3302 Final Project Doctor Visits

Charles Costanzo

2023-04-12

Create a variable table

```
variable_table <- data.frame(Variable = c("visits","gender","age","income","illness","reduced",
                          "health","private","freepoor","freerepat","nchronic",
                          "lchronic"),
            Description = c("Number of doctor visits in past 2 weeks.",
                            "Factor indicating gender.",
                            "Age in years divided by 100.",
                            "Annual income in tens of thousands of dollars.",
                            "Number of illnesses in past 2 weeks.",
                            "Number of days of reduced activity in past 2 weeks
                            due to illness or injury.",
                            "General health questionnaire score using Goldberg's method.",
                            "Factor. Does the individual have private health insurance?",
                            "Factor. Does the individual have free government
                            health insurance due to low income?",
                            "Factor. Does the individual have free government
                            health insurance due to old age, disability or veteran status?",
                            "Factor. Is there a chronic condition not limiting activity?",
                            "Factor. Is there a chronic condition limiting activity?"))

# Create table for export
knitr::kable(variable_table) %>%
  kable_styling(full_width = F,
                font_size = 12,
                position = "left")
```

Variable	Description
visits	Number of doctor visits in past 2 weeks.
gender	Factor indicating gender.
age	Age in years divided by 100.
income	Annual income in tens of thousands of dollars.
illness	Number of illnesses in past 2 weeks.
reduced	Number of days of reduced activity in past 2 weeks due to illness or injury.
health	General health questionnaire score using Goldberg's method.
private	Factor. Does the individual have private health insurance?
freepoor	Factor. Does the individual have free government health insurance due to low income?
freerepat	Factor. Does the individual have free government health insurance due to old age, disability
nchronic	Factor. Is there a chronic condition not limiting activity?
lchronic	Factor. Is there a chronic condition limiting activity?

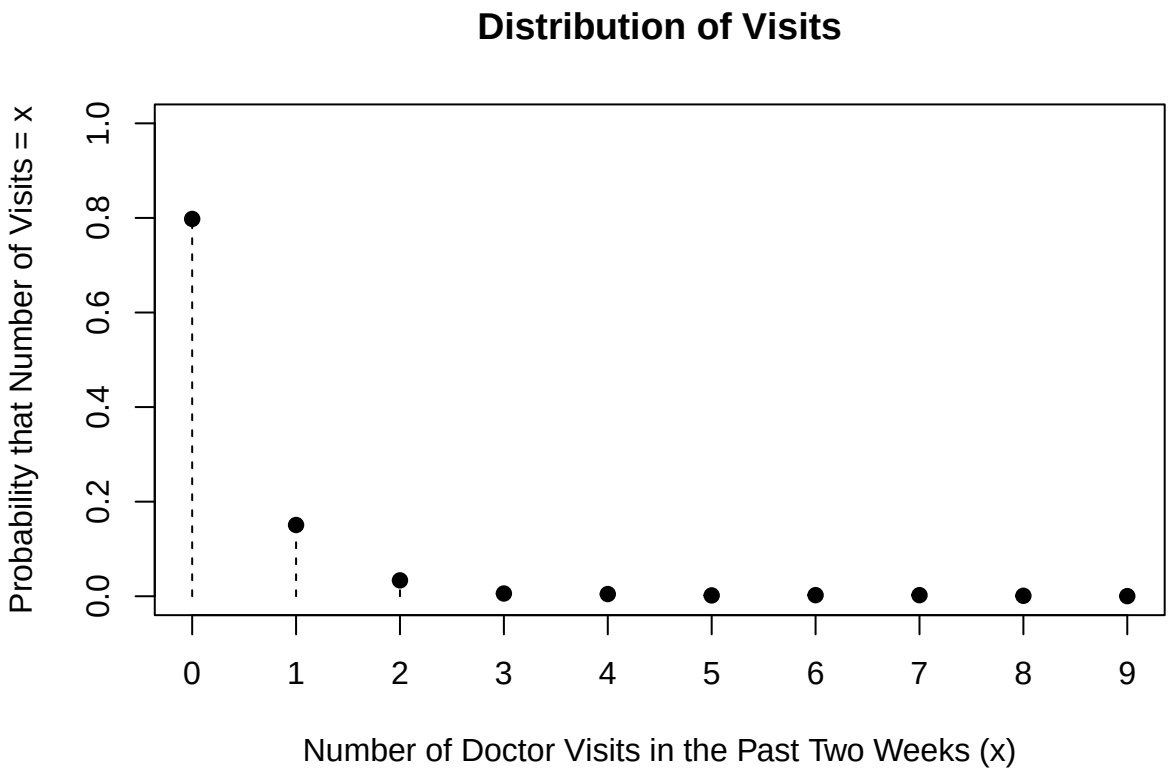
Create density plot for visits

```
library(ggthemes)
p <- ggplot(data = doctor) +
  geom_density(aes(doctor$visits)) +
  xlab("Number of Doctor Visits in Past 2 Weeks") +
  ylab("Density") +
  ggtitle("Density of Dependent Variable") +
  scale_x_continuous(labels = c(0, 1,2,3,4,5,6,7,8,9,10),
                     breaks = c(0, 1,2,3,4,5,6,7,8,9,10),
                     lim = c(-1,10)) +
  theme_hc() + scale_colour_hc()

# theme(rect = element_rect(fill = "transparent"),
#         panel.background = element_rect(fill = "transparent",
#         colour = NA_character_))
```

```
poisson_y <- doctor %>%
  group_by(visits) %>%
  summarize(count = n(),
            prob = count/5190)

x_vec <- poisson_y$visits # vector of sample elements
y_vec <- poisson_y$prob # probability that count = x

i <- 0
plot(x_vec, y_vec, ylim = c(0, 1), pch = 19,
     xlab = "Number of Doctor Visits in the Past Two Weeks (x)",
     ylab = "Probability that Number of Visits = x", xaxt = "n",
     main = "Distribution of Visits") +
  axis(side = 1, at = x_vec, tick = TRUE)
```

```
## numeric(0)
```

```
for (i in 1:length(x_vec)) { # connect dots to x axis
  segments(x_vec[i], 0, x_vec[i], y_vec[i], lty = 2)
}
```



