STAT 516 sp 2026 final exam

150 minutes, three pages of handwritten notes allowed, no calculators

1.

Students in two statistics classes were asked to record their height, the lengths of their index and pinky fingers, and their shoe size along with the shoe size gender ("m" or "w"). Suppose we classify a student as "tall" if he or she has a height greater than 5'9". The data are read into R with the code below:

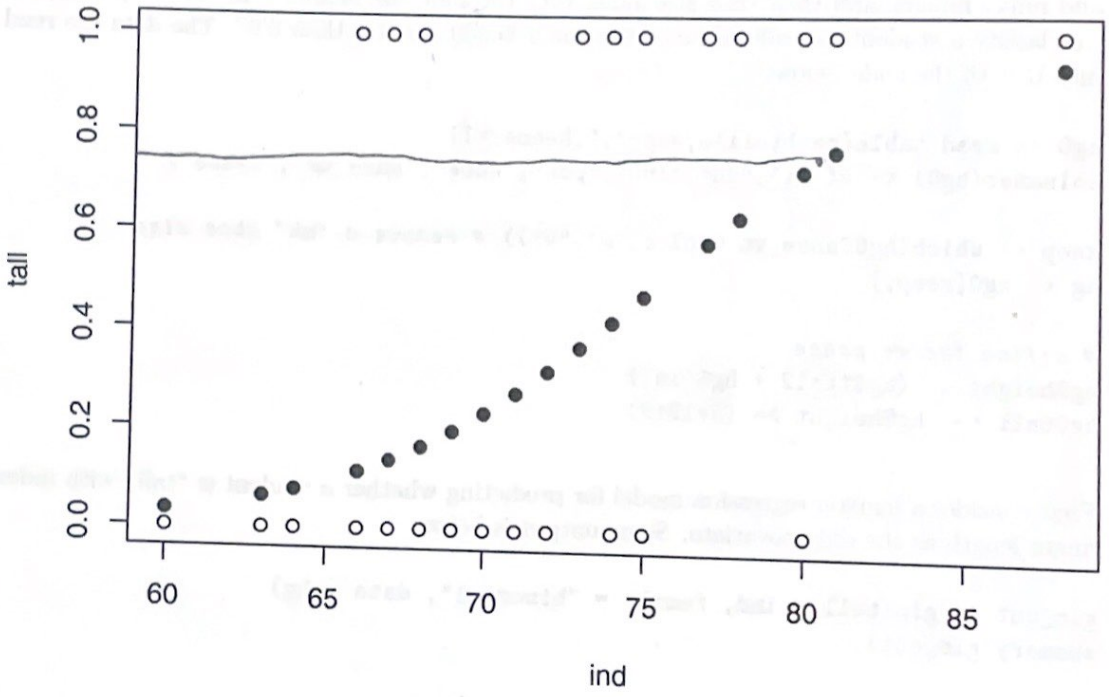
```
hg0 <- read.table(pathtofile,sep=",",header=T)
colnames(hg0) <- c("ft","in","ind","pnk","shoe","shoe_wm","class")

keep <- which(hg0$shoe_wm %in% c("m","w")) # remove a "uk" shoe size
hg <- hg0[keep,]

# define the response
hg$height <- (hg$ft*12 + hg$`in`)
hg$tall <-  hg$height >= (5*12+9)
```

First consider a logistic regression model for predicting whether a student is "tall" with index finger length as the only covariate. Some output is below.

```
glm_out <- glm(tall ~ ind, family = "binomial", data = hg)
summary(glm_out)
```

```
Call:
glm(formula = tall ~ ind, family = "binomial", data = hg)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -16.45603    5.62719  -2.924  0.00345 **
ind           0.21825    0.07707   2.832  0.00463 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 57.286  on 44  degrees of freedom
Residual deviance: 45.702  on 43  degrees of freedom
AIC: 49.702

Number of Fisher Scoring iterations: 4

plot(tall ~ ind, data = hg)
points(glm_out$fitted.values ~ hg$ind, pch = 19)
```



(a) According to the output above, describe carefully the estimated relationship between index finger length and the probability that a person is "tall". In particular, describe the change in the probability of being "tall" corresponding to an increase in index finger length of 1 mm.

For an increase in index finger length of 1mm, the odds of being tall are expected to increase by a factor of $e^{.21825}$.

