p8130_final_jn2761

jiaying Ning

12/11/2020

Problem 1

(a)

- **Outcome**:
  - Antibodies : level of antibodies at 6-month
- **Predictor**:
  - age : age (years) of patients at baseline
  - Vaccine 1
  - Vaccine 2
  - Placebo
- **equation of the linear regression model**
  - `Antibodies(i) = beta0 + beta1*age(i) + beta2*I(Trt=Vaccine1)(i) + beta3*(Trt=Vaccine2)(i)+erro`
  - i=1,2,. . . ,30
  - With placebo as reference category
- **distribution assumption imposed on the error terms**
  - Error terms should be uncorrelated anf follow a normal distribution with mean = 0 and a constant variance. N(0,var)

(b)

Problem 2

(a)

$$salaries(hat) = 17.847+1.103*Years(hat)+0.322*Publications(hat)+1.593*Gender(hat)+1.289*Grants(hat)$$

(b)

$$slopeGender + (-)t(30, 0.975) * se(Gender)$$

```
qt(0.975,30)
```

```
## [1] 2.042272
```

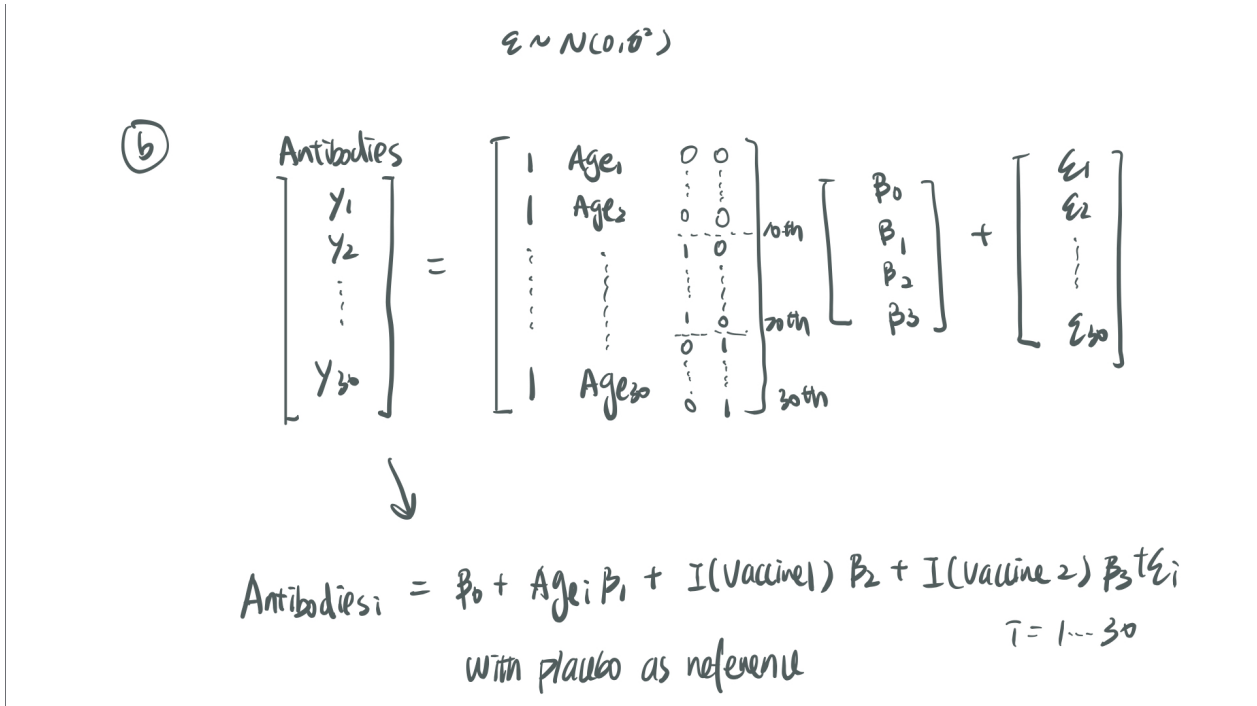


Figure 1: Matrix notation

```
gender_upper=1.593+(qt(0.975,30)*0.688)
gender_lower=1.593-(qt(0.975,30)*0.688)
```

- 95% CI for the coefficient associated with 'Gender' is (0.1879166,2.9980834)

(c)