Save plot as a .png file in "plots" folder

```
ggsave("visits_density.png", p, bg="transparent",
       path = paste(getwd(),"/plots",sep=""))
```

```
## Saving 6.5 x 4.5 in image
```

Create distribution plot for `private` vs `insurance`

```
p2 <- ggplot(data = doctor) +
  geom_violin(aes(x = freepoor,
                  y = income,
                  fill = as.factor(freepoor))) +
  geom_boxplot(aes(x = freepoor,
                   y = income),
               width = 0.1) +
  xlab("") +
  ylab("Annual Income (Tens of Thousands of Australian Dollars)") +
  ggtitle("Parallel Box and Violin Plots for Income by Free Low Income Insurance") +
  theme_hc() + scale_fill_manual(name = "Free Government Insurance due to Low Income",
    values = c("#6B9080","#A4C3B2")) +
  coord_flip()
```

Save plot as a .png file in "plots" folder

```
ggsave("private_income_boxplot.png", p2, bg="transparent",
       path = paste(getwd(),"/plots",sep=""))
```

```
## Saving 6.5 x 4.5 in image
```

Summarize `visits` values

```
doctor %>%
  group_by(as.factor(visits)) %>%
  summarize(count = n())
```

```
## # A tibble: 10 x 2
##    `as.factor(visits)` count
##    <fct>               <int>
##  1 0                    4141
##  2 1                     782
##  3 2                     174
##  4 3                      30
##  5 4                      24
##  6 5                       9
##  7 6                      12
##  8 7                      12
##  9 8                       5
## 10 9                       1
```

Get mean number of doctor visits by insurance type

```
mean_visits_insur <- doctor %>%
  group_by(private, freepoor, freerepat) %>%
  summarize(mean_visits = mean(visits, na.rm = TRUE))
```

```
## `summarise()` has grouped output by 'private', 'freepoor'. You can override
## using the `.groups` argument.
```

```
mean_visits_insur
```

```
## # A tibble: 4 x 4
## # Groups:   private, freepoor [3]
##   private freepoor freerepat mean_visits
##   <fct>   <fct>    <fct>           <dbl>
## 1 no      no       no              0.218
## 2 no      no       yes             0.467
## 3 no      yes      no              0.158
## 4 yes     no       no              0.295
```

```
mean_visits_insur_table <- data.frame(
  Insurance = c("Private","Free Government Insurance due to Low Income",
      "Free Government Insurance due to Old Age, Disability, or Veteran Status",
      "No Insurance"),
  `Mean Doctor Visits` =
    c(mean_visits_insur$mean_visits[mean_visits_insur$private == "yes"],
      mean_visits_insur$mean_visits[mean_visits_insur$freepoor == "yes"],
      mean_visits_insur$mean_visits[mean_visits_insur$freerepat == "yes"],
mean_visits_insur$mean_visits[mean_visits_insur$private == "no" & mean_visits_insur$freepoor == "no" & m
    ))

mean_visits_insur_table$Mean.Doctor.Visits <-
  round(mean_visits_insur_table$Mean.Doctor.Visits, 3)

names(mean_visits_insur_table) <- c("Insurance Type", "Mean Doctor Visits (past 2 weeks)")

knitr::kable(mean_visits_insur_table) %>%
  kable_styling(full_width = F,
                font_size = 16,
                position = "left")
```

Insurance Type	Mean D
Private	
Free Government Insurance due to Low Income	
Free Government Insurance due to Old Age, Disability, or Veteran Status	
No Insurance	

