Logistic regression

Kwang-Yeol Park

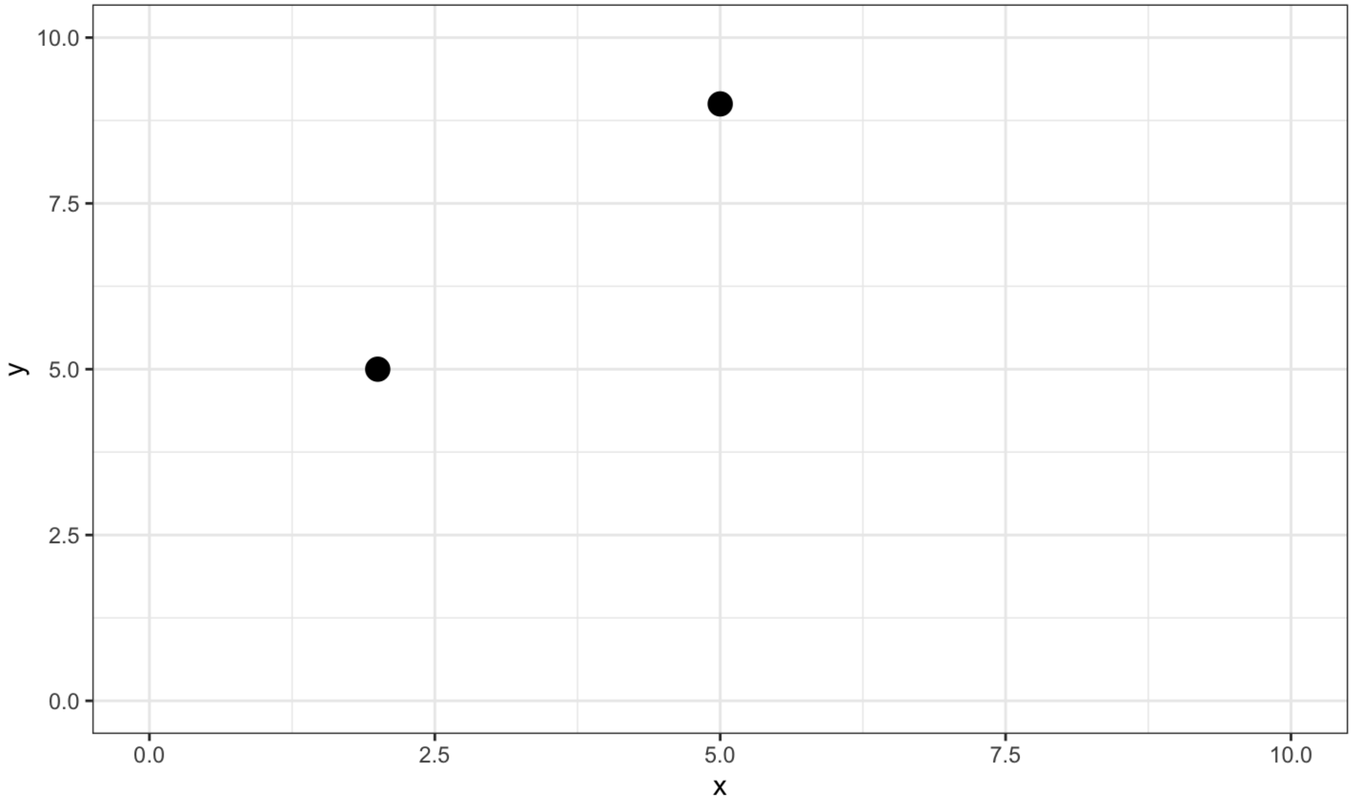
Objectives

- You can do or understand at least one of the following goals

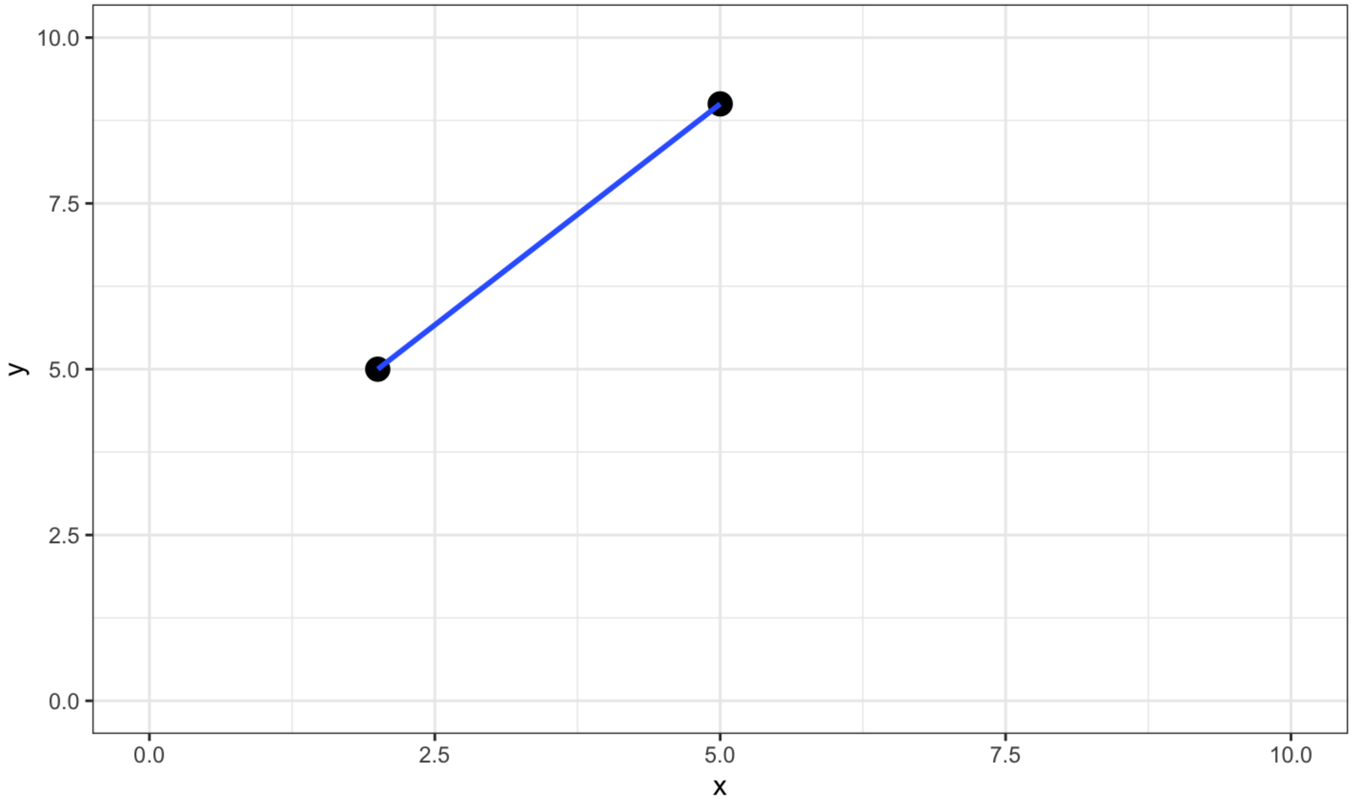
1. You can perform logistic regression test using R
2. You can interpret the result of logistic regression
3. You can understand R^2 in linear regression
4. You can tell correlation coefficient from regression coefficient

