BMI 713: Computational Statistics for Biomedical Sciences

Assignment 7

Simple Linear Regression

1. To study the relationship between a father’s height and his son’s height, Karl Pearson (1857-1936) collected the data of heights from 1078 father-son pairs.

(a). Get the dataset by the following R commands:

```r
install.packages("UsingR")
library(UsingR)
data(father.son)
```

Then the data frame `father.son` contains the 1078 observations on 2 variables: `fheight` (father’s height in inches, x) and `sheight` (adult son’s height in inches, y).

(b). Draw a scatter plot of son’s height versus father’s height. Does the relationship appear linear?

**Sol’n.** From the scatter plot below, we can see that the son’s height tends to increase as the father’s height increases.

```r
> plot(fheight, sheight, xlab="Father’s height (in)", ylab="Son’s height (in)",
    xlim=c(58,78), ylim=c(58,80), bty="l", pch=20)
```

(c). Fit the simple linear regression of son’s height on father’s height. What are the estimated regression coefficients, \(a\) and \(b\), respectively?

**Sol’n.** Denote the 1078 father-son pairs of observations as \((x_1, y_1), \ldots, (x_n, y_n)\), where \(n = 1078\).

We will fit the linear regression model of son’s height \(y\) on father’s height \(x\):

\[
y = \alpha + \beta x + e, \quad e \sim N(0, \sigma^2)
\]
Fit the linear model by the method of least squares, and the estimated regression coefficients are:

\[
b = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\text{Cov}(x,y)}{\text{Var}(x)} = 0.514 \quad \text{(slope)},
\]

and

\[
a = \bar{y} - b\bar{x} = 33.89 \quad \text{(intercept)}.
\]

We can also fit the linear regression model in R by the function `lm`:

```R
> m <- lm(sheight ~ fheight)
> summary(m)
```

**Call:**

```
lm(formula = sheight ~ fheight)
```

**Residuals:**

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-8.877151</td>
<td>-1.514415</td>
<td>0.007896</td>
<td>4.415194</td>
<td>8.968479</td>
</tr>
</tbody>
</table>

**Coefficients:**

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | 33.88660 | 1.83235 | 18.49    | <2e-16 *** |
| fheight   | 0.51409   | 0.02705 | 19.01    | <2e-16 *** |

---

**Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

Residual standard error: 2.437 on 1076 degrees of freedom
Multiple R-squared: 0.2513, Adjusted R-squared: 0.2506
F-statistic: 361.2 on 1 and 1076 DF, p-value: < 2.2e-16

(d). Add the regression line \( y = a + bx \) to the plot in (b).

**Sol'n.**

```R
> abline(lm(sheight ~ fheight), lty=1, lwd=2)
```
(e). Calculate Pearson correlation coefficient $r$ between father’s height and son’s height. Perform a proper test to test the null hypothesis $\rho = 0$, where $\rho$ is the population correlation coefficient.

**Sol’n.** Pearson correlation coefficient $r$ is defined as

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right)}} = \frac{\text{Cov}(x,y)}{\sqrt{\text{Var}(x)\text{Var}(y)}}$$

We can calculate it in R:

```r
> r <- cov(fheight, sheight) / (sd(fheight) * sd(sheight))
> r
[1] 0.5013383
```

or using `cor` function in R:

```r
> cor(fheight, sheight)
[1] 0.5013383
```

To test the null hypothesis $\rho = 0$, we need to first calculate the standard error of $r$:

$$SE(r) = \sqrt{\frac{1-r^2}{n-2}} = 0.026$$

Under the normality assumption and null hypothesis, the $t$-statistic

$$t = \frac{r - 0}{SE(r)} = \frac{r}{\sqrt{\frac{n-2}{1-r^2}}}$$

follows a $t$ distribution, with $n - 2 = 1076$ degrees of freedom.

For $r = 0.501$, the test statistic is $t = 19$, which gives the $p$-value less than $1e-10$. Thus, we reject the null hypothesis $\rho = 0$.