Sex	Mean Doctor Visits (past 2 weeks)
Male	0.2363344
Female	0.3619541

Get mean number of doctor visits by gender

```
mean_visits_gender <- doctor %>%
  group_by(gender) %>%
  summarize(Mean.Doctor.Visits = mean(visits, na.rm = TRUE))

mean_visits_gender_table <- data.frame(
  Gender = c("Male","Female"),
  Mean.Doctor.Visits =
    c(mean_visits_gender$Mean.Doctor.Visits[mean_visits_gender$gender == "male"],
      mean_visits_gender$Mean.Doctor.Visits[mean_visits_gender$gender == "female"]))

names(mean_visits_gender_table) <- c("Sex", "Mean Doctor Visits (past 2 weeks)")

knitr::kable(mean_visits_gender_table) %>%
  kable_styling(full_width = F,
                font_size = 30,
                position = "center")
```

Model Description

Let $\mathbf{Y}$ be our vector of data containing the number of doctor visits within the past two weeks, `visits`.

$$\text{Assume } \{Y_1 \ldots, Y_{n=5190}\} \overset{ind}{\sim} Poisson(\lambda_i)$$

We will use the log link.

Poisson GLM:

$$\eta_i = g(\lambda_i) = log(\lambda_i) = \mathbf{x}_i^\intercal \boldsymbol{\beta}, \;\; i = 1, \ldots, n = 5190$$

where

- η_i is the estimated log mean number of doctor visits in the past two weeks for individual i (according to their specific covariate values $\mathbf{x}_i$)
- $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,p=12})^\intercal$ are the p covariates (including an intercept) for individual i
- $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{p=12})$ is our unknown coefficient vector.

run first Poisson model for visits

```
model1 <- glm(visits ~ .,
              family = "poisson",
              data = doctor)
summary(model1)
```

```
##
## Call:
## glm(formula = visits ~ ., family = "poisson", data = doctor)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9502  -0.6858  -0.5747  -0.4852   5.7055
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.097821   0.101554 -20.657  < 2e-16 ***
## genderfemale  0.156490   0.056139   2.788  0.00531 **
## age           0.279123   0.165981   1.682  0.09264 .
## income       -0.187416   0.085478  -2.193  0.02834 *
## illness       0.186156   0.018263  10.193  < 2e-16 ***
## reduced       0.126690   0.005031  25.184  < 2e-16 ***
## health        0.030683   0.010074   3.046  0.00232 **
## privateyes    0.126498   0.071552   1.768  0.07707 .
## freepooryes  -0.438462   0.179799  -2.439  0.01474 *
## freerepatyes  0.083640   0.092070   0.908  0.36365
## nchronicyes   0.117300   0.066545   1.763  0.07795 .
## lchronicyes   0.150717   0.082260   1.832  0.06692 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 5634.8  on 5189  degrees of freedom
## Residual deviance: 4380.1  on 5178  degrees of freedom
## AIC: 6735.7
##
## Number of Fisher Scoring iterations: 6
```

Create Analysis of Deviance Table

```
anova(model1, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: visits
##
## Terms added sequentially (first to last)
##
##
##           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                       5189     5634.8
## gender     1    68.64      5188     5566.2 < 2.2e-16 ***
## age        1   118.90      5187     5447.3 < 2.2e-16 ***
## income     1    12.42      5186     5434.9 0.0004239 ***
## illness    1   354.29      5185     5080.6 < 2.2e-16 ***
## reduced    1   673.63      5184     4406.9 < 2.2e-16 ***
## health     1     9.57      5183     4397.4 0.0019818 **
## private    1     3.95      5182     4393.4 0.0468745 *
## freepoor   1     7.96      5181     4385.5 0.0047840 **
## freerepat  1     1.25      5180     4384.2 0.2644339
## nchronic   1     0.74      5179     4383.5 0.3897203
## lchronic   1     3.34      5178     4380.1 0.0676851 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conduct a Chi-Square Test at the $\alpha = 0.05$ level to demonstrate that the `freerepat`, `nchronic`, and `lchronic` variables should be removed from the model.

H_0 : Reduced Model (without `freerepat`, `nchronic`, and `lchronic`) is good enough vs.

H_1 : Full Model (including all predictor variables) is Needed

- $D_{full} = 4380.1$
- $D_{reduced} = 4385.5$

Test Statistic:

$$\Delta D = D_{reduced} - D_{full} = 4385.5 - 4380.1$$
$$\Rightarrow \Delta D = 5.4$$

Under $H_0, \Delta D \sim \chi^2_{pFull-pReduced=12-9}$

$$\Rightarrow \Delta D \sim \chi^2_3$$

Rejection Rule: Reject if $\Delta D >$ `qchisq(1-0.05,3)` $= 7.814728$

Since our Test Statistic $\Delta D = 5.4 < 7.8$, we fail to reject the null hypothesis at the $\alpha = 0.05$ significance level.

p value: `pchisq(5.4, 3, lower.tail = FALSE)` $= 0.1447436$

Conclusion:

The reduced model is good enough (that does not include `freerepat`, `nchronic`, and `lchronic`) and provides a better fit to the data set than the full model that includes those three covariates. We will now drop these covariates and fit a reduced model.

