STAT 456 Final Project

Justin Papagelis

Table of contents

Project Description **2**
1. Reproducible Data Analysis . . . 2
2. Research Presentation . . . 3
Submission and timeline . . . 3

1 Reproducible Data Analysis **4**
1.1 Context and questions of interest . . . 4
1.2 Data wrangling and exploration . . . 4
1.3 Model building and diagnostics . . . 7
1.4 Model summary and conclusions . . . 11

Project Description

Your second project will follow a similar structure as your first, however *you* will be responsible for finding an appropriate dataset for analysis. To gain experience beyond STAT 230, the primary outcome of the dataset should be either count data or positive, continuous data, unless you decide to try your hand at working with correlated, continuous data. Potential data sources include:

- Data Dryad
- our textbook (**GLMsData** package)
- Alan Agresti's textbook data (see me for the textbooks if needed)
  - (**CatDataAnalysis** package)
  - Introduction to CDA data, also available on Github
- Harvard Dataverse

You are welcome to reproduce or semi-reproduce analyses from published articles, but all work should be your own.

1. Reproducible Data Analysis

The first item will be organized documentation of your data analysis, which should include, at a minimum:

- data exploration, including justification of choices
- notes on model building
  - It is not necessary to include code and output for every single model you considered; instead, it is okay to to only keep the final model and any relevant versions and document the decisions and reasoning you took to get there. Documentation would include the name of the plot, test, or other diagnostic procedure used to make a decision, justification for using the chosen procedure (e.g., why an F test instead of a chi-square?) and the results of that procedure as well as any other reasoning used (e.g., scientific justification).
- full diagnostic analysis of your final model(s), including justification of choices made in the process (e.g., why that particular type of residuals?)
- relevant tables, figures, etc. for summarizing your final model(s) and drawing conclusions

The narration does not need to be in paragraph form, but I want to be able to assess how you are using your theoretical knowledge to make decisions in practice.

2. Research Presentation

The second item will be a presentation in the IMRD format that should be accessible to a broader audience (no more than 15 minutes). You should be prepared to discuss the statistical details after during Q&A.

***I**ntroduction*: Introduce the context of the data (study design), the question(s) of interest, and any corresponding hypotheses.

***M**ethods*: Describe relevant variables of interest (how were they measured? what do we need to know about them?), summarize any relevant variables **except for the outcome**. Then describe your primary statistical methods for the final analysis, any additional information we need to know about how you processed variables for analysis (e.g., did you have to transform any variables?), and which software you used to do it: using R software [version 4.2, @RCoreTeam2022].

***R**esults*: Describe and interpret the results of your statistical analyses, including any relevant graphs and/or tables. The results should not contain any additional commentary or explanations.

***D**iscussion*: Summarize and comment on your main findings in a way that is accessible to a broad audience. How did the findings compare with what you expect or what you know of relevant literature? What statistical challenges did you have?

The presentation can be generated using any method you prefer (Quarto, Powerpoint, Google slides, etc.).

Submission and timeline

You will complete the reproducible data analysis process for your final chosen dataset within this Quarto document. Please rename the Quarto document before you begin to include your name at the end of the file name (you may further customize the file name to your own liking).

- Friday, April 28: Choice of datasets narrowed down
- Wednesday, May 3: Questions of interest specifically defined (share in class)
  - This will require having wrangled and explored the data to verify your question(s) can be answered by the data on hand.
- Monday, May 9: Presentations and discussion! Push the final Quarto document, output, and slides to your GitHub repo by the start of class. If your slides are web-based, please include the URL within this document.

1 Reproducible Data Analysis

1.1 Context and questions of interest

The dataset *mutationfreq* contains the somatic cell mutant frequencies at the hprt locus of the X-chromosome measured with the T-lymphocyte cloning assay in a healthy pediatric population. The study (conducted in 1993), which included 49 subjects (29 females and 20 males) under 18 years of age, aimed to see the effects of variables such as age (`Age`), sex (`Sex`), and unselected cloning efficiency (`Ceff`) on the thioguanine-resistant mutation frequency (`Mfreq`).

The question of interest of this project is to determine if there is evidence of a significant difference between male and female children in the mutation frequency. I will also explore the effects that the other covariates have on the mutation frequency.

