STAT0028 Week 1 Exercises

Andrew Wu

2024-10-06

Question 1

a) The linear model is given by

$$y_{ij} = \mu_1 x_{1ij} + \mu_2 x_{2ij} + e_{ij},$$

where y_{ij} is the $j^{th} \in \{1, \dots, 20\}$ value in group $i \in \{1, 2\}$, $x_{kij} = \delta_{ki}$ are indicator variables, μ_i are the means of their respective types and e_{ij} are iid Gaussian error terms with constant variance. The equations in matrix form are

$$\mathbf{y} = X\mu + \mathbf{e},$$

which we define in the code block below and proceed to compute the ols estimates by fitting an intercept-free model.

```
type1 <- c(3.03,  5.53,  5.60,  9.30,  9.92, 12.51, 12.95, 15.21, 16.04, 16.84)
type2 <- c(3.19,  4.26,  4.47,  4.53,  4.67,  4.69, 12.78,  6.79,  9.37, 12.75)
y <- c(type1, type2)
X <- cbind(c(rep(1, 10), rep(0, 10)), c(rep(0, 10), rep(1, 10)))

engine.lm <- lm(y ~ 0 + X)
summary(engine.lm)
```

```
##
## Call:
## lm(formula = y ~ 0 + X)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.663 -2.333 -1.083  3.094  6.147
##
## Coefficients:
##    Estimate Std. Error t value Pr(>|t|)
## X1   10.693      1.345   7.948 2.69e-07 ***
## X2    6.750      1.345   5.017 8.95e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.255 on 18 degrees of freedom
## Multiple R-squared:  0.8307, Adjusted R-squared:  0.8119
## F-statistic: 44.17 on 2 and 18 DF,  p-value: 1.141e-07
```

Now, we perform a two sample t-test at a 5% power level. We must assume the groups have equal variance due to the constant variance assumption of the linear model:

```
t.test(type1, type2, alternative='two.sided', var.equal=TRUE, conf.level=0.95)
```

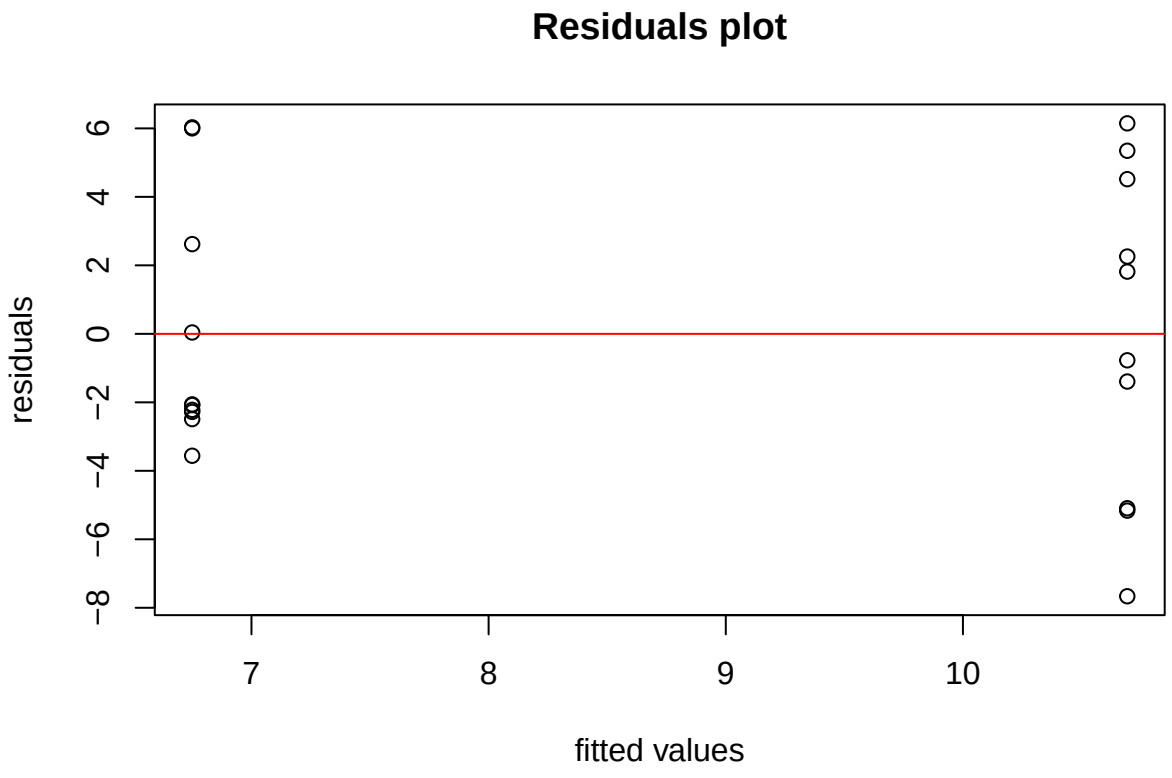
```
##
##  Two Sample t-test
##
## data:  type1 and type2
## t = 2.0723, df = 18, p-value = 0.05288
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.05444249  7.94044249
## sample estimates:
## mean of x mean of y
##    10.693     6.750
```

