Poisson Regression

MSBA7002: Business Statistics

Contents

Contents

1. Bikeshare Data

1.1 Fit OLS regression

Let us first fit a naive OLS regression as a benchmark for comparison.

```
attach(Bikeshare)
dim(Bikeshare)
```

```
## [1] 8645   15
```

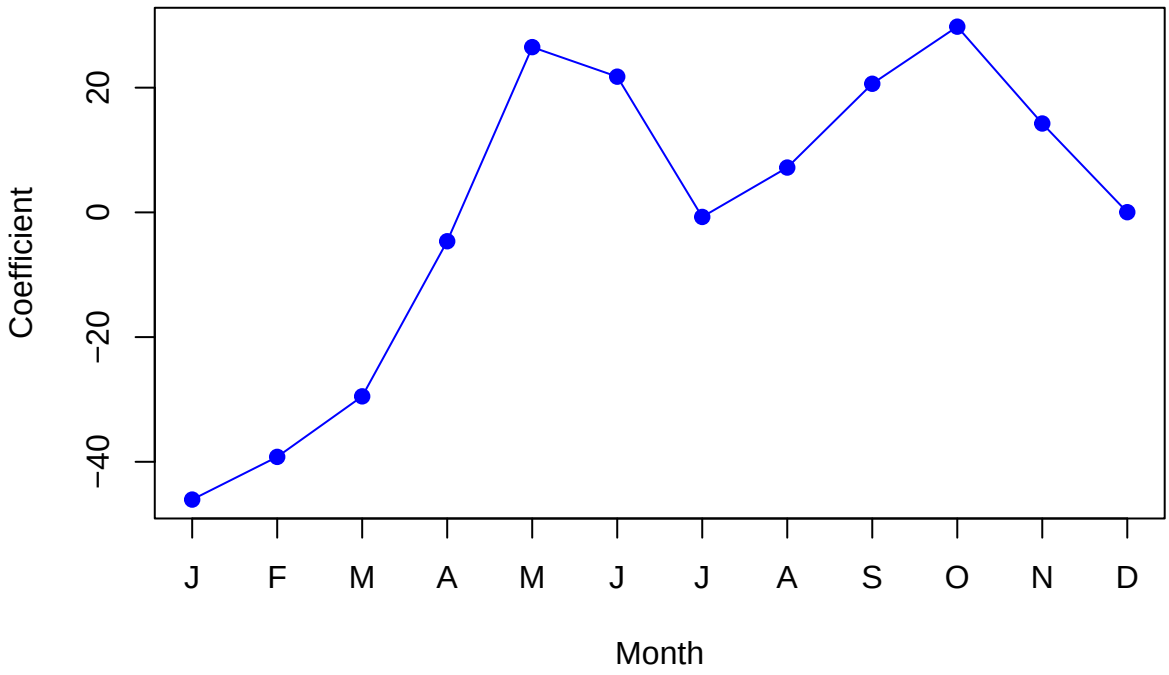
```
names(Bikeshare)
```

```
##  [1] "season"     "mnth"       "day"        "hr"         "holiday"
##  [6] "weekday"    "workingday" "weathersit" "temp"       "atemp"
## [11] "hum"        "windspeed"  "casual"     "registered" "bikers"
```

