Statistical Computing Coursework 3

December 2022
Student I.D. 10459970

1 Birth Length Data

We are given `birth_length.txt`, a file containing 42 records in 4 columns of data, which correspond to:

1. **length** y: the length of a newborn baby, in inches.
2. **mheight** m: the height of the newborn's mother, in inches.
3. **fheight** f: the height of the newborn's father, in inches.
4. **smoker** s: a binary variable indicating the mother's smoking status.

A clinician proposes the following linear model to predict a newborn's length:

$$y_i = \alpha + \beta m_i + \gamma f_i + \varepsilon_i, \qquad (i = 1, 2, ...42),$$

where the errors ε_i are assumed to be observations of independent and identically distributed zero mean random variables.

Using R, I calculated the least-squares estimates of α, β and γ:

```
> install.packages("scatterplot3d") # Install
> library("scatterplot3d") # load
> install.packages("boot") # Install
> library("boot") # load
#set directory; commented out for marking
#setwd("\\\\nask.man.ac.uk\\home$\\Documents\\R")
#import data
> table <- read.table("birth_length.txt", header = TRUE, sep = "", dec = ".");
```

```
#retrieve columns
> length <- table$length;
> mother_height <- table$mheight;
> father_height <- table$fheight;
> smoker <- table$smoker;
#scatterplot3d(mheight, fheight, length)
> fit <- lm(length ~ mheight + fheight, data = table);
> print(fit$coefficients)

(Intercept)     mheight     fheight
 6.65996234  0.17164827  0.03128298
```

As shown above, we have the estimates for the coefficients α (Intercept), β (mheight), and γ (fheight). The residuals ε are shown in a histogram below:

```
> hist(fit$residuals, nclass = 15, freq = FALSE,
+ main = "Histogram of residuals", xlab = "residuals (inches)")
> curve(dnorm(x, m=0, sd=1), col="darkblue", lwd=2, add=TRUE, yaxt="n")
```

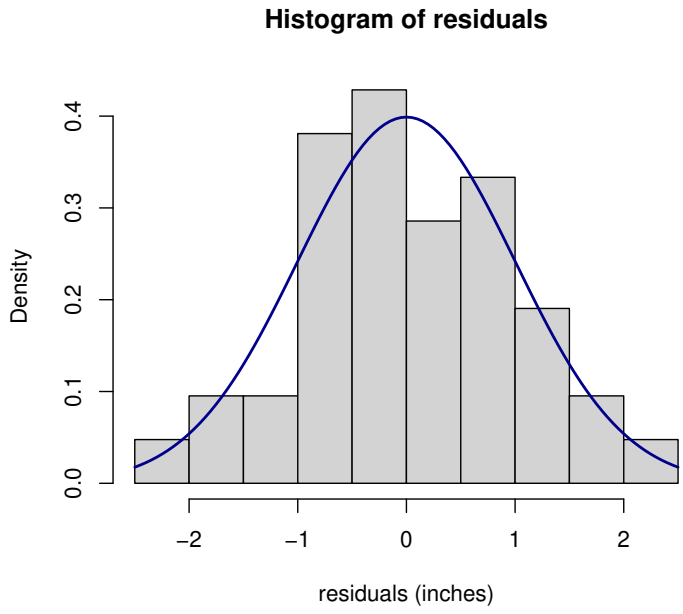

Figure 1: A histogram of the residuals ε of the least-squares method applied to birth length data, with a normal distribution fitted.

Despite the limited data we are given, Fig. 1 fits reasonably well to a normal distribution. Although other distributions may also fit well, residuals of an individual's height are typically modelled to be normally distributed[1], which merits this assumption here.

2 Bootstrap Residuals

For a sufficiently large sample, these estimates would be unlikely to be far off the true population least-squares estimates, presenting sufficient evidence to suggest a correlation between the variables mheight and fheight, and length. Unfortunately, few newborns were available when these data were taken and consequently the least-squares method results is rendered less useful by large potential variance between this sample's characteristics and those of the true population. To calculate a sampling distribution of the coefficients, we can use the bootstrap residuals method. In this context, the bootstrap method calculates the least-squares method on 42 randomly chosen samples of our data 1000 times to produce an estimate for the values of the coefficients. Importantly, when we sample our data, we replace the sample chosen every time, which is a resampling method allowing estimates of the accuracy of our estimates. Below is an implementation of the bootstrap residuals method in R to calculate a sampling distribution of $\hat{\gamma}$:

```
> n <- nrow(table);
> B <- 1000;
> errors <- residuals(fit);
#y fitted to equation
> y_fitted = fitted(fit);
#residuals matrix to be filled
> bs_resid <- matrix(0, nrow = B, ncol = 3);
> colnames(bs_resid) <- c("alpha", "beta", "gamma")
#create bootstrap samples
> for(i in 1:B){
+   #42 random selections from residuals
+   e_star <- sample(errors, n, replace = TRUE);
+   #fit residuals to samples
+   y_star <- y_fitted + e_star;
+   #perform least squares on selections
+   fit_star <- lm(y_star ~  mother_height + father_height);
+   #fill residuals column with bootstrap residuals
+   bs_resid[i, ] <- coef(fit_star);
+ }
```

```
> hist(bs_resid[ , 3], xlab = "gamma", freq = FALSE,
+ main = "Histogram of residuals of gamma")

#estimate mean and standard error of father_height
> estimates <- apply(bs_resid, 2, function(father_height)
+ c(mean = mean(father_height), sd = sd(father_height)))

#bias of father's height; mean of residuals - least squares of sample
> estimates[1, 3] - fit$coefficients[3]
      fheight
0.0009475225

#standard error of beta
> estimates[2, 3]
[1] 0.05563766

#Proportion of residuals of beta below zero
> length(bs_resid[, 3][bs_resid[, 3] < 0]) / B
[1] 0.273
```

