Linear Regression and Generalized Regression - An R tutorial

Jieying Jiao

12/1/2019

Linear regression

For n observations in total, responses are y_i and corresponding measurements (covariates) are $\mathbf{x_i}$, with dimension p. We want to find the underlining relationship between the covariates and response.

Independent data

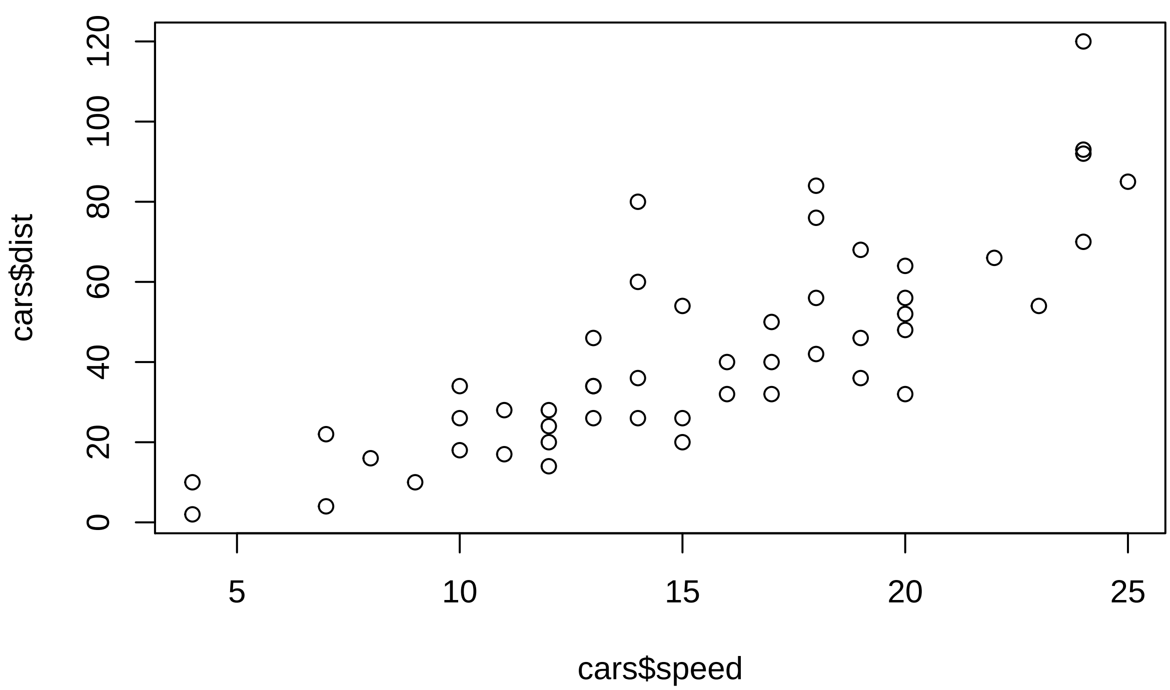
Model setup

$$
\begin{aligned}
y_i &= \beta_0 + \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i \\
\varepsilon_i &\overset{iid}{\sim} N(0, \sigma_\varepsilon^2)
\end{aligned}
\tag{1}
$$

For categorical covariates, use dummy variables. A categorial covariate with a different levels can be transformed into $a - 1$ dummy variables.