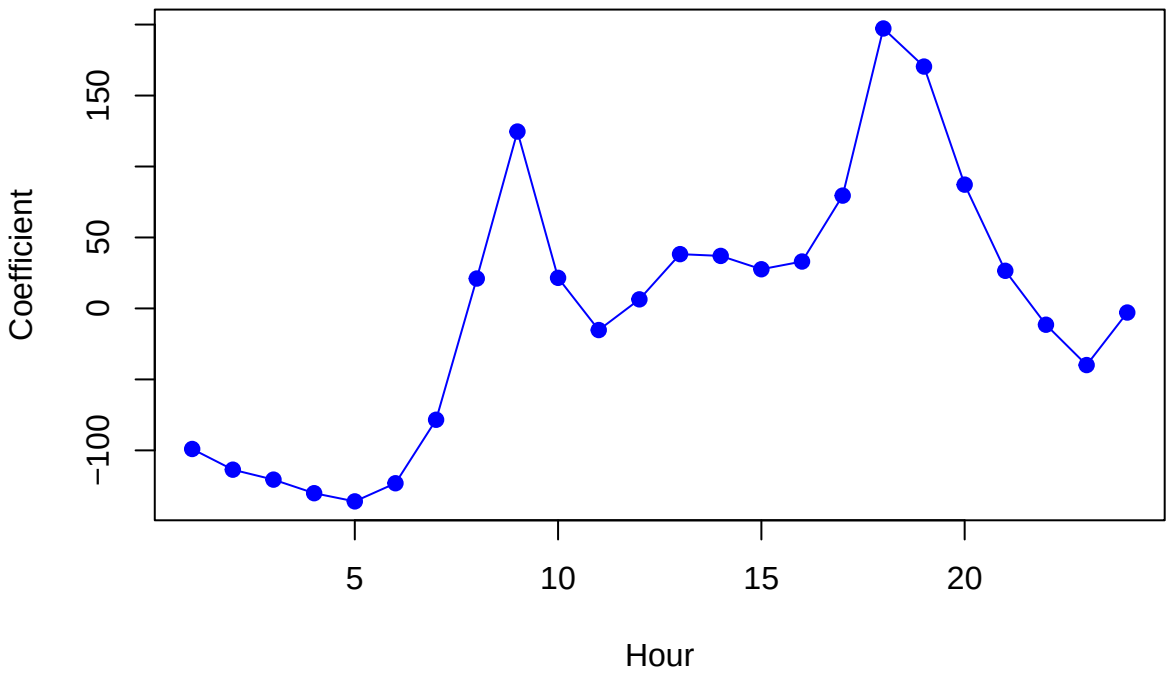
```
contrasts(Bikeshare$hr) = contr.sum(24)
contrasts(Bikeshare$mnth) = contr.sum(12)

mod.lm <- lm(bikers ~ mnth + hr + workingday + temp + weathersit,
             data = Bikeshare)
#summary(mod.lm)
```

```
coef.months <- c(coef(mod.lm)[2:12], 0) - sum(coef(mod.lm)[2:12])/12
# demean to make the total coef of 12 months to be zero
plot(coef.months, xlab = "Month", ylab = "Coefficient", xaxt = "n",
     col = "blue", pch = 19, type = "o")
axis(side = 1, at = 1:12, labels = c("J", "F", "M", "A", "M", "J",
                                     "J", "A", "S", "O", "N", "D"))
```




```
coef.hours <- c(coef(mod.lm)[13:35], 0) - sum(coef(mod.lm)[13:35])/24
# demean to make the total coef of 24 hours to be zero
plot(coef.hours, xlab = "Hour", ylab = "Coefficient",
     col = "blue", pch = 19, type = "o")
```



