

Data check:

no duplicate IDs ✓

sex only 1 or 2 ✓

group only 1 or 2 ✓

GA-ultra value of 73 → need to set to missing

GA-est ✓

Bweight looks ok ✓

PPnum ✓

PD-pre ✓

PD-post ✓

I decided to change the value of GA-ultra = 73 to missing because I assume that a pregnancy would not be viable at that length of time. While there are some very low birth weights, there are enough of them that it would be bad practice to set them to missing.

1.b. mean ppnum ≈ 1.50

1.b. Poisson(1.5)

$$P(X \leq 3) = \sum_{x=0}^3 \frac{e^{-1.5}(1.5)^x}{x!} \approx .93436$$

calc check: $e^{-1.5} + e^{-1.5}(1.5) + \frac{e^{-1.5}(1.5)^2}{2} + \frac{e^{-1.5}(1.5)^3}{6} \approx .93436$

$$P(X > 3) = P(4 \text{ or more}) = 1 - .93436 = \boxed{.06564} \%$$

observed # = 0

$$\text{expected } \# = 453 (.06564) = 29.73492 \approx \boxed{30}$$

<4	≥4
observed	453
expected	423



need to do a χ^2 test

$H_0: \pi_1 = \pi_2$ where π_1 is true probability of women having at least 4 previous births (expected)

$H_A: \pi_1 \neq \pi_2$

π_2 : same but for observed

test statistic: 31.0274 $\sim \chi^2$

P-value = < .0001

reject H_0

We have sufficient evidence that the expected # of women with at least 4 previous births as modeled by the Poisson(1.5) distribution differs from the observed number. Thus, the Poisson(1.5) model is not a good approximation.

10c. using the K-S test

H_0 : Ultrasound GA follows a normal distribution

H_A : Ultrasound GA does not follow a normal distribution

D = .177425

P-value = < .0100

reject H_0

We have sufficient evidence that ultrasound GA does not follow a normal distribution.

We can also see this from the histogram which is extremely left-skewed and not bell-shaped.

1. d.

		GA Ultra		
		(0, 37)	[37, 40)	[40, ∞)
(0, 37)		41	11	0
GA Est	[37, 40)	7	216	27
[40, ∞)		0	34	116

note: 1 missing

$$H_0: K = 0$$

where K is the true agreement between the two GA measurements

$$\bar{K} = .6898$$

$$95\% \text{ C.I.} : (.6267, .7529)$$

↑
moderate-to-substantial agreement

$$Z = 18.6935 \quad p\text{-value} < .0001 \quad \text{reject } H_0$$

We have sufficient evidence that the true agreement between the two GA measurements is different than 0.

There is moderate-to-substantial agreement among the measurements.

i.e. χ^2 test of trend

$$H_0: p_0 = p_1 = p_2 = p_3$$

$$H_A: p_0 \leq p_1 \leq p_2 \leq p_3 \text{ or } p_0 \geq p_1 \geq p_2 \geq p_3$$

with at least one strict inequality

* using ultrasound GA to categorize preterm vs. not
since it's more accurate

$$Z = .0637 \text{ so } \chi^2 = (.0637)^2 = .00405769$$

p-value = .9492 fail to reject H_0

We do not have sufficient evidence that
the risk of preterm delivery varies
monotonically with the number of
previous pregnancies.

1 of 6

		preterm birth	
		1	0
		27	194
treatment group		2	210

$$\chi^2 = .8773 \quad p\text{-value} = .3489 \quad \text{fail to reject } H_0$$

$$H_0: \pi_1 = \pi_2$$

$$H_A: \pi_1 \neq \pi_2$$

We do not have sufficient evidence that the risk/probability of preterm birth differs between the treatment groups.

log. $H_0: M_{\text{base}} - M_{\text{after}} = 0$

$$H_A: M_{\text{base}} - M_{\text{after}} \neq 0$$

normality assumption met

$$t_{175} = .09 \quad p\text{-value} = .9286 \quad \text{fail to reject } H_0$$

We do not have sufficient evidence that the population mean average pocket depth change is different than 0. AKA we don't have evidence that average pocket depth changed after treatment in those women that received prenatal care.

