Homework 6 with SOLUTIONS

***Directions.*** This homework pertains to materials on logistic regression in Lesson 6. The assignment should be typed, with your name on the document, and with properly labeled computer output. I suggest you attach your SAS/R input code at the end of your file clearly indicating the problem it corresponds to. If you choose to collaborate, the write-up should be your own. Please show your work! Upload the file to the HW6 Dropbox on ANGEL.

1. *9pts* Consider the saturated model for a 2 by 2 table under multinomial sampling. There are two variables, $X$ and $Y$ , each taking values 0 or 1. The saturated model has three degrees of freedom. Inside it is the independence model with its two degrees of freedom. Suppose we represent these degrees of freedom using three parameters: $\alpha$, the log odds that $Y = 1$ ignoring X,$\gamma$, the log odds that $X = 1$ ignoring $Y$ ,and $\delta$, the log odds ratio.

   (a) Explain how any independence model (a multinomial distribution $(p_{11}, p_{12}, p_{21}, p_{22})$ such that $p_{11}p_{22} = 1$) can be represented by some choice of $\alpha$ and $\gamma$, but fixing $\delta = 0$.

   **Solution:**

| | $Y=0$ | $Y=1$ | |
|---|---|---|---|
| $X=0$ | $p_{11}$ | $p_{12}$ | $p_{1+}$ |
| $X=1$ | $p_{21}$ | $p_{22}$ | $p_{2+}$ |
| | $p_{+1}$ | $p_{+2}$ | |

   From the problem we know that

$$\alpha = log(\frac{p_{+2}}{1-p_{+2}})$$
$$\gamma = log(\frac{p_{2+}}{1-p_{2+}})$$
$$\delta = log(\frac{p_{11}p_{22}}{p_{12}p_{21}}).$$

   It is easy to show that

$$p_{+1} = \frac{1}{1+e^{\alpha}}, p_{+2} = \frac{e^{\alpha}}{1+e^{\alpha}},$$
$$p_{1+} = \frac{1}{1+e^{\gamma}}, p_{2+} = \frac{e^{\gamma}}{1+e^{\gamma}}.$$

   Under the independence assumption, we have $p_{ij} = p_{i+}p_{+j}, i, j = 1, 2$. Therefore, under the independence assumption, the cell probabilities can

be represented in terms of $\alpha$ and $\gamma$ as follows

$$
\begin{aligned}
p_{11} &= \frac{1}{1+e^{\alpha}} \cdot \frac{1}{1+e^{\gamma}}, \\
p_{12} &= \frac{1}{1+e^{\gamma}} \cdot \frac{e^{\alpha}}{1+e^{\alpha}}, \\
p_{21} &= \frac{e^{\gamma}}{1+e^{\gamma}} \cdot \frac{1}{1+e^{\alpha}}, \\
p_{22} &= \frac{e^{\gamma}}{1+e^{\gamma}} \cdot \frac{e^{\alpha}}{1+e^{\alpha}}
\end{aligned}
$$

,

(b) Show that any multinomial distribution $(p_{11}, p_{12}, p_{21}, p_{22})$ can be represented by some choice of $\alpha, \gamma, \delta$ (now $\delta$ must be allowed to be nonzero).

**Solution:** Note that $p_{ij} \neq p_{i+}p_{+j}, i, j = 1, 2$. We know that

$$
\begin{aligned}
p_{11} + p_{21} = \frac{1}{1+e^{\alpha}},\ p_{12} + p_{22} &= \frac{e^{\alpha}}{1+e^{\alpha}}, \\
p_{11} + p_{12} = \frac{1}{1+e^{\gamma}}, p_{21} + p_{22} &= \frac{e^{\gamma}}{1+e^{\gamma}} \\
\delta = log(\frac{p_{11}p_{22}}{p_{12}p_{21}}), & \\
p_{11} + p_{12} + p_{21} + p_{22} = 1. &
\end{aligned}
$$

This is a system of equations with unique solutions the cell probabilities, $p_{ij}, i, j = 1, 2$. There are one-to-one functions $\psi_{ij}$ such that $p_{ij} = \psi_{ij}(\alpha, \gamma, \delta)$.

