:  AIC=847.95
## math ~ gender + race + ses + schtyp + prog + gender:race + gender:ses +
##     gender:schtyp + race:ses + race:prog + ses:schtyp + schtyp:prog
##
##                 Df Sum of Sq   RSS    AIC
## - gender:race    3    126.52 10204 844.44
## - ses:schtyp     2     62.18 10140 845.18
## - gender:schtyp  1      1.80 10079 845.99
## - race:ses       6    559.76 10637 846.76
## - schtyp:prog    2    198.43 10276 847.85
## <none>                       10078 847.95
## - gender:ses     2    305.63 10383 849.93
## - race:prog      6    830.36 10908 851.79
##
## Step:  AIC=844.44
## math ~ gender + race + ses + schtyp + prog + gender:ses + gender:schtyp +
##     race:ses + race:prog + ses:schtyp + schtyp:prog
##
##                  Df Sum of Sq   RSS    AIC
## - ses:schtyp     2     85.20 10289 842.11
## - gender:schtyp  1      1.36 10205 842.47
## - race:ses       6    578.45 10782 843.47
## - schtyp:prog    2    192.88 10397 844.19
## <none>                       10204 844.44
## - gender:ses     2    315.99 10520 846.54
## - race:prog      6    783.17 10987 847.23
##
## Step:  AIC=842.11
## math ~ gender + race + ses + schtyp + prog + gender:ses + gender:schtyp +
##     race:ses + race:prog + schtyp:prog
##
##                  Df Sum of Sq   RSS    AIC
## - gender:schtyp  1      3.51 10293 840.18
## - race:ses       6    545.29 10835 840.44
## <none>                       10289 842.11
## - gender:ses     2    290.78 10580 843.68
## - schtyp:prog    2    313.59 10603 844.11
## - race:prog      6    813.51 11103 845.33
##
## Step:  AIC=840.18
## math ~ gender + race + ses + schtyp + prog + gender:ses + race:ses +
##     race:prog + schtyp:prog
##
##                Df Sum of Sq   RSS    AIC
## - race:ses     6    542.16 10835 838.44
## <none>                     10293 840.18
## - gender:ses   2    293.72 10586 841.80
## - schtyp:prog  2    310.28 10603 842.12
## - race:prog    6    813.13 11106 843.38
##
## Step:  AIC=838.44
## math ~ gender + race + ses + schtyp + prog + gender:ses + race:prog +
##     schtyp:prog
##
##                Df Sum of Sq   RSS    AIC
## - race:prog    6    492.77 11328 835.34
## <none>                     10835 838.44
## - gender:ses   2    236.90 11072 838.77
## - schtyp:prog  2    264.78 11100 839.27
##
## Step:  AIC=835.34
## math ~ gender + race + ses + schtyp + prog + gender:ses + schtyp:prog
##
##                Df Sum of Sq   RSS    AIC
## - schtyp:prog  2    214.96 11543 835.10
## <none>                     11328 835.34
## - gender:ses   2    318.31 11646 836.88
## - race         3   1116.18 12444 848.13
##
## Step:  AIC=835.1
## math ~ gender + race + ses + schtyp + prog + gender:ses
##
##               Df Sum of Sq   RSS    AIC
## - schtyp      1      27.18 11570 833.57
## <none>                    11543 835.10
## - gender:ses  2    314.23 11857 836.47
## - race        3   1131.77 12674 847.80
## - prog        2   2876.09 14419 875.59
##
## Step:  AIC=833.57
## math ~ gender + race + ses + prog + gender:ses
##
##               Df Sum of Sq   RSS    AIC
## <none>                    11570 833.57
## - gender:ses  2    301.89 11872 834.72
## - race        3   1125.16 12695 846.13
## - prog        2   2883.26 14453 874.07
##
## Call:
## lm(formula = math ~ gender + race + ses + prog + gender:ses,
##     data = hsb)
##
## Coefficients:
##          (Intercept)             gendermale              raceasian
##              53.7268                -2.0658                 8.9369
##         racehispanic              racewhite                 seslow
##               0.7998                 6.0392                -3.3854
##            sesmiddle            proggeneral           progvocation
##              -3.9952                -6.0894                -9.0475
##    gendermale:seslow  gendermale:sesmiddle
##               0.1996                 5.1173
```