```r
model1_reduced <- glm(visits ~ gender + age + income + illness + reduced + health +
    private + freepoor,
    family = "poisson",
    data = doctor)
summary(model1_reduced)
```

```
##
## Call:
## glm(formula = visits ~ gender + age + income + illness + reduced +
##     health + private + freepoor, family = "poisson", data = doctor)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0180  -0.6811  -0.5772  -0.4916   5.6590
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.072446   0.100191 -20.685  < 2e-16 ***
## genderfemale  0.167591   0.055604   3.014 0.002578 **
## age           0.437894   0.137070   3.195 0.001400 **
## income       -0.203978   0.084206  -2.422 0.015420 *
## illness       0.196366   0.017603  11.155  < 2e-16 ***
## reduced       0.127994   0.004905  26.097  < 2e-16 ***
## health        0.032854   0.009961   3.298 0.000973 ***
## privateyes    0.087156   0.053501   1.629 0.103304
## freepooryes  -0.465788   0.176364  -2.641 0.008265 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 5634.8  on 5189  degrees of freedom
## Residual deviance: 4385.5  on 5181  degrees of freedom
## AIC: 6735
##
## Number of Fisher Scoring iterations: 6
```

Check for overdispersion

```
var(doctor$visits)
```

```
## [1] 0.6370176
```

```
mean(doctor$visits)
```

```
## [1] 0.3017341
```

```
dispersiontest(model1, alternative = "greater")
```

```
##
##  Overdispersion test
##
## data:  model1
## z = 6.5386, p-value = 3.105e-11
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
## dispersion
##   1.415602
```

There is indeed overdispersion, so run a quasipoisson model

Run quasipoisson model

```
model2_reduced <- glm(visits ~ gender + age + income + illness + reduced + health +
    private + freepoor,
    family = "quasipoisson",
    data = doctor)
summary(model2_reduced)
```

```
##
## Call:
## glm(formula = visits ~ gender + age + income + illness + reduced +
##     health + private + freepoor, family = "quasipoisson", data = doctor)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0180  -0.6811  -0.5772  -0.4916   5.6590
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.072446   0.115325 -17.970  < 2e-16 ***
## genderfemale  0.167591   0.064003   2.618  0.00886 **
## age           0.437894   0.157775   2.775  0.00553 **
## income       -0.203978   0.096926  -2.104  0.03539 *
## illness       0.196366   0.020262   9.692  < 2e-16 ***
## reduced       0.127994   0.005645  22.672  < 2e-16 ***
## health        0.032854   0.011465   2.865  0.00418 **
## privateyes    0.087156   0.061583   1.415  0.15705
## freepooryes  -0.465788   0.203005  -2.294  0.02180 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 1.324931)
##
##     Null deviance: 5634.8  on 5189  degrees of freedom
## Residual deviance: 4385.5  on 5181  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 6
```

Also try a zero-inflated Poisson since there are a lot of zeros (see density plot)

```
model3 <- zeroinfl(visits ~ .,
    data = doctor)
summary(model3)
```

```
##
## Call:
## zeroinfl(formula = visits ~ ., data = doctor)
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -1.5991 -0.4452 -0.2880 -0.1896 11.5788
##
## Count model coefficients (poisson with log link):
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.549671   0.143841  -3.821 0.000133 ***
## genderfemale -0.020143   0.071572  -0.281 0.778378
## age          -0.002373   0.217526  -0.011 0.991296
## income       -0.214095   0.109938  -1.947 0.051485 .
## illness       0.044165   0.024656   1.791 0.073259 .
## reduced       0.082782   0.005993  13.813  < 2e-16 ***
## health        0.022713   0.011261   2.017 0.043699 *
## privateyes   -0.021801   0.096733  -0.225 0.821692
## freepooryes  -0.382340   0.241160  -1.585 0.112870
## freerepatyes -0.211985   0.118573  -1.788 0.073809 .
## nchronicyes  -0.011807   0.092675  -0.127 0.898626
## lchronicyes   0.001584   0.102220   0.015 0.987639
##
## Zero-inflation model coefficients (binomial with logit link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.59903    0.29413   8.836  < 2e-16 ***
## genderfemale -0.45847    0.17082  -2.684  0.00728 **
## age          -1.31175    0.51221  -2.561  0.01044 *
## income       -0.07151    0.24158  -0.296  0.76721
## illness      -0.44176    0.08320  -5.309 1.10e-07 ***
## reduced      -1.24859    0.23896  -5.225 1.74e-07 ***
## health       -0.07918    0.03876  -2.043  0.04105 *
## privateyes   -0.42978    0.19703  -2.181  0.02916 *
## freepooryes   0.31229    0.51329   0.608  0.54292
## freerepatyes -1.25153    0.31946  -3.918 8.94e-05 ***
## nchronicyes  -0.11607    0.19917  -0.583  0.56006
## lchronicyes  -0.44751    0.30582  -1.463  0.14338
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 47
## Log-likelihood: -3181 on 24 Df
```

Run a model to see what variables impact whether or not someone has private insurance