(c) Explain how to get the functional form of logistic regression for $Y$ predicted by $X$ from this set-up (hint: compute $Pr(Y = 1|X = 1)$ as a function of $\alpha, \gamma, \delta$). Which of our parameters $\alpha, \gamma, \delta$ correspond to the $\beta_0$ and $\beta_1$ in binary logistic regression with a single categorical predictor?

**Solution:** The logistic regression model is

$$logit(P(Y = 1|X = i)) = \beta_0 + \beta_1 i.$$

where $i = 0, 1$. If $X = 0$, then logit$(P(Y = 1|X = 0)) = \beta_0$. If $X = 1$, then logit$(P(Y = 1|X = 1)) = \beta_0 + \beta_1$. Hence, $\beta_1 = logit(P(Y = 1|X = 1)) - logit(P(Y = 1|X = 0))$. Now,

$$\beta_0 = logit(\frac{P(Y = 1|X = 0)}{P(Y = 0|X = 0)}) = log(p_{12}) - log(p_{11})$$

Furthermore, one can easily show that

$$\beta_1 = log(\frac{p_{22}p_{11}}{p_{12}p_{21}}) = \delta$$

Under the assumption that X and Y are independent $P(Y = 1|X = i) = P(Y = 1)$ for all $i = 0, 1$. Hence, $\alpha = \beta_0$.

2. *9pts* Let $Y_1, Y_2, ..., Y_N$ be independent random variables. For each of the following models, state whether it is or is not a generalized linear model. If it is a GLM, then what is the link function and the systematic component. If it is not a GLM, explain why is it not a GLM. The basic intro to GLMs was given in Lesson 5.

   (a) $Y_i = \alpha + \log(\beta_1 + \beta_2 x_i) + e_i$ where $e_i \sim N(0, \sigma^2)$

   **Solution:** Assume all the greek letters represent parameters.

   No. There is no way to separate $\beta_1$ and $\beta_2$ in here.

   (b) $Y_i = \exp\{\alpha + \log(\beta_1)x_{1i} + \beta_2 \log(x_{2i})\}$ where $Y_i$ is a Poisson random variable.

   **Solution:** Assume all the greek letters represent parameters.

   Yes.

   The random component is the Poisson distribution which is an exponential family.
   The link function $g(\mu) = log(\mu)$ where $\mu = E(Y)$
   The linear predictor: $\alpha + \gamma x_{1i} + \beta_2 log(x_{2i})$ where $\gamma = log(\beta_1)$

   (c) $\pi_i = \frac{\exp\{\alpha+\beta e^{x_i}\}}{1+\exp\{\alpha+\beta e^{x_i}\}}$ where $Y_i$ is binomial with probability $\pi_i$ and $n$ trials.

   **Solution:** Assume all the greek letters represent parameters.

   Yes this is the logistic regression.

   The random component is the binomial distribution which belongs to the exponential family.
   The link function is the logit function i.e logit($\pi_i$)
   The linear predictor $\alpha + \beta e^{x_i}$

3. *25pts* Children of ages 0-15 were classified according to whether they carried Streptococcus pyogenes and according to the size of their tonsils; see homework 4.

   Let $\pi_i$ be the probability of a child being a *carrier* of *Streptococcus* and $X_i$ *tonsil size*.

| | Tonsil size | | |
|---|---|---|---|
| | Normal | Slightly enlarged | Very enlarged |
| Carrier | 19 | 29 | 24 |
| Non-carrier | 497 | 560 | 269 |

(a) Fit the logistic regression model. Write the model in terms of $\pi_i$, $\beta$'s, and $X'$s. Make sure you are modeling the "event" of "being a carrier". How many dummy variables do you have? How many parameters does your saturated model have? What did you choose as you baseline(reference) level for the explanatory variable.

**Solution:** The logistic regression model, based on the SAS/R output below, is as follows:

$$\pi = \exp(B_0 + B_1X_1 + B_2X_2)/[1 + \exp(B_0 + B_1X_1 + B_2X_2)],$$

or equivalently,

$$logit(\pi) = B_0 + B_1X_1 + B_2X_2.$$

There are 2 dummy variables, $X_1$ and $X_2$. $X_1 = \{1$ if tonsils=slightly enlarged, 0 otherwise$\}$ and $X_2 = \{1$ if tonsils=very enlarged, 0 otherwise$\}$. If the tonsils are normal, $X_1$ and $X_2$ are 0. The model has 3 parameters, $B_0$, $B_1$, and $B_2$.
The baseline reference is children with normal size tonsils.

