DAT 401 Project

Yasna Kazimi

2025-09-11

Worked with: "Rahila Jafari"

Quesion 1

```
play_once <- function(){
doors <- c(1:3)

prize_door <- sample(doors, 1)

# selects a door and assign it to picked_door
picked_door <- sample(doors, 1)
#print(picked_door)

# opens a goat door
if (picked_door == prize_door){
opened_door   <- sample(setdiff(doors, c(picked_door, prize_door)), 1)}else{
opened_door <- setdiff(doors, c(picked_door, prize_door))
}
# opens a prize door
switched_door <- setdiff(doors, c(picked_door, opened_door))

#try out the game
return(c(picked_door == prize_door, switched_door == prize_door))
}

# each will play for 10000 times
# play the Monty Hall game without switching
n <- 10000


results <- sapply(1:n,function(j) play_once())


rowMeans(results)
```

```
## [1] 0.336 0.664
```

If you stick with your original door, you win about one-third of the time, but if you switch, you win about two-thirds of the time. The key idea is that Monty's action—opening a door he knows has a goat—gives you useful information. By removing one losing door, he effectively transfers the entire 2/3 chance that your first guess was wrong onto the only unopened door. That's why switching dramatically increases your chances of

winning and is the clearly better strategy.

Question 2

```
# in terms of negative loss
u <- function(a,x) {result <- -5*(x <= a)*(a-x) - 35*(x > a)*(x-a); return (result)}
# Overall distribution
p <- c(4,15,35,5,5,5,5,20,3,3)/100

x <- 1:10


# Forecast probabilities
pr <- 0.20
p_rain  <- c(0,0,0,0,0,1,2,15,1,1)/20
p_sun   <- c(4,15,35,5,5,4,3,5,2,2)/80


# Expectded utility
gstar_r <- which.max(sapply (1:10, function(a) sum(p_rain*u(a,1:10))))

gstar_s <- which.max(sapply (1:10, function(a) sum(p_sun*u(a,1:10))))

ustar_r <-  max(sapply (1:10, function(a) sum(p_rain*u(a,1:10))))
ustar_s <-  max(sapply (1:10, function(a) sum(p_sun*u(a,1:10))))

ustar_r*0.2 + ustar_s*0.8
```

```
## [1] -19.9
```

```
max(sapply (1:10, function(a) sum(p*u(a,1:10))))
```

```
## [1] -20.3
```

When I assume I know the weather in advance, I can pick one optimal order for rainy days and a different optimal order for sunny days. After weighting these two best decisions by the probabilities of rain (20%) and sun (80%), my expected utility comes out to about –19.9. When I do not know the weather, I must choose a single order quantity that works across all days using the overall distribution of demand. In that case, the best expected utility I can get is about –20.3. Since these are losses, higher is better, so –19.9 is better than –20.3. This means having the weather forecast gives us a small but real improvement in expected utility.

Question 3

Answer the following statements TRUE or FALSE, providing a succinct explanation of your reasoning.

(a) You roll two fair three-sided dice. The probability the two dice show the same number is 1/3.
TRUE. There are $3 \times 3 = 9$ outcomes; matching pairs $(1,1), (2,2), (3,3)$ are 3, so $3/9 = 1/3$.

(b) If events A and B are independent and $P(A) > 0$ and $P(B) > 0$, then $P(A \cap B) > 0$.
TRUE. Independence gives $P(A \cap B) = P(A)P(B)$. A product of two positive numbers is positive.

(c) If two events A and B are independent and $P(A) > 0$ and $P(B) > 0$, then A and B cannot be mutually exclusive.

TRUE. Mutually exclusive means $P(A \cap B) = 0$, but independence yields $P(A \cap B) = P(A)P(B) > 0$. Contradiction.

(d) If $P(A \cap B) \geq 0.40$ then $P(A) \leq 0.40$.
FALSE. Always $P(A \cap B) \leq P(A)$, so $P(A) \geq 0.40$ (not ≤ 0.40). Example: $P(A) = 0.8, P(B) = 0.5 \Rightarrow P(A \cap B) = 0.4$.

Question 4

Prove the following properties of expectation directly from the definition.You may assume a finite, discrete sample space for simplicity.

(a)

$$E(a + X) = a + E(X),$$

$$E(a + X) = \sum_x (a + X(x))\, p(x).$$

$$E(a + X) = \sum_x a\, p(x) \;+\; \sum_x X(x)\, p(x).$$

$$E(a + X) = a \sum_x p(x) \;+\; E(X).$$

$$\sum_x p(x) = 1,$$

$$E(a + X) = a + E(X).$$

(b) Adding a constant:

$$\begin{aligned} E[a + X] &= \sum_{s \in S} \big(a + X(s)\big)\, p(s) \\ &= \sum_{s \in S} a\, p(s) \;+\; \sum_{s \in S} X(s)\, p(s) \\ &= a \sum_{s \in S} p(s) \;+\; E[X] \\ &= a \cdot 1 + E[X] \\ &= a + E[X] \end{aligned}$$

(c) Show that:

$$E(X + Y) = E(X) + E(Y).$$

$$E(X + Y) = \sum_s (X(s) + Y(s))\, p(s).$$

$$E(X + Y) = \sum_s X(s) p(s) \;+\; \sum_s Y(s) p(s).$$

$$E(X + Y) = E(X) + E(Y).$$

Question 5

	$X=0$	$X=1$
$Y=0$	0.5	0.2
$Y=1$	0.2	0.1

We verify that $0.5+0.2+0.2+0.1=1$.

$$E(XY)=\sum_x\sum_y xy\,p(x,y)$$

$$E(XY)=1\cdot 1\cdot 0.1=0.1.$$

$$E(X)=0(0.5+0.2)+1(0.2+0.1)=0.3, \qquad E(Y)=0(0.5+0.2)+1(0.2+0.1)=0.3.$$

$$E(X)E(Y)=0.3\times 0.3=0.09.$$

$$E(XY)=0.1\neq 0.09=E(X)E(Y)$$

Question 6

$$E(XY)=\sum_x\sum_y xyP(X=x,Y=y)$$

$$E(X)=\sum_x xP(X=x)$$

$$E(Y)=\sum_y yP(Y=y)$$

$$\Rightarrow E(XY)=E(X)E(Y).$$

Q7.

I considered a finite, discrete sample space omega = {1,2,3,4} with probabilities p = (0.1, 0.3, 0.2, 0.4) that will equal to 1.

$$V(aX)=E[(aX)^2]-\big(E[aX]\big)^2$$

(a)

$$=E(a^2X^2)-E(aX)\cdot E(aX)$$
$$=a^2E(X^2)-aE(X)\cdot aE(X)$$
$$=a^2E(X^2)-a^2E(X)\cdot E(X)$$
$$=a^2E(X^2)-a^2[E(X)]^2$$
$$=a^2[E(X^2)-E(X)^2]\Rightarrow V(aX)=a^2V(X)$$

(b) V(a + X) = V(X)

$$V(a+X)=E(a+X)^2-[E(a+X)]^2$$
$$=E(a^2+X^2+2aX)-E(a+X)\cdot E(a+X)$$
$$=E(a^2)+E(X^2)+E(2aX)-[E(a)+E(X)]\cdot[E(a)+E(X)]$$

$$E(a^2) = a^2, \quad E(2aX) = 2aE(X), \quad E(a) = a$$

$$= a^2 + E(X^2) + 2aE(X) - [a + E(X)] \cdot [a + E(X)]$$

$$= a^2 + E(X^2) + 2aE(X) - [a^2 + [E(X)]^2 + 2aE(X)]$$
$$= a^2 + E(X^2) + 2aE(X) - a^2 - [E(X)]^2 - 2aE(X)$$
$$E(X^2) - [E(X)]^2 \Rightarrow V(X)$$

(c) V(aX + bY) = aˆ2V(X) + bˆ2V(Y) + 2abCov(X, Y)

$$V(aX + bY) = E(aX + bY)^2 - [E(aX + bY)]^2$$
$$= E(a^2X^2 + b^2Y^2 + 2abXY) - E(aX + bY) \cdot E(aX + bY)$$
$$= E(a^2X^2) + E(b^2Y^2) + E(2abXY) - [aE(X) + bE(Y)] \cdot [aE(X) + bE(Y)]$$
$$= a^2E(X^2) + b^2E(Y^2) + 2ab(XY) - [a^2[E(X)]^2 + abE(X)E(Y) + b^2[E(Y)]^2 + abE(XE(Y))]$$
$$= a^2E(X^2) + b^2E(Y^2) + 2abE(XY) - [a^2[E(X)]^2 + b^2[E(Y)]^2 + 2abE(X)E(Y)]$$
$$= a^2E(X^2) + b^2E(Y^2) + 2abE(XY) - a^2[E(X)]^2 - b^2[E(Y)]^2 - 2abE(X)E(Y)$$
$$= a^2[E(X^2) - E(X)^2] + b^2[E(X^2) - E(X)^2] + 2ab[E(XY) - E(X)E(Y)]$$
$$Cov(X, Y) = [E(XY) - E(X)E(Y)] \Rightarrow a^2V(X) + b^2V(Y) + 2abCov(X, Y)$$

