Class Activity: Hypothesis Testing and Model Selection

STAT 340: Applied Regression

Women's Labor Force Participation (1976)

The following data were drawn from the 1976 U.S. Panel Study of Income Dynamics; the response variable is married women's labor force participation.

- `lfp`: wife's labor force participation (factor, `no`, `yes`).
- `k5`: number of children ages 5 and younger (0-3, few 3s).
- `k618`: number of children ages 6 to 18 (0-8, few > 15).
- `age`: wife's age in years (30-60).
- `wc`: wife's college attendance (factor, `no`, `yes`).
- `hc`: husband's college attendance (factor, `no`, `yes`).
- `lwg`: log of wife's estimated wage rate.
- `inc`: family income excluding wife's income ($1000s).

```
library("car")
```

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

```
## Fit model with all possible explanatory variables as main effects
mroz.mod <- glm(lfp ~ k5 + k618 + age + wc + hc + lwg + inc, family=binomial(link=logit), data=Mroz)
S(mroz.mod)
```

```
## Call: glm(formula = lfp ~ k5 + k618 + age + wc + hc + lwg + inc, family =
##           binomial(link = logit), data = Mroz)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.182140   0.644375   4.938 7.88e-07 ***
## k5          -1.462913   0.197001  -7.426 1.12e-13 ***
## k618        -0.064571   0.068001  -0.950 0.342337
## age         -0.062871   0.012783  -4.918 8.73e-07 ***
## wcyes        0.807274   0.229980   3.510 0.000448 ***
## hcyes        0.111734   0.206040   0.542 0.587618
## lwg          0.604693   0.150818   4.009 6.09e-05 ***
## inc         -0.034446   0.008208  -4.196 2.71e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1029.75  on 752  degrees of freedom
## Residual deviance:  905.27  on 745  degrees of freedom
##
##  logLik      df     AIC     BIC
## -452.63       8  921.27  958.26
##
## Number of Fisher Scoring iterations: 4
```

```
##
## Exponentiated Coefficients and Confidence Bounds
##                Estimate     2.5 %     97.5 %
## (Intercept) 24.0982799 6.9377228 87.0347916
## k5           0.2315607 0.1555331  0.3370675
## k618         0.9374698 0.8200446  1.0710837
## age          0.9390650 0.9154832  0.9625829
## wcyes        2.2417880 1.4347543  3.5387571
## hcyes        1.1182149 0.7467654  1.6766380
## lwg          1.8306903 1.3689201  2.4768235
## inc          0.9661401 0.9502809  0.9814042
```

```
## Fit model that excludes k5 and k618 (update can be used to fit a nested model)
mroz.mod.2 <- update(mroz.mod, . ~ . - k5 - k618)
S(mroz.mod.2)
```

```
## Call: glm(formula = lfp ~ age + wc + hc + lwg + inc, family = binomial(link =
##           logit), data = Mroz)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.809890   0.451079   1.795  0.07258 .
## age         -0.016979   0.009689  -1.752  0.07970 .
## wcyes        0.652428   0.215562   3.027  0.00247 **
## hcyes        0.028581   0.195488   0.146  0.88376
## lwg          0.615726   0.145266   4.239 2.25e-05 ***
## inc         -0.032803   0.007639  -4.294 1.75e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1029.75  on 752  degrees of freedom
## Residual deviance:  971.75  on 747  degrees of freedom
##
##  logLik      df     AIC     BIC
## -485.88       6  983.75 1011.50
##
## Number of Fisher Scoring iterations: 4
##
## Exponentiated Coefficients and Confidence Bounds
##              Estimate     2.5 %    97.5 %
## (Intercept) 2.2476601 0.9296539 5.4579991
## age         0.9831645 0.9646121 1.0019892
## wcyes       1.9201981 1.2626460 2.9429640
## hcyes       1.0289932 0.7010201 1.5099374
## lwg         1.8510006 1.4005106 2.4791496
## inc         0.9677297 0.9529277 0.9819626
```

```
## The anova() function performs analysis of deviance,
## specifying test gives us a p-value
anova(mroz.mod.2, mroz.mod, test="Chisq")
```

```
## Analysis of Deviance Table
##
```

```
## Model 1: lfp ~ age + wc + hc + lwg + inc
## Model 2: lfp ~ k5 + k618 + age + wc + hc + lwg + inc
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       747     971.75
## 2       745     905.27  2   66.485 3.655e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Linear hypothesis - computes general Wald chi-square tests:
linearHypothesis(mroz.mod, c("k5", "k618"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## k5 = 0
## k618 = 0
##
## Model 1: restricted model
## Model 2: lfp ~ k5 + k618 + age + wc + hc + lwg + inc
##
##   Res.Df Df  Chisq Pr(>Chisq)
## 1    747
## 2    745  2 55.163  1.051e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
linearHypothesis(mroz.mod, "k5=k618")
```

```
## Linear hypothesis test
##
## Hypothesis:
## k5 - k618 = 0
##
## Model 1: restricted model
## Model 2: lfp ~ k5 + k618 + age + wc + hc + lwg + inc
##
##   Res.Df Df  Chisq Pr(>Chisq)
## 1    746
## 2    745  1 49.479  2.005e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(a) Write out the models associated with `mroz.mod` and `mroz.mod.2`, respectively.

(b) Write down the null and alternative hypotheses associated with `anova(mroz.mod.2, mroz.mod, test="Chisq")`.

(c) Test the hypothesis $H_0 : \beta_{k5} = \beta_{k618} = 0$ versus $H_A :$ at least one coefficient is not equal to 0. Interpret the result in the context of the problem.

(d) The `S(model)` function prints the summary of the model fit (`summary(model)`), and it also gives us estimate effects (on the odds scale) and the associated 95% confidence interval. Consider the estimated effect for `hc`, husband's college attendance (yes/no). Interpret the estimate and associated 95% confidence interval in the context of the problem. This interpretation follows the structure of that for a linear model.