(b) Report the estimated model, and more specifically report the estimated coefficients and standard errors. Interpret $\exp(\beta_1)$ and report 95% confidence interval for $\exp(\beta_1)$. How about $\beta_2$?

**Solution:**
The model is as given in (a). The parameters and their coefficients are as follows:
$B_0 = -3.2642$ with (se=0.2338)
$B_1 = 0.3035$ with (se=0.3015)
$B_2 = 0.8475$ with (se=0.3163)

$\exp(B_n)$ is the odds ratio of the $B_n$-th term and is given in the SAS output.

The odds ratio for $B_1$ is 1.355 with a 95%CI of (0.750, 2.446). Because the interval includes 1.000, we do not have sufficient evidence to conclude (at the 95% confidence level) that children with slightly enlarged tonsils can predict if the child is a carrier of the pyogenes compared to children with normal sized tonsils.

The odds ratio for B2 is 2.334 with a 95%CI of (1.256,4.338). The confidence interval does not include 1.000, and we can state that children with very enlarged tonsils are about 2.5 times more likely to be a carrier of the disease than those with only normal sized tonsils.

(c) Does the model appear to fit the data? Perform an overall goodness-of-fit test and comment on the results. Is this test trustworthy?

**Solution:** $H_0 : B_1 = B_2 = 0$ vs $H_a$: Either $B_1$ or $B_2$ is non-zero. Null hypothesis is to assume that the tonsil size is not predictive to whether a person is a carrier or not. The model fits the data perfectly. Because this model is saturated, the goodness of fit statistics $X^2$ and $G^2$ from this model are both zero. If you look at the Likelihood Ratio, Score, and Wald tests, all 3 have a p-value of less than .05, thus indicating a good fit. Given that these tests agree, the model seems to be trustworthy.

(d) How do results from this model compare to your conclusions from homework 4?

**Solution:** The results agree because tonsil size and carrier were associated with each other, and a person with the largest tonsil size is more likely of being a carrier.

(e) What is the predicted probability of a child being a carrier if she has "very enlarged" tonsils?

**Solution:**

P(A child being a carrier $\mid$ having very enlarged tonsils) = $\frac{\exp(-3.2642+0.8475)}{1+\exp(-3.2642+0.8475)} = 0.081908$.

**R Code/Output:**

```
>> T=factor(c("normal","Slightly enlarged", "Very enlarged"))
> carrier=c(19,29,24)
> noncarrier=c(497,560,269)
> count=cbind(carrier,noncarrier)
> count
     carrier noncarrier
[1,]      19        497
[2,]      29        560
[3,]      24        269
> Tsenlarged=(T=="Slightly enlarged")
> Tveryenlarged=(T=="Very enlarged")
> tonsil.fit=glm(count~Tsenlarged+Tveryenlarged,
family=binomial('logit'))
> summary(tonsil.fit)
```

```
Call:
glm(formula = count ~ Tsenlarged + Tveryenlarged, family
= binomial("logit"))

Deviance Residuals:
[1]  0  0  0

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -3.2642     0.2338 -13.964  < 2e-16 ***
TsenlargedTRUE      0.3035     0.3015   1.007  0.31412
TveryenlargedTRUE   0.8475     0.3163   2.680  0.00737 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7.3209e+00  on 2  degrees of freedom
Residual deviance: 8.3267e-14  on 0  degrees of freedom
AIC: 20.851

Number of Fisher Scoring iterations: 3

>
> X2 = sum(residuals(tonsil.fit,type="pearson")^2)
> X2
[1] 3.095579e-28
> 1-pchisq(X2,tonsil.fit$df.residual)
[1] 0
>
```