Question 8

Derive the identity:V(X) = E(Xˆ2) - E(X)ˆ2

$$V(X) = E(X - E(X))^2$$

$$= E(X^2 - 2XE(X) + (E(X))^2)$$

$$= E(X^2) - 2E(X)E(X) + (E(X))^2$$

$$= E(X^2) - E(X)^2$$

Question 9

```r
p <- c(0.1, 0.3, 0.2, 0.4)

# Random variables
X <- c(-1, 0, 2, 4)

E <- function(z) sum(z * p)
Var <- function(z) E(z^2) - E(z)^2


mu <- E(X)
sigma <- sqrt(Var(X))
Z <- (X - mu)/sigma
EZ <- sum(Z * p)       # should be 0
VZ <- Var(Z)          # should be 1

cat("E[Z] =", EZ, "\n")           # The outcome would not be zero but VERY small number.
```

```
## E[Z] = -1.665335e-16
```

```r
cat("Var(Z) =", VZ, "\n")
```

```
## Var(Z) = 1
```

Question 10

Correlation Formula

$$\text{corr}(X, Y) = \frac{E[XY] - E[X]E[Y]}{\sigma_X \sigma_Y}.$$

Part (a):

$$\text{Cov}(X, Y) = E[XY] - E[X]E[Y] = 0,$$

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = 0.$$

Part (b):

$$E[X] = 0, \quad E[Y] = E[X^2] = \frac{1}{3}, \quad E[XY] = E[X^3] = 0.$$

$$\text{Cov}(X, Y) = E[XY] - E[X]E[Y] = 0,$$

$$\rho_{X,Y} = 0.$$

Question 11

Proof: Law of Iterated Expectation

$$E[X] = \sum_x x\, p(x)$$

$$p(x) = \sum_y p(x, y) = \sum_y p(x \mid y)p(y)$$

$$E[X] = \sum_x x \sum_y p(x \mid y)p(y)$$

$$E[X] = \sum_y p(y) \left(\sum_x x p(x \mid y) \right)$$

$$E[X \mid Y = y] = \sum_x x p(x \mid y)$$

$$E[X] = \sum_y p(y) E[X \mid Y = y]$$

Finally:

$$E[X] = E(E[X \mid Y])$$

Question 12

```r
# Probabilities
p <- c(small = 0.2, medium = 0.5, large = 0.3)

# Payoffs (millions)
standard <- c(small = 20, medium = 30, large = 50)
horizontal <- c(small = -20, medium = 40, large = 90)

# (a) Means
E_standard <- sum(p * standard)
E_horizontal <- sum(p * horizontal)

# (b) Variance & SD
Var_standard <- sum(p * (standard - E_standard)^2)
Var_horizontal <- sum(p * (horizontal - E_horizontal)^2)
sd_standard <- sqrt(Var_standard)
sd_horizontal <- sqrt(Var_horizontal)


# (d) EVPI
# best payoff in each state
best_in_state <- pmax(standard, horizontal)
E_with_perfect_info <- sum(p * best_in_state)
EVPI <- E_with_perfect_info - max(E_standard, E_horizontal)
```

```r
# Print results
cat(sprintf("Means (millions): Standard = %.2f, Horizontal = %.2f\n", E_standard, E_horizontal))
```

```
## Means (millions): Standard = 34.00, Horizontal = 43.00
```

```r
cat(sprintf("Variance: Standard = %.3f, Horizontal = %.3f\n", Var_standard, Var_horizontal))
```

```
## Variance: Standard = 124.000, Horizontal = 1461.000
```

```r
cat(sprintf("SD (millions): Standard = %.2f, Horizontal = %.2f\n", sd_standard, sd_horizontal))
```

```
## SD (millions): Standard = 11.14, Horizontal = 38.22
```

```r
cat(sprintf("E with perfect info = %.2f million; EVPI = %.2f million\n", E_with_perfect_info, EVPI))
```

```
## E with perfect info = 51.00 million; EVPI = 8.00 million
```

Part (c)

A useful way to compare the two drilling strategies is to think about how the company approaches risk. If the company is **risk-averse**, it cares most about avoiding losses. Using the Maximin rule—which selects the option with the best worst-case outcome—the Standard Drill is preferred: its minimum payoff is \$20 million, while the Horizontal Drill's worst outcome is a loss of \$20 million. A cautious decision-maker would therefore choose the Standard Drill.

If the company is **risk-seeking**, it focuses on the highest possible payoff. Under the Maximax rule—which selects the option with the greatest upside—the Horizontal Drill is more attractive, since it can return up to \$90 million, compared to a maximum of \$50 million for the Standard Drill.

A **risk-neutral** company bases its decision on expected value, and since the Horizontal Drill has the higher average payoff, it would again be the preferred option.

In summary:

- **Risk-averse:** choose the Standard Drill (safer minimum payoff).
- **Risk-seeking:** choose the Horizontal Drill (higher upside).
- **Risk-neutral:** choose the Horizontal Drill (higher expected value).

Part (d)

From my calculations, if I had perfect geological information that told me exactly how much oil was at the site, my expected payoff would be \$51 million. Without that information, the best I can do is choose the drilling method with the highest expected payoff, which is the horizontal drill at \$43 million. The difference between these two values is \$8 million, and that's what the perfect information is actually worth to me. In other words, having a guaranteed, accurate forecast would improve my expected outcome by \$8 million. So the most I'd ever be willing to pay for a geological evaluation that reveals the true state of the oil site is \$8 million; anything more than that wouldn't be worth it.

Question 13

S: email is spam
$\bar{S}$: email is not spam
B: email contains "bigger"

$$P(S) = \frac{550}{100000} = 0.005, \quad P(\bar{S}) = 0.9945$$
$$P(B \mid S) = 0.5, \quad P(B \mid \bar{S}) = 0.02$$

Using Bayes' Rule:

$$P(S \mid B) = \frac{P(B \mid S)P(S)}{P(B \mid S)P(S) + P(B \mid \bar{S})P(\bar{S})}.$$
$$P(S \mid B) = \frac{(0.5)(0.005)}{(0.5)(0.005) + (0.02)(0.9945)} = \frac{0.0025}{0.0025 + 0.01989} = \frac{0.0025}{0.0224} \approx 0.1215.$$

The probability that this new email is spam, given it contains **"bigger"**, is approximately:

$$P(S \mid B) = 0.1215$$

Question 14

M: miscarriage,
L: low-risk pregnancy,
H: high-risk pregnancy

$$P(M) = 0.15, \quad P(H) = 0.02, \quad P(L) = 0.98, \quad P(M \mid H) = 0.80.$$

(a)

Law of Total Probability:

$$P(M) = P(M \mid L)P(L) + P(M \mid H)P(H).$$
$$0.15 = P(M \mid L)(0.98) + (0.80)(0.02).$$

$$P(M \mid L) = \frac{0.15 - 0.016}{0.98} = 0.1367.$$

(b)

Bayes' Rule

$$P(H \mid M) = \frac{P(M \mid H)P(H)}{P(M)} = \frac{(0.80)(0.02)}{0.15} = 0.1067.$$

(c)

$$P(M_2 \mid M_1) = P(M_2 \mid M_1, L)P(L \mid M_1) + P(M_2 \mid M_1, H)P(H \mid M_1).$$

$$P(M_2 \mid M_1, L) = P(M \mid L), \quad P(M_2 \mid M_1, H) = P(M \mid H).$$
$$P(M_2 \mid M_1) = (0.80)(0.1067) + (0.1367)(0.8933)$$
$$P(M_2 \mid M_1) = 0.08536 + 0.12214 \approx 0.2075$$

After a woman has one miscarriage, the chance of having another one becomes higher than the usual population average. The reason is that the first miscarriage gives us new information about her risk level. Even though only 2% of all pregnancies belong to the high-risk group, high-risk pregnancies have an 80% miscarriage rate. So when a miscarriage happens, the probability that the woman belongs to this high-risk group increases. Once we update this risk, the overall chance of a second miscarriage becomes roughly 20.7%, which is higher than the original 15% miscarriage rate in the general population. In short: the first miscarriage shifts some probability toward being in the high-risk category, and that change is what raises the likelihood of a second miscarriage.