1.b. normality assumption met
equal variances failed $F_{175,191} = 1.37$ $p = .0351$

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_A: \mu_1 - \mu_2 \neq 0$$

μ_1 = pop. mean average diff
in PD for group 1 (prenatal trat)

$$t_{366} \sim 3.65 \text{ p-value} = .0003$$

reject H_0

We have sufficient evidence that there is a diff.
in population mean average pocket depth
difference between the treatment groups.

$$\bar{X}_{\text{diff}} = .00338$$

$$\bar{X}_{\text{adiff}} = -.1730$$

It was better to get the
treatment prenatal \rightarrow saw improvement
those with treatment afterwards
had worse PD

1.b. While the treatment did not seem to significantly
help when we just look at the prenatal
group, it is important to note that receiving
prenatal treatment did help decrease PD
to some extent and those who did not
receive it just got worse over time.

Therefore, the treatment is not very
effective which may have been the
reason why there wasn't a difference
in birthweight. However, it is still preferred
women receive prenatal treatment if
possible.

$$2. a. i. y = \beta_1 I_{\text{cln1}} + \beta_2 I_{\text{cln2}} \approx y = XB + \epsilon$$

$$y: 168 \times 1 \quad \epsilon: 168 \times 1 \quad X = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \quad 168 \times 2 \quad B = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \quad 2 \times 1$$

2. a. ii. $H_0: \beta_1 = \beta_2 = 0$
 $H_A: \text{at least one } \beta_i \text{ is different from 0}$

$$C = [1 \ -1] \quad \Theta = [0]$$

$\hat{\beta}_1 \approx 63.32$ The mean percentage of baseline tumor volume remaining 30 days after treatment for those with clinical subtype HR+/HER2+ receiving chemotherapy is about 63.32%.

$\hat{\beta}_2 \approx 90.07$ The mean percentage of baseline tumor volume remaining 30 days after treatment for those with clinical subtype HR-/HER2+ receiving chemotherapy is about 90.07%.

$$\text{test statistic} = 30.77 \sim F_{1,166}$$

$$\text{p-value} = <.0001 \text{ reject } H_0$$

We have sufficient evidence that the population mean % of baseline tumor volume remaining 30 days after treatment for patients receiving chemotherapy differs across the clinical subtypes. Since the group mean for HR+/HER2+ clinical subtype is much lower, we have evidence that this treatment is more effective for those with HR+/HER2+ clinical subtype compared to those with HR-/HER2+ clinical subtype.

2.a. iii.

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_A: \sigma_1^2 \neq \sigma_2^2$$

$$F = 1.05 \sim F_{46, 120}$$

P-value = .8148 fail to reject H_0

We do not have sufficient evidence that the variance in chemotherapy response differs between the clinical subtype groups. While this assumption (homogeneity of variance) seems to be met, other assumption violations may be at play. My next step would be to check the normality assumption.

2.a. iv. $H_0: \frac{1}{2}B_1 = B_2$ interpretation!

$$H_A: \frac{1}{2}B_1 \neq B_2$$

01/18/19

Q.B. i. two-way ANOVA problem

Clinical Subtype	Treatment	
	Chemo	Hormone
HR+/HER2+	63.32 ¹¹	82.33 ¹²
HR-/HER2+	90.07 ²¹	79.69 ²²
	M ₁₁	M ₁₂

reference:
HR+/HER2+
hormone

reference cell coding

$$Y = M + d_1 + \beta_1 + \gamma_{11} = M + d_{HR-/HER2+} + \beta_{Chemo} + \gamma_{HR-/HER2+, Chemo}$$

M = mean response for HR+/HER2+ and hormone

d₁ = mean response difference for HR+/HER2+ and HR-/HER2+ for hormone (subtype effect)

β_1 = mean response difference for chemo and hormone for HR+/HER2+ (trt effect)

