STAT 516 sp 2026 exam 02

75 minutes, two pages of handwritten notes allowed, no calculators

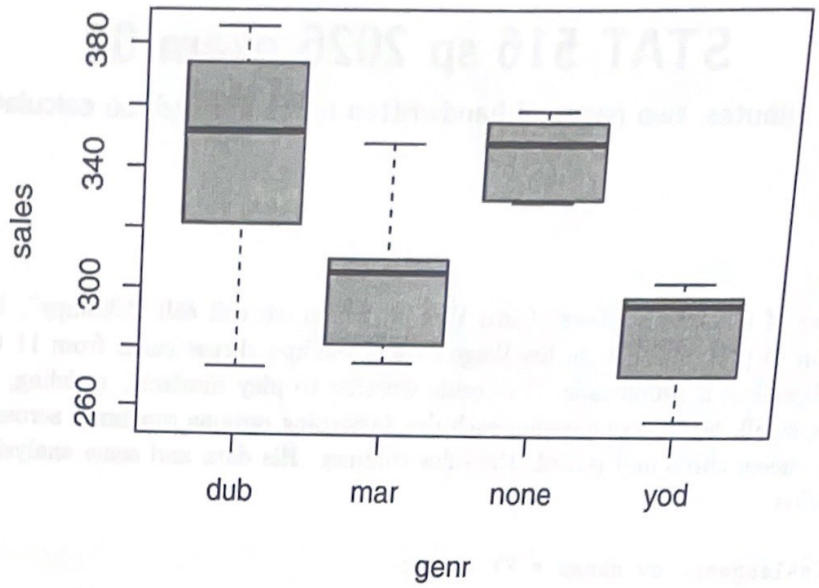
1.

The owner of Chompy's Cheese Curd Wagon, whom we will call "Chompy", has obtained permission to play music from his Wagon while vending cheese curds from 11 to 2 daily at the north end of a promenade. To decide whether to play mariachi, yodeling, dub-step, or no music at all, he tries one option each day (assigning options randomly across days) while vending cheese curds and records the sales volumes. His data and some analysis output are given below:

```
print(salesgenr,row.names = F)
```

```
 day genr    sales
   1 none 347.8170
   2  dub 273.5590
   3  dub 373.7479
   4 none 358.3151
   5 none 329.5143
   6  mar 305.5093
   7  yod 255.3859
   8  yod 297.2734
   9  mar 309.8852
  10  mar 281.2064
  11 none 354.6366
  12  mar 347.8584
  13  yod 270.8383
  14  dub 321.3628
  15  mar 274.9311
  16  yod 302.1497
  17  dub 351.6953
  18  dub 385.3534
  19  yod 294.6229
  20 none 328.6520
```

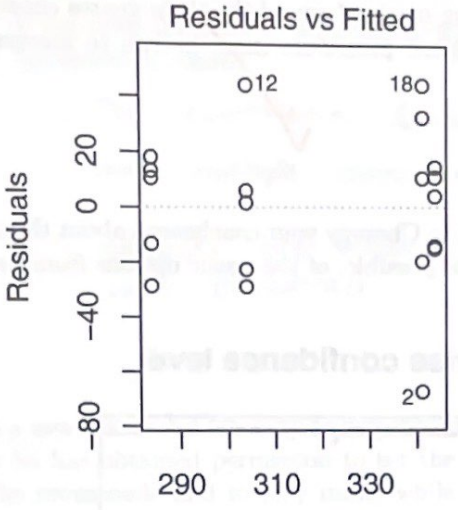
```
boxplot(sales ~ genr, data = salesgenr)
```

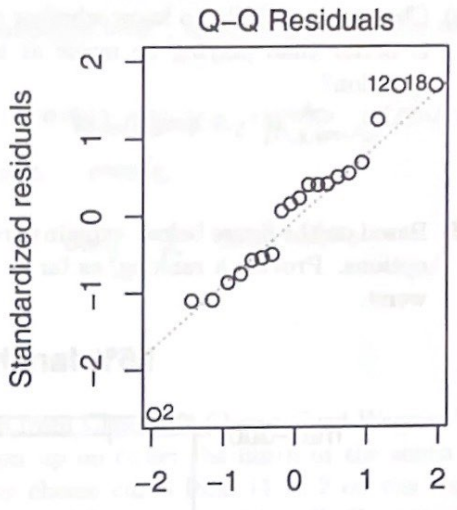



```
lm_out <- lm(sales ~ genr, data = salesgenr)

anova(lm_out)

Analysis of Variance Table

Response: sales
          Df Sum Sq Mean Sq F value  Pr(>F)
genr       ?  12761  4253.6  4.9314 0.01299 *
Residuals ??  13801   862.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

par(mfrow=c(1,2))
plot(lm_out,which = 1,add.smooth=F)
plot(lm_out,which = 2)
```

