Regression Prelim

2024-05-22

Time: 90 minutes

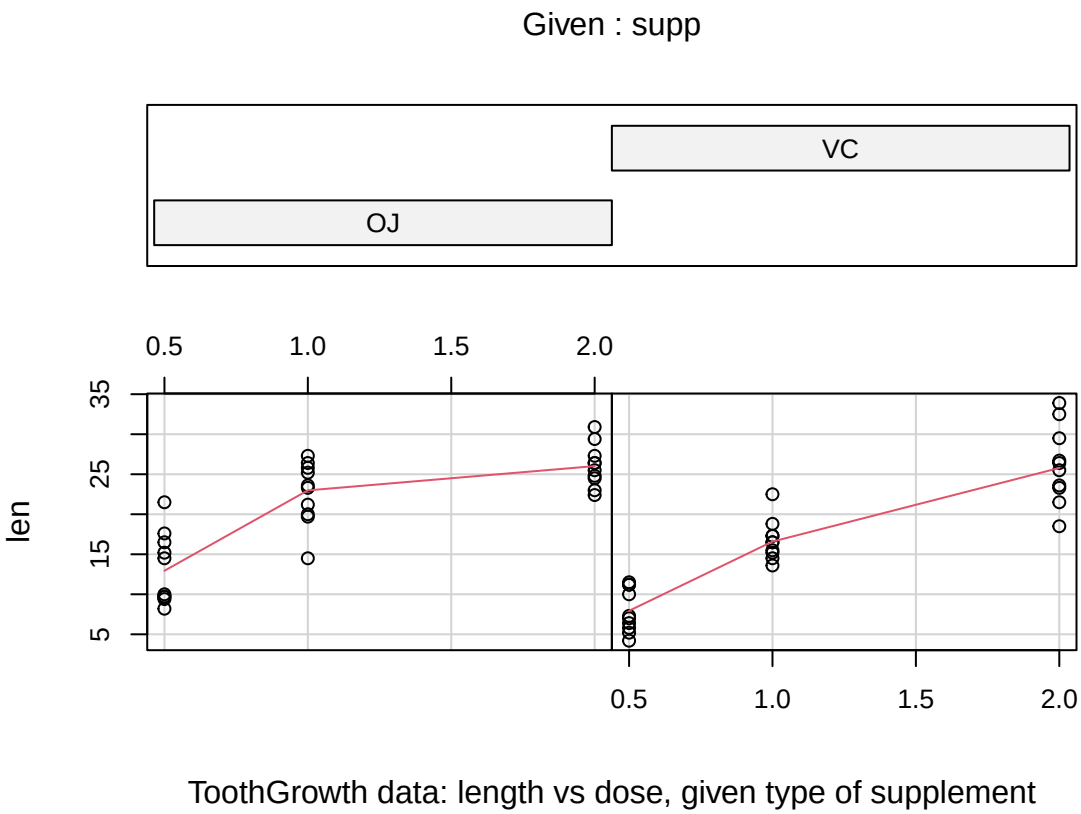
Total possible points: 60

Question 1 (35 pts)

The question concerns the dataset of "The Effect of Vitamin C on Tooth Growth in Guinea Pigs Description".

