Modeling Senior Theses

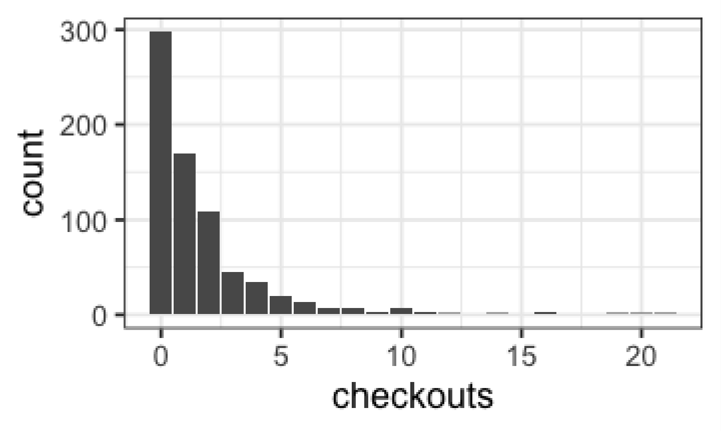
Case study: senior theses

```
theses <- read.csv("../data/sample_theses.csv")
str(theses)
```

```
## 'data.frame':    724 obs. of  3 variables:
##  $ year     : int  1984 2004 1939 1992 1941 1989 1993
##  $ checkouts: int  3 2 1 1 0 0 0 1 1 7 ...
##  $ division : Factor w/ 5 levels "ARTS","HSS","LL",..:
```

Question 1: What is the average number of times a thesis is checked out? (Description)

```
ggplot(theses, aes(x = checkouts)) +
  geom_bar() +
  theme_bw(base_size = 18)
```



```
theses %>%
  summarize(mean(checkouts),
            median(checkouts))
```

```
##   mean(checkouts) median(checkouts)
## 1        1.668508                 1
```

Question 1: What is the average number of times a thesis is checked out? (Inference)

Confidence interval on one mean (approximation method)

$$\bar{x} \pm t^* \times s/\sqrt{n}$$

```
stats <- theses %>%
  summarize(n     = n(),
            x_bar = mean(checkouts),
            s     = sd(checkouts))
stats$x_bar + c(1, -1) * qt(.025, df = stats$n - 1) *
  stats$s / sqrt(stats$n)
```

```
## [1] 1.479991 1.857025
```

Question 1: What is the average number of times a thesis is checked out? (Inference)

Confidence interval on one mean (computational method)

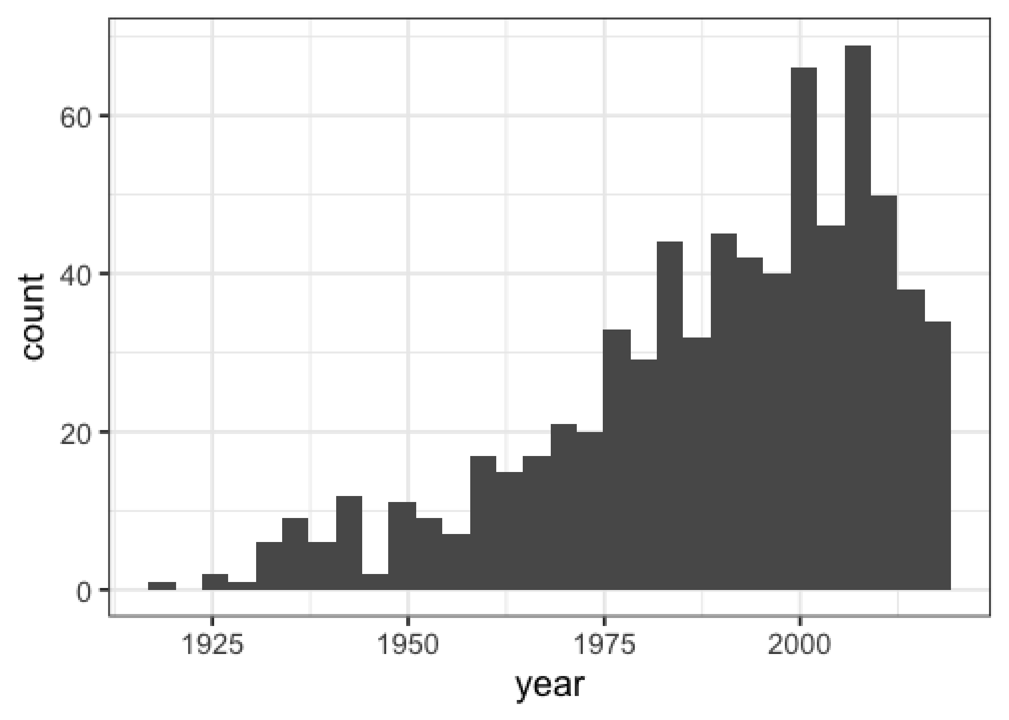
```
se_boot <- theses %>%
  specify(response = checkouts) %>%
  generate(500, type = "bootstrap") %>%
  calculate(stat = "mean") %>%
  summarize(sd(stat)) %>%
  pull()
```

```
stats$x_bar + c(-1, 1) * 2 * se_boot
```

```
## [1] 1.466855 1.870162
```

