STA2101 Assignment 1*

Ahmed Nawaz Amanullah †

October 2020

1

1.1 Problem

In a political poll, a random sample of n registered voters are to indicate which of two candidates they prefer. State a reasonable model for these data, in which the population proportion of registered voters favouring Candidate A is denoted by θ. Denote the observations $Y_1, ..., Y_n$.

1.2 Solution

My humble suggestion would be to try a Bernoulli distribution, that is let the observations $Y_1, ..., Y_n$ be a random sample from a Bernoulli distribution with θ denoting the probability of a registered voter favouring Candidate A. That is independently for $i = 1, ..., n$,

$$Y_i \sim B(1, \theta) \tag{1}$$

$$P(y_i|\theta) = \theta^{y_i}(1-\theta)^{1-y_i} \tag{2}$$

for $y_i = 0$ or $y_i = 1$, and zero otherwise.

†with help from the Overleaf team

2

2.1 Problem

A medical researcher conducts a study using twenty-seven litters of cancer-prone mice. Two members are randomly selected from each litter, and all mice are subjected to daily doses of cigarette smoke. For each pair of mice, one is randomly assigned to Drug A and one to Drug B. Time (in weeks) until the first clinical sign of cancer is recorded.

2.1.1 (a)

State a reasonable model for these data. Remember, a statistical model is a set of assertions that partly specify the probability distribution of the observable data. For simplicity, you may assume that the study continues until all the mice get cancer, and that log time until cancer has a normal distribution.

2.1.2 (b)

What is the parameter space for your model?

2.2 Solution

2.2.1 (a)

Let the observations of log time until cancer for mice administered drug A be $X_1, ..., X_{27}$ and the observations of log time until cancer for mice administered drug B be $Y_1, ..., Y_{27}$. Then, let $X_1, ..., X_{27}$ be a random sample from a normal distribution with expected value μ_A and variance σ_A^2. μ_A and variance σ_A^2 are unknown constants. Likewise, let $Y_1, ..., Y_{27}$ be a random sample from a normal distribution with expected value μ_B and variance σ_B^2. μ_B and variance σ_B^2 are unknown constants.

2.2.2 (b)

$$\{(\mu_A, \mu_B, \sigma_A^2, \sigma_B^2) : -\infty < \mu_A < \infty, -\infty < \mu_B < \infty, \sigma_A^2 > 0, \sigma_B^2 > 0\} \quad (3)$$

3

3.1 Problem

Suppose that volunteer patients undergoing elective surgery at a large hospital are randomly assigned to one of three different pain killing drugs, and one week after surgery they rate the amount of pain they have experienced on a scale from zero (no pain) to 100 (extreme pain).

3.1.1 (a)

State a reasonable model for these data. For simplicity, you may assume normality.

3.1.2 (b)

What is the parameter space?

3.2 Solution

3.2.1 (a)

Let the three pain killing drugs being administered be A, B and C. Let the pain rating of patients administered drug A be $X_1, ..., X_{n_A}$, where n_A is the number of patients administered drug A (which is a known constant). Likewise, let the pain rating of patients administered drug B be $Y_1, ..., Y_{n_B}$, where n_B is the number of patients administered drug B (which is a known constant). Finally, let the pain rating of patients administered drug C be $W_1, ..., W_{n_C}$, where n_C is the number of patients administered drug C (which is a known constant).

Then, let $X_1, ..., X_{n_A}$ be a random sample from a normal distribution with expected value μ_A and variance σ_A^2. μ_A and variance σ_A^2 are unknown constants. Likewise, let $Y_1, ..., Y_{n_B}$ be a random sample from a normal distribution with expected value μ_B and variance σ_B^2. μ_B and variance σ_B^2 are unknown constants. Finally, let $W_1, ..., W_{n_C}$ be a random sample from a normal distribution with expected value μ_C and variance σ_C^2. μ_C and variance σ_C^2 are unknown constants.

3.2.2 (b)

$$\{(\mu_A, \mu_B, \mu_C, \sigma_A^2, \sigma_B^2, \sigma_C^2) : 0 < \mu_i < 100, \sigma_i^2 > 0, i = A, B, C\} \tag{4}$$

4

4.1 Problem

Let $X_1, ..., X_n$ be a random sample (meaning independent and identically distributed) from a distribution with density $f(x) = \frac{\theta}{x^{\theta+1}}$ for $x > 1$, where $\theta > 0$.

4.1.1 (a)

Find the maximum likelihood estimator of θ. Show your work. The answer is a formula involving $X_1, ..., X_n$.

4.1.2 (b)

Suppose you observe these data: $1.37, 2.89, 1.52, 1.77, 1.04, 2.71, 1.19, 1.13, 15.66, 1.43$. Calculate the maximum likelihood estimate. My answer is 1.469102.

4.2 Solution

4.2.1 (a)

First, let's set up the likelihood function,

$$L(\theta) = \prod_{i=1}^{n} \frac{\theta}{x_i^{\theta+1}} = \frac{\theta^n}{(\prod_{i=1}^{n} x_i)^{\theta+1}} \tag{5}$$

Then, derive the log-likelihood function from this,

$$l(\theta) = \log\left(\frac{\theta^n}{(\prod_{i=1}^{n} x_i)^{\theta+1}}\right) = n \log\theta - (\theta + 1)\sum_{i=1}^{n} \log x_i \tag{6}$$

Finally, differentiate the log-likelihood function with respect to θ, set to zero, and solve,

$$\frac{\partial l(\theta)}{\partial \theta} = \frac{n}{\theta} - \sum_{i=1}^{n} \log x_i = 0 \tag{7}$$

$$\theta = \frac{n}{\sum_{i=1}^{n} \log x_i} \tag{8}$$

Thus, we have the MLE, $\hat{\theta} = \frac{n}{\sum_{i=1}^{n} \log x_i}$.

N.B. In the above derivation and throughout this document log is the natural logarithm, unless otherwise stated.