The response is the length of odontoblasts (cells responsible for tooth growth) in 60 guinea pigs. Each animal received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day) by one of two delivery methods, orange juice or ascorbic acid (a form of vitamin C and coded as VC).

```
library(graphics)
coplot(len ~ dose | supp, data = ToothGrowth, panel = panel.smooth,
       xlab = "ToothGrowth data: length vs dose, given type of supplement")
```



```r
X <- model.matrix(len ~ dose + supp, ToothGrowth)
rct_lm_result <- lm(len ~ dose + supp, ToothGrowth)
rct_lm_summary <- summary(rct_lm_result)
Y <- matrix(ToothGrowth$len, ncol = 1)
n <- nrow(ToothGrowth)
```

```r
# showing the design matrix X
tibble::as_tibble(X)
```

```
## # A tibble: 60 x 3
##    `(Intercept)`  dose suppVC
##            <dbl> <dbl>  <dbl>
##  1             1   0.5      1
##  2             1   0.5      1
##  3             1   0.5      1
##  4             1   0.5      1
##  5             1   0.5      1
##  6             1   0.5      1
##  7             1   0.5      1
##  8             1   0.5      1
##  9             1   0.5      1
## 10             1   0.5      1
## # i 50 more rows
```

```r
# showing the response
tibble::as_tibble(Y)
```

```
## # A tibble: 60 x 1
##       V1
##    <dbl>
##  1   4.2
##  2  11.5
##  3   7.3
##  4   5.8
##  5   6.4
##  6  10
##  7  11.2
##  8  11.2
##  9   5.2
## 10   7
## # i 50 more rows
```

```r
# showing the summary of linear regression model
rct_lm_summary
```

```
##
## Call:
## lm(formula = len ~ dose + supp, data = ToothGrowth)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.600 -3.700  0.373  2.116  8.800
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.2725     1.2824   7.231 1.31e-09 ***
## dose          9.7636     0.8768  11.135 6.31e-16 ***
## suppVC       -3.7000     1.0936  -3.383   0.0013 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.236 on 57 degrees of freedom
## Multiple R-squared:  0.7038, Adjusted R-squared:  0.6934
## F-statistic: 67.72 on 2 and 57 DF,  p-value: 8.716e-16
```

```r
# showing ANOVA table
anova(rct_lm_result)
```

```
## Analysis of Variance Table
##
## Response: len
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## dose       1 2224.30 2224.30 123.989 6.314e-16 ***
## supp       1  205.35  205.35  11.447  0.001301 **
## Residuals 57 1022.56   17.94
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(lm(len ~ supp, ToothGrowth))
```

```
## Analysis of Variance Table
##
## Response: len
##           Df Sum Sq Mean Sq F value  Pr(>F)
## supp       1  205.4  205.35  3.6683 0.06039 .
## Residuals 58 3246.9   55.98
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# (X'X)^{-1}X'Y
beta <- solve(t(X) %*% X) %*% t(X) %*% Y
beta
```

```
##                  [,1]
## (Intercept)  9.272500
## dose         9.763571
## suppVC      -3.700000
```

```r
# b'X'Y - 1/n Y'JY = Y'(H - 1/n J)Y
# J is the matrix of all ones: J = matrix(1, nrow=n, ncol=n)
t(beta) %*% t(X) %*% Y - 1 / n * t(Y) %*% matrix(1, nrow=n, ncol=n) %*% Y
```

```
##          [,1]
## [1,] 2429.654
```

```r
# H = X(X'X)^{-1}X'
H <- X %*% solve(t(X) %*% X) %*% t(X)
```

```r
# SSE = e'e = Y'Y - b'X'Y = Y'(I - H)Y
SSE <- t(Y) %*% (diag(n) - H) %*% Y
SSE
```

```
##          [,1]
## [1,] 1022.555
```

```r
# s^2{e} = MSE (I - H)
# Get the F^* test statistic
# F_star <- (SSR / 1) / (SSE / (n - 3))
```

```r
# s^2{b} = MSE(X'X)^{-1}
# where MSE = SSE / (n - p), and we have 3 parameters in this case:
s_b_squared <- as.numeric(SSE / (n - 3)) * solve(t(X) %*% X)
s_b_squared
```

```
##             (Intercept)       dose     suppVC
## (Intercept)   1.6444599 -0.8969781 -0.5979854
## dose         -0.8969781  0.7688384  0.0000000
## suppVC       -0.5979854  0.0000000  1.1959708
```

In the question, the regression model we refer to is the full model above.

a. State the model assumption. Give the estimated regression function.