γ_{11} = treatment effect by clinical subtype

cell means coding

$$Y = \gamma_{Clin1, Chemo} + \gamma_{Clin1, Hor} + \gamma_{Clin2, Chemo} + \gamma_{Clin2, Hor}$$

$$H_0: M_{11} - M_{21} = 0 \quad \text{and} \quad M_{12} - M_{22} = 0$$

$$H_A: M_{11} - M_{21} \neq 0 \quad \text{or} \quad M_{12} - M_{22} \neq 0$$

$$M_{11} = M + \beta_{Chemo} = 82.33 - 19.01 = 63.32$$

$$M_{12} = M = 82.33$$

$$M_{21} = M + d_{HR-/HER2+} + \beta_{Chemo} + \gamma = 82.33 - 2.64 - 19.01 + 29.39 \\ = 90.07$$

$$M_{22} = M + d_{HR-/HER2+} = 82.33 - 2.64 = 79.69$$

$$MTC = \begin{bmatrix} \text{indph} & \text{indpc} & \text{indmh} & \text{indmc} \\ 0 & 1 & 0 & -1 \\ 1 & 0 & -1 & 0 \end{bmatrix} \quad \theta_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

using cell means
to test

$$\beta = \begin{bmatrix} \gamma_{1,0} \\ \gamma_{1,1} \\ \gamma_{2,0} \\ \gamma_{2,1} \end{bmatrix} \rightarrow$$

2. b. i. H_0 : $\mu_{11} - \mu_{12} = 0$ and $\mu_{21} - \mu_{22} = 0$
Overall by subtype

H_A : $\mu_{11} - \mu_{12} \neq 0$ or $\mu_{21} - \mu_{22} \neq 0$

$$C = \begin{bmatrix} \text{indph} & \text{indpc} & \text{indmh} & \text{indmc} \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \quad \Theta_0 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$F_{2,297} = 14.87 \quad p\text{-value} < .0001$$

$$\beta = \begin{bmatrix} \gamma_{1,0} \\ \gamma_{1,1} \\ \gamma_{2,0} \\ \gamma_{2,1} \end{bmatrix}$$

diff for HR+/HER2+ : $F_{1,298} = 23.49 \quad p\text{-value} < .0001$
diff for HR-/HER2+ : $F_{1,298} = 6.26 \quad p\text{-value} = .0129$

Based on the analysis and the group means/interactions, those with clinical subtype HR+/HER2+ are better off getting chemotherapy and those with clinical subtype HR-/HER2+ are better off getting hormone therapy. We know this because there are significant differences for both clinical subtypes and smaller mean responses are preferred.



$$2.b.ii. R^2_{\text{clinical model}} = .917930$$

$$R^2_{\text{molecular model}} = .958006$$

No, we cannot use R^2 to choose between the two models are not nested. However, we could use AIC or BIC to evaluate which model is more effective.

2.b.iii. we cannot carry out this test because we would need the candidate (molecular) model to be nested within the previous (clinical) model to get an appropriate F test statistic and p-value.

3. low birth weight = 1
normal birth weight = 0

3.a. for non-hypertensive

H_0 : low birth weight in child 1 is not associated /

independent with low birth weight of the second child

H_A : low birth weight in child 1 is associated with low birth weight of the second child

would prefer to use Fisher's Exact Test due to small sample sizes but will use χ^2 test for independence since I can actually use the formulas

$H_0: \pi_1 = \pi_2$

$H_A: \pi_1 \neq \pi_2$

where π_1 is the true proportion of LBW second children when first born had LBW

and π_2 is for second children without LBW

non-hypertensive group

<u>expected</u>	and born not LBW	and born LBW
1st born not LBW	$\frac{(70)(80)}{100} = 56$	$\frac{(70)(20)}{100} = 14$
1st born f. LBW	$\frac{(30)(80)}{100} = 24$	$\frac{(30)(20)}{100} = 6$

$$\chi^2_{\text{obs}} = \frac{(60-56)^2}{56} + \frac{(10-14)^2}{14} + \frac{(20-24)^2}{24} + \frac{(10-6)^2}{6} = 4.76$$

p-value = $1 - .970871452 = .0291$

(don't know if this is right)

reject H_0 , 1st and 2nd child birth weight not independent

hypertensive group:

expected table		2nd - not LBW	2nd c-LBW
1st born not LBW		$\frac{(60)(70)}{100} = 42$	$\frac{(60)(30)}{100} = 18$
1st born LBW		$\frac{(70)(40)}{100} = 28$	$\frac{(30)(40)}{100} = 12$

$$\chi^2 = \frac{(48-42)^2}{42} + \frac{(12-18)^2}{18} + \frac{(22-28)^2}{28} + \frac{(18-12)^2}{12} = 7.14$$

reject H_0

We have sufficient evidence that low birth rate of the first child and low birth rate of the second child are not independent for mothers with hypertension.

