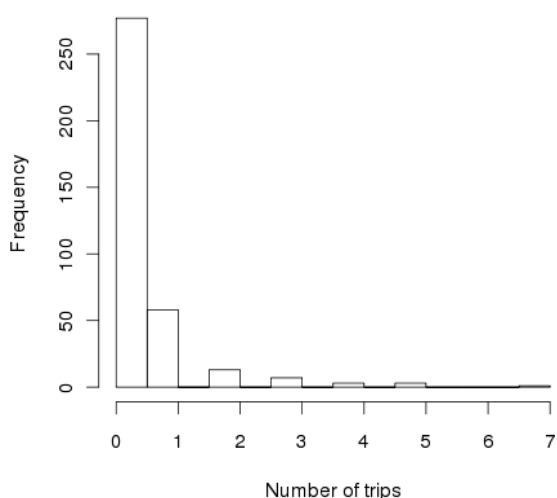


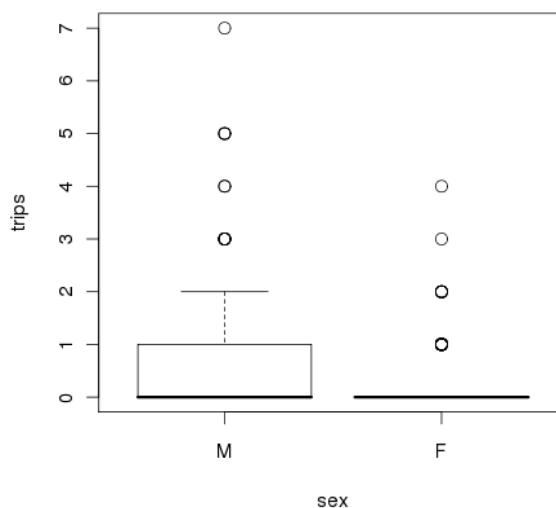
Individual Level Data

The data described in the following graphs comes from a survey of Brazilian tourists in the Atlantic Coastal Forest of Brazil (see Simonoff, p. 53, exercise 4). The aim of the study was to relate the number of of trips to adventure tourism areas in the region to characteristics of the tourists, including age, sex and monthly household income, and estimated travel cost (in U.S. dollars) of the respondent.

