STA 4320 CHAP 3.2.2

He Jiang

Sep 2024

Sec 3.2.2

Advertising dataset

```
fpath = getwd()
Advertising = read.csv(paste0(fpath, "/Advertising.csv"))
y = Advertising$sales
n = nrow(Advertising)
```

R multiple regression by specifying the data frame

```
res = summary( lm(sales ~ TV + radio + newspaper, data = Advertising) )
res
```

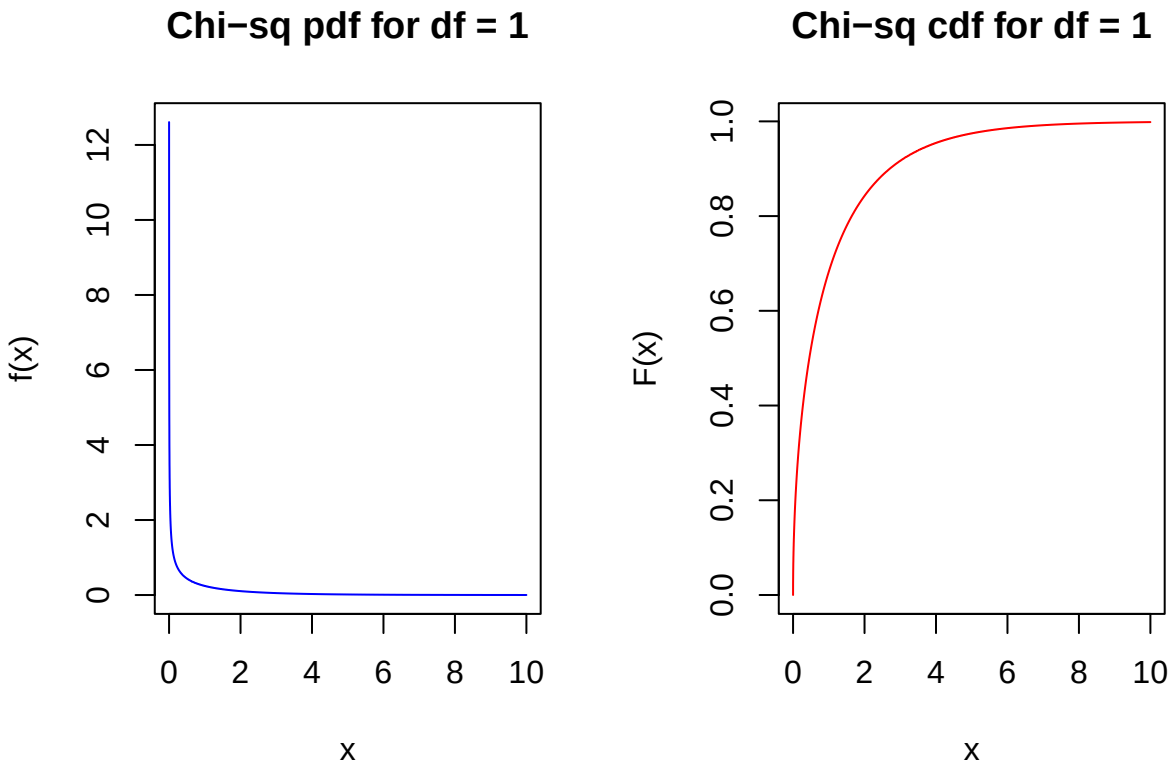
```
##
## Call:
## lm(formula = sales ~ TV + radio + newspaper, data = Advertising)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8277 -0.8908  0.2418  1.1893  2.8292
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.938889   0.311908   9.422   <2e-16 ***
## TV           0.045765   0.001395  32.809   <2e-16 ***
## radio        0.188530   0.008611  21.893   <2e-16 ***
## newspaper   -0.001037   0.005871  -0.177     0.86
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.686 on 196 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8956
## F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

Chisq Distribution

See the pdf and cdf

```
# degree of freedom
df = 1

# a sequence of values to evaluate the chisq density
x = seq(0, 10, length.out = 10001)
```

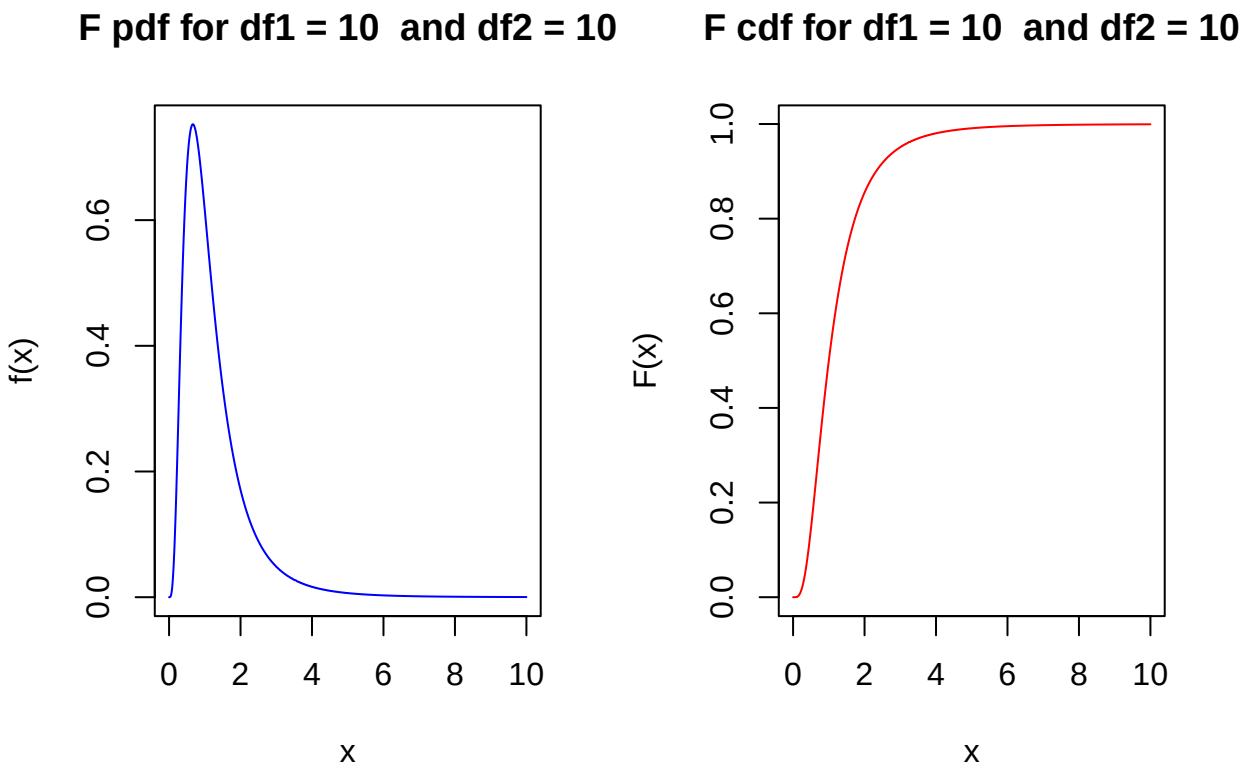
```r
# plot two pictures together, in one row and two columns
par(mfrow = c(1,2))

