

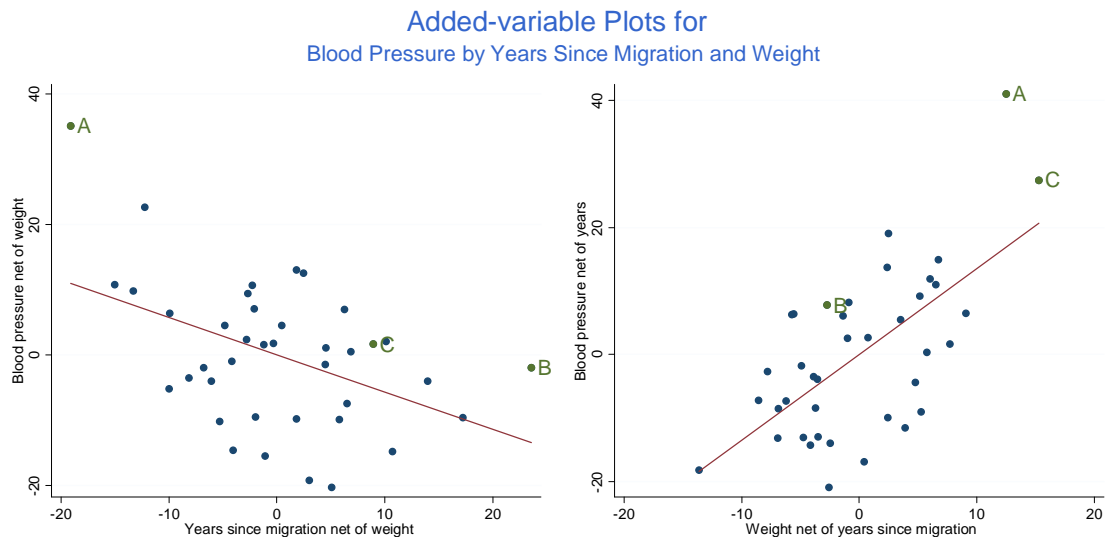
Mid Term Solutions

October 28, 2009

Please answer each question in a separate booklet. Take the 5% critical value of the normal distribution to be 2. For t , F and χ^2 tests report the statistic and the corresponding degrees of freedom. There's no need to calculate P -values.

[1] Blood Pressure in Peru (50%) Part a is worth 10 pts, parts b to i are 5 each

A team of anthropologist measured the systolic blood pressure of Peruvian Indians who had migrated from the mountains to towns and cities at much lower altitude. They expected that over time their blood pressure would decline, but were surprised to find that the correlation between blood pressure and time since migration was not significant. Then they had the brilliant idea of adjusting for weight. The figure below shows added variable plots for blood pressure as a function of each predictor net of the other.



Here are a few selected results of their analysis:

- Mean blood pressure was 127.3 mmHg. The standard deviation was 13.110.
- The correlation between blood pressure and years since migration was -0.0875.
- The regression of blood pressure on weight yielded an estimate of σ of 11.338. This estimate is also known as the root MSE.
- The additive model with both predictors had an R-squared of 0.4208 and the following estimated equation:

$$E(\text{Blood pressure}) = 50.32 + 1.354 \text{ weight} - 0.572 \text{ years}$$

- I added an interaction between weight and years, but it had a t statistic of 0.77, so I dropped it.

(a) Calculate a complete hierarchical analysis of variance table with the sums of squares and d.f. due to weight, years since migration, the interaction and the residual, in that order. If you get stuck use a RSS of 6,500 for the null model and skip the interaction, which is not needed later.

The RSS for the null model comes from the standard deviation: $13.11^2 \times 38 = 6,531.1$. Weight has a RSS of $11.338^2 \times 37 = 4,756.4$ so it explains 1,774.7. The two predictors together explain $0.4208 \times 6,531.1 = 2,748.3$ so years since migration has added 973.6 and the new RSS is 3,782.8. Now for the hard part: the square of the t -test for the interaction is an F statistic: $0.77^2 = x / ((3,782.8 - x) / 35)$ where x is the interaction SS. This gives $x = 3,782.8 \times 0.593 / (35 + 0.593) = 63.0$, so the final RSS is 3,719.8 on 35 d.f. This leads to the following hierarchical anova:

Source	SS	df
Weight	1,774.7	1
Years since migration	973.6	1
Interaction	63.0	1
Residual	3,719.8	35

(b) Test the hypothesis that blood pressure is independent of weight.

