# ST623 Midterm

Ericka Smith

## Problem 1

### Part A

**i)**

```
## [1] 231.8282
```

**ii)**

```
## [1] 231.8282
```

**iii)**

```
## [1] 137.8554
```

**iv)**

```
## [1] 137.8554
```

### Part B

```
##
## Call:
## lm(formula = response ~ family_size + birth_order + sex_seq,
##     data = birth_long)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1045 -1.6665  0.0798  1.3244  3.8590
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    47.68810    1.22212  39.021  < 2e-16 ***
## family_size    -4.50786    0.25483 -17.690  < 2e-16 ***
## birth_order2-3  6.42768    0.72637   8.849 7.02e-12 ***
## birth_order3-4 10.70452    0.81916  13.068  < 2e-16 ***
## birth_order4-5 14.91679    0.97036  15.372  < 2e-16 ***
## birth_order5-6 19.77071    1.27415  15.517  < 2e-16 ***
## sex_seqFM      -0.02667    0.77852  -0.034    0.973
```

```
## sex_seqMF       0.24000    0.77852   0.308    0.759
## sex_seqMM      -0.06000    0.77852  -0.077    0.939
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.132 on 51 degrees of freedom
## Multiple R-squared:  0.9015, Adjusted R-squared:  0.886
## F-statistic: 58.32 on 8 and 51 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = response ~ family_size + birth_order + sex_seq,
##     data = birth_reversed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1045 -1.6665  0.0798  1.3244  3.8590
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    47.68810    1.22212  39.021  < 2e-16 ***
## family_size    -4.50786    0.25483 -17.690  < 2e-16 ***
## birth_order3-2  6.42768    0.72637   8.849 7.02e-12 ***
## birth_order4-3 10.70452    0.81916  13.068  < 2e-16 ***
## birth_order5-4 14.91679    0.97036  15.372  < 2e-16 ***
## birth_order6-5 19.77071    1.27415  15.517  < 2e-16 ***
## sex_seqFM       0.24000    0.77852   0.308    0.759
## sex_seqMF      -0.02667    0.77852  -0.034    0.973
## sex_seqMM      -0.06000    0.77852  -0.077    0.939
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.132 on 51 degrees of freedom
## Multiple R-squared:  0.9015, Adjusted R-squared:  0.886
## F-statistic: 58.32 on 8 and 51 DF,  p-value: < 2.2e-16
```

These models are the same except for the interpretation of them. The sex sequence is not significant in either model, only the birth order is. Since the birth order has the same levels just with different titles, this makes sense. For the second model you would talk about the distance in time from the second child to the first child. In the second model you'd talk about the distance in teim from the first child to the second child. These are the same values.

## Part C

```
##
## Call:
## lm(formula = response ~ family_size + birth_order + sex_seq,
##     data = birth_long)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1045 -1.6665  0.0798  1.3244  3.8590
##
```

```
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    47.68810    1.22212  39.021  < 2e-16 ***
## family_size    -4.50786    0.25483 -17.690  < 2e-16 ***
## birth_order2-3  6.42768    0.72637   8.849 7.02e-12 ***
## birth_order3-4 10.70452    0.81916  13.068  < 2e-16 ***
## birth_order4-5 14.91679    0.97036  15.372  < 2e-16 ***
## birth_order5-6 19.77071    1.27415  15.517  < 2e-16 ***
## sex_seqFM      -0.02667    0.77852  -0.034    0.973
## sex_seqMF       0.24000    0.77852   0.308    0.759
## sex_seqMM      -0.06000    0.77852  -0.077    0.939
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.132 on 51 degrees of freedom
## Multiple R-squared:  0.9015, Adjusted R-squared:  0.886
## F-statistic: 58.32 on 8 and 51 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = log(response) ~ family_size + birth_order + sex_seq,
##     data = birth_long)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.074799 -0.038529 -0.003535  0.035363  0.095885
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.919386   0.028241 138.783   <2e-16 ***
## family_size    -0.136100   0.005889 -23.112   <2e-16 ***
## birth_order2-3  0.200386   0.016785  11.938   <2e-16 ***
## birth_order3-4  0.329782   0.018929  17.422   <2e-16 ***
## birth_order4-5  0.457131   0.022423  20.386   <2e-16 ***
## birth_order5-6  0.596873   0.029443  20.272   <2e-16 ***
## sex_seqFM      -0.001996   0.017990  -0.111    0.912
## sex_seqMF       0.005719   0.017990   0.318    0.752
## sex_seqMM      -0.002301   0.017990  -0.128    0.899
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04927 on 51 degrees of freedom
## Multiple R-squared:  0.9405, Adjusted R-squared:  0.9312
## F-statistic: 100.8 on 8 and 51 DF,  p-value: < 2.2e-16
```

I don't notice anything particularly unusual. I tried a transformation and found that transforming the response variable results in a higher adjusted R-squared and a higher F-statistic. The untransformed model is still a good fit but the transformed one is better.