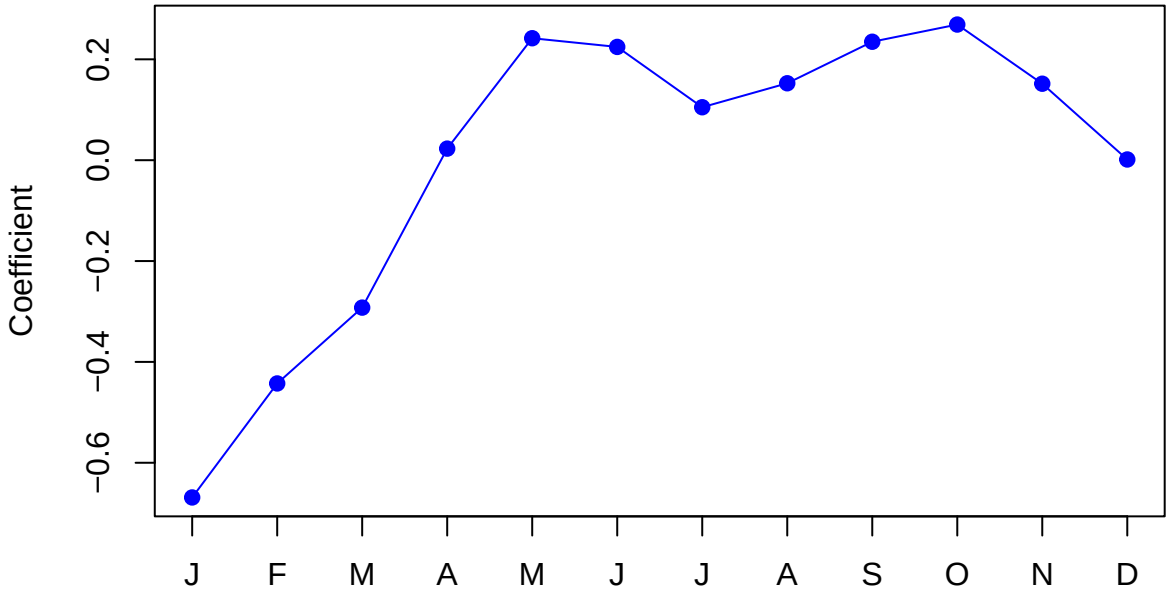
1.2 Fit Poisson regression

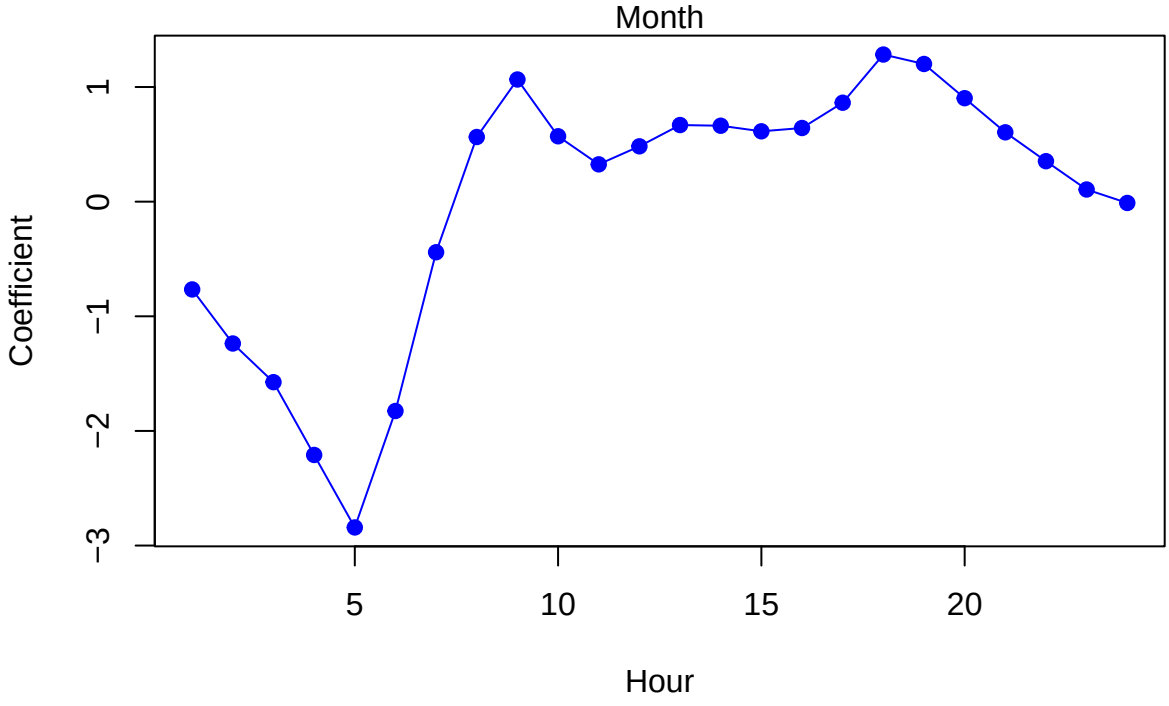
Now, we consider instead fitting a Poisson regression model to the Bikeshare data. Very little changes, except that we now use the function glm() with the argument family = poisson to specify that we wish to fit a Poisson regression model.

```r
mod.pois <- glm(bikers ~ mnth + hr + workingday + temp + weathersit,
                data = Bikeshare, family = poisson)
summary(mod.pois)
```

```
##
## Call:
## glm(formula = bikers ~ mnth + hr + workingday + temp + weathersit,
##     family = poisson, data = Bikeshare)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -20.7574   -3.3441   -0.6549    2.6999   21.9628
##
## Coefficients:
##                              Estimate Std. Error  z value Pr(>|z|)
## (Intercept)                  4.118245   0.006021  683.964  < 2e-16 ***
## mnth1                       -0.670170   0.005907 -113.445  < 2e-16 ***
## mnth2                       -0.444124   0.004860  -91.379  < 2e-16 ***
## mnth3                       -0.293733   0.004144  -70.886  < 2e-16 ***
## mnth4                        0.021523   0.003125    6.888 5.66e-12 ***
## mnth5                        0.240471   0.002916   82.462  < 2e-16 ***
## mnth6                        0.223235   0.003554   62.818  < 2e-16 ***
## mnth7                        0.103617   0.004125   25.121  < 2e-16 ***
## mnth8                        0.151171   0.003662   41.281  < 2e-16 ***
## mnth9                        0.233493   0.003102   75.281  < 2e-16 ***
## mnth10                       0.267573   0.002785   96.091  < 2e-16 ***
## mnth11                       0.150264   0.003180   47.248  < 2e-16 ***
## hr1                         -0.754386   0.007879  -95.744  < 2e-16 ***
## hr2                         -1.225979   0.009953 -123.173  < 2e-16 ***
## hr3                         -1.563147   0.011869 -131.702  < 2e-16 ***
## hr4                         -2.198304   0.016424 -133.846  < 