GroupI_HW3

Cortinovis, Cvetinovic, Savarin, Stromieri

Contents

FSDS - Chapter 6

Ex 6.12

For the UN data file at the book's website (see Exercise 1.24), construct a multiple regression model predicting Internet using all the other variables. Use the concept of multicollinearity to explain why adjusted R^2 is not dramatically greater than when GDP is the sole predictor. Compare the estimated GDP effect in the bivariate model and the multiple regression model and explain why it is so much weaker in the multiple regression model.

```
## Loading required package: MASS
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

Solution

Ex 6.14

The data set30 Crabs2 at the book's website comes from a study of factors that affect sperm traits of male horseshoe crabs. A response variable, SpermTotal, is the log of the total number of sperm in an ejaculate. It has y = 19.3 and s = 2.0. The two explanatory variables used in the R output are the horseshoe crab's carapace width (CW, mean 18.6 cm, standard deviation 3.0 cm), which is a measure of its size, and color (1 = dark, 2 = medium, 3 = light), which is a measure of adult age, darker ones being older.

(a) Using the results shown, write the prediction equation and interpret the parameter estimates.

(b) Explain the differences in what is tested with the F statistic (i) for the overall model, (ii) for the factor(Color) effect, (iii) for the interaction term. Interpret each.

Solution

(a)
(b)

Ex 6.30

When the values of y are multiplied by a constant c, from their formulas, show that s_y and $\hat{\beta}_1$ in the bivariate linear model are also then multiplied by c. Thus, show that $r = \hat{\beta}_1(s_x/s_y)$ does not depend on the units of measurement.

Solution

Ex 6.42

You can fit the quadratic equation $E(Y) = \beta_0 + \beta_1 1x + \beta_2 x^2$ by fitting a multiple regression model with $x1 = x$ and $x2 = x^2$.

(a) Simulate 100 independent observations from the model $Y = 40.0 - 5.0x + 0.5x^2 + \epsilon$, where X has a uniform distribution over $[0, 10]$ and $\epsilon \sim N(0, 1)$. Plot the data and fit the quadratic model. Report how the fitted equation compares with the true relationship.

(b) Find the correlation between x and y and explain why it is so weak even though the plot shows a strong relationship with a large R^2 value for the quadratic model.

Solution

(a)
(b)

Ex 6.52

F statistics have alternate expressions in terms of R2 values.

(a) Show that for testing $H_o : \beta 1 = \cdots = \beta_p = 0$,

$$F = \frac{(TSS - SSE)/p}{SSE/[n - (p + 1)]} = \frac{R^2/p}{(1 - R^2)/[n - (p + 1)]}.$$

Explain why larger values of R^2 yield larger values of F.

(b) Show that for comparing nested linear models,

$$F = \frac{(SSE_0 - SSE_1)/(p_1 - p_0)}{SSE_1/[n - (p_1 + 1)]} = \frac{(R_1^2 - R_0^2)/(p_1 - p_0)}{(1 - R_1^2)/[n - (p_1 + 1)]}$$

Solution

(a)
(b)

FSDS - Chapter 7

Ex 7.4

Analogously to the previous exercise, randomly sample 30 X observations from a uniform in the interval (-4,4) and conditional on $X = x$, 30 normal observations with $E(Y) = 3.5x^3 - 20x^2 + 0.5x + 20$ and $\sigma = 30$. Fit polynomial normal GLMs of lower and higher order than that of the true relationship. Which model would you suggest? Repeat the same task for $E(Y) = 0.5x^3 - 20x^2 + 0.5x + 20$ (same σ) several times. What do you observe? Which model would you suggest now?

Solution

Ex 7.20

In the Crabs data file introduced in Section 7.4.2, the variable y indicates whether a female horseshoe crab has at least one satellite (1 = yes, 0 = no).

(a) Fit a main-effects logistic model using weight and categorical color as explanatory variables. Conduct a significance test for the color effect, and construct a 95% confidence interval for the weight effect.

(b) Fit the model that permits interaction between color as a factor and weight in their effects, showing the estimated effect of weight for each color. Test whether this model provides a significantly better fit.

(c) Use AIC to determine which models seem most sensible among the models with (i) interaction, (ii) main effects, (iii) weight as the sole predictor, (iv) color as the sole predictor, and (v) the null model.

Solution

(a) We first load the dataset and perform some light data exploration