Model Fitting

```r
plot(cars$speed, cars$dist)
```



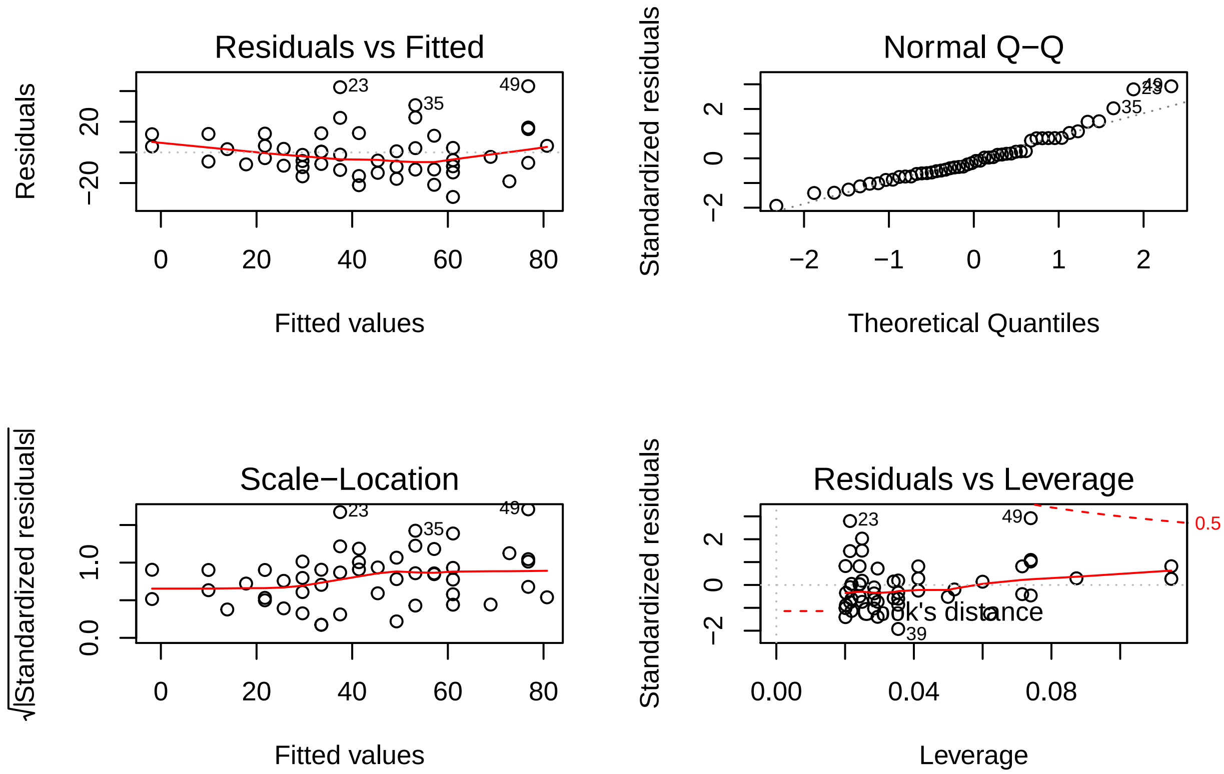
```r
mod1 <- lm(dist ~ speed, data=cars)
summary(mod1)
```

```
##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.069  -9.525  -2.272   9.215  43.201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.5791     6.7584  -2.601   0.0123 *
## speed         3.9324     0.4155   9.464 1.49e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.38 on 48 degrees of freedom
## Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
## F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

$$
\begin{aligned}
\hat{y}_i &= \hat{\beta}_0 + \boldsymbol{x}^\top \hat{\boldsymbol{\beta}} \\
r_i &= \hat{\varepsilon}_i = y_i - \hat{y}_i
\end{aligned} \tag{2}
$$

Model diagnostics