4.2.2 (b)

By executing the following lines of code in R, we confirm that we got the same answer.

```
# Observed data
X <- c(1.37, 2.89, 1.52, 1.77, 1.04, 2.71, 1.19, 1.13, 15.66, 1.43)
# number of obs
n <- length(X)
# natural log of obs
lnX <- log(X)
# compute MLE
theta_hat <- n/sum(lnX)
# display MLE
sprintf('mle estimate, theta_hat: %.6f', theta_hat)
```

5 True or False

Label each statement below True or False. Write "T" or "F" beside each statement. Assume the $\alpha = 0.05$ significance level.

5.1 (a)

F The p-value is the probability that the null hypothesis is true.

5.2 (b)

F The p-value is the probability that the null hypothesis is false.

5.3 (c)

F In a study comparing a new drug to the current standard treatment, the null hypothesis is rejected. We conclude that the new drug is ineffective.

5.4 (d)

F If $p > .05$ we reject the null hypothesis at the .05 level.

5.5 (e)

T If $p < .05$ we reject the null hypothesis at the .05 level.

5.6 (f)

F The greater the p-value, the stronger the evidence against the null hypothesis.

5.7 (g)

F In a study comparing a new drug to the current standard treatment, $p > .05$. We conclude that the new drug and the existing treatment are not equally effective.

5.8 (h)

T The 95% confidence interval for β_3 is from -0.26 to 3.12. This means $P\{-0.26 < \beta_3 < 3.12\} = 0.95$.

6

6.1 Problem

Let $Y_1, ..., Y_n$ be a random sample from a normal distribution with mean μ and variance σ^2, so that $T = \frac{\sqrt{n}(\bar{Y}-\mu)}{S} \sim t(n-1)$. This is something you don't need to prove, for now.

6.1.1 (a)

Derive a $(1-\alpha)100\%$ confidence interval for μ. "Derive" means show all the high school algebra. Use the symbol $t_{\alpha/2}$ for the number satisfying $Pr(T > t_{\alpha/2}) = \alpha/2$.

6.1.2 (b)

A random sample with $n = 23$ yields $\bar{Y} = 2.57$ and a sample variance of $S^2 = 5.85$. Using the critical $t_{0.025} = 2.07$, give a 95% confidence interval for μ. The answer is a pair of numbers.

6.1.3 (c)

Test $H_0 : \mu = 3$ at $\alpha = 0.05$.

i. Give the value of the T statistic. The answer is a number.

ii. State whether you reject H_0, Yes or No.

iii. Can you conclude that μ is different from 3? Answer Yes or No.

iv. If the answer is Yes, state whether $\mu > 3$ or $\mu < 3$. Pick one.

6.1.4 (d)

Show that using a t-test, $H_0 : \mu = \mu_0$ is rejected at significance level α if and only the $(1-\alpha)100\%$ confidence interval for μ does not include μ_0. The problem is easier if you start by writing the set of T values for which H_0 is *not* rejected.

6.1.5 (e)

In Question 6b, does this mean $Pr1.53 < \mu < 3.61 = 0.95$? Answer Yes or No and briefly explain.

6.2 Solution

6.2.1 (a)

$$P\{t_{n-1,1-\alpha/2} < T < t_{n-1,\alpha/2}\} = 1 - \alpha \tag{9}$$

$$P\left\{t_{n-1,1-\alpha/2} < \frac{\sqrt{n}(\bar{Y} - \mu)}{S} < t_{n-1,\alpha/2}\right\} = 1 - \alpha \tag{10}$$

$$P\left\{t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}} < \bar{Y} - \mu < t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right\} = 1 - \alpha \tag{11}$$

$$P\left\{\bar{Y} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{Y} - t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}}\right\} = 1 - \alpha \tag{12}$$

$$P\left\{\bar{Y} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{Y} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right\} = 1 - \alpha \tag{13}$$

Thus, the $(1-\alpha)100\%$ confidence interval for μ is: $\bar{Y} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{Y} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}$.

6.2.2 (b)

Inserting the numbers provided in the confidence interval formula derived previously, we get the confidence interval: $1.53 < \mu < 3.61$. This computation was done in R using the following code:

```
# Question 6 b computation
# Sample mean
Y_bar <- 2.57
# number of obs
n <- 23
# sample standard dev
s <- sqrt(5.85)
# critical value
t <- 2.07
# compute lower and upper bounds as derived in (a)
```

```
lower <- Y_bar - t*(s/sqrt(n))
upper <- Y_bar + t*(s/sqrt(n))
# display bounds
sprintf('conf interval: %.2f < mu < %.2f', lower, upper)
```

6.2.3 (c)

i. We compute $T = \frac{\sqrt{n}(\bar{Y}-\mu)}{S}$ with the numbers provided. We thus get the value of the T statistic as -0.85. We carried out this computation in R using the following code:

```
# Question 6 c i computation
# Sample mean
Y_bar <- 2.57
# number of obs
n <- 23
# sample standard dev
s <- sqrt(5.85)
# null hypothesis mean
mu <- 3
# test statistic computation
t <- (sqrt(n)*(Y_bar - mu))/s
#display computation
sprintf('T: %.2f', t)
```

ii. No.

iii. No.

iv. N/A.

6.2.4 (d)

Making use of the hint, let's start by writing out the set of T values for which H_0 is *not* rejected. We know if the T value computed is within the critical value bounds, H_0 is *not* rejected. That is,

$$t_{n-1,1-\alpha/2} < T < t_{n-1,\alpha/2} \tag{14}$$

$$t_{n-1,1-\alpha/2} < \frac{\sqrt{n}(\bar{Y} - \mu_0)}{S} < t_{n-1,\alpha/2} \tag{15}$$

$$\bar{Y} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} < \mu_0 < \bar{Y} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} \tag{16}$$

So we glean from this that H_0 is *not* rejected if μ_0 is within the interval $\bar{Y} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} < \mu_0 < \bar{Y} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}$, which we observe is the $(1-\alpha)100\%$ confidence interval for μ we derived in 6 (a). From this we get that the $(1-\alpha)100\%$ confidence interval for μ does not include μ_0 if H_0 is rejected (the contrapositive argument). Also if the $(1-\alpha)100\%$ confidence interval for μ does not include μ_0, then the computed T value is not within the critical value bounds, and if that is so, then H_0 is rejected, thus completing our proof.

