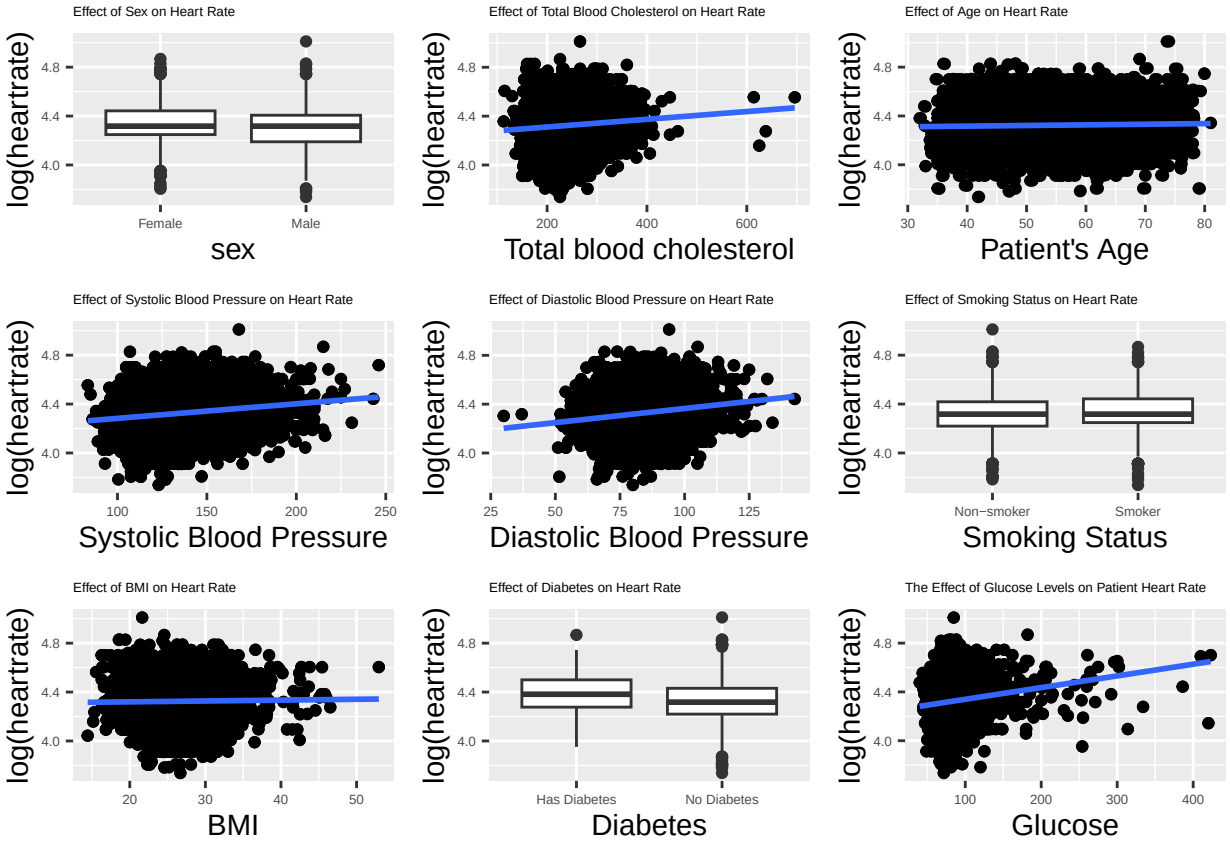
Cardiovascular Health

Jillian Warburton and Mary Curtis

2023-03-16

1. Create exploratory plots of looking at the relationship between `log(HEARTRTE)` (the response variable) and some of the explanatory variables. Comment on any general relationships you see from the data.

```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



It does not appear that there is a significant difference in heart rate between the two sexes. Total blood cholesterol has a weak, positive, linear relationship with heart rate. It does not appear that there is a

relationship between a patient's age and their heart rate? There do appear to be a few older patients that have a much higher heart rate. Systolic blood pressure has a moderately strong linear relationship with heart rate. The same can be said of diastolic blood pressure. There does not seem to be a significant different in heart rate between smokers and non-smokers. BMI has a very weak relationship between heart rate. However, it appears that those with higher BMIs generally have higher heart rates. Those that have diabetes seem to have a higher heart rate on average than those without diabetes. Glucose appears to have a strong, positive linear relationship with heart rate.

2. Fit an independent MLR model with a linear effect of all variables except `RANDID` and `PERIOD`. Explore the residuals to see if there is evidence of correlation within a patients from period to period (visit to visit).