Question 2: Is the distribution of theses uniform in time? (Description)

```
ggplot(theses, aes(x = year)) +
  geom_histogram() +
  theme_bw(base_size = 18)
```



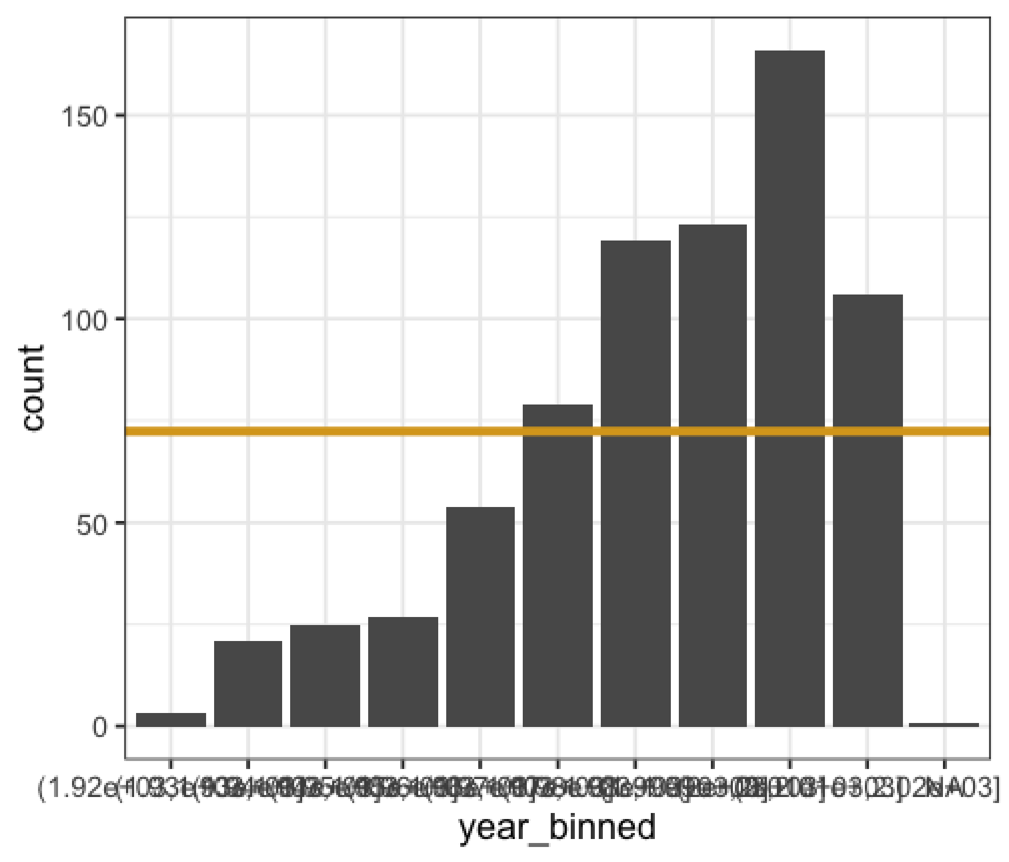
Question 2: Is the distribution of theses uniform in time? (Inference)

Chi-squared goodness of fit test

```
bd <- data.frame(year_binned =
                   cut(theses$year, seq(from = 1920,
                                        to = 2020,
                                        by = 10)))
ggplot(bd, aes(x = year_binned)) +
  geom_bar() +
  geom_hline(yintercept = stats$n/10,
             color = "goldenrod", lwd = 2) +
  theme_bw(base_size = 18)
```

Question 2: Is the distribution of theses uniform in time? (Inference)