(a) $\begin{bmatrix} 3 \\ 16 \end{bmatrix}$

25/25





(a) Replace the question marks in the ANOVA table with the correct numbers.

(b) Write the null and alternate hypotheses for which the F value in the ANOVA table is the test statistic: State your conclusion with respect to these hypotheses.

H_0: there is no difference between the means for each treatment group

H_1: At least one of the treatment group means is different

conclusion: reject H_0

(c) Comment on whether you believe the assumptions of the one-way ANOVA model are satisfied.

✓ 1. Experimental design: results independent of each other

✗ 2. Residuals vs Fitted Plot: variance for each treatment group do NOT appear equal

✓ 3. Q-Q Plot: Residuals appear normally distributed

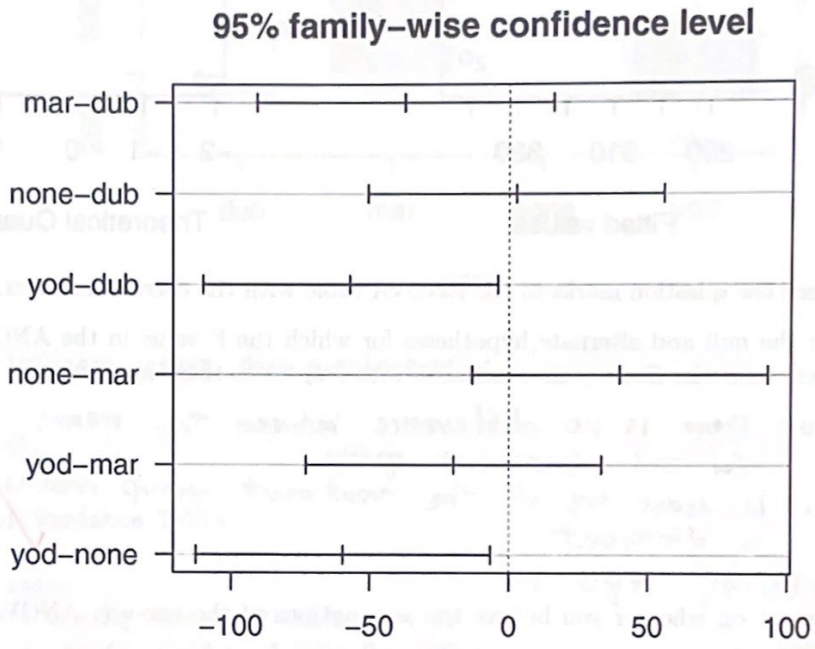
(d) Chompy would like to rank the music genres as well as the option of no music in order to maximize revenue. How many comparisons between means must he make to obtain a ranking, and what procedure must he run to make these comparisons?

- $\binom{4}{2} = 6$ comparisons
- Use Tukey's HSD

(e) Chompy would like to know whether playing music of any of the three genres considered is better than playing no music at all. What procedure must he run to answer this question?

Dunnett's method

(f) Based on the figure below, explain carefully to Chompy your conclusions about the music options. Provide a ranking, as far as this is possible, of the music options from best to worst.



- According to the figure, there appears to be no significant difference between any of the types of music, except that yodelling is worse than both dubstep & no music.
- Only definitive conclusion is that yodelling is either the 1st or second worst option

(g) What does the phrase "95% family-wise confidence level", appearing in the title of the figure above, mean?

