Homework 4 - P8130 Biostatistics Method I

Yongyan Liu (yl6107)

2025-11-25

Problem 1 (10 points)

A new device has been developed which allows patients to evaluate their blood sugar levels. The most widely device currently on the market yields widely variable results. The new device is evaluated by 25 patients having nearly the same distribution of blood sugar levels yielding the following data:

125, 123, 117, 123, 115, 112, 128, 118, 124, 111, 116, 109, 125, 120, 113, 123, 112, 118, 121, 118, 122, 115, 105, 118, 131

Part (a)

Is there significant ($\alpha = 0.05$) evidence that median blood sugar readings was less than 120 in the population from which the 25 patients were selected? Use the sign test and report the test statistic and p-value.

Solution:

Hypotheses:

- $H_0 : M = 120$ (median blood sugar = 120)
- $H_a : M < 120$ (median blood sugar < 120)

Summary:

- Number above 120 (n_+): 10
- Number below 120 (n_-): 14
- Number equal to 120 (ties): 1
- Sample size excluding ties: $n = 10 + 14 = 24$

Under H_0, the number of positive signs follows $B \sim \text{Binomial}(n = 24, p = 0.5)$.

For a one-sided test ($H_a : M < 120$), we calculate:

$$p\text{-value} = P(B \leq 10) = \sum_{k=0}^{10} \binom{24}{k} (0.5)^{24}$$

```
blood_sugar <- c(125, 123, 117, 123, 115, 112, 128, 118, 124, 111, 116, 109, 125,
                 120, 113, 123, 112, 118, 121, 118, 122, 115, 105, 118, 131)

n_plus <- sum(blood_sugar > 120)
n_minus <- sum(blood_sugar < 120)
```

```r
n <- n_plus + n_minus

p_value_sign <- pbinom(n_plus, size = n, prob = 0.5)
p_value_sign
```

```
## [1] 0.2706281
```

Test Statistic: $B = 10$ (number of positive signs)

P-value: $P(B \leq 10) = 0.2706$

Since p-value $= 0.2706 > 0.05$, we fail to reject H_0. There is insufficient evidence that the median blood sugar reading is less than 120.

Part (b)

Is there significant ($\alpha = 0.05$) evidence that median blood sugar readings was less than 120 in the population from which the 25 patients were selected? Use the Wilcoxon signed-rank test and report the test statistic and p-value.

Solution:

Hypotheses:

- $H_0 : M = 120$
- $H_a : M < 120$

```r
wilcox_result <- wilcox.test(blood_sugar, mu = 120, alternative = "less", exact = FALSE)
wilcox_result
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  blood_sugar
## V = 112.5, p-value = 0.1447
## alternative hypothesis: true location is less than 120
```

Test Statistic: $T^+ = V = 112.5$

P-value: 0.1447

Since p-value $= 0.1447 > 0.05$, we fail to reject H_0. There is insufficient evidence that the median blood sugar reading is less than 120.

Problem 2 (10 points)

Human brains have a large frontal cortex with excessive metabolic demands compared with the brains of other primates. However, the human brain is also three or more times the size of the brains of other primates. Is it possible that the metabolic demands of the human frontal cortex are just an expected consequence of greater brain size? For this problem, use the provided data file entitled "Brain data".

```r
brain_data <- read_excel("Brain data.xlsx")
brain_data$`Brain mass (g)` <- as.numeric(brain_data$`Brain mass (g)`)
nonhuman <- brain_data[brain_data$Species != "Homo sapiens", ]
human <- brain_data[brain_data$Species == "Homo sapiens", ]
```

Part (a)