```
data_lm <- lm(data = data, formula = HEARTRTE ~ SEX + TOTCHOL + AGE + SYSBP + DIABP +
                CURSMOKE + BMI + DIABETES + BPMEDS + GLUCOSE)

cor(matrix(data = data_lm$residuals, nrow = 1734, byrow = T))
```

```
##           [,1]      [,2]      [,3]
## [1,] 1.0000000 0.4448351 0.3809072
## [2,] 0.4448351 1.0000000 0.4882654
## [3,] 0.3809072 0.4882654 1.0000000
```

```
# where 1734 is the number of individuals in the data set
```

This correlation matrix clearly shows that there is correlation within patients from period to period.

3. To determine an appropriate correlation structure to use, fit a longitudinal MLR model with an AR1, MA1 and general symmetric correlation matrix within each patient but independent across patients. Compare the model fits using AIC (which can be extracted from a `gls()` object using `AIC()`).

```
ar1 <- gls(model = log(HEARTRTE) ~ SEX + TOTCHOL + AGE + SYSBP + DIABP + CURSMOKE +
             BMI + DIABETES + BPMEDS + GLUCOSE,
    data = data,
    correlation = corAR1(form = ~PERIOD|RANDID),
    method = "ML")

AIC(ar1)
```

```
## [1] -5868.104
```

```
# MA1
ma1 <- gls(model = log(HEARTRTE) ~ SEX + TOTCHOL + AGE + SYSBP + DIABP + CURSMOKE +
             BMI + DIABETES + BPMEDS + GLUCOSE,
    data = data,
    correlation = corARMA(q = 1, form = ~PERIOD|RANDID),
    method = "ML")

AIC(ma1)
```

```
## [1] -5647.617
```

```
symm <- gls(model = log(HEARTRTE) ~ SEX + TOTCHOL + AGE + SYSBP + DIABP + CURSMOKE +
              BMI + DIABETES + BPMEDS + GLUCOSE,
    data = data,
    correlation = corSymm(form=~1:3|RANDID),
    method = "ML")