## Part D

This data is really interesting for a few reasons. First, that the sex sequence doesn't have any significant effect on the mean interval in months between successive births in the same family. Family size and birth

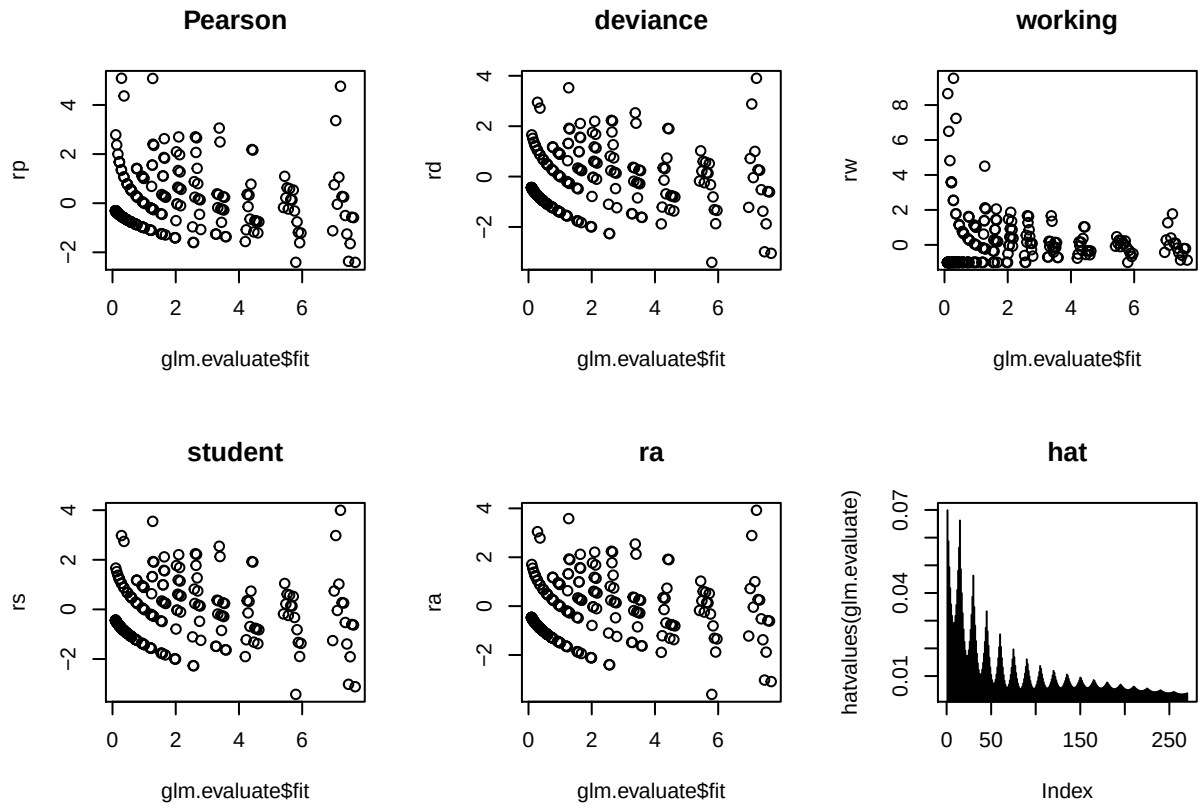
order however are both significant ($\alpha = 0.05$). As family size increases we see a decrease in mean interval between successive births. Conversely, it seems that there is a greater time duration between later birth orders (i.e. 5-6) as compared to earlier birth orders (i.e. 2-3).

