Estimating Covariance with Bootstrap

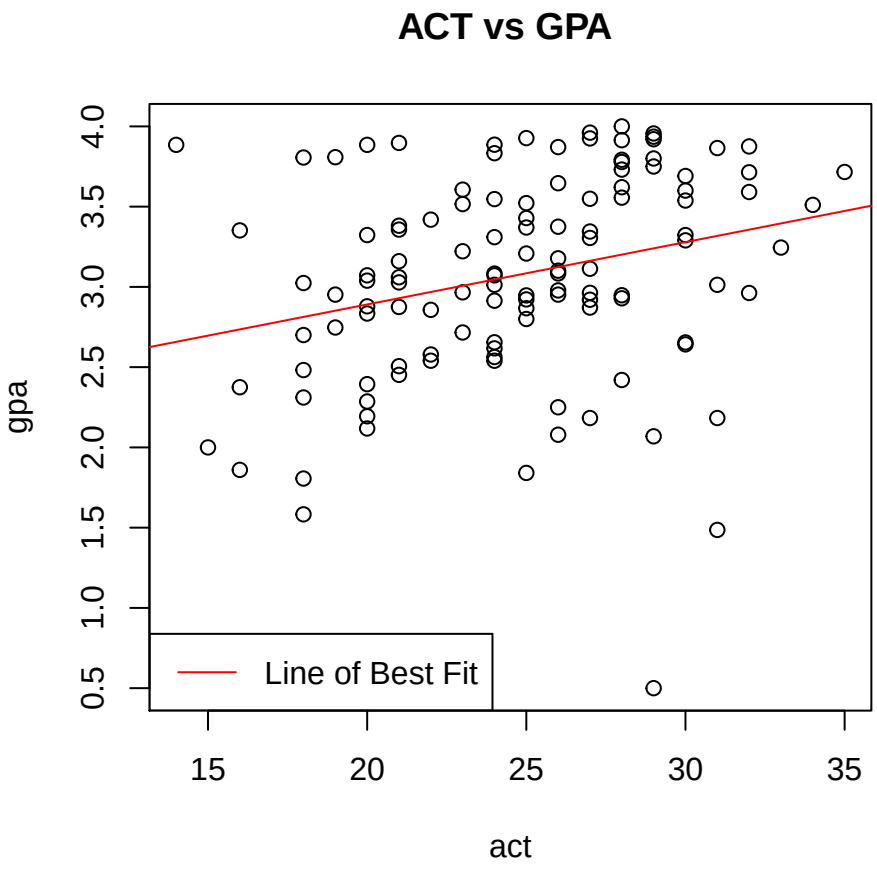
Jaime

10/27/2020

```
library(boot)
library(caret)
library(bestglm)
```

1. Consider the gpa data stored in the gpa.csv file available on eLearning. The data consist of GPA at the end of freshman year (gpa) and ACT test score (act) for randomly selected 120 students from a new freshman class.

(a) Make a scatterplot of gpa against act and comment on the strength of linear relationship between the two variables.



Interpreting Results

From the previous scatter plot it is evident to see:

- The points are not very close to each other or the mean, in other words words there is a weak relationship.
- The slope is positive increasing, hence there exists a linear relationship.

We can conclude that there is a **Weak Linear Relationship** between ACT and GPA.

(b) Let p denote the population correlation between gpa and act. Provide a point estimate of p, bootstrap estimates of bias and standard error of the point estimate, and 95% confidence interval computed using percentile bootstrap. Interpret the results.

The correlation between GPA and ACT is:

```
## [1] 0.2694818
```

```
#Creating function to manually calculate bootstrap for every x,y values
get_correlation <- function(data, index) {
  gpa.index <- gpa_dataset$gpa[index]
  act.index <- gpa_dataset$act[index]
  result <- cor(gpa.index, act.index)
  return(result)
}

#Running correlation 1000 for x,y values in dataset
gpa_boot <- boot(gpa_dataset, get_correlation, R = 1000, sim = "ordinary", stype = "i")
```

```
#Estimate for correlation of ACT vs GPA based on 1000 bootrap samples
correlation_gpa_act_bootstrap <- mean(gpa_boot$t)
correlation_gpa_act_bootstrap
```

Point Estimate of p:

```
## [1] 0.2677492
```

```
# Bootstrap Estimate of Bias
correlation_gpa_act_bootstrap - correlation_gpa_act
```

Bootstrap Estimates of Bias

```
## [1] -0.001732573
```

```
# Standard Error of Point Estimate - bootstrap replicate of the result of calling statistic.
sd(gpa_boot$t)
```

Standard Error of Point Estimate P

```
## [1] 0.1075271
```

```
# 95% confidence interval
boot.ci(gpa_boot, type = "perc")
```

Confidence Interval using Bootstrap

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = gpa_boot, type = "perc")
##
## Intervals :
## Level     Percentile
## 95%   ( 0.0594,  0.4838 )
## Calculations and Intervals on Original Scale
```

Interpreting Results

Actual p: 0.2694818

Using Bootstrap Estimate p: 0.2747637

Standard Error of p: 0.1102458

We are 95% for that true correlation values (p) will be between 0.0645, 0.4879.

(c) Fit a simple linear regression model for predicting gpa on the basis of act. Provide the least square estimates of the regression coefficients, standard errors of the estimates, and 95% confidence intervals of the coefficients. Perform model diagnostics to verify the model assumptions and comment on the results.