AIC(symm)
```

```
## [1] -5939.495
```

The symmetric correlation model has the lowest AIC value. Thus, I will go with this one.

4. Write out your model for analyzing the Tachycardia data in terms of parameters. Explain and interpret any parameters associated with the model.

The model for this data set is $\mathbf{y} \sim \mathcal{MVN}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{B})$, with those variables expanded below.

$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$

$\boldsymbol{\epsilon} \sim \mathcal{MVN}(\mathbf{0}, \sigma^2\mathbf{B})$

Specifically, this is a constant, general symmetric correlation structure within `RANDID` (Patient):

Notation: $\mathbf{y} = \begin{bmatrix} \text{heart rate}_1 \\ \text{heart rate}_2 \\ \vdots \\ \text{heart rate}_{5202} \end{bmatrix}$

where heart rate_i is the measured heart rate of a patient.

$$\mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,10} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,10} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{5202,1} & x_{5202,2} & \dots & x_{5202,10} \end{bmatrix}$$

where $x_{i,1} = \begin{cases} 1 & \text{if patient } i \text{ is Male} \\ 0 & \text{if patient } i \text{ is Female} \end{cases}$.

where $x_{i,2} =$ Total blood cholesterol of patient i.

where $x_{i,3} =$ Age of patient i.

where $x_{i,4} =$ Systolic blood pressure of patient i.

where $x_{i,5} =$ Diastolic blood pressure of patient i.

where $x_{i,6} = \begin{cases} 1 & \text{if patient } i \text{ is a smoker} \\ 0 & \text{otherwise} \end{cases}$.

where $x_{i,7} =$ Body Mass Index of patient i.

where $x_{i,8} = \begin{cases} 1 & \text{if patient } i \text{ has diabetes} \\ 0 & \text{otherwise} \end{cases}$.

where $x_{i,9} = \begin{cases} 1 & \text{if patient } i \text{ is on medication} \\ 0 & \text{otherwise} \end{cases}$.

where $x_{i,10}$ = Glucose level of patient i.

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_{\text{Sex}} \\ \beta_{\text{Totchol}} \\ \beta_{\text{Age}} \\ \beta_{\text{Sysbp}} \\ \beta_{\text{Diabp}} \\ \beta_{\text{Cursmoke}} \\ \beta_{\text{Bmi}} \\ \beta_{\text{Diabetes}} \\ \beta_{\text{Bpmeds}} \\ \beta_{\text{Glucose}} \end{bmatrix}.$$

where β_{Totochol} is the change in a patient's heart rate for one unit increase in the patient's total blood cholesterol, holding all else constant.

$$\mathbf{B} = \begin{bmatrix} \mathbf{R} & 0 & 0 & \dots & 0 \\ 0 & \mathbf{R} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & 0 & 0 \\ 0 & 0 & 0 & \dots & \mathbf{R} \end{bmatrix}.$$

$$\mathbf{R} = \begin{bmatrix} 1 & \rho(t_1, t_2) & \rho(t_1, t_3) \\ \rho(t_2, t_1) & 1 & \rho(t_2, t_3) \\ \rho(t_3, t_1) & \rho(t_3, t_2) & 1 \end{bmatrix},$$

where $\rho(y_{t_1}, y_{t_2}) = \rho(|t_1 - t_2|)$ is the correlation between an observation at time t_1 and an observation at time t_2,

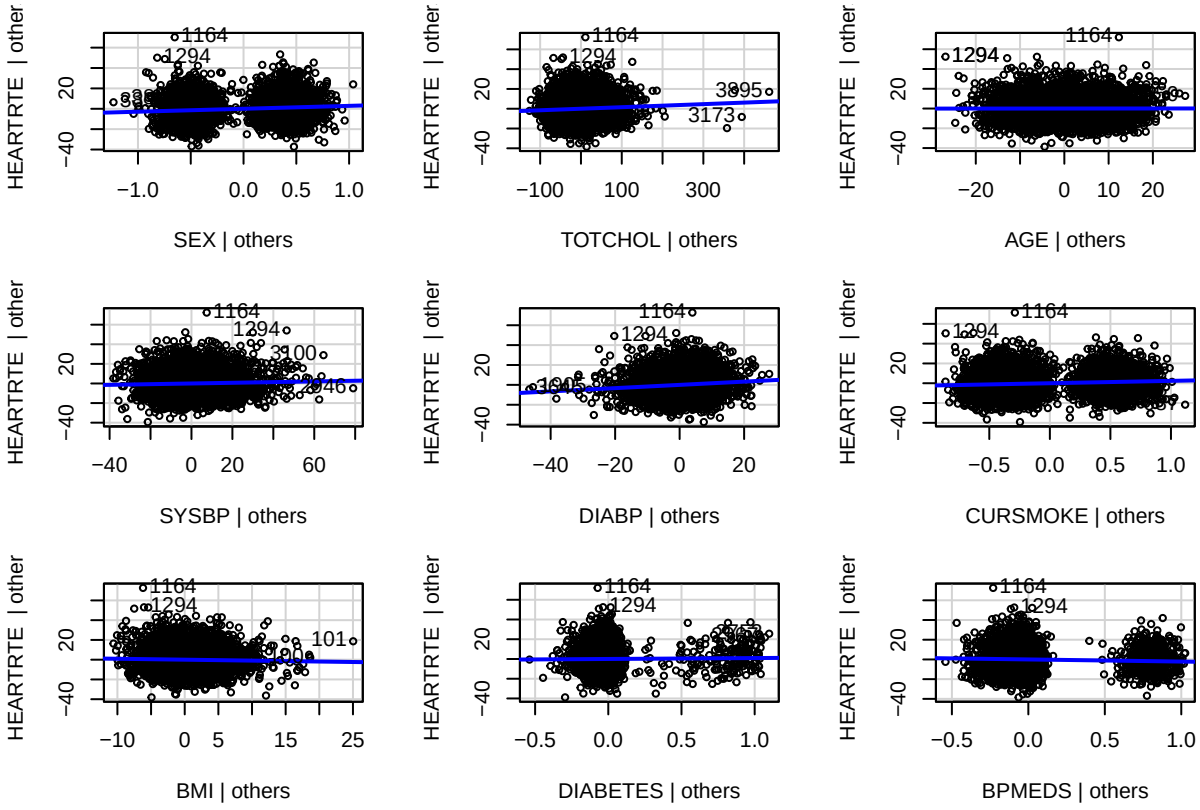
where $t = 1, 2, 3$ is the time period (visit) at which the patient was evaluated.

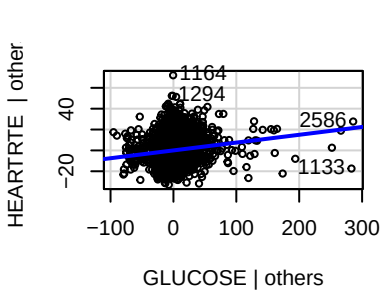
5. Fit your longitudinal model and validate any assumptions you made to fit the model.