Question 15

(a) Probability of exactly 5 correct out of 7:

$$P(X = 5) = \binom{7}{5}(0.5)^7 = 21 \times \left(\frac{1}{2}\right)^7 = \frac{21}{128} \approx 0.164$$

(b) Probability of 5 or more correct out of 7:

$$P(Y \geq 5) = \sum_{k=5}^{7} \binom{7}{k} \left(\frac{1}{2}\right)^7 = P(Y = 5) + P(Y = 6) + P(Y = 7) \approx 0.2265625.$$

(c) After 28 trials the friend gets 20 correct. Under $p = \frac{1}{2}$,

$$\text{Total number of trials increases to } n = 28.$$
$$Y = \text{number of correct choices} = 0, 1, 2, \ldots, 27, 28.$$

$$p = \frac{1}{2}, \qquad Y \sim \text{Binomial}(28, \tfrac{1}{2}).$$

$$P(Y = y) = \binom{28}{y} \left(\tfrac{1}{2}\right)^y \left(\tfrac{1}{2}\right)^{28-y} = \binom{28}{y} \left(\tfrac{1}{2}\right)^{28}.$$
$$P(Y \geq 20) = P(Y = 20) + P(Y = 21) + \cdots + P(Y = 28).$$

$$\begin{aligned} P(Y \geq 20) = &\binom{28}{20} \left(\tfrac{1}{2}\right)^{20} + \binom{28}{21} \left(\tfrac{1}{2}\right)^{21} + \binom{28}{22} \left(\tfrac{1}{2}\right)^{22} \\ &+ \binom{28}{23} \left(\tfrac{1}{2}\right)^{23} + \binom{28}{24} \left(\tfrac{1}{2}\right)^{24} + \binom{28}{25} \left(\tfrac{1}{2}\right)^{25} \\ &+ \binom{28}{26} \left(\tfrac{1}{2}\right)^{26} + \binom{28}{27} \left(\tfrac{1}{2}\right)^{27} + \binom{28}{28} \left(\tfrac{1}{2}\right)^{28}. \end{aligned}$$

$$P(Y \geq 20) \approx 0.0178$$

So the probability of observing 20 or more correct by chance (if $p = 1/2$) is about **1.78%**.

(d) Although the observed proportions are the same (5/7 and 20/28 both approximately 71.4%), the probabilities under the null differ because the sampling variability depends on sample size:

- With a **small sample size** ($n = 7$), the binomial distribution is wide and extreme proportions are not very unlikely. As a result, $P(Y \geq 5)$ is fairly large (approximately 22.7%).
- With a **larger sample size** ($n = 28$), observing the same proportion (around 71%) is much less likely under $p = 1/2$. The binomial distribution becomes much more concentrated around its mean np, so deviations of this size are far more significant, giving a much smaller tail probability (about 1.78%).

This illustrates the importance of **sample size** in statistical evidence: larger samples reduce sampling variability and make the same observed proportion more or less surprising under the null hypothesis (law of large numbers, sampling variability, statistical power).

Question 16

$$\mu_A = 3\%,\ \sigma_A = 1\%, \qquad \mu_B = 10\%,\ \sigma_B = 10\%, \qquad \mu_C = 15\%,\ \sigma_C = 20\%.$$

Correlations:

$$\rho_{A,B} = 0, \quad \rho_{A,C} = 0, \quad \rho_{B,C} = 0.7.$$

(a) Expected Values

$$Y_1 = 0.25X_A + 0.75X_B, \qquad Y_2 = 0.5X_A + 0.5X_C, \qquad Y_3 = 0.5X_B + 0.5X_C.$$

$$E[Y] = w_A\mu_A + w_B\mu_B + w_C\mu_C.$$

$$E[Y_1] = 0.25(3) + 0.75(10) = 8.25\%,$$
$$E[Y_2] = 0.5(3) + 0.5(15) = 9\%,$$
$$E[Y_3] = 0.5(10) + 0.5(15) = 12.5\%.$$

(b) Variances

Variance of a two-asset portfolio:

$$\sigma_Y^2 = w_1^2\sigma_1^2 + w_2^2\sigma_2^2 + 2w_1w_2\rho_{12}\sigma_1\sigma_2.$$

Compute each:

1. $Y_1 = 0.25X_A + 0.75X_B$ ($\rho_{A,B} = 0$):
$$\sigma_{Y_1}^2 = (0.25)^2(1^2) + (0.75)^2(10^2) = 0.0625 + 56.25 = 56.3125$$
$$\sigma_{Y_1} = 7.50\%.$$

2. $Y_2 = 0.5X_A + 0.5X_C$ ($\rho_{A,C} = 0$):
$$\sigma_{Y_2}^2 = (0.5)^2(1^2) + (0.5)^2(20^2) = 0.25 + 100 = 100.25$$
$$\sigma_{Y_2} = 10.01\%.$$

3. $Y_3 = 0.5X_B + 0.5X_C$
   ($\rho_{B,C} = 0.7$):

$$\sigma^2_{Y_3} = (0.5)^2(10^2) + (0.5)^2(20^2) + 2(0.5)(0.5)(0.7)(10)(20)$$
$$= 25 + 100 + 70 = 195,$$
$$\sigma_{Y_3} = 13.96\%.$$

(c) 95% Ranges

Assuming normality, approximately 95% of outcomes lie within $\mu_Y \pm 1.96\sigma_Y$.

$$Y_1 : 8.25 \pm 1.96(7.50) = [-6.45,\ 22.95]\%,$$
$$Y_2 : 9 \pm 1.96(10.01) = [-10.62,\ 28.62]\%,$$
$$Y_3 : 12.5 \pm 1.96(13.96) = [-15.86,\ 40.86]\%.$$

Portfolio Y_1 (A & B) has the smallest variance due to the uncorrelated assets.
Portfolio Y_3 (B & C) has the highest expected return but also the greatest risk because the assets are strongly correlated.

Question 17

Suppose that historically my quantitative exam scores follow a normal distribution with mean $\mu = 85$ and standard deviation $\sigma = 5$. I received three scores below 75 on the past three exams, but I do **not** believe my underlying ability has changed.
Under this assumption, each exam score is an independent draw from the same distribution. The expected value of my next score is therefore the population mean:

$$E[X_{\text{next}} \mid X_1, X_2, X_3] = E[X_{\text{next}}] = \mu = 85.$$

Even though I recently performed below average, if my ability has not changed, my best prediction for the next score (the mean of the distribution) remains **85%**.

Question 18

I make putts with probability $p_1 = 0.7$; My brother with $p_2 = 0.8$.

(a)

The probability that I make my **first three** putts and **miss** the fourth:

$$P = (0.7)^3(0.3) = 0.1029.$$

(b)

The probability that I make **exactly three** putts in my first four attempts:

$$P(Y = 3) = \binom{4}{3}(0.7)^3(0.3) = 4(0.7)^3(0.3) = 0.4116.$$

(c)

The probability that my brother makes **more than half** of his first four putts (i.e., at least 3 successes):

$$P(Y \geq 3) = \binom{4}{3}(0.8)^3(0.2) + \binom{4}{4}(0.8)^4 = 4(0.512)(0.2) + 0.4096 = 0.8192.$$

(d)

$X \sim \text{Binomial}(100, 0.7)$ = my number of makes,
$Y \sim \text{Binomial}(100, 0.8)$ = my brother's number of makes.

Expected amounts:

$$E[X] = 100(0.7) = 70, \qquad E[Y] = 100(0.8) = 80.$$

$$E[5(Y - X)] = 5(E[Y] - E[X]) = 5(80 - 70) = 5(10) = 50.$$

So **I owe my brother \$50 on average**.

(e)

Used the **normal approximation** to the binomial.

$$X \approx N(100(0.7),\, 100(0.7)(0.3)), \qquad Y \approx N(100(0.8),\, 100(0.8)(0.2)).$$

$$D \sim N(\mu_D, \sigma_D^2),$$
$$\mu_D = 100(0.7 - 0.8) = -10, \qquad \sigma_D^2 = 100(0.7)(0.3) + 100(0.8)(0.2) = 21 + 16 = 37.$$

$$Z = \frac{0 - (-10)}{\sqrt{37}} = \frac{10}{6.083} \approx 1.64,$$

$$P(X - Y > 0) = P(Z > 1.64) \approx 0.05.$$

So there is roughly a **5% chance I win** the contest.

(Using R: `1 - pnorm(0, mean = -10, sd = sqrt(37))`*)*

Question 19

A professional Baseball player with drinking problem.