N.B. Alternatively if H_0 is rejected, then T value is in critical region region, and if so μ_0 does not lie in the confidence interval (can show this algebraically). This argument can be used in place of the contrapositive argument used above.

6.2.5 (e)

Yes, as that is the 95% confidence interval for μ we generated in 6 (b), and by the definition of the confidence interval, there is a 95% probability for the confidence interval to contain the unknown population parameter μ.

7

7.1 Problem

Let $Y_1, ..., Y_n$ be a random sample from a distribution with mean μ and standard deviation σ.

7.1.1 (a)

Show that the sample variance $S^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}$ is an unbiased estimator of σ^2.

7.1.2 (b)

Denote the sample standard deviation by $S = \sqrt{S^2}$. Assume that the data come from a continuous distribution, so that $Var(S) > 0$. Using this fact, show that S is a *biased* estimator of σ.

7.2 Solution

7.1.1 (a)

We want to show $E(S^2) = \sigma^2$. Consider,

$$E[(n-1)S^2] = (n-1)E[S^2] = E\left[\sum_{i=1}^{n}(Y_i - \bar{Y})^2\right] \tag{17}$$

$$(n-1)E[S^2] = E\left[\sum_{i=1}^{n}(Y_i - \mu + \mu - \bar{Y})^2\right] \tag{18}$$

$$(n-1)E[S^2] = E\left[\sum_{i=1}^{n}[(Y_i - \mu)^2 - 2(Y_i - \mu)(\bar{Y} - \mu) + (\bar{Y} - \mu)^2]\right] \tag{19}$$

$$(n-1)E[S^2] = E\left[\sum_{i=1}^{n}(Y_i - \mu)^2 - 2(\bar{Y} - \mu)\sum_{i=1}^{n}(Y_i - \mu) + n(\bar{Y} - \mu)^2\right] \tag{20}$$

$$(n-1)E[S^2] = E\left[\sum_{i=1}^{n}(Y_i - \mu)^2 - 2n(\bar{Y} - \mu)^2 + n(\bar{Y} - \mu)^2\right] \tag{21}$$

$$(n-1)E[S^2] = \sum_{i=1}^{n} E\left[(Y_i - \mu)^2\right] - nE\left[(\bar{Y} - \mu)^2\right] \tag{22}$$

$$(n-1)E[S^2] = \sum_{i=1}^{n} \sigma^2 - nE\left[(\bar{Y} - \mu)^2\right] \tag{23}$$

$$(n-1)E[S^2] = n\sigma^2 - nE\left[\bar{Y}^2 - 2\bar{Y}\mu + \mu^2\right] \tag{24}$$

$$(n-1)E[S^2] = n\sigma^2 - n(E[\bar{Y}^2] - 2\mu E[\bar{Y}] + \mu^2) \tag{25}$$

$$(n-1)E[S^2] = n\sigma^2 - n(\sigma^2_{\bar{Y}} + E[\bar{Y}]^2 - 2\mu E[\bar{Y}] + \mu^2) \tag{26}$$

$$(n-1)E[S^2] = n\sigma^2 - n(\sigma^2_{\bar{Y}} + \mu^2 - 2\mu^2 + \mu^2) \tag{27}$$

$$(n-1)E[S^2] = n\sigma^2 - n\sigma^2_{\bar{Y}} \tag{28}$$

$$(n-1)E[S^2] = n\sigma^2 - n\frac{1}{n^2}\sum_{i=1}^{n}\sigma_{Y_i}^2 \tag{29}$$

$$(n-1)E[S^2] = (n-1)\sigma^2 \tag{30}$$

$$E[S^2] = \sigma^2 \tag{31}$$

This completes our proof of sample variance S^2 being an unbiased estimator of σ^2 as required.

N.B. Since $Y_1, ..., Y_n$ are a random sample we could use the variance sum rule for independent random variables above for the necessary step $\sigma_{\bar{Y}}^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sigma_{Y_i}^2$.

7.2.2 (b)

We want to show that S is a *biased* estimator of σ. That is, $E[S] \neq \sigma$. From (a), we know $E[S^2] = \sigma^2$. Using this, consider,

$$\sigma^2 = E[S^2] = Var(S) + E[S]^2 \tag{32}$$

$$E[S]^2 = \sigma^2 - Var(S) \tag{33}$$

$$E[S] = \sqrt{\sigma^2 - Var(S)} < \sigma \tag{34}$$

The inequality in the last line follows as we were provided $Var(S) > 0$. It follows from $E[S] < \sigma$ that $E[S] \neq \sigma$, with which we have shown that S is a *biased* estimator of σ as required.

8

8.1 Problem

Let $Y_1, ..., Y_n$ as a a random sample from a Bernoulli distribution with parameter θ. That is, independently for $i = 1, ..., n$, $P(y_i|\theta) = \theta^{y_i}(1-\theta)^{1-y_i}$ for $y_i = 0$ or $y_i = 1$, and zero otherwise. A large-sample test of $H_0 : \theta = \theta_0$ uses the test statistic

$$Z_1 = \frac{\sqrt{n}(\bar{Y} - \theta_0)}{\sqrt{\theta_0(1-\theta_0)}} \tag{35}$$

When the null hypothesis is true, the distribution of Z_1 is approximately standard normal. What if the null hypothesis is false? In the following, you may use the following fact without proof: the Central Limit Theorem implies that $\hat{\theta} = \bar{Y}$ is approximately normal with mean θ and variance $\frac{\theta(1-\theta)}{n}$.

8.1.1 (a)

Derive the approximate large-sample cumulative distribution function of Z_1 when θ is not necessarily equal to θ_0. Derive means show the high school algebra. Use $\Phi(.)$ to denote the cumulative distribution function of a standard normal.

8.1.2 (b)

Give an expression for the power of a two-sided, size α test of $H_0 : \theta = \theta_0$. Let $z_{\alpha/2}$ denote the point with $\alpha/2$ of the standard normal above it. For example, $z_{0.025} = 1.96$.

8.1.3 (c)

Suppose we are testing $H_0 : \theta = \frac{1}{2}$ at $\alpha = 0.05$, with $n = 100$. If the true value of θ is 0.55, what is the power of the test? The answer is a number between zero and one. I used R's **pnorm** function to calculate standard normal probabilities.