```r
par(mfrow= c(2, 2))
plot(mod1)
```



- Checking for linear trend: if there is any other trend missing here;
- Checking for normal assumption;
- Checking for equal variance (heteroscedasticity problem);
- Checking for influential observations.

Panel (longitudinal) data – Linear mixed effects model

Data involves repeated observations over time on different individuals. Such data are clustered, and observations on same individuals should be correlated. So the independent data model is not suitable. Let $y_{i,j}$ being the j_{th} observation on i_{th} subject.

$$\begin{aligned} y_{ij} &= \beta_0 + \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + \tau_i + \varepsilon_{ij} \\ \tau_i &\overset{iid}{\sim} N(0, \sigma_\tau^2) \\ \varepsilon_{ij} &\overset{iid}{\sim} N(0, \sigma_\varepsilon^2) \end{aligned} \tag{3}$$

Under this model setup, we have:

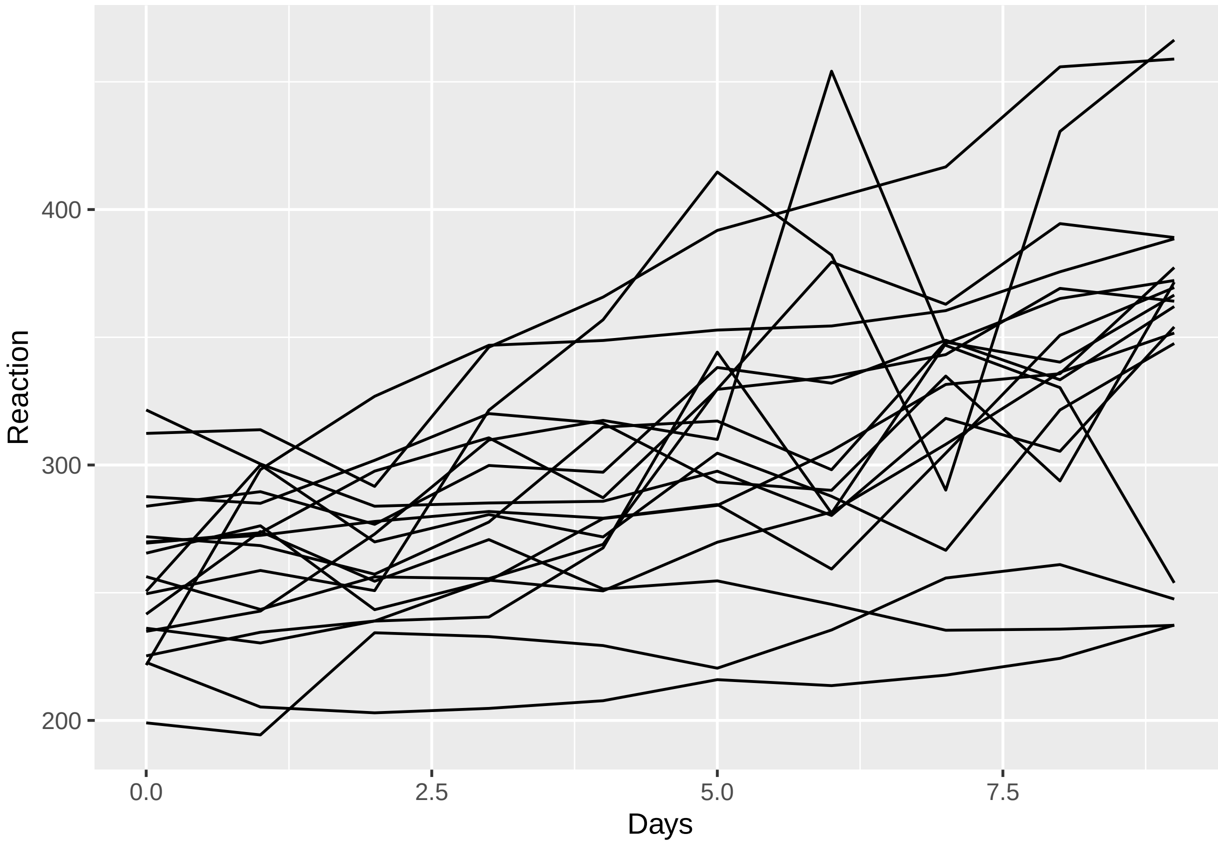
$$\begin{aligned} Cov(y_{ij}, y_{i'j'}) &= 0 \\ Cov(y_{ij}, y_{ij'}) &= Var(\tau_i) = \sigma_\tau^2 \\ Var(y_{ij}) &= \sigma_\varepsilon^2 + \sigma_\tau^2 \end{aligned} \tag{4}$$

Model Visualization