(a) Overall Probabilty of getting a hit:

$$P(hit) = P(hit \mid D)P(D) + P(hit \mid S)P(S) + P(hit \mid H)P(H)$$

$$P(hit) = (0.05)(0.15) + (0.30)(0.80) + (0.21)(0.05) = 0.258$$

(b) Hungover and got a hit:

$$P(H \mid \text{hit}) = \frac{P(\text{hit} \mid H)\, P(H)}{P(\text{hit})}$$

$$P(H - over \mid Hit) = \frac{(0.21)(0.05)}{0.258} = 0.0407$$

(c) He had 3 at-bats and got 2 hits, the probability that he was sober:

$$P(2 \mid s) = \binom{3}{2} p_s^2 (1 - p_s).$$

$$P(2 \mid \text{drunk}) = 3(0.05)^2(0.95) = 0.007125,$$
$$P(2 \mid \text{sober}) = 3(0.30)^2(0.70) = 0.189,$$
$$P(2 \mid \text{hung}) = 3(0.21)^2(0.79) = 0.104517.$$

$$P(2) = 0.15(0.007125) + 0.80(0.189) + 0.05(0.104517)$$
$$= 0.00106875 + 0.1512 + 0.00522585$$
$$\approx 0.1575.$$

Thus,

$$\boxed{P(\text{exactly 2 hits}) \approx 0.1575.}$$

$$P(X = 2) = 0.00035625 + 0.1512 + 0.00522585 \approx 0.1567821$$

(d)

$$P(\text{sober} \mid X = 2) = \frac{0.00522585}{0.1567821} \approx 0.033.$$

Qestion 20

(a)

$$Z = \frac{-1.5 - 0}{0.4} = -3.75, \qquad p = P(Z < -3.75) \approx 8.84 \times 10^{-5}.$$

The p-value is about $8.84 \times 10^{-5} < 0.05$. The new drug reduced headache duration by an average of 1.5 minutes compared to the placebo. If the drug actually had no effect, we would expect the average difference to be around zero, with onlyrandom variation from sample to sample. But the observed result is 3.75 standard errors below zero, which is extremely far into the tail of the normal distribution. The chance of getting a result this extreme just by luck is about 8.8×10^{-5}, which is almost zero. Because we are only interested in whether the new drug *reduces* headache duration (and not whether it might increase it), a one-sided test is appropriate. Overall, this very small p-value tells us that the observed improvement is highly unlikely to be due to chance, providing strong evidence that the new medication genuinely works better than the placebo.

(b)

Yes, the new drug is statistically significantly better than the placebo at the 5% level. The p-value we obtained in part (a) was extremely small (approximately 8.8×10^{-5}), which is far below the 0.05 threshold. This means that seeing an improvement as large as the one observed would be highly unlikely if the drug had no real effect. Therefore, the study provides strong evidence that the new drug reduces headache duration more effectively than the placebo.

(c)

Since the p-value in this study is extremely small, we already have very strong statistical evidence that the new drug reduces headache duration. Collecting more data is not strictly necessary to establish efficacy, because the current sample of 100 patients already provides a precise estimate with a large and highly significant effect. However, additional data could still be useful for refining the estimated size of the improvement or for checking that the effect is consistent across different patient groups, but it is not required to demonstrate that the drug works.

(d)

Based on the analysis, I would recommend continuing development of the drug because the statistical evidence strongly suggests it reduces headache duration. However, the actual improvement is only about 1.5 minutes, so more data are needed to confirm whether this benefit is clinically meaningful. A larger follow-up study would help verify the effect size and ensure the drug is safe across different types of patients. Overall, the results justify moving forward, but not making any final decisions yet.

Question 21

The XKCD cartoon is making fun of how scientific results can be misleading when researchers test lots of different hypotheses without adjusting for it. In the comic, scientists test whether each of 20 different jelly bean colors causes acne, and every color shows “no link” except green — a result that appears “significant” but is actually just a false positive caused by random chance. When you run many tests at the 5% significance level, you expect some false alarms purely through luck, yet the comic shows this one fluke being treated as a real discovery and turned into a dramatic headline. The joke highlights how easy it is to “find” significance when you look in enough places, and how this misunderstanding can lead to false claims in the news when people don’t appreciate the multiple-testing problem.

Question 22

$$X = N(\mu, \sigma^2)$$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$f(x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}\right)$$

$$\mathrm{f}(x_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}\right)$$

$$\mathrm{M}MLE = \arg\max \mu \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}\right)$$

$$\mathrm{argMax} = \log\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}\right)\right)$$

$$\frac{d}{dm} = \sum_{i=1}^{n} \frac{1}{2} log(2\pi\sigma^2) - \sum_{i=1}^{n} \frac{1}{2}(\frac{(x_i-\mu)^2}{\sigma^2})$$

$$\sum_{i=1}^{n} \frac{(x-\mu)}{\sigma^2} = 0$$

$$\frac{\sum_{i=1}^{n} x_i}{\sigma^2} = \frac{n\mu}{\sigma^2}$$

$$\Rightarrow \sum_{i=1}^{n} x_i = n\mu \Rightarrow \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Question 23

$$S = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \qquad \text{where} \qquad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

$$\begin{aligned}
S &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i^2 - 2X_i\bar{X} + \bar{X}^2) \\
&= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \frac{2\bar{X}}{n}\sum_{i=1}^{n} X_i + \frac{1}{n}\sum_{i=1}^{n}\bar{X}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - 2\bar{X}^2 + \bar{X}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2.
\end{aligned}$$

$$
\begin{aligned}
E[S] &= E\left[\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2\right] \\
&= E\left[\frac{1}{n}\sum_{i=1}^{n} X_i^2\right] - E[\bar{X}^2] \\
&= \frac{1}{n}\sum_{i=1}^{n} E[X_i^2] - E[\bar{X}^2] \\
&= E[X^2] - E[\bar{X}^2].
\end{aligned}
$$

$$
E[\bar{X}^2] = \mathrm{Var}(\bar{X}) + (E[\bar{X}])^2 = \frac{\mathrm{Var}(X)}{n} + (E[X])^2.
$$

$$
\mathrm{Var}(X) = E[X^2] - (E[X])^2 \quad \Rightarrow \quad E[X^2] = \mathrm{Var}(X) + (E[X])^2.
$$

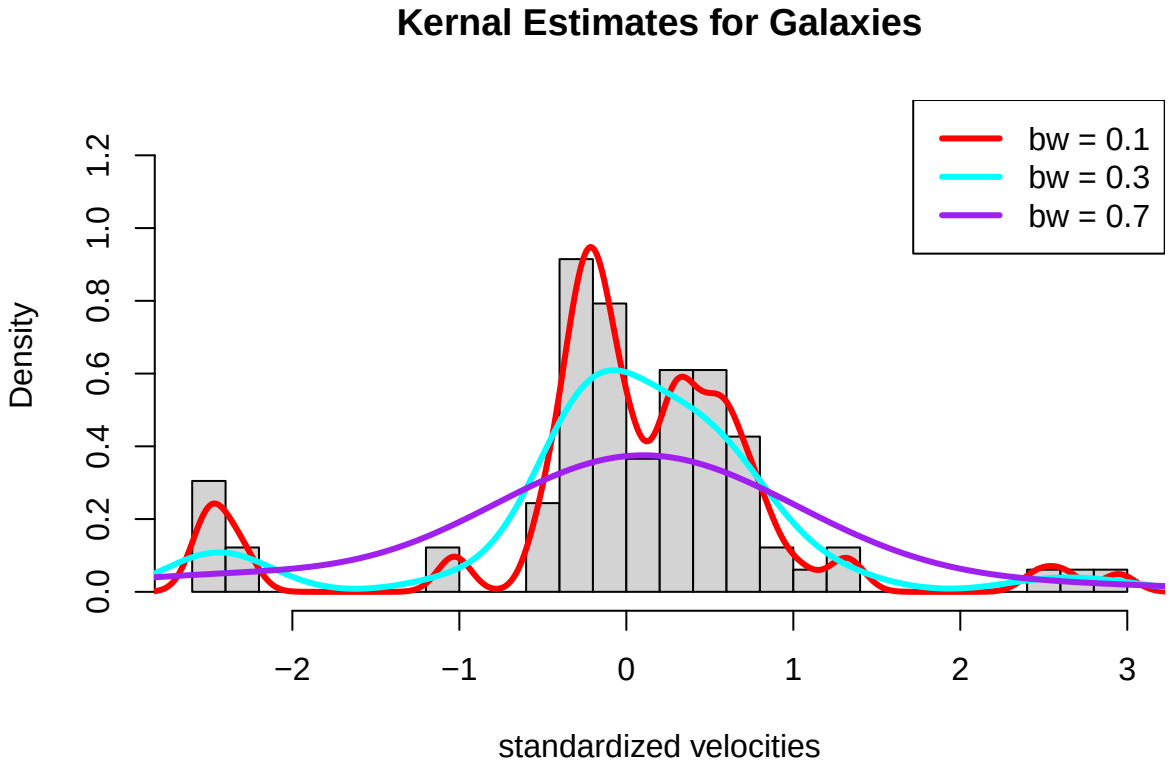
$$
\begin{aligned}
E[S] &= \mathrm{Var}(X) + (E[X])^2 - \left(\frac{\mathrm{Var}(X)}{n} + (E[X])^2\right) \\
&= \mathrm{Var}(X)\left(1 - \frac{1}{n}\right) \\
&= \frac{n-1}{n}\,\mathrm{Var}(X).
\end{aligned}
$$

$$
S' = \frac{n}{n-1}S = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2,
$$

$$
E[S'] = \frac{n}{n-1}E[S] = \frac{n}{n-1} \cdot \frac{n-1}{n}\,\mathrm{Var}(X) = \mathrm{Var}(X).
$$

Question 24

```r
library(MASS)
data(galaxies)

galaxies <- galaxies - mean(galaxies)
galaxies <- galaxies / sd(galaxies)

# histogram background
hist(galaxies, 30, freq = FALSE,
     ylim = c(0, 1.3),
     main = "Kernal Estimates for Galaxies",
     xlab = "standardized velocities")

# density() with different bandwidths
d1 <- density(galaxies, bw = 0.1)
d2 <- density(galaxies, bw = 0.3)
d3 <- density(galaxies, bw = 0.8)

# overlay curves
lines(d1, col = "red", lwd = 3)
lines(d2, col = "cyan", lwd = 3)
lines(d3, col = "purple", lwd = 3)


legend("topright",
       legend = c("bw = 0.1", "bw = 0.3", "bw = 0.7"),
       col = c("red", "cyan", "purple"),
       lty = 1, lwd = 3)
```



When we change the bandwidth in a kernel density estimate, we're basically changing how "tightly" or "loosely" we smooth the data. A small bandwidth makes the curve very spiky — it tries to follow every little bump in the data, even the random noise. A medium bandwidth gives a nicer balance: the curve is smoother but still shows the main shapes and peaks. A large bandwidth, however, smooths things so much that important details disappear, and everything starts to blend together.