```
model4 <- glm(private ~ .,
              family = "binomial",
              data = doctor)
summary(model4)
```

```
##
## Call:
## glm(formula = private ~ ., family = "binomial", data = doctor)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3081  -0.9428  -0.0001   0.8459   1.7941
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.80457    0.11925 -15.133  < 2e-16 ***
## visits         0.07415    0.05661   1.310  0.19026
## genderfemale   0.81767    0.07437  10.995  < 2e-16 ***
## age            3.72581    0.24176  15.411  < 2e-16 ***
## income         0.68901    0.10145   6.792 1.11e-11 ***
## illness        0.02544    0.03207   0.793  0.42767
## reduced       -0.02688    0.01620  -1.659  0.09714 .
## health        -0.02151    0.01938  -1.110  0.26706
## freepooryes  -18.41287  418.75057  -0.044  0.96493
## freerepatyes -20.26393  190.61808  -0.106  0.91534
## nchronicyes    0.24067    0.08150   2.953  0.00315 **
## lchronicyes    0.28532    0.13868   2.057  0.03964 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7126.7  on 5189  degrees of freedom
## Residual deviance: 4660.6  on 5178  degrees of freedom
## AIC: 4684.6
##
## Number of Fisher Scoring iterations: 17
```

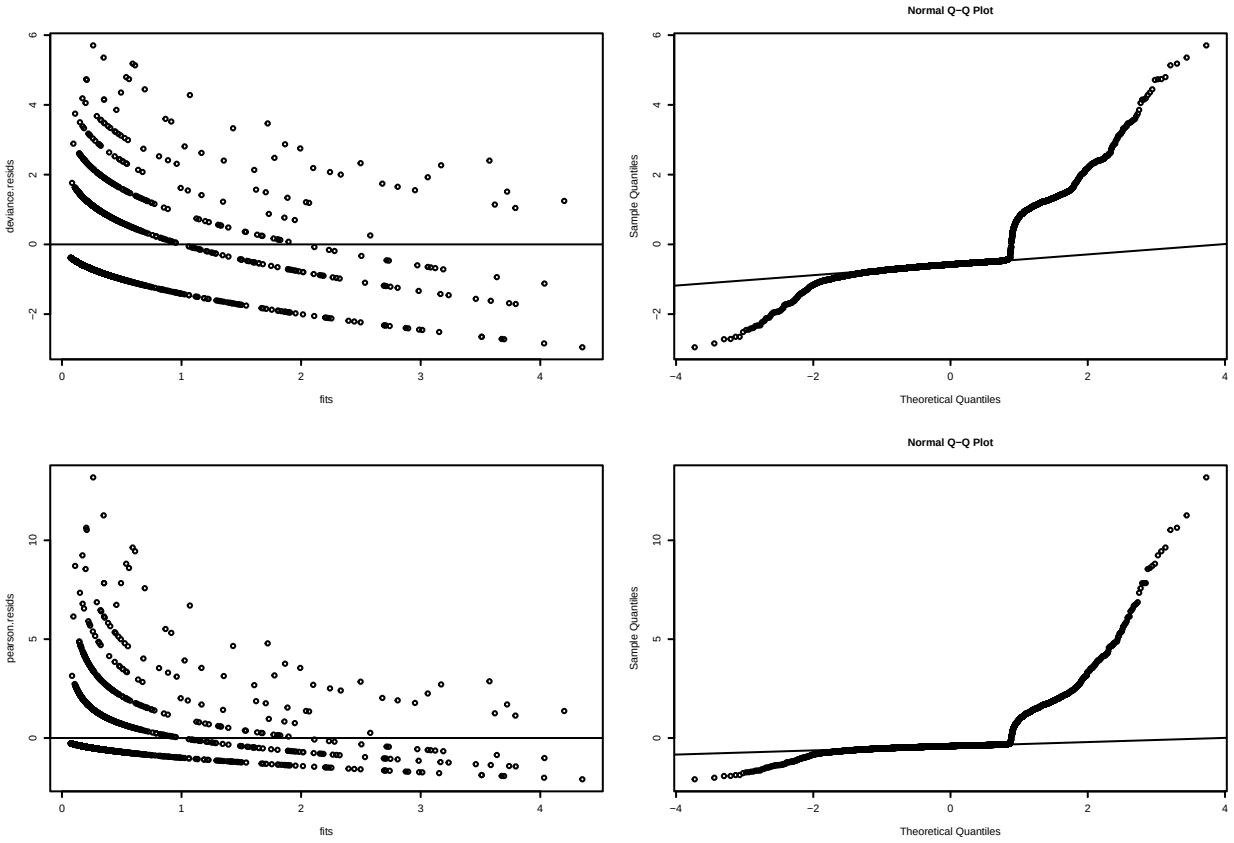
Create stargazer tables for models