1.2 Data wrangling and exploration

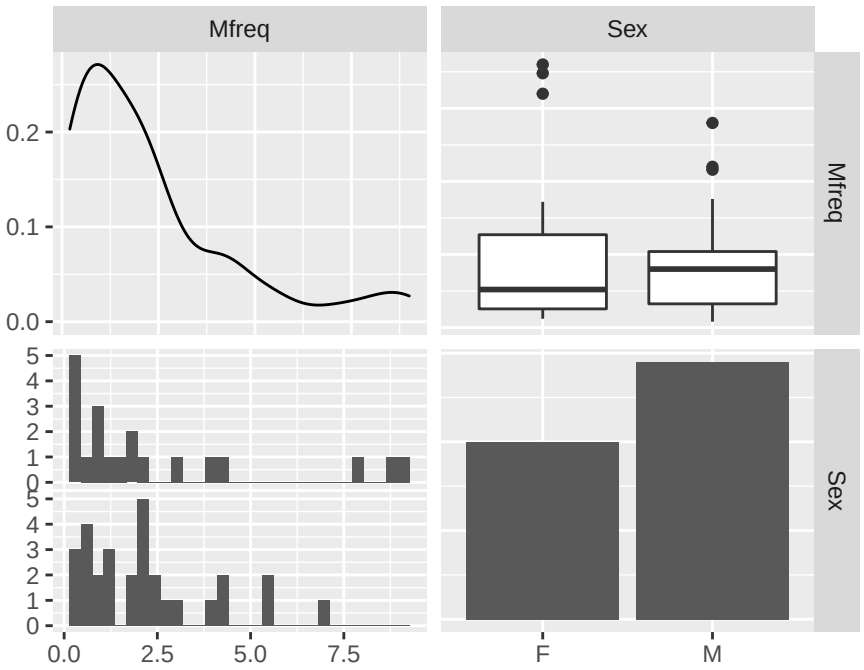
```
data(mutantfreq)
summary(mutantfreq)
```

```
     Donor    Sex         Age              Ceff            Mfreq
 BF1    : 1   F:20   Min.   : 0.08   Min.   :0.0700   Min.   :0.200
 BF10   : 1   M:29   1st Qu.: 1.30   1st Qu.:0.2000   1st Qu.:0.700
 BF12   : 1          Median : 4.70   Median :0.3100   Median :1.800
 BF13   : 1          Mean   : 5.67   Mean   :0.3488   Mean   :2.331
 BF14   : 1          3rd Qu.: 9.00   3rd Qu.:0.4300   3rd Qu.:2.900
 BF15   : 1          Max.   :15.20   Max.   :0.9100   Max.   :9.000
 (Other):43
```