# Problem 2

## Part A

**Model 1: linear in $x_1$ and $x_2$**

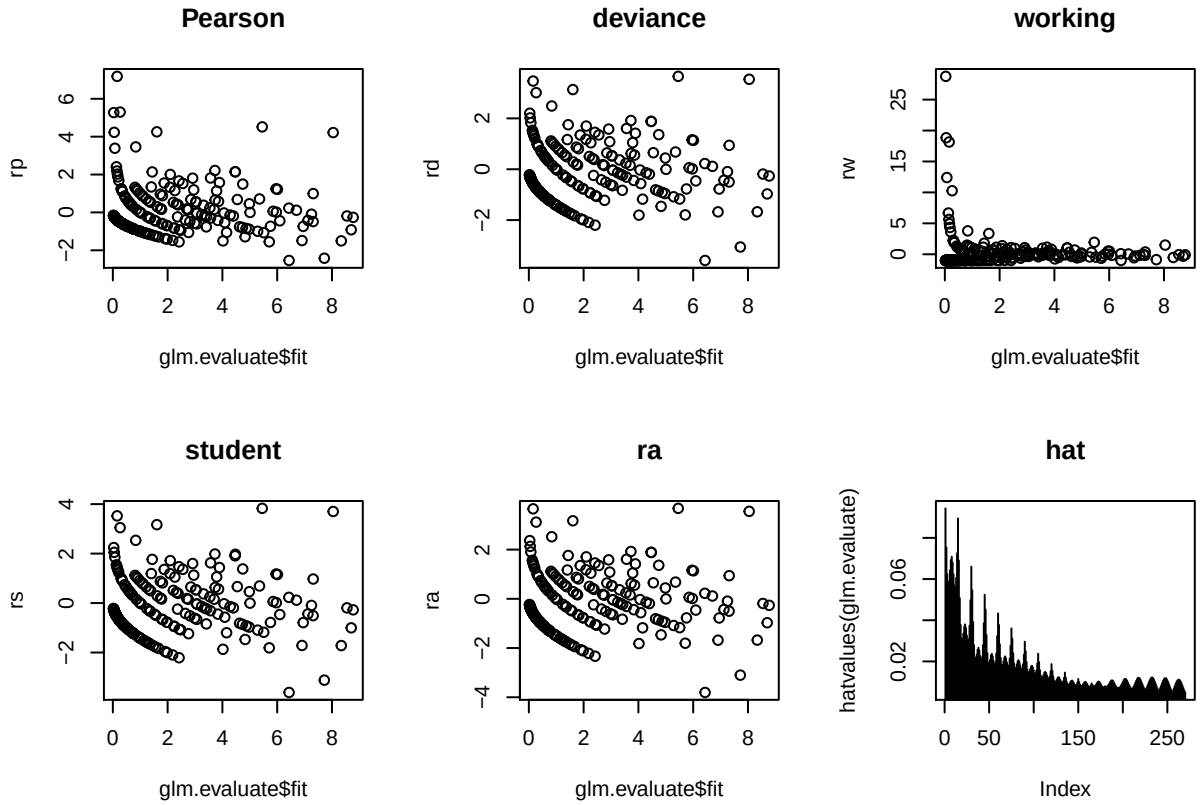
```
##
## Call:
## glm(formula = count ~ redshift + brightness, family = poisson(link = log),
##     data = galaxy_long)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4054  -0.9347  -0.5103   0.3473   3.9018
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.506166   0.214054 -11.708   <2e-16 ***
## redshift    -0.006946   0.010502  -0.661    0.508
## brightness   0.252852   0.013059  19.363   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 906.85  on 269  degrees of freedom
## Residual deviance: 350.52  on 267  degrees of freedom
## AIC: 759.55
##
## Number of Fisher Scoring iterations: 5
```




**Model 2: quadratic in $x_1$ and $x_2$**

```
##
## Call:
## glm(formula = count ~ poly(redshift, 2) + poly(brightness, 2),
##     family = poisson(link = log), data = galaxy_long)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.5872  -0.8601  -0.4175   0.3781   3.6533
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -0.3326     0.1119  -2.973  0.00295 **
## poly(redshift, 2)1     -0.6399     0.8499  -0.753  0.45147
## poly(redshift, 2)2     -4.6912     0.8323  -5.637 1.73e-08 ***
## poly(brightness, 2)1   25.0705     2.0837  12.032  < 2e-16 ***
## poly(brightness, 2)2   -2.7495     1.2523  -2.196  0.02812 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 906.85  on 269  degrees of freedom
## Residual deviance: 310.76  on 265  degrees of freedom
```