Chi-squared goodness of fit test



Question 2: Is the distribution of theses uniform in time? (Inference)

Chi-squared goodness of fit test

```
(obs <- table(bd$year_binned))
```

```
##
## (1.92e+03,1.93e+03] (1.93e+03,1.94e+03] (1.94e+03,1.95e+03]
##                   3                  21                  25
## (1.95e+03,1.96e+03] (1.96e+03,1.97e+03] (1.97e+03,1.98e+03]
##                  27                  54                  79
## (1.98e+03,1.99e+03]     (1.99e+03,2e+03]     (2e+03,2.01e+03]
##                 119                 123                 166
## (2.01e+03,2.02e+03]
##                 106
```

```
(exp <- length(bd$year_binned)/10)
```

```
## [1] 72.4
```

```
(chisq_obs <- sum((obs - exp)^2/exp))
```

```
## [1] 369.7541
```

Question 2: Is the distribution of theses uniform in time? (Inference)

Chi-squared goodness of fit test (approximation method)

```r
pchisq(chisq_obs, df = length(obs) - 1, lower.tail = FALSE)
```

```
## [1] 3.851317e-74
```

Question 2: Is the distribution of theses uniform in time? (Inference)

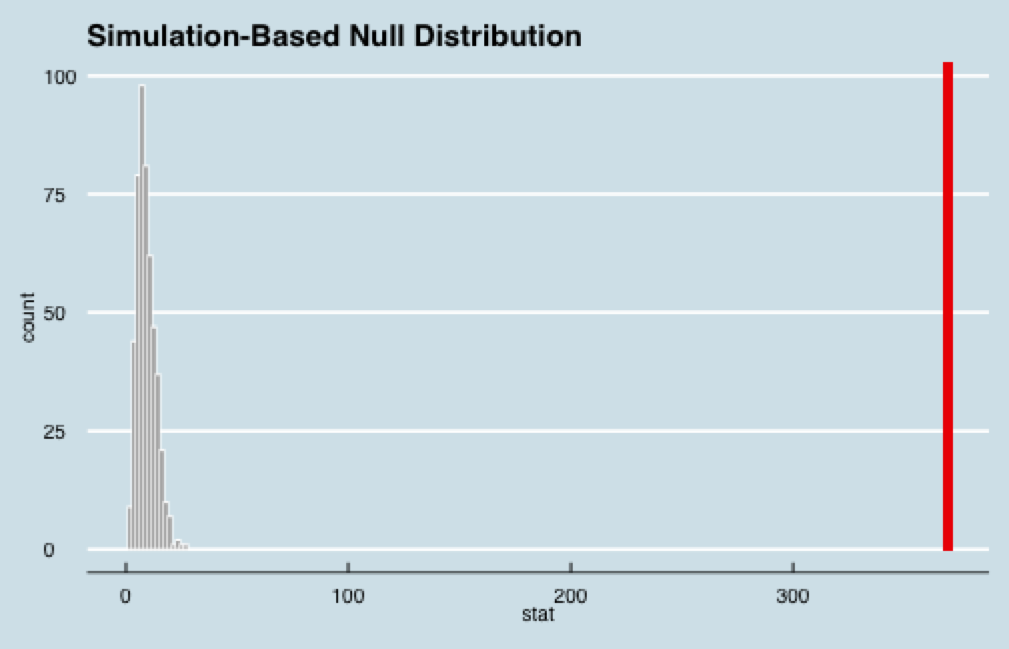
Chi-squared goodness of fit test (computational method)

```
levels(bd$year_binned) <- 1:10
null <- bd %>%
  specify(response = year_binned) %>%
  hypothesize(null = "point", p = c("1" = .1,
                                    "2" = .1,
                                    "3" = .1,
                                    "4" = .1,
                                    "5" = .1,
                                    "6" = .1,
                                    "7" = .1,
                                    "8" = .1,
                                    "9" = .1,
                                    "10" = .1)) %>%
  generate(500, type = "simulate") %>%
  calculate(stat = "Chisq")
```

Question 2: Is the distribution of theses uniform in time? (Inference)

Chi-squared goodness of fit test (computational method)

```
library(ggthemes)
visualize(null) +
  shade_p_value(obs_stat = chisq_obs, direction = "right") +
  theme_economist()
```



Question 3: What is the relationship between the age of a thesis and the number of checkouts? (Description)

```
head(theses)
```