Question 25

```
# Loading the dataset
df <- read.table("~/Downloads/MidCity.txt", header = TRUE)
head(df)
```

```
##   Home Nbhd Offers SqFt Brick Bedrooms Bathrooms  Price
## 1    1    2      2 1790    No        2         2 114300
## 2    2    2      3 2030    No        4         2 114200
## 3    3    2      1 1740    No        3         2 114800
## 4    4    2      3 1980    No        3         2  94700
## 5    5    2      3 2130    No        3         3 119800
## 6    6    1      2 1780    No        3         2 114600
```

```
df$Brick <- factor(df$Brick)
```

(a) Is there a premium for Brick houses?

```
model1 <- lm( Price ~ Brick, data = df)
summary(model1)
```

```
##
## Call:
## lm(formula = Price ~ Brick, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -52858 -16758  -3564  18781  63431
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   121958       2594  47.024  < 2e-16 ***
## BrickYes       25811       4528   5.701 8.02e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24050 on 126 degrees of freedom
## Multiple R-squared:  0.205,  Adjusted R-squared:  0.1987
## F-statistic:  32.5 on 1 and 126 DF,  p-value: 8.023e-08
```

Yes, there is a premium for brick houses. In the simple model where price is explained only by Brick, the coefficient is about $25,811, meaning brick houses sell for roughly $26k more than non-brick houses on average. However, this premium changes when we fit models with more predictors (like SqFt, bedrooms, bathrooms, neighborhood). This is because the simple model gives brick credit for differences that are actually due to other home features. Once those factors are included, the estimated brick effect adjusts, since we are now comparing brick and non-brick homes that are similar in size and quality.

(b) Is there a premium for houses in Neighborhood 3?

```
df$Brick <- factor(df$Brick)
df$Nbhd  <- factor(df$Nbhd)

model_b <- lm(Price ~ factor(Nbhd), data = df)
summary(model_b)
```

```
##
## Call:
## lm(formula = Price ~ factor(Nbhd), data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -42931 -12310  -1643  11251  51905
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     110154       2693  40.904  < 2e-16 ***
## factor(Nbhd)2    15077       3787   3.981 0.000116 ***
## factor(Nbhd)3    49140       3929  12.508  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17860 on 125 degrees of freedom
## Multiple R-squared:  0.565,  Adjusted R-squared:  0.558
## F-statistic: 81.16 on 2 and 125 DF,  p-value: < 2.2e-16
```

When I treat neighborhood rating as a categorical variable (which is the correct way to handle quality ratings), the model shows that houses in Neighborhood 3 sell for about $49,140 more than houses in Neighborhood 1. This number comes directly from the regression coefficient for factor(Nbhd)3. In other words, being in the highest-quality neighborhood adds roughly fifty thousand dollars to a home's price, on average. This makes sense because higher-rated neighborhoods generally have better amenities, better location, and higher demand, which pushes prices up.

(c) Is there an extra premium for Brick houses in Neighborhood 3?

```
df$Brick <- factor(df$Brick)
df$Nbhd  <- factor(df$Nbhd)

model3 <- lm(Price ~ Brick * Nbhd + SqFt + Bedrooms + Bathrooms + Offers, data = df)
summary(model3)
```

```
##
## Call:
## lm(formula = Price ~ Brick * Nbhd + SqFt + Bedrooms + Bathrooms +
##     Offers, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27225.1  -5219.0   -273.7   4297.4  27507.2
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3695.511   8829.382   0.419  0.67631
```

```
## BrickYes       12093.056   4082.168   2.962  0.00369 **
## Nbhd2          -1317.656   2679.849  -0.492  0.62385
## Nbhd3          16980.797   3437.529   4.940 2.60e-06 ***
## SqFt              53.745      5.686   9.453 3.96e-16 ***
## Bedrooms        4777.216   1586.397   3.011  0.00318 **
## Bathrooms       6457.287   2160.867   2.988  0.00341 **
## Offers         -8381.770   1068.248  -7.846 2.15e-12 ***
## BrickYes:Nbhd2  2668.449   5068.893   0.526  0.59957
## BrickYes:Nbhd3 11933.197   5341.027   2.234  0.02735 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9847 on 118 degrees of freedom
## Multiple R-squared:  0.8752, Adjusted R-squared:  0.8657
## F-statistic: 91.94 on 9 and 118 DF,  p-value: < 2.2e-16
```

Yes, there is an extra premium for brick houses in Neighborhood 3. The interaction term BrickYes:Nbhd is about $9,439, meaning brick homes in Neighborhood 3 sell for roughly nine thousand dollars more than we would expect from simply adding the overall brick effect and the overall neighborhood effect. This interaction term is statistically significant, so the combination of being both brick and located in the highest-rated neighborhood provides an additional boost to price beyond the separate effects of brick and neighborhood quality.

(d)

```r
df <- read.table("~/Downloads/MidCity.txt", header = TRUE)
df$Brick <- factor(df$Brick)
df$Nbhd  <- factor(df$Nbhd)



df$Nbhd_combined <- ifelse(df$Nbhd == 3, "3", "1or2")

model_nbhd2 <- lm(
  Price ~ SqFt + Bedrooms + Bathrooms + Brick + factor(Nbhd_combined) + Offers,
  data = df
)

summary(model_nbhd2)
```

```
##
## Call:
## lm(formula = Price ~ SqFt + Bedrooms + Bathrooms + Brick + factor(Nbhd_combined) +
##     Offers, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26810.5  -5953.6   -266.5   5662.9  26793.0
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             3067.471   8746.712   0.351 0.726423
## SqFt                      52.149      5.572   9.359 5.44e-16 ***
## Bedrooms                4070.005   1570.921   2.591 0.010751 *
```

```
## Bathrooms                7810.698   2109.060   3.703 0.000322 ***
## BrickYes                17058.771   1942.805   8.780 1.28e-14 ***
## factor(Nbhd_combined)3 21937.572   2482.393   8.837 9.39e-15 ***
## Offers                 -8019.003   1013.011  -7.916 1.32e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9995 on 121 degrees of freedom
## Multiple R-squared:  0.8682, Adjusted R-squared:  0.8616
## F-statistic: 132.8 on 6 and 121 DF,  p-value: < 2.2e-16
```

(d) Does combining neighborhoods 1 and 2 diminish the ability of the model to predict house prices?

To see whether combining neighborhoods 1 and 2 affects the model's predictive ability, I refit the regression using a new variable that collapses these two neighborhoods into a single category ("1or2") while leaving neighborhood~3 separate. In this combined model, the residual standard error increases to **9995** and the R^2 value drops to **0.8682**. Both of these changes indicate a worse fit compared to the original model, which had a lower residual error and a slightly higher R^2. This means that keeping neighborhoods 1 and 2 separate captures important differences in housing prices that are lost when they are merged. Therefore, combining these neighborhoods diminishes the model's ability to explain and predict house prices accurately.