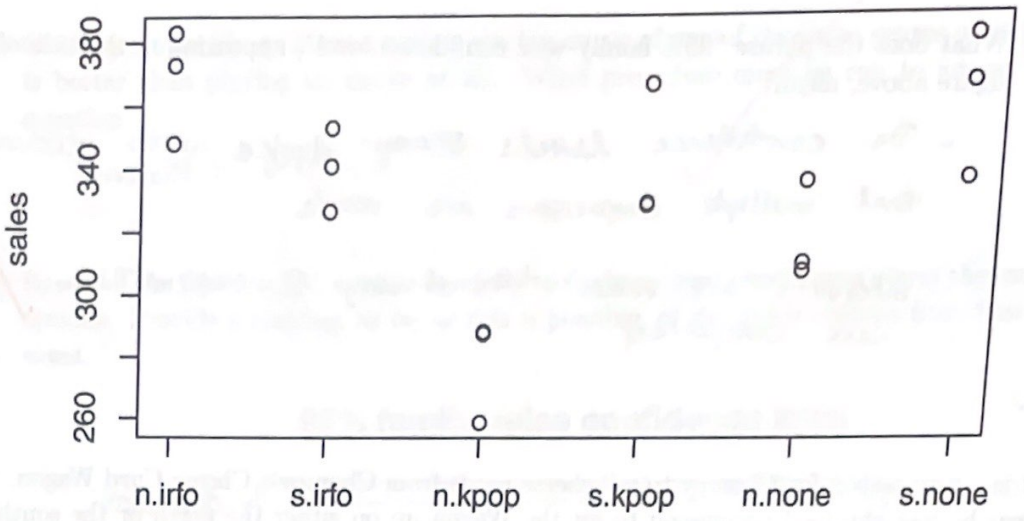
- The confidence levels shown take into account that multiple comparisons are made
- intervals are wider than if only 2 covariates were compared ✓

2.

It is a new season for Chompy to sell cheese curds from Chompy's Cheese Curd Wagon: This year he has obtained permission to set the Wagon up on either the north or the south end of the promenade and to play music while selling cheese curds from 11 to 2 on any day of the week. This year he considers playing K-pop, irish folk, or no music at all. He sets up an experiment to find the best combination of location and music for maximizing sales. His data as well as some analysis output are below:

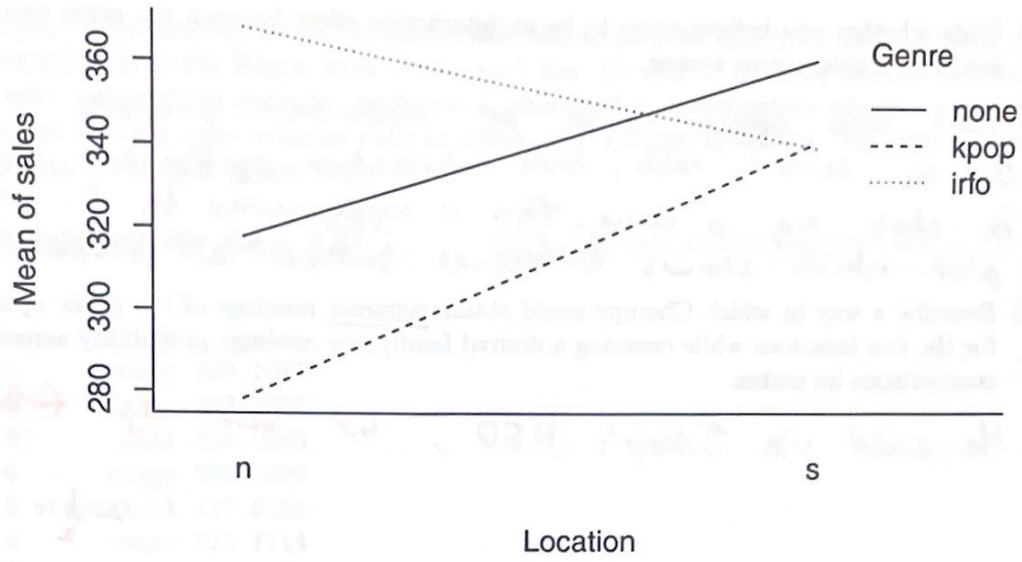
day	genr	loc	sales
1	kpop	n	288.1713
2	none	n	335.9089
3	none	n	310.0643
4	kpop	s	329.3373
5	irfo	s	340.9375
6	kpop	n	287.5502
7	irfo	n	383.0008
8	none	s	336.6392
9	irfo	n	348.2206
10	irfo	s	353.0560
11	irfo	s	326.6296
12	kpop	s	328.7751
13	kpop	n	258.6882
14	irfo	n	372.8953
15	none	s	366.9432
16	none	s	381.4332
17	none	n	307.3991
18	kpop	s	366.6999

```
stripchart(sales ~ loc + genr,data = salesgenrloc, vertical=T, pch = 1)
```




```
lm_out <- lm(sales ~ loc + genr + loc:genr, data = salesgenrloc)

summary(lm_out)
anova(lm_out)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     368.04      10.60  34.727 2.07e-13 ***
locs            -27.83      14.99  -1.857  0.08803 .
genrkpop        -89.90      14.99  -5.998 6.23e-05 ***
genrnone        -50.25      14.99  -3.353  0.00575 **
locs:genrkpop    91.30      21.20   4.307  0.00102 **
locs:genrnone    71.71      21.20   3.383  0.00544 **

Analysis of Variance Table

Response: sales
(b) Df          Df Sum Sq Mean Sq F value   Pr(>F)
1   loc          ? 3161.5  3161.5  9.3826 0.009842 **
2   genr         ? 6114.3  3057.1  9.0728 0.003979 **
2   loc:genr     ? 6930.9  3465.4 10.2845 0.002502 **
12  Residuals ?? 4043.5   337.0
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2(3)(3-1) = 6(7)