```
# Loading the dataset
Crabs <- read.table("http://stat4ds.rwth-aachen.de/data/Crabs.dat", header=TRUE)
head(Crabs)
```

```
##   crab sat y weight width color spine
## 1    1   8 1   3.05  28.3     2     3
## 2    2   0 0   1.55  22.5     3     3
## 3    3   9 1   2.30  26.0     1     1
## 4    4   0 0   2.10  24.8     3     3
## 5    5   4 1   2.60  26.0     3     3
## 6    6   0 0   2.10  23.8     2     3
```

```
attach(Crabs)

unique(color)
```

```
## [1] 2 3 1 4
```

Given that there're 4 different colors for the `color` feature we fit the following logistic model:

$$logit(E(Y)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

X_1 refers to the `weight` while X_2, X_3 and X_4 are all dummy variables for the categorical variable `color`.

```
# Fitting a logistic model with weight and categorial color
fit <- glm(y ~  weight + factor(color), family = binomial(link = logit))
summary(fit)
```

```
##
## Call:
## glm(formula = y ~ weight + factor(color), family = binomial(link = logit))
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -3.2572     1.1985  -2.718  0.00657 **
## weight           1.6928     0.3888   4.354 1.34e-05 ***
## factor(color)2   0.1448     0.7365   0.197  0.84410
## factor(color)3  -0.1861     0.7750  -0.240  0.81019
## factor(color)4  -1.2694     0.8488  -1.495  0.13479
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 225.76  on 172  degrees of freedom
## Residual deviance: 188.54  on 168  degrees of freedom
## AIC: 198.54
##
## Number of Fisher Scoring iterations: 4
```

Using `anova()` function we test for the color effect:

```
anova(fit, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: y
##
## Terms added sequentially (first to last)
##
##
##               Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                            172     225.76
## weight         1  30.0214       171     195.74 4.273e-08 ***
## factor(color)  3   7.1949       168     188.54   0.06594 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Given that we obtained a rather low drop in the residual deviance when we added the `color` variable, and that the p.value is greater than 0.05, we can suspect that the `color` to be marginally impactful for modelling whether a female horseshoe crab has at least one satellite when paired with the main effect of `weight`.

We can compute the 95% confidence interval for the weight effect using Wald test, remembering that we must then compute the exponential for this value since the effect is measured in odds of $E(Y)$.

```
ConfInterval <- fit$coefficients[2] + c(-1,1)* summary(fit)$coefficients[2,2] * qnorm(0.975)
exp(ConfInterval)
```

```
## [1]  2.536345 11.645635
```

(b) For our second model we fit:

$$\begin{aligned} logit(E(Y)) &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_1 X_2 + \beta_6 X_1 X_3 + \beta_7 X_1 X_4 \\ &= \beta_0 + \beta_1 X_1 + X_2(\beta_2 + X_1\beta_5) + X_3(\beta_3 + X_1\beta_6) + X_4(\beta_4 + X_1\beta_4) \end{aligned}$$

```
fit2 <- glm(y ~  weight + factor(color) + weight:factor(color), family = binomial(link = logit))
summary(fit2)
```

```
##
## Call:
## glm(formula = y ~ weight + factor(color) + weight:factor(color),
##     family = binomial(link = logit))
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -1.6203     4.8909  -0.331    0.740
```

```
## weight                   1.0483     1.8929   0.554    0.580
## factor(color)2          -0.8320     5.0311  -0.165    0.869
## factor(color)3          -6.2964     5.5165  -1.141    0.254
## factor(color)4           0.4335     5.4046   0.080    0.936
## weight:factor(color)2    0.3613     1.9559   0.185    0.853
## weight:factor(color)3    2.7065     2.2284   1.215    0.225
## weight:factor(color)4   -0.8536     2.1551  -0.396    0.692
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 225.76  on 172  degrees of freedom
## Residual deviance: 181.66  on 165  degrees of freedom
## AIC: 197.66
##
## Number of Fisher Scoring iterations: 5
```

Again we can use the `anova()` function to test whether we were able to gain more information on the process

```
anova(fit2, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: y
##
## Terms added sequentially (first to last)
##
##
##                      Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                   172     225.76
## weight                1  30.0214       171     195.74 4.273e-08 ***
## factor(color)         3   7.1949       168     188.54   0.06594 .
## weight:factor(color)  3   6.8860       165     181.66   0.07562 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can again see a rather high p.value and a small drop in the residual deviance, leading us to believe that the additional variables are not needed. This result doesn't surprise us given that this model has an AIC score of 197.66 and the previous one is, for model checking purposes, identical at 198.54.

(c) Computing the AIC in the order given by the exercise we see:

```
scores <- matrix( c(fit2$aic, fit$aic, AIC(glm(y~ weight, family = binomial())),
                    AIC(glm(y~factor(color), family = binomial())), AIC(glm(y~ 1, family = binomial()))


scores
```

```
##                              AIC
## With interaction         197.6563
## With main effects        198.5423
## Weight as sole predictor 199.7371
## Color as sole predictor  220.0608
## Null model               227.7585
```

Given that for the first three models the AIC scores is similar, we can follow Occam's razor theory and choose the model with the least amount of explanatory variables between those: $logit(E(Y)) = \beta_0 + \beta_1 X_1$ with X_1

being the `weight` with an AIC of 199.7371.

Ex 7.26

A headline in The Gainesville Sun (Feb. 17, 2014) proclaimed a worrisome spike in shark attacks in the previous two years. The reported total number of shark attacks in Florida per year from 2001 to 2013 were 33, 29, 29, 12, 17, 21, 31, 28, 19, 14, 11, 26, 23. Are these counts consistent with a null Poisson model? Explain, and compare aspects of the Poisson model and negative binomial model fits.

Solution

```
SharksAttacks <- c(33, 29, 29, 12, 17, 21, 31, 28, 19, 14, 11, 26, 23)
Attacks <- data.frame(SharksAttacks)

# Null poisson model
NullPo <- glm(SharksAttacks ~ 1, data = Attacks, family = poisson)
summary(NullPo)
```

```
##
## Call:
## glm(formula = SharksAttacks ~ 1, family = poisson, data = Attacks)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.11522    0.05842   53.32   <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: 31.392  on 12  degrees of freedom
## Residual deviance: 31.392  on 12  degrees of freedom
## AIC: 97.129
##
## Number of Fisher Scoring iterations: 4
```

We now compute the mean and the variance for the `SharksAttack` data:

```
mean(SharksAttacks)
```

```
## [1] 22.53846
```

```
var(SharksAttacks)
```

```
## [1] 55.76923
```

The values $\overline{y} = 22.53846$ and $s^2 = 55.76923$ suggest overdispersion (a poisson model implies that $var(Y) = E(Y) = \mu$), therefore we can believe the null poisson model not to be adequate. Now we test a negative binomial null model

```
# Null negative binomial model
NullNegBin <- glm.nb(SharksAttacks ~ 1, link = log, data = Attacks)
summary(NullNegBin)
```

```
##
## Call:
## glm.nb(formula = SharksAttacks ~ 1, data = Attacks, link = log,
##     init.theta = 15.49441181)
##
```

```
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.11522    0.09153   34.03   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(15.4944) family taken to be 1)
##
##     Null deviance: 13.363  on 12  degrees of freedom
## Residual deviance: 13.363  on 12  degrees of freedom
## AIC: 92.608
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  15.5
##           Std. Err.:  10.5
##
##  2 x log-likelihood:  -88.608
```

The AIC for the negative binomial is only slightly better being about 5 units lower, the estimated dispersion parameter is $\frac{1}{k} = 1/15.5 \approx 0.064$ but with a standard error of 10.5 that we can impute to a very small sample size. Since the dispersion parameter is small we can conclude that the negative binomial variance $\mu + \frac{\mu^2}{k}$ is similar to the poisson variance, therefore there isn't much difference between the two models fit.

DAAG - Chapter 8

Ex 6

As in the previous exercise, the function poissonsim() allows for experimentation with Poisson regression. In particular, poissonsim() can be used to simulate Poisson responses with log-rates equal to $a + bx$, where a and b are fixed values by default.

(a) Simulate 100 Poisson responses using the model

$$log\lambda = 2 - 4x$$

for $x = 0, 0.01, 0.02, ..., 1.0$. Fit a Poisson regression model to these data, and compare the estimated coefficients with the true coefficients. How well does the estimated model predict future observations?

(b) Simulate 100 Poisson responses using the model

$$log\lambda = 2 - bx$$

where b is normally distributed with mean 4 and standard deviation 5. [Use the argument slope.sd=5 in the poissonsim() function.] How do the results using the poisson and quasipoisson families differ?

Solution

(a)
(b)