(b) Give an expression for the estimated probability that a person with index finger length equal to 80 mm will be "tall". You do not need to evaluate your expression. Use the plot above to find an approximation to the estimated probability.

exp $\hat{\beta}_0 + \hat{\beta}_1 x$ = odds

$\frac{\text{odds}}{1+\text{odds}}$ $\frac{e^{-16.4560 + 0.21825(80)}}{1 + e^{-16.4560 + 0.21825(80)}}$

plot est ≈ 0.73 or 73%

(c) Suppose a person randomly sampled from among all people with a certain index finger length will be "tall" with probability 0.75. What are the corresponding *odds* of such a person's being "tall"? odds $= \frac{p}{1-p} \cdot \frac{0.75}{0.25} =$ 3:1 odds

(d) Explain why we do not look at a normal quantile-quantile plot of the differences $Y_i - \hat{\pi}_i$, $i = 1, \ldots, n$ in logistic regression.

The data is not continuous, it is a binary class, so it would look there were problems / inherent structure that we were missing due to patterns that are naturally a part of the data. This assumption cannot be checked in the same ways.

(e) Suppose the model is used to obtain, for each of ten *new* students, whose data were not used to fit the model, estimated probabilities of being "tall", resulting in the table below. The table includes the index finger lengths, the true response values (1 if truly "tall", 0 if not), and the estimated probabilities according to the fitted model of the ten new students.

	Y_{new}	$\hat{\pi}_{\text{new}}$	x_{new}
✗	0	0.83	68
✓	1	0.68	72
✓	1	0.89	66
✓	0	0.23	81
✓	1	0.63	73
✓	1	0.89	66 → ignore
✓	1	0.58	74
✓	0	0.47	76
✗ ✓	0	0.52	75
✓	1	0.80	69

Suppose we use the classifier $\hat{Y}_{new} = 1$ if $\hat{\pi}_{new} \geq 1/2$ and $\hat{Y}_{new} = 0$ if $\hat{\pi}_{new} < 1/2$ to classify these ten new students as "tall" or not. Treating these ten students as a testing data set, give:

i. The mis-classification rate. 2/10 → both FP ✓

ii. The true positive rate. All TPs found 100% ✓

iii. The false positive rate. 50%. $\frac{2}{4}$ negatives incorrectly class ✓

Now we add some additional predictors to the model:

```
glm2_out <- glm(tall ~ ind + pnk + shoe_wm + shoe, data = hg)
summary(glm2_out)
```

```
Call:
glm(formula = tall ~ ind + pnk + shoe_wm + shoe, data = hg)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.05796    0.65979   0.088 0.930442
ind         -0.02461    0.01519  -1.620 0.113042
pnk          0.02096    0.01284   1.632 0.110491
shoe_wmw    -0.43299    0.11121  -3.894 0.000366 ***
shoe         0.11040    0.03223   3.425 0.001434 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.08983291)

    Null deviance: 10.0000  on 44  degrees of freedom
Residual deviance:  3.5933  on 40  degrees of freedom
AIC: 25.963
```

] → not sig anymore

```
Number of Fisher Scoring iterations: 2
```

(f) According to the model with the additional predictors, how many times more likely is a person wearing an "m" shoe size to be "tall" compared to a person wearing a "w" size, if these people have otherwise the same characteristics?

log odds

$\beta_{shoe\_wm} = -0.43299$ odds ratio $= e^{\beta} = e^{-0.43299}$ → this is for women so take inverse

A person wearing an m shoe size is $e^{0.43299}$ more likely to be tall. ✓

(g) Give the value of the full-reduced model test statistic for testing whether the new variables `pnk`, `shoe_wm`, and `shoe` contribute significantly to predicting whether someone is "tall". In addition, give the degrees of freedom of the chi-squared distribution from which we obtain the p value associated with the test statistic.