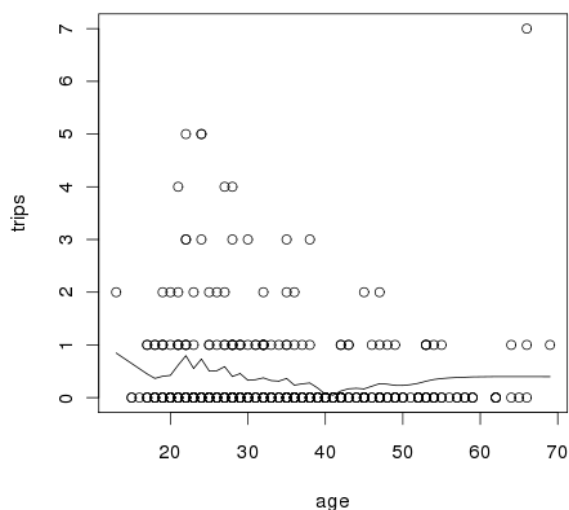
Histogram of trips



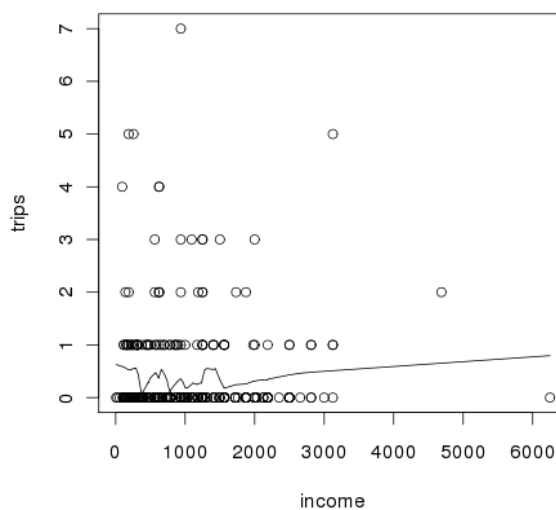
trips vs. sex



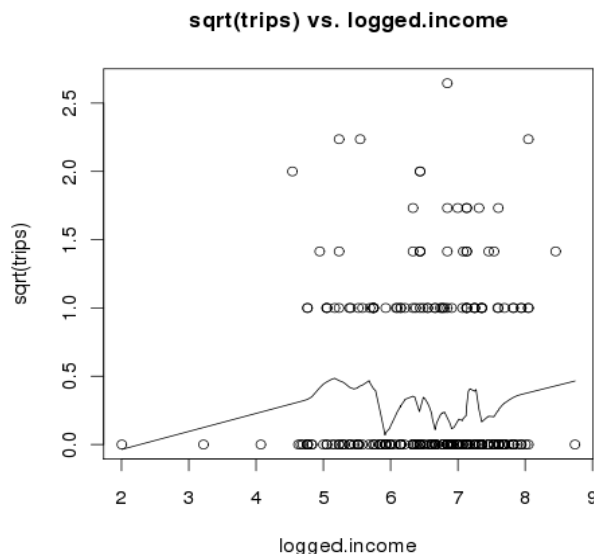
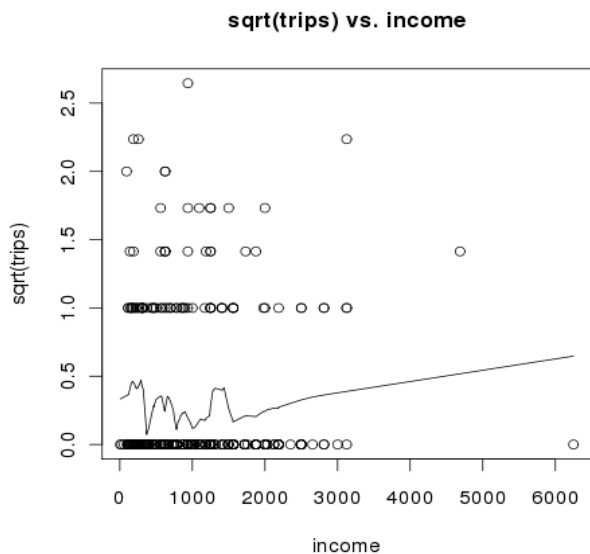
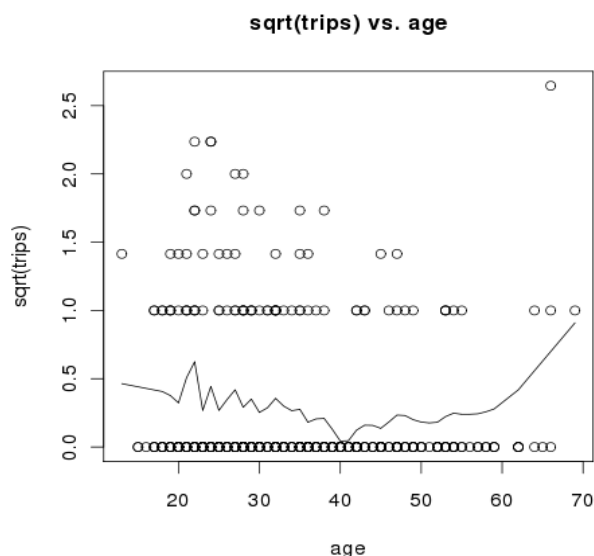
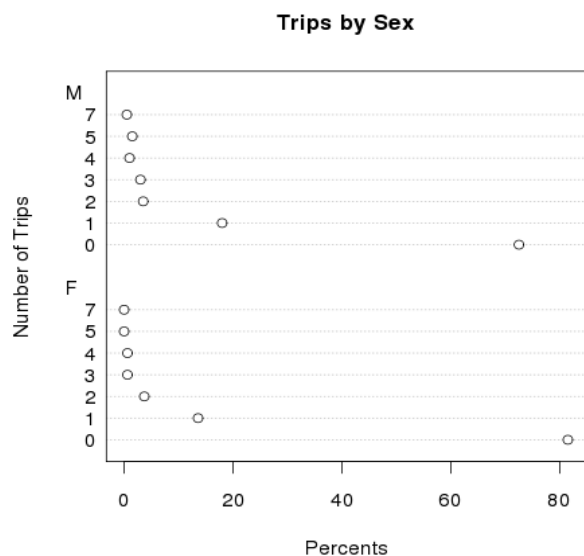
trips vs. age