Hypothesis

- H0:There is no linear association exists between publication and faculty salaries (in thousands of dollars).beta_publication=0
- H1:There is linear association exists between publication and faculty salaries (in thousands of dollars).beta_publication!=0

Calculations

$$(beta_publication - 0)/se(publication) = 0.322/0.521 = 0.6180422$$

Decision

```
qt(0.975,30)
```

```
## [1] 2.042272
```

- **Critical Value**: t~(30,0.975) = 2.042272
- **Decision Rule**:
  - Reject H0: if |t| > t(30,0.975)
  - Fail to reject H0:|t| < t(30,0.975)

Conclusion

- For the current data, we have test stats of 0.6180422 for slope. Since t_stats:0.6180422 < critical value 2.042272, we fail to reject the null and conlcude that there is no significant linear association exists between publication and faculty salaries (in thousands of dollars).

comment - Now we know that there is no significant linear association exists between publication and faculty salaries, we can conclude that the the most significant variables in predicting salary is the success in obtaining grant support.

(d)

- We could not generalize our findings to the entire academic community (i.e., US) because even though we yield some significance result from our current data, all of out samples are drawn from the MSPH faculty which might not represent all faculty in the united states. Differences between department and location might make our current conlusion not applicable.

Problem 3

Among the three model, I would recommend model B.

- Mallow's Cp Criteria: Cp compares the predictive ability for each model to the full modes, we want to choose Cp values that is <= number of parameter. From the three models, Model A's Cp value is larger than the number of parameter. Model B and C both have Cp smaller than the number of parameter, but model B have Cp value that are closer to the number of parameter. Also Model B have smaller number of paramter but similar Cp, which meet the goal of model selection: "get a small number of variables while maintaining the same predictive ability"
- Adjusted R-squared: Adjusted R^2 tell us how well the variance of outcome is explained by the relationship with predictors. Therefore, we will want to have models that have larger adjusted R-squared value. Overall, Model B have the largest adjusted R-squared value.
- MSE:
  - We first compare the MSE between models and see that all three models have similar MSE. Model B and Model C have smaller MSE than Model1.
  - We then compare MSE on Testing and Training data within each model. Which I calculate the difference between MSE and MSPE within each model. the difference between MSE and MSPE is also smallest in Model B comparing to other models.

```
0.078-0.045
```

```
## [1] 0.033
```

```
0.072-0.043
```

```
## [1] 0.029
```

```
0.079-0.042
```

```
## [1] 0.037
```

- Therefore, overall, I would recommend model B.

Problem 4

(a)

```
#Import Data
hospital_df=
   read.csv("./data/Hospital.csv")
```

fit regression model

fit a simple linear regression with length of stay (LOS) as the outcomes and number of beds (BEDS) as only predictor. Use R to address the following points:

```
  linear = lm(LOS ~ BEDS, data=hospital_df)
```

summary

```
summary(linear)
```

```
##
## Call:
## lm(formula = LOS ~ BEDS, data = hospital_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8291 -1.0028 -0.1302  0.6782  9.6933
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 8.6253643  0.2720589  31.704  < 2e-16 ***
## BEDS        0.0040566  0.0008584   4.726 6.77e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.752 on 111 degrees of freedom
## Multiple R-squared:  0.1675, Adjusted R-squared:   0.16
## F-statistic: 22.33 on 1 and 111 DF,  p-value: 6.765e-06
```

95% CI for the true slope

$$slope + (-)t(111, 0.975) * se(beta)$$

```r
confint(linear,level=0.95)
```

```
##                   2.5 %      97.5 %
## (Intercept) 8.086261517 9.164467086
## BEDS        0.002355649 0.005757623
```

- 95% CI for the true slope is (0.002355649,0.005757623)