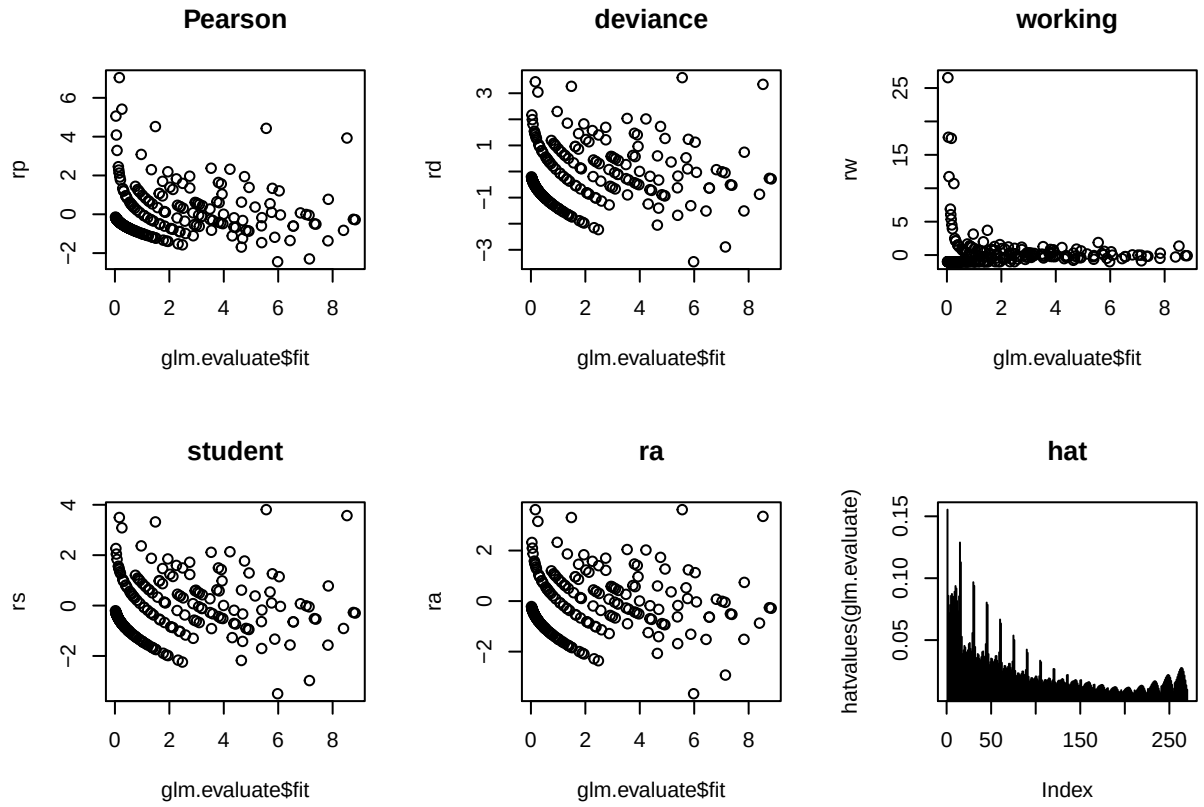
```
## AIC: 723.79
##
## Number of Fisher Scoring iterations: 6
```




**Model 3: cubic in $x_1$ and $x_2$**

```
##
## Call:
## glm(formula = count ~ poly(redshift, 3) + poly(brightness, 3),
##     family = poisson(link = log), data = galaxy_long)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4580  -0.8438  -0.4202   0.3714   3.5977
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -0.33264    0.11944  -2.785  0.00535 **
## poly(redshift, 3)1    -0.95278    0.88104  -1.081  0.27951
## poly(redshift, 3)2    -4.76961    0.83829  -5.690 1.27e-08 ***
## poly(redshift, 3)3    -1.20097    0.82329  -1.459  0.14463
## poly(brightness, 3)1 24.96456    2.45534  10.167  < 2e-16 ***
## poly(brightness, 3)2 -2.61719    2.06872  -1.265  0.20583
## poly(brightness, 3)3 -0.09819    1.22952  -0.080  0.93635
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 906.85  on 269  degrees of freedom
## Residual deviance: 308.62  on 263  degrees of freedom
## AIC: 725.65
##
## Number of Fisher Scoring iterations: 6
```




