Goodness of Fit tests

Steve Hulac

4/1/2019

Load Packages and Data

```r
library(lubridate)
library(tidyverse)

all_variables <- read_csv('all_vars.csv')
```

Filter Outliers and Tabluate

```r
all_vars_no_outliers <- all_variables %>%
  # filter(Hour == 15) %>%
  filter(n < 40)

vol_freq_table <- all_vars_no_outliers %>%
  count(Hourly_Calls = n) %>%
  rename(Observed = nn) %>%
  mutate(Proportion = Observed / sum(Observed))
```

Testing Function

```r
#Once you've created your new variable (and named it Ei) you can run this on the new frequency table to check both the chi-s
quared test and to visualize it.
test <- function(tib) {
  #Chi-Squared Test
  tib %>%
    mutate(chiq = (Observed - Ei)^2 / Ei) %>%
    summarise(Tstat = sum(chiq)) %>%
    pull() -> Tstat
  nrow(tib)-1 -> df
  p <- pchisq(Tstat, df, lower.tail = F) #chisq p-value calc

  tib %>%
    ggplot(aes(x = Hourly_Calls, y = Observed)) +
    geom_col() +
    geom_line(aes(y = Ei)) +
    ggtitle(paste0("p value: ", p, "        T = ", Tstat)) +
    labs(caption = "remember: low p means reject good fit i.e. :(")