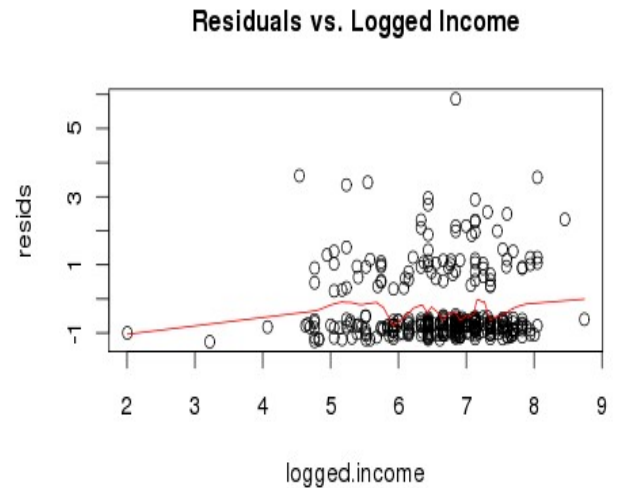
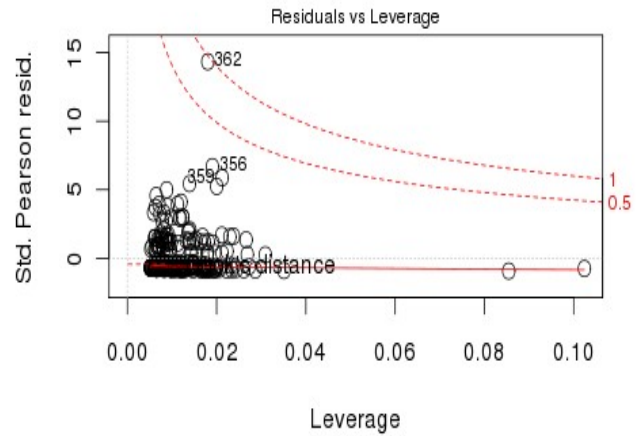
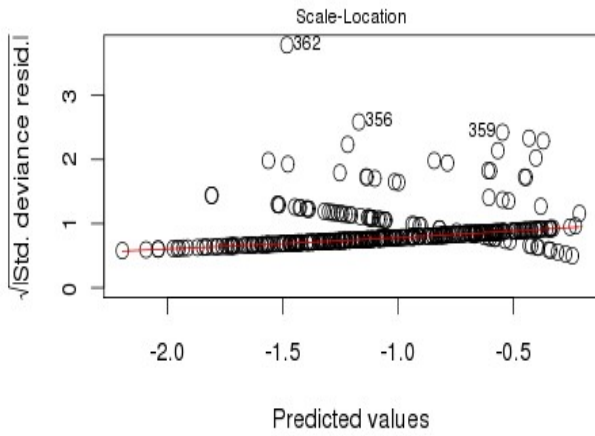
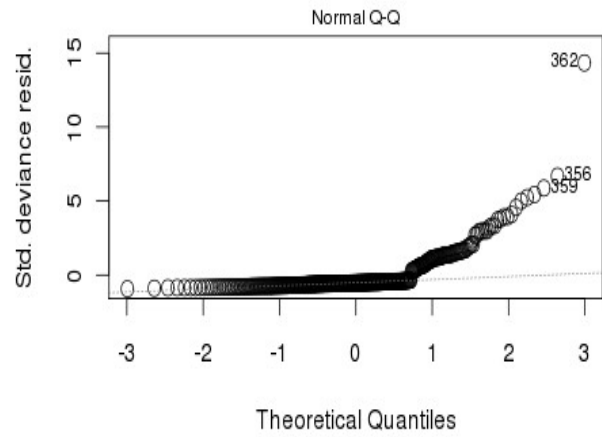
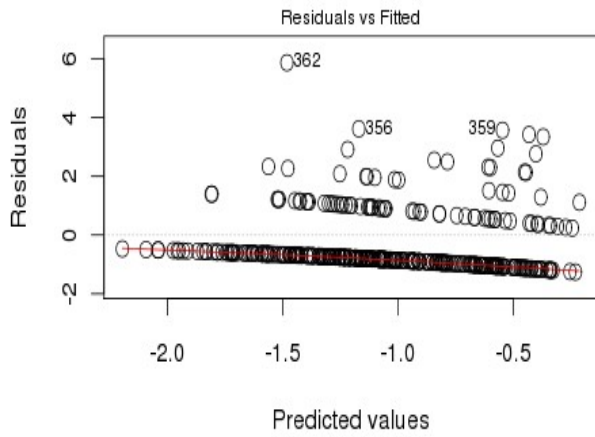
trips vs. income



The plots below are chosen to be more appropriate to the discrete nature of the response variable. The dot plot summarizes percentages. In the scatter plots, a square root transformation has been applied to the number of trips. This transformation is oftent helpful when examining count data – it is the variance stabilizing transformation for variables where the variance is proportional to the mean. Note that the distribution of income has outliers, which will have high leverage values. A log transformation is one solution to this issue.