## Part B

```
## [1] 759.5474
```

```
##  (Intercept)     redshift   brightness
## -2.506166110 -0.006945706  0.252851911
```

```
## [1] 723.7929
```

```
##          (Intercept)   poly(redshift, 2)1   poly(redshift, 2)2
##           -0.3326490           -0.6399081           -4.6912272
## poly(brightness, 2)1 poly(brightness, 2)2
##           25.0705480           -2.7494577
```

```
## [1] 725.6511
```

```
##          (Intercept)   poly(redshift, 3)1   poly(redshift, 3)2
##          -0.33263718          -0.95277613          -4.76960789
##   poly(redshift, 3)3 poly(brightness, 3)1 poly(brightness, 3)2
##          -1.20097114           24.96455848          -2.61718556
## poly(brightness, 3)3
##          -0.09819248
```

Based on the plots, the estimated regression coefficients and AIC values, the quadratic and cubic models appear most appropriate. The quadratic model just barely edges the cubic model out though, in part because the coefficients make a little more intuitive sense but also because the AIC is slightly smaller. Here is a summary of the fit of the model:

```
##
## Call:
## glm(formula = count ~ redshift + brightness, family = poisson(link = log),
##     data = galaxy_long)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4054  -0.9347  -0.5103   0.3473   3.9018
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.506166   0.214054 -11.708   <2e-16 ***
## redshift    -0.006946   0.010502  -0.661    0.508
## brightness   0.252852   0.013059  19.363   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 906.85  on 269  degrees of freedom
## Residual deviance: 350.52  on 267  degrees of freedom
## AIC: 759.55
##
## Number of Fisher Scoring iterations: 5
```

# Problem 3

## Part A

```
##
## Call:
## glm(formula = byss_prev ~ ., family = binomial, data = cotton)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7584  -0.2312  -0.1500  -0.1168   3.3936
##
## Coefficients:
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -7.8987     0.4617 -17.109  < 2e-16 ***
## racew        -0.2463     0.2061  -1.195  0.23203
## sexm          0.2590     0.2116   1.224  0.22095
## smoking       0.6292     0.1931   3.259  0.00112 **
## employment    0.3856     0.1069   3.607  0.00031 ***
## dust          1.3751     0.1155  11.901  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1477.2  on 5418  degrees of freedom
## Residual deviance: 1224.2  on 5413  degrees of freedom
## AIC: 1236.2
##
## Number of Fisher Scoring iterations: 7
```

Based on these results smoking, employment, and dust are significant predictors for being affected by bysinnosis ($\alpha = 0.05$, $p_{smoking} = 0.00112$, $p_{employment} = 0.00031$, $p_{dust} < 2 * 10^{-16}$)

The residual degrees of freedom is calculated by taking $df_{saturated} - df_{proposed}$ where $df_{saturated}$ is for the saturated model, which assumes $n = 5419$ parameters, and $df_{proposed}$ is for the proposed model and assumes $p + 1 = 5 + 1 = 6$ parameters. So we get:

$$df_{saturated} - df_{proposed} = 5419 - 6 = 5413$$

## Part B

The regression coefficient of sex is 0.25290. This can be interpreted as the expected change in log odds (log odds of being affected as compared to not being affected) for a a male as opposed to a female. The odds ratio can the be calculated by exponentiating this value to get 1.287754, which means that we expect to see about **28.8% increase in the odds of being affected by bysinnosis for males over females.**

A 90% confidence interval for the odds ratio (males vs. females) of contracting byssinosis is as follows:

$$0.2590 \pm 1.65 \frac{0.2116}{\sqrt{(5419)}}$$

$$0.2590 \pm 0.004743$$

$$(0.254257, 0.263743)$$

$$(e^{0.254257}, e^{0.263743})$$

$$(1.289503, 1.301794)$$

## Part C

The regression coefficient of dust is 1.3751. This can be interpreted as the expected change in log odds (log odds of being affected as compared to not being affected) for a one step increase in dust level. The odds ratio can the be calculated by exponentiating this value to get 3.955472, which means that **we expect each one step increase in dustiness of work environment level to have an almost 4 times excess risk of being affected by bysinnosis.**

A 90% confidence interval for the excess risk of contracting byssinosis due to increased dustiness is as follows:

$$1.3751 \pm 1.65 \frac{0.1155}{\sqrt{(5419)}}$$

$$1.3751 \pm 0.00259$$

$$(1.37251, 1.37769)$$

$$(e^{1.37251}, e^{1.37769})$$

$$(3.945241, 3.96573)$$

## Part D

Based on this analysis I can conclude that increased smoking habits and increased length of employment are risk factors for bysinnosis. By far the predictor with the greatest effect though is the dustiness of the work environment.