}
```
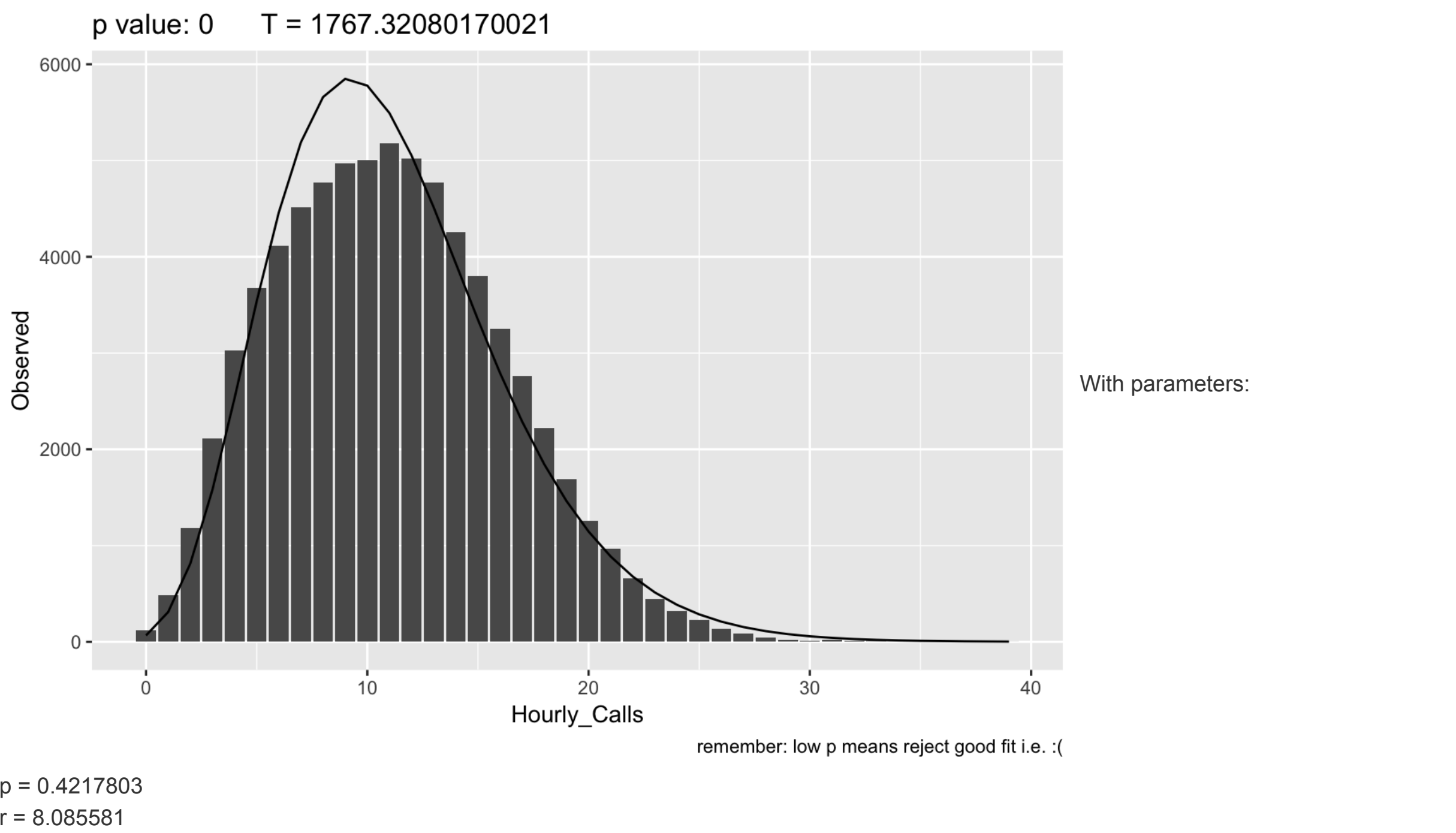
Negative Binomial Test

```r
mu <- mean(all_vars_no_outliers$n)
var <- var(all_vars_no_outliers$n)
p <- mu / var
r <- (mu * p) / (1 - p)

negbin <- vol_freq_table %>%
  mutate(Ei = dnbinom(x = Hourly_Calls, size = r, prob = p) * sum(.$Observed))

test(negbin)
```
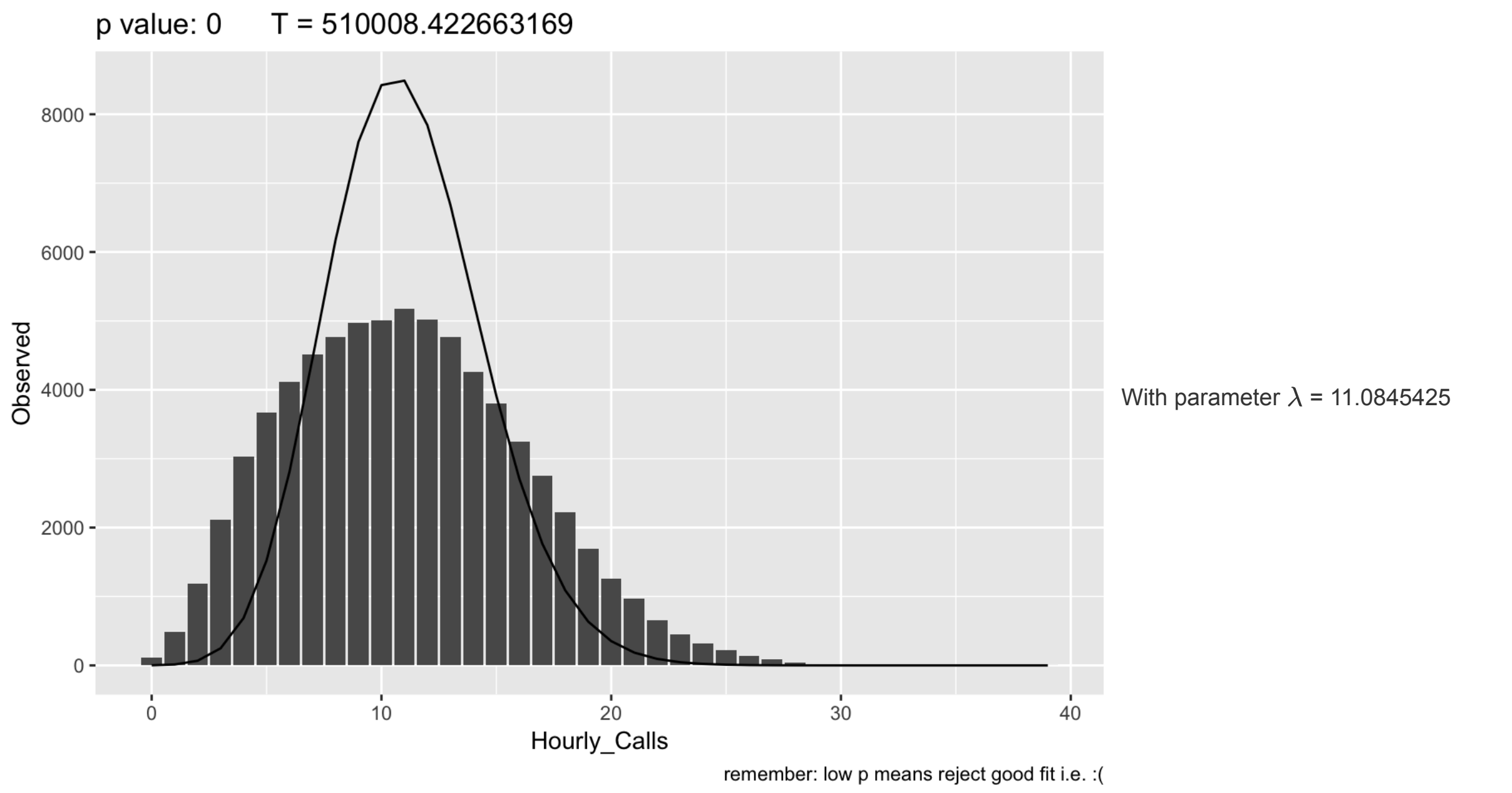

With parameters:

remember: low p means reject good fit i.e. :(

p = 0.4217803
r = 8.085581

This fits if you just select one hour. For instance, uncomment `filter(Hour == 15) %>%` above and this will then fit.

Poisson Test

```r
pois <- vol_freq_table %>%
  mutate(Ei = dpois(x = Hourly_Calls, lambda = mu)  * sum(.$Observed))

test(pois)
```


With parameter λ = 11.0845425

remember: low p means reject good fit i.e. :(

Weibull?

`dweibull()` exists, but I'm having a hard time estimating the model parameters. I found
http://interstat.statjournals.net/YEAR/2000/articles/0010001.pdf but I'm getting tired.

Zipf?

There is a package `zipfR` but I haven't been able to find a PDF function in it, let alone estimate the model parameters.