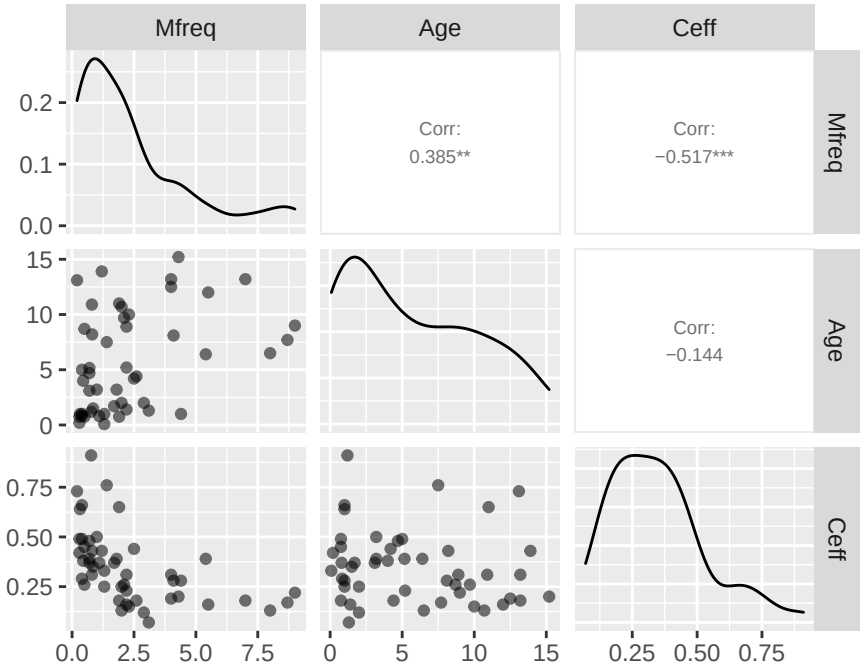
I plotted the relationships that the categorical and the quantitative predictors have with our variable of interest.

```
# Categorical predictors
mutantfreq |>
  dplyr::select(Mfreq, Sex) |>
  GGally::ggpairs(progress = FALSE,
                  lower = list(discrete = "colbar"),
                  upper = list(discrete = "facetbar"))
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

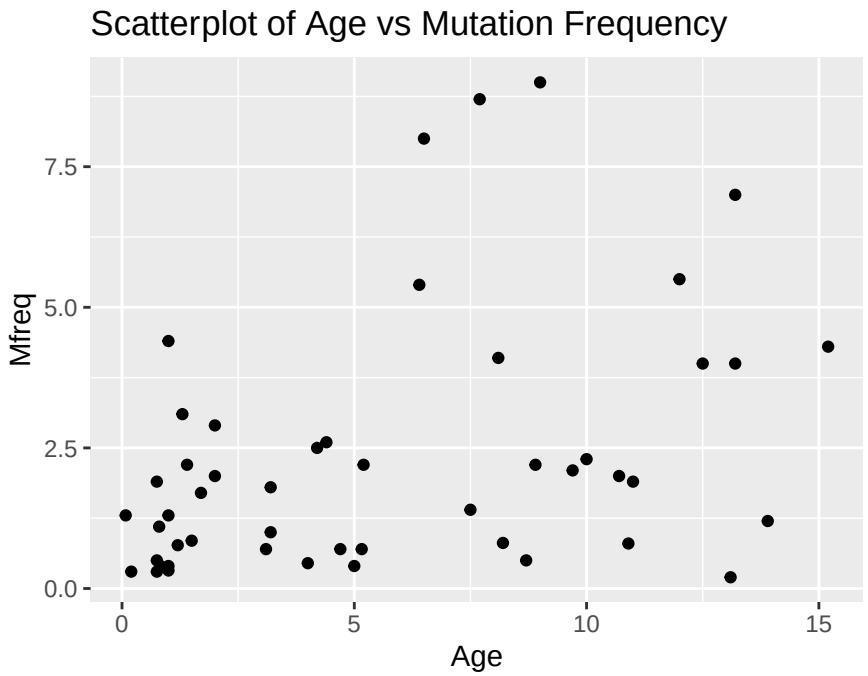



```
# Quantitative predictors
mutantfreq |>
  select(Mfreq, Age, Ceff) |>
  GGally::ggpairs(progress = FALSE,
                  aes(alpha = 0.7),
                  upper = list(continuous = wrap("cor", size = 2.5)),
                  lower = list(combo = "facetdensity"))
```

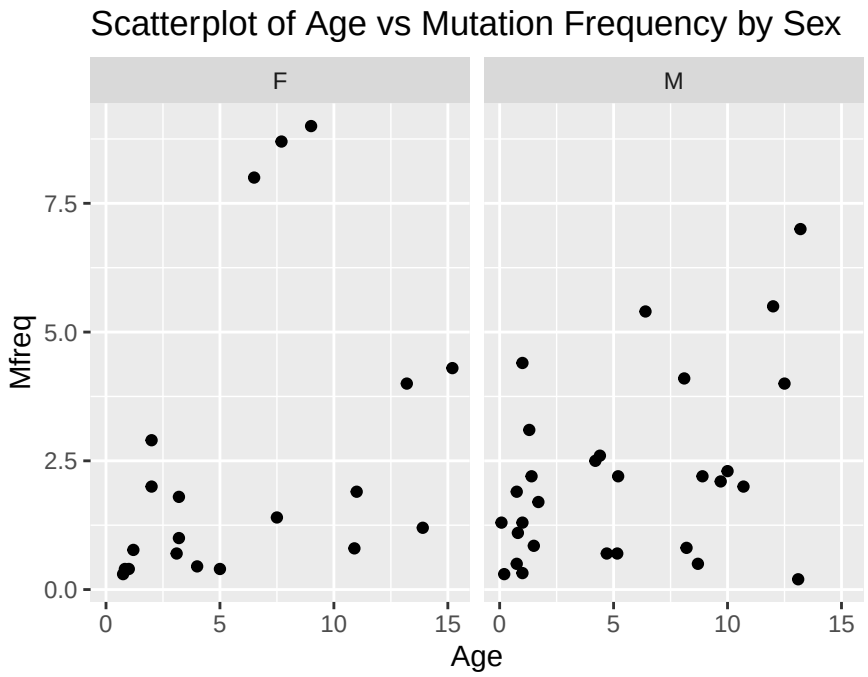


Looking at the boxplots between mutation frequency and sex, there doesn't appear to a significant difference between males and females. From the scatterplots age and mutation frequency have a weak positive correlation while cloning efficiency and mutation frequency have a strong negative correlation. I will keep an eye on using age and mutation frequency for our later models.