**SAS code/Output:**

```
OPTIONS NODATE NONUMBER CENTER;

data prob1;
input size $ y n;
cards;
normal 19 516
slightly 29 589
very_enl 24 293
;
run;
```

```
proc logistic data=prob1 descending;
class size (ref=first) / param=ref;
model y/n = size /scale=none lackfit;
output out=predict pred=prob;
run;
```

| Analysis of Maximum Likelihood Estimates | | | | | | |
|---|---|---|---|---|---|---|
| Parameter | | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
| Intercept | | 1 | -3.2642 | 0.2338 | 194.9848 | <.0001 |
| s | slightly | 1 | 0.3035 | 0.3015 | 1.0133 | 0.3141 |
| s | very_enl | 1 | 0.8012 | 0.3191 | 6.3047 | 0.0120 |

| Model Fit Statistics | | | |
|---|---|---|---|
| | | Intercept and Covariates | |
| Criterion | Intercept Only | Log Likelihood | Full Log Likelihood |
| AIC | 563.508 | 561.149 | 20.813 |
| SC | 568.751 | 576.878 | 36.541 |
| -2 Log L | 561.508 | 555.149 | 14.813 |

4. *20pts* Moritz and Satariano (*J. Clin. Epidemiology,* **46***: 443-454, 1993*) used logistic regression to predict whether the stage of breast cancer as diagnosis was advanced or local for a sample of 444 middle-aged and elderly women. A table referring to a particular set of demographic factors reported the estimated odds ratio for the effect of living arrangement (three categories) as 2.02 for spouse vs. alone and 1.71 for others vs. alone; it reported the effect of income (three categories) as 0.72 for \$10,000-24,999 vs. <\$10,000 and 0.41 for \$25,000+ vs <\$10,000. Estimate the odds ratios for the third pair of categories for each factor.

   **Solution:** Let $X$ denote the predictor of living arrangement with three categories (alone,spouse, others) and $Z$ the predictor of income with three categories $(< 10,000, 10,000-24,999, 25,000+)$. Let $\pi_{ik}$ the probability the the

cancer is advanced. The logistic regression model with alone and $< 10,000$ as the baseline levels is:

$$log(\frac{\pi_{ik}}{1-\pi_{ik}}) = \beta_0 + \beta_i^X + \beta_k^Z \quad i,k = 1,2,3 \tag{1}$$

$\beta_1^X = \beta_1^Z = 0$

Thus, for any income level $exp(\beta_i^X)$ is the odds ratio of breast cancer between the category i of X and the baseline (alone) and $exp(\beta_a^X - \beta_b^X)$ is the odds ratio of breast cancer between the a and b categories of X. Therefore, we have that $exp(\beta_2^X) = 2.02$ and $exp(\beta_3^X) = 1.71$. This implies that $\beta_2^X = 0.703$ and $\beta_3^X = 0.536$ . Hence the odds ratio of breast cancer for spouse vs others is $exp(\beta_2^X - \beta_2^X) = exp(0.703 - 0.536) = exp(0.167) = 1.182$

Similarly, for any living arrangement $exp(\beta_2^Z) = 0.72$ and $exp(\beta_3^Z) = -0.89$. Thus we have that $\beta_2^Z = -0.329$ and $\beta_3^Z = -0.89$. Thus the odds ratio of breast cancer of $10,000 - 24,999$ vs $25,000+$ is $exp(\beta_2^Z - \beta_3^Z) = exp(-0.329 + 0.89) = 1.75$

**Alternatively** The response variable is Breast Cancer Stage {local, advanced}

The explanatory variables are:
Living arrangement {alone, spouse, others}
Income {<10K, 10K-25K, >25K}

Given:
(odds of advanced—others)/(odds of advanced—alone)= 1.71
(odds of advanced—spouse)/(odds of advanced—alone)=2.02

We are to find the odds estimate of Living arrangement (Others vs Spouse), which is
(odds of advanced—others)/(odds of advanced—spouse)=1.71/2.02=0.85

Given:
(odds of advanced |10-25)/(odds of advanced | <10k)=0.72
(odds of advanced | >25k)/(odds of advanced | <10k)=0.41

We are to find the odds estimate of income (>25k vs 10-25k), which is
(odds of advanced | >25k)/(odds of advanced |10-25k)=0.41/0.72=0.57

5. *25pts* Montana Economic Outlook. Use the files from homework 5. This 1992 Montana poll asked a random sample of Montana residents whether their personal financial status

was the worse, the same, or better than a year ago, and whether they thought the state economic outlook was better over the next year.