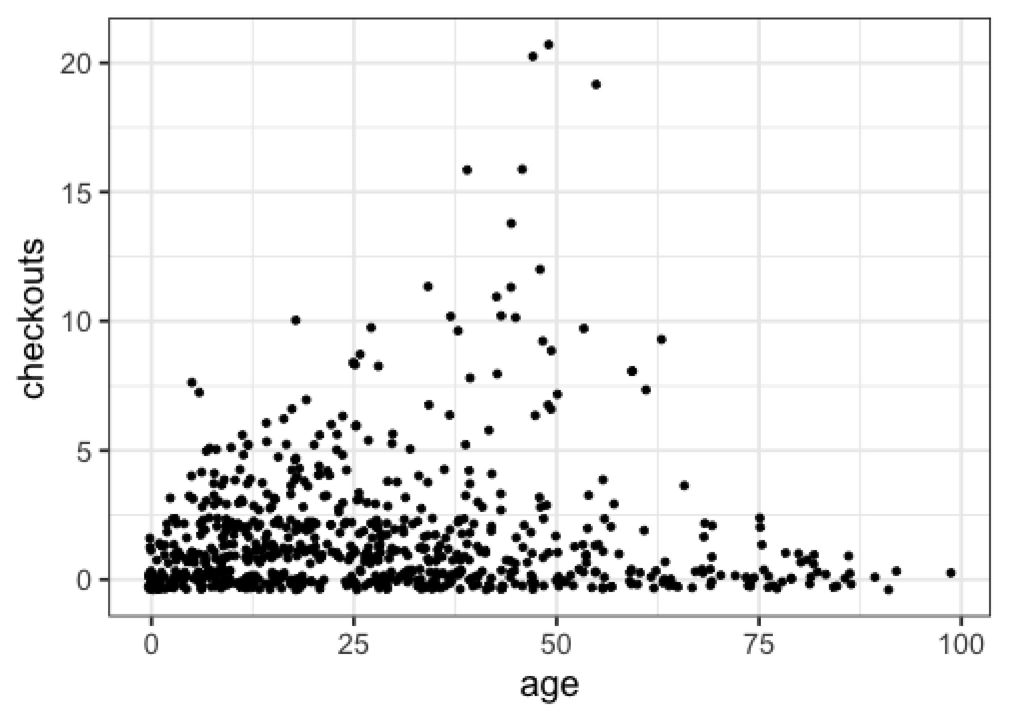
```
##   year checkouts division
## 1 1984         3      HSS
## 2 2004         2       LL
## 3 1939         1     PRPL
## 4 1992         1     PRPL
## 5 1941         0      HSS
## 6 1989         0      HSS
```

```
theses <- theses %>%
  mutate(age = 2017 - year)
head(theses)
```

```
##   year checkouts division age
## 1 1984         3      HSS  33
## 2 2004         2       LL  13
## 3 1939         1     PRPL  78
## 4 1992         1     PRPL  25
## 5 1941         0      HSS  76
## 6 1989         0      HSS  28
```

Question 3: What is the relationship between the age of a thesis and the number of checkouts? (Description)

```r
ggplot(theses, aes(x = age, y = checkouts)) +
  geom_jitter() +
  theme_bw(base_size = 18)
```