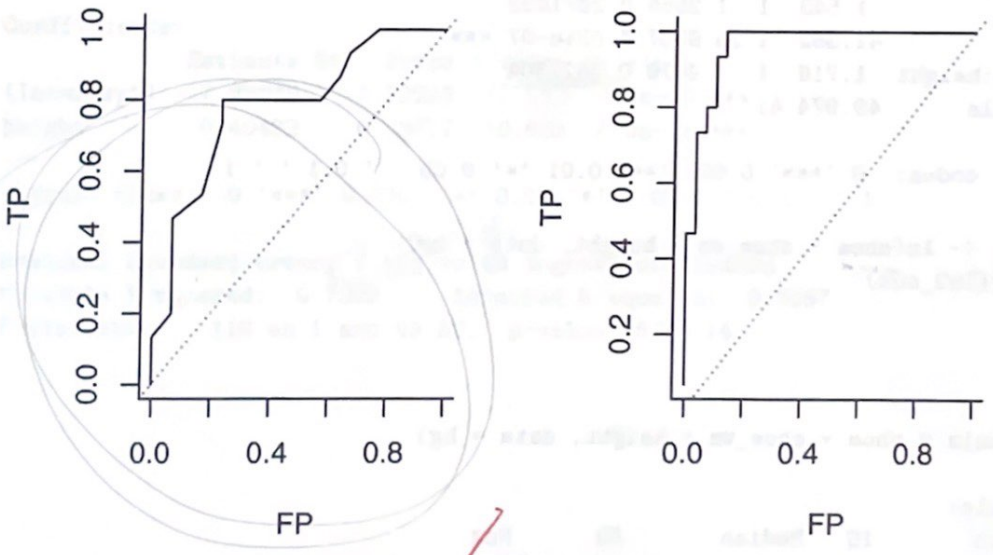
43-40=3 degrees freedom? ✓

Calculated w/ deviance

$F_{stat} = Dev(full) - Dev(reduced) = 3.5933 - 45.702$

or its $= Dev(red) - Dev(full) = 45.702 - 3.5933$

(h) The two panels below show ROC curves for the two fitted models. Each curve is based on predicting the response values for the same set of observations used to fit the model. Which curve corresponds to the model with only the index finger length as the predictor?



✓

(i) What word of caution would you give to someone comparing the two models based on these ROC curves?

Beware of overfitting with the model on the right hand side. Ensure that adequate testing is done using training data to fit and testing data to compare performance. ✓

2.

Using the same data as in the previous question, three models are fitted for predicting a person's shoe size based on his or her other characteristics:

```
lm_out <- lm(shoe ~ shoe_wm + height + shoe_wm:height, data = hg)
library(car)
Anova(lm_out, type = "III")
```

```
Anova Table (Type III tests)

Response: shoe
              Sum Sq Df F value    Pr(>F)
(Intercept)   15.553  1 12.7601 0.0009221 ***
shoe_wm        1.543  1  1.2656 0.2671433
height        41.382  1 33.9507 7.631e-07 ***
shoe_wm:height  1.716  1  1.4076 0.2422804
Residuals     49.974 41
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lm2_out <- lm(shoe ~ shoe_wm + height, data = hg)
summary(lm2_out)
```

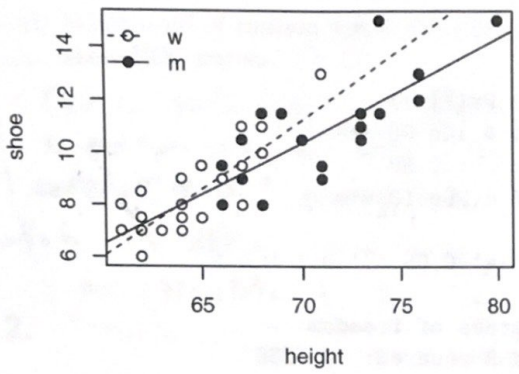
```
Call:
lm(formula = shoe ~ shoe_wm + height, data = hg)

Residuals:
     Min       1Q   Median       3Q      Max
-1.75874 -0.96734 -0.00074  0.88270  2.91610
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -20.60328    3.89513  -5.289 4.14e-06 ***
shoe_wmw      0.45059    0.48710   0.925     0.36
height        0.44172    0.05463   8.085 4.26e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.109 on 42 degrees of freedom
Multiple R-squared:  0.7383,    Adjusted R-squared:  0.7258
F-statistic: 59.23 on 2 and 42 DF,  p-value: 5.964e-13
```

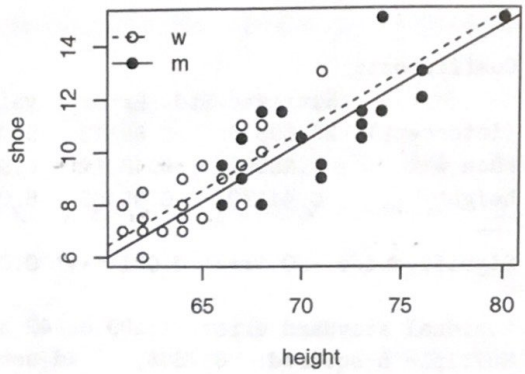
```
lm3_out <- lm(shoe ~ height, data = hg)
summary(lm3_out)
```