```
aggregate(sales ~ loc, mean, data = salesgenrloc)

  loc    sales
1   n 321.3221
2   s 347.8279

aggregate(sales ~ genr, mean, data = salesgenrloc)

  genr    sales
1 irfo 354.1233
2 kpop 309.8703
3 none 339.7313
```

(a) Use the output to compute the mean sales volume on days Chompy sold cheese curds at the south end of the promenade while playing Irish folk music. Do any arithmetic necessary to obtain your answer.

368.04 − 27.83 = 340.21

(b) Give the value in the position of each question mark in the ANOVA table.

		Df
loc	(2−1) =	1
genr	(3−1) =	2
loc:genr	(2−1)(3−1) =	2
err	2(3)(3−1) =	12

(c) State whether you believe there to be an interaction effect between the genre and the location. Explain your answer.

I believe that there is an interaction effect between genre and location because the null assumption is there is no interaction but because the p-val for the interaction is smaller than the alpha level of 0.05, we reject the null. Additionally, the interaction plot shows differing slopes and lines crossing which further backs up our conclusion that there is an interaction when combined with the low p-value.

(d) Describe a way in which Chompy could obtain separate rankings of the music options for the two locations while ensuring a desired familywise coverage probability across all comparisons he makes.

Chompy can use Bonferroni correction and by dividing alpha by 2 and proceeding as if he is doing 2 separate procedures in comparisons, he can ensure a desired family wise coverage probability.

(e) Give the marginal means $\bar{Y}_{i\cdot\cdot}$ for $i = 1, \ldots, a$ for the location factor.

$\bar{Y}_{1\cdot\cdot} = 321.3221$

$\bar{Y}_{2\cdot\cdot} = 347.8279$

321.3221
+ 347.8279
669.1500

334.5750
2 / 669.1500

(f) Based on the marginal means $\bar{Y}_{\cdot j\cdot}$ for genres $j = 1, \ldots, b$, Chompy's buddy tells him he should always play Irish folk music in order to maximize revenue. Respond carefully to this advice.

$\bar{Y}_{\cdot 1\cdot} = 354.1233$

$\bar{Y}_{\cdot 2\cdot} = 309.8703$

$\bar{Y}_{\cdot 3\cdot} = 339.7313$

Although irish folk music had the highest marginal mean we must proceed with caution because there was an interaction present we must look at the simple effects as the effect of A changes based on B.

3.

Another year has passed. This season Chompy takes a vacation, leaving Chompy's Cheese Curd Wagon to be run by several part-time operators, whom he instructs to park the Wagon

everyday in the same place on the promenade and to play no music; a single operator vends cheese curds from the Wagon from 11 to 2 each day. By and by Chompy returns and reviews the sales volumes. In doing so, he becomes interested in investigating whether some of the variation in daily sales volumes owes to differences between operators. His data as well as some analysis output appear below:

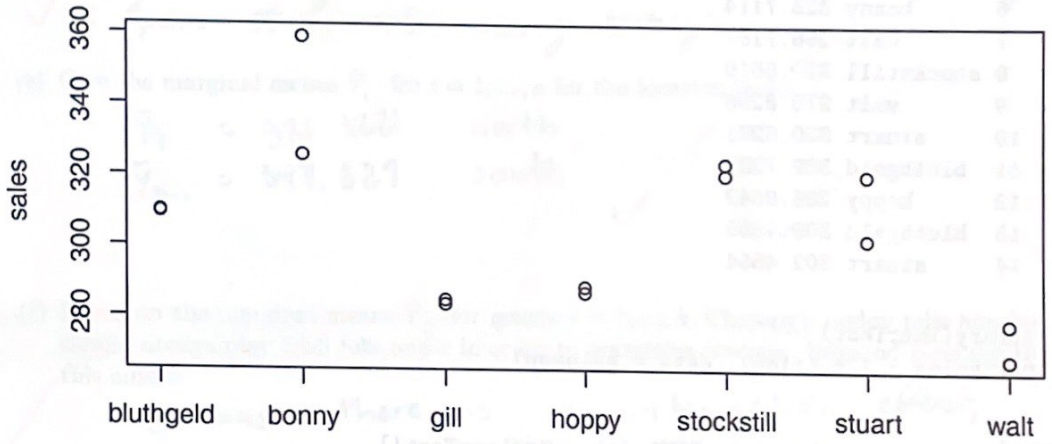
```
print(salesop,row.names = F)
```