## Problem 4

$X \sim \text{Beta}(rv, r(1-v))$, $r$ known

$$f(x) = \frac{\Gamma(r)}{\Gamma(rv)\Gamma(r(1-v))} x^{rv-1}(1-x)^{r(1-v)-1} \qquad 0<x<1$$

$$E[X] = \frac{rv}{rv+r-rv} = \frac{rv}{r} = v$$

$$\text{Var}[X] = \frac{rv(r)(1-v)}{r^2(r+1)} = \frac{v(1-v)}{r+1}$$

$$f(x) = \exp\left\{\log(\Gamma(r)) - \log(\Gamma(rv)) - \log(\Gamma(r-rv)) + (rv-1)\log x + (r-rv-1)\log(1-x)\right\}$$

$$f(x) = \exp\left\{\frac{\log(\Gamma(r)) + rv\log x + r\log(1-x)}{\log(\Gamma(rv)\Gamma(r-rv)) + \log x + rv\log(1-x) + \log(1-x)}\right\} = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right)$$

(1) $y = x$

(2) $\theta = \log\left(\frac{u}{1-u}\right)$

(3) $\phi = \frac{1}{r}$

(4) $a(\phi) =$

$b(\theta) =$

$c(y,\phi) =$

(5) $u = v$

(6) $g(u) = \log\left(\frac{u}{1-u}\right)$

# Appendix - Code

### Problem 1

```
#Load and manipulate data
birth <- read.table("C:/Users/erick/Downloads/birth.txt", quote="\"", stringsAsFactors=TRUE)
# fix column names
names(birth) <- c("family_size", "birth_order", "MM", "MF", "FM", "FF")
# make sex sequence a variable
birth_long <- birth %>% pivot_longer(c("MM", "MF", "FM", "FF"), names_to = "sex_seq", values_to = "resp
```

### Part A

i)

```
modi <- lm(response~family_size + birth_order + sex_seq, data=birth_long)
deviance(modi)
```

ii)

```
#First create a dataset with birth order reversed.
birth_reversed <- birth_long %>%
  mutate(birth_order = recode(birth_order,
                              "1-2" = "2-1",
                              "2-3" = "3-2",
                              "3-4" = "4-3",
                              "4-5" = "5-4",
                              "5-6" = "6-5"),
         sex_seq = recode(sex_seq,
                          "MF" = "FM",
                          "FM" = "MF"))
```

```
#Now find the residual sum of squares.
modii <- lm(response~family_size + birth_order + sex_seq, data=birth_reversed)
deviance(modii)
```

iii)

```
modiii <- lm(response~family_size*birth_order + sex_seq, data=birth_long)
deviance(modiii)
```

iv)

```
modiv <- lm(response~family_size*birth_order + sex_seq, data=birth_reversed)
deviance(modiv)
```

```
summary(modi)
summary(modii)
```

**Part B**

```
summary(modi)
logmodi <- lm(log(response)~family_size + birth_order + sex_seq, data=birth_long)
summary(logmodi)
# summary(modii)
# logmodiii <- lm(log(response)~family_size*birth_order + sex_seq, data=birth_reversed)
# summary(logmodiii)
```

**Part C**

**Problem 2**

```
#Load Data:
galaxy <- read.table("~/generalized_regression_models/midterm/galaxy.txt", quote="\"", comment.char="")
# change row and column names to fit given data
row.names(galaxy) <- seq(18, 1, by=-1)
names(galaxy) <- seq(1, 15, by=1)
# format for glm()
galaxy_long <- galaxy %>% pivot_longer(c("1", "2", "3", "4",
                                         "5", "6", "7", "8",
                                         "9", "10", "11", "12",
                                         "13", "14", "15"),
                                       names_to = "redshift",
                                       values_to = "count") %>%
  mutate(brightness = rep(18:1, each=15))
galaxy_long$redshift <- as.integer(galaxy_long$redshift)
```