Linear regression

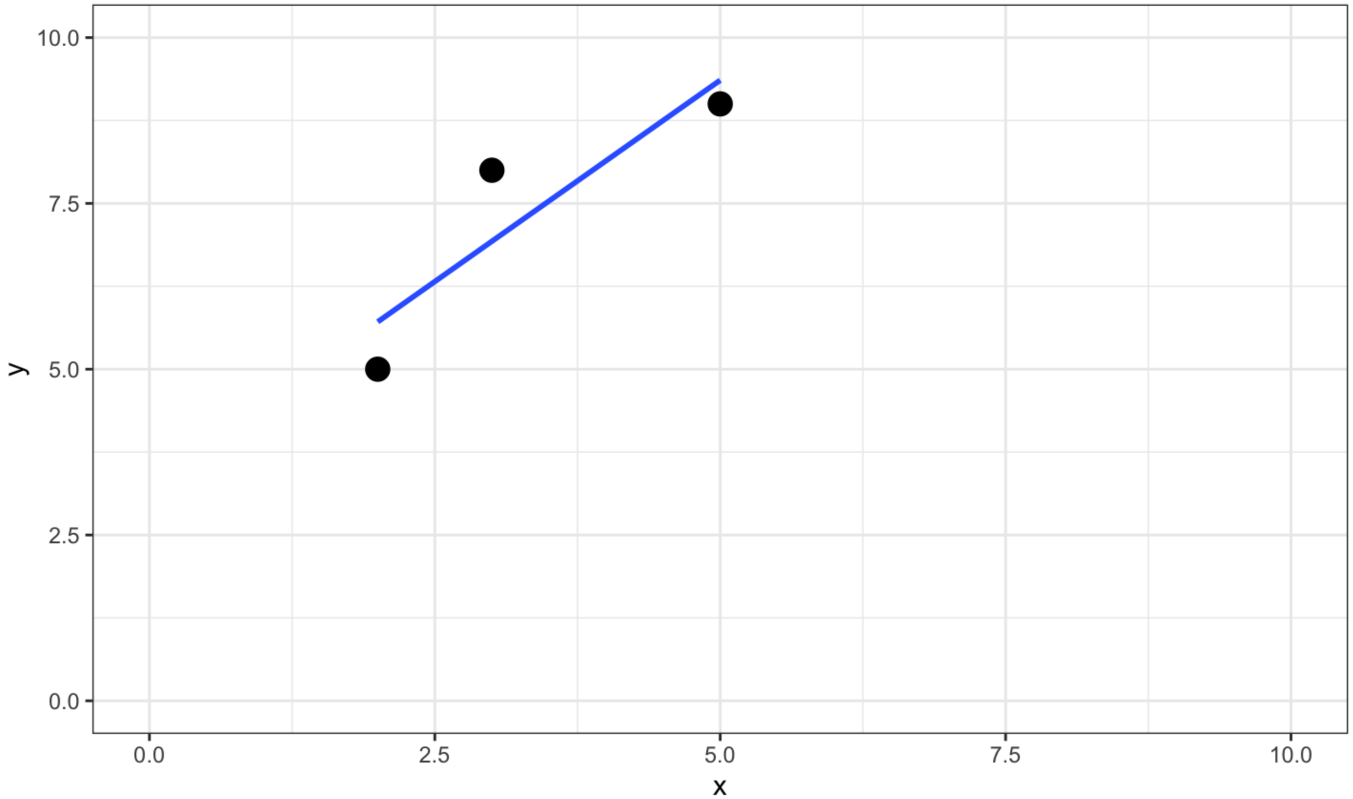
Motivation



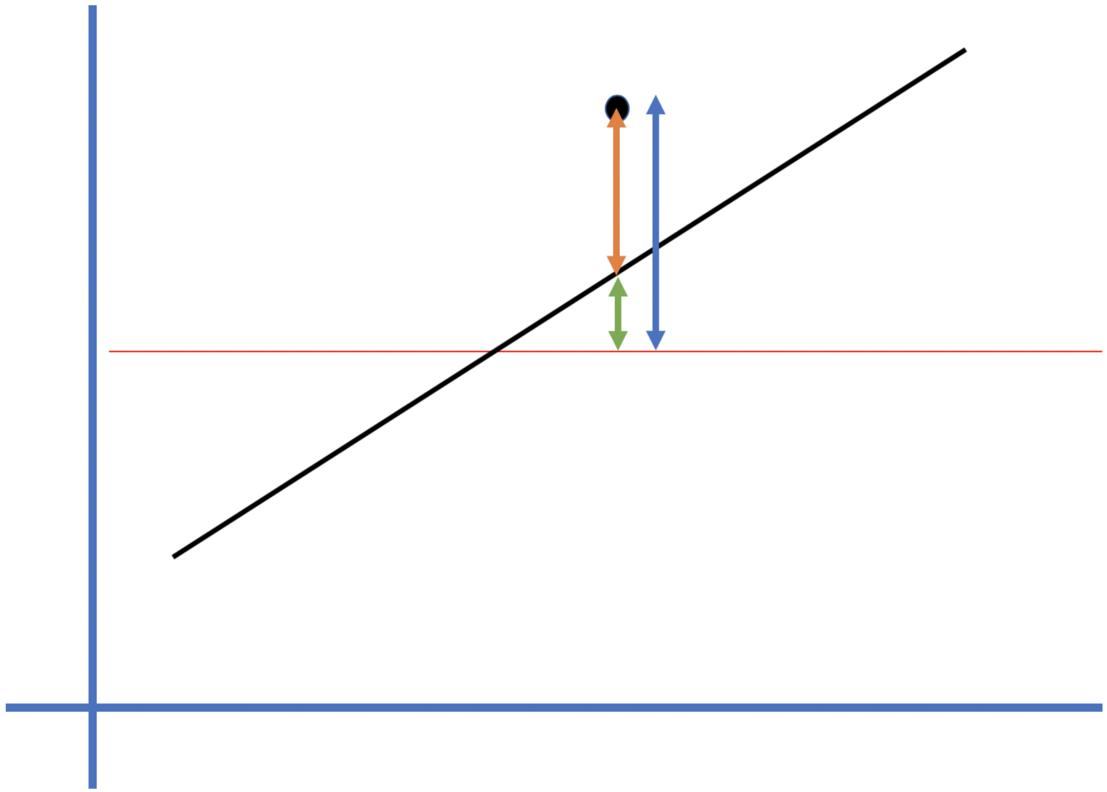
Motivation



Motivation



Linear regression: Ordinary least square method



Linear regression

- Dependent variable (continuous variable) can be explained by independent variables (continuous or categorical variables)