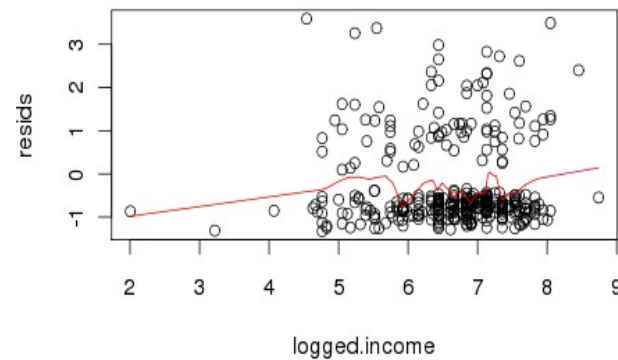
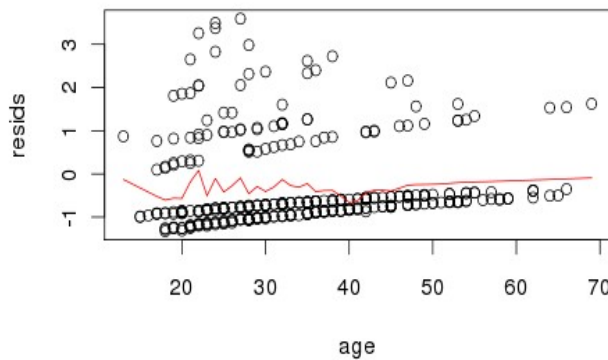
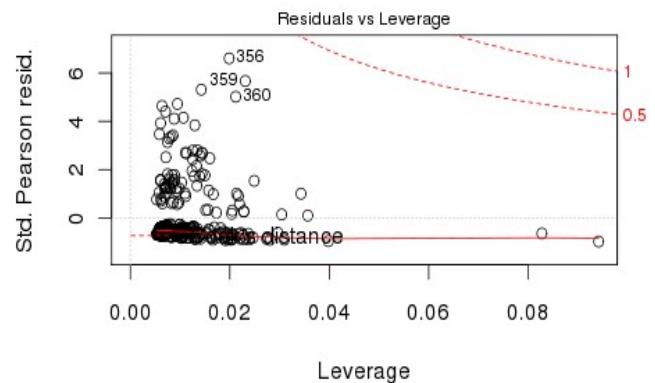
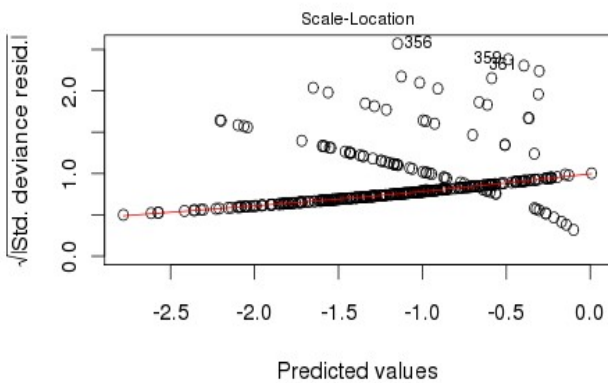
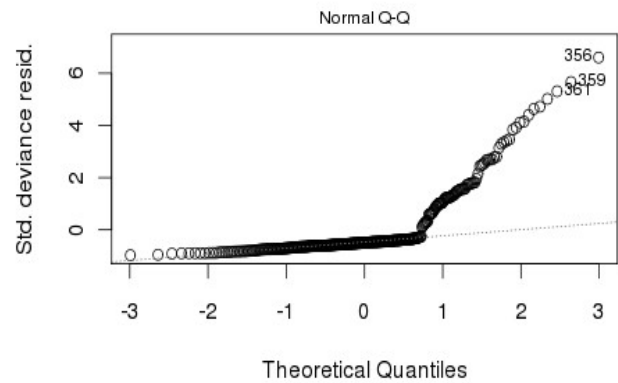
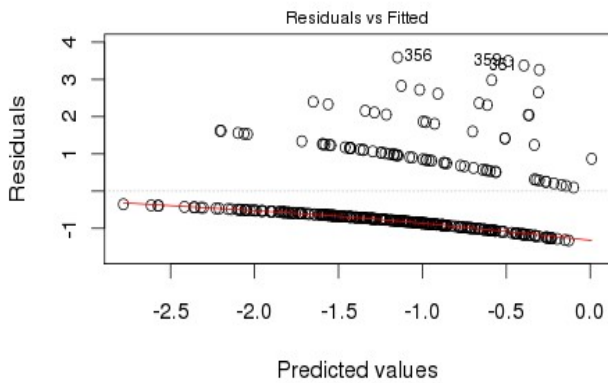
Diagnostic plots are given on the next page for a Poisson regression model with `trips ~ sex + age + logged.income`