b. Obtain a 95% percent interval estimate of the mean length of odontoblasts whose a dose of 1 mg/day vitamin C and orange juice delivery. Interpret your confidence interval.

c. The investigators decided to add to the experiment another guinea pig who will be given a dose of 1 mg/day vitamin C and orange juice delivery. Predict its length of odontoblasts using a 95% prediction interval. Interpret your prediction interval.

d. Give the SSR and the SSE.

e. Give the proportion of variance explained by the model. What is it called in the printed results of R?

f. We consider extra sums of squares. Please set up the ANOVA table. In the table, the columns are Source of Variation; SS (Sum of Squares); df (degree of freedom); and MS (Mean Square). The Source of Variation include Reg (Regression); X_1; $X_2|X_1$; Error; and Total. Here we use X_1 to denote `dose` and X_2 to denote `supp`.

g. Conduct an F test to determine whether X_2 (`supp`) can be dropped from the regression model given that X_1 (`dose`) is retained. Use $\alpha = 0.05$. State the alternatives, decision rule, and conclusion. Explain the results to non-statisticians that are interested in the scientific question.

Question 2 (10 pts)

Consider the one-way ANOVA model:

$$X_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad j = 1, \ldots, n_i, \, i = 1, \ldots, m,$$

where μ and α_i are unknown parameters and ϵ_{ij} are independent and identically distributed random variables with mean 0. Let

$$X = (X_{11}, \ldots, X_{1n_1}, \ldots, X_{m1}, \ldots, X_{mn_m})^T,$$

$$\epsilon = (\epsilon_{11}, \ldots, \epsilon_{1n_1}, \ldots, \epsilon_{mn_m})^T,$$

and $\beta = (\mu, \alpha_1, \ldots, \alpha_m)^T$.

a. Provide the matrix X.

b. Provide the matrix $X^T X$, and describe the linear space generated by the columns of $X^T X$ by giving the form of the vectors, $l = (l_0, l_1, \ldots, l_m)^T \in \mathcal{R}^{m+1}$, in the linear space.

Question 3 (15 pts)

We consider the model $Y_{ij} = (x_{ij} - \bar{x_i})\gamma_i + \epsilon_{ij}$, $i = 1, 2$, $j = 1, \ldots, n_i$, where $n_i > 0$, $\bar{x_i} \equiv (1/n_i)\sum_j x_{ij}$, and μ_i and γ_i are scalar parameters. Suppose that x_{ij} are known scalars which are not all equal for each $i = 1, 2$. Further, suppose that $\{e_{ij},\, i = 1, 2,\, j = 1, \ldots, n_i\}$ are assumed to be independent and identically distributed such that $\epsilon_{ij} \sim N(0, \sigma^2)$, where σ^2 is a scalar parameter.

a. Let $\beta = (\mu_1, \gamma_1, \mu_2, \gamma_2)^T$. We wish to write this model as $Y = X\beta + \epsilon$, where $Y = (Y_{11}, \ldots, Y_{1n_1}, Y_{21}, \ldots, Y_{2n_2})^T$, $\epsilon = (\epsilon_{11}, \cdots, \epsilon_{1n_1}, \epsilon_{21}, \cdots, \epsilon_{2n_2})^T$, and X is an appropriately defined matrix.

   (i) Provide the matrix X.

   (ii) Provide $\hat{\beta}$, which is the least squares estimate of β.

b. (i) Specify a column vector a such that $a^T\beta = (\gamma_1 - \gamma_2)$.

   (ii) Suppose σ^2 is known. Derive the distribution of $a^T\hat{\beta}$ and give a $(1-\alpha)$-level confidence interval for $\gamma_1 - \gamma_2$.

   (iii) Suppose σ^2 is unknown. Give a statistic to test $H_0 : \gamma_1 = \gamma_2$ and indicate its distribution under H_0.

Answer to Question 1

a.

b.

c.

d.

e.

f.

g.

Answer to Question 2

a.

b.

Answer to Question 3

a.

b.