$$y_i = \beta_0 + \beta_1 * x_1 + \epsilon_i$$

$$\hat{\beta_0} = \overline{y} - \hat{\beta_1} * \overline{x}$$

$$\hat{\beta_1} = \frac{\sum((x_i - \overline{x}) * (y_i - \overline{y})}{\sum((x_i - \overline{x})^2)}$$

- Assumptions
  - $\epsilon_i \sim N(0, \sigma^2)$
  - Y: normal distribution
  - Y: equal variance around the mean

Sample size (rule of thumb)

- Multiple linear regression
  - 20 subjects per independent variable
- Multiple logistic regression
  - 10 outcomes per independent variable
  - Number of outcome is either $Y = 1$ or $Y = 0$ which is smaller

How to perform linear regression

1. Draw plot
2. Perform regression test
3. Check the model fit
4. Check the βs
5. Check multicollinearity

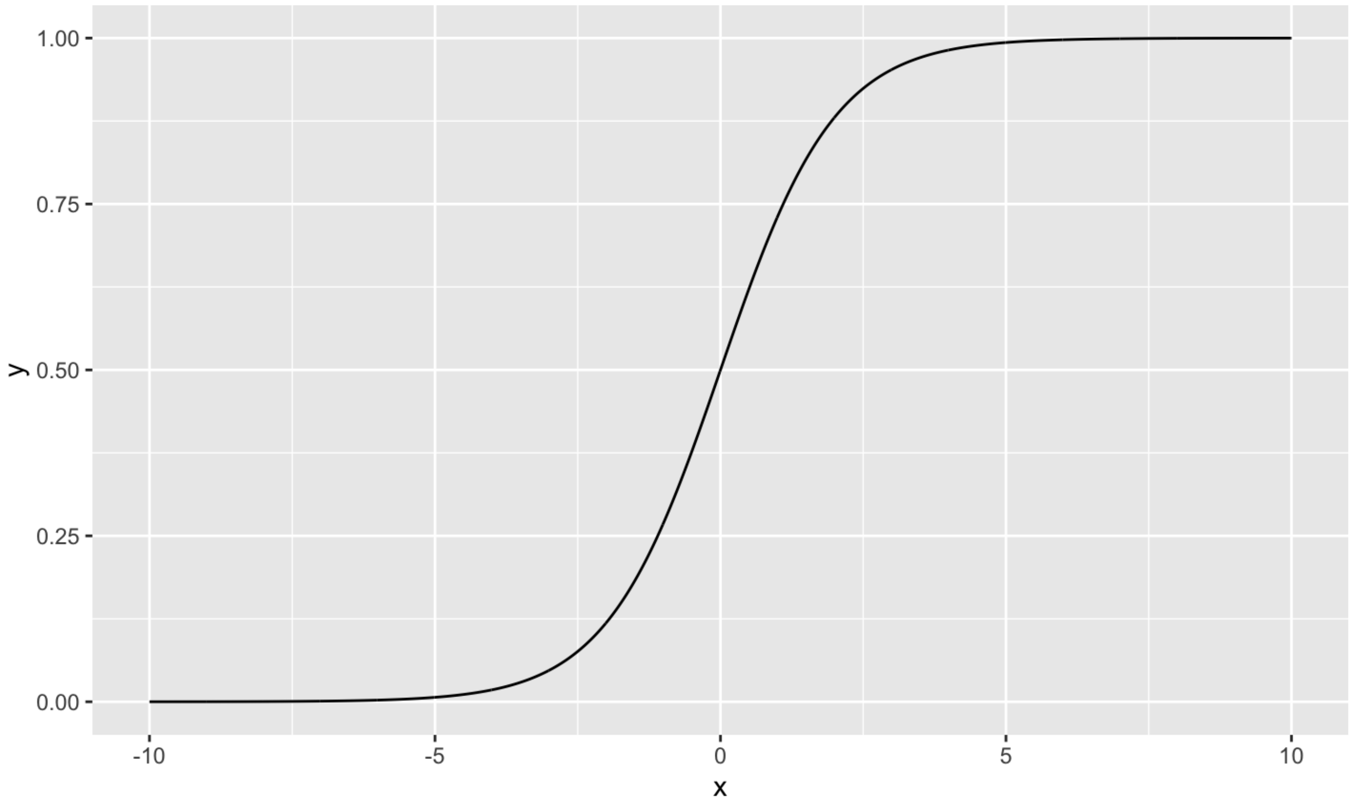
3. Model fit

- 결정계수: coefficient of determination: R^2
  - squared correlation coefficient between y_i and $\hat{y}$
  - how much variance of y can be explained by model.
- F test
  - H_0: all βs are 0