```
data_gls <- symm
```

Linearity

```
avPlots(data_lm, ask = FALSE)
```


HEARTRTE | other
SEX | others
TOTCHOL | others
AGE | others
SYSBP | others
DIABP | others
CURSMOKE | others
BMI | others
DIABETES | others
BPMEDS | others



This assumption appears to be met as there are no obvious non-linear trends.

Independence

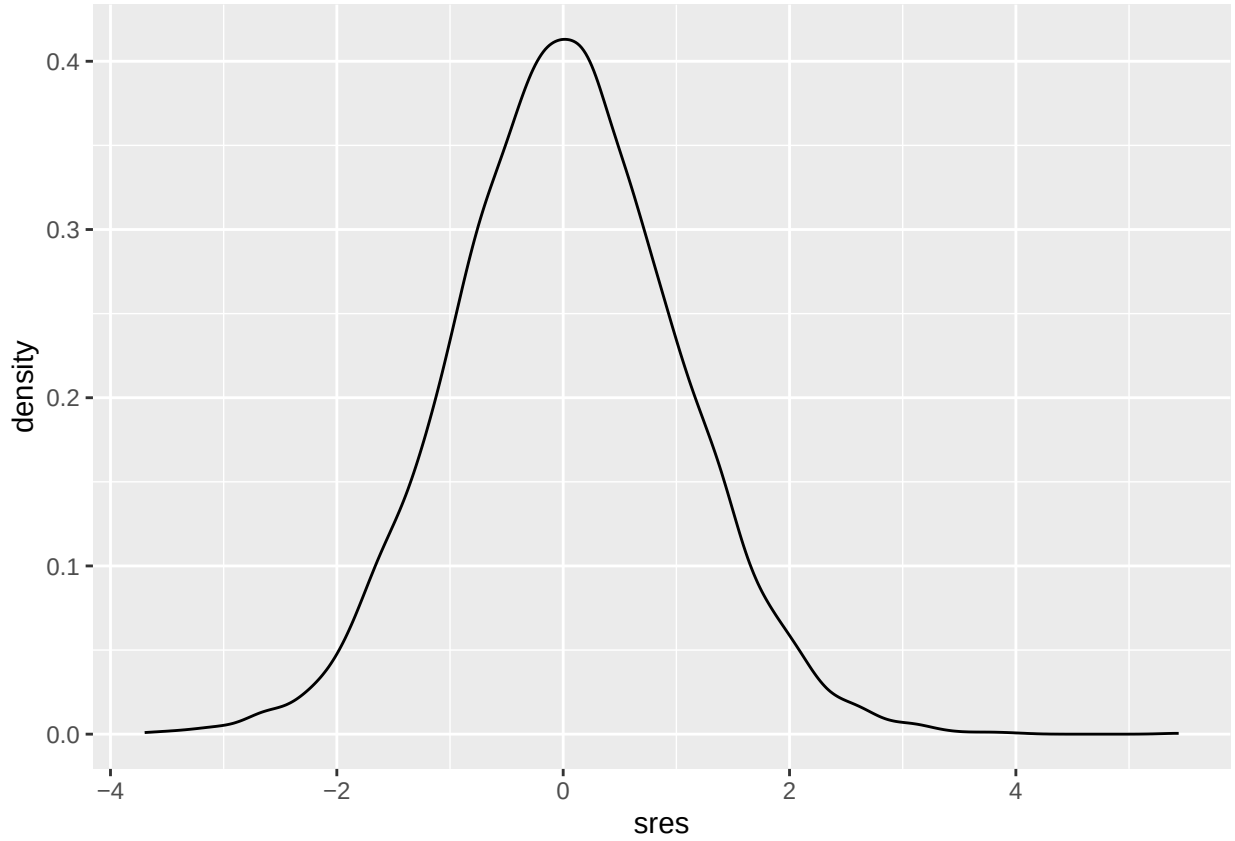
```
sres <- stdres.gls(data_gls)
cor(matrix(data = sres, nrow = 1734, byrow = T))
```

