chapter04_inference

Let's examine the effect of one percent increase in batting avg. onto the winning ratio.

```
library(Lahman)
a = subset(Teams, yearID>2000)
attach(a)
a$avg = H/AB
a$wp = W/G
lm(wp~avg, a)
```

```
##
## Call:
## lm(formula = wp ~ avg, data = a)
##
## Coefficients:
## (Intercept)          avg
##     -0.0924       2.2763
```

pwr.f2.test gives a minimum samples for linear regression

```
library(pwr)
pwr.f2.test(1, NULL, 0.01, 0.05, 0.95)
```

```
##
##      Multiple regression power calculation
##
##               u = 1
##               v = 1299.395
##              f2 = 0.01
##       sig.level = 0.05
##           power = 0.95
```

Do linear regression

```
a = subset(Teams, lgID == 'AL' | lgID=='NL')
b = sample(1:nrow(a), 1302)
c = a[b,]
c$avg = c$H / c$AB
c$wp = c$W/c$G
d = lm(wp~avg, data=c)
summary(d)
```

```
##
## Call:
## lm(formula = wp ~ avg, data = c)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.300902 -0.054594  0.006242  0.054711  0.250172
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.04762    0.03327  -1.431    0.153
## avg          2.08129    0.12702  16.386   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07887 on 1300 degrees of freedom
## Multiple R-squared:  0.1712, Adjusted R-squared:  0.1705
## F-statistic: 268.5 on 1 and 1300 DF,  p-value: < 2.2e-16
```

Inspection of the effectivess of new variable If it is useless and no impact to explain our model, we can not reject the Hypothesis.

H0 : existing model = new model existing model => Winning ratio = beta_0 + error new model => winning ratio = beta_0 + beta_1*batting_avg + error

```
summary.aov(d)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## avg            1  1.670  1.6700   268.5 <2e-16 ***
## Residuals   1300  8.086  0.0062
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

so, we can reject the H0 and we can inspect new model is more effective than existing one.

confidence interval and predicted interval

```
e = data.frame(avg = 0.270)
predict(d, e, level=0.95, interval = 'confidence')
```

```
##         fit       lwr       upr
## 1 0.5143324 0.5095366 0.5191283
```

```
e = data.frame(avg = 0.270)
predict(d, e, level=0.95, interval = 'predict')
```

```
##         fit       lwr      upr
## 1 0.5143324 0.3595359 0.669129
```