# plot the pdf
plot(x, dchisq(x, df),
     main = paste("Chi-sq pdf for df =", df),
     ylab = "f(x)",
     pch = 16,
     cex = 0.01)

lines(x, dchisq(x, df),
      col = "blue")

# plot the cdf
plot(x, pchisq(x, df),
     main = paste("Chi-sq cdf for df =", df),
     ylab = "F(x)",
     pch = 16,
     cex = 0.01)

lines(x, pchisq(x, df),
      col = "red")
```



F Distribution

See the pdf and cdf

```r
# degree of freedom
df_1 = 10
df_2 = 10

# a sequence of values to evaluate the chisq density
x = seq(0, 10, length.out = 10001)

# plot two pictures together, in one row and two columns
par(mfrow = c(1,2))

# plot the pdf
plot(x, df(x, df_1, df_2),
     main = paste("F pdf for df1 =", df_1, " and df2 =", df_2),
     ylab = "f(x)",
     pch = 16,
     cex = 0.01)

lines(x, df(x, df_1, df_2),
      col = "blue")

# plot the cdf
plot(x, pf(x, df_1, df_2),
     main = paste("F cdf for df1 =", df_1, " and df2 =", df_2),
     ylab = "F(x)",
     pch = 16,
     cex = 0.01)

lines(x, pf(x, df_1, df_2),
      col = "red")
```



Simple linear regression

Simple linear regression

F stat is t stat squared

```
summary( lm(Advertising$sales ~ Advertising$TV))
```

```
##
## Call:
## lm(formula = Advertising$sales ~ Advertising$TV)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3860 -1.9545 -0.1913  2.0671  7.2124
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    7.032594   0.457843   15.36   <2e-16 ***
## Advertising$TV 0.047537   0.002691   17.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.259 on 198 degrees of freedom
## Multiple R-squared:  0.6119, Adjusted R-squared:  0.6099
## F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

Computing the p value

```
1 - pf(312.1, 1, 198)
```

```
## [1] 0
```

Test for a subset of variables

Exercise 1 Full model

```
reg = lm(sales ~ TV + radio + newspaper, data = Advertising)
res = summary( reg )

y = Advertising$sales
RSS = sum( (y - predict(reg, data.frame(Advertising[,2:4])))^2 )
TSS = sum( (y - mean(y))^2 )

res
```

```
##
## Call:
## lm(formula = sales ~ TV + radio + newspaper, data = Advertising)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8277 -0.8908  0.2418  1.1893  2.8292
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.938889   0.311908   9.422   <2e-16 ***
```

```
## TV           0.045765   0.001395  32.809   <2e-16 ***
## radio        0.188530   0.008611  21.893   <2e-16 ***
## newspaper   -0.001037   0.005871  -0.177     0.86
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.686 on 196 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8956
## F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

Model without newspaper

```
reg_0 = lm(sales ~ TV + radio, data = Advertising)

y = Advertising$sales
RSS_0 = sum( (y - predict(reg_0, data.frame(Advertising[,2:3])))^2 )
n = nrow(Advertising)
p = 3
q = 1

# F_0 should be the corresponding t value squared
F_0 = ( (RSS_0 - RSS) / q ) / ( RSS / (n - p - 1) )

summary( reg_0 )
```

```
##
## Call:
## lm(formula = sales ~ TV + radio, data = Advertising)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.7977 -0.8752  0.2422  1.1708  2.8328
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.92110    0.29449   9.919   <2e-16 ***
## TV           0.04575    0.00139  32.909   <2e-16 ***
## radio        0.18799    0.00804  23.382   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.681 on 197 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8962
## F-statistic: 859.6 on 2 and 197 DF,  p-value: < 2.2e-16
```

Exercise 2 Model with only TV (and intercept)

```
reg_0 = lm(sales ~ TV, data = Advertising)

y = Advertising$sales
RSS_0 = sum( (y - predict(reg_0, data.frame(TV = Advertising[,2])))^2 )
n = nrow(Advertising)
p = 3
q = 2
```

```r
F_0 = ( (RSS_0 - RSS) / q ) / ( RSS / (n - p - 1) )
F_0
```

```
## [1] 272.0407
```

```r
1 - pf(F_0, q, n-p-1)
```

```
## [1] 0
```