```
##              [,1]         [,2]         [,3]
## [1,] 1.000000000  0.010689325  0.005603369
## [2,] 0.010689325  1.000000000 -0.004182278
## [3,] 0.005603369 -0.004182278  1.000000000
```

This assumption is now met as the correlations between visits is much, much lower.

Normality

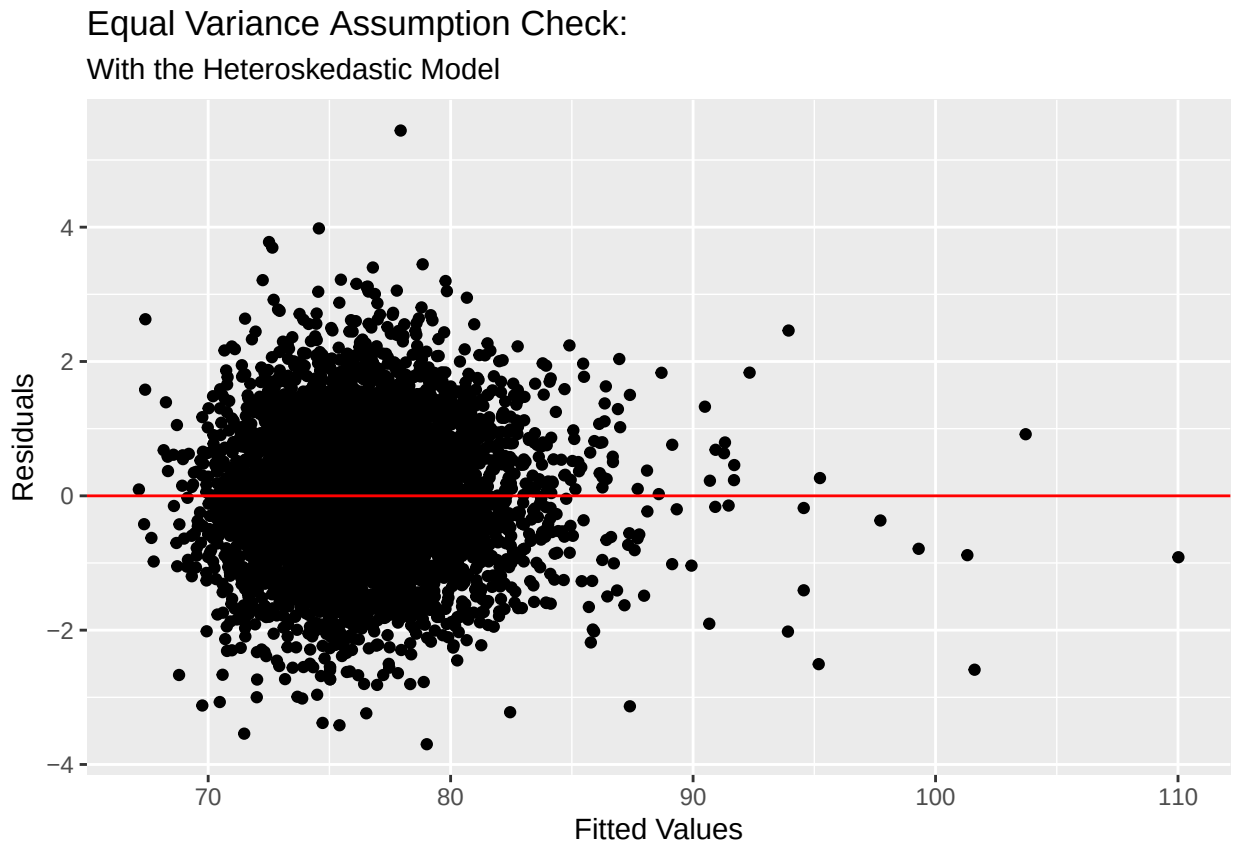
```
ggplot(mapping = aes(x = sres)) +
  geom_density()
```



This density plot demonstrates that the residuals are sufficiently normally distributed.

Equal Variance

```
ggplot(data, mapping = aes(x=data_lm$fitted.values, y=sres)) +
  geom_point() +
  xlab('Fitted Values') +
  ylab('Residuals') +
  ggtitle('Equal Variance Assumption Check:',
          subtitle ='With the Heteroskedastic Model') +
  geom_hline(yintercept = 0, col = "red")
```



For the most part, it appears that this assumption is met. However, for relatively few number of higher fitted values, the variance appears to be smaller.

6. Is DIABETES a risk factor for Tachycardia? Justify your answer and explain any effect of DIABETES on heart rate (include uncertainty in your conclusions).

