

36-707: Regression Analysis Homework 4 Solutions

Fall 2007

Problem 1

Forward Stepwise gives AST and Urea. Backward gives Calving, Daysrec, Urea, Myopathy, Inflatat. Either of these was acceptable.

Problem 2

We find that the equation of the logistic regression is $\text{logit}(p) = -3.3 + 1.25 \cdot \text{Intensity}$. To test whether intensity is significant in predicting mouth movement, we can look at the t-statistic for intensity. We see it is extremely large at 11.13 and so we conclude that intensity is not independent of mouth movement. You should have also included a plot with the logistic regression curve on it.

Problem 3

The scatter plot matrix of the six predictors shows that there are clear differences in the Counterfeit ($Y=0$) and Genuine ($Y=1$) measurements. With the exception of the predictor Diagonal, it appears that all of the variables have a joint effect since the prevalence for $Y=0$ vs $Y=1$ changes along a diagonal. The predictor diagonal does not seem to have a joint effect since the prevalence for $Y=0$ vs $Y=1$ seems to change along the vertical. This means that only the predictor Diagonal will have an effect on the probability of being counterfeit. This makes sense since the measurements are intrinsically connected - ie if I move an image closer to the bottom, it will automatically be further from the top. The variable Diagonal seems to capture all of the other variables and makes Top, Bottom, Left, Right,

and Length all unnecessary. We'll check this theory with a logistic regression.

Indeed, we see that the algorithm does not converge which indicates a problem with multicollinearity. If we run the regression with only diagonal, we get a highly significant coefficient for Diagonal. I also accepted removing diagonal from the regression and running some sort of stepwise that eliminates clearly dependent variables like left or right. (Note I am not including the scatter plot matrix but you should have included it in your homework.)

Problem 4

a. I get $Y = 2.94 + 4.01x$. You should have a plot with this line.

b. I get $Y = 2.85 + 1.93x$. You should have a plot with this line. This is clearly biased downward since the slope is so much smaller than 4.

$$c. \tilde{\beta} = \hat{\beta}_1(b) * \frac{\hat{\sigma}_w^2}{\hat{\sigma}_w^2 - \sigma_u} = 3.93 \approx 4.01 = \hat{\beta}_1(a)$$

This slope is very similar to the slope from part a and much better than that of part b so we conclude the correction worked.

Problem 5

a. Draw a scatter plot.

b. We get $E[\text{length} | \text{Age} = t] = 192.8(1 - \exp(-.41(t - .08)))$.

c. CI = (170.3, 232.4)

Problem 6

a. not shown.

b. I find the best bandwidth is $h = 134$.

c. I find the best bandwidth is $h = 228$.

d. I find the best bandwidth is ≈ 5 .

e. Legendre polynomials give $d = 3$. You could use others.