STA255 Week-12

Shahriar Shams

30/03/2020

Review of week-10

- Relationship among quantitative variables
- Pearson correlation coefficient
- Least square regression
- Regression under Normal distribution
  - Parameter estimation
  - Interpretation of regression parameters
  - Properties of estimators of regression parameters
  - Confidence interval/t-test for the slope
  - Sum of squares decomposition

Learning goals

- Regression: Quantitative Y, Categorical X
- Some useful R codes

Regression: Quantitative Y, Categorical X

- Assume we have response variable Y which is quantitative.
- And we have predictor X which is categorical
- We want to check whether X and Y are related or not.
- Let's assume a simple case where X only has two categories. (For example, male and female)
- We can create what's known as **dummy variables**.
- Let, $X_m = 1,$ if Male and $X_m = 0,$ if Female

- A hypothetical data will look like this:

Y	Sex (X)	X_m
10	Male	1
12	Male	1
8	Female	0
9	Female	0
...	...	...
...	...	...

- X_m is the numerical representation of the categorical variable Sex.
- Recall the Normality assumption from previous week.
  - we can write, $Y|X \sim N(\beta_1 + \beta_2 X_m,\ \sigma^2)$
  - Therefore, $E[Y|X] = \beta_1 + \beta_2 X_m$
  - $E[Y|X = Female] = \beta_1$
  - $E[Y|X = Male] = \beta_1 + \beta_2$

- Subtracting one from the other, we get

$$\beta_2 = E[Y|X = Male] - E[Y|X = Female]$$

- If we want to test whether the two group averages are equal or not that's same as testing $H_0 : \beta_2 = 0$
- Likelihood contribution of each of the y_i's with $X = 0$ will be

$$(2\pi\sigma^2)^{-1/2} exp[-\frac{1}{2\sigma^2}(y_i - \beta_1)^2]$$

- Likelihood contribution of each of the y_i's with $X = 1$ will be

$$(2\pi\sigma^2)^{-1/2} exp[-\frac{1}{2\sigma^2}(y_i - \beta_1 - \beta_2)^2]$$

- We can maximize the entire likelihood and calculate the estimates of β_1 and β_2
  - In this example(try it on your own),

$$\hat{\beta}_1 = \bar{y}_{\text{ for the female group}}$$

$$\hat{\beta}_2 = \bar{y}_{\text{ for the male group}} - \bar{y}_{\text{ for the female group}}$$

A numerical example

Suppose we have measurements from two groups: Group1 and Group2.

```
# Let's enter the data in R
G1=c(79.98,80.04,80.02,80.04,80.03,80.03,80.04,79.97,
     80.05,80.03,80.02,80.00,80.02)
G2=c(80.02,79.94,79.98,79.97,79.97,80.03,79.95,79.97)

mean(G1)
```

```
## [1] 80.02077
```

```
mean(G2)
```

```
## [1] 79.97875
```

```
# Converting into what is known as "long" format data
Y=c(G1,G2)
# Creating the dummy variable
X_G1=c(rep(1,length(G1)),rep(0,length(G2)))

# Let's see how the data looks
cbind(Y,X_G1)
```

```
##           Y X_G1
##  [1,] 79.98    1
##  [2,] 80.04    1
##  [3,] 80.02    1
##  [4,] 80.04    1
##  [5,] 80.03    1
##  [6,] 80.03    1
##  [7,] 80.04    1
##  [8,] 79.97    1
##  [9,] 80.05    1
## [10,] 80.03    1
## [11,] 80.02    1
## [12,] 80.00    1
## [13,] 80.02    1
## [14,] 80.02    0
## [15,] 79.94    0
## [16,] 79.98    0
## [17,] 79.97    0
## [18,] 79.97    0
## [19,] 80.03    0
## [20,] 79.95    0
## [21,] 79.97    0
```

```r
# Fitting a linear model
m=lm(Y~X_G1)
summary(m)
```

```
##
## Call:
## lm(formula = Y ~ X_G1)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.050769 -0.008750 -0.000769  0.019231  0.051250
##
## Coefficients:
##             Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 79.978750   0.009521 8399.914  < 2e-16 ***
## X_G1         0.042019   0.012101    3.472  0.00255 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02693 on 19 degrees of freedom
## Multiple R-squared:  0.3882, Adjusted R-squared:  0.356
## F-statistic: 12.06 on 1 and 19 DF,  p-value: 0.002551
```

```r
# Confidence intervals of regression parameters
confint(m,level = 0.95)
```

```
##                   2.5 %      97.5 %
## (Intercept) 79.95882153 79.99867847
## X_G1         0.01669058  0.06734788
```

