1 Mean depth of the ocean

The **depths** dataset shown below contains the depths of the ocean in the **alt** column and whether the sampled location is in the northern or southern hemisphere given by the **South** column (South=0 is the north and South=1 is the south). There are 200 samples from the north and 200 from the south.

```
head(depths)
```

```
##           X        lon       lat  alt water South
## 41995 41995  -87.21236 59.290367  190     1     0
## 11151 11151 -122.33034  5.554558 4167     1     0
## 43640 43640 -148.54790 36.237464 5447     1     0
## 8615   8615  -24.92364 21.625967 5063     1     0
## 8126   8126  177.18458 13.880370 5634     1     0
## 16548 16548   48.88215  3.229250 3691     1     0
```

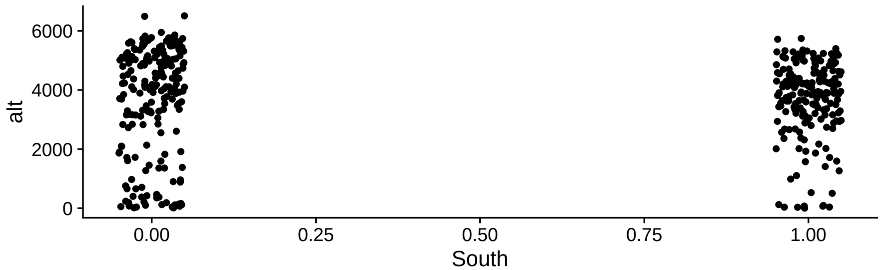
```
dim(depths)
```

```
## [1] 400   6
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3683.52      78.71    46.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1574 on 399 degrees of freedom
```

1. A linear regression output is shown. What type of regression is this?
2. What is the sample size?
3. What is the parameter of interest?
4. How many determinants are there for the parameter of interest?
5. Give the regression equation in terms of population parameters. Define each of the parameters in your model.
6. Provide the `R` code used to fit the regression equation.
7. What does the `Estimate` for `(Intercept)` represent? Would it be possible to calculate this value without running a regression? If yes, how?
8. How is the `Std. Error` calculated?
9. Show how the `t value` was calculated. State the implied hypothesis test.
10. Show how the `p-value` was calculated.
11. Calculate a 95% confidence interval for the parameter of interest. State the assumptions you used to calculate this CI.

2 Mean depth of the ocean in northern and southern hemisphere




```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3643.08     111.42  32.698   <2e-16 ***
## South          80.88     157.56   0.513    0.608
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1576 on 398 degrees of freedom
## Multiple R-squared: 0.0006617,Adjusted R-squared: -0.001849
## F-statistic: 0.2635 on 1 and 398 DF,  p-value: 0.608
```

```
t.test(alt ~ South, data = depths, var.equal = TRUE)
```

```
##  Two Sample t-test with alt by South
## t = -0.5133, df = 398, p-value = 0.608
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -390.6487  228.8787
## sample estimates:
## mean in group 0 mean in group 1
##        3643.080        3723.965
```

1. A linear regression output is shown. What is the parameter of interest?
2. How many determinants are there for the parameter of interest?
3. Give the regression equation in terms of population parameters. Define each of the parameters in your model.
4. Provide the `R` code used to fit the regression equation.
5. What does the `Estimate` for `South` represent? Would it be possible to calculate this value without running a regression? If yes, how?
6. Show how the `t value` for `South` was calculated. State the implied hypothesis test.
7. Show how the `p-value` was calculated. Interpret this p-value in the context of the problem.
8. Calculate a 95% confidence interval for the parameter of interest. State the assumptions you used to calculate this CI.

3 Ratio depth of the ocean in northern and southern hemisphere

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.20058    0.03058 268.144   <2e-16 ***
## South        0.02196    0.04278   0.513    0.608
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 2482673)
##
##     Null deviance: 988758010  on 399  degrees of freedom
## Residual deviance: 988103771  on 398  degrees of freedom
## AIC: 7029.1
##
## Number of Fisher Scoring iterations: 5
```

1. A linear regression output is shown. What is the parameter of interest?
2. Give the regression equation in terms of population parameters. Define each of the parameters in your model.
3. Provide the `R` code used to fit the regression equation.
4. What does the `Estimate` for `South` represent? Would it be possible to calculate this value without running a regression? If yes, how?
5. State the implied hypothesis test. Show how the `p-value` was calculated. Interpret this p-value in the context of the problem.
6. Calculate a 95% confidence interval for the % difference between the depths of the ocean in the South vs the North.
7. Are the depths of the ocean different in the South compared to the North? Explain.

4 Student drinking

A professor asked her sophomore students, How many drinks do you typically have per session (A drink is defined as one 12-ounce beer, one 4-ounce glass of wine, or one 1-ounce shot of liquor.) Some of the students didnt drink. Below are two fitted regressions based on the responses of the female and male students who did drink. `gender=1` is male and `gender=0` is female.