To test the null hypothesis in R:

```r
> cor.test(fheight, sheight)
```

Pearson's product-moment correlation

data:  fheight and sheight
t = 19.0062, df = 1076, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
  0.4552586 0.5447396
sample estimates:
cor
  0.5013383
```

(f). What is the 95% confidence interval for the slope coefficient $\beta$?

**Sol’n.** To construct the confidence interval, we first need to determine the standard error of $b$

$$SE(b) = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n}(y_i - \hat{y})^2 / \sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Compute it in R:

```r
> SE.b <- sqrt(1/(1078-2) * sum((sheight-fitted(m))^2) / sum((fheight-mean(fheight))^2))
> SE.b
[1] 0.02704874
```
Under the normality assumption, the t-statistic
\[ t = \frac{b - \beta}{SE(b)} \]
follows a t-distribution, with (n-2) degrees of freedom.
Then, 95% confidence interval for \( \beta \) is:
\[ b \pm t_{0.025, \infty} \times SE(b) = 0.51409 \pm 1.962 \times 0.02705 = (0.461, 0.567) \]

(g). Calculate the coefficient determination \( R^2 \). What does the \( R^2 \) statistic mean?

**Sol’n.** The coefficient of determination \( R^2 \) is calculated as
\[ R^2 = \frac{\text{Reg SS}}{\text{Total SS}} = \frac{\sum (\hat{y} - \bar{y})^2}{\sum (y_i - \bar{y})^2} \]

In R:
```r
> R.squared <- sum((fitted(m) - mean(sheight))^2) / sum((sheight - mean(sheight))^2)
> R.squared
[1] 0.2513401
```
Here the \( R^2 \) statistic is the proportion of the total response variation explained by the explanatory variable in the linear regression model.

(h). Draw a residual plot. Are the residuals normally distributed with constant variance?

**Sol’n.** The model assumptions of normal distribution and constant variance seem valid based on the residual plot below.

![Residual Plot](image-url)
(i). What are the estimated means of son's height given that his father's height is 72, 75, 60, and 63 inches, respectively?

(Notice that sons of tall fathers tended to be tall, but on average not as tall as their fathers. Similarly, sons of short fathers tended to be short, but on average not as short as their fathers. This phenomenon was first described by Sir Francis Galton, as "regression towards mediocrity", where the term regression came from. The regression effect – phenomenon of regression toward the mean – appears in any test-retest situation.)

**Sol’n.** The estimated means of son's heights are 70.9, 72.4, 64.7, 66.3 inches, given that his father's height is 72, 75, 60, and 63 inches, respectively.

(j). Given a father’s height, we can use simulation method to construct the 100(1-\(\alpha\))% confidence interval for the mean of his son’s height. First draw 1000 samples each of size 1078 with replacement from the 1078 pairs of father-son heights, then from each sample fit a linear regression model by the method of least squares, and compute the estimated mean of son’s height.

What are the mean and standard deviation of these 1000 simulated values?

Sort these 1000 estimated means in ascending order. Denote the 25th largest as \(h_{25}\) and the 975th largest as \(h_{975}\), which are our estimates of the 0.025 and 0.975 quantiles of the sampling distribution for the mean of son’s height. Then the 100(1-\(\alpha\))% confidence interval for the mean of the son’s height is \((h_{25}, h_{975})\). Compute the 95% confidence interval for the mean of son’s height if his father is 72 inches tall.

**Sol’n.** Use the following loop to run 1000 simulations in R:

```r
> n <- 1078
> h.father <- 72
> h.son <- rep(NA, 1000)
> for (i in 1:1000) {
+ v <- sample(1:n, n, replace=TRUE)
+ fheight.sim <- fheight[v]
+ sheight.sim <- sheight[v]
+ b.sim <- cov(fheight.sim, sheight.sim) / var(fheight.sim)
+ a.sim <- mean(sheight.sim) - b.sim * mean(fheight.sim)
+ h.son[i] <- a.sim + h.father * b.sim
+ }
> mean(h.son) # mean of the 1000 simulated values
[1] 70.89778
> sd(h.son) # standard deviation of the 1000 simulated values
[1] 0.1346589
> h.son.sort <- sort(h.son)
> h.son.sort[25] # the estimated 0.025 quantile
[1] 70.63136
> h.son.sort[975] # the estimated 0.975 quantile
[1] 71.1718
```

Thus, the estimated 95% confidence interval for the mean of son’s height is (70.63, 71.17) inches, given that the father’s height is 72 inches.
Contingency Table

2. In an investigation of the association between smoking habit and lung cancer, lung cancer patients and controls were obtained. The patients and controls were matched for age, sex, and community. The data are shown in the table below. (Data from P. Notani and L. D. Sanghvi, “A Retrospective Study of Lung Cancer in Bomby”, Br. J. Cancer 29(6): 477-482, 1974.)

<table>
<thead>
<tr>
<th></th>
<th>Lung Cancer</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smokers</td>
<td>413</td>
<td>318</td>
</tr>
<tr>
<td>Nonsmokers</td>
<td>107</td>
<td>201</td>
</tr>
</tbody>
</table>

(a). What is the type of this study in terms of study design?

**Sol’n.** *This is a case-control (or retrospective) study.*

(b). Calculate the odds of lung cancer for smokers, the odds of lung cancer for nonsmokers, and the ratio of two odds.

**Sol’n.** The odds of lung cancer for smokers are 413/318 = 1.299, and the odds of lung cancer for nonsmokers are 107/201 = 0.532. The ratio of two odds is 1.299/0.532 = 2.44; that is, the odds of having lung cancer for smokers are estimated to be 2.44 times as large as the odds of having lung cancer for nonsmokers.

(c). Calculate the 95% confidence interval for the odds ratio.

**Sol’n.** To get a confidence interval for the odds ratio, construct a confidence interval for the log of the odds ratio and take the antilogarithm of the endpoints.

The log of the estimated odds ratio is \( \ln(\hat{OR}) = \ln(2.44) = 0.89 \).

The variance of the log odds ratio is estimated as

\[
\text{Var}[\ln(\hat{OR})] = \frac{1}{413} + \frac{1}{318} + \frac{1}{107} + \frac{1}{201} \approx 0.0199
\]

95% confidence interval for the log odds ratio is

\[
0.89 \pm 1.96 \times \sqrt{0.0199} = 0.61 \text{ to } 1.17
\]

95% confidence interval for the odds ratio is

\[
\exp(0.61) \text{ to } \exp(1.67); \text{ or } 1.84 \text{ to } 3.22
\]

3. The Salk polio vaccine trials of 1954 included a double-blind experiment in which elementary school children of consenting parents were assigned at random to injection with the Salk vaccine of with a placebo. Both treatment and control groups were set at 200,000 because the target disease, infantile paralysis, was uncommon (but greatly feared). (Data from J. M. Tanur et al., *Statistics: A Guide to the Unknown*, San Francisco: Holden-Day, 1972.)

<table>
<thead>
<tr>
<th>Infantile paralysis victim?</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Placebo</td>
<td>142</td>
<td>199,858</td>
</tr>
<tr>
<td>Salk polio vaccine</td>
<td>56</td>
<td>199,944</td>
</tr>
</tbody>
</table>

(a). Is this a randomized experiment or a cohort study?

**Sol’n.** *This is a randomized experiment.*
(b). Calculate the proportion of infantile paralysis victims among placebo group, and the proportion of infantile paralysis victims among vaccine group, respectively.

**Sol’n.** The proportion of infantile paralysis victims among placebo group is 
\[ \hat{p}_1 = \frac{142}{200000} = 0.00071, \]
and
the proportion of infantile paralysis victims among vaccine group is 
\[ \hat{p}_2 = \frac{56}{200000} = 0.00028. \]

(c). What is the risk difference? Calculate the 95% confidence interval for the risk difference.

**Sol’n.** The risk difference is 
\[ \hat{p}_1 - \hat{p}_2 = 0.00071 - 0.00028 = 0.00043. \]

To construct the confidence interval, first we need to estimate the standard error of risk difference:
\[
SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = \sqrt{\frac{0.00071(1-0.00071)}{200000} + \frac{0.00028(1-0.00028)}{200000}} = 0.0000703
\]

Then the 95% confidence interval is
\[ 0.00043 \pm 1.96 \times 0.0000703 = 0.00029 \text{ to } 0.00057. \]

(d). What is the relative risk? Calculate the 95% confidence interval for the relative risk.

**Sol’n.** The relative risk is
\[ RR = \frac{\hat{p}_1}{\hat{p}_2} = \frac{0.00071}{0.00028} = 2.536. \]

To get a confidence interval for the relative risk, we need to construct a confidence interval for the log of the relative risk and then take the antilogarithm of the endpoints.

The log of the estimated relative risk is 
\[ \ln(\hat{R}R) = \ln(2.536) = 0.93. \]

The variance of the log relative risk is estimated as
\[
Var[\ln(\hat{R}R)] = \frac{199858}{142 \times 200000} + \frac{199944}{56 \times 200000} = 0.025
\]

95% confidence interval for the log relative risk is
\[ 0.93 \pm 1.96 \times \sqrt{0.025} = 0.62 \text{ to } 1.24 \]

95% confidence interval for the relative risk is
\[ \exp(0.62) \text{ to } \exp(1.24) = 1.86 \text{ to } 3.46. \]