The test statistic is $F = 1,774.7 / (\frac{4,756.4}{37}) = 13.8$ on 1 and 37 d.f. (We reject.)

(c) Interpret the coefficient of years in the additive model and test its significance. Can we be confident that this relationship is linear?

One is tempted to conclude that blood pressure declines 0.572 mmHG for each year living at lower altitudes as long as the person maintains his or her weight. A more cautious interpretation with cross-sectional data is that migrants who have live longer at lower altitudes tend to have lower blood pressure than newer arrivals with the same weight, with an expected difference of 0.572 mmHg per year. To test significance we build an F test: $F = 973.6 / (3,783.9 / 36) = 9.26$ on 1 and 36 d.f. (We reject the hypothesis of no differences in blood pressure by years after adjusting for weight.) To check linearity we look at the first panel of the added-variable plot: the relationship looks reasonably linear.

(d) What's the correlation between the two variables in the left panel of the added variable plot?

That would be the partial correlation between blood pressure and years since migration controlling for weight. We see that time since migration explains 973.6 of the 4,756.4 that weight left unexplained. The proportion explained is 0.2047 and its square root with the same sign as the regression coefficient is the partial correlation: -0.452 . (Compare to -0.087 before adjusting for weight, a *suppressor* variable.)

(e) Estimate the expected blood pressure of two Indians of average weight, 63 Kg, one who just migrated and another who migrated 40 years ago (that's close to the range in the sample).

From the estimated model we get $50.32 + 1.354 \times 63 = 135.6$ for a new arrival, and the same *minus* $40 \times 0.572 = 22.9$, or 112.7 for someone who migrated 40 years ago.

(f) If you calculated the expected blood pressure for each person in the sample, what would be the linear correlation between observed and expected blood pressure?

The linear correlation between observed and predicted is the square root of R^2 , $\sqrt{0.4209} = 0.649$.

(g) The graphs identify three individuals, A, B and C. One of these has the largest residual by a clear margin. Which one is it? Explain why in a short sentence.

Clearly A, whose blood pressure is almost 40 mmHG higher than one would expect from his or her weight and years since migration. The added-variable plot shows that nobody else is that far away from the regression plane.

(h) The observation with the highest leverage turned out not to have much influence at all. Which one is it? *Hint*: there are two suspects, both migrated a long time ago, but one of them gained a lot more weight, and thus acquired leverage.

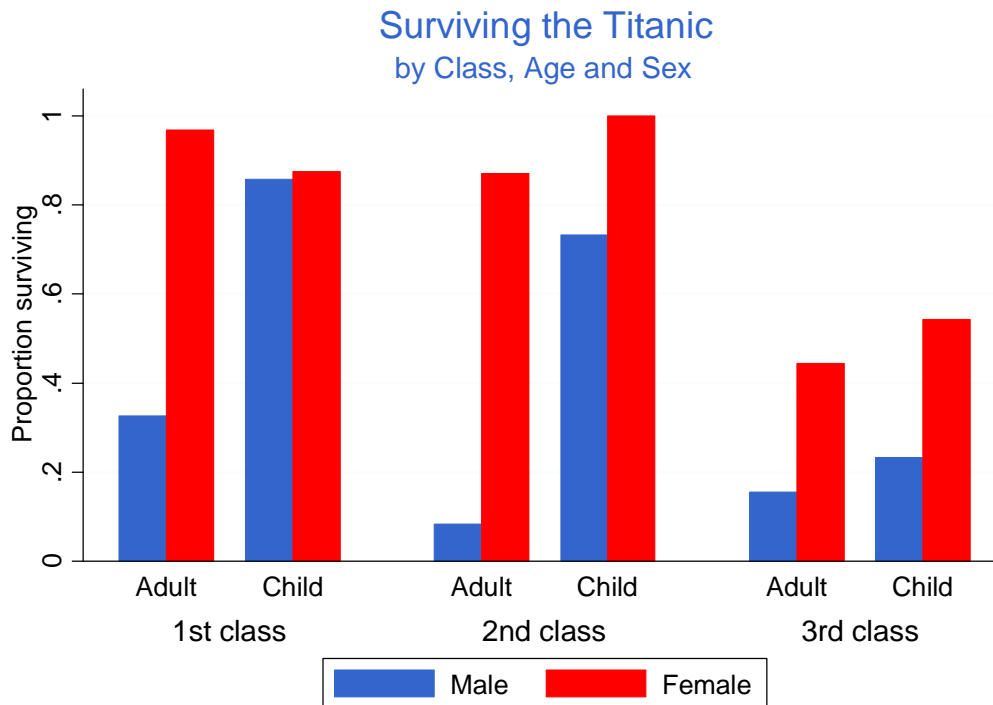
The fact that it is not influential rules out A. You might think that one panel suggests B and the other C, but leverage comes only from the values of the predictors. The winner is C, who migrated long ago but also weights about 15 Kg more than expected given time since migration.