$$3. b. \text{ odds ratio} = \frac{(10)(60)}{(10)(20)} = \boxed{3.00 \quad \text{non-hypertensive}}$$

$$\text{Odds ratio} = \frac{(18)(48)}{(12)(22)} = \boxed{3.27 \quad \text{hypertensive}}$$

$$\text{Upper 95% C.I. : } e^{[\ln(\text{OR}) + 1.96 \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}]}$$

$$\text{Lower 95% C.I. : } e^{[\ln(\text{OR}) - 1.96 \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}]}$$

non-hypertensive

$$e^{[\ln(3) + 1.96 \sqrt{\frac{1}{10} + \frac{1}{20} + \frac{1}{10} + \frac{1}{60}}]} = 8.25$$

$$e^{[\ln(3) - 1.96 \sqrt{\frac{1}{10} + \frac{1}{20} + \frac{1}{10} + \frac{1}{60}}]} = 1.09$$

$$\boxed{95\% \text{ C.I.} = (1.09, 8.25)}$$

hypertensive

$$e^{[\ln(3.27) + 1.96 \sqrt{\frac{1}{18} + \frac{1}{22} + \frac{1}{12} + \frac{1}{48}}]} = 7.95$$

$$e^{[\ln(3.27) - 1.96 \sqrt{\frac{1}{18} + \frac{1}{22} + \frac{1}{12} + \frac{1}{48}}]} = 1.35$$

$$\boxed{95\% \text{ C.I.} = (1.35, 7.95)}$$

3.0.C. nonhypertensive:

$$P_1 = \text{probability of LBW for 1st child} = \frac{30}{100} = .3$$

$$P_2 = \text{probability of LBW for 2nd child} = \frac{20}{100} = .2$$

$$H_0: \pi_1 = \pi_2$$

$$H_A: \pi_1 \neq \pi_2$$

$$C_{.05} = \sum |z| > 1.963$$

$$Z = \frac{-3 - .2 - 0}{\sqrt{\frac{(.3)(.7)}{100} + \frac{(.2)(.8)}{100}}} = 1.64$$

fail to reject H_0

We do not have sufficient evidence that the true probability of LBW differs for the first and second born child in mothers without hypertension.

hypertensive:

$$P_1 = \frac{40}{100} = .4$$

$$P_2 = \frac{30}{100} = .3$$

$$Z = \frac{.4 - .3 - 0}{\sqrt{\frac{(1.4)(.6)}{100} + \frac{(.3)(.7)}{100}}} = 1.49$$

fail to reject H_0

We do not have sufficient evidence that the true probability/risk of LBW differs for the first and second born child in mothers with hypertension.



3.a ..

3.e. $y = \beta_1 + \beta_2 x$

$x=0$ non-hypertensive
 $x=1$ hypertensive

non-hypertensive : 30 out of 100 first born LBW

hypertensive : 40 out of 100 first borns LBW

Parameter	Estimate	Std. Error
β_1		
β_2	.4418	

OR for β_2 : $\frac{41.6}{31.7} = 1.25 = e^{\hat{\beta}_2}$

estimate for $\beta_2 = \ln(e^{\hat{\beta}_2}) = .4418$

3-f. γ_3 : The log odds of LBW for the first child is about γ_3 greater than the log odds of LBW for the second child, on average.

γ_4 : The log odds of LBW when the mother is hypertensive and it is the first born child is γ_4 less than the log odds of LBW for any other mother-child combination.

However, log odds is hard for many to understand, so we will interpret this another way.

Since γ_3 is positive, it means that the probability of LBW is higher for the first child than it is for the second child.

Since γ_4 is negative, a mother that is hypertensive and is having their first child actually decreased the probability of that child having a low birth rate.

The question is exactly opposite to how I think so my results are the opposite.

3.9. Using the 400 births (full data) could be problematic because each woman would be responsible for 2 observations/births.

This would mean that the observations are not independent and would thus, violate the independence assumption.

Assuming structure:

Mom ID	First born	Hypertension	LBW
1	1	1	1
1	0	1	1
2	1	0	0
2	0	0	0
:			