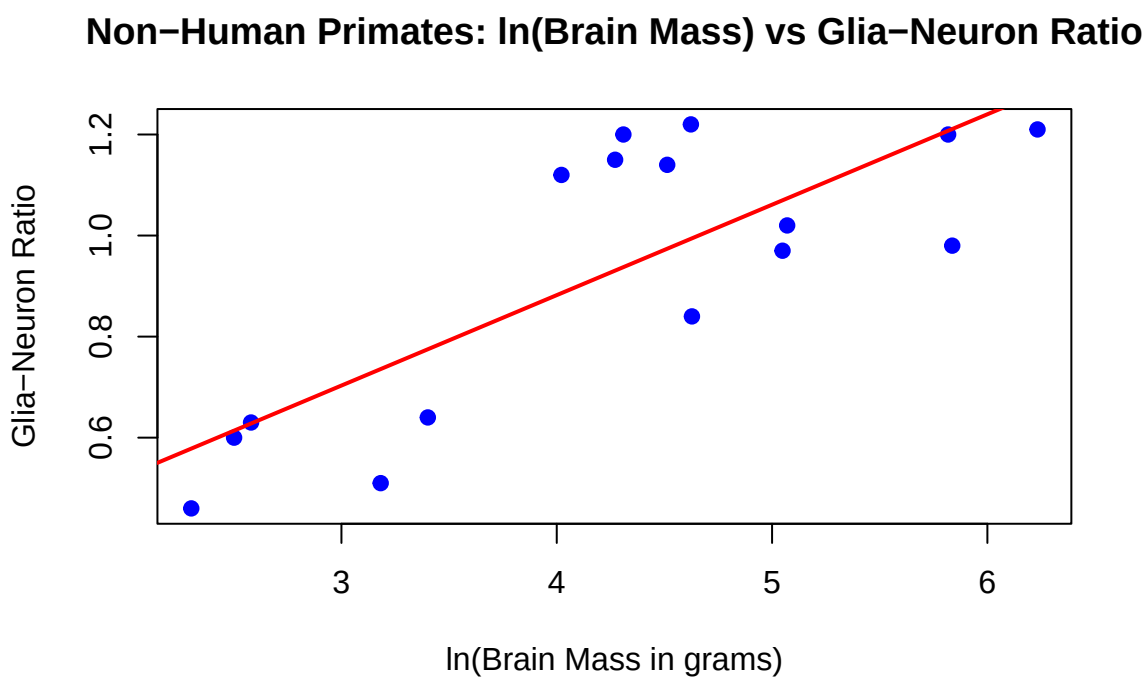
Using only the non-human data, make a scatterplot of (natural) log of brain mass on the X-axis and glia-neuron ratio as outcome. Then fit the corresponding regression model and write an expression for the fitted regression line.

Solution:

```r
nonhuman$log_mass <- log(nonhuman$`Brain mass (g)`)

plot(nonhuman$log_mass, nonhuman$`Glia-neuron ratio`,
     xlab = "ln(Brain Mass in grams)", ylab = "Glia-Neuron Ratio",
     main = "Non-Human Primates: ln(Brain Mass) vs Glia-Neuron Ratio",
     pch = 19, col = "blue")

model <- lm(`Glia-neuron ratio` ~ log_mass, data = nonhuman)
abline(model, col = "red", lwd = 2)
```



Regression Coefficient Formulas:

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}, \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

```r
coef_model <- coef(model)
coef_model
```

```
## (Intercept)    log_mass
##   0.1662434   0.1789595
```

Fitted Regression Equation:

$$\widehat{\text{Glia-Neuron}} \text{ Ratio} = 0.1662 + 0.179 \times \ln(\text{Brain Mass})$$

Part (b)

Using the nonhuman primate relationship, what is the predicted glia-neuron ratio for humans, given their brain mass?

Solution:

Human brain mass: 1373.3 grams

$$\ln(\text{Human Brain Mass}) = \ln(1373.3) = 7.225$$

Predicted Glia-Neuron Ratio:

$$\hat{Y} = 0.1662 + 0.179 \times 7.225 = 1.4592$$

```r
human_log_mass <- log(human$`Brain mass (g)`)
predicted <- predict(model, newdata = data.frame(log_mass = human_log_mass))
actual <- human$`Glia-neuron ratio`
```

- **Predicted:** 1.4592
- **Actual:** 1.65
- **Difference (Actual - Predicted):** 0.1908

Part (c)

Construct a 95% prediction interval corresponding to the prediction made in part (b). Based on this, does the human brain have an excessive glia-neuron ratio for its mass compared with other primates?

Solution:

The 95% prediction interval formula:

$$\hat{Y}_0 \pm t_{n-2,0.975} \times s_e \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{XX}}}$$

```r
pred_int <- predict(model, newdata = data.frame(log_mass = human_log_mass),
                    interval = "prediction", level = 0.95)
pred_int
```

```
##        fit      lwr      upr
## 1 1.459221 1.005942 1.912499
```

95% Prediction Interval: [1.0059, 1.9125]

Conclusion: The actual human glia-neuron ratio (1.65) falls **within** the 95% prediction interval [1.0059, 1.9125]. Based on this interval, there is **not enough evidence** to conclude that the human brain has an excessive glia-neuron ratio for its mass compared to other primates.