(i) A quick run of the Box-Cox command in Stata finds a maximum log-likelihood of -143.24 at the m.l.e. $\hat{\lambda} = -0.96$, but also reports a log-likelihood of -144.54 for $\lambda = 1$. Do we need to transform the data? If so, which transformation would be appropriate?

No, we don't need to transform the outcome. Twice the difference in log-likelihoods between the optimal and no transformation is $2(144.54 - 143.23) = 2.40$. Under the null hypothesis of no transformation this is a chi-squared on one d.f. and is clearly not significant, so we fail to reject.

[2] Surviving the Titanic (50%) *Each part is worth 5 pts*

Harrell has an interesting dataset with information on the survival of 1046 individual passengers on the Titanic. Here we focus on three variables that are closely associated to survival: age, sex, and passenger class. To test the old dictum “women and children first” we will use an indicator for females and an indicator for passengers under age 18. We retain the distinction between 1st, 2nd and 3rd class passengers. This is what the data look like:



We have 12 groups of passengers. The overall percent surviving the disaster is 40.7%, but the chances of survival vary by age, sex and class. To examine this variation I fitted logistic regression models treating the 12 group as independent binomial observations.

(a) The null model has a deviance of 487.57. What’s the maximum likelihood estimate of the constant? What are the odds of survival?

The odds of survival are $\frac{0.407}{1-0.407} = 0.686$ and $\log(.686) = -0.376$ is the m.l.e. of the constant.

(b) The model with a dummy variable for females has a model chi-squared of 309.65 and an estimated equation of $\text{logit}(\text{pr}(\text{survive})) = -1.354 + 2.453 \text{ female}$. Interpret the coefficient of female in terms of an odds ratio and test its significance.

The odds ratio is $e^{2.453} = 11.6$, so the odds of survival of women are 11.6 times those of men. To test (the obvious) significance we need to compare this model with the null, which is what the model chi-squared does. We get $\chi^2 = 309.65$ on one d.f., which is highly significant.

(c) Adding a dummy for children reduces the deviance by 3.83. The coefficient of child is 0.400 with a standard error of 0.208. What can you say with 95% confidence about the odds of survival of children compared to adults of the same sex?

A 95% CI for the logit coefficient is $0.400 \pm 2(0.208)$ with bounds -0.008 and 0.808 . Exponentiating these values we get bounds of 0.99 and 2.24 for the odds ratio. We are 95% confident that the odds of survival for children are between one and two and a quarter times those of adults. Yes, that includes one. (This is a lot more informative than just saying that there is no significant difference.)

(d) The additive model with all three factors, including a dummy for children and dummies for 2nd and 3rd class, has a deviance of 73.53 and the following estimated equation:

$$\text{logit}(\text{pr}(\text{survive})) = -0.399 + 2.485 \text{ female} + 0.903 \text{ child} - 0.986 \text{ second} - 1.894 \text{ third}$$

Interpret the coefficients of second and third class in terms of odds ratios and test their significance. How would you test if this model fits the data?