AGE = 1 under 35, 2 35-54, 3 55 and over
SEX = 0 male, 1 female
INC = yearly income: 1 under $20K, 2 20-35$K, 3 over $35K
POL = 1 Democrat, 2 Independent, 3 Republican
AREA = 1 Western, 2 Northeastern, 3 Southeastern Montana
FIN = Financial status 1 worse, 2 same, 3 better than a year ago
STAT = State economic outlook 0 better, 1 not better than a year ago

(a) Model the feelings on the economic outlook as a function of gender. Give you logistic regression model. Report the estimated coefficients and the standard errors. Does the model appear to fit the data well? Who is more likely to have a positive outlook, men or women?

**Solution:**

The logistic regression model is:

$$\pi = \exp(B_0 + B_1 Sex)/[1 + \exp(B_0 + B_1 Sex)],$$

where Sex=$\{$1 if female, 0 if male$\}$, therefore Male is the reference class. Or equivalently, $logit(\pi) = B_0 + B_1 Sex$.

The coefficients, along with the standard errors are:

$B_0 = 0.3469 (se = 0.2040)$
$B_1 = 0.6563 (se = 0.3221)$

Odds ratios of female v.s. male for having a positive outlook is 1.928, the confidence interval for odds ratio is (1.025, 3.624). P-value for sex is 0.0416, so this covariate is significant. Therefore females are significantly more likely than men to have a positive outlook on the economy.
If we compare the current model with the intercept-only model, $\Delta$G2 =4.2 with df=1 and p-value of 0.0391, so we can see that this model is a better fit. However, more covariates might be needed.

For the overall goodness-of-fit test, we need to compare the current model with the saturated model, which is measured by **Residual Deviance in R** and $-2logL$ **in SAS**. From the R output below we see that the Residual Deviance $\Delta$G2 = 229.7 with df = 179. The corresponding p-value is $0.006278678$. Thus we reject the null hypothesis that the current model fits the data well. We need a more complicated model to fit the data.

**R Code:**

```
> montana=read.csv('montana.csv',head=T)
> data.clean <-data.frame(montana[montana$Stat!=".",])
> result=glm(Stat~Sex,family=binomial("logit"),
data=data.clean)
> result

Call:  glm(formula = Stat ~ Sex, family = binomial("logit"),
data = data.clean)

Coefficients:
(Intercept)          Sex
    -0.3469      -0.6564

Degrees of Freedom: 180 Total (i.e. Null);  179 Residual
Null Deviance:     233.9
Residual Deviance: 229.7  AIC: 233.7
>
```

**SAS Code/Output:**

```
PROC LOGISTIC data = montana;class sex (ref=first) /
param=ref;model stat = sex / scale=none aggregate;
output out=predict pred=phat reschi=pearson
resdev=deviance;run;
```

| Analysis of Maximum Likelihood Estimates | | | | | |
|---|---|---|---|---|---|
| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
| Intercept | 1 | 0.3469 | 0.2040 | 2.8899 | 0.0891 |
| sex | 1 | 0.6563 | 0.3221 | 4.1517 | 0.0416 |

| Odds Ratio Estimates | | | |
|---|---|---|---|
| Effect | Point Estimate | 95% Wald Confidence Limits | |
| sex | 1.928 | 1.025 | 3.624 |

| Model Fit Statistics | | |
|---|---|---|
| Criterion | Intercept Only | Intercept and Covariates |
| AIC | 235.939 | 233.684 |
| SC | 239.138 | 240.081 |
| -2 Log L | 233.939 | 229.684 |

(b) Fit the full logistic regression model (i.e., STAT is a response and ALL other variables as predictors). Does your model fit the data well? Report the appropriate goodness-of-fit statistics, and (if relevant) the residuals and influence points.

**Solution:** In this part all parameters (but not their interactions) are used to model economic outlook. From the output below, all covariates except Sex are not important. Therefore, we may use a simpler model without all those variables.

The residuals vs. linear predicted plot followed also shows certain lack-of-fit for the model. There are a few points with absolute values of residuals greater than 2. Deviance is 152.57 with df=122, which leads to p-value=0.03. This indicates the lack-of-fit of the model. Pearson goodness-of-fit is 131.16 with df=122, which leads to p-value=0.27. The two statistics do not agree with each other.