Question 26

```
df <- read.csv("~/Downloads/oakland_As.txt", header = TRUE)
head(df)
```

```
##   TICKET OPP POS GB DOW TEMP PREC TOG TV PROMO NOBEL WKEND OD DH
## 1  24415   2   5  1   4   57    0   2  0     0     0     0  1  0
## 2   5729   2   3  1   5   66    0   2  0     0     0     1  0  0
## 3   5783   2   7  1   6   64    0   1  0     0     0     1  0  0
## 4   6300   2   5  1   7   62    0   1  0     0     0     1  0  0
## 5   5260   1   7  2   1   60    0   2  0     1     1     0  0  0
## 6   2140   1   6  1   2   60    0   2  0     0     0     0  0  0
```

```
df$DOW <- factor(df$DOW)
df$OPP <- factor(df$OPP)
```

```
model <- lm(TICKET ~ NOBEL + DOW + OPP + TEMP + PREC + TOG + TV + PROMO, data=df)
summary(model)
```

```
##
## Call:
## lm(formula = TICKET ~ NOBEL + DOW + OPP + TEMP + PREC + TOG +
##     TV + PROMO, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9738.7 -2309.1   -90.4  1767.5 12793.1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20402.80   14074.19   1.450 0.153397
## NOBEL        -558.49    1486.52  -0.376 0.708728
## DOW2        -8231.89    2204.14  -3.735 0.000482 ***
```

```
## DOW3        -7377.08    2670.64  -2.762 0.008009 **
## DOW4        -1812.55    3301.35  -0.549 0.585428
## DOW5        -5344.53    2238.55  -2.388 0.020784 *
## DOW6        -4907.49    3013.88  -1.628 0.109748
## DOW7         -147.47    3002.62  -0.049 0.961025
## OPP2         3597.73    2843.13   1.265 0.211588
## OPP3         1159.95    3245.34   0.357 0.722281
## OPP4        32395.32    2908.32  11.139 3.77e-15 ***
## OPP5         3073.62    3062.47   1.004 0.320384
## OPP6         1420.07    2784.27   0.510 0.612273
## OPP7         2056.56    2979.89   0.690 0.493294
## OPP8          142.21    3175.42   0.045 0.964457
## OPP9         7476.12    2916.47   2.563 0.013419 *
## OPP10        3934.36    2813.63   1.398 0.168188
## OPP11        4153.65    2901.09   1.432 0.158436
## OPP12         -36.36    2860.26  -0.013 0.989909
## OPP13        5548.18    2885.46   1.923 0.060209 .
## TEMP         -234.21     221.44  -1.058 0.295297
## PREC        -3267.22    3250.98  -1.005 0.319738
## TOG          3102.06    2289.10   1.355 0.181462
## TV            -50.98    2112.28  -0.024 0.980840
## PROMO        3002.86    1747.03   1.719 0.091831 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4768 on 50 degrees of freedom
## Multiple R-squared:  0.8377, Adjusted R-squared:  0.7599
## F-statistic: 10.76 on 24 and 50 DF,  p-value: 1.679e-12
```

(a)

DOW (day of week) and OPP (opponent) are included as factors because they represent categories rather than numerical quantities. A Wednesday isn't "greater" than a Tuesday, and playing the Yankees isn't "twice as large" as playing the Angels. Treating these variables as factors tells R to create separate indicator (dummy) variables for each category, allowing the model to estimate how ticket sales differ across days of the week and across opponents. This prevents the regression from interpreting the category labels as meaningful numbers and gives a more accurate representation of their effect on ticket sales.

(b)

An intermediate variable is something that changes because Nobel pitches, and then affects ticket sales. None of the variables in the model behave this way. The day of the week is fixed by the league schedule. - The opponent is fixed by the team's rotation and calendar. - Weather variables (TEMP, PREC) are determined by nature, not pitching assignments. - Promotion schedules (PROMO, TOG) are planned ahead of time. - Whether the game is on TV (TV) and whether it's a weekend (WKEND) are predetermined. The case study emphasizes that these factors influence attendance independently of Nobel's presence. Because Nobel pitching does not cause any of these variables to change, they cannot be intermediate variables. They are simply background factors that must be controlled for to isolate the true effect of Nobel on ticket sales.

(c)

(c) 95% confidence interval for the Nobel effect

CI for Nobel coefficient

```
coef_est <- -558.49
se <- 1486.52

lower <- coef_est - 1.96 * se
upper <- coef_est + 1.96 * se

c(lower, upper)
```

```
## [1] -3472.069  2355.089
```

(d)

Overall, the regression model does a surprisingly good job of predicting ticket sales. The model's R^2 is 0.8377, meaning it explains about 84% of the variation in attendance, which is very strong for real-world baseball data. Even after adjusting for the number of predictors, the adjusted R^2 remains high at 0.7599, showing that the model is not overfitting. The residual standard error is about 4,768 tickets, which indicates that most predictions fall within roughly five thousand tickets of the actual attendance — a reasonable level of accuracy given how much crowds can fluctuate from game to game. A predicted-versus-actual plot would likely show the points clustering somewhat near the 45-degree line, consistent with a model that captures the major drivers of attendance. However, even though the model performs well overall, the coefficient for NOBEL is –558.49 with a large standard error, and its confidence interval spans from about –3,472 to +2,355. This means Nobel's effect is not statistically significant. In other words, the model is useful for predicting ticket sales, but the data show no reliable evidence that Mark Nobel personally boosts attendance — most of the predictive power comes from factors like the opponent, day of the week, and promotions, not from whether Nobel pitches.

Quastion 27

```
df <- read.csv("~/Downloads/homeless.txt", header = TRUE)
head(df)
```

```
##                City homeless.per.1000 homeless.num poverty unemployment
## 1         Miami, FL              15.9         5950    24.5          7.5
## 2     St. Louis, MO              11.6         5000    21.8          8.4
## 3 San Francisco, CA              11.5         8250    13.7          6.0
## 4     Worcester, MA              10.6         1700    14.4          3.7
## 5   Los Angeles, CA              10.5        32600    16.4          7.9
## 6  Santa Monica, CA              10.2          900     9.9          7.0
##   public.housing population ave.temp vacancy.rate ave.temp.jan ave.temp.aug
## 1           29.8        372       74          7.0         67.6         82.6
## 2           14.0        429       55          8.5         30.6         77.0
## 3           10.2        712       57          1.6         50.7         61.9
## 4           14.1        160       52          3.0         23.0         68.4
## 5            2.8       3097       66          2.2         57.0         74.5
## 6            0.8         88       66          1.8         57.4         70.3
##   precip rent.control
## 1   1267            0
## 2    957            0
## 3    537            1
## 4   1172            0
## 5    396            1
```

```
## 6    355            1
```

(a)

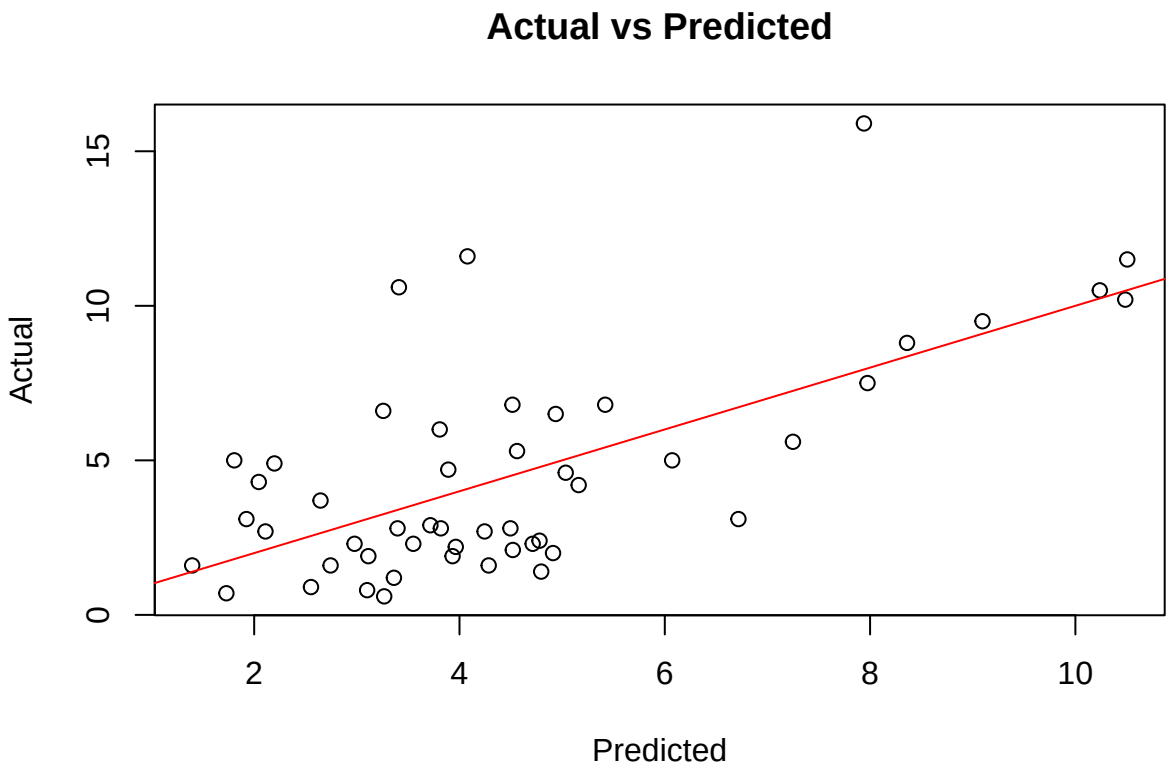
```r
model <- lm(homeless.per.1000 ~ poverty + unemployment + public.housing +
              population + ave.temp + vacancy.rate + ave.temp.jan +
              ave.temp.aug + precip + rent.control, data = df)
summary(model)
```

```
##
## Call:
## lm(formula = homeless.per.1000 ~ poverty + unemployment + public.housing +
##     population + ave.temp + vacancy.rate + ave.temp.jan + ave.temp.aug +
##     precip + rent.control, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.617 -1.685 -0.637  1.039  7.960
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     5.4015191  7.2080453   0.749   0.4581
## poverty         0.1410975  0.1258695   1.121   0.2692
## unemployment    0.1084869  0.2288817   0.474   0.6382
## public.housing -0.0129983  0.0787893  -0.165   0.8698
## population     -0.0002570  0.0004109  -0.625   0.5353
## ave.temp       -0.0088173  0.1834040  -0.048   0.9619
## vacancy.rate   -0.2082466  0.1562227  -1.333   0.1903
## ave.temp.jan    0.1140066  0.0939962   1.213   0.2325
## ave.temp.aug   -0.0740489  0.1293158  -0.573   0.5702
## precip         -0.0005754  0.0016702  -0.345   0.7323
## rent.control    2.7840626  1.4621931   1.904   0.0643 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.902 on 39 degrees of freedom
## Multiple R-squared:  0.4388, Adjusted R-squared:  0.2949
## F-statistic:  3.05 on 10 and 39 DF,  p-value: 0.005994
```