8.1.4 (d)

Suppose again that the true value of θ is 0.55. What is the smallest value of the sample size n such that the power of the test will be at least 0.80? The answer is an integer.

Here are a few comments on this question. The use of **pnorm** tells you that you are *not* estimating power by simulation. I used a **while** loop; look it up if you need to. My answer is greater than 700, but less than 800.

Bring your R printout for this question to the quiz. The printout must show all input and output.

8.2 Solution

8.2.1 (a)

The derivation of the approximate large-sample cumulative distribution function of Z_1, $F_{Z_1}(z) = P(Z_1 \leq z)$, when θ is not necessarily equal to θ_0 is as follows,

$$F_{Z_1}(z) = P(Z_1 \leq z) = P\left(\frac{\sqrt{n}(\bar{Y} - \theta_0)}{\sqrt{\theta_0(1-\theta_0)}} \leq z\right) \tag{36}$$

$$F_{Z_1}(z) = P(Z_1 \leq z) = P\left(\bar{Y} \leq z\sqrt{\frac{\theta_0(1-\theta_0)}{n}} + \theta_0\right) \tag{37}$$

$$F_{Z_1}(z) = P(Z_1 \leq z) = P\left(\frac{\sqrt{n}(\bar{Y} - \theta)}{\sqrt{\theta(1-\theta)}} \leq z\sqrt{\frac{\theta_0(1-\theta_0)}{\theta(1-\theta)}} + \frac{\sqrt{n}(\theta_0 - \theta)}{\sqrt{\theta(1-\theta)}}\right) \tag{38}$$

$$F_{Z_1}(z) = P(Z_1 \leq z) \approx \Phi\left(z\sqrt{\frac{\theta_0(1-\theta_0)}{\theta(1-\theta)}} + \frac{\sqrt{n}(\theta_0 - \theta)}{\sqrt{\theta(1-\theta)}}\right) \tag{39}$$

8.2.2 (b)

We note that power is the probability that the null hypothesis is rejected provided an alternative hypothesis is true. We want the power of a two-sided, size α test of $H_0 : \theta = \theta_0$, where say the alternative hypothesis $H_a : \theta = \theta_a$ is true. We get this making use of our result from (a), substituting $\theta = \theta_a$ in the derived expression, as follows,

$$\text{Power} = F_{Z_1}(z_{1-\alpha/2}) + (1 - F_{Z_1}(z_{\alpha/2})) \tag{40}$$

$$\begin{aligned}\text{Power} \approx & 1 + \Phi\left(z_{1-\alpha/2}\sqrt{\frac{\theta_0(1-\theta_0)}{\theta_a(1-\theta_a)}} + \frac{\sqrt{n}(\theta_0 - \theta_a)}{\sqrt{\theta_a(1-\theta_a)}}\right) \\ & - \Phi\left(z_{\alpha/2}\sqrt{\frac{\theta_0(1-\theta_0)}{\theta_a(1-\theta_a)}} + \frac{\sqrt{n}(\theta_0 - \theta_a)}{\sqrt{\theta_a(1-\theta_a)}}\right)\end{aligned} \tag{41}$$

8.2.3 (c)

Plugging in the provided values in R, we get a power of 0.17. This computation was done in R using the following code:

```
# Question 8 c computation
# null hypothesis theta
theta_o <- 1/2
# size of test is 0.05; critical value corresponding to this
alpha <- 0.05
z_c <- -qnorm(alpha/2)
# sample size
n <- 100
# true proportion
theta <- 0.55
# compute power
# intermediate terms
x1 <- sqrt(n)*((theta_o-theta)/sqrt(theta*(1-theta)))
x2 <- z_c*sqrt((theta_o*(1-theta_o))/(theta*(1-theta)))
# power
power <- 1-pnorm(x1+x2)+pnorm(x1-x2)
# display computation
sprintf('Power: %.2f', power)
```

N.B. In the code implementation, we make use of the relation $z_{1-\alpha/2} = -z_{\alpha/2}$.

8.2.4 (d)

We want to solve for the value of sample size n that gives a power of 0.80. That is we want the value of n such that

$$\begin{aligned} 0.80 =& 1 + \Phi\left(z_{0.975}\sqrt{\frac{0.5(1-0.5)}{0.55(1-0.55)}} + \frac{\sqrt{n}(0.5-0.55)}{\sqrt{0.55(1-0.55)}} \right) \\ & - \Phi\left(z_{0.025}\sqrt{\frac{0.5(1-0.5)}{0.55(1-0.55)}} + \frac{\sqrt{n}(0.5-0.55)}{\sqrt{0.55(1-0.55)}} \right) \end{aligned} \tag{42}$$

In R, we can simply iterate through the range provided until we arrive at the value of n that gives a power of at least n. An alternative is making use of Newton's method, which we made use of to get a value of $n = 783$, which is between 700 and 800 as specified by the professor. This computation was done in R using the following code:

```
# Question 8 d computation
# null hypothesis theta
theta_o <- 1/2
# size of test is 0.05; critical value corresponding to this
alpha <- 0.05
z_c <- -qnorm(alpha/2)
# true proportion
theta <- 0.55
# beta (power)
beta <- 0.80
# get least sample size -
#  need to implement Newton's method for this
# least sample size Newton's method
# initial estimate (prof said between 700 and 800 so choose)
n <- 750
# intermediate terms
x1 <- sqrt(n)*((theta_o-theta)/sqrt(theta*(1-theta)))
x2 <- z_c*sqrt((theta_o*(1-theta_o))/(theta*(1-theta)))
x3 <- (theta_o-theta)/(2*sqrt(n*theta*(1-theta)))
# next estimate
f_n <- 1-beta-pnorm(x1+x2)+pnorm(x1-x2)
fp_n <- dnorm(x1-x2)*x3-dnorm(x1+x2)*x3
n_ip1 <- n - (f_n/fp_n)
# use while loop with accepted tolerance on when to stop
tol <- 0.1
while((n_ip1-n)>tol){
        n <- n_ip1
        # intermediate terms
        x1 <- sqrt(n)*((theta_o-theta)/sqrt(theta*(1-theta)))
        x2 <- z_c*sqrt((theta_o*(1-theta_o))/(theta*(1-theta)))
        x3 <- (theta_o-theta)/(2*sqrt(n*theta*(1-theta)))
        # next estimate
```

```
        f_n <- 1-beta-pnorm(x1+x2)+pnorm(x1-x2)
        fp_n <- dnorm(x1-x2)*x3-dnorm(x1+x2)*x3
        n_ip1 <- n - (f_n/fp_n)
}
# get accepted tolerance estimate
n <- n_ip1
# print estimated least sample size giving at least required power
# display computation
sprintf('Least sample size: %.0f', n)
```