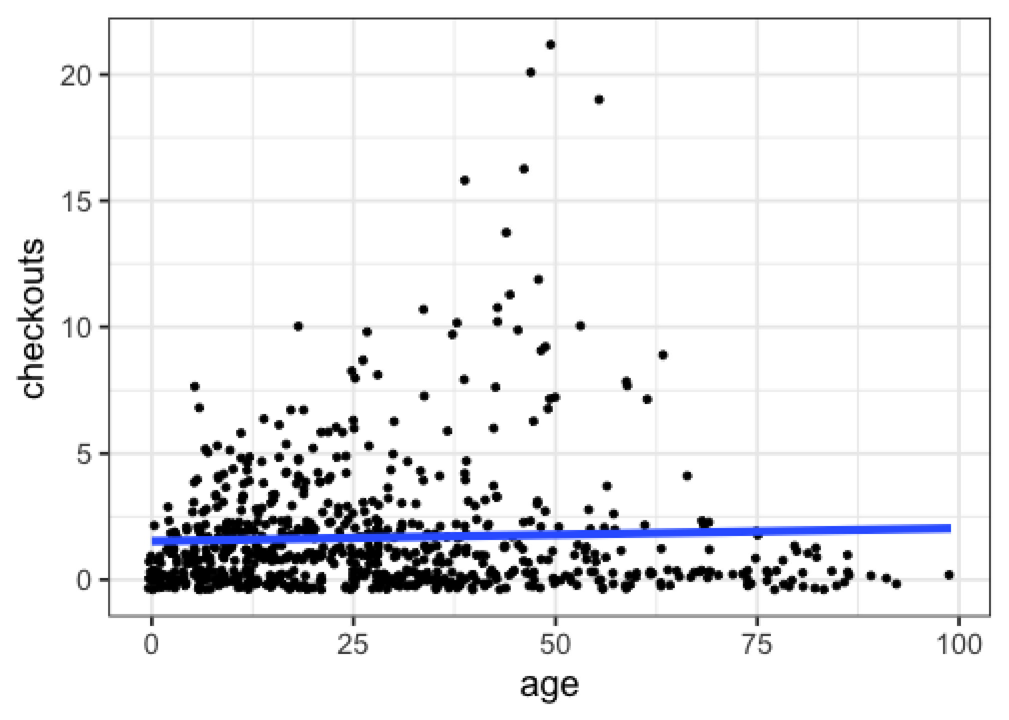
Question 3: What is the relationship between the age of a thesis and the number of checkouts? (Inference)

```
m1 <- lm(checkouts ~ age, data = theses)
summary(m1)
```

```
##
## Call:
## lm(formula = checkouts ~ age, data = theses)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0341 -1.6141 -0.6653  0.4064 19.2220
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.527038   0.158518   9.633   <2e-16 ***
## age         0.005122   0.004567   1.122    0.262
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.583 on 722 degrees of freedom
## Multiple R-squared:  0.001739,	Adjusted R-squared:  0.0003566
## F-statistic: 1.258 on 1 and 722 DF,  p-value: 0.2624
```

Question 3: What is the relationship between the age of a thesis and the number of checkouts? (Inference)

```
ggplot(theses, aes(x = age, y = checkouts)) +
  geom_jitter() +
  theme_bw(base_size = 18) +
  stat_smooth(method = "lm", se = FALSE, lwd = 2)
```



