

chapter04_inference

```
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
## -0.34992 -0.05183  0.00477  0.05310  0.28032
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.03715    0.03385  -1.098   0.273
## avg          2.04722    0.12888  15.884 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.07872 on 1300 degrees of freedom
## Multiple R-squared:  0.1625, Adjusted R-squared:  0.1619
## F-statistic: 252.3 on 1 and 1300 DF,  p-value: < 2.2e-16
```

F-statistics can be derived that between group deviation divided by within-group deviation. Inspection of the effectiveness of new variable. If it is useless and has no impact to explain our model, we can not reject the Hypothesis.

H_0 : existing model = new model
existing model \Rightarrow Winning ratio = β_0 + error
new model \Rightarrow winning ratio = $\beta_0 + \beta_1 \cdot \text{batting_avg} + \text{error}$

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

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## avg              1  1.564   1.5636   252.3 <2e-16 ***
## Residuals    1300   8.056   0.0062
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

so, we can reject the H_0 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 0.5155973 0.5108728 0.5203219
```

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

```
##          fit          lwr          upr
## 1 0.5155973 0.3610894 0.6701053
```