```
res_final <- lm(math ~ gender + race + ses + prog + gender:ses, data = hsb)
par(mfrow = c(2, 2))
plot(res_final)
```



```r
summary(res_final)
```

```
## 
## Call:
## lm(formula = math ~ gender + race + ses + prog + gender:ses, 
##     data = hsb)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -20.6637  -5.7672   0.2553   5.2340  25.2253 
## 
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)    
## (Intercept)            53.7268     2.3461  22.901  < 2e-16 ***
## gendermale             -2.0658     2.0740  -0.996  0.32050    
## raceasian               8.9369     2.9972   2.982  0.00324 ** 
## racehispanic            0.7998     2.3973   0.334  0.73904    
## racewhite               6.0392     1.9597   3.082  0.00237 ** 
## seslow                 -3.3854     2.0929  -1.618  0.10741    
## sesmiddle              -3.9952     1.8749  -2.131  0.03439 *  
## proggeneral            -6.0894     1.4372  -4.237 3.54e-05 ***
## progvocation           -9.0475     1.3940  -6.490 7.34e-10 ***
## gendermale:seslow       0.1996     3.2281   0.062  0.95077    
## gendermale:sesmiddle    5.1173     2.6221   1.952  0.05246 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.824 on 189 degrees of freedom
## Multiple R-squared:  0.3376,Adjusted R-squared:  0.3025
## F-statistic: 9.631 on 10 and 189 DF,  p-value: 6.52e-13
```

The final model contains the effect of gender, race, socioeconomic class, high school program, and the interaction effect between gender and socioeconomic class. Everything seems to be fine with the residual plots. Let us write down mathematically this final model. To this end, let us define the following dummy variables:

$$\text{gender}_{male,i} = \begin{cases} 1, \text{if student } i \text{ is male,} \\ 0, \text{ otherwise.} \end{cases}$$

$$\text{race}_{asian,i} = \begin{cases} 1, \text{if student } i \text{ is asian,} \\ 0, \text{ otherwise.} \end{cases}$$

$$\text{race}_{hispanic,i} = \begin{cases} 1, \text{if student } i \text{ is hispanic,} \\ 0, \text{ otherwise.} \end{cases}$$

$$\text{race}_{white,i} = \begin{cases} 1, \text{if student } i \text{ is white,} \\ 0, \text{ otherwise.} \end{cases}$$

$$\text{ses}_{low,i} = \begin{cases} 1, \text{if student } i \text{ is from a low socioeconomic class ,} \\ 0, \text{ otherwise.} \end{cases}$$

$$\text{ses}_{middle,i} = \begin{cases} 1, \text{if student } i \text{ is from a middle socioeconomic class ,} \\ 0, \text{ otherwise.} \end{cases}$$

$$\text{prog}_{general,i} = \begin{cases} 1, \text{if student } i \text{ is in a general high school program ,} \\ 0, \text{ otherwise.} \end{cases}$$

$$\text{prog}_{vocation,i} = \begin{cases} 1, \text{if student } i \text{ is in a vocational high school program ,} \\ 0, \text{ otherwise.} \end{cases}$$

So, the baseline for gender is female, for race is african american, for socioeconomic class is high, and for high school program is academic. We can then write the model as

$$\begin{aligned} y_i &= \mu + \alpha_{male}\text{gender}_{male,i} + \beta_{asian}\text{race}_{asian,i} + \beta_{hispanic}\text{race}_{hispanic,i} + \beta_{white}\text{race}_{white,i} \\ &\quad + \gamma_{low}\text{ses}_{low,i} + \gamma_{middle}\text{ses}_{middle,i} + \delta_{general}\text{prog}_{general,i} + \delta_{vocation}\text{prog}_{vocation,i} \\ &\quad + \nu_{male,low}\text{gender}_{male,i}\text{ses}_{low,i} + \nu_{male,middle}\text{gender}_{male,i}\text{ses}_{middle,i} + \epsilon_i, \end{aligned}$$

with $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$. In this model μ represents the expected math score for an african american female, from a high socio economic status, who is enrolled in an academic high school program.