```r
# install.packages("lme4")
# install.packages("ggplot2")
library(lme4)
```

```
## Loading required package: Matrix
```

```r
library(ggplot2)
ggplot(aes(x = Days, y = Reaction), data = sleepstudy) + geom_line(aes(group = Subject))
```



Model Fitting

```r
mod2 <- lmer(Reaction ~ Days + ( 1 | Subject), data = sleepstudy)
summary(mod2)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Reaction ~ Days + (1 | Subject)
##    Data: sleepstudy
##
## REML criterion at convergence: 1786.5
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.2257 -0.5529  0.0109  0.5188  4.2506
##
```

```
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Subject  (Intercept) 1378.2   37.12
##  Residual              960.5   30.99
## Number of obs: 180, groups:  Subject, 18
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept) 251.4051     9.7467   25.79
## Days         10.4673     0.8042   13.02
##
## Correlation of Fixed Effects:
##      (Intr)
## Days -0.371
```

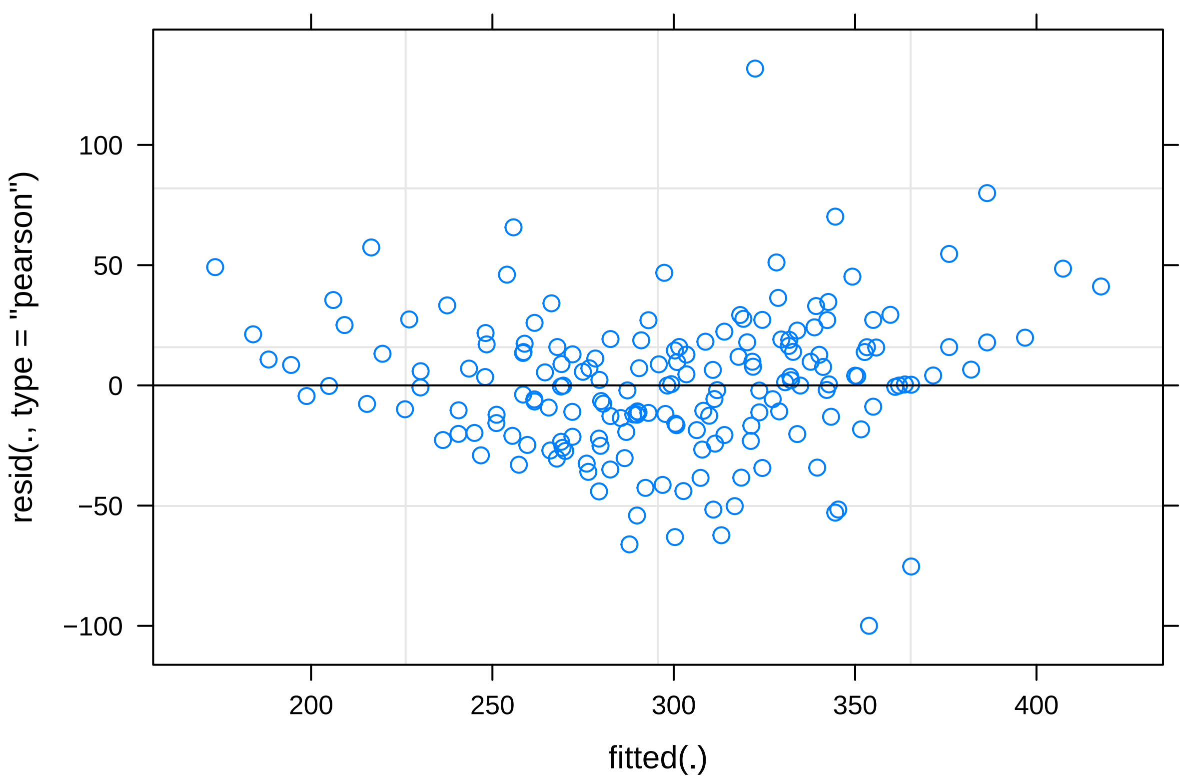
Model diagnostics