```
 day         op    sales
   1      bonny 359.1077
   2       gill 283.5982
   3       gill 284.8243
   4      hoppy 288.5896
   5 stockstill 323.8126
   6      bonny 325.7114
   7       walt 268.7187
   8 stockstill 320.5619
   9       walt 278.8296
  10     stuart 320.6271
  11  bluthgeld 309.7233
  12      hoppy 286.9642
  13  bluthgeld 309.4359
  14     stuart 302.4554
```

```
library(lmerTest)
lmer(sales ~ 1 + (1|op), data = salesop)
```

```
Linear mixed model fit by REML ['lmerModLmerTest']
Formula: sales ~ 1 + (1 | op)
   Data: salesop
REML criterion at convergence: 114.8403
Random effects:
 Groups   Name        Std.Dev.
 op       (Intercept) 22.83
 Residual             10.56
Number of obs: 14, groups:  op, 7
Fixed Effects:
(Intercept)
      304.5
```

```
anova(lm(sales ~ op, data = salesop))

Analysis of Variance Table

Response: sales
          Df Sum Sq Mean Sq F value   Pr(>F)
op         6 6921.5 1153.58  10.336 0.003491 **
Residuals  7  781.3  111.61
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

stripchart(sales ~ op, data = salesop, pch = 1, vertical = T)
```



(a) Suppose Chompy regards each operator as having been drawn randomly from a population of qualified applicants, in which the distribution of Aptitude for Vending Cheese Curds is distributed according to a bell-shaped curve. Write down a mathematical model for the sales volume Y_{ij} on day j under operator i for $i = 1, \ldots, a$ and $j = 1, \ldots, n_i$, where a is the number of operators, and n_i is the number of days on which operator i operated the Wagon. Explain each term in your model.

$Y_{ij} = \mu + A_i + \varepsilon_{ij}$

sales volume on day j when operator i is the operator

μ: the fixed overall mean of sales

A_i: random effect of operator i on sales

ε_{ij}: noise in the data causing one day with operator i to differ from another day with the same operator i

(b) Find the number 22.83 in the above output. Interpret this number. Of what is this number an estimate?