2e-16 ***
## hr5                         -2.830484   0.022538 -125.586  < 2e-16 ***
## hr6                         -1.814657   0.013464 -134.775  < 2e-16 ***
## hr7                         -0.429888   0.006896  -62.341  < 2e-16 ***
## hr8                          0.575181   0.004406  130.544  < 2e-16 ***
## hr9                          1.076927   0.003563  302.220  < 2e-16 ***
## hr10                         0.581769   0.004286  135.727  < 2e-16 ***
## hr11                         0.336852   0.004720   71.372  < 2e-16 ***
## hr12                         0.494121   0.004392  112.494  < 2e-16 ***
## hr13                         0.679642   0.004069  167.040  < 2e-16 ***
## hr14                         0.673565   0.004089  164.722  < 2e-16 ***
## hr15                         0.624910   0.004178  149.570  < 2e-16 ***
## hr16                         0.653763   0.004132  158.205  < 2e-16 ***
## hr17                         0.874301   0.003784  231.040  < 2e-16 ***
## hr18                         1.294635   0.003254  397.848  < 2e-16 ***
## hr19                         1.212281   0.003321  365.084  < 2e-16 ***
## hr20                         0.914022   0.003700  247.065  < 2e-16 ***
## hr21                         0.616201   0.004191  147.045  < 2e-16 ***
## hr22                         0.364181   0.004659   78.173  < 2e-16 ***
## hr23                         0.117493   0.005225   22.488  < 2e-16 ***
## workingday                   0.014665   0.001955    7.502 6.27e-14 ***
## temp                         0.785292   0.011475   68.434  < 2e-16 ***
## weathersitcloudy/misty      -0.075231   0.002179  -34.528  < 2e-16 ***
## weathersitlight rain/snow   -0.575800   0.004058 -141.905  < 2e-16 ***
```

```
## weathersitheavy rain/snow -0.926287   0.166782   -5.554 2.79e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1052921  on 8644  degrees of freedom
## Residual deviance:  228041  on 8605  degrees of freedom
## AIC: 281159
##
## Number of Fisher Scoring iterations: 5
```