```
# stargazer(model1,model1_reduced,model2_reduced, type = "html", out = "count_models.html")

# stargazer(model4, type = "html", out = "logistic_model.html")
```

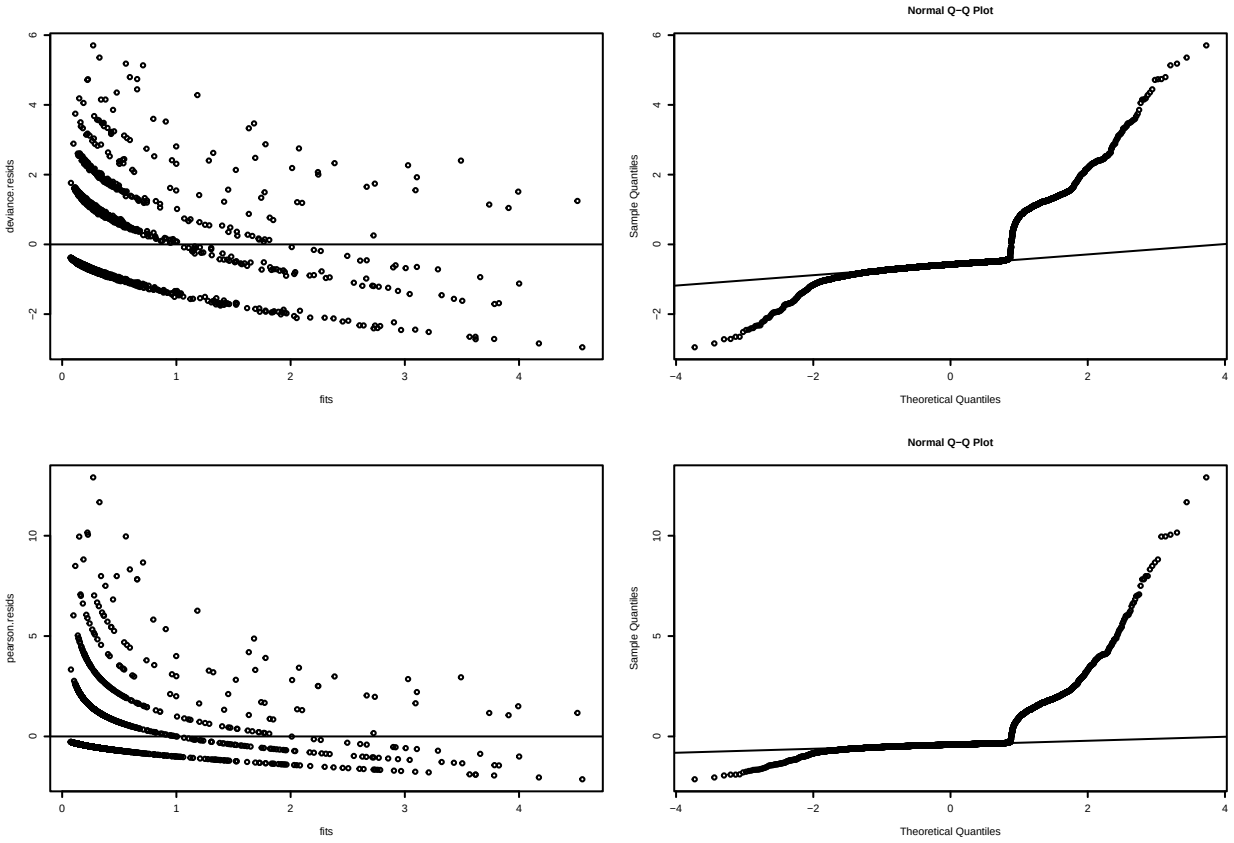
Create a function to run diagnostics on the count models