Part (d)

Is there anything to be cautious about when using the non-human data to make predictions for humans? Explain your answer.

Solution:

```
range_mass <- range(nonhuman$`Brain mass (g)`)
range_mass
```

```
## [1]  10.0 509.2
```

Cautions:

1. The human brain mass (1373.3 g) is far outside the range of non-human primate brain masses (10 to 509.2 g). Predictions outside the observed data range are less reliable.
2. We assume the linear relationship between log(brain mass) and glia-neuron ratio continues beyond the observed range, which may not hold.
3. With only 16 non-human observations, there is considerable uncertainty in the regression estimates.

Problem 3 (20 points)

For this problem, you will be using data HeartDisease.csv. The investigator is mainly interested if there is an association between total cost (in dollars) of patients diagnosed with heart disease and the number of emergency room (ER) visits. The model may need to be adjusted for other factors, including age, gender, number of complications that arose during treatment, and duration of treatment condition.

```
heart <- read.csv("HeartDisease.csv")
```

Part (a)

Generate appropriate descriptive statistics for all variables of interest (continuous and categorical).

Solution:

1. Continuous Variables: totalcost, age, duration

```
summary(heart[, c("totalcost", "age", "duration")])
```

```
##    totalcost            age           duration
##  Min.   :    0.0   Min.   :24.00   Min.   :  0.00
##  1st Qu.:  161.1   1st Qu.:55.00   1st Qu.: 41.75
##  Median :  507.2   Median :60.00   Median :165.50
##  Mean   : 2800.0   Mean   :58.72   Mean   :164.03
##  3rd Qu.: 1905.5   3rd Qu.:64.00   3rd Qu.:281.00
##  Max.   :52664.9   Max.   :70.00   Max.   :372.00
```

- **totalcost:** Ranges from $0 to $52,664.90 with median $507.2 and mean $2800. The large difference between mean and median suggests right-skewness.
- **age:** Patients range from 24 to 70 years old, with mean age 58.7 years.
- **duration:** Treatment duration ranges from 0 to 372 days, with mean 164 days.

2. Discrete Variables (>10 unique values): ERvisits, interventions, comorbidities

```
# ERvisits (17 unique values)
cat("ERvisits - Range:", range(heart$ERvisits), " Mean:", round(mean(heart$ERvisits), 2), "\n")
```

```
## ERvisits - Range: 0 20  Mean: 3.43
```

```
# Interventions (32 unique values)
cat("Interventions - Range:", range(heart$interventions), " Mean:", round(mean(heart$interventions), 2)
```

```
## Interventions - Range: 0 47  Mean: 4.71
```

```
# Comorbidities (32 unique values)
cat("Comorbidities - Range:", range(heart$comorbidities), " Mean:", round(mean(heart$comorbidities), 2)
```

```
## Comorbidities - Range: 0 60  Mean: 3.77
```

- **ERvisits:** Patients had between 0 and 20 ER visits, with an average of 3.43 visits.
- **interventions:** Number of interventions ranged from 0 to 47, averaging 4.71.
- **comorbidities:** Comorbid conditions ranged from 0 to 60, with mean 3.77.

3. Discrete Variables (<10 unique values): drugs, complications

```
# Drugs (9 unique values)
table(heart$drugs)
```

```
##
##   0   1   2   3   4   5   6   7   9
## 610  89  49  19   9   5   4   2   1
```

```
# Complications (3 unique values)
table(heart$complications)
```

```
##
##   0   1   3
## 745  42   1
```

- **drugs:** Most patients received 0 drugs (610 patients, 77.4%), with a maximum of 9 drugs.
- **complications:** The majority had no complications (745 patients, 94.5%). Only 43 patients had 1 or more complications.

