Multivariate HW07

Luke Beebe

2024-04-03

```
data("Heights", package="alr4")
dim(Heights)
```

```
## [1] 1375    2
```

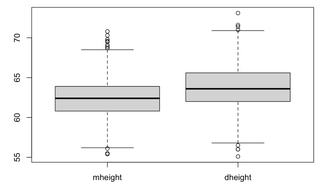
```
head(Heights)
```

```
##   mheight dheight
## 1    59.7    55.1
## 2    58.2    56.5
## 3    60.6    56.0
## 4    60.7    56.8
## 5    61.8    56.0
## 6    55.5    57.9
```

1

Are daughters' heights on average higher than mothers' heights? Investigate this question by proposing a statistical model for the data that includes a parameter relevant to the question. Comment on whether your assumptions are satisfied or not and what kind of test you should do. Report the results of your statistical analysis, and comment on the conclusions you have drawn.

```
boxplot(Heights)
```




```
t.test(Heights$dheight, Heights$mheight, paired=T)
```

```
##
##  Paired t-test
##
## data:  Heights$dheight and Heights$mheight
## t = 19.184, df = 1374, p-value < 2.2e-16
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  1.165499 1.431010
## sample estimates:
## mean difference
##        1.298255
```

```
shapiro.test(Heights$dheight-Heights$mheight)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Heights$dheight - Heights$mheight
## W = 0.99835, p-value = 0.2048
```

There is a statistically significant difference between daughters' heights and mothers' heights, daughters' heights being higher. After glossing over the boxplots of our data I investigated this question by using a pairwise t-test to compare the means. I used a t-test over a z-test because we do not know the population standard deviation. After finding the difference between the two groups significant (we rejected the null), I checked our normal assumptions with a Shapiro-Wilke test and found our data to be approximately normal.

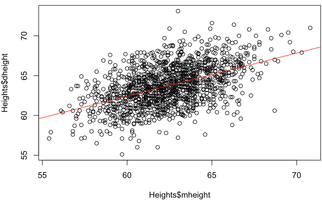
2

What is the relationship between a mother's height and her daughter's height? Propose a linear model that contains parameters relevant to this question, and comment on the appropriateness of the model.

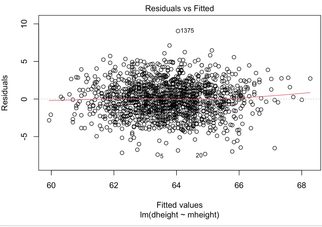
```
model <- lm(dheight ~ mheight, data = Heights)
summary(model)
```

```
##
## Call:
## lm(formula = dheight ~ mheight, data = Heights)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.397 -1.529  0.036  1.492  9.053
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 29.91744    1.62247   18.44   <2e-16 ***
## mheight      0.54175    0.02596   20.87   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.266 on 1373 degrees of freedom
## Multiple R-squared:  0.2408, Adjusted R-squared:  0.2402
## F-statistic: 435.5 on 1 and 1373 DF,  p-value: < 2.2e-16
```

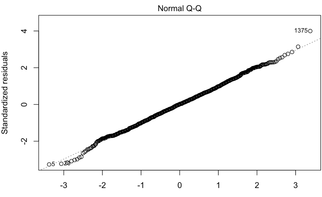
```
plot(Heights$mheight, Heights$dheight)
abline(model, col='red')
```




```
plot(model, which=1)
```




```
plot(model, which=2)
```



According to our linear model, **there is a statistically significant, positive, linear relationship between a mother's height and her daughter's height**, following the equation: dheight = 29.9174 + 0.54175 * mheight

Based on the residual and Q-Q plots, we see that the residuals our model produces are approximately normal in their distribution.

3

Fit the model and report the results of fitting the model, including parameter estimates and standard errors where appropriate. Comment on the results.

```
model <- lm(dheight ~ mheight, data = Heights)
summary(model)
```

```
##
## Call:
## lm(formula = dheight ~ mheight, data = Heights)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.397 -1.529  0.036  1.492  9.053
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 29.91744    1.62247   18.44   <2e-16 ***
## mheight      0.54175    0.02596   20.87   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.266 on 1373 degrees of freedom
## Multiple R-squared:  0.2408, Adjusted R-squared:  0.2402
## F-statistic: 435.5 on 1 and 1373 DF,  p-value: < 2.2e-16
```

As above, the equation derived with our model is: **dheight = 29.9174 + 0.54175 * mheight**
This signifies that for every inch of mheight (mother's height), there is an expected increase of 0.5417 of dheight (daughter's height), added to 29.9174.
The p-values suggest the estimates of mheight, intercept are significant (are not 0). The residual standard error (standard deviation of the residuals) is 2.266 on 1373 degrees of freedom, which seems in line given our data. The R-squared value is 0.2408, meaning 24.08% of the variability in dheight is explained by mheight.

4

Compute a 95% confidence interval for the slope parameter (hint, use the qt function to help with this).

```
slope = 0.54175
se = 0.02596
margin = qt(0.975, 1373) * se
lower = slope - margin
upper = slope + margin
paste(lower, upper)
```

```
## [1] "0.490824442344626 0.592675557655374"
```

The 95% CI for the parameter of slope is approximately **(0.4908, 0.5927)**

5

Compute predicted values and prediction intervals for daughter's height when mother's height is 64 inches and 68 inches, respectively.

```
predict(model, data.frame(mheight=c(64,68)), interval="prediction", level=0.95)
```

```
##       fit      lwr      upr
## 1 64.58925 60.14112 69.03737
## 2 66.75623 62.29985 71.21262
```

The prediction interval for: **64 inches = (60.14112, 69.03737)**, **68 inches = (62.29985, 71.21262)**