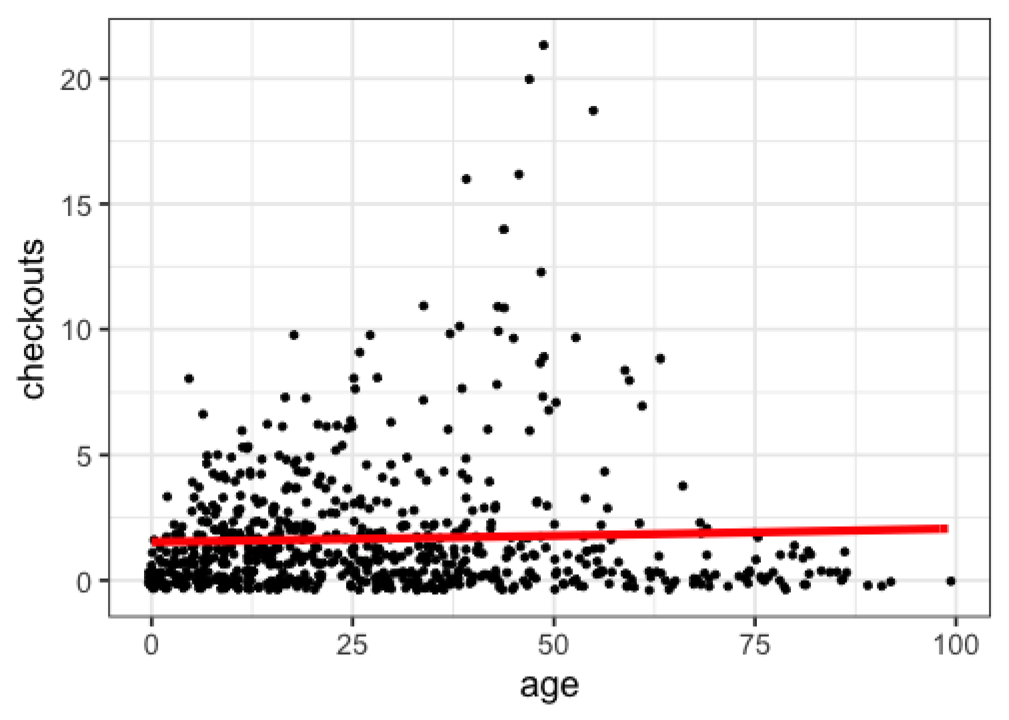
boardwork

Poisson Regression

```r
m2 <- glm(checkouts ~ age, data = theses, family = "poisson")
summary(m2)
```

```
## 
## Call:
## glm(formula = checkouts ~ age, family = "poisson", data = theses)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.0300  -1.7962  -0.5550   0.3091   8.0823  
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept) 0.427501   0.048115   8.885   <2e-16 ***
## age         0.002984   0.001331   2.242   0.0249 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for poisson family taken to be 1)
## 
##     Null deviance: 2117.9  on 723  degrees of freedom
## Residual deviance: 2113.0  on 722  degrees of freedom
## AIC: 3251.8
## 
```

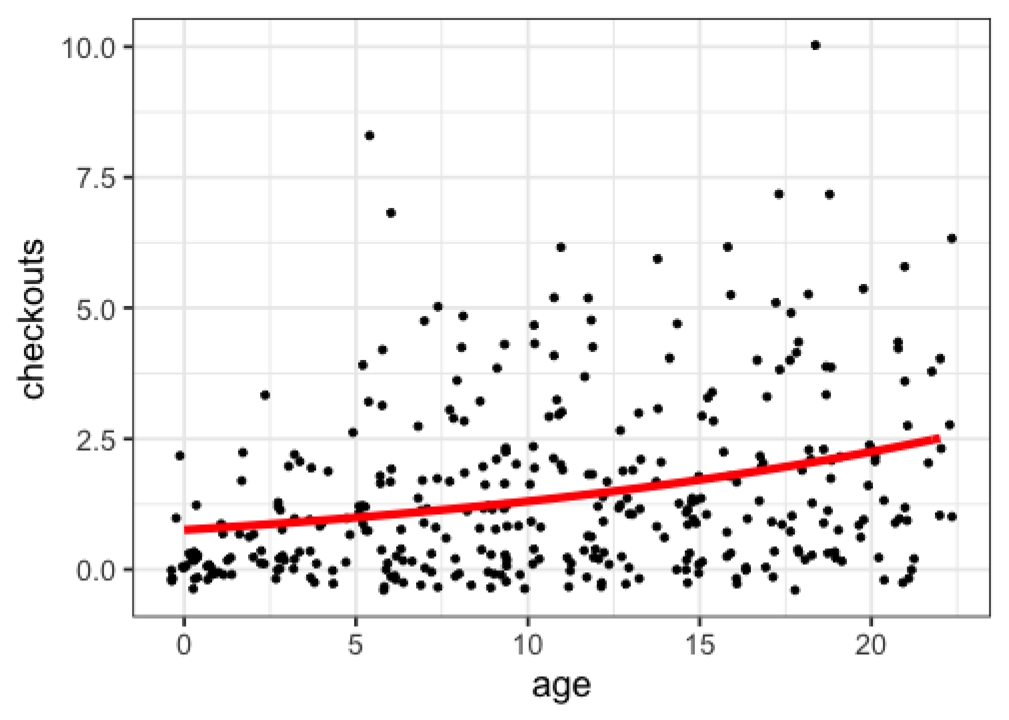
```
ggplot(theses, aes(x = age, y = checkouts)) +
  geom_jitter() +
  theme_bw(base_size = 18) +
  stat_function(fun = function(age) {exp(coef(m2)[1] +
                                         coef(m2)[2] * age)},
                color = "red", lwd = 2)
```



```r
t2 <- theses %>%
  filter(year > 1994)
m2 <- glm(checkouts ~ age, data = t2, family = "poisson")
summary(m2)
```

```
## 
## Call:
## glm(formula = checkouts ~ age, family = "poisson", data = t2)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.1820  -1.3725  -0.4017   0.6289   4.3983  
## 
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -0.277988   0.102643  -2.708  0.00676 ** 
## age          0.054541   0.007247   7.526 5.24e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for poisson family taken to be 1)
## 
##     Null deviance: 689.91  on 350  degrees of freedom
## Residual deviance: 631.55  on 349  degrees of freedom
## AIC: 1195.1
## 
## Number of Fisher Scoring iterations: 5
```

```
ggplot(t2, aes(x = age, y = checkouts)) +
  geom_jitter() +
  theme_bw(base_size = 18) +
  stat_function(fun = function(age) {exp(coef(m2)[1] +
                                          coef(m2)[2] * age)},
                color = "red", lwd = 2)
```


checkouts
age