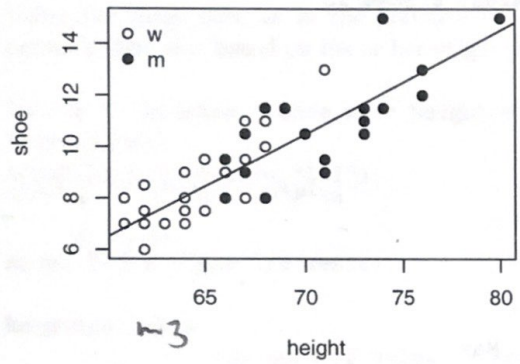
```
Call:
lm(formula = shoe ~ height, data = hg)

Residuals:
     Min       1Q   Median       3Q      Max
-1.88114 -0.90522 -0.04743  0.64294  2.90441

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.86078    2.52213  -7.082  9.8e-09 ***
height        0.40482    0.03727  10.863  6.6e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.108 on 43 degrees of freedom
Multiple R-squared:  0.7329,    Adjusted R-squared:  0.7267
F-statistic:   118 on 1 and 43 DF,  p-value: 6.6e-14
```

m1



m2





m3

(a) Does shoe size appear to have the same relationship to height for both gender sizings? Justify your answer carefully.

(p = 0.24228)

Yes, the interaction term is not significant, which means that the slope of shoe size can be shared across both genders.

(b) Does it appear necessary to allow different intercepts for the regression lines in the "w" and "m" groups? Justify your answer carefully.

No, the shoe_www feature was also not significant, indicating that there is no need to differentiate between the two and change the intercept.

in model with interaction + without
p77 0.052 (0.26, 0.36)

(c) What is the difference in the heights of the two lines (at any value of height) in the upper right plot?

0.44172.

from lm2_out

(d) Give an estimate of the error term variance and explain where you obtained it.

the error term variance in all these three models are pretty close.

$(\hat{\sigma})^2 = \frac{49.974}{41} \approx (1.109)^2 \approx (1.108)^2 \approx 1.21$

in lm_out in lm2_out in lm3_out

(e) Based on all the output above, what shoe size would you guess for a person with height 65? You need not compute an answer exactly, but may give an approximation. Explain your answer.

$\hat{Y} = (-17.86078) + (0.40482) \cdot 65$

from the plot in the lower left, we can approximate the shoe size is 8.4

from the questions aboved, we can assume heighs works significantly for the shoe size. so I omit the gender and use the model 3.

(f) How many observations are in the data set?

plot to estimate it.

N=45

3.

(a) Explain the essential difference between Type I and Type III sums of squares.

depends on order in which factor

Type I sums of squares works sequentially in anova() function, which is default in R.

Type III sums of squares measures contribution of each main effect in presence of interaction, does not depend on factors' order.

(b) What can go wrong with method of moments estimators of variance components in a model with random effects?

With method of moments estimators of variance components in a model with random effects you can get $\sigma_A^2 < 0$ which is not good because a squared value can never be negative.

(c) When should you choose to include a random effect in your model?

You should include a random effect in the model when you think there could be variability based on the levels of a random factor in the model.

(d) What is the advantage of having a balanced design in a two-way factorial experiment?

The advantage of having a balanced design in a two-way factorial design is that there is a clean decomposition of sums of squares. In the balanced design $SS_{trt} = SS_A + SS_B + SS_{AB}$ but in unbalanced we cannot assume the overall mean has the same weight so $SS_{trt} \neq SS_A + SS_B$

(e) In a one-way ANOVA model with 8 treatment groups, how many comparisons are entailed in making "all pairwise comparisons" of the treatment group means?