```
ggplot(data = mutantfreq, aes(x = Age, y = Mfreq)) +
  geom_point() +
  labs(title = "Scatterplot of Age vs Mutation Frequency")
```




```
ggplot(data = mutantfreq, aes(x = Age, y = Mfreq)) +
  geom_point() +
  facet_wrap(~Sex) +
  labs(title = "Scatterplot of Age vs Mutation Frequency by Sex")
```



It seems that the mutation frequency has an increasing mean-variance relationship with age so a Gamma GLM might be useful.

1.3 Model building and diagnostics

Since our data is positive and continuous, I will start by using a Gamma GLM.

I will start out with a saturated model. Since ϕ is unknown and is estimated using the Pearson estimator (default in R), I will be using F-tests. I will be starting off using the logarithmic link function.

```
glm.mf <- glm(Mfreq ~ Sex*Age*Ceff,family=Gamma(link = "log"), data = mutantfreq)
anova(glm.mf, test="F")
```

```
Analysis of Deviance Table

Model: Gamma, link: log

Response: Mfreq

Terms added sequentially (first to last)


             Df Deviance Resid. Df Resid. Dev       F     Pr(>F)
NULL                           48     43.285
Sex           1   0.2223        47     43.063  0.3679   0.547481
Age           1   6.9307        46     36.132 11.4725   0.001569 **
```

```
Ceff          1  12.2242        45     23.908 20.2349 0.00005538 ***
Sex:Age       1   0.2532        44     23.654  0.4192   0.520949
Sex:Ceff      1   0.6761        43     22.978  1.1192   0.296282
Age:Ceff      1   0.3873        42     22.591  0.6411   0.427932
Sex:Age:Ceff  1   0.0047        41     22.586  0.0078   0.929943
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

None of the interaction terms were significant since they all had p-values of their corresponding F-tests greater than 0.05. Sex also was not significant. We will remove them from our model.

```
glm.mf2 <- glm(Mfreq ~ Age+Ceff,family=Gamma(link = "log"), data = mutantfreq)
anova(glm.mf2,test="F")
```

```
Analysis of Deviance Table

Model: Gamma, link: log

Response: Mfreq

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev      F     Pr(>F)
NULL                    48     43.285
Age   1   7.1512        47     36.134 13.231  0.0006938 ***
Ceff  1  11.7411        46     24.393 21.723 0.00002718 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We will also try doing the same model we had above using an inverse Gaussian distribution in case the mean-variance relationship increases faster than that of a gamma distribution

```
glm.mf.inv <- glm(Mfreq ~ Age + Ceff,family=inverse.gaussian(link = "log"), data = mutantfre
anova(glm.mf.inv,test="F")
```

```
Analysis of Deviance Table

Model: inverse.gaussian, link: log

Response: Mfreq

Terms added sequentially (first to last)
```

```
     Df Deviance Resid. Df Resid. Dev      F    Pr(>F)
NULL                    48     30.718
Age   1   3.6599        47     27.058 11.154 0.0016700 **
Ceff  1   5.7050        46     21.353 17.387 0.0001334 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
msummary(glm.mf.inv)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.85822    0.27143   3.162 0.002774 **
Age          0.08358    0.02701   3.094 0.003352 **
Ceff        -1.99568    0.48867  -4.084 0.000175 ***