Observation 362, which has a case id of #412, is listed below

```
trips sex age income
#412   7  M  66 937.5
```

Diagnostics for an analysis excluding this case are given below:



A case should not be deleted just because it is an outlier or influential. In this situation, the 7 trips reported for case #412 appears so completely aberrant that I did so, resulting in the fit below:

Call: glm(formula = trips ~ sex + age + logged.income, family = poisson)

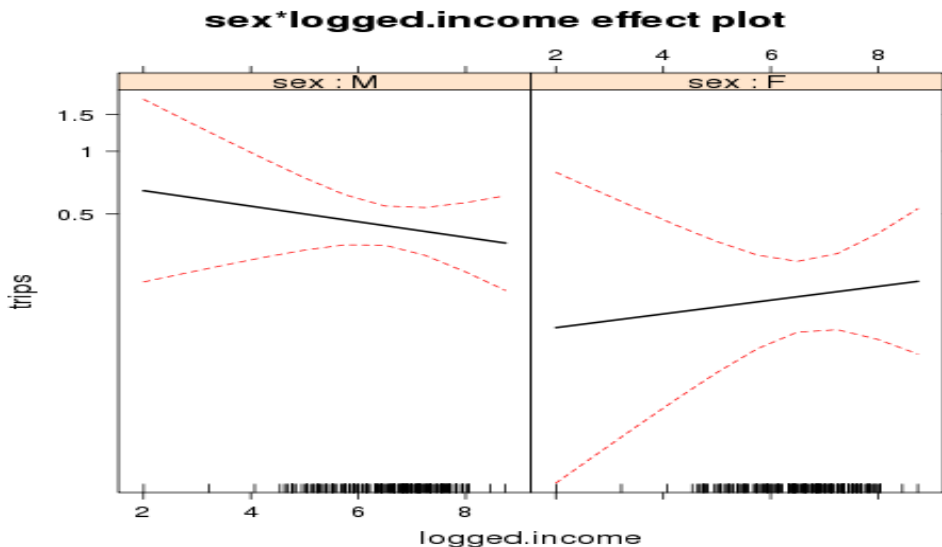
Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.763473	0.645409	1.183	0.236837
sexF	-0.671287	0.188941	-3.553	0.000381 ***
age	-0.039681	0.008792	-4.513	6.39e-06 ***
logged.income	-0.037189	0.095166	-0.391	0.695959

Null deviance: 429.34 on 360 degrees of freedom
 Residual deviance: 394.57 on 357 degrees of freedom
 AIC: 593.73

Mainly for the purposes of illustration, I have considered a model with interaction terms between sex and the remaining variables. If interactions are to be considered for biological data, sex and/or age interactions are typically the first that need to be explored. The effects plot (achieved using the `effects` package) reveals a potential interaction between sex and income.

The overall



contributions to the interactions above can be assessed through significance testing, or more appropriately, by consideration of the predictive properties of the model as measured by AIC. In this instance we use the AIC-based C_p statistic.

	Resid.	Df	Resid. Dev	Df	Deviance	P(> Chi)	Cp
1		357	394.5672	NA	NA	NA	402.5672
2		355	393.3296	2	1.23758	0.5385958	405.329

Dealing with Overdispersion

In the following we examine potential overdispersion in the shark attack and Brazil trip data and methods to accommodate it when it is evident.