Logistic regression

Sigmoid function

$$S(t) = \frac{1}{1 + exp(-t)}$$



Logistic function to logistic regression

$$f(x) = \frac{L}{1 + exp(-k(x - x0))}$$

- general form of sigmoid function

$$f(x) = \frac{exp(\beta_0 + \beta_1 * x_1)}{1 + exp(\beta_0 + \beta_1 * x_1)}$$

$$ln(Odds) = ln(\frac{P(Y = 1)}{1 - P(Y = 1)}) = \beta_0 + \beta_1 * x_1$$

- ln(odds) = logit

$$logit(p(x)) = ln(\frac{p}{1 - p}) = \beta_0 + \beta_1 * x_i + \epsilon_i$$

OR (odds ratio)

$$ln(Odds) = ln(\frac{P(Y=1)}{1-P(Y=1)}) = \beta_0 + \beta_1 * x_1$$

- if $x_1 = 1$

$$ln(Odds(x_1 = 1)) = \beta_0 + \beta_1$$

- if $x_1 = 0$

$$ln(Odds(x_1 = 1)) = \beta_0$$

$$ln(\frac{Odds(x_1 = 1)}{Odds(x_1 = 0)}) = ln(Odds(x_1 = 1)) - ln(Odds(x_1 = 0)) = \beta_1$$

$$OR = exp(\beta_1)$$

Logistic regression

- Dependent variable (categorical variable) can be explained by independent variables (continuous or categorical variables)
- Maximum likelihood estimation (cf> Ordinary least square method in linear regression)
- Hosmer and Lemeshow Goodness of fit test
- Nagelkerke R^2 : SPSS
- Cox-Snell R^2 : SAS and SPSS
    - https://statisticalhorizons.com/r2logistic

Sample size (rule of thumb)

- Multiple linear regression
  - 20 subjects per independent variable
- Multiple logistic regression
  - 10 outcomes per independent variable
  - Number of outcome is either $Y = 1$ or $Y = 0$ which is smaller

How to perform logistic regression

1. Check the variables and the number of event
2. Bivariate analyses
3. Perform regression test
4. Check the model fit
5. Check the βs
6. Check multicollinearity