Since the p-value exceeds 0.05, we are unable to reject the null hypothesis that the means of the two groups are equal. Notice that the means of each group are exactly the ols estimates.

b)

We first create a residuals plot to validate the assumptions of iid errors with constant variance

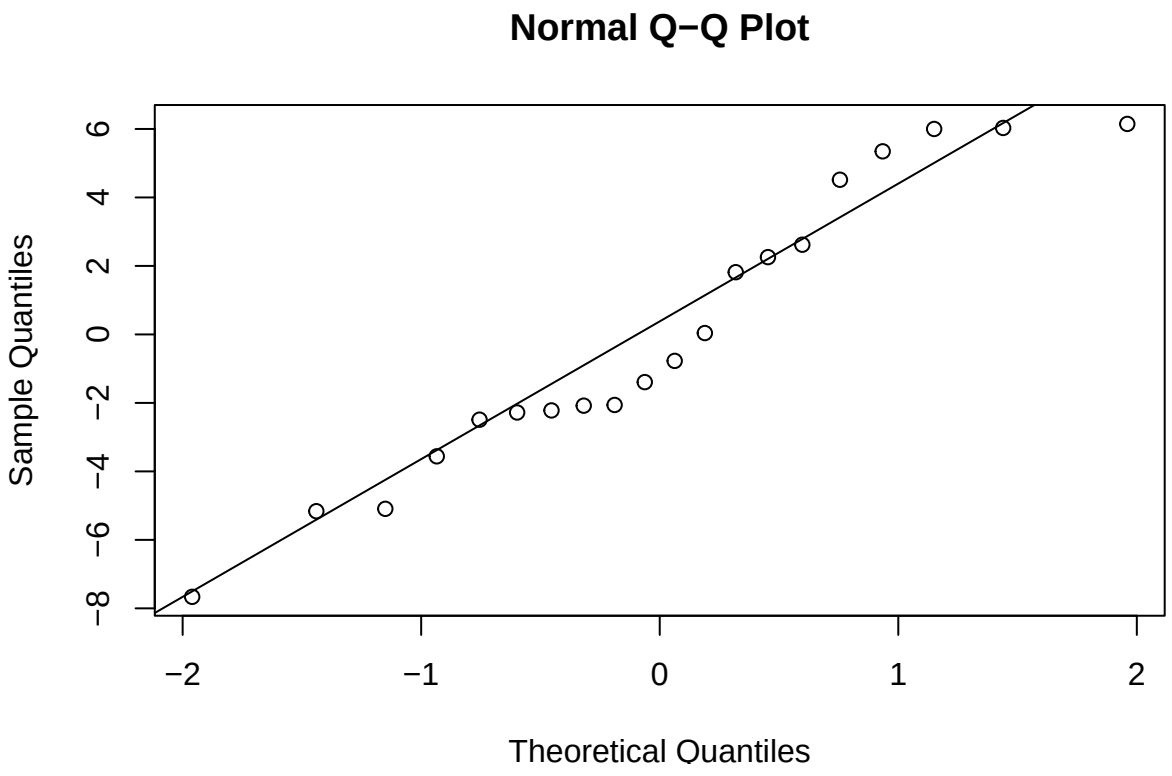
```
y.fit <- fitted(engine.lm)
residuals <- resid(engine.lm)
plot(y.fit, residuals, xlab="fitted values", ylab="residuals", main="Residuals plot")
abline(h=0, col="red")
```



The residuals separate into two distinct groups based on each type as our predictors are dummy variables. There is no clear pattern that would suggest the independence assumption is violated, although it is debatable whether the residuals of type I have 0 mean. Furthermore, the spread of the data appears to be slightly different, but this is inconclusive from our small sample.