```
fit <- lm(drinks ~ gender, data = drinks)
summary(fit)
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.2947     0.2837  15.138  < 2e-16 ***
## gender        2.2238     0.4182   5.318  3.2e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.765 on 174 degrees of freedom
## Multiple R-squared: 0.1398,Adjusted R-squared: 0.1348
## F-statistic: 28.28 on 1 and 174 DF,  p-value: 3.197e-07
```

```
fit <- glm(drinks ~ gender, data = drinks, family = gaussian(link=log))
summary(fit)
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.45739    0.06606  22.062  < 2e-16 ***
## gender       0.41726    0.08115   5.142 7.27e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 7.646385)
##
##     Null deviance: 1546.7  on 175  degrees of freedom
## Residual deviance: 1330.5  on 174  degrees of freedom
## AIC: 861.48
##
## Number of Fisher Scoring iterations: 5
```

1. Two linear regression outputs are shown. For each, what is the parameter of interest?
2. For each, give the regression equation in terms of population parameters. Define each of the parameters in your model.
3. What does the `Estimate` for `gender` represent in each regression? Would it be possible to calculate this value without running a regression? If yes, how?
4. State the implied hypothesis test. Show how the `p-value` was calculated. Interpret this p-value in the context of the problem.
5. Calculate a 95% confidence interval for the % difference between the number of drinks in males vs females.
6. Are the number of drinks different in males vs. females? Explain.

5 Breastfeeding and respiratory infection I

A total of 189,612 person-years of follow up were accumulated over the course of the study: 151,690 among infants who were being breastfed and 37,922 among infants not being breastfed. Over the course of follow up the investigators identified 514,230 incident cases of respiratory infection among breastfeeding infants and 140,312 among non-breastfeeding infants. Calculate the crude incidence rate difference and 95% CI comparing infants who were not breastfed with those who were.

```
fit <- glm(cases ~ -1 + PT + PT:not_breastfed, family = poisson(link = identity))
summary(fit)
```

```
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## PT               3.390006   0.004727  717.10   <2e-16 ***
## PT:not_breastfed 0.310010   0.010951   28.31   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance:        Inf  on 2  degrees of freedom
## Residual deviance: 1.1195e-10  on 0  degrees of freedom
## AIC: 32.678
##
## Number of Fisher Scoring iterations: 2
```

1. Give the implied regression equation in terms of population parameters. Define each of the parameters in your model.
2. What does the `Estimate` for `PT` and `PT:not_breastfed` each represent?
3. Explain why there is a `-1` in the `glm` formula.

6 Breastfeeding and respiratory infection II

A total of 189,612 person-years of follow up were accumulated over the course of the study: 151,690 among infants who were being breastfed and 37,922 among infants not being breastfed. Over the course of follow up the investigators identified 514,230 incident cases of respiratory infection among breastfeeding infants and 140,312 among non-breastfeeding infants. Calculate the crude incidence rate difference and 95% CI comparing infants who were not breastfed with those who were. We are interested in calculating the incidence rate ratio and 95% CI comparing infants who were not breastfed with those who were.

1. What is the parameter of interest?
2. Give the regression equation in terms of the population parameters including the parameter of interest. Define each of the parameters in your model.
3. How would you fit this model in `R`? First show what the data looks like, and then provide the `R` code to fit the regression model given in part 2.
4. The fitted regression equation is given by:

   $$\widehat{\log(\mu)} = 1.22 + 0.0875 \cdot NBF + \log(PT)$$

   where μ is the expected number of cases of respiratory infection, $NBF = 1$ if not breastfed and 0 otherwise, and PT is the person time in years. Calculate the fitted values, i.e., the expected number of cases, for both the not breastfed and breastfed group. Do you notice anything in particular about these fitted values? If yes, explain.

7 Malaria control with bednets

See the 2018 Lancet article *Efficacy of Olyset Duo, a bednet containing pyriproxyfen and permethrin, versus a permethrin-only net against clinical malaria in an area with highly pyrethroid-resistant vectors in rural Burkina Faso: a cluster-randomised controlled trial* (`Bednets.pdf` in `A9` folder of myCourses) by Tiono et. al. exposure=1 is the new bednet, and exposure=0 is the existing bednet. A poisson regression with log link is fitted to the data and provides the following output:

```
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.68314    0.02432  28.092  < 2e-16 ***
## exposure    -0.26687    0.03286  -8.121 4.62e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1381.2  on 23  degrees of freedom
## Residual deviance: 1316.0  on 22  degrees of freedom
## AIC: 1476.7
##
## Number of Fisher Scoring iterations: 5
```

1. What is the regression equation in terms of population parameters? Define all your parameters.
2. What is the fitted regression equation for the expected counts of malaria cases?
3. What is the rate ratio and 95% CI comparing PPF-treated (exposure=1) to Standard long-lasting insecticidal (exposure=0) nets?
4. Do you expect this model to be a good fit for the data in Table 2? Explain why or why not.

8 Population mortality rates in Denmark

We can fit the following simple (multiplicative) rate ratio model to the patterns of mortality rates for 1980-1984 and 2000-2004. The reference cell is females 70-74, 1980-84. R = rate. M = multiplier.

Year	Age	Female (F)			Male (M)			
1980-1984	70-74	R_F			R_F		$\times M_M$	
	75-79	R_F	$\times M_{75}$		R_F	$\times M_{75}$	$\times M_M$	
	80-84	R_F	$\times M_{80}$		R_F	$\times M_{80}$	$\times M_M$	
	85-89	R_F	$\times M_{85}$		R_F	$\times M_{85}$	$\times M_M$	
2000-2004	70-74	R_F		$\times M_{20y}$	R_F		$\times M_M$	$\times M_{20y}$
	75-79	R_F	$\times M_{75}$	$\times M_{20y}$	R_F	$\times M_{75}$	$\times M_M$	$\times M_{20y}$
	80-84	R_F	$\times M_{80}$	$\times M_{20y}$	R_F	$\times M_{80}$	$\times M_M$	$\times M_{20y}$
	85-89	R_F	$\times M_{85}$	$\times M_{20y}$	R_F	$\times M_{85}$	$\times M_M$	$\times M_{20y}$

Year	Age	Female_deaths	Female_PT	Female_rate	Male_deaths	Male_PT	Male_rate
1980-1984	70-74	15989	586882.8	0.0272439	23810	456908.21	0.0521111
1980-1984	75-79	20838	454142.7	0.0458843	24707	300318.92	0.0822692
1980-1984	80-84	24073	297678.6	0.0808691	20319	167303.51	0.1214499
1980-1984	85-89	20216	147771.7	0.1368057	13524	74295.83	0.1820291
2000-2004	70-74	13912	521561.9	0.0266737	17360	436994.92	0.0397259
2000-2004	75-79	19731	471945.5	0.0418078	22477	341362.82	0.0658449
2000-2004	80-84	25541	369989.9	0.0690316	22992	217929.72	0.1055019
2000-2004	85-89	27135	226798.1	0.1196439	17444	104009.58	0.1677153
2005-2009	70-74	12179	540568.6	0.0225300	15782	472012.84	0.0334355
2005-2009	75-79	17273	444474.2	0.0388616	19547	344351.34	0.0567647
2005-2009	80-84	23513	363534.1	0.0646789	21781	230530.24	0.0944822
2005-2009	85-89	26842	237877.3	0.1128397	17811	114485.04	0.1555749

Rate =	______	×______	×______	×______	×______	×______
		if 75-79	if 80-84	if 85-89	if male	if 2000-04
$\log[Rate] =$	______	+______	+______	+______	+______	+______
		if 75-79	if 80-84	if 85-89	if male	if 2000-04
$\log[Rate] =$	______	+______	+______	+______	+______	+______
		$\times$ I_{75-79}	$\times$ I_{80-84}	$\times$ I_{85-89}	$\times$ I_{male}	$\times$ $I_{2000-04}$

where each 'I' is a (0/1) indicator of the category in question. By using both the 0 and 1 values of each I, this 6-parameter equation produces a fitted value for each of the $4 \times 2 \times 2 = 16$ cells.

1. How many determinants are there for the mortality rate? How many parameters are required to represent these determinants?
2. Estimate the parameters using a calculator, and fill in the blanks in the regression equations.
3. Interpret the estimate for the M_{80} parameter.
4. Fill in the blanks: The corresponding regression equation would reparametrize ______ parameters as a function of ______ parameters.

9 Kidney stone removal procedures 1

The 1986 BMJ article *Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy* by Charig et. al, was a study designed to compare different methods of treating kidney stones in order to establish which was the most cost effective and successful. The procedure, either open surgery, or percutaneous nephrolithotomy (PN, a keyhole surgery procedure), was defined to be successful if stones were eliminated or reduced to less than 2 mm after three months. The study collected cases of kidney stones treated at a particular UK hospital during 1972-1985. The counts of successes for the two surgical procedures were:

	Unsuccessful	Successful	Total
Open surgery	77	273	350
PN	61	289	350
Total	138	562	700