this is the standard deviation of treatment groups. σ_A
it is restricted maximum likelihood. to make sure σ_A always greater than zero.

(c) Addressing yourself to Chompy, interpret carefully the output in the ANOVA table, speaking to his questions about whether operator-to-operator differences contribute significantly to variation in daily sales volumes from Chompy's Cheese Curd Wagon.

Since the p-value is 0.003491, quite small, so we can ~~assume~~ ~~that~~ conclude that operator-to-operator differences contribute significantly to variation in daily sales volumes.

4.

Below is a subsample from a data set containing the sale price and several attributes of homes sold in Miami (The data were found at https://www.openml.org/search?type=data&status=active&id=43093&sort=runs). Some analysis output follows:

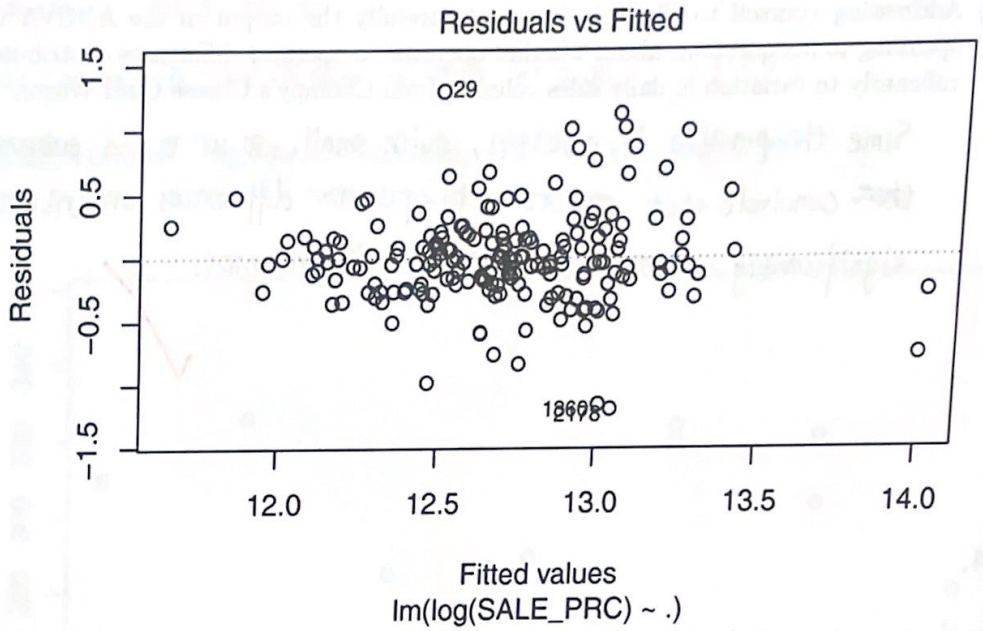
```
head(miami)
```

	SALE_PRC	LND_SQFOOT	RAIL_DIST	OCEAN_DIST	WATER_DIST	CNTR_DIST	SUBCNTR_DI
10675	160000	8515	2504.8	46927.4	25647.9	149322.1	99754.3
2858	240000	6360	6353.5	17180.9	1311.6	17028.3	12032.1
13702	237800	3927	524.5	39230.3	20568.9	100575.6	49513.3
5400	315000	10340	3728.3	14707.0	10026.1	76385.7	26901.7
9888	200000	7500	24530.0	24630.2	17535.8	123424.5	75166.1
526	193000	5000	6241.1	22058.7	1824.6	43289.0	42772.6

	HWY_DIST	age	avno60plus
10675	3585.2	19	0
2858	1467.6	70	0

```
13702  16548.9  18          0
5400    9311.5  44          0
9888     698.8  25          0
526     6697.8  70          0
```

```
lm_all <- lm(log(SALE_PRC)~., data = miami)
plot(lm_all,which = 1,add.smooth=F)
```




```
library(car)
vif(lm_all)
```

```
LND_SQFOOT  RAIL_DIST OCEAN_DIST WATER_DIST  CNTR_DIST SUBCNTR_DI   HWY_DIST
  1.141211   1.360817   2.994495   4.018879   6.957380   5.332055   1.347611
       age avno60plus
  1.514359   1.042170
```

```
library(leaps)
regsubsets_out <- regsubsets(log(SALE_PRC) ~ ., data = miami)
summary(regsubsets_out)
```