4. Binary Variable: gender

```r
table(heart$gender)
```

```
##
##   0   1
## 608 180
```

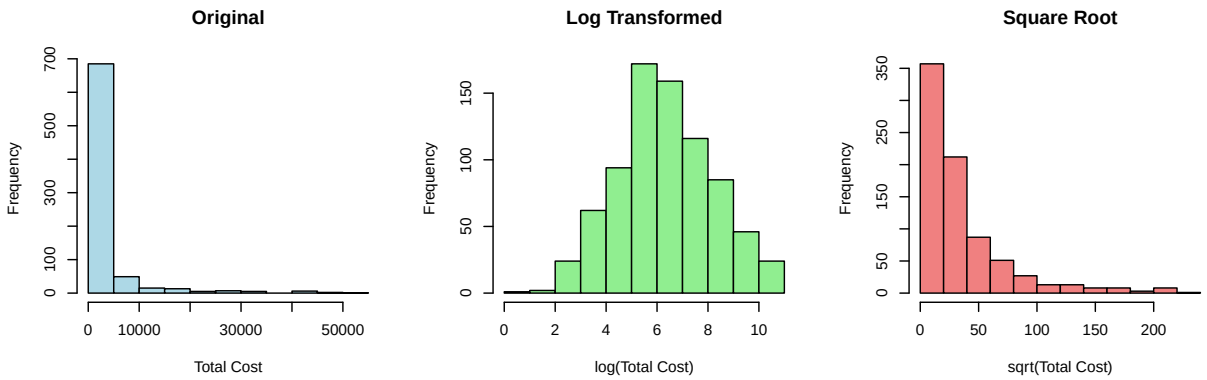
- **gender:** 608 patients (77.2%) are coded as 0, and 180 patients (22.8%) are coded as 1.

Part (b)

Investigate the shape of the distribution for variable totalcost and try different transformations, if necessary.

Solution:

```r
par(mfrow = c(1, 3))
hist(heart$totalcost, main = "Original", xlab = "Total Cost", col = "lightblue")
hist(log(heart$totalcost), main = "Log Transformed",
     xlab = "log(Total Cost)", col = "lightgreen")
hist(sqrt(heart$totalcost), main = "Square Root", xlab = "sqrt(Total Cost)", col = "lightcoral")
```



We should use **log transformation** because:

1. The original distribution is highly right-skewed
2. Log transformation produces a more symmetric, approximately normal distribution

```r
heart$log_cost <- ifelse(heart$totalcost > 0, log(heart$totalcost), NA)
```

Part (c)

Create a new variable called comp_bin by dichotomizing the complications variable: 0 if no complications, and 1 otherwise.

Solution:

$$\text{comp\_bin} = \begin{cases} 0 & \text{if complications} = 0 \\ 1 & \text{if complications} \geq 1 \end{cases}$$

```r
heart$comp_bin <- ifelse(heart$complications == 0, 0, 1)
table(heart$comp_bin)
```

```
##
##   0   1
## 745  43
```

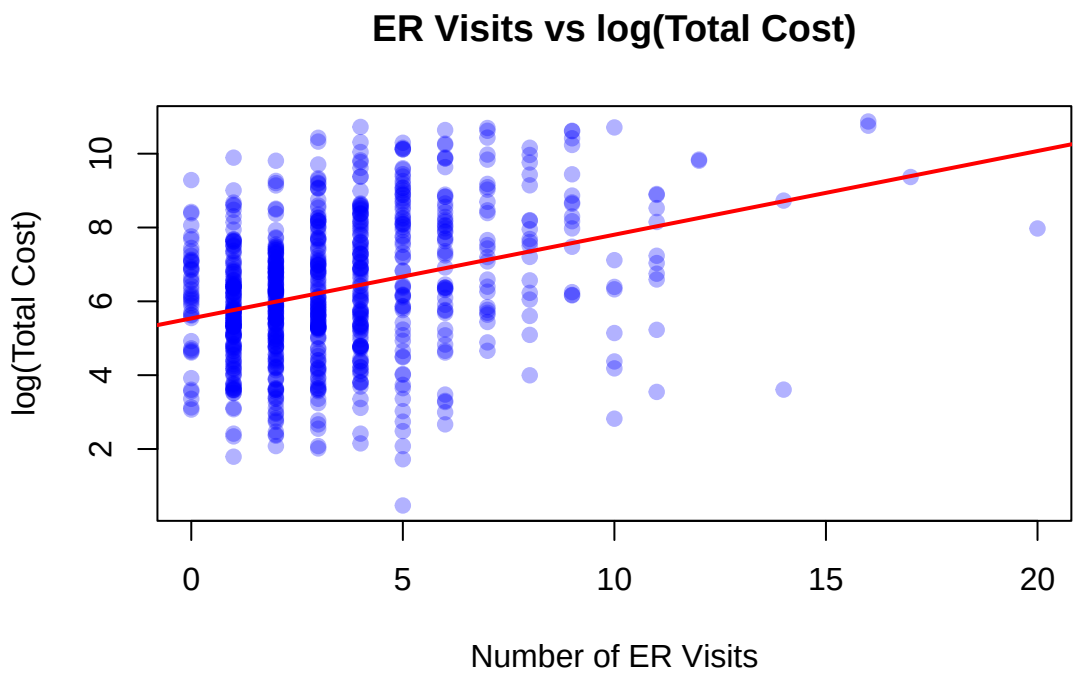
Part (d)