9

9.1 Problem

In the *centered* linear regression model, sample means are subtracted from the explanatory variables, so that values above average are positive and values below average are negative. Here is a version with one explanatory variable. For $i = 1, ..., n$, let $y_i = \beta_0 + \beta_1(x_i - \bar{x}) + \epsilon_i$, where

- β_0 and β_1 are unknown constants (parameters).
- x_i are known, observed constants.
- $\epsilon_1, ..., \epsilon_n$ are unobservable random variables with $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2$ and $Cov(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$.
- σ^2 is an unknown constant (parameter).
- $y_1, ..., y_n$ are observable random variables.

9.1.1 (a)

What is $E(y_i)$? What is $Var(y_i)$?

9.1.2 (b)

Prove that $Cov(y_i, y_j) = 0$. Use the following definition: $Cov(U, V) = E(U - E(U))(V - E(V))$.

9.1.3 (c)

If ϵ_i and ϵ_j are independent (not just uncorrelated), then so are y_i and y_j, because functions of independent random variables are independent. Proving this in full generality requires advanced definitions, but in this case the functions are so simple that we can get away with an elementary definition. Let X_1 and X_2 be independent random variables, meaning $PX_1 \leq x_1, X_2 \leq x_2 = PX_1 \leq x_1 PX_2 \leq x_2$ for all real x_1 and x_2. Let $Y_1 = X_1 + a$ and $Y_2 = X_2 + b$, where a and b are constants. Prove that Y_1 and Y_2 are independent.

9.1.4 (d)

In *least squares estimation*, we observe random variables $y_1, ..., y_n$ whose distributions depend on a parameter θ, which could be a vector. To estimate θ, write the expected value of y_i as a function of θ, say $E_\theta(y_i)$, and then estimate θ by the value that gets the observed data values as close as possible to their expected values. To do this, minimize

$$Q = \sum_{i=1}^{n} (y_i - E_\theta(y_i))^2 \tag{43}$$

The value of θ that makes Q as small as possible is the least squares estimate. Using this framework, find the least squares estimates of β_0 and β_1 for the centered regression model. The answer is a pair of formulas. Show your work.

9.1.5 (e)

Because of the centering, it is possible to verify that the solution actually *minimizes* the sum of squares Q, using only single-variable second derivative tests. Do this part too.

9.1.6 (f)

How about a least squares estimate of σ^2?

9.1.7 (g)

You know that the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ must be unbiased, but show it by calculating their expected values for this particular case.

9.1.8 (h)

Calculate $\hat{\beta}_0$ and $\hat{\beta}_1$ for the following data. Your answer is a pair of numbers.

x	8	7	7	9	4
y	9	13	9	8	6

I get $\hat{\beta}_1 = \frac{1}{2}$.

9.1.9 (i)

Going back to the general setting (not just the numerical example with $n = 5$), suppose the ϵ_i are normally distributed.

i. What is the distribution of y_i?

ii. Write the log likelihood function.

iii. Obtain the maximum likelihood estimates of β_0 and β_1; don't bother with σ^2. The answer is a pair of formulas. *Don't do more work than you have to!* As soon as you realize that you have already solved this problem, stop and write down the answer.

9.1.10 (j)

Still for this centered model with a single explanatory variable, suppose we centered the y_i values too. In this case what is the least squares estimate of β_0? Show your work.

9.2 Solution

9.2.1 (a)

$$E(y_i) = E(\beta_0 + \beta_1(x_i - \bar{x}) + \epsilon_i) \tag{44}$$
$$E(y_i) = \beta_0 + \beta_1(x_i - \bar{x}) + E(\epsilon_i) \tag{45}$$
$$E(y_i) = \beta_0 + \beta_1(x_i - \bar{x}) + 0 \tag{46}$$
$$E(y_i) = \beta_0 + \beta_1(x_i - \bar{x}) \tag{47}$$

$$Var(y_i) = Var(\beta_0 + \beta_1(x_i - \bar{x}) + \epsilon_i) \tag{48}$$
$$Var(y_i) = Var(\epsilon_i) \tag{49}$$
$$Var(y_i) = \sigma^2 \tag{50}$$

9.2.2 (b)

$$Cov(y_i, y_j) = E\{(y_i - E(y_i))(y_j - E(y_j))\} \tag{51}$$
$$Cov(y_i, y_j) = E\{(\beta_0 + \beta_1(x_i - \bar{x}) + \epsilon_i - \beta_0 - \beta_1(x_i - \bar{x}))(y_j - E(y_j))\} \tag{52}$$
$$Cov(y_i, y_j) = E\{\epsilon_i(y_j - E(y_j))\} \tag{53}$$
$$Cov(y_i, y_j) = E\{\epsilon_i\epsilon_j\} \tag{54}$$
$$Cov(y_i, y_j) = E\{(\epsilon_i - 0)(\epsilon_j - 0)\} \tag{55}$$
$$Cov(y_i, y_j) = E\{(\epsilon_i - E(\epsilon_i))(\epsilon_j - E(\epsilon_j))\} \tag{56}$$
$$Cov(y_i, y_j) = Cov(\epsilon_i, \epsilon_j) \tag{57}$$

And we are provided that for $i \neq j$, $Cov(\epsilon_i, \epsilon_j) = 0$, thus completing our proof that $Cov(y_i, y_j) = Cov(\epsilon_i, \epsilon_j) = 0$.