```
Subset selection object
```

```
Call: regsubsets.formula(log(SALE_PRC) ~ ., data = miami)
9 Variables  (and intercept)
           Forced in Forced out
LND_SQFOOT     FALSE      FALSE
RAIL_DIST      FALSE      FALSE
OCEAN_DIST     FALSE      FALSE
WATER_DIST     FALSE      FALSE
CNTR_DIST      FALSE      FALSE
SUBCNTR_DI     FALSE      FALSE
HWY_DIST       FALSE      FALSE
age            FALSE      FALSE
avno60plus     FALSE      FALSE
1 subsets of each size up to 8
Selection Algorithm: exhaustive
         LND_SQFOOT RAIL_DIST OCEAN_DIST WATER_DIST CNTR_DIST SUBCNTR_DI
1  ( 1 ) "*"        " "       " "         " "        " "       " "
2  ( 1 ) " "        " "       " "         " "        " "       "*"
3  ( 1 ) "*"        " "       " "         " "        " "       "*"
4  ( 1 ) "*"        " "       "*"         " "        " "       "*"
5  ( 1 ) "*"        " "       "*"         " "        " "       "*"
6  ( 1 ) "*"        " "       "*"         "*"        " "       "*"
7  ( 1 ) "*"        " "       "*"         "*"        " "       "*"
8  ( 1 ) "*"        "*"       "*"         "*"        " "       "*"
         HWY_DIST age avno60plus
1  ( 1 ) " "      " " " "
2  ( 1 ) " "      "*" " "
3  ( 1 ) " "      "*" " "
4  ( 1 ) " "      "*" " "
5  ( 1 ) "*"      "*" " "
6  ( 1 ) "*"      "*" " "
7  ( 1 ) "*"      "*" "*"
8  ( 1 ) "*"      "*" "*"
```

```
model1 <- lm(log(SALE_PRC) ~ LND_SQFOOT + RAIL_DIST + CNTR_DIST + SUBCNTR_DI, data = miami)
model2 <- lm(log(SALE_PRC) ~ LND_SQFOOT + RAIL_DIST, data = miami)
model3 <- lm(log(SALE_PRC) ~ LND_SQFOOT + RAIL_DIST + OCEAN_DIST, data = miami)

anova(model1)
```

```
Analysis of Variance Table

Response: log(SALE_PRC)
```

```
              Df Sum Sq Mean Sq F value    Pr(>F)
LND_SQFOOT     1  7.835  7.8351 39.4070 2.185e-09 ***
RAIL_DIST      1  0.358  0.3585  1.8030    0.1809
CNTR_DIST      1  0.241  0.2411  1.2124    0.2722
SUBCNTR_DI     1  3.521  3.5209 17.7088 3.925e-05 ***
Residuals    195 38.771  0.1988
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

p=4

n-(p+1)=195

n-(4+1)=195

n-5=195

n=200

```
anova(model2)

Analysis of Variance Table

Response: log(SALE_PRC)
              Df Sum Sq Mean Sq F value    Pr(>F)
LND_SQFOOT     1  7.835  7.8351 36.2899 8.208e-09 ***
RAIL_DIST      1  0.358  0.3585  1.6604    0.1991
Residuals    197 42.533  0.2159
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

anova(model3)

Analysis of Variance Table

Response: log(SALE_PRC)
              Df Sum Sq Mean Sq F value    Pr(>F)
LND_SQFOOT     1  7.835  7.8351 36.4780 7.621e-09 ***
RAIL_DIST      1  0.358  0.3585  1.6690    0.1979
OCEAN_DIST     1  0.434  0.4341  2.0208    0.1567
Residuals    196 42.099  0.2148
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(a) Which predictor variable appears to be most highly correlated with the other predictor variables? Explain your answer.

CNTR-DIST because the VIF is the largest, which is an indicator of correlation w/ other variables.

(b) What predictor variables are included in the best three-predictor model according to the `regsubsets()` function?

LND_SQFOOT, SUBCNTR_DL, age ✓

(c) Give an expression for the statistic for testing whether `model1` and `model2` differ in their ability to predict the sale price of a house. State the relevant null and alternate hypotheses and give the rejection rule.

Full vs Reduced F test, where m_1 is the full model and m_2 is the reduced,

$H_0: SSE_f = SSE_r$

$H_a: SSE_f \neq SSE_r$

reject if $F_{crit} < \frac{(SSE_r - SSE_f)/s}{SSE_f/(n-(p+1))}$ ← params full − reduced

$\frac{(42,533 - 38,771)/2}{38,771/195}$ ✓ correct

(d) How can we compare `model2` and `model3` in order to decide which one is "better"? Give at least one way and explain why this comparison is different from the comparison of `model1` and `model2`.

* R^2_{adj} better by 1 pred, not true for 1,2

If we just care about more than SSE, we can instead use AIC which accounts for model complexity. This factors in SSE and also # preds, which is why it is often used as a selection criterion in stepwise adding/removal.

If only care about SSE, same as above but m_2 is full, m_3 is red.

(e) How many observations are in this data set?

$n - (p+1) =$ denom

n_1 ~~n-5~~

$n - 5 = 195$

$n = 200$ ✓