```r
library(car)
```

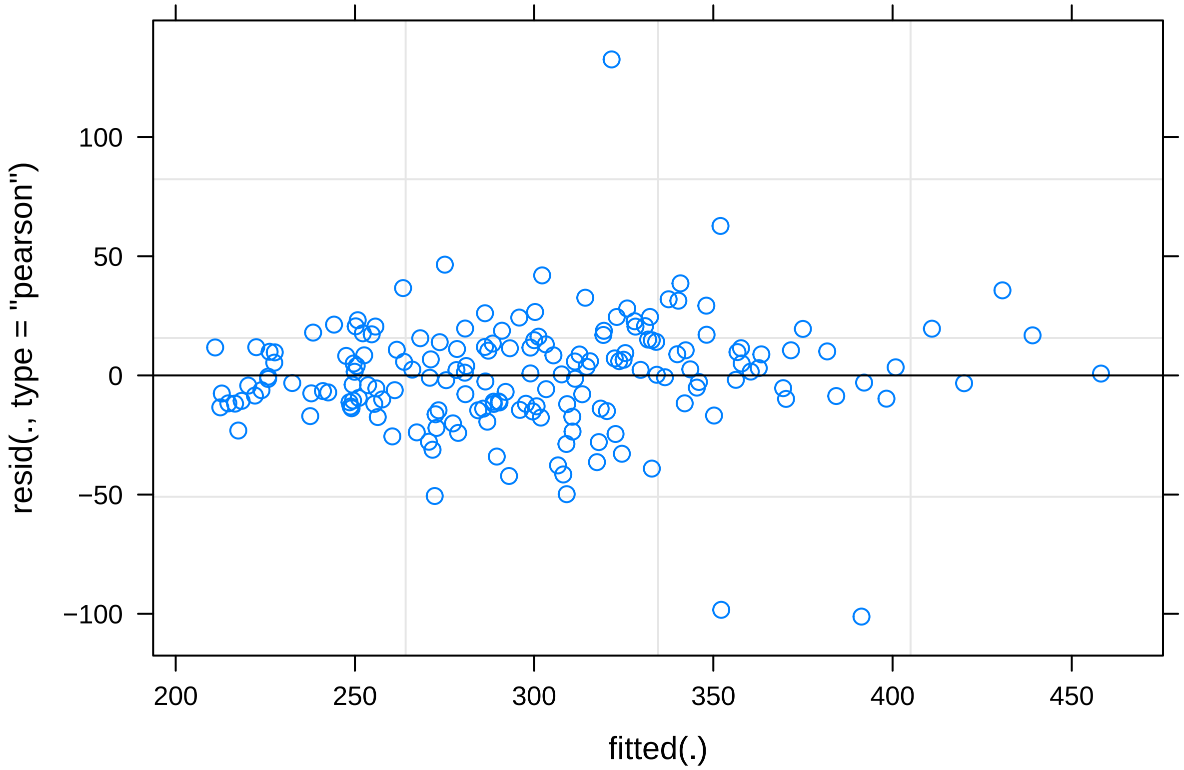
```
## Loading required package: carData
```