9.2.3 (c)

$$P\{Y_1 \leq y_1, Y_2 \leq y_2\} = P\{X_1 + a \leq y_1, X_2 + b \leq y_2\} \tag{58}$$
$$P\{Y_1 \leq y_1, Y_2 \leq y_2\} = P\{X_1 \leq y_1 - a, X_2 \leq y_2 - b\} \tag{59}$$
$$P\{Y_1 \leq y_1, Y_2 \leq y_2\} = P\{X_1 \leq y_1 - a\}P\{X_2 \leq y_2 - b\} \tag{60}$$
$$P\{Y_1 \leq y_1, Y_2 \leq y_2\} = P\{X_1 + a \leq y_1\}P\{X_2 + b \leq y_2\} \tag{61}$$
$$P\{Y_1 \leq y_1, Y_2 \leq y_2\} = P\{Y_1 \leq y_1\}P\{Y_2 \leq y_2\} \tag{62}$$

Having shown $P\{Y_1 \leq y_1, Y_2 \leq y_2\} = P\{Y_1 \leq y_1\}P\{Y_2 \leq y_2\}$, we have completed our proof that Y_1 and Y_2 are independent as required.

9.2.4 (d)

Let the least square estimates of β_0 and β_1 be b_0 and b_1 respectively. In our case, $\theta = [b_0, b_1]'$ and $E_\theta(y_i) = b_0 + b_1(x_i - \bar{x})$. Then consider,

$$Q = \sum_{i=1}^{n} (y_i - E_\theta(y_i))^2 \tag{63}$$

$$Q = \sum_{i=1}^{n} (y_i - b_0 - b_1(x_i - \bar{x}))^2 \tag{64}$$

First differentiate with respect to b_0 and set to 0,

$$\frac{\partial Q}{\partial b_0} = \sum_{i=1}^{n} -2(y_i - b_0 - b_1(x_i - \bar{x})) = 0 \tag{65}$$

$$nb_0 = \sum_{i=1}^{n} (y_i - b_1(x_i - \bar{x})) \tag{66}$$

$$nb_0 = n\bar{y} - b_1(n\bar{x} - n\bar{x}) \tag{67}$$

$$b_0 = \bar{y} \tag{68}$$

Next, differentiate with respect to b_1 and set to 0,

$$\frac{\partial Q}{\partial b_1} = \sum_{i=1}^{n} (-2(y_i - b_0 - b_1(x_i - \bar{x}))(x_i - \bar{x})) = 0 \tag{69}$$

$$b_0 \sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} y_i(x_i - \bar{x}) - b_1 \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{70}$$

$$b_0(n\bar{x} - n\bar{x}) = \sum_{i=1}^{n} y_i(x_i - \bar{x}) - b_1 \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{71}$$

$$0 = \sum_{i=1}^{n} y_i(x_i - \bar{x}) - b_1 \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{72}$$

$$b_1 \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} y_i(x_i - \bar{x}) \tag{73}$$

$$b_1 = \frac{\sum_{i=1}^{n} y_i(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{74}$$

The least squares estimates are thus $b_0 = \bar{y}$ and $b_1 = \frac{\sum_{i=1}^{n} y_i(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$.

9.2.5 (e)

First, carry out the single-variable second derivative test with respect to b_0,

$$\frac{\partial^2 Q}{\partial b_0^2} = \sum_{i=1}^{n} -2(0-1-0) = 2n \tag{75}$$

Since $n > 0$, $\frac{\partial^2 Q}{\partial b_0^2} = 2n > 0$, which indicates a local minimum by the single-variable second derivative test.

Next, carry out the single-variable second derivative test with respect to b_1,

$$\frac{\partial^2 Q}{\partial b_1^2} = \sum_{i=1}^{n}(-2(0-0-(x_i-\bar{x}))(x_i-\bar{x})) = 2\sum_{i=1}^{n}(x_i-\bar{x})^2 \tag{76}$$

Since $(x_i-\bar{x})^2 \geq 0$, $\frac{\partial^2 Q}{\partial b_1^2} = 2\sum_{i=1}^{n}(x_i-\bar{x})^2 \geq 0$, which indicates a local minimum by the single-variable second derivative test.

Making use of the second partial derivatives test (since this is multivariable), we further calculate,

$$\frac{\partial^2 Q}{\partial b_1 \partial b_0} = \sum_{i=1}^{n} -2(0-0-(x_i-\bar{x})) = 2\sum_{i=1}^{n}(x_i-\bar{x}) = 0 \tag{77}$$

$$\frac{\partial^2 Q}{\partial b_0 \partial b_1} = \frac{\partial^2 Q}{\partial b_1 \partial b_0} = 0 \tag{78}$$

Afterwards calculating second derivative test discriminant,

$$D = \frac{\partial^2 Q}{\partial b_0^2}\frac{\partial^2 Q}{\partial b_1^2} - \frac{\partial^2 Q}{\partial b_1 \partial b_0}\frac{\partial^2 Q}{\partial b_0 \partial b_1} \tag{79}$$

$$D = (2n)(2\sum_{i=1}^{n}(x_i-\bar{x})^2) - 0 = 4n\sum_{i=1}^{n}(x_i-\bar{x})^2 \geq 0 \tag{80}$$

Since $\frac{\partial^2 Q}{\partial b_0^2} = 2n > 0$ and $D \geq 0$, the least square estimates corresponds to a local minimum of Q by the second partial derivatives test as required.

9.2.6 (f)

I confess I am at a loss with regards to answering this. The method prescribed in (d) can not be seemingly applied to this, seeing as the expected value of y_i can't be expressed as a function of σ^2. The best response I have is writing the unbiased estimate of σ^2 using the least squares estimates, b_0 and b_1, that was derived earlier as follows,

$$S^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} \tag{81}$$

$$S^2 = \frac{\sum_{i=1}^{n}(y_i - b_0 - b_1(x_i - \bar{x}))^2}{n-2} \tag{82}$$

$$S^2 = \frac{\sum_{i=1}^{n}\left(y_i - \bar{y} - \left(\frac{\sum_{i=1}^{n} y_i(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)(x_i - \bar{x})\right)^2}{n-2} \tag{83}$$

$$S^2 = \frac{\sum_{i=1}^{n}\left((y_i - \bar{y})^2 - 2(y_i - \bar{y})\left(\frac{\sum_{i=1}^{n} y_i(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)(x_i - \bar{x}) + \left(\frac{\sum_{i=1}^{n} y_i(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)^2 (x_i - \bar{x})^2\right)}{n-2} \tag{84}$$

$$S^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 - 2(\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}))\left(\frac{\sum_{i=1}^{n} y_i(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right) + \frac{(\sum_{i=1}^{n} y_i(x_i-\bar{x}))^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}{n-2} \tag{85}$$

$$S^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 - 2(\sum_{i=1}^{n} y_i(x_i - \bar{x}))\left(\frac{\sum_{i=1}^{n} y_i(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right) + \frac{(\sum_{i=1}^{n} y_i(x_i-\bar{x}))^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}{n-2} \tag{86}$$

$$S^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 - \frac{(\sum_{i=1}^{n} y_i(x_i-\bar{x}))^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}{n-2} \tag{87}$$