Fit a simple linear regression (SLR) model with totalcost (original or transformed, based on your decision in part (b)) as the outcome variable and ERvisits as predictor. This should include a scatterplot and results of the regression analysis, with appropriate comments on significance and interpretation of the slope estimate.

Solution:

```r
plot(heart$ERvisits, heart$log_cost,
     xlab = "Number of ER Visits", ylab = "log(Total Cost)",
     main = "ER Visits vs log(Total Cost)", pch = 19, col = rgb(0,0,1,0.3))

slr <- lm(log_cost ~ ERvisits, data = heart)
abline(slr, col = "red", lwd = 2)
```




```r
summary(slr)
```

```
##
## Call:
## lm(formula = log_cost ~ ERvisits, data = heart)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.2013 -1.1265  0.0191  1.2668  4.2797
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.53771    0.10362   53.44   <2e-16 ***
## ERvisits     0.22672    0.02397    9.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.772 on 783 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.1026, Adjusted R-squared:  0.1014
## F-statistic:  89.5 on 1 and 783 DF,  p-value: < 2.2e-16
```

Fitted Regression Equation:

$$\log(\text{Total Cost}) = 5.5377 + 0.2267 \times \text{ERvisits}$$

- **Intercept** ($\hat{\beta}_0 = 5.5377$): Expected log(total cost) when ERvisits = 0
- **Slope** ($\hat{\beta}_1 = 0.2267$): For each additional ER visit, log(total cost) increases by 0.2267 on average

Significance: ERvisits is highly significant ($p < 0.001$). The positive slope indicates more ER visits are associated with higher costs.

Part (e)

Fit a multiple linear regression (MLR) with comp_bin and ERvisits as predictors.

(I) Test for Interaction

Test for an interaction between comp_bin and ERvisits. Give your conclusions and interpret the results.

Solution:

```r
model_int <- lm(log_cost ~ ERvisits * comp_bin, data = heart)

model_int
```

```
##
## Call:
## lm(formula = log_cost ~ ERvisits * comp_bin, data = heart)
##
## Coefficients:
##       (Intercept)           ERvisits           comp_bin  ERvisits:comp_bin
##           5.49899            0.21125            2.17969           -0.09927
```

```r
summary(model_int)$coefficients
```

```
##                       Estimate Std. Error   t value      Pr(>|t|)
## (Intercept)        5.49898897 0.10348511 53.137971 1.360876e-261
## ERvisits           0.21124855 0.02453446  8.610281  3.991332e-17
## comp_bin           2.17969246 0.54603655  3.991843  7.174153e-05
## ERvisits:comp_bin -0.09926527 0.09483207 -1.046748  2.955397e-01
```

Interaction term: $\hat{\beta}_3 = -0.0993$, p-value $= 0.2955$

The interaction is **not significant** ($p > 0.05$). The effect of ER visits on log(total cost) does not differ between patients with and without complications.

(II) Test for Confounding

Test whether comp_bin is a confounder of the relationship between totalcost and ERvisits. Give your conclusions and interpret the results.