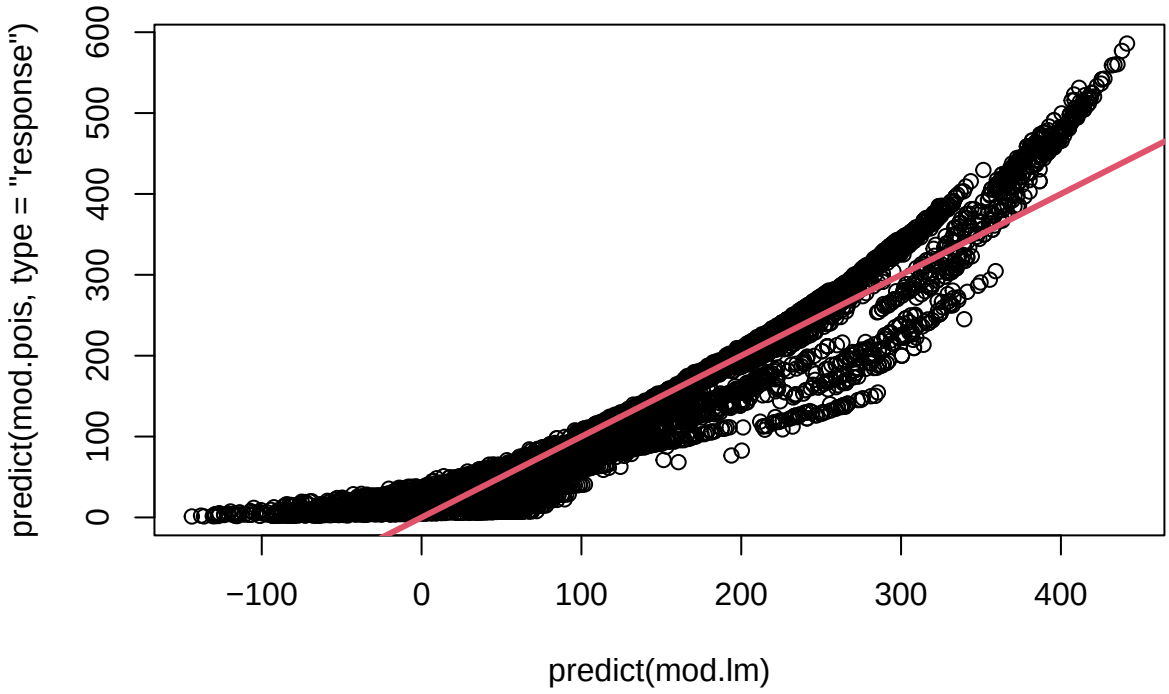




1.3 Compare predictions

We can once again use the predict() function to obtain the fitted values (predictions) from this Poisson regression model. However, we must use the argument type = "response" to specify that we want R to output $exp(\hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p)$ rather than $\hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p$, which it will output by default.

```
plot(predict(mod.lm), predict(mod.pois, type = "response"))
abline(0, 1, col = 2, lwd = 3)
```



The predictions from the Poisson regression model are correlated with those from the linear model; however, the former are non-negative. As a result the Poisson regression predictions tend to be larger than those from the linear model for either very low or very high levels of ridership.

```
cor(predict(mod.lm), Bikeshare$bikers)
```

```
## [1] 0.8212995
```

```
cor(predict(mod.pois, type = "response"), Bikeshare$bikers)
```

```
## [1] 0.8566638
```

The correlation of predicitons from Poisson regression with the truth is indeed higher.

2. The Children Ever Born Data

2.1 EDA

```
ceb <- read.table('ceb.txt')
# names(ceb)

ceb$dur <- factor(ceb$dur, levels = c("0-4", "5-9", "10-14", "15-19", "20-24", "25-29"))
ceb$res <- factor(ceb$res)
ceb$educ <- factor(ceb$educ, levels = c("none", "lower", "upper", "sec+"))
ceb$y <- round(ceb$y)

contrasts <- list(dur = contr.treatment(levels(ceb$dur), base=1),
                  res = contr.treatment(levels(ceb$res), base=2),
                  educ = contr.treatment(levels(ceb$educ), base=1))
summary(ceb)
```

```
##     dur        res        educ         mean            var
##  0-4  :12   rural:23   none :18   Min.   :0.500   Min.   : 0.0000
##  5-9  :12   Suva :24   lower:18   1st Qu.:2.252   1st Qu.: 0.9325
##  10-14:12   urban:23   upper:18   Median :3.645   Median : 1.9450
##  15-19:12              sec+ :16   Mean   :3.696   Mean   : 3.2960
##  20-24:12                         3rd Qu.:5.263   3rd Qu.: 4.2925
##  25-29:10                         Max.   :7.810   Max.   :12.6000
##        n                y
##  Min.   :  1.00   Min.   :   2.0
##  1st Qu.: 13.00   1st Qu.:  38.0
##  Median : 24.50   Median :  77.5
##  Mean   : 38.24   Mean   : 151.5
##  3rd Qu.: 47.00   3rd Qu.: 151.5
##  Max.   :195.00   Max.   :1459.0
```

These are the data from Fiji on children ever born[1]. The dataset has 70 rows representing grouped individual data. Each row has entries for:

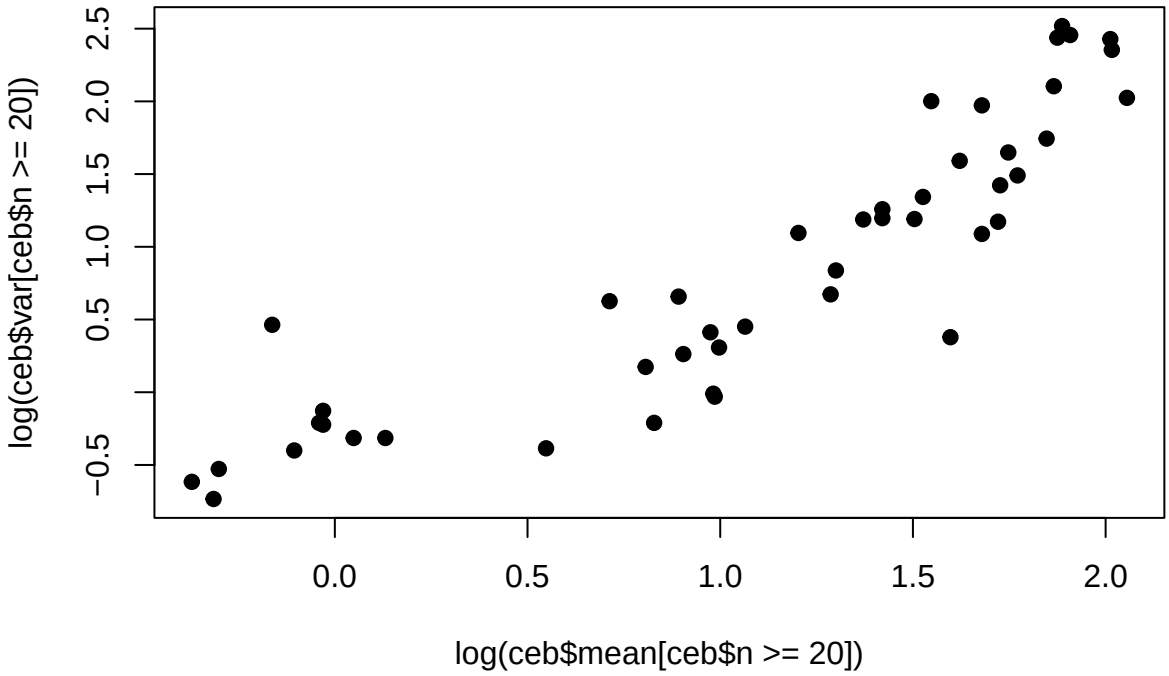
- The cell number (1 to 71, cell 68 has no observations),
- marriage duration (1=0-4, 2=5-9, 3=10-14, 4=15-19, 5=20-24, 6=25-29),
- residence[2] (1=Suva, 2=Urban, 3=Rural),
- education (1=none, 2=lower primary, 3=upper primary, 4=secondary+),
- mean number of children ever born (e.g. 0.50),
- variance of children ever born (e.g. 1.14), and
- number of women in the cell (e.g. 8).

[1]The data was orignially downloaded from the website https://data.princeton.edu/wws509/datasets/#ceb with notes https://data.princeton.edu/wws509/notes/c4.pdf

[2]Suva is the capital city of Fiji, Urban means other urban areas except Suva.

Let us check the mean-variance relationship first to make sure Poisson could be a reasonable model to use.

```
plot(log(ceb$mean[ceb$n >= 20]), log(ceb$var[ceb$n >= 20]), pch=19)
```



We plot the variance versus the mean for all cells in the table with at least 20 observations. For convenience we used a log-log scale. Clearly, the assumption of constant variance used by OLS is not valid. Although the variance is not exactly equal to the mean, it is not far from being proportional to it. Thus, we conclude that we can do far more justice to the data by fitting Poisson regression models than by clinging to ordinary linear models.

2.2 Fit Poisson regression

The dataset only contains grouped data instead of babies each individual woman has. But we only summarized mean of number of babies, variance of number of babies and the sample size of women with identical predictors in one row.