**R code/Output**

```
> result=glm(Stat~Age+Sex+Inc+Pol+Area+Fin,
family=binomial("logit"),data=montana)
Call:  glm(formula = Stat ~ Age + Sex + Inc + Pol
+ Area + Fin, family = binomial("logit"), data = montana)

Coefficients:
(Intercept)      Age1         Age2         Age3          Sex
    25.8234   -11.5286     -11.8782     -10.9570      -0.1625
 Inc1         Inc2         Inc3         Pol1
-1.1061       1.7615       1.3365       2.1240
  Pol2         Pol3         Area         Fin1         Fin2
  0.9502      -0.2151      -0.3188     -10.4314     -13.1570

Degrees of Freedom: 208 Total (i.e. Null);  194 Residual
Null Deviance:     164.6
Residual Deviance: 130.6  AIC: 160.6
```

**SAS Code/Output:**

```
PROC LOGISTIC data = montana;class age sex inc pol area
fin(ref=first)/param=ref;model stat = age sex inc pol area fin/
scale=none aggregate;output out=predict pred=phat
reschi=pearson resdev=deviance;run;
```

| Parameter | | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|---|
| Intercept | | 1 | -0.3646 | 0.6728 | 0.2937 | 0.5879 |
| AGE | 35-54 | 1 | 0.3645 | 0.4900 | 0.5534 | 0.4569 |
| AGE | <35 | 1 | -0.7197 | 0.4720 | 2.3245 | 0.1274 |
| SEX | FEMALE | 1 | 0.8169 | 0.3957 | 4.2613 | 0.0390 |
| INC | 20-35$K | 1 | -0.1761 | 0.4459 | 0.1560 | 0.6929 |
| INC | < $20K | 1 | -0.0986 | 0.5273 | 0.0350 | 0.8516 |
| POL | DEMOCRAT | 1 | 0.5093 | 0.4251 | 1.4358 | 0.2308 |
| POL | INDEPENDENT | 1 | 0.8580 | 0.5383 | 2.5399 | 0.1110 |
| AREA | NORTHEASTERN | 1 | 0.8681 | 0.4928 | 3.1027 | 0.0782 |
| AREA | SOUTHWESTERN | 1 | -0.4831 | 0.4294 | 1.2658 | 0.2606 |
| FIN | SAME | 1 | 0.4702 | 0.4675 | 1.0117 | 0.3145 |
| FIN | WORSE | 1 | 0.9880 | 0.4649 | 4.5161 | 0.0336 |

Deviance and Pearson Goodness-of-Fit Statistics

| Criterion | Value | DF | Value/DF | Pr > ChiSq |
|---|---|---|---|---|
| Deviance | 152.5753 | 122 | 1.2506 | 0.0317 |
| Pearson | 131.1609 | 122 | 1.0751 | 0.2693 |

(c) Now select Backward Elimination as a model selection algorithm under MODEL options in SAS (`selection=backward`, see also Water Level Study Handouts); in R use *step* function or consider alternatives such as *stepAIC* in package MASS. You should be able to choose backward elimination in other software as well. How does the model in step 1 differ from your model in step 0? Give the final model you get from your procedure. What can you say about its fit?

**Solution:**

In Step 1, the variable INC is removed because it is not a significant variable when adjusted for the other variable in the model. The final model also removes the variable POL, leaving us with AGE SEX AREA FIN. From step 0 to step 1, the G2 changes from 182.394 to 182.551. The final model has a G2 = 185.426 with df=163-7=156. It fits the data well. Again, from output, you should use the measurement $-2logL$ as the G2 statistic of the overall goodness-of-fit.