Solution:

A variable is a confounder if adjusting for it changes the coefficient by more than 10%.

```r
model_adj <- lm(log_cost ~ ERvisits + comp_bin, data = heart)
beta_crude <- coef(slr)["ERvisits"]
beta_adj <- coef(model_adj)["ERvisits"]
pct_change <- abs((beta_adj - beta_crude) / beta_crude) * 100
```

- Crude $\hat{\beta}_{\text{ERvisits}}$: 0.2267
- Adjusted $\hat{\beta}_{\text{ERvisits}}$: 0.2046
- Percent change: $\frac{|0.2046-0.2267|}{|0.2267|} \times 100\% = 9.76\%$

comp_bin is **not a confounder** (change $= 9.76\% < 10\%$).

(III) Should comp_bin Be Included?

Should comp_bin be included as a covariate in the model (along with ERvisits)? Explain your reasoning.

Solution:

```r
summary(model_adj)$coefficients["comp_bin", ]
```

```
##     Estimate   Std. Error      t value     Pr(>|t|)
## 1.685863e+00 2.749421e-01 6.131699e+00 1.378861e-09
```

Reasoning:

1. Interaction: Not significant
2. Confounding: $< 10\%$ change - not a confounder
3. Statistical significance: comp_bin is significant ($p < 0.05$)
4. Clinical relevance: Complications are clinically meaningful

comp_bin should be included because it is statistically significant and clinically important.

Part (f)

Use your choice of model in part (e) and add additional covariates (age, gender, and duration of treatment).

(I) Fit MLR with All Covariates

Fit a MLR, provide the fitted regression equation, and give the interpretation of each estimated parameter. Which variables seem to be significant?

Solution:

```
mlr <- lm(log_cost ~ ERvisits + comp_bin + age + gender + duration, data = heart)
summary(mlr)
```

```
##
## Call:
## lm(formula = log_cost ~ ERvisits + comp_bin + age + gender +
##     duration, data = heart)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0823 -1.0555 -0.1352  0.9533  4.3462
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.0449619  0.5063454  11.938  < 2e-16 ***
## ERvisits     0.1757486  0.0223189   7.874 1.15e-14 ***
## comp_bin     1.4921110  0.2554883   5.840 7.65e-09 ***
## age         -0.0221376  0.0086023  -2.573   0.0103 *
## gender      -0.1176181  0.1379809  -0.852   0.3942
## duration     0.0055406  0.0004848  11.428  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.605 on 779 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.268,  Adjusted R-squared:  0.2633
## F-statistic: 57.03 on 5 and 779 DF,  p-value: < 2.2e-16
```

Fitted Regression Equation:

$$\log(\text{Cost}) = 6.04 + (0.1757 * \text{ERvisits}) + (1.4921 * \text{comp\_bin}) + (-0.0221 * \text{age}) + (-0.1176 * \text{gender}) + (0.0055 * \text{duration})$$

Variable	Estimate	Interpretation (holding others constant)
ERvisits	0.1757	Each ER visit increases log(cost) by 0.1757
comp_bin	1.4921	Having complications increases log(cost) by 1.4921
age	-0.0221	Each year of age changes log(cost) by -0.0221
gender	-0.1176	Gender effect on log(cost)
duration	0.0055	Each unit of duration increases log(cost) by 0.0055

Significant Variables ($p < 0.05$): ERvisits, comp_bin, duration, age

(II) Compare SLR and MLR Models

Compare the SLR and MLR models using the appropriate testing procedure. Which model would you use to address the investigator's objective and why?

Solution:

Partial F-test formula:

$$F = \frac{(SSE_{\text{reduced}} - SSE_{\text{full}})/(df_{\text{reduced}} - df_{\text{full}})}{SSE_{\text{full}}/df_{\text{full}}}$$

```
anova_result <- anova(slr, mlr)
anova_result
```

```
## Analysis of Variance Table
##
## Model 1: log_cost ~ ERvisits
## Model 2: log_cost ~ ERvisits + comp_bin + age + gender + duration
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    783 2459.8
## 2    779 2006.5  4     453.3 43.996 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model Comparison:

Model	R^2	Adjusted R^2
SLR	0.1026	0.1014
MLR	0.268	0.2633

F-statistic: 44, **P-value:** < 0.001

We should use the **MLR model** because:

1. Partial F-test is significant ($p < 0.001$) - additional covariates improve fit
2. Higher R^2 (0.268 vs 0.1026)
3. Controls for confounders (age, gender, duration, complications)
4. Provides more accurate estimate of ERvisits effect on total cost