```
## Registered S3 methods overwritten by 'car':
##   method                          from
##   influence.merMod                lme4
##   cooks.distance.influence.merMod lme4
##   dfbeta.influence.merMod         lme4
##   dfbetas.influence.merMod        lme4
```

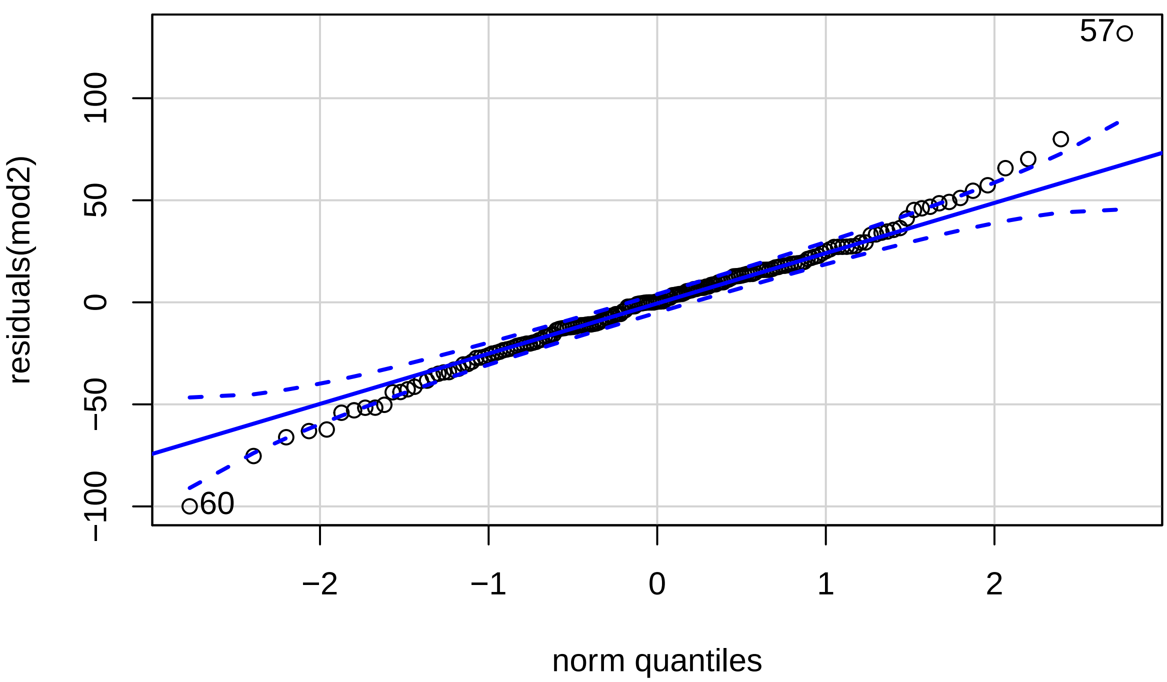
```r
plot(mod2)
```



```r
mod3 <- lmer(Reaction ~ Days + (Days | Subject), data = sleepstudy)
plot(mod3)
```

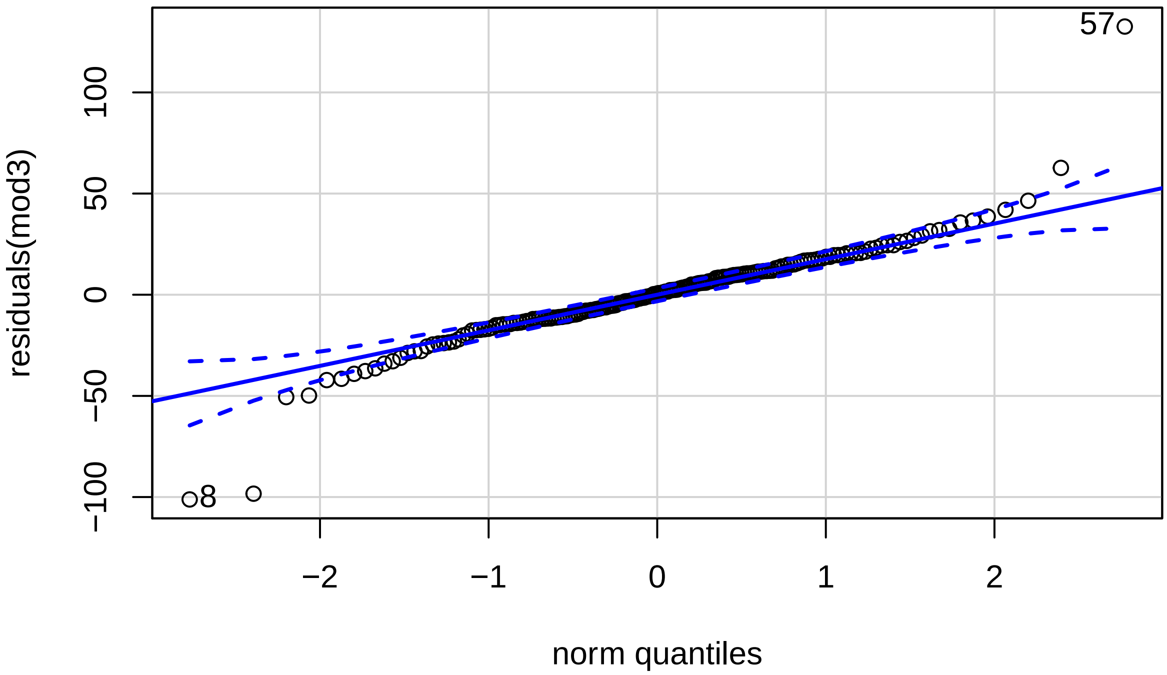