```
poisson_diagnostics <- function (model){
  fits <- fitted(model)
  deviance.resids <- resid(model1) ## deviance is the default
  par(mfrow = c(2,2), cex = 0.3)
  plot(fits, deviance.resids); abline(h = 0)
  qqnorm(deviance.resids); qqline(deviance.resids)

  pearson.resids <- resid(model, type = "pearson")
  plot(fits, pearson.resids);abline(h = 0)
  qqnorm(pearson.resids) ; qqline(pearson.resids)
}
```

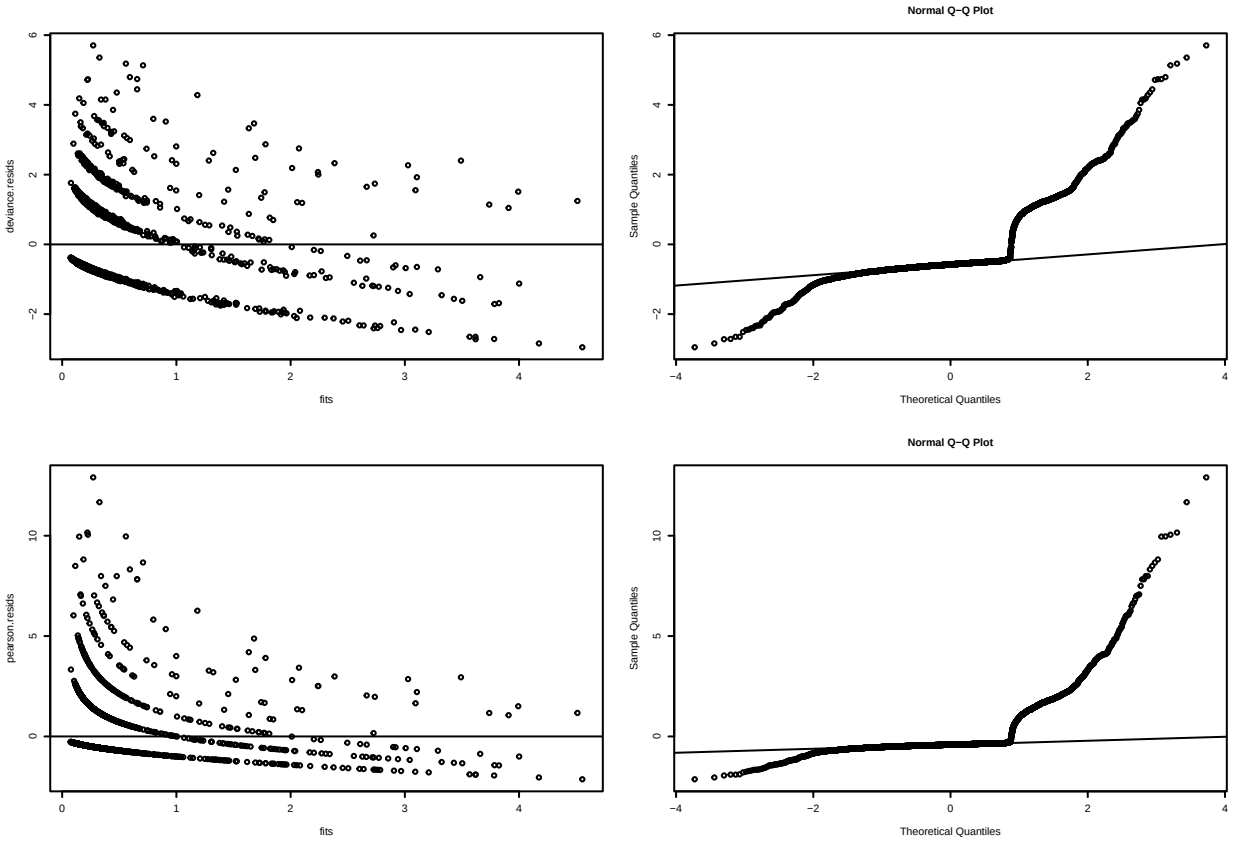
```
poisson_diagnostics(model1)
```



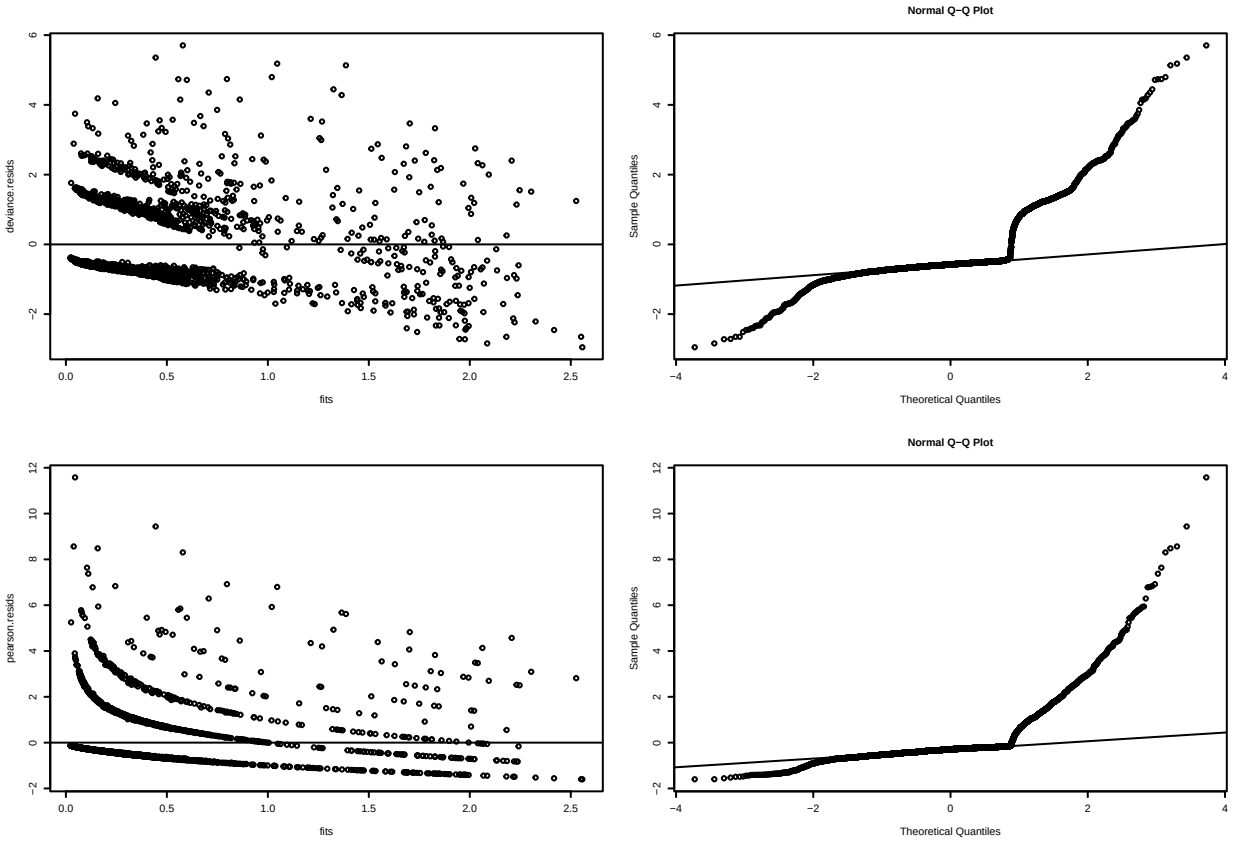
```
poisson_diagnostics(model1_reduced)
```