Re-examination of Shark model

```
> llfitSeg <- glm( attacks ~ offset(log(population)) + (year > 1985)
+ upTo1985 + after1985, family=poisson)
> muHat <- fitted(llfitSeg)
# Dispersion estimate from residual based chi-square statistic
> estDisper <- sum((attacks - muHat)^2/muHat)/50
> cat("Estimated dispersion is", signif(estDisper,3) , fill=TRUE)
```

Estimated dispersion is 2.12

```
> t.test(((attacks - fitted(llfitSeg))^2 -obs)/fitted(llfitSeg))
```

One Sample t-test

data: z1

t = 2.277, df = 53, p-value = 0.02685

...

Fit quasipoisson model to accommodate overdispersion

```
> summary(glm( attacks ~ offset(log(population)) + (year > 1985) +
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-51.01906	23.13117	-2.206	0.03203	*
year > 1985TRUE	-123.19614	55.82345	-2.207	0.03194	*
upTo1985	0.01862	0.01173	1.587	0.11874	
after1985	0.08054	0.02548	3.161	0.00267	**

(Dispersion parameter for quasipoisson family taken to be 2.115509)

Test significance of 3rd order polynomial using quasi-likelihood

```
> qFitPoly <- glm( attacks ~ offset(log(population)) + poly(year,3),
+ family=quasipoisson)
```

```
> anova(qFitPoly,test="F")
Analysis of Deviance Table
Model: quasipoisson, link: log
Response: attacks
```

	Df	Deviance	Resid.	Df	Resid. Dev	F	Pr(>F)
NULL				53	176.934		
poly(year, 3)	3	69.756		50	107.177	10.963	1.206e-05 ***

```
# Re-examination of Brazil trips model
```

```
> llFit <- glm( trips ~ sex + age + logged.income , family=poisson)
> t.test(((trips - fitted(llFit))^2 - obs)/fitted(llFit))
```

One Sample t-test

```
data: z1
t = 3.494, df = 360, p-value = 0.0005352
...
```

```
> summary(glm( trips ~ sex + age + logged.income ,
family=quasipoisson))
```

(see next page)

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.76347	0.80601	0.947	0.344164
sexF	-0.67129	0.23596	-2.845	0.004698 **
age	-0.03968	0.01098	-3.614	0.000345 ***
logged.income	-0.03719	0.11885	-0.313	0.754527

(Dispersion parameter for quasipoisson family taken to be 1.559586)

```
# Here's a function definition that "automates" the score test
The arguments are obs= the response variable (vector of counts) and
mod= the object returned by the the glm() command.
```

```
dispScore.test <- function(obs,mod) {
z1 <- ((obs - fitted(mod))^2 - obs)/fitted(mod)
t.test(z1) }
```

Accounting for overdispersion using the Negative Binomial Model

```
> library(MASS)
> nbFitSeg <- glm.nb( attacks ~ offset(log(population)) + (year >
1985) +
+   upTo1985 + after1985)
>
> summary(nbFitSeg)
```

Call:

```
glm.nb(formula = attacks ~ offset(log(population)) + (year >
1985) + upTo1985 + after1985, init.theta = 6.90522872916435,
link = log)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.7325	-0.8635	-0.2523	0.2918	2.3645

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-55.17955	19.79594	-2.787	0.00531	**
year > 1985TRUE	-129.29020	65.31419	-1.980	0.04776	*
upTo1985	0.02073	0.01005	2.063	0.03910	*
after1985	0.08569	0.03123	2.744	0.00607	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(6.9052) family taken to be 1)

Null deviance: 93.516 on 53 degrees of freedom

Residual deviance: 63.226 on 50 degrees of freedom

AIC: 270.62

Number of Fisher Scoring iterations: 1

Theta: 6.91
Std. Err.: 3.28

2 x log-likelihood: -260.622

Brazilian Tourist Data

```
summary(glm.nb( trips ~ sex + age + logged.income ))
```

Call:

```
glm.nb(formula = trips ~ sex + age + logged.income, init.theta =
0.557509081541586,
link = log)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.0288	-0.7680	-0.6395	-0.4198	2.2879

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.77625	0.88722	0.875	0.381617
sexF	-0.68302	0.24230	-2.819	0.004820 **
age	-0.03899	0.01106	-3.526	0.000421 ***
logged.income	-0.04178	0.12897	-0.324	0.745964

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(0.5575) family taken to be 1)

Null deviance: 252.14 on 360 degrees of freedom
Residual deviance: 231.20 on 357 degrees of freedom
AIC: 553.83

Number of Fisher Scoring iterations: 1

Theta: 0.558
Std. Err.: 0.153

2 x log-likelihood: -543.830