```r
qqPlot(residuals(mod2))
```




```
## [1] 57 60
```

```
qqPlot(residuals(mod3))
```




```
## [1] 57  8
```

Generalized linear regression

Binary data - Logistic regression

When response y_i is a binary variable, which only takes value $\{0, 1\}$, we usualy use Binary distribution to model it: $y_i \sim \text{Bernoulli}(p_i)$. Parameter p is the success probability, which takes value in the interval $[0, 1]$. We want to see how the potential covariates influence the success probability:

$$
\begin{aligned}
y_i &\overset{\text{inde}}{\sim} \text{Bernoulli}(p_i) \\
\text{logit}(p_i) &= \log \frac{p_i}{1 - p_i} = \eta_i = \beta_0 + \boldsymbol{x}_i^\top \boldsymbol{\beta}
\end{aligned} \tag{5}
$$

```
library(ISLR)
mod4 <- glm(Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 + Volume, data = Smarket,
            family = binomial)
summary(mod4)
```

```
##
## Call:
## glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
##     Volume, family = binomial, data = Smarket)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.446  -1.203   1.065   1.145   1.326
##
```

```
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.126000   0.240736  -0.523    0.601
## Lag1        -0.073074   0.050167  -1.457    0.145
## Lag2        -0.042301   0.050086  -0.845    0.398
## Lag3         0.011085   0.049939   0.222    0.824
## Lag4         0.009359   0.049974   0.187    0.851
## Lag5         0.010313   0.049511   0.208    0.835
## Volume       0.135441   0.158360   0.855    0.392
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1731.2  on 1249  degrees of freedom