**R code/Output**

```
>
> step.montana.fit = step(result,direction="backward",trace=F)
> step.montana.fit$anova
    Step Df  Deviance Resid. Df Resid. Dev      AIC
1         NA        NA       194   130.5947 160.5947
2  - Age   3 2.0014246       197   132.5962 156.5962
3  - Pol   3 2.0897749       200   134.6859 152.6859
4 - Area   1 0.9168689       201   135.6028 151.6028
> summary(step.montana.fit)

Call:
glm(formula = Stat ~ Sex + Inc + Fin, family = binomial("logit"),
    data = montana)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2992   0.1401   0.3895   0.6074   1.2759
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   13.8027  1455.3977   0.009   0.9924
Sex           -0.9726     0.4838  -2.010   0.0444 *
Inc1           1.7351     0.7415   2.340   0.0193 *
Inc2           1.2742     0.6295   2.024   0.0430 *
Inc3           1.7634     0.7173   2.458   0.0140 *
Fin1         -10.4580  1455.3980  -0.007   0.9943
Fin2         -12.9968  1455.3976  -0.009   0.9929
Fin3         -13.0585  1455.3977  -0.009   0.9928
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 164.64  on 208  degrees of freedom
Residual deviance: 135.60  on 201  degrees of freedom
AIC: 151.6

Number of Fisher Scoring iterations: 14
```

**SAS Code/Output:**

```
PROC LOGISTIC data = montana;class age sex inc pol area
 fin (ref=first) / param=ref;model stat = age sex inc
 pol area fin
/selection=backward lackfit;output out=predict pred=phat
reschi=pearson resdev=deviance;run;

                   Value         DF     p-value
Deviance            185.4262     155    0.0481
Pearson             160.2650     155    0.3694
```

Summary of Backward Elimination

| Step | Effect Removed | DF | Number In | Wald Chi-Square | Pr > ChiSq | Variable Label |
|---|---|---|---|---|---|---|
| 1 | INC | 2 | 5 | 0.1563 | 0.9248 | YEARLY INCOME |
| 2 | POL | 2 | 4 | 2.8270 | 0.2433 | PARTY AFFILIATION |

Type 3 Analysis of Effects

| Effect | DF | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|
| AGE | 2 | 6.7364 | 0.0345 |
| SEX | 1 | 5.0584 | 0.0245 |
| AREA | 2 | 7.7126 | 0.0211 |
| FIN | 2 | 6.7997 | 0.0334 |

Analysis of Maximum Likelihood Estimates

| Parameter | | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|---|
| Intercept | | 1 | -1.8351 | 0.5728 | 10.2640 | 0.0014 |
| AGE | <35 | 1 | 1.1684 | 0.4560 | 6.5654 | 0.0104 |
| AGE | >55 | 1 | 0.4463 | 0.4658 | 0.9180 | 0.3380 |
| SEX | MALE | 1 | 0.8568 | 0.3810 | 5.0584 | 0.0245 |
| AREA | SOUTHWESTERN | 1 | 1.2847 | 0.4628 | 7.7058 | 0.0055 |
| AREA | WESTERN | 1 | 0.7360 | 0.4678 | 2.4747 | 0.1157 |
| FIN | SAME | 1 | -0.7105 | 0.4387 | 2.6235 | 0.1053 |

6. *12 pts* Consider a $2 \times 3 \times 3$ table representing variables $A, B, C$, where we use logistic regression to estimate the probability of success for the first variable ($A$).

   (a) What are the degrees of freedom for the full (saturated) model? Write down the logistic regression equation.

   **Solution:** Let $B_1$ and $B_2$ the indicator variables for B and $C_1$, $C_2$ the indicator variables for C. The saturated model is

   $$\pi = \beta_0 + \beta_1 B_1 + \beta_2 B_2 + \gamma_1 C_1 + \gamma_2 C_2 + (\beta\gamma)_{11} B_1 C_1 + (\beta\gamma)_{12} B_1 C_2 + (\beta\gamma)_{21} B_2 C_1 + (\beta\gamma)_{22} B_2 C_2$$

   where $\pi$ is the probability of success for the response $A$. The degrees of freedom for the full model is 0.

   (b) Write the logistic regression equation for the model predicting $A$ from $B$ alone. How many degrees of freedom will the goodness of fit test have?

   **Solution:** The logistic model is

   $$\pi = \beta_0 + \beta_1 B_1 + \beta_2 B_2$$

The model will have 6 degrees of freedom.

(c) Write the logistic regression equation for the model predicting $A$ from $B$ and $C$ with a no-three-way-interaction model (i.e., no interaction term). How many degrees of freedom will the goodness of fit test have?

**Solution:** The logistic model is

$$\pi = \beta_0 + \beta_1 B_1 + \beta_2 B_2 + \gamma_1 C_1 + \gamma_2 C_2$$

The model will have 4 degrees of freedom.