STA 108 Discussion 4: Anova and Correlation Model

Yuanyuan Li

Reference: Textbook Chapter 2.7-2.11.

1. Obtain the Anova table

Analysis of Variance (ANOVA) is another perspective of Regression Analysis. It is based on the partitioning of sums of squares and degrees of freedom associated with the response variable Y.

$$
\begin{aligned}
Y_i - \bar{Y} &= Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y}, \\
\text{SSTO} &= \text{SSE} + \text{SSR}, \\
\text{Total Sum of Squares:} \quad \text{SSTO} &= \sum (Y_i - \bar{Y})^2, \ df = n-1, \\
\text{Error Sum of Squares:} \quad \text{SSE} &= \sum (Y_i - \hat{Y}_i)^2, \ df = n-2, \ \text{MSE} = \frac{SSE}{n-2} \\
\text{Regression Sum of Squares:} \quad \text{SSR} &= \sum (\hat{Y}_i - \bar{Y})^2, \ df = 1, \ \text{MSR} = \frac{SSR}{1}
\end{aligned}
$$

Let us use the built-in dataset `cars` as an example. The ANOVA table can be obtained from `anova` function. The first row shows the d.f. of SSR, SSR and MSR, the second rows shows the d.f. of SSE, SSE, and MSE. In order to print the anova table as a formal table, you can use the `kable` function from `knitr` package.

```
Y = cars$dist
X = cars$speed
n = length(X)
fit=lm(Y~X)
library(knitr)#install.packages("knitr") to install it
kable(anova(fit))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X	1	21185.46	21185.4589	89.56711	0
Residuals	48	11353.52	236.5317		

F-test of $\beta_1 = 0$ v.s. $\beta_1 \neq 0$

$$F = \frac{MSR}{MSE}$$

The distribution of F under the null hypothesis is $F_{1,n-2}$. Reject H_0 when $F > F(1-\alpha; 1, n-2)$.

We can use either the two-tailed t test or the one-tailed F test for testing $\beta_1 = 0$ versus $\beta_1 \neq 0$. The t test,

however, is more flexible since it can be used for one-sided alternatives involving $\beta_1(><)0$, while the F test cannot.

```
#Compute manually:
y_hat = fit$fitted.values
SSTO = sum((Y-mean(Y))^2)
SSE = sum((Y-y_hat)^2)
SSR = sum((y_hat-mean(Y))^2)
MSR = SSR/(1)
MSE = SSE/(n-2)
Fstatistic = MSR/MSE
pvalue = pf(Fstatistic, 1, n-2, lower.tail = F)
result=data.frame(Source=c("Regression", "Error", "Total"),
                  Df=c(1, n-2,n-1), SS=c(SSR, SSE, SSTO),
                  MS=c(MSR, MSE,NA), F_value=c(Fstatistic,NA,NA),
                  p_value=c(pvalue,NA,NA))
kable(result)
```

Source	Df	SS	MS	F_value	p_value
Regression	1	21185.46	21185.4589	89.56711	0
Error	48	11353.52	236.5317		
Total	49	32538.98			

2. Obtain Coefficient of determination (R^2)

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO},\ 0 \leq R^2 \leq 1.$$

We may interpret R^2 as the proportionate reduction of total variation associated with the use of the predictor variable X.

```
#From the output of model fit
summary(fit)$r.squared
```

```
## [1] 0.6510794
```

```
#Or usiing formula
Rsquare = SSR/SSTO
Rsquare
```

```
## [1] 0.6510794
```

Interpretation: 65% variation in "distance" (Y) is explained/reduced by the use of predictor variable "speed" (X).

3. Inferences on Correlation Coefficients (ρ)

Regression model: X values are assumed as known constants, estimate $E(Y|X)$ and β_1, β_0.

Correlation Model: Both X and Y are considered as random variables, (X, Y) has a bivariate-normal distribution. Estimate the *coefficient of correlation* ρ, where

$$\rho = \text{cor}(X, Y) = \frac{E\{(X - \mu_X)(Y - \mu_Y)\}}{\text{sd}(X)\text{sd}(Y)}, \mu_X = E(X),\ \mu_Y = E(Y).$$

Correlation test: $H_0 : \rho = 0$ v.s. $H_a : \rho \neq 0$ (or $\rho < 0$, $\rho > 0$). The test statistic is

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

which has a distribution t_{n-2} under the null hypothesis. Reject H_0 when $|t| > t_{n-2;1-\alpha/2}$ for a two-sided test.

4. Obtain Pearson product-moment correlation coefficient (r)

The Pearson product-moment correlation coefficient (r) is the point estimator for ρ.

Approach 1:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