Exponentiating the coefficients we get odds ratios of $e^{-0.986} = 0.373$ and $e^{-1.894} = 0.150$, so the odds of survival for passengers in 2nd class are 63% lower, and in 3rd class are 85% lower, than for 1st class passengers of the same sex and age group (adult or child). To test significance of these coefficients we need to compare this model with the model that includes only the dummies for female and child, which we can do from the deviances of $487.57 - 309.65 - 3.84 = 174.08$ and 73.53. The difference gives a $\chi^2 = 100.55$ on 7 d.f. (12 groups minus 5 parameters), which happens to be significant.

(e) Estimate the probability of survival for adult males travelling in first and in third class, convert to odds, and compute the odds ratio.

The logits are -0.399 for first class and $-0.399 - 1.894 = -2.293$. The probabilities are $\text{logit}^{-1}(-0.399) = 0.402$ and $\text{logit}^{-1}(-2.293) = 0.092$. The odds are $e^{-0.399} = 0.671$ and $e^{-2.293} = 0.100$, so the odds ratio is $\frac{0.100}{0.671} = 0.150$, the same value calculated more directly in d. (It was probably easier to compute first the odds and then the probabilities as $\frac{0.671}{1+0.671} = 0.401$ and $\frac{0.1}{1+0.1} = 0.091$.)

(f) If you did the calculation for adult females, would you get the same difference in probabilities? The same odds ratio? Explain. No need to calculate anything.

I would get exactly the same odds ratio because the model is additive in the log-odds, and there is no interaction between sex and class. I would not get the same difference in probabilities because adding the coefficient for females changes the odds and hence the probabilities and their difference, while keeping the odds ratio the same. (The probabilities for adult females are 0.890 and 0.548, with a difference of -0.342. For adult males they were 0.402 and 0.092 with a difference of -0.301.)

(g) Estimate the approximate marginal effect of travelling in third class compared to first class for an average passenger.

The marginal effect is $\pi(1 - \pi)\beta$, in this case $-0.407(1 - 0.407)1.894 = -0.457$ or a difference of 45.7 percentage points. (The actual difference was -0.309 for adult males, -0.301 for adult females, -0.424 for boys and -0.203 for girls. The approximation is not very good because we are dealing with a dummy variable and the effect is large.)

(h) I added an interaction between sex and age, creating a dummy for females under 18 (calculated as the product of female and child). The model with this additional variable has a deviance of 59.38 and estimated equation

$$\text{logit}(\text{pr}(\text{survive})) = -0.485 + 2.780 \text{ female} + 1.584 \text{ child} - 1.535 \text{ femXchild} - 1.030 \text{ second} - 1.947 \text{ third}$$

Is the interaction term significant? Interpret carefully the effects of female and child taking into account the interaction. *Hint*: women do better than men and children better than adults. What happens if a passenger is both female and a child?

Yes, the interaction is significant; comparing this model with the additive we get a χ^2 of $73.53 - 59.38 = 14.15$ on one d.f. Exponentiating the coefficients we get $e^{2.780} = 16.1$, $e^{1.584} = 4.87$ and $e^{-1.535} = 0.215$. I would interpret these results saying that the odds of survival of girls and women are 16 to 17 times those of adult men traveling in the same class, whereas the odds for boys are 5 times those of men in the same class. (The odds ratio is 16.1 for women and $e^{2.780+1.584-1.535} = 16.1 \times 4.87 \times 0.215 = 16.9$ for girls. These are close enough that they can be combined; girls get essentially the same advantage as women, as the interaction term offsets most of the child advantage.)

(i) It turns out that passenger class interacts with female and with child. Summarize in two sentences what this would mean in terms of odds ratios.

It means that the class disparities as measured by the odds ratios would differ by sex and by adult or child status. Alternatively, the odds ratios representing increased odds of survival depending on being a female and/or a child rather than an adult would differ by passenger class.

(j) What would you need to know to determine if you can trust the model deviances to assess goodness of fit?

The asymptotics depend on the size of the 12 groups going to infinity, so we would need assurances that the groups are large. Cochran's rule of thumb would require at least one survivor and one fatality per group and at least 5 survivors and five fatalities in 10 of the 12 groups. (We barely make it.)