STK4900 Oblig 2

Daniel Heinesen, daniehei

10. april 2019

1 Problem 1

1.1 a)

We have a data set where the outcome is whether the female crab has one or more satellites $y = 1$, or none $y = 0$. We are looking for a regression that can, given the covariants, give us a probability that the female crab have satellites. This means that we are looking for a regression model that gives us a probability for a binary outcome. The best choice for such a model is a logistic regression model

$$p(x_1, ..., x_p) = \frac{\exp(\beta_0 + \beta_1 x_1 + ... + \beta_n x_n)}{1 + \exp(\beta_0 + \beta_1 x_1 + ... + \beta_n x_n)}, \tag{1}$$

where p is the desired probability, x_i the covariates and β_i their fitted coefficients.

Is that the best way to describe β_i?

1.2 b)

We want to find the odds ratio between crabs that differ with one centimetre in with. We know that with width as the only covariant, we define the odds as

$$\frac{p(x)}{1 - p(x)} = \exp(\beta_0 + \beta_1 x). \tag{2}$$

From this we get the odds ratio for a difference in one centimetre

$$OR = \frac{p(x+1)/[1 - p(x+1)]}{p(x)/[1 - p(x)]} = \frac{\exp(\beta_0 + \beta_1 \cdot (x+1))}{\exp(\beta_0 + \beta_1 x)} = \exp(\beta_1 \cdot 1) = \exp(\beta_1). \tag{3}$$

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-12.3508	2.6287	-4.70	0.0000
width	0.4972	0.1017	4.89	0.0000

Tabell 1: Summary of the logical regression done the satellite crabs, with width as the only covariant.

We get the β's from the logical regression found in table 1. From this we see that

$$OR = \exp(0.4972) = 1.64. \tag{4}$$

This means that odds of a female crab having satellite males increase by 64% if the width of female increases by 1 cm.

If $p(1) = p(x+1)$ and $p(0) = p(x)$ are small, we can approximate the odds ratio with the relative risk $RR = p(1)/p(0)$. In this case we have that $p(1) = 6.8 \cdot 10^{-6}$ and $p(0) = 4.1 \cdot 10^{-6}$, which means that both are very small. This means that we can assume that $RR \approx OR$. And comparing them $OR = 1.6367$ and $RR = 1.6487$ we see that this is correct.

Since we know that

$$t = \frac{\beta}{se(\beta)} \tag{5}$$

is close to normally distributed, we can use this to find a confidence interval for the odds ratio. We simply do this by calculating the confidence interval for β_1 and then taking exp of this interval.

From tab. 2 we see that we get a confidence interval $CI = (1.35, 2.01)$. Since 1 is outside of this interval, we can say that width gives an significant increase in the probability of satellites.

	expcoef	lower	upper
(Intercept)	4.33e-06	2.50-08	0.00075
width	1.64	1.35	2.01

Tabell 2: Confidence interval for the odds ratio for crabs differing by one cm in width.

1.3 c)

1.4 d)

1.5 e)

2 Problem 2

2.1 a)

We have an outcome, medals, which is a count outcome, and we therefore assume that the count is distributed with a Poisson distribution $Y_i \sim Po(\lambda_i)$. Such a distribution is parametrized with a parameter λ, the rate. In our data we will have that this rate is dependent on several covariants, so we need a way to determine this dependence on the covariants. It is here we use Poisson regression. Given n independent subjects, we have that

- y_i is the count of the i^{th} subject
- x_{ij} is the j^{th} covariant for the i^{th} subject

We then define our model as

$$\lambda_i = \lambda(x_{1,i}, ..., x_{p,i}) = \exp(\beta_0 + \beta_1 x_{1,i} + ... + \beta x_{p,i}). \quad (6)$$

It is this model we want to fit to the medal counts. But there is something we have to be careful with: A country with a higher number of athletes will most likely have more medals than a country with fewer athletes. So instead of the medal count following the distribution $Y_i \ Po(\lambda_i)$ we instead say that they follow the distribution $Y_i \ Po(w_i\lambda_i)$, where w_i is the number of athletes representing the country. But how do use this in our model? We can see this from taking the expected value

$$E[Y_i] = w_i\lambda_i = w_i \exp(\beta_0 + \beta_1 x_{1,i} + ... + \beta x_{p,i}) = \exp(\log(w_i) + \beta_0 + \beta_1 x_{1,i} + ... + \beta x_{p,i}). \quad (7)$$

This means that to compensate for this imbalance of athletes, we can fit out model with $\log(w_i)$ as a covariant. This is what we call an offset. This we already *Log.athletes* in our data, we can just use this as the offset in our regression.

2.2 b)

To find a fit for our model we are going to use two methods. The first is to fit a model with all the covariant and see which of them are significant. The other method is to add one covariant after an other and use a two-way ANOVA to see if the addition is significant.

From table3 we see the summary of the fit with all covariant. We see that with p-values close to 0

3 Problem 3

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.8623	0.3191	-8.97	0.0000
Log.population	0.0275	0.0315	0.87	0.3831
GDP.per.cap	-0.0149	0.0032	-4.65	0.0000
Total1996	0.0118	0.0016	7.36	0.0000

Tabell 3: Summary of a Poisson regression with all the covariants. *Log.popilation* is the logarithm of the nation's population size per 1000, *GDP.per.cap* is the GDP per capita and $Total1996$ is the medal count for the previous Olympic Games.