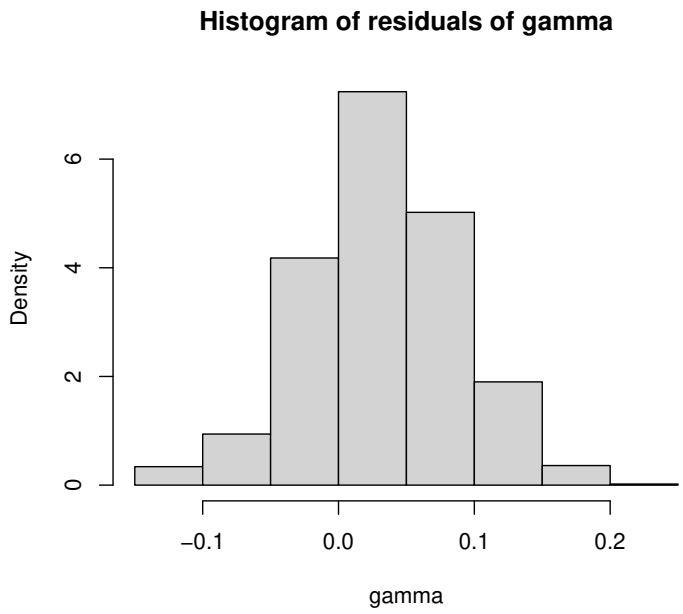

Figure 2: A histogram of the residuals of γ from applying the bootstrap residuals method to birth length data.

From these bootstrap residuals, we can see that the bias is a small proportion of the calculated value, showing a small difference between the average bootstrap value and the least-squares of the sample. But, the standard error is comparatively large, resulting in 27% of calculated values of $\hat{\gamma}$ being on the other side of zero. This suggests that we have insufficient data to determine that a newborn's length correlates with the father's height.

3 Bootstrap-t

With these data, we can also use the bootstrap-t method to produce confidence intervals for the estimate of $\hat{\beta}$, which allows us to determine if it is likely that the birth length of a newborn is correlated with the mother's height:

```
#mean of beta from data
> mean_beta = mean(fit$coefficients[2]);
#standard deviation of beta from data
> sd_beta = summary(fit)$coefficients[2 , 2];
```

```
#bootstrap residuals again with extra column for z-score calculation
> bs_resid_beta <- matrix(0, nrow = B, ncol = 4);
> colnames(bs_resid_beta) <- c("alpha", "beta", "gamma", "Zscore_beta")
> for(i in 1:B){
+   e_star <- sample(errors, n, replace = TRUE);
+   y_star <- y_fitted + e_star;
+   fit_star <- lm(y_star ~  mother_height + father_height);
+   bs_resid_beta[i,1:3] <- coef(fit_star);
+   #calculation of z-score
+   bs_resid_beta[i,4] <- (fit_star$coefficients[2] - mean_beta
+   ) / summary(fit_star)$coefficients[2, 2];
+ }

#2.5th and 97.5th percentiles of z-scores
> low_interval <- sort(bs_resid_beta[ , 4])[25];
> high_interval <- sort(bs_resid_beta[ , 4])[975];

#convert z-score percentiles into beta values
> c(mean_beta - high_interval * sd_beta, mean_beta - low_interval * sd_beta);
[1] 0.03375923 0.31337835
```

Since the 95% confidence interval excludes 0, I conclude that it is implausible that $\beta = 0$, suggesting that there is sufficient evidence to assume a correlation between the birth length of a newborn and their mother's height.