Import DB

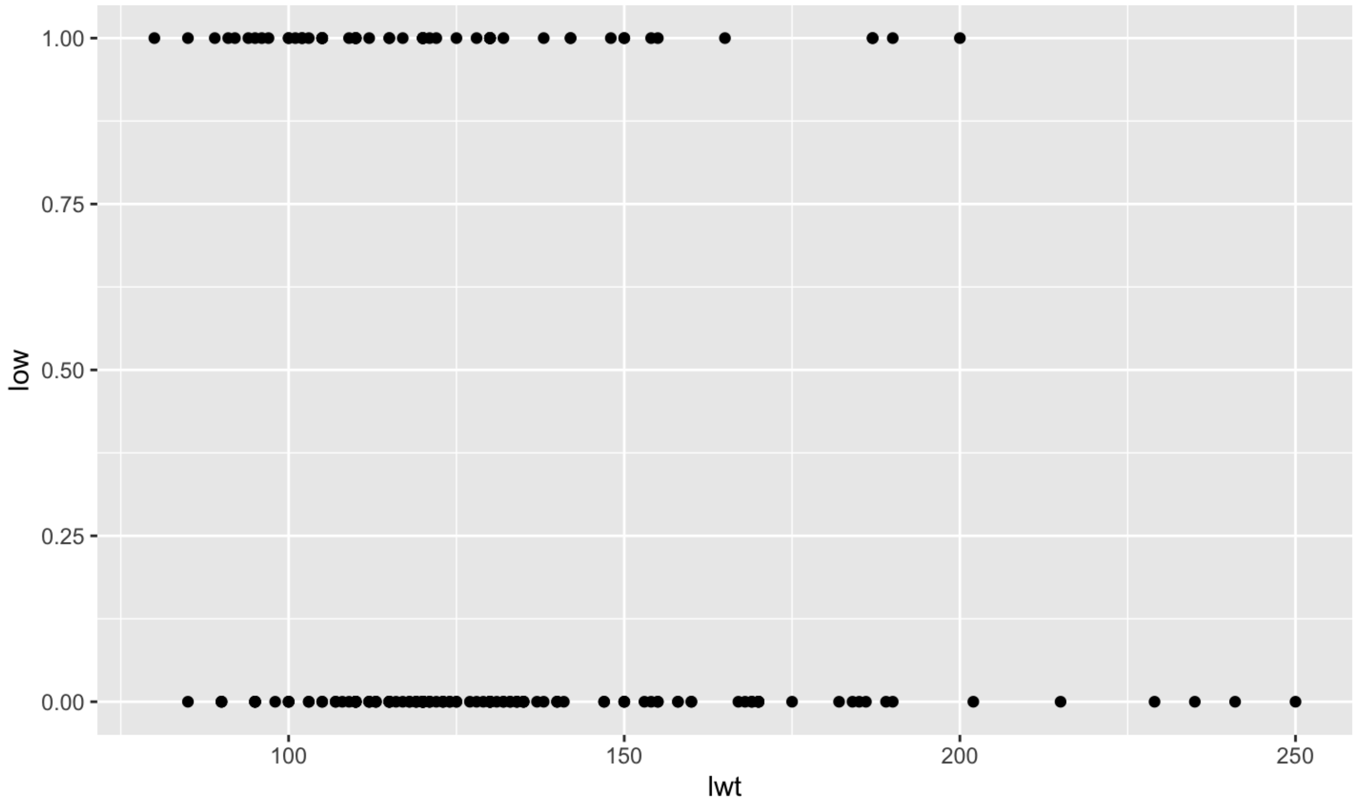
```
library(MASS)
data("birthwt")
```

1. Check the variable

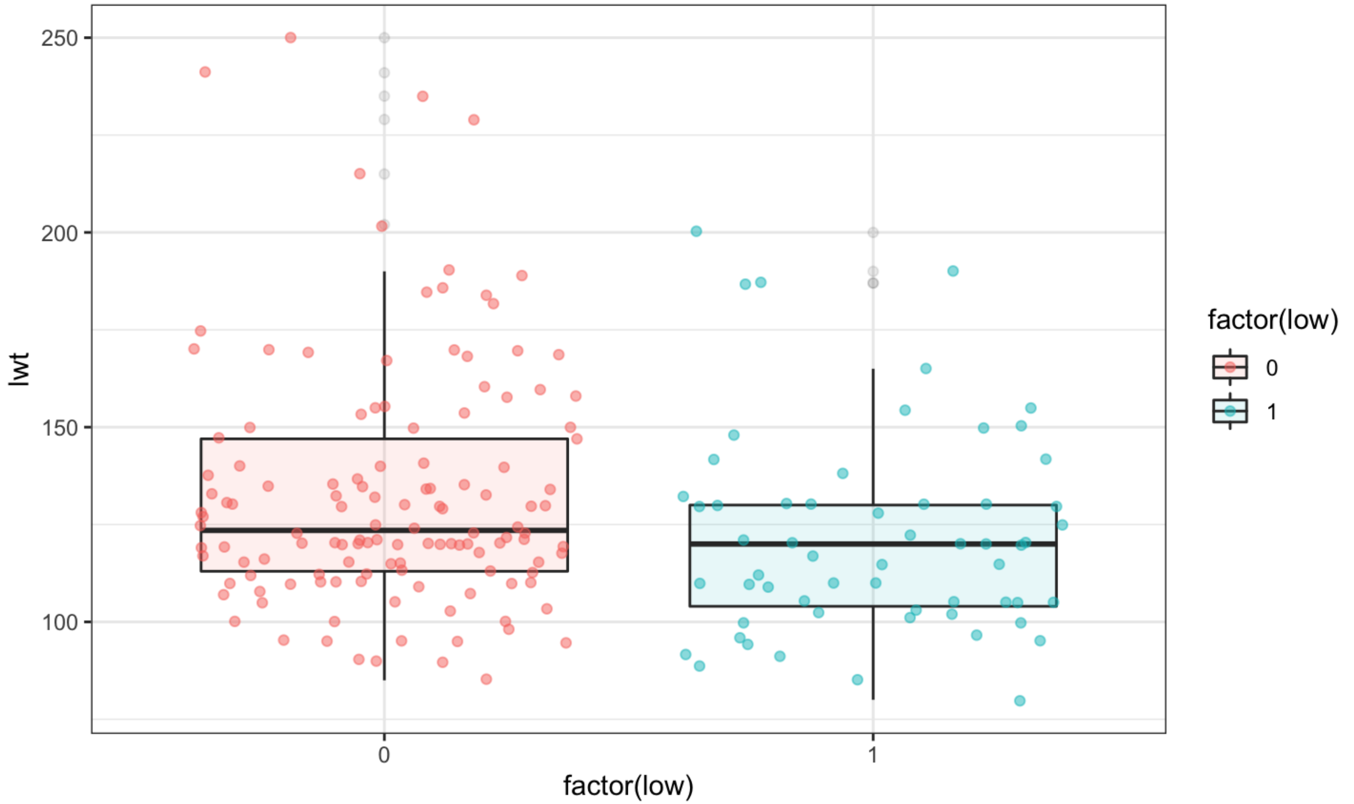
```
table(birthwt$low)
```

```
##
##   0   1
## 130  59
```