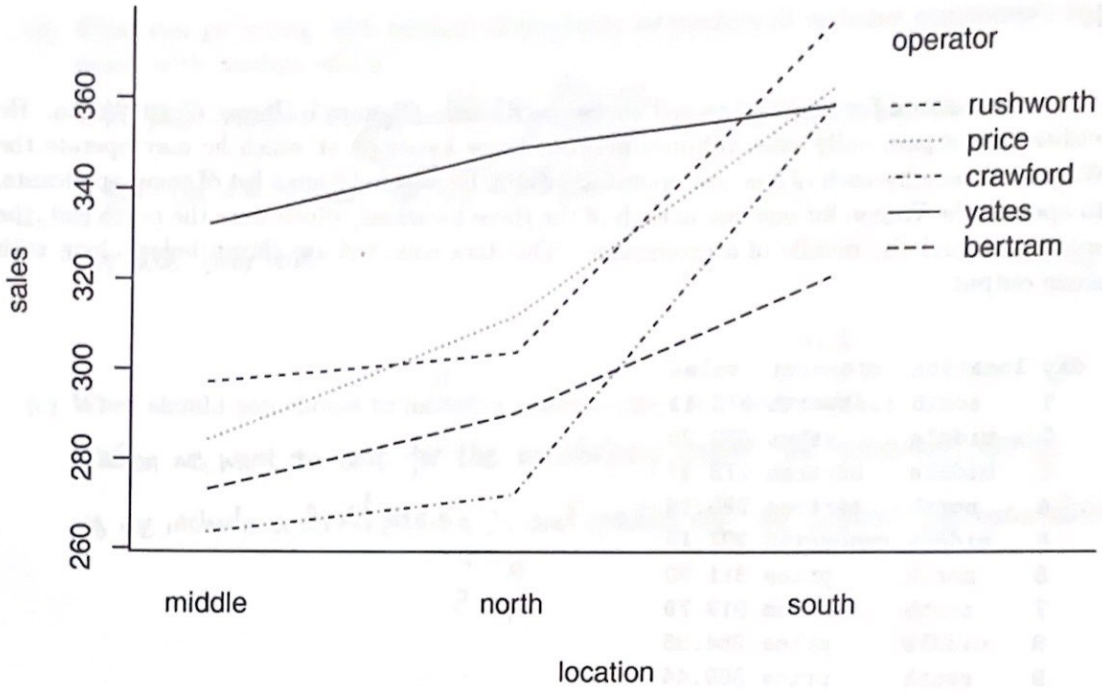
8 choose 2 $= \binom{8}{2} = \frac{8!}{2!\,6!} = \frac{8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1 \,(6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1)}$

$= \frac{56}{2} = \boxed{28}$

4.

It is a new season for Chompy to sell cheese curds from Chompy's Cheese Curd Wagon. He wishes to compare daily sales volumes between three locations at which he may operate the Wagon. He assigns each of five new operators, whom he selected from a list of eager applicants, to operate the Wagon for one day in each of the three locations, which were the north end, the south end, and the middle of a promenade. The data collected are shown below along with some output.

day	location	operator	sales
1	south	rushworth	375.11
2	middle	yates	332.20
3	middle	bertram	273.27
4	north	bertram	290.18
5	middle	rushworth	297.15
6	north	price	311.90
7	south	bertram	319.79
8	middle	price	284.35
9	south	price	360.44
10	north	crawford	272.03
11	middle	crawford	263.66
12	north	yates	348.52
13	north	rushworth	303.66
14	south	yates	356.66
15	south	crawford	357.02




```
lm_out <- lm(sales ~ location + operator, data = sales_oploc)

anova(lm_out)

Analysis of Variance Table

Response: sales
           Df  Sum Sq Mean Sq F value   Pr(>F)
location    ? 11067.6  5533.8 18.4606 0.001006 **
operator    ?  5362.5  1340.6  4.4723 0.034319 *
Residuals   ?  2398.1   299.8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2 ← location
4 ← operator
8 ← Residuals

(a) What kind of experimental design is Chompy following in his experiment?

Randomized Complete Block Design (RCBD)

(b) Replace the question marks in the above output with the correct degrees of freedom values.

	df		
location	$a-1$	$= 3-1$	$= 2$
operator	$b-1$	$= 5-1$	$= 4$
residuals	$(a-1)(b-1)$	$= (3-1)(5-1)$	$= 8$

(c) State the null and alternate hypotheses for which 4.4723 (found in the output) is the value of the test statistic.

$H_0: \sigma^2_{operator} = 0$

$H_1: \sigma^2_{operator} > 0$

(d) State the null and alternate hypotheses for which 18.4606 (found in the output) is the value of the test statistic. $\mu_i = \mu + \overline{\tau}_i$

$H_0: \mu_{north} = \mu_{south} = \mu_{middle}$

H_1: not all locations yield the same mean sales

(e) Chompy is wondering whether some operators tend to achieve greater sales at some locations than in others; that is, he wants to know if there are any particular operator/location combinations which could significantly increase sales. He asks if you can find any evidence of this in the data he collected. Answer him as well as you can.

Based on the interaction plot, this might be plausible. There are intersections and some converging slopes, which means interactions might be present. Overall, slopes stay pretty consistent, as performance, regardless of operator improves as you go from the middle, to north, then south. However, the few intersections we do see suggest the following:

1. Rushworth is ~~[illegible]~~ worse than Price in the north, and Price is worse in the middle and the south.
2. Bertram is much worse than ~~[illegible]~~ Crawford in the south, but better elsewhere.
3. Yates is much better than the field in the middle & north, but worse than 2 in the south.

This

Although this information might be valuable, it could also be due to noise. The interaction term isn't tested, so we are unaware of this information's significance. These conclusions should be taken with a grain of salt.