Let's do a t-test that we learned in Week-8 (slides 14 and 15)

```r
t.test(G1,G2, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  G1 and G2
## t = 3.4722, df = 19, p-value = 0.002551
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.01669058 0.06734788
## sample estimates:
## mean of x mean of y
##  80.02077  79.97875
```

- Will we be able to use this if our X variable has more than two categories?
  - Absolutely!
  - We will just need more dummy variables.
  - For example, if we have a variable called *program of study* which has three categories (Undergrad, Masters and PhD) we will need two dummy variables.

Program (X)	X_u	X_m
Undergrad	1	0
Masters	0	1
PhD	0	0

- But then we have multiple linear regression which you will learn in STA302, STA303 or similar courses.

Some useful R codes

Singe sample t-test/confidence interval

Taken from lecture notes (week-8(page-6), week-9(page-4)) (4.7, 5.5, 4.4, 3.3, 4.6, 5.3, 5.2, 4.8, 5.7, 5.3) $\overset{iid}{\sim} N(\mu, \sigma^2)$ with both μ and σ^2 unknown

Calculate the 0.95-confidence interval for μ.

Test $H_0 : \mu = 5$ vs $H_a : \mu \neq 5$ at level of significance, $\alpha = 0.05$

```
x=c(4.7, 5.5, 4.4, 3.3, 4.6, 5.3, 5.2, 4.8, 5.7, 5.3)

# Alternative, mu not equal 5
t.test(x, mu=5, conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  x
## t = -0.54545, df = 9, p-value = 0.5987
## alternative hypothesis: true mean is not equal to 5
## 95 percent confidence interval:
##  4.382325 5.377675
## sample estimates:
## mean of x
##      4.88
```

```r
# Alternative, mu<5
t.test(x, mu=5, conf.level = 0.95, alternative="less")
```

```
##
##  One Sample t-test
##
## data:  x
## t = -0.54545, df = 9, p-value = 0.2993
## alternative hypothesis: true mean is less than 5
## 95 percent confidence interval:
##      -Inf 5.283285
## sample estimates:
## mean of x
##      4.88
```

```r
# Alternative, mu>5
t.test(x, mu=5, conf.level = 0.95, alternative="greater")
```

```
##
##  One Sample t-test
##
## data:  x
## t = -0.54545, df = 9, p-value = 0.7007
## alternative hypothesis: true mean is greater than 5
## 95 percent confidence interval:
##  4.476715      Inf
## sample estimates:
## mean of x
##      4.88
```

Singe proportion Z-test/confidence interval

Taken from Week-10 (day-1) page 2 [proportion of COVID-19 cases]

```r
prop.test(37, 1008, p=0.03) # try alternative="less" or "greater"
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  37 out of 1008, null probability 0.03
## X-squared = 1.336, df = 1, p-value = 0.2477
## alternative hypothesis: true p is not equal to 0.03
## 95 percent confidence interval:
##  0.02632657 0.05075226
## sample estimates:
```

```
##          p
## 0.03670635
```

Paired t-test

Taken from Week-10 (day-2) page-6 [before after measurements]

```
x=c(10.19, 7.92, 6.67, 12.22, 8.21, 8.26, 13.06, 8.20, 9.83, 5.94)
y=c(7.00, 7.53, 6.45, 1.31, 5.42, 2.81, 6.60, 0.55, 3.13, 5.00)

t.test(x, y, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  x and y
## t = 3.9873, df = 9, p-value = 0.003171
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.933984 7.006016
## sample estimates:
## mean of the differences
##                    4.47
```

Homework

For Regression with categorical X

Take the data given in Exercise 27 (page 507) and do a pooled t-test(assuming population variances are equal). Now fit a model with one dummy variable and compared your numbers from the regression to the numbers from t-test. (If you have time do it by hand or code it yourself without using the functions t.test() or lm())