## Residual deviance: 1727.6  on 1243  degrees of freedom
## AIC: 1741.6
##
## Number of Fisher Scoring iterations: 3
```

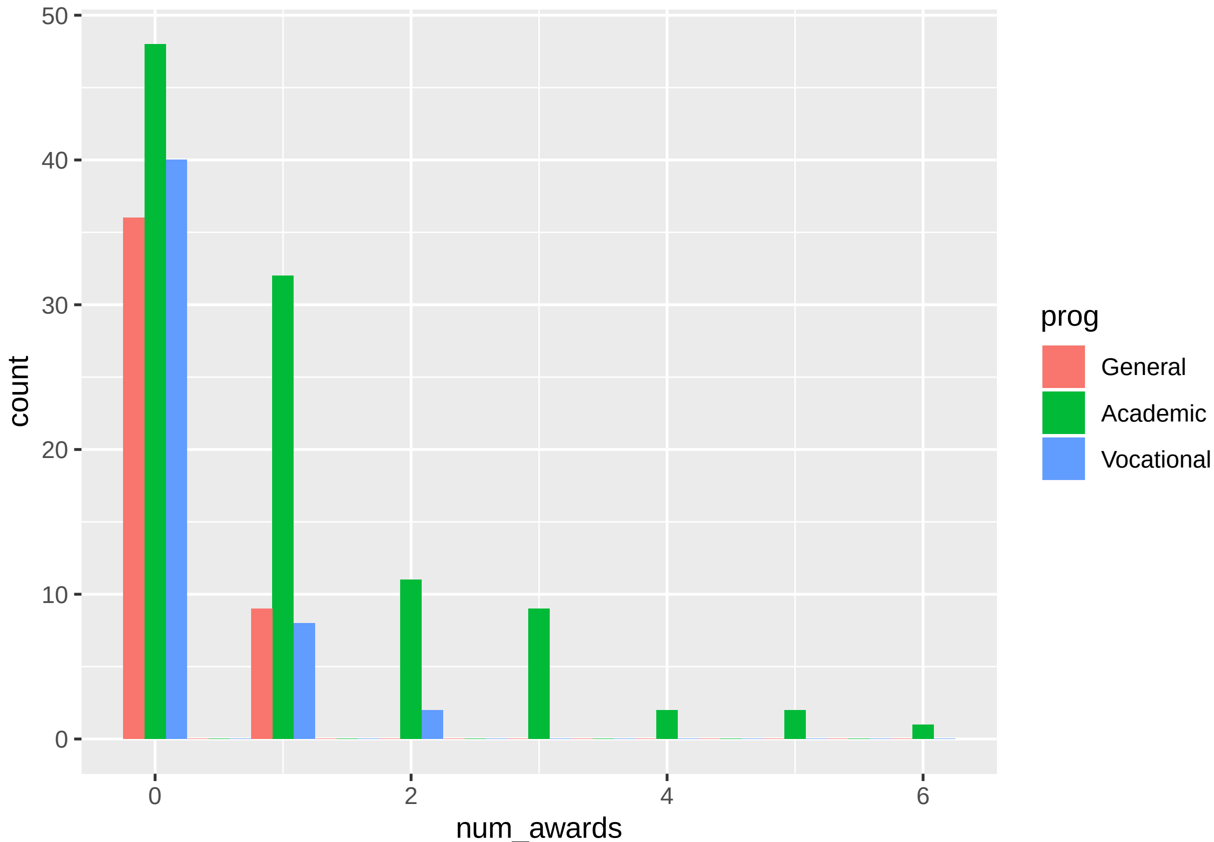
In order to do prediction, we need a threshold k (usually 0.5), such that when $\hat{p}_i > k$, $\hat{y}_i = 1$.

Count data - Poisson regression

When we have count data, Poisson distribution is usually assumed: $y_i \sim \text{Poisson}(\lambda_i)$:

$$\begin{aligned} y_i &\overset{\text{inde}}{\sim} \text{Poisson}(\lambda_i) \\ \log \lambda_i &= \eta_i = \beta_0 + \boldsymbol{x}_i^\top \boldsymbol{\beta} \end{aligned} \tag{6}$$

```r
p <- read.csv("https://stats.idre.ucla.edu/stat/data/poisson_sim.csv")
p <- within(p, {
  prog <- factor(prog, levels=1:3, labels=c("General", "Academic",
                                            "Vocational"))
  id <- factor(id)
})
ggplot(p, aes(num_awards, fill = prog)) +
  geom_histogram(binwidth=.5, position="dodge")
```




```r
mod5 <- glm(num_awards ~ prog + math, family="poisson", data=p)
summary(mod5)
```

```
##
## Call:
## glm(formula = num_awards ~ prog + math, family = "poisson", data = p)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2043  -0.8436  -0.5106   0.2558   2.6796
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -5.24712    0.65845  -7.969 1.60e-15 ***
## progAcademic    1.08386    0.35825   3.025  0.00248 **
## progVocational  0.36981    0.44107   0.838  0.40179
## math            0.07015    0.01060   6.619 3.63e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 287.67  on 199  degrees of freedom
## Residual deviance: 189.45  on 196  degrees of freedom
## AIC: 373.5
##
## Number of Fisher Scoring iterations: 6
```