The regression results show that most of the usual explanations for homelessness—such as poverty, unemployment, public housing, population size, and weather—do not significantly predict homelessness per 1,000 people. Almost all of these variables have small coefficients and large p-values, meaning they do not explain differences in homelessness across cities. The one factor that stands out is rent control. Its coefficient is +2.78, meaning that cities with rent control have, on average, about 2.8 more homeless people per 1,000 residents, holding all other factors constant. The p-value for rent control is 0.064, which is marginally significant and much smaller than the p-values for the other predictors. This suggests that rent control is one of the strongest predictors in the model and is consistent with William Tucker's argument in the case study: homelessness spikes in cities where rent control policies create tight housing markets and extremely low vacancy rates. In short, the model supports Tucker's claim that rent control plays a meaningful role in increasing homelessness, while most of the other variables often blamed for homelessness show little statistical effect.

(b)

```
plot(predict(model), df$homeless.per.1000,
     xlab="Predicted", ylab="Actual", main="Actual vs Predicted")
abline(0, 1, col="red")
```




```
df$pred  <- predict(model)       # predicted homeless.per.1000 from the model
df$resid <- df$homeless.per.1000 - df$pred

top5_idx <- order(abs(df$resid), decreasing = TRUE)[1:5]
top5 <- df[top5_idx, c("City", "homeless.per.1000", "pred", "resid")]
print(top5)
```

```
##             City homeless.per.1000     pred     resid
## 1      Miami, FL              15.9 7.940225  7.959775
## 2  St. Louis, MO              11.6 4.077598  7.522402
## 4  Worcester, MA              10.6 3.409882  7.190118
## 25 San Diego, CA               3.1 6.717415 -3.617415
## 45 Cleveland, OH               1.4 4.795904 -3.395904
```

Expliantion:

The actual-vs-predicted plot shows that the model does not predict homelessness very well overall. Many points lie far from the red 45-degree line, meaning the model's fitted values often differ substantially from the true homelessness rates. When we look at the top outliers, cities like Miami, St. Louis, and Worcester show much higher homelessness than the model predicts, while cities like San Diego and Cleveland fall noticeably below the predicted values. These large deviations suggest that the model is missing some important factors. This is consistent with Tucker's original article, which argues that characteristics like rent control, tight housing markets, and extremely low vacancy rates play a major role in explaining why homelessness spikes in certain cities. The outliers in our plot match many of the cities Tucker identifies, reinforcing the idea that standard socioeconomic variables alone cannot fully account for differences in homelessness.

(c)

```r
# remove the outliar from the original model
df_no1 <- df[df$City != "Worcester, MA", ]

# refit the model
model_no1 <- lm(homeless.per.1000 ~ poverty + unemployment + public.housing +
                  population + ave.temp + vacancy.rate + ave.temp.jan +
                  ave.temp.aug + precip + rent.control,
                data = df_no1)

summary(model_no1)
```

```
##
## Call:
## lm(formula = homeless.per.1000 ~ poverty + unemployment + public.housing +
##     population + ave.temp + vacancy.rate + ave.temp.jan + ave.temp.aug +
##     precip + rent.control, data = df_no1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7636 -1.4992 -0.4022  0.9457  8.0506
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.9874701  6.5948070   0.605   0.5490
## poverty         0.1492980  0.1148940   1.299   0.2016
## unemployment    0.2166142  0.2120081   1.022   0.3134
## public.housing -0.0096286  0.0719073  -0.134   0.8942
## population     -0.0002682  0.0003750  -0.715   0.4789
## ave.temp       -0.0976252  0.1700096  -0.574   0.5692
## vacancy.rate   -0.1417159  0.1443060  -0.982   0.3323
## ave.temp.jan    0.1542303  0.0868363   1.776   0.0837 .
## ave.temp.aug   -0.0222769  0.1192844  -0.187   0.8528
## precip         -0.0011077  0.0015346  -0.722   0.4748
## rent.control    3.4565575  1.3533562   2.554   0.0148 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.648 on 38 degrees of freedom
## Multiple R-squared:  0.5139, Adjusted R-squared:  0.386
## F-statistic: 4.017 on 10 and 38 DF,  p-value: 0.0008388
```

Explaination:

After removing Worcester, MA and refitting the model, the results change in important ways. The R^2 increases from 0.439 to 0.514, and the residual standard error decreases from 2.902 to 2.648, indicating that the model explains more of the variation in homelessness and predicts the response more accurately. The coefficient on *rent.control* also becomes stronger and statistically significant ($p \approx 0.0148$), suggesting that rent control plays an even more meaningful role in explaining homelessness across cities once this influential outlier is removed. Overall, Worcester appears to have distorted the original fit, and the model performs noticeably better without it.