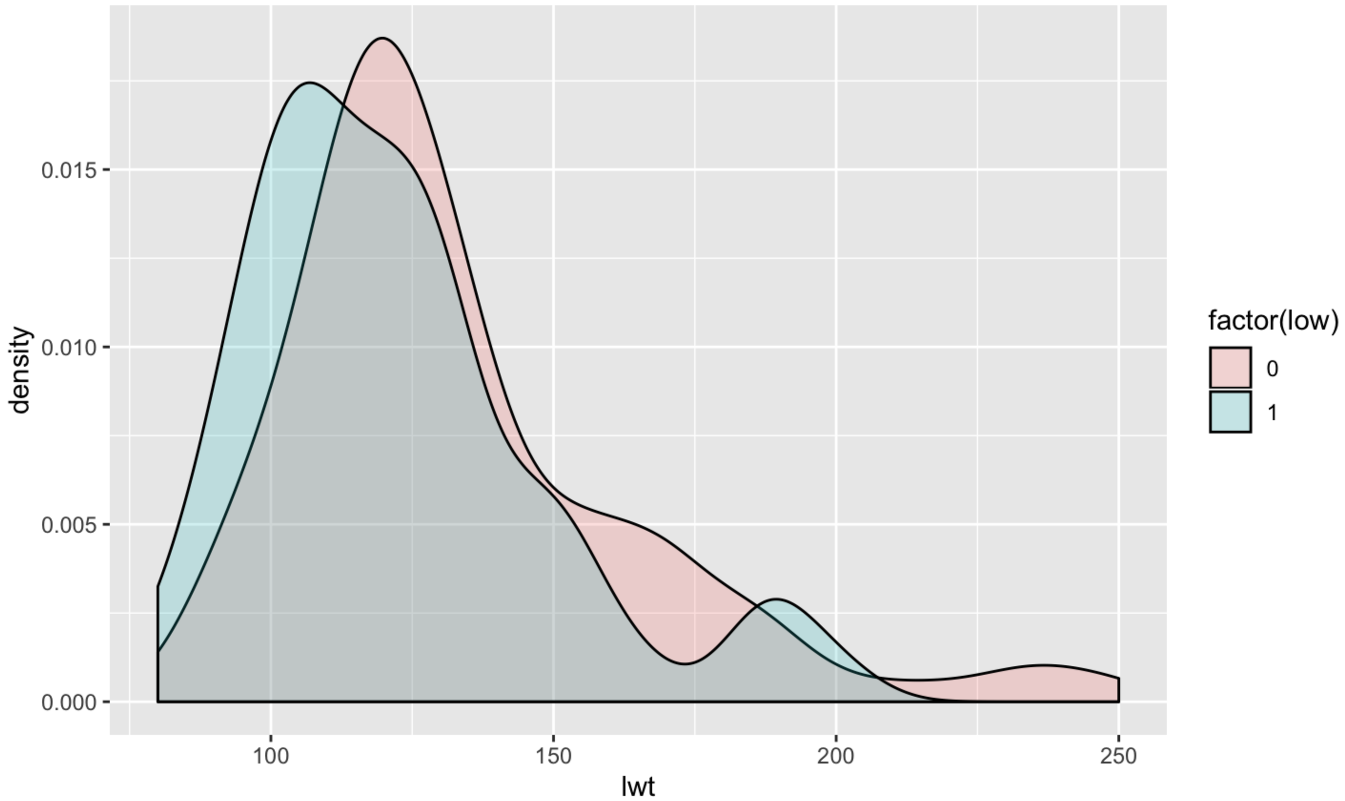
```
ggplot(birthwt, aes(x = lwt, y = low)) + geom_point()
```



```
ggplot(birthwt, aes(x = factor(low), y = lwt, fill = factor(low))) +
  geom_boxplot(alpha = 0.1) +
  geom_jitter(aes(color = factor(low)), alpha = 0.5) +
  theme_bw()
```



```r
ggplot(birthwt, aes(x =  lwt, fill = factor(low))) +
  geom_density(alpha = 0.2)
```