4 Mother and Father's height

We can also use the bootstrapping method in order to determine confidence intervals for the relation between two variables. Using *R,* I calculated the 95% confidence intervals for the value of EM - EF:

```
> mean_m = mean(mother_height);
> sd_m = sd(mother_height);
> mean_f = mean(father_height);
> sd_f = sd(father_height);
> bs_t <- matrix(0, nrow = B, ncol = 1);
> colnames(bs_t) <- c("zscore_m-f");
> for(i in 1:B){
+   m_star <- sample(mother_height, n, replace = TRUE);
+   f_star <- sample(father_height, n, replace = TRUE);
```

```
+   #calculation of z-score for the combined EM-EF value
+   bs_t[i] <- ((mean(m_star) - mean(f_star)) - (mean_m
+   - mean_f)) / sqrt(sd(m_star)^2+sd(f_star)^2);
+ }
> low_interval_m_f <- sort(bs_t)[25];
> high_interval_m_f <- sort(bs_t)[975];
#convert z-score percentiles back into EM-EF values
> c((mean_m - mean_f) - high_interval_m_f
+ * sqrt(sd_m^2 + sd_f^2), (mean_m - mean_f)
+ - low_interval_m_f * sqrt(sd_m^2 + sd_f^2));
[1] -7.594846 -5.227351
```

Since the 95% confidence intervals aren't even remotely close to 0, we can determine that it is very implausible that $\mathrm{E}M = \mathrm{E}F$. Our results suggest instead that the average height of the father lies between between 5.22 and 7.59 inches taller than the average height of the mother.

We can also use bootstrapping methodology to attempt to see if the smoking status of the mother will affect the length of the newborn. By splitting the dataset into two conditionally on the mother's smoking status, we can use bootstrapping to estimate the confidence intervals of $\theta = E(Y|S = 0) - E(Y|S = 1)$:

```
> smoke = c();
> non_smoke = c();
#Split the length of newborns data by smoking status of mother
> for(i in 1:n){
+   if(table[i, 4] == 0){
+     non_smoke <- append(non_smoke, table[i, 1]);
+   } else{
+     smoke <- append(smoke, table[i, 1]);
+   }
+ }
> theta_1 = mean(smoke);
> theta_0 = mean(non_smoke);
> expectation = theta_0 - theta_1;
> bs_t <- matrix(0, nrow = B, ncol = 1);
> colnames(bs_t) <- c("zscore_smoke");
> for(i in 1:B){
+   #take smaller bootstrap samples from both lists
+   smoke_star <- sample(smoke, length(smoke), replace = TRUE);
+   non_smoke_star <- sample(non_smoke, length(non_smoke),
```

```
+   replace = TRUE);
+   bs_t[i] <- ((mean(non_smoke_star) - mean(smoke_star)) -
+   (expectation)) /
+   sqrt(sd(non_smoke_star)^2+sd(smoke_star)^2);
+ }
#convert list of z-scores back into values of theta hat
> bs_t <- expectation - bs_t * sqrt(sd(smoke)^2 + sd(non_smoke)^2)
#count proportion of estimates above zero
> sum(bs_t > 0) / B;
[1] 0.92
> hist(bs_t, main = "Histogram of bootstrapping estimates of theta-hat",
xlab = "theta-hat", freq = FALSE)
```

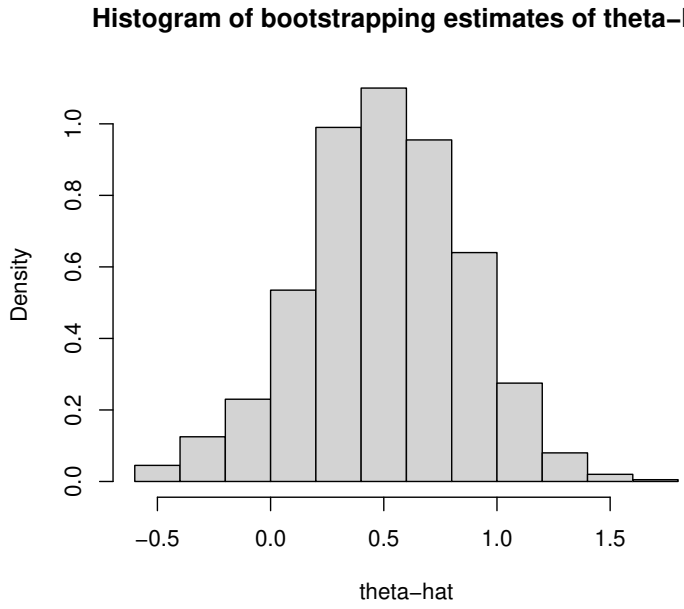

Figure 3: A histogram of the bootstrap estimates of theta hat.

Our data suggests that there is a 92% that $\hat{\theta} > 0$, which is not sufficient to reject the null hypothesis. This goes to show that although bootstrapping is a powerful tool for estimating the confidence intervals of results when working with small datasets, there is no substitute for taking enough measurements to draw conclusive results, especially when the difference may be small. In the case of the difference in height of

mothers and fathers, not much data was necessary to draw a conclusion due to a very consistent and large difference in values. However, the smoking status of the mother didn't cause a big enough difference to draw any conclusions from our data. In reality, smoking during pregnancy has been linked to a lower birth length, along with a slew of other harmful conditions, and tobacco companies have been using inconclusiveness to their advantage in order to sell more products whilst being aware of the health risks themselves[2].

5 bibliography

1: Slavskii, S.A., Kuznetsov, I.A., Shashkova, T.I. et al (2021). *The limits of normal approximation for adult height.* https://doi.org/10.1038/s41431-021-00836-7
2: Brandt AM. (2012). *Inventing conflicts of interest: a history of tobacco industry tactics. Am J Public Health.* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3490543/