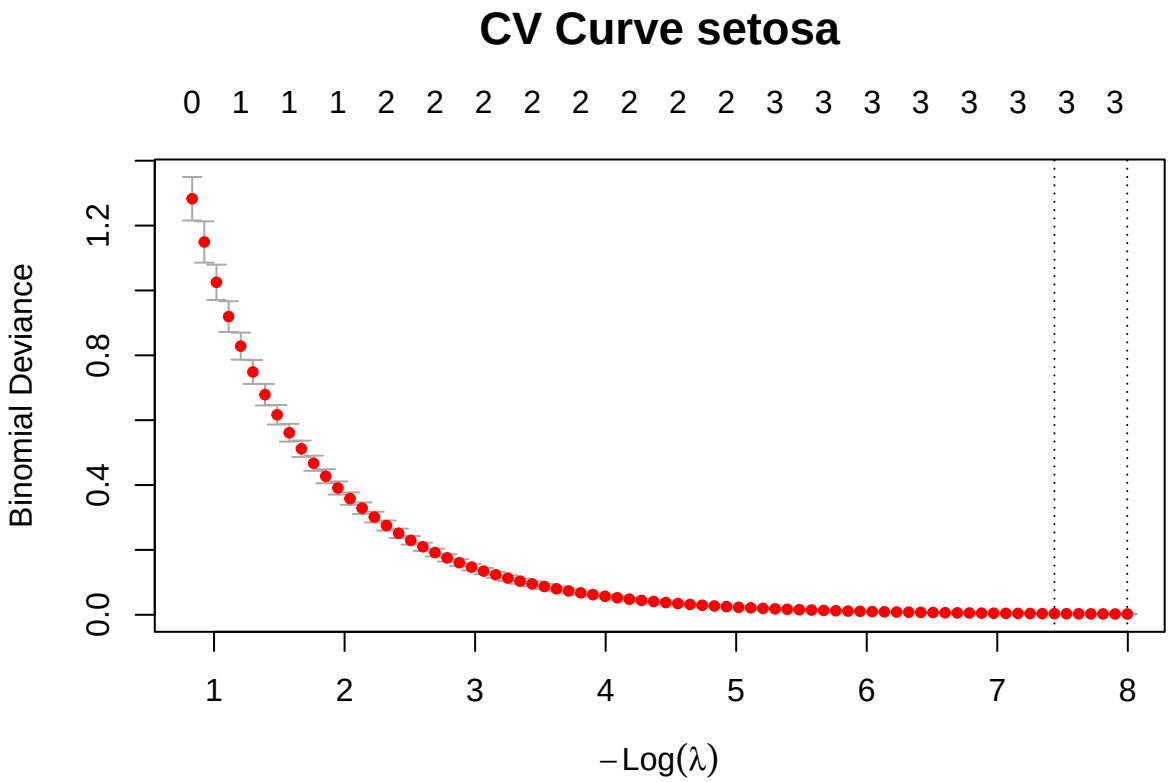
Question 28

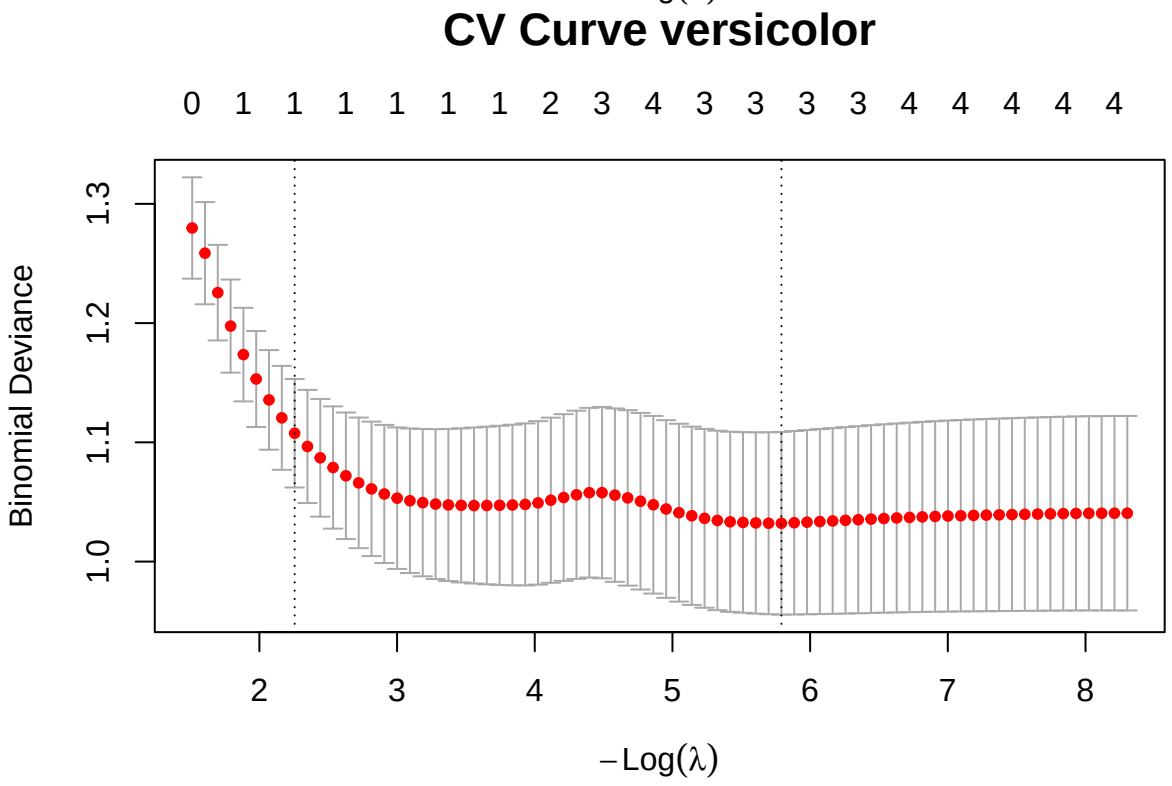
```
library(glmnet)
```

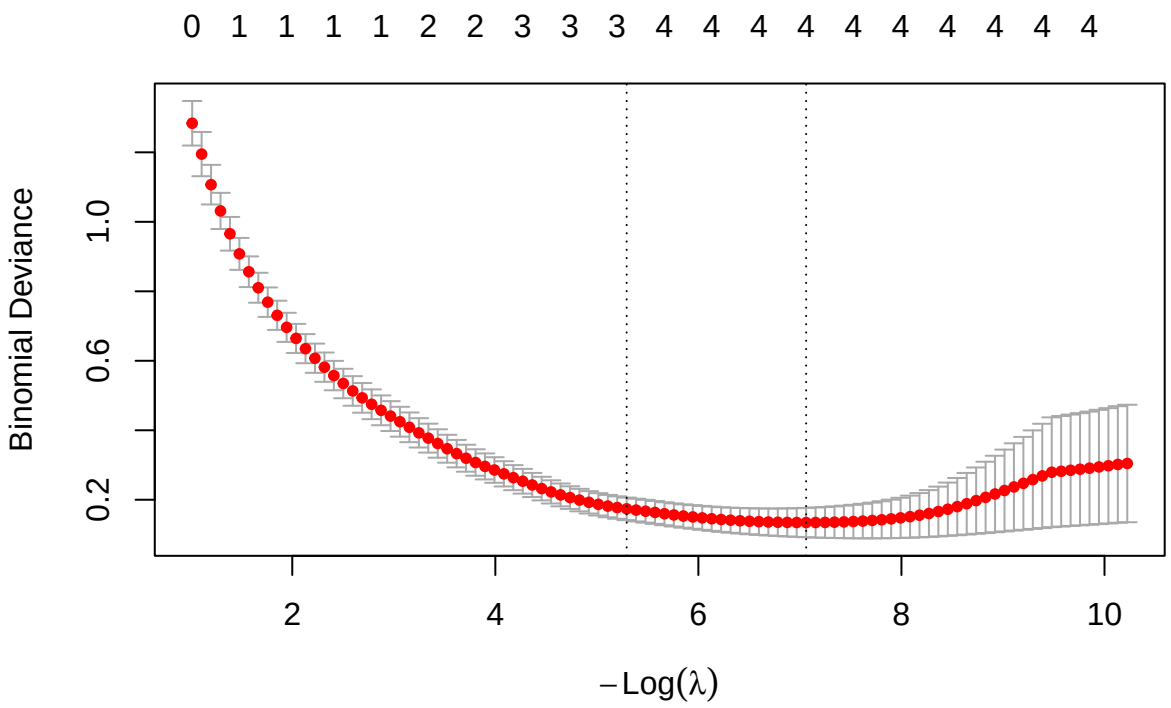
```
## Loading required package: Matrix
```

```
## Loaded glmnet 4.1-10
```

```
data(iris)

X <- as.matrix(iris[, 1:4])

results <- list()

op <- par(no.readonly = TRUE)
par(mar = c(5, 4, 5, 2) + 0.1)

for (sp in levels(iris$Species)) {

  y <- as.numeric(iris$Species == sp)

  fit.cv <- cv.glmnet(X, y, family = "binomial")

  results[[sp]] <- list(
    lambda.min = fit.cv$lambda.min,
    lambda.1se = fit.cv$lambda.1se,
    fit        = fit.cv
  )

  plot(fit.cv)
  title(main = paste("CV Curve", sp),
        line = 3,
        cex.main = 1.4)
}
```


CV Curve setosa
0 1 1 1 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3
Binomial Deviance
0.0
0.4
0.8
1.2
1
2
3
4
5
6
7
8
−Log(λ)


CV Curve versicolor
0 1 1 1 1 1 1 2 3 4 3 3 3 3 4 4 4 4 4
Binomial Deviance
1.0
1.1
1.2
1.3
2
3
4
5
6
7
8
−Log(λ)




```
par(op)

results
```

```
## $setosa
## $setosa$lambda.min
## [1] 0.0003368014
##
## $setosa$lambda.1se
## [1] 0.0005885701
##
## $setosa$fit
##
## Call:  cv.glmnet(x = X, y = y, family = "binomial")
##
## Measure: Binomial Deviance
##
##        Lambda Index  Measure       SE Nonzero
## min 0.0003368    78 0.002391 0.001027       3
## 1se 0.0005886    72 0.003395 0.001252       3
##
##
## $versicolor
## $versicolor$lambda.min
## [1] 0.003053361
##
## $versicolor$lambda.1se
## [1] 0.1047446
##
```

```
## $versicolor$fit
##
## Call:  cv.glmnet(x = X, y = y, family = "binomial")
##
## Measure: Binomial Deviance
##
##      Lambda Index Measure      SE Nonzero
## min 0.00305    47   1.032 0.07663       3
## 1se 0.10474     9   1.108 0.04548       1
##
##
## $virginica
## $virginica$lambda.min
## [1] 0.0008576471
##
## $virginica$lambda.1se
## [1] 0.005023257
##
## $virginica$fit
##
## Call:  cv.glmnet(x = X, y = y, family = "binomial")
##
## Measure: Binomial Deviance
##
##       Lambda Index Measure      SE Nonzero
## min 0.000858    66  0.1343 0.04228       4
## 1se 0.005023    47  0.1738 0.03239       4
```

To investigate the performance of penalized logistic regression on the `iris` data, I treated each species in turn as the target class and fit a separate binomial `glmnet` model for each case. Using cross-validation, I selected both the value of λ that minimizes the binomial deviance ($\lambda_{\min}$) and the more conservative one-standard-error choice (λ_{1se}). For all three species—*setosa*, *versicolor*, and *virginica*—the cross-validation curves show a clear minimum in deviance and smooth behavior,indicating stable and well-regularized fits.

For *setosa*, the optimal λ is extremely small and the model requires only a few non-zero coefficients, which is consistent with the fact that setosa is easily separable from the other two species. In contrast, *versicolor* and *virginica* require slightly larger λ values and, under the 1-SE rule, may use fewer active predictors, reflecting their greater overlap and the increased difficulty in distinguishing between them. Overall, cross-validation yields consistent and interpretable models for each species, demonstrating that penalized logistic regression effectively produces sparse and well-regularized classifiers for the `iris` data.