Are we able to still run Poisson regression? Yes, but using the nice "offset" feature of "glm" function. Suppose the l-th woman in a group has Y_l babies. The group total is $Y = \sum_{l=1}^{n} Y_l$ and we have n women in this group. In Poisson regression, if we know each Y_l, we assume

$$\log E[Y_l] = X'\beta.$$

Note that all Y_l's have shared predictors X. Then we have

$$\log E[Y] = \log E[\sum_{l=1}^{n} Y_l] = \log \sum_{l=1}^{n} E[Y_l] = \log nE[Y_l] = \log n + X'\beta.$$

This means if we only observe Y, but fail to observe each Y_l, we can still treat Y as the observed count data, but with an additional offset term $\log n$. Let us now add the offset term and fit the Poisson regression.

```
fit.ceb <- glm(y ~ dur + res + educ + offset(log(n)),
               family=poisson, data=ceb, contrasts=contrasts)
summary(fit.ceb)
```

```
##
```

```
## Call:
## glm(formula = y ~ dur + res + educ + offset(log(n)), family = poisson,
##     data = ceb, contrasts = contrasts)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2960  -0.6641   0.0725   0.6336   3.6782
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.11710    0.05491  -2.132 0.032969 *
## dur5-9       0.99693    0.05274  18.902  < 2e-16 ***
## dur10-14     1.36940    0.05107  26.815  < 2e-16 ***
## dur15-19     1.61376    0.05119  31.522  < 2e-16 ***
## dur20-24     1.78491    0.05121  34.852  < 2e-16 ***
## dur25-29     1.97641    0.05003  39.501  < 2e-16 ***
## resrural     0.15166    0.02833   5.353 8.63e-08 ***
## resurban     0.11242    0.03250   3.459 0.000541 ***
## educlower    0.02297    0.02266   1.014 0.310597
## educupper   -0.10127    0.03099  -3.268 0.001082 **
## educsec+    -0.31015    0.05521  -5.618 1.94e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 3731.852  on 69  degrees of freedom
## Residual deviance:   70.665  on 59  degrees of freedom
## AIC: 522.14
##
## Number of Fisher Scoring iterations: 4
```

2.3 Interpret results

- What is the expected number of children for a Suvanese woman with no education who have been married 0-4 years?

$$\exp\{-0.1173\} = 0.89.$$

- As we move from duration 0-4 to 5-9 the log of the mean increases by almost one, which means that the number of CEB gets multiplied by $\exp\{0.9977\} = 2.71$. By duration 25-29, women in each category of residence and education have $\exp\{1.977\} = 7.22$ times as many children as they did at duration 0-4.
- The effects of residence show that Suvanese women have the lowest fertility. At any given duration since first marriage, women living in other urban areas have 12% larger families ($exp\{0.1123\} = 1.12$) than Suvanese women with the same level of education. Similarly, at any fixed duration, women who live in rural areas have 16% more children ($\exp\{0.1512\} = 1.16$), than Suvanese women with the same level of education.
- Finally, we see that higher education is associated with smaller family sizes net of duration and residence. At any given duration of marriage, women with upper primary education have 10% fewer kids, and women with secondary or higher education have 27% fewer kids ($1 - \exp\{-0.3096\} = 0.27$), than women with no education who live in the same type of place of residence.

Do we have interaction effect between marital duration and education here?

From the model itself, we did not include any interaction term. However, the model is additive in the log

scale. In the original scale the model is multiplicative, and postulates relative effects which translate into different absolute effects depending on the values of the other predictors.

Consider the effect of education. Women with secondary or higher education have 27% fewer kids than women with no education.

```
newdata <- data.frame(dur=rep(levels(ceb$dur), each=2), res=rep('Suva',12),
                      educ=rep(c('none','sec+'), 6), n=rep(1,12))

newdata$prediction <- predict(fit.ceb, newdata=newdata, type = "response")
newdata
```

```
##      dur  res educ n prediction
## 1    0-4 Suva none 1  0.8894987
## 2    0-4 Suva sec+ 1  0.6523026
## 3    5-9 Suva none 1  2.4105084
## 4    5-9 Suva sec+ 1  1.7677157
## 5  10-14 Suva none 1  3.4983737
## 6  10-14 Suva sec+ 1  2.5654879
## 7  15-19 Suva none 1  4.4667458
## 8  15-19 Suva sec+ 1  3.2756312
## 9  20-24 Suva none 1  5.3005682
## 10 20-24 Suva sec+ 1  3.8871043
## 11 25-29 Suva none 1  6.4192929
## 12 25-29 Suva sec+ 1  4.7075068
```

Note that the effect of 27% fewer kids means different number of kids when the marital duration changes. If we had used OLS regression for these data we would have ended up with a large number of interaction effects to accommodate the fact that residence and educational differentials increase with marital duration.