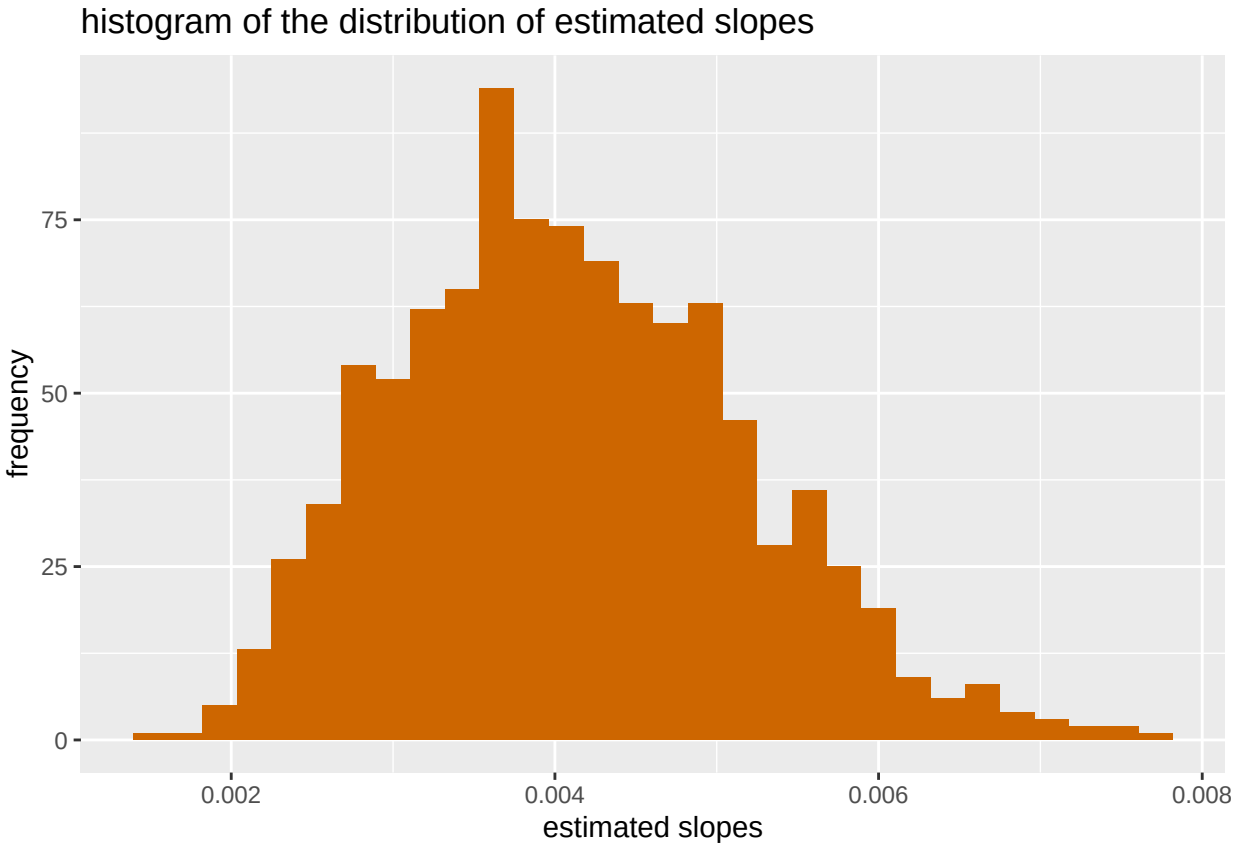
(b)

```r
library(infer)
boot_dist <- hospital_df %>%
specify(formula = LOS ~ BEDS) %>%
generate(reps = 1000, type = "bootstrap") %>%
calculate(stat = "slope")
# Save the estimated slope values (1,000)
boot_dist_slope <- boot_dist[[2]]
```

(c)

```r
library(ggplot2)
 boot_dist%>%
  ggplot(aes(x = stat)) + geom_histogram(fill="darkorange3")+labs(
    titles = "histogram of the distribution of estimated slopes",
    x = "estimated slopes",
    y = "frequency"
  )
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Comment: In order to make inferences, the sampling distribution of slope should follow normal distribution. From the hisrogram we see that the estimated slope roughly follow a normal distribution, thus we conclude that after bootstrap, the distribuiton of slope does approximate the shape of the sampling distribution of beta1 which meet the assumption of normal distribuion for parameters.

(d)

```
lower=quantile(boot_dist_slope,0.025)
upper=quantile(boot_dist_slope,0.975)
```

- The 95% CI for the true slope is (0.0022891,0.0063338)

(e)

- Before Bootstrap:95% CI for the true slope is (0.002355649,0.005757623)
- After bootstrap:The 95% CI for the true slope is (0.0022891,0.0063338)

Overall, both confidence interval have similar value, meaning the orginal data does represent the undelying distribution. To be exact, the 95% CI for the true slope after bootstrap has wider interval comparing to the 95% CI before bootstrap.