(Dispersion parameter for inverse.gaussian family taken to be 0.3281226)

    Null deviance: 30.718  on 48  degrees of freedom
Residual deviance: 21.353  on 46  degrees of freedom
AIC: 165.81

Number of Fisher Scoring iterations: 22
```

AIC

```
c("Gamma:"=AIC(glm.mf2), "Inverse Gaussian:"=AIC(glm.mf.inv) )
```

```
              Gamma: Inverse Gaussian:
           156.5464          165.8087
```

BIC

```
c("Gamma:"=BIC(glm.mf2), "Inverse Gaussian:"=BIC(glm.mf.inv) )
```

```
              Gamma: Inverse Gaussian:
           164.1137          173.3759
```

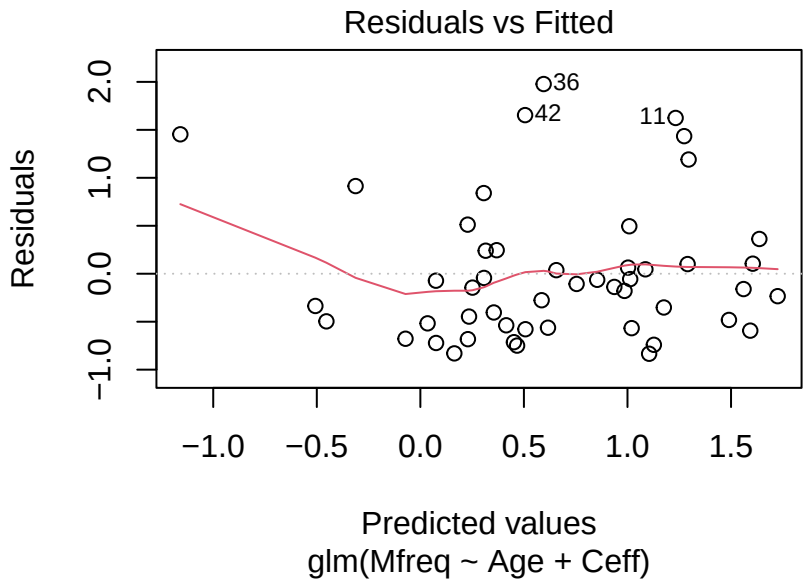
The model using the Gamma distribution has the lowest AIC and BIC so I will be using that model as my final model.

```
final_model <- glm.mf2
```

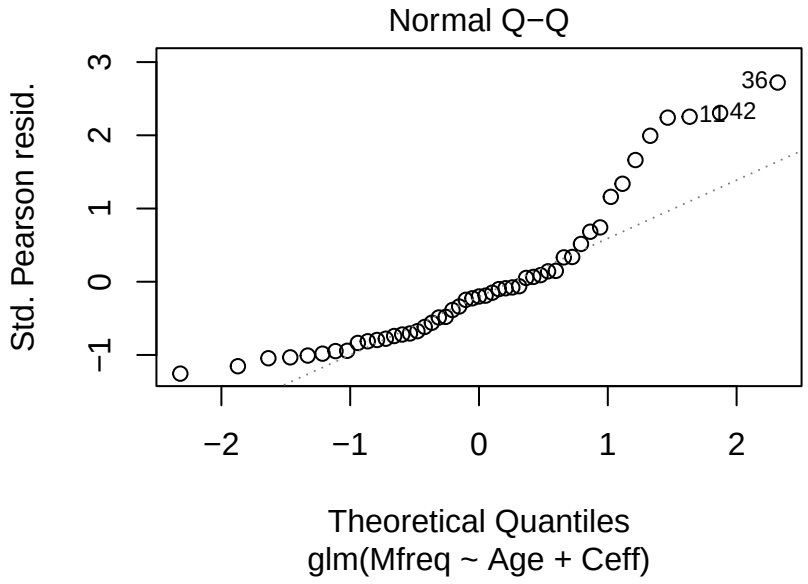
Checking Diagnostics

I use Pearson residuals when checking the residuals vs fitted plot because I am using a response variable that is continuous.

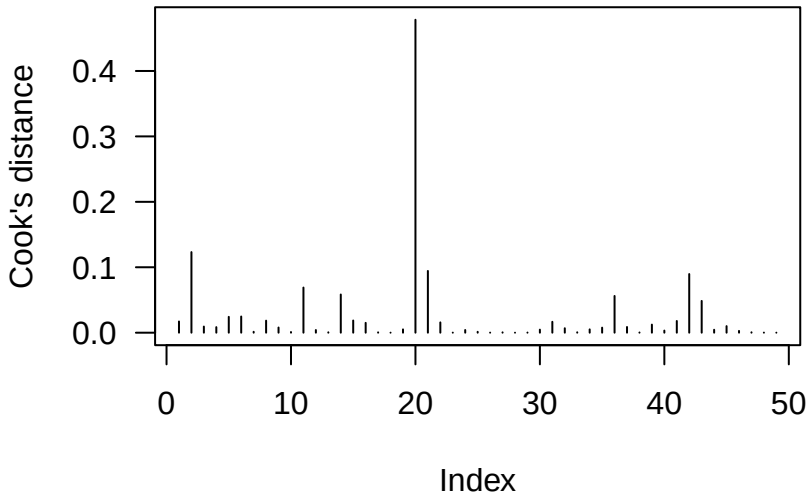
```r
# residuals vs fitted
plot(final_model, which = 1)
```