```
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.5556     0.1409 -11.040   <2e-16 ***
## open          0.2899     0.1911   1.517    0.129
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2.3148e+00  on 1  degrees of freedom
## Residual deviance: 9.9920e-15  on 0  degrees of freedom
## AIC: 15.696
##
## Number of Fisher Scoring iterations: 3
```

1. A logistic regression output with logit link is shown. What is the parameter of interest?
2. Give the regression equation in terms of population parameters. Define each of the parameters in your model.
3. Fill in the blanks: The corresponding regression equation would reparametrize ______ parameters as a function of ______ parameters.
4. What should the data look like so that it can be used in a regression routine?
5. Provide the `R` code used to fit the regression equation.
6. Provide the fitted regression equation for the log odds.
7. Provide the fitted regression equation for the odds.
8. Provide the fitted regression equation for the risk.
9. What does the `Estimate` for `open` represent? Would it be possible to calculate this value without running a regression? If yes, how?
10. State the implied hypothesis test for `open`. Show how the `p-value` was calculated. Interpret this p-value in the context of the problem.
11. Calculate a 95% confidence interval for the parameter of interest.
12. What is the risk of unsuccessful surgery in the open surgery group? in the PN group?

10 Kidney stone removal procedures 2

Below are the same outcomes tabulated by the size of the kidney stone (smaller than 2cm/at least 2cm in diameter):

$<$ 2cm	Unsuccessful	Successful	Total
Open surgery	6	81	87
PN	36	234	270
Total	42	315	357

$\geq$ 2cm	Unsuccessful	Successful	Total
Open surgery	71	192	263
PN	25	55	80
Total	96	247	343

```
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.9366     0.1704 -11.361  < 2e-16 ***
## open         -0.3572     0.2291  -1.559    0.119
## size          1.2606     0.2390   5.274 1.33e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 33.1239  on 3  degrees of freedom
## Residual deviance:  1.0082  on 1  degrees of freedom
## AIC: 26.355
##
## Number of Fisher Scoring iterations: 3
```

1. A logistic regression output with logit link is shown (open=1 is open surgery and 0 otherwise, size=1 is $\geq 2cm$ and 0 otherwise). Give the regression equation in terms of population parameters. Define each of the parameters in your model.
2. Fill in the blanks: The corresponding regression equation would reparametrize ______ parameters as a function of ______ parameters.
3. What should the data look like so that it can be used in a regression routine?
4. Provide the `R` code used to fit the regression equation.
5. Provide the fitted regression equation for the log odds.
6. Provide the fitted regression equation for the odds.
7. Provide the fitted regression equation for the risk.
8. Interpret the `Estimate` for `open`.
9. Calculate a 95% confidence interval for the `open` parameter.
10. What is the risk of unsuccessful surgery in the open surgery group with kidney stones less than 2cm?

11 Diabetes cohort data 1

	Dead	Censored	Total
Type II	218	326	544
Type I	105	253	323
Total	323	579	902