```
cor(X,Y)
```

```
## [1] 0.8068949
```

```
#or
sum((X-mean(X))*(Y-mean(Y)))/sqrt(sum( (X-mean(X) )^2)*sum((Y-mean(Y))^2))
```

```
## [1] 0.8068949
```

Approach 2: Mathematically, correlation coefficient (r) happens to be equal to the square root of Coefficient of determination (R^2) in regression model.

$$r = \pm\sqrt{R^2} = \begin{cases} \sqrt{R^2}, & \text{if } \hat{\beta}_1 > 0. \\ -\sqrt{R^2}, & \text{if } \hat{\beta}_1 < 0. \end{cases} \quad (1)$$

```
b1hat=fit$coefficients[2]
r=ifelse(b1hat>0, sqrt(Rsquare),-sqrt(Rsquare))
r
```

```
##         X
## 0.8068949
```

Finish the test in part 3:

```
t=r*sqrt(n-2)/sqrt(1-r^2)
t
```

```
##       X
## 9.46399
```

```r
alpha=0.05
critical.value= qt(1-alpha/2,df=n-2)
critical.value
```

```
## [1] 2.010635
```

Since $|t| > t_{n-2;1-\alpha/2}$, we can reject H_0, which means we have enough evidence that X and Y are correlated under $\alpha = 0.05$.

You can also use the R built-in function `cor.test` to get the result of this correlation test. You can find the t-value, p-value, confidence interval in the output of this function.

```r
cor.test(X,Y)
```

```
##
##  Pearson's product-moment correlation
##
## data:  X and Y
## t = 9.464, df = 48, p-value = 1.49e-12
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6816422 0.8862036
## sample estimates:
##       cor
## 0.8068949
```

Appendix: Simulation for different R^2

```r
par(mfrow=c(2,2))#arrange plots 2*2
#R^2=0:
X= rep(c(-1,0,1),3)
Y= rep(1:3,each=3)
plot(X,Y,xlim=c(-2,2),ylim=c(0,4),main="R square=0")
fit1= lm(Y~X)
summary(fit1)#R^2 is Multiple R-squared in the output
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##     -1     -1      0      1      1
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.000e+00  3.086e-01   6.481  0.00034 ***
## X           -4.532e-17  3.780e-01   0.000  1.00000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.9258 on 7 degrees of freedom
## Multiple R-squared:  3.287e-32,  Adjusted R-squared:  -0.1429
## F-statistic: 2.301e-31 on 1 and 7 DF,  p-value: 1
```

```
abline(fit1,col="red")
#increase points to make R^2>0
X=c(X,-2,2)#increase two data points
Y=c(Y, 0,4)
plot(X,Y,xlim=c(-2,2),ylim=c(0,4),main="add points")
fit2= lm(Y~X)
summary(fit2)#R^2 is Multiple R-squared in the output
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5714 -0.7143  0.0000  0.7143  1.5714
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.0000     0.3086   6.481 0.000114 ***
## X             0.5714     0.2736   2.089 0.066298 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.024 on 9 degrees of freedom
## Multiple R-squared:  0.3265, Adjusted R-squared:  0.2517
## F-statistic: 4.364 on 1 and 9 DF,  p-value: 0.0663
```

```
abline(fit2,col="red")
#reduce points to make R^2>0
X=c(-1,0,1)#increase two data points
Y=c(1, 2,3)
plot(X,Y,xlim=c(-2,2),ylim=c(0,4), main="reduce points")
fit3= lm(Y~X)
summary(fit3)#R^2 is Multiple R-squared in the output
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
## 1 2 3
## 0 0 0
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)        2          0     Inf   <2e-16 ***
## X                  1          0     Inf   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0 on 1 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic:   Inf on 1 and 1 DF,  p-value: < 2.2e-16
```

```
abline(fit3,col="red")
```