9.2.7 (g)

First consider least squares estimator $\hat{\beta}_0$ or b_0,

$$E(b_0) = E(\bar{y}) \tag{88}$$

$$E(b_0) = E\left(\frac{\sum_{i=1}^{n} y_i}{n}\right) \tag{89}$$

$$E(b_0) = \frac{\sum_{i=1}^{n} E(y_i)}{n} \tag{90}$$
$$E(b_0) = \frac{\sum_{i=1}^{n} (\beta_0 + \beta_1(x_i - \bar{x}))}{n} \tag{91}$$
$$E(b_0) = \frac{n\beta_0 + \beta_1 \sum_{i=1}^{n} (x_i - \bar{x})}{n} \tag{92}$$
$$E(b_0) = \frac{n\beta_0}{n} \tag{93}$$
$$E(b_0) = \beta_0 \tag{94}$$

Next consider least squares estimator $\hat{\beta}_1$ or b_1,

$$E(b_1) = E\left(\frac{\sum_{i=1}^{n} y_i(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\right) \tag{95}$$
$$E(b_1) = \frac{\sum_{i=1}^{n} (E(y_i))(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{96}$$
$$E(b_1) = \frac{\sum_{i=1}^{n} (\beta_0 + \beta_1(x_i - \bar{x}))(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{97}$$
$$E(b_1) = \frac{(\beta_0 \sum_{i=1}^{n} (x_i - \bar{x})) + \beta_1 \sum_{i=1}^{n} (x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{98}$$
$$E(b_1) = \frac{0 + \beta_1 \sum_{i=1}^{n} (x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{99}$$
$$E(b_1) = \beta_1 \tag{100}$$

9.2.8 (h)

Plugging in the provided data in our derived formulae we get, $\hat{\beta}_0 = b_0 = 9.00$ and $\hat{\beta}_1 = b_1 = 0.50$ (the latter matching the provided answer). This computation was done in R using the following code:

```
# Question 9 h computation
# data input
x <- c(8,7,7,9,4)
y <- c(9,13,9,8,6)

# b0
b0 <- mean(y)
```

```
# b1
# Sxy
Sxy <- sum((y)*(x-mean(x)))
# Sxx
Sxx <- sum((x-mean(x))^2)
# b1
b1 <- Sxy/Sxx
# print computation
sprintf('b0: %.2f, b1: %.2f', b0, b1)
```

9.2.9 (i)

i. $y_i = \beta_0 + \beta_1(x_i - \bar{x}) + \epsilon_i \sim N(\beta_0 + \beta_1(x_i - \bar{x}), \sigma^2)$

N.B. y_i differs from ϵ_i by a constant, so distribution for y_i is just the distribution of ϵ_i shifted).

ii. First, obtain the likelihood function,

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \beta_0 - \beta_1(x_i - \bar{x}))^2}{2\sigma^2}} \tag{101}$$

Then, proceed with obtaining the log likelihood function,

$$l(\beta_0, \beta_1) = \log(L(\beta_0, \beta_1)) \tag{102}$$

$$l(\beta_0, \beta_1) = \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \beta_0 - \beta_1(x_i - \bar{x}))^2}{2\sigma^2}}\right) \tag{103}$$

$$l(\beta_0, \beta_1) = \sum_{i=1}^{n} \left(-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(y_i - \beta_0 - \beta_1(x_i - \bar{x}))^2}{2\sigma^2}\right) \tag{104}$$

$$l(\beta_0, \beta_1) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1(x_i - \bar{x}))^2}{2\sigma^2} \tag{105}$$

iii. First, obtain MLE of β_0 by differentiating log-likelihood with respect to β_0 and setting to 0,

$$\frac{\partial l}{\partial \beta_0} = -\frac{\sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1(x_i - \bar{x}))(-1)}{2\sigma^2} = 0 \tag{106}$$

$$\frac{\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1(x_i - \bar{x}))}{\sigma^2} = 0 \tag{107}$$

$$n\beta_0 = \sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n}(x_i - \bar{x}) \tag{108}$$

$$n\beta_0 = n\bar{y} - \beta_1(0) \tag{109}$$

$$\beta_0^{\mathrm{ML}} = \bar{y} \tag{110}$$

Net, obtain MLE of β_1 by differentiating log-likelihood with respect to β_1 and setting to 0,

$$\frac{\partial l}{\partial \beta_1} = -\frac{\sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1(x_i - \bar{x}))(-(x_i - \bar{x}))}{2\sigma^2} = 0 \tag{111}$$

$$\frac{\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1(x_i - \bar{x}))(x_i - \bar{x})}{\sigma^2} = 0 \tag{112}$$

$$\beta_1 \sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x}) = \sum_{i=1}^{n} y_i(x_i - \bar{x}) - \beta_0 \sum_{i=1}^{n}(x_i - \bar{x}) \tag{113}$$

$$\beta_1 \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} y_i(x_i - \bar{x}) - \beta_0(0) \tag{114}$$

$$\beta_1^{\mathrm{ML}} = \frac{\sum_{i=1}^{n} y_i(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{115}$$

N.B. We observe that the ML estimates are the same as the LS estimates.