```
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.8794     0.1161  -7.576 3.58e-14 ***
## type          0.4770     0.1454   3.282  0.00103 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1.0978e+01  on 1  degrees of freedom
## Residual deviance: 1.4033e-13  on 0  degrees of freedom
## AIC: 16.858
##
## Number of Fisher Scoring iterations: 2
```

1. A logistic regression output with logit link is shown. What is the parameter of interest?
2. Give the regression equation in terms of population parameters. Define each of the parameters in your model.
3. Fill in the blanks: The corresponding regression equation would reparametrize ______ parameters as a function of ______ parameters.
4. What should the data look like so that it can be used in a regression routine?
5. Provide the `R` code used to fit the regression equation.
6. Provide the fitted regression equation for the log odds.
7. Provide the fitted regression equation for the odds.
8. Provide the fitted regression equation for the risk.
9. What does the `Estimate` for `type` represent? Would it be possible to calculate this value without running a regression? If yes, how?
10. Calculate a 95% confidence interval for the parameter of interest.
11. What is the risk of death for males aged 45 living with Type II diabetes?
12. What is the risk of death for females aged 57 living with Type II diabetes?

12 Diabetes cohort data 2

Below are the same outcomes tabulated by age:

≤ 40	Dead	Censored	Total
Type II	0	15	15
Type I	1	129	130
Total	1	144	145

> 40	Dead	Censored	Total
Type II	218	311	529
Type I	104	124	228
Total	322	435	757

```
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -4.9525     1.0036  -4.935 8.02e-07 ***
## type         -0.1816     0.1595  -1.139    0.255
## age           4.7781     1.0108   4.727 2.28e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 133.81237  on 3  degrees of freedom
## Residual deviance:   0.18471  on 1  degrees of freedom
## AIC: 20.745
##
## Number of Fisher Scoring iterations: 5
```

1. A logistic regression output with logit link is shown (type=1 is Type II and 0 otherwise, age=1 is Age > 40 and 0 otherwise). Give the regression equation in terms of population parameters. Define each of the parameters in your model.
2. Fill in the blanks: The corresponding regression equation would reparametrize ______ parameters as a function of ______ parameters.
3. What should the data look like so that it can be used in a regression routine?
4. Provide the `R` code used to fit the regression equation.
5. Provide the fitted regression equation for the log odds.
6. Provide the fitted regression equation for the odds.
7. Provide the fitted regression equation for the risk.
8. Interpret the `Estimate` for `type`.
9. Calculate a 95% confidence interval for the `type` parameter.
10. What is the risk of death for an individual who is 35 years old with Type II diabetes? Why isnt the risk for this individual equal to 0?

13 Cesarean section and transmission of HIV

To evaluate the relation between elective cesarean section and vertical (mother-to-child) transmission of human immunodeficiency virus type 1 (HIV-1), the authors performed a meta-analysis using data on individual patients from 15 prospective cohort studies.

Mode of Delivery	Covariate: No. of periods of antiretroviral therapy	Covariate: Advanced maternal disease	Covariate: Low birth weight of infant (<2500 g)	No. of Mother–Child Pairs	No. of HIV-1–Infected Children
Elective cesarean	0	No	No	372	30
Other	0	No	No	3850	652
Elective cesarean	0	Yes	No	28	5
Other	0	Yes	No	303	74
Elective cesarean	0	No	Yes	110	17
Other	0	No	Yes	767	196
Elective cesarean	0	Yes	Yes	27	4
Other	0	Yes	Yes	114	40
Elective cesarean	1 or 2	No	No	41	0
Other	1 or 2	No	No	441	49
Elective cesarean	1 or 2	Yes	No	23	3
Other	1 or 2	Yes	No	186	33
Elective cesarean	1 or 2	No	Yes	7	0
Other	1 or 2	No	Yes	83	22
Elective cesarean	1 or 2	Yes	Yes	10	3
Other	1 or 2	Yes	Yes	54	19
Elective cesarean	3	No	No	124	2
Other	3	No	No	878	49
Elective cesarean	3	Yes	No	34	1
Other	3	Yes	No	208	24
Elective cesarean	3	No	Yes	25	0
Other	3	No	Yes	109	11
Elective cesarean	3	Yes	Yes	8	1
Other	3	Yes	Yes	38	6

	Exposed	Unexposed	Total
Cases	a_i	b_i	M_{1i}
Controls	c_i	d_i	M_{0i}
Total	N_{1i}	N_{0i}	T_i

$$OR_{MH} = \frac{\sum_i \frac{a_i d_i}{T_i}}{\sum_i \frac{b_i c_i}{T_i}}$$

$$RR_{MH} = \frac{\sum_i \frac{a_i N_{0i}}{T_i}}{\sum_i \frac{b_i N_{1i}}{T_i}}$$

1. You only have access to a hand calculator. You are asked to provide a evidence on whether Elective cesarean is protective for HIV vertical transmission or not. Explain what you would do.
2. Several regression outputs are shown below. For each, provide the regression equation in terms of the population parameters being fit.
3. When applicable, explain what the `Estimate` for `caesarian` represents.
4. What statement can you make about c-sections and HIV vertical transmission based on these models? Provide statistical evidence to support your answer.