**Part A** Model 1: linear in $x_1$ and $x_2$

```
# Create Model
linear <- glm(count ~
                redshift +
                brightness,
              family=poisson(link=log),
              data=galaxy_long)
summary(linear)
```

```
# Residual Analysis
glm.evaluate=linear

rp=resid(glm.evaluate, "pearson")
```

```
rd=resid(glm.evaluate, "deviance")
rw=resid(glm.evaluate, "working")
rs=rstudent(glm.evaluate)
ra=3*(glm.evaluate$y^{2/3}-glm.evaluate$fit^{2/3})/glm.evaluate$fit^{1/6}/2

par(mfrow=c(2,3))
plot(glm.evaluate$fit,rp, main="Pearson")
plot(glm.evaluate$fit,rd, main = "deviance")
plot(glm.evaluate$fit,rw, main="working")
plot(glm.evaluate$fit,rs, main="student")
plot(glm.evaluate$fit,ra, main="ra")
plot(hatvalues(glm.evaluate), type="h", main="hat")
```

Model 2: quadratic in $x_1$ and $x_2$

```
#Create Model:
quadratic <- glm(count ~
                   poly(redshift, 2) +
                   poly(brightness, 2),
                 family=poisson(link=log),
                 data=galaxy_long)
summary(quadratic)
```

```
#Residual Analysis:
glm.evaluate=quadratic

rp=resid(glm.evaluate, "pearson")
rd=resid(glm.evaluate, "deviance")
rw=resid(glm.evaluate, "working")
rs=rstudent(glm.evaluate)
ra=3*(glm.evaluate$y^{2/3}-glm.evaluate$fit^{2/3})/glm.evaluate$fit^{1/6}/2

par(mfrow=c(2,3))
plot(glm.evaluate$fit,rp, main="Pearson")
plot(glm.evaluate$fit,rd, main = "deviance")
plot(glm.evaluate$fit,rw, main="working")
plot(glm.evaluate$fit,rs, main="student")
plot(glm.evaluate$fit,ra, main="ra")
plot(hatvalues(glm.evaluate), type="h", main="hat")
```

Model 3: cubic in $x_1$ and $x_2$

```
# Create Model:
cubic <- glm(count ~
               poly(redshift, 3) +
               poly(brightness, 3),
             family=poisson(link=log),
             data=galaxy_long)
summary(cubic)
```

```
#Residual Analysis:
glm.evaluate=cubic
```

```
rp=resid(glm.evaluate, "pearson")
rd=resid(glm.evaluate, "deviance")
rw=resid(glm.evaluate, "working")
rs=rstudent(glm.evaluate)
ra=3*(glm.evaluate$y^{2/3}-glm.evaluate$fit^{2/3})/glm.evaluate$fit^{1/6}/2

par(mfrow=c(2,3))
plot(glm.evaluate$fit,rp, main="Pearson")
plot(glm.evaluate$fit,rd, main = "deviance")
plot(glm.evaluate$fit,rw, main="working")
plot(glm.evaluate$fit,rs, main="student")
plot(glm.evaluate$fit,ra, main="ra")
plot(hatvalues(glm.evaluate), type="h", main="hat")
```

```
linear$aic
coef(linear)


quadratic$aic
coef(quadratic)

cubic$aic
coef(cubic)
```

```
summary(linear)
```

**Part B**

**Problem 3**

```
#Load and manipulate data:
cotton <- read.table("~/generalized_regression_models/midterm/cotton.txt", quote="\"")
names(cotton) <- c("affected", "not_affected", "race", "sex", "smoking", "employment", "dust")
cotton <- cotton %>%
  mutate(race = recode(race,
                       "1" = "w",
                       "2" = "nw"),
         sex = recode(sex,
                      "1" = "m",
                      "2" = "f")) %>%
  pivot_longer(c("affected", "not_affected"), names_to = "byss_prev", values_to = "count") %>%
  mutate(byss_prev = recode(byss_prev,
                            "affected" = "1",
                            "not_affected" = "0")) %>%
  uncount(weights = count)
cotton$byss_prev <- as.integer(cotton$byss_prev)
```

```
fit <- glm(byss_prev~., family = binomial, data=cotton)
summary(fit)
```

**Part A**