Using a qqplot, it appears the residuals are not quite normally distributed - there is positive skew (negative at 0 normal quantile) and greater kurtosis (curvature at tails), though this is again inconclusive based on the small sample.

```
plot.new()
qqnorm(residuals)
qqline(residuals)
```



Question 2

a)

Predicting house prices given a set of features/predictors. The response is the house price, the features may include quantities such as area, number of bedrooms, number of storeys. The goal is to predict an approximate price given the relevant features and this could be used in helping real estate agents estimate the prices/commissions.

b)

Drawing parallels between crime and various statistics. The response is the average level of crime in a postcode, the variables could be things such as average income, number of schools, number of parks. The goal of the investigation is to assess what variables correlate to the amount of crime in an area without necessarily establishing causality.

c)

Quantifying the causal effect of a new drug on an illness. The response could be the size of a tumour, the features would be the dosage of a drug. The goal of fitting the model and analysis is to quantify the effectiveness of the drug and the appropriate dosages.

Question 3

a)

The sum of squares loss function is given by

$$\ell(\beta) = \sum_i \left(y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2})\right)^2$$

Differentiating with respect to each parameter and setting to zero, we have

$$\begin{aligned}
\frac{\partial \ell}{\partial \beta_0} &= -2 \sum_i \left(y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2})\right) = 0 \\
\frac{\partial \ell}{\partial \beta_1} &= -2 \sum_i x_{i1} \left(y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2})\right) = 0 \\
\frac{\partial \ell}{\partial \beta_2} &= -2 \sum_i x_{i2} \left(y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2})\right) = 0
\end{aligned}$$

b)

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} \\ & \vdots & \\ 1 & x_{n1} & x_{n2} \end{bmatrix}, \quad \beta = (\beta_0, \beta_1, \beta_2)^\top, \quad \mathbf{y} = (y_1, \ldots, y_n)^\top.$$

Evaluating both sides of $X^\top X \hat{\beta} = X^\top \mathbf{y}$,

$$\begin{bmatrix} n & \sum_i x_{i1} & \sum_i x_{i2} \\ \sum_i x_{i1} & \sum_i x_{i1}^2 & \sum_i x_{i1} x_{i2} \\ \sum_i x_{i2} & \sum_i x_{i1} x_{i2} & \sum_i x_{i2}^2 \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 1 \\ x_{11} & \cdots & x_{n1} \\ x_{12} & \cdots & x_{n2} \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

which when expanded and moved to one side yields

$$\begin{bmatrix} \sum_i \left(y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2})\right) \\ \sum_i x_{i1} \left(y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2})\right) \\ \sum_i x_{i2} \left(y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2})\right) \end{bmatrix} = \mathbf{0}$$

exactly the normal equations found in part a) once the constants are divided out.

Question 4

Let $\mathbf{y} = (y_1, y_2)^\top, \mu = (\mu_1, \mu_2)^\top$, $\sigma_i = \sqrt{\text{Var} y_i}$, $\rho = \text{Corr}(y_1, y_2)$. Then,

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix},$$

and

$$\det \Sigma = \sigma_1^2 \sigma_2^2 (1-\rho^2), \quad \Sigma^{-1} = \frac{1}{\det \Sigma} \begin{bmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{bmatrix}.$$

Since we have a quadratic form on a symmetric matrix,

$$\begin{aligned}(\mathbf{y}-\mu)^\top \Sigma^{-1} (\mathbf{y}-\mu) &= \frac{1}{\det(\Sigma)} \left(\sigma_2^2 (y_1-\mu_1)^2 + \sigma_1^2 (y_2-\mu_2)^2 - 2\rho\sigma_1\sigma_2 (y_1-\mu_1)(y_2-\mu_2) \right) \\ &= \frac{1}{1-\rho^2} \left(\frac{(y_1-\mu_1)^2}{\sigma_1^2} + \frac{(y_2-\mu_2)^2}{\sigma_2^2} - 2\rho \frac{y_1-\mu_1}{\sigma_1} \cdot \frac{y_2-\mu_2}{\sigma_2} \right).\end{aligned}$$

which upon substitution into the exponential along with the determinant yields the desired density function:

$$\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)} \left[\frac{(y_1-\mu_1)^2}{\sigma_1^2} + \frac{(y_2-\mu_2)^2}{\sigma_2^2} - 2\rho \frac{y_1-\mu_1}{\sigma_1} \cdot \frac{y_2-\mu_2}{\sigma_2} \right] \right)$$