```r
# QQ plot
plot(final_model, which = 2)
```




```r
# Cook's distance
plot(cooks.distance(final_model), ylab="Cook's distance", las=1, type="h")
```




```
# other outlier measures
im<-influence.measures(final_model)
colSums(im$is.inf)
```

```
dfb.1_  dfb.Age dfb.Ceff    dffit    cov.r   cook.d      hat
     0        0        0        2        1        0        2
```

```
# check for multicollinearity
car::vif(final_model)
```

```
      Age     Ceff
1.021291 1.021291
```

Independence is assumed from child to child based on the specifics of the study. The residuals vs fitted plot doesn't show any trends and the data seems dispersed evenly above and below 0 except for a couple potential outliers. Therefore, the linearity condition is satisfied. The Q-Q plot shows that using a gamma GLM is reasonable and the Normality condition is satisfied because most of the points lie along the diagonal. There is slight concern as some points deviate from the diagonal at higher values, but I'll keep using this model and proceed with caution. There do not seem to be ouliers since none have a Cook's Distance high enough to be influential. Some points are flagged by other measures, but I'm not too concerned because most of the measures did not identify any point we need to be worried about. I also verified that there is no multicollinearity between our variables and since they have VIF values around 1, there is no cause for concern there either. Therefore, I'll proceed with this model.

1.4 Model summary and conclusions

```
msummary(final_model)
```

```
Coefficients:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)  1.18090    0.28005   4.217  0.000115 ***
Age          0.07087    0.02356   3.008  0.004253 **
Ceff        -2.66467    0.58534  -4.552 0.0000388 ***

(Dispersion parameter for Gamma family taken to be 0.5404957)

    Null deviance: 43.285  on 48  degrees of freedom
Residual deviance: 24.393  on 46  degrees of freedom
AIC: 156.55

Number of Fisher Scoring iterations: 9
```

```r
tidy(final_model, exponentiate = TRUE, conf.int = TRUE) |>
  kable(booktabs = TRUE, digits = 3, col.names = c(
    "Term", "Estimate", "Std. Error", "Statistic", "p-value", "Conf. Low", "Conf. High"
  )) %>%
  kable_styling(latex_options = "HOLD_position")
```

Term	Estimate	Std. Error	Statistic	p-value	Conf. Low	Conf. High
(Intercept)	3.257	0.280	4.217	0.000	1.929	5.645
Age	1.073	0.024	3.008	0.004	1.021	1.130
Ceff	0.070	0.585	-4.552	0.000	0.026	0.200

We are 95% confident that an increase in one year of age would cause the mean mutation frequency to increase by 2% to 13%, after controlling for other variables. Additionally, we are 95% confident that an increase in one unit of cloning efficiency would cause the mean mutation frequency to decrease between 80% to 97%, after controlling for other variables in the model.

Overall, age and cloning efficiency had the most significant effect on somatic cell mutant frequencies at the hprt locus of the X-chromosome in a healthy pediatric population. There was no significant evidence of a difference in the mutation frequencies between male and female children.