```
## Call:
## glm(formula = cbind(n.hivpos, n.hivneg) ~ 1, family = binomial(link = logit),
##     data = ds)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.671      0.031     -54   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 293.95  on 23  degrees of freedom
## Residual deviance: 293.95  on 23  degrees of freedom
## AIC: 387.8
##
## Number of Fisher Scoring iterations: 4
```

```
## Call:
## glm(formula = cbind(n.hivpos, n.hivneg) ~ caesarian, family = binomial(link = logit),
##     data = ds)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.606      0.032   -50.2   <2e-16 ***
## caesarian     -0.815      0.132    -6.2    7e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 293.95  on 23  degrees of freedom
## Residual deviance: 247.78  on 22  degrees of freedom
## AIC: 343.7
##
## Number of Fisher Scoring iterations: 4
```

```
## Call:
## glm(formula = cbind(n.hivpos, ds$n.hivneg) ~ caesarian + ART1or2 +
##     ART3 + m.advancedHIV + c.LBW, family = binomial(link = logit),
##     data = ds)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -1.608      0.041   -39.6   <2e-16 ***
## caesarian       -0.852      0.134    -6.3    2e-10 ***
## ART1or2         -0.362      0.106    -3.4    6e-04 ***
## ART3            -1.178      0.114   -10.3   <2e-16 ***
## m.advancedHIV    0.535      0.090     6.0    3e-09 ***
## c.LBW            0.581      0.075     7.8    9e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 293.945  on 23  degrees of freedom
## Residual deviance:  18.393  on 18  degrees of freedom
## AIC: 122.3
##
## Number of Fisher Scoring iterations: 4
## Call:
## glm(formula = cbind(n.hivpos, ds$n.hivneg) ~ caesarian + ART1or2 +
##     ART3 + m.advancedHIV + c.LBW, family = binomial(link = log),
##     data = ds)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -1.793      0.034   -53.2   <2e-16 ***
## caesarian       -0.720      0.119    -6.0    2e-09 ***
## ART1or2         -0.278      0.087    -3.2    0.001 **
## ART3            -1.016      0.104    -9.8   <2e-16 ***
## m.advancedHIV    0.409      0.068     6.0    2e-09 ***
## c.LBW            0.453      0.057     7.9    2e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 293.945  on 23  degrees of freedom
## Residual deviance:  21.295  on 18  degrees of freedom
## AIC: 125.2
##
## Number of Fisher Scoring iterations: 5
```

14 Smoking among women in Whickham, UK

Consider the following *age stratified* mortality data (Rothman, Table 1-2) from a study that looked at smoking habits of residents of Whickham, England, in the period 1972-1974 and then tracked the survival over the next 20 years of those who were interviewed.

		Smoking		
Age	Vital Status	Yes	No	Total
18-24	Dead	2	1	3
	Alive	53	61	114
	Risk	0.04	0.02	0.03
25-34	Dead	3	5	8
	Alive	121	152	273
	Risk	0.02	0.03	0.03
35-44	Dead	14	7	21
	Alive	95	114	209
	Risk	0.13	0.06	0.09
45-54	Dead	27	12	39
	Alive	103	66	169
	Risk	0.21	0.15	0.19
55-64	Dead	51	40	91
	Alive	64	81	145
	Risk	0.44	0.33	0.39
65-74	Dead	29	101	130
	Alive	7	28	35
	Risk	0.81	0.78	0.79
75+	Dead	13	64	77
	Alive	0	0	0
	Risk	1.00	1.00	1.00

1. Calculate the crude odds ratio for smoking vs. not smoking.
2. You are asked to provide evidence on whether smoking affects mortality. Provide a full analysis plan, including a regression model with population parameters, the parameter of interest, what the data should look like to fit in a regression routine and the `R` code for the regression routine. State concretely what you would provide as evidence.