3. We observe that, among students of the same gender, belonging to the same socioeconomic class, and enrolled in the same high school program, the expected difference in math scores is 8.94 points between asian and african american students, 0.80 points between hispanic and african american students, and 6.04 points between white and african american students.

   In a similar fashion, among students of the same gender and race and belonging to the same socioeconomic class, the expected difference in math scores is -6.1 points between students enrolled in a general and in an academic high school program and of -9.05 points between students enrolled in a vocational and in an academic high school program.

   In turn, any comments regarding the effect of gender on math scores are dependent on the socioeconomic class, and vice versa. For instance, among students of the same race and enrolled in the same high school program, the expected difference in math scores between a male and a female student, both belonging to a high socioeconomic class (baseline) is -2.07 points. Further, among students of the same race and enrolled in the same high school program, the expected difference in math scores between a male and a female student, both belonging to a low socioeconomic class is $-2.0658 + 0.1996 = -1.87$ points.

4. Neither is a GLM: in both case the random response variable (rather than its expected value) is determined by the linear predictor. This sort of model can not be written in standard GLM form.

5. Taking logs yields

$$\log(\mu_i) = \log(n) + \log(\gamma_k) + \log(\alpha_j) \text{ if } y_i \text{ is gender } k \text{ and faith } j.$$

   Defining $\tau = \log(n)$, $\beta_k = \log(\gamma_k)$ and $\delta_j = \log(\alpha_j)$, we get

$$\log(\mu_i) = \tau + \beta_k + \delta_j \text{ if } y_i \text{ is gender } k \text{ and faith } j.$$

   with $Y_i \underset{\text{indep.}}{\sim} \text{Poi}(\mu_i)$ — clearly a GLM. If the data are ordered FB, FN, MB, MN (using obvious initials) then the linear predictor is

$$\begin{pmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \tau \\ \beta_1 \\ \beta_2 \\ \delta_1 \\ \delta_2 \end{pmatrix}.$$

   Obviously this is not identifiable, but by setting $\beta_1 = \delta_1 = 0$ (therefore removing columns 2 and 4 of the design matrix)

$$\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \tau \\ \beta_2 \\ \delta_2 \end{pmatrix}.$$

6. A 'logistic regression' model might be appropriate (as used for the heart attack data in the notes). In particular, viewing each of the 6 o-rings at each launch temperature as independent binomial trials, with a probability of failure that depends on temperature, gives:

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_1 + \beta_2 \texttt{temp}_i, \quad y_i \overset{\text{indep.}}{\sim} \text{binom}(p_i, 6),$$

where p_i denotes the probability of o-ring failure/damage at temperature temp_i and y_i denotes the number of damaged o-rings, out of 6, at temperature temp_i. So the model matrix would have i^{th} row $[1, \texttt{temp}_i]$.

7. Let us start by writing $f(y;\mu)$ as $\exp\{\log f(y;\mu)\}$

$$\log f(y;\mu) = -\log\mu - \frac{1}{\mu}y \Rightarrow \exp\{\log f(y;\mu)\} = \exp\left\{-\log\mu - \frac{1}{\mu}y\right\} = \exp\left\{\frac{y(-1/\mu) - \log\mu}{1} + 0\right\}.$$

This is in the exponential family of distributions with:

- $\theta = -\frac{1}{\mu} \Rightarrow \mu = -\frac{1}{\theta}$.
- $b(\theta) = \log\mu = \log\left(\frac{1}{-\theta}\right) = -\log(-\theta)$.
- $a(\phi) = \phi = 1$.
- $c(y,\phi) = 0$.