9.2.10 (j)

Our new model is $y_i - \bar{y} = \beta_0 + \beta_1(x_i - \bar{x}) + \epsilon_i$ for $i = 1, ..., n$. So in this case, following the same process prescribed in (d),

$$Q = \sum_{i=1}^{n}(y_i - \bar{y} - E_\theta(y_i - \bar{y}))^2 \tag{116}$$

$$Q = \sum_{i=1}^{n} (y_i - \bar{y} - b_0 - b_1(x_i - \bar{x}))^2 \tag{117}$$

Differentiate with respect to b_0 and set to 0,

$$\frac{\partial Q}{\partial b_0} = \sum_{i=1}^{n} -2(y_i - \bar{y} - b_0 - b_1(x_i - \bar{x})) = 0 \tag{118}$$

$$nb_0 = \sum_{i=1}^{n} (y_i - \bar{y} - b_1(x_i - \bar{x})) \tag{119}$$

$$nb_0 = n\bar{y} - n\bar{y} - b_1(n\bar{x} - n\bar{x}) \tag{120}$$

$$b_0 = 0 \tag{121}$$

So in this case, LS estimate is $b_0 = 0$.

10

10.1 Problem

Consider the centered *multiple* regression model

$$y_i = \beta_0 + \beta_1(x_{i,1} - \bar{x_1}) + ... + \beta_{p-1}(x_{i,p-1} - \bar{x_{p-1}}) + \epsilon_i \tag{122}$$

with the usual details.

10.1.1 (a)

What is $E_{\boldsymbol{\beta}}(y_i)$?

10.1.2 (b)

What is the least squares estimate of β_0? Show your work.

10.1.3 (c)

For an ordinary uncentered regression model, what is the height of the least squares plane at the point where all x variables are equal to their sample mean values?

10.2 Solution

10.2.1 (a)

$$E_{\boldsymbol{\beta}}(y_i) = \beta_0 + \beta_1(x_{i,1} - \bar{x_1}) + ... + \beta_{p-1}(x_{i,p-1} - \bar{x_{p-1}}) + E(\epsilon_i) \tag{123}$$

$$E_{\boldsymbol{\beta}}(y_i) = \beta_0 + \beta_1(x_{i,1} - \bar{x_1}) + ... + \beta_{p-1}(x_{i,p-1} - \bar{x_{p-1}}) + 0 \tag{124}$$

$$E_{\boldsymbol{\beta}}(y_i) = \beta_0 + \beta_1(x_{i,1} - \bar{x_1}) + ... + \beta_{p-1}(x_{i,p-1} - \bar{x_{p-1}}) \tag{125}$$

10.2.2 (b)

Following the procedure outlined in the previous question,

$$Q = \sum_{i=1}^{n}(y_i - E_{\boldsymbol{\beta}}(y_i))^2 \tag{126}$$

$$Q = \sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1(x_{i,1} - \bar{x_1}) + ... + \beta_{p-1}(x_{i,p-1} - \bar{x_{p-1}})))^2 \tag{127}$$

Now, differentiate Q with respect to β_0 and set to 0 and solve for LS estimate,

$$\frac{\partial Q}{\partial \beta_0} = \sum_{i=1}^{n} 2(y_i - (\beta_0 + \beta_1(x_{i,1} - \bar{x_1}) + ... + \beta_{p-1}(x_{i,p-1} - \bar{x_{p-1}})))(-1) = 0 \tag{128}$$

$$n\beta_0 = \sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n}(x_{i,1} - \bar{x_1}) - ... - \beta_{p-1} \sum_{i=1}^{n}(x_{i,p-1} - \bar{x_{p-1}}) \tag{129}$$

$$n\beta_0 = n\bar{y} - 0 - ... - 0 \tag{130}$$

$$\beta_0^{\text{LS}} = \bar{y} \tag{131}$$

10.2.3 (c)

The ordinary uncentered regression model is written as follows,

$$y_i = \beta_0 + \beta_1 x_{i,1} + ... + \beta_{p-1} x_{i,p-1} + \epsilon_i \tag{132}$$

Rearranging the centered model,

$$y_i = \beta_0 + \beta_1(x_{i,1} - \bar{x_1}) + ... + \beta_{p-1}(x_{i,p-1} - \bar{x_{p-1}}) + \epsilon_i \tag{133}$$

$$y_i = (\beta_0 - \beta_1\bar{x_1} - ... - \beta_{p-1}\bar{x_{p-1}}) + \beta_1 x_{i,1} + ... + \beta_{p-1}x_{i,p-1} + \epsilon_i \tag{134}$$

We essentially get the uncentered model, thus we can assume the OLS parameter estimates are the same, apart from the intercept term. So to arrive at our answer we simply plug in the sample mean values into the centered model,

$$\widehat{y_i^{\text{UC}}} = (b_0^C - b_1^C\bar{x_1} - ... - b_{p-1}^C\bar{x_{p-1}}) + b_1^C\bar{x_1} + ... + b_{p-1}^C\bar{x_{p-1}} = b_0^C = \bar{y} \tag{135}$$

N.B. In the computation above the abbreviation UC corresponds to uncentered and C corresponds to centered.

From the computation above we observe that for an ordinary uncentered regression model, the height of the least squares plane at the point where all x variables are equal to their sample mean values is $\widehat{y_i^{\text{UC}}} = \bar{y}$.

11

11.1 Problem

Suppose that volunteer patients undergoing elective surgery at a large hospital are randomly assigned to one of three different pain killing drugs, and one week after surgery they rate the amount of pain they have experienced on a scale from zero (no pain) to 100 (extreme pain). Write a multiple regression model for these detail; specify how the explanatory variables are defined.

11.2 Solution

Suppose the drugs are labelled drugs 1, 2 and 3. Then, we humbly suggest this model,

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \epsilon_i \tag{136}$$

Where,

- β_0, β_1 and β_2 are unknown constants (parameters).
- $x_{i,1}$ and $x_{i,2}$ are known, observed constants. Specifically $x_{i,1}$ is 1 if patient is administered drug 1 and 0 otherwise, while $x_{i,2}$ is 1 if patient is administered drug 2 and 0 otherwise.
- $\epsilon_1, ..., \epsilon_n$ are unobservable random variables with $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2$ and $Cov(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$.
- σ^2 is an unknown constant (parameter).
- $y_1, ..., y_n$ are observable random variables. Specifically it is the patient's pain rating on a scale from zero (no pain) to 100 (extreme pain).