2. Bivariate analysis

- Dependent variable: categorical
- if independent variable: categorical
    - χ^2 test
    - Fisher’s exact test
- if independent variabls: continuous
    - Student’s t test
    - Mann-Whitney or Wilcoxon rank sum test

3. Perform logistic regression

- birthwt shuuld not have missing values.

```
result <- glm(low ~ lwt + age + ht + smoke, data = birthwt,
    family = binomial)
```

4. Model fit

- Hosmer & Lemeshow goodness of fit

```
library(ResourceSelection)
```

```
## ResourceSelection 0.3-5   2019-07-22
```

```
hoslem.test(result$y, fitted(result))
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  result$y, fitted(result)
## X-squared = 4.0767, df = 8, p-value = 0.8501
```

- Nagelkerke R^2

```
library(fmsb)
NagelkerkeR2(result)
```

```
## $N
## [1] 189
##
## $R2
## [1] 0.1344158
```

5. β or $exp(\beta)$

```
summary(result)
```

```
## 
## Call:
## glm(formula = low ~ lwt + age + ht + smoke, family = binomial, 
##     data = birthwt)
## 
## Deviance Residuals: 
##    Min      1Q  Median      3Q     Max  
## -1.692  -0.842  -0.657   1.188   2.219  
## 
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)   
## (Intercept)  1.766856   1.052266   1.679  0.09313 . 
## lwt         -0.016955   0.006623  -2.560  0.01047 * 
## age         -0.035687   0.033365  -1.070  0.28480   
## ht           1.788156   0.685882   2.607  0.00913 **
## smoke        0.679020   0.331939   2.046  0.04079 * 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

formula

$$f(x) = \frac{exp(\beta_0 + \beta_1 * x_1)}{1 + exp(\beta_0 + \beta_1 * x_1)}$$

$$f(x) = \frac{exp(-0.02 * lwt + 1.79 * ht + 0.68 * smoke)}{1 + exp(-0.02 * lwt + 1.79 * ht + 0.68 * smoke)}$$

Odds ratio

$$ln(Odds) = ln(\frac{P(Y=1)}{1-P(Y=1)}) = \beta_0 + \beta_1 * x_1$$

$$ln(\frac{Odds(ht=1)}{Odds(ht=0)}) = ln(Odds(ht=1)) - ln(Odds(ht=1)) = 1.79$$

$$OR = exp(1.79) = 5.978418$$

result of logistic regression

```
coef(result)
```

```
## (Intercept)         lwt         age          ht       smoke 
##  1.76685576 -0.01695511 -0.03568724  1.78815623  0.67902040
```

```
exp(coef(result))
```

```
## (Intercept)         lwt         age          ht       smoke 
##   5.8524230   0.9831878   0.9649420   5.9784194   1.9719451
```

```
confint(result)
```

```
## Waiting for profiling to be done...
```

```
##                   2.5 %       97.5 %
## (Intercept) -0.24303022  3.899499577
## lwt         -0.03078896 -0.004668962
## age         -0.10299314  0.028367348
## ht           0.48215833  3.228623435
## smoke        0.02958821  1.334845690
```

```
confint(result) %>% exp()
```

```
## Waiting for profiling to be done...
```

```
##                 2.5 %     97.5 %
## (Intercept) 0.7842478 49.3777332
## lwt         0.9696802  0.9953419
## age         0.9021332  1.0287735
## ht          1.6195662 25.2448818
## smoke       1.0300303  3.7994096
```

6. Multicollinearity

```
library(car)
vif(result)
```

```
##      lwt      age       ht    smoke
## 1.158363 1.019751 1.137391 1.000308
```

- vif > 2.5 problematic
- vif > 10 serious

```r
library(performance)
check_collinearity(result)
```

```
## # Check for Multicollinearity
##
## Low Correlation
##
##  Parameter  VIF Increased SE
##        lwt 1.16         1.08
##        age 1.02         1.01
##         ht 1.14         1.07
##      smoke 1.00         1.00
```

```
x <- check_collinearity(result)
plot(x)
```