Note that in fact we already knew from the lectures that the exponential distribution belongs to the exponential family of distributions since it is a particular case of a Gamma distribution with $\nu = 1$. Now, moving to the second part of the question. We have proved that $E[Y] = b'(\theta)$ and in this case $b'(\theta) = -\frac{1}{\theta} = \mu$. We have also proved that $\text{var}(Y) = \frac{\phi b''(\theta)}{\omega}$ and in this case $\omega = \phi = 1$ and $b''(\theta) = \frac{1}{\theta^2}$ leading to $\text{var}(Y) = \frac{1}{\theta^2} = \mu^2$. Notice that in this case $\text{V}(\mu) = b''(\theta)/\omega = \mu^2 = \text{var}(Y)$.

8. (a) By definition,

$$\mu \equiv \mathbb{E}[Y] = \sum_{y=0}^{1} y f(y;p) = 0 \times (1-p) + 1 \times p = p.$$

(b) We can rewrite the probability mass function of the Bernoulli distribution in terms of its mean by simply replacing p by μ. We thus have that

$$\log f(y;\mu) = y\log\mu + (1-y)\log(1-\mu),$$

leading to

$$\begin{aligned}\exp\{\log f(y;\mu)\} &= \exp\{y\log\mu + (1-y)\log(1-\mu)\}\\ &= \exp\{y\log\mu + \log(1-\mu) - y\log(1-\mu)\}\\ &= \exp\left\{y\log\left(\frac{\mu}{1-\mu}\right) + \log(1-\mu)\right\}\\ &= \exp\left\{\frac{y\log\left(\frac{\mu}{1-\mu}\right) - [-\log(1-\mu)]}{1} + 0\right\}.\end{aligned}$$

This is in the exponential family of distributions with:

- $\theta = \log\left(\frac{\mu}{1-\mu}\right) \Rightarrow \mu = \frac{e^\theta}{1+e^\theta}$.
- $b(\theta) = -\log(1-\mu) = \log(1+e^\theta)$.
- $a(\phi) = \phi = 1$.
- $c(y,\phi) = 0$.

9. As before, let's write $f(y;\mu,\gamma) = \exp\{\log f(y;\mu,\gamma)\}$. We have that

$$\begin{aligned}\log f(y;\mu,\gamma) &= \frac{1}{2}\log(\gamma) - \frac{1}{2}\log(2\pi y^3) - \frac{\gamma(y-\mu)^2}{2\mu^2 y}\\ &= \frac{1}{2}\log(\gamma) - \frac{1}{2}\log(2\pi y^3) - \frac{\gamma(y^2-2\mu y+\mu^2)}{2\mu^2 y}\\ &= \frac{1}{2}\log(\gamma) - \frac{1}{2}\log(2\pi y^3) - \frac{\gamma y}{2\mu^2} + \frac{\gamma}{\mu} - \frac{\gamma}{2y}\end{aligned}$$

and so

$$\begin{aligned}\exp\{\log f(y;\mu,\gamma)\} &= \exp\left\{\frac{1}{2}\log(\gamma) - \frac{1}{2}\log(2\pi y^3) - \frac{\gamma y}{2\mu^2} + \frac{\gamma}{\mu} - \frac{\gamma}{2y}\right\}\\ &= \exp\left\{\frac{y[-1/(2\mu^2)] - (-1/\mu)}{1/\gamma} + \left(\frac{1}{2}\log\gamma - \frac{1}{2}\log(2\pi y^3) - \frac{\gamma}{2y}\right)\right\}\end{aligned}$$

This is in the exponential family of distributions with:

- $\theta = -\frac{1}{2\mu^2}$ (note that $\theta < 0$) $\Rightarrow \mu = \sqrt{-\frac{1}{2\theta}}$.
- $b(\theta) = -\frac{1}{\mu} = -\sqrt{-2\theta}$.
- $a(\phi) = \phi = \frac{1}{\gamma}$.
- $c(y,\phi) = \frac{1}{2}\log\gamma - \frac{1}{2}\log(2\pi y^3) - \frac{\gamma}{2y} = -\frac{1}{2}\log(\gamma^{-1}2\pi y^3) - \frac{\gamma}{2y} = -\frac{1}{2}\left[\log(\phi 2\pi y^3) + \frac{1}{\phi y}\right]$.