```
#Fitting a Model
full_model <- lm(gpa ~ act, data = gpa_dataset)

#Least Square Estimates and SE of estimates
summary(full_model)
```

Fitting a Model and getting LSE and SE of estimates:

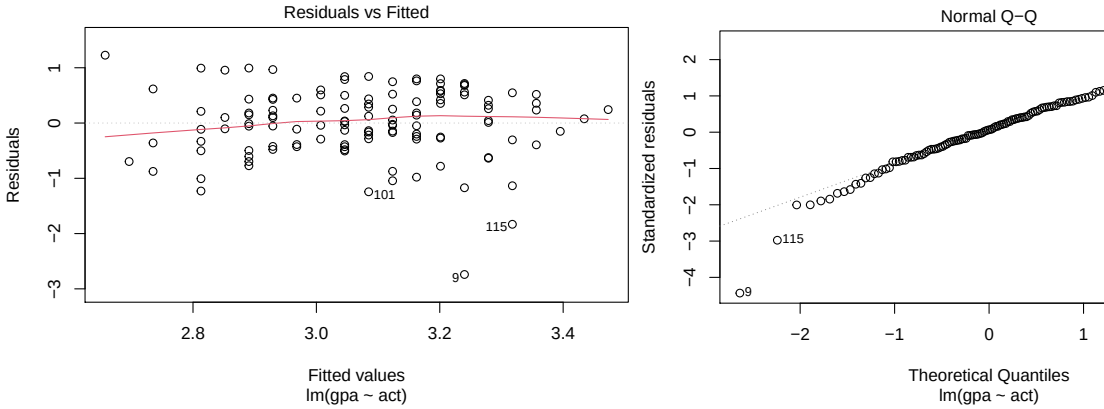
```
##
## Call:
## lm(formula = gpa ~ act, data = gpa_dataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.74004 -0.33827  0.04062  0.44064  1.22737
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.11405    0.32089   6.588  1.3e-09 ***
## act          0.03883    0.01277   3.040  0.00292 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6231 on 118 degrees of freedom
## Multiple R-squared:  0.07262,    Adjusted R-squared:  0.06476
## F-statistic:  9.24 on 1 and 118 DF,  p-value: 0.002917
```

```r
# 95% confidence interval
confint(full_model, level = 0.95, method = "percentile")
```

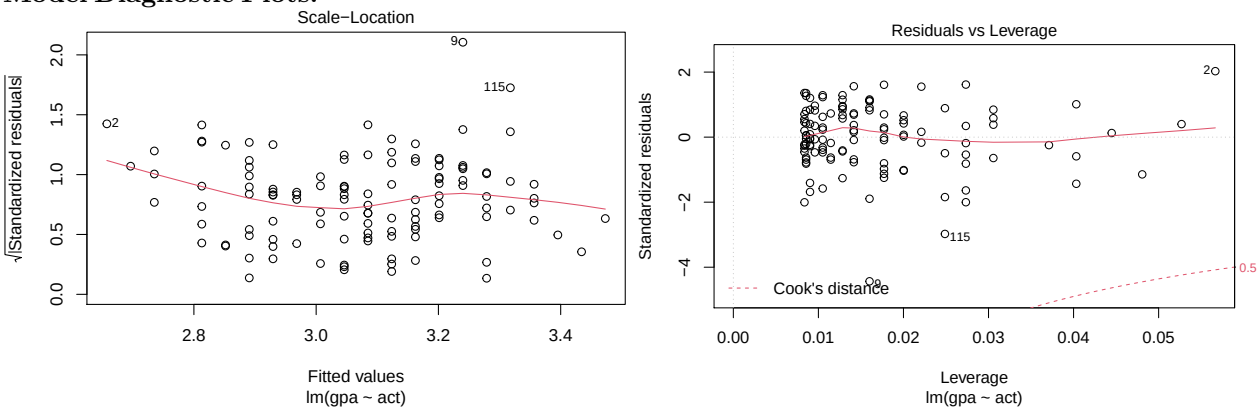
Performing a 95% confidence intervals of the coefficients:

```
##                  2.5 %     97.5 %
## (Intercept) 1.47859015 2.74950842
## act         0.01353307 0.06412118
```

```r
plot(full_model)
```



Model Diagnostic Plots:



Interpreting Results

From the Diagnostic Plots we can see:

- **Residuals vs Fitted.**
- **Normal Q-Q.** The residuals look fairly normally distributed. With the exception of a couple of outliers.
- **Residuals vs Leverage.**

(d) Use nonparametric bootstrap to compute the standard errors and 95% confidence intervals (using percentile bootstrap) mentioned in part (c) and compare the two sets of results.

```r
#Function to compute coefficients of line of best fit 1000 times
coefficients_line_of_best_fit <- function(data, index) {
  coefficients_for_sample <- coef(lm(gpa ~ act, data = gpa_dataset, subset = index))[2]
```

```
  return(coefficients_for_sample)
}

boot_regression_coefficients <- boot(gpa_dataset, coefficients_line_of_best_fit, R = 1000)
```

```
#coefficient estimate of gpa
coefficients_bootstrap <- mean(boot_regression_coefficients$t)
coefficients_bootstrap
```

Coefficient Estimate Using Bootstrap

```
## [1] 0.03932953
```

```
# standard error estimate
sd(boot_regression_coefficients$t)
```

Standard Error Using Bootstrap

```
## [1] 0.01467852
```

```
# 95% confidence interval
boot.ci(boot_regression_coefficients, type = "perc")
```

Confidence Interval Using Bootstrap

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot_regression_coefficients, type = "perc")
##
## Intervals :
## Level     Percentile
## 95%   ( 0.0110,  0.0676 )
## Calculations and Intervals on Original Scale
```

Interpreting Results

Least Square Estimates

- Coefficient of ACT: 0.03883
- Standard Error of ACT: 0.01277
- Conf. Interval of ACT: 0.01353307 0.06412118

Bootstrap Estimates

- Coefficient of ACT: 0.03924538
- Standard Error of ACT: 0.01443574
- Conf. Interval of ACT: 0.0096, 0.0671