```
poisson_diagnostics(model2_reduced)
```



```
poisson_diagnostics(model3)
```



Count models for Doctor Visits

Table 1:

	Dependent variable:		
	visits		
	Poisson	*negative binomial*	*zero-inflated count data*
	(1)	(2)	(3)
genderfemale	0.156***	0.216***	−0.020
	(0.056)	(0.070)	(0.072)
age	0.279*	0.331	−0.002
	(0.166)	(0.208)	(0.218)
income	−0.187**	−0.156	−0.214*
	(0.085)	(0.104)	(0.110)
illness	0.186***	0.215***	0.044*
	(0.018)	(0.024)	(0.025)
reduced	0.127***	0.144***	0.083***
	(0.005)	(0.007)	(0.006)
health	0.031***	0.038***	0.023**
	(0.010)	(0.014)	(0.011)
privateyes	0.126*	0.116	−0.022
	(0.072)	(0.086)	(0.097)
freepooryes	−0.438**	−0.497**	−0.382
	(0.180)	(0.211)	(0.241)
freerepatyes	0.084	0.146	−0.212*
	(0.092)	(0.116)	(0.119)
nchronicyes	0.117*	0.098	−0.012
	(0.067)	(0.079)	(0.093)
lchronicyes	0.151*	0.183*	0.002
	(0.082)	(0.103)	(0.102)
Constant	−2.098***	−2.276***	−0.550***
	(0.102)	(0.123)	(0.144)
Observations	5,190	5,190	5,190
Log Likelihood	−3,355.850	−3,199.838	−3,180.927
θ		0.930*** (0.087)	
Akaike Inf. Crit.	6,735.701	6,423.676	
Note:			*p<0.1; **p<0.05; ***p<0.01

Logistic Regression Model for Private Insurance

Table 2:

	Dependent variable:
	private
visits	0.074
	(0.057)
genderfemale	0.818∗ ∗ ∗
	(0.074)
age	3.726***
	(0.242)
income	0.689***
	(0.101)
illness	0.025
	(0.032)
reduced	−0.027*
	(0.016)
health	−0.022
	(0.019)
freepooryes	−18.413
	(418.751)
freerepatyes	−20.264
	(190.618)
nchronicyes	0.241***
	(0.082)
lchronicyes	0.285**
	(0.139)
Constant	−1.805***
	(0.119)
Observations	5,190
Log Likelihood	−2,330.324
Akaike Inf. Crit.	4,684.648
Note:	*p<0.1; **p<0.05; ***p<0.01