```
#difference between diabetes and no diabetes for a patient
a.matrix <- c(0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0)
test1 <- glht(data_gls, linfct = t(a.matrix), rhs = 0, alternative = "greater")
summary(test1)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: gls(model = log(HEARTRTE) ~ SEX + TOTCHOL + AGE + SYSBP + DIABP +
##     CURSMOKE + BMI + DIABETES + BPMEDS + GLUCOSE, data = data,
##     correlation = corSymm(form = ~1:3 | RANDID), method = "ML")
##
## Linear Hypotheses:
##        Estimate Std. Error z value Pr(>z)
## 1 <= 0  0.01084    0.01213   0.894  0.186
## (Adjusted p values reported -- single-step method)
```

```
confint(data_gls, level = 0.95)
```

```
##                      2.5 %        97.5 %
## (Intercept)  3.86767623802  3.9851886234
## SEX          0.02657927946  0.0499874797
## TOTCHOL      0.00005633516  0.0002587811
## AGE          0.00039055100  0.0015229511
## SYSBP        0.00017062676  0.0008101834
## DIABP        0.00073954066  0.0018467089
## CURSMOKE     0.01977056388  0.0395766221
## BMI         -0.00135122424  0.0015201189
## DIABETES    -0.01293307603  0.0346174332
## BPMEDS      -0.03575960389 -0.0066310911
## GLUCOSE      0.00056359733  0.0009273436
```

To determine if `DIABETES` is a risk factor for Tachycardia, we test if β_{Diabetes} has an effect on `log(HEARTRTE)` with a t-test. We set the hypotheses for the t-test at $H_0 : \beta_{\text{Diabetes}} = 0$, i.e., diabetes is **not** a risk factor for Tachycardia, and at $H_A : \beta_{\text{Diabetes}} > 0$, i.e. diabetes **is** a risk factor for Tachycardia. To test this effect, we compare the difference of a patient with diabetes to the same patient without diabetes. The result of this difference in a t-test produces a p-value of 0.186. We conclude there is not enough evidence to support considering diabetes a risk factor for Tachycardia, and fail to reject the null hypothesis. We are 95% confident that the true value of β_{Diabetes} on `log(HEARTRTE)` is between -0.0129 and 0.0346.

7. What is the expected difference in heart rate for a female patient with at age 35 who is a smoker vs. an older female of 45 but not a smoker (assume the other characteristics are the same)? What does this say about the effect of smoking?

```
b.matrix <- c(1, 0, 0, 35, 0, 0, 1, 0, 0, 0, 0)
test2 <- glht(data_gls, linfct = t(b.matrix), rhs = 0, alternative = "less")
summary(test2)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: gls(model = log(HEARTRTE) ~ SEX + TOTCHOL + AGE + SYSBP + DIABP +
##     CURSMOKE + BMI + DIABETES + BPMEDS + GLUCOSE, data = data,
##     correlation = corSymm(form = ~1:3 | RANDID), method = "ML")
##
## Linear Hypotheses:
##        Estimate Std. Error z value Pr(<z)
## 1 >= 0  3.98959    0.02663   149.8      1
## (Adjusted p values reported -- single-step method)
```

```
c.matrix <- c(1, 0, 0, 45, 0, 0, 0, 0, 0, 0, 0)
test3 <- glht(data_gls, linfct = t(c.matrix), rhs = 0, alternative = "less")
summary(test3)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
```

```
##
## Fit: gls(model = log(HEARTRTE) ~ SEX + TOTCHOL + AGE + SYSBP + DIABP +
##     CURSMOKE + BMI + DIABETES + BPMEDS + GLUCOSE, data = data,
##     correlation = corSymm(form = ~1:3 | RANDID), method = "ML")
##
## Linear Hypotheses:
##        Estimate Std. Error z value Pr(<z)
## 1 >= 0  3.96949    0.02741   144.8      1
## (Adjusted p values reported -- single-step method)
```

To determine the expected difference in `log(HEARTRTE)` for a female patient with at age 35 who is a smoker vs. an older female of 45 but not a smoker, assuming the other characteristics are the same, we must conduct two t-tests checking if $H_A : \boldsymbol{\alpha}'\boldsymbol{\beta} < 0$ where $\boldsymbol{\alpha}' = (1, 0, 0, 35 \text{ or } 45, 0, 0, 1 \text{ or } 0, 0, 0, 0, 0)$. These t-tests produce p-values close to 1, meaning there was no meaningful expected difference between the two patients, and we fail to reject the null hypothesis. These results suggest that the effect of older age is similar to the effect of smoking on `log(HEARTRTE)`.