STA841 HW4

Dipesh Gautam

November 5, 2015

1 Problem 1

In this problem we're looking at the spatial correlation of cancer rates using the data from $66x66$ pixel map of Eastern US.

1.1

We've plotted the incidence map of the cancer rates in the Eastern US in Fig. 1

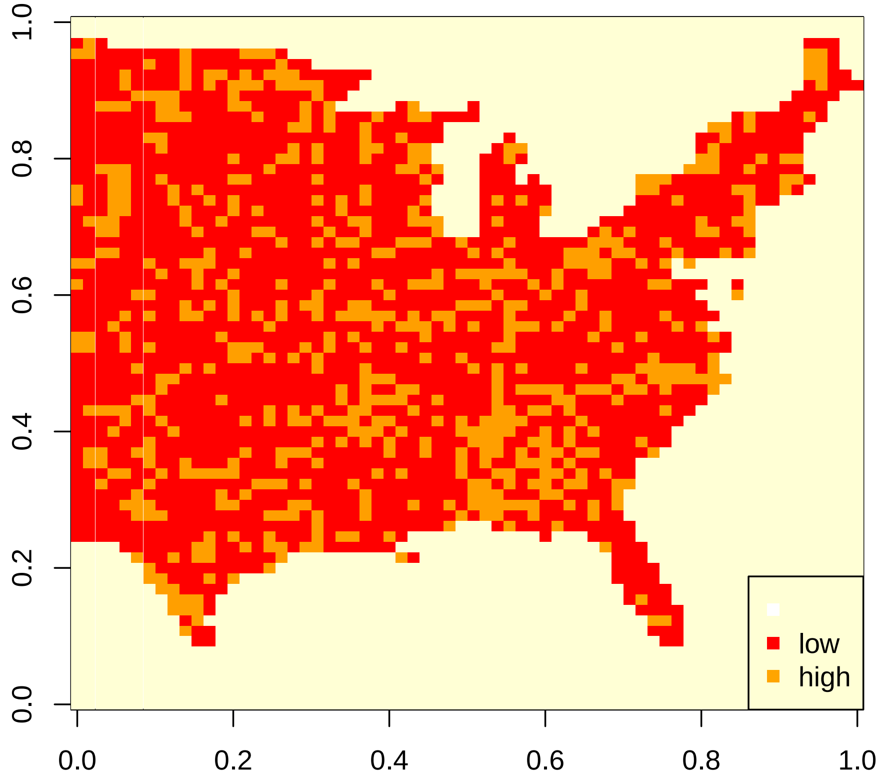

Figure 1: Incidence map of cancer rates in the Eastern US

1.2

To check for spatial correlation in cancer rates, we fit the following autologistic model:

$$logit(P(Y)) = \alpha + \beta(Y_E + Y_S + Y_W + Y_N)$$

We notice that all the pixels do not have all four neighbors. To easily handle this missing data, we removed all the pixels without all four neighbors from analysis. We lost only about 10% of the data in doing so. Since we're losing just 10% from about 2300 data points, the drop is justified.
The summary of the model is shown below.

```
Call:
glm(formula = cbind(y, 1 - y) ~ y1, family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2633  -0.7690  -0.6378   1.2732   1.8399

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.48913    0.08264 -18.020  < 2e-16 ***
y1           0.42221    0.05215   8.096 5.67e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2334.9  on 2002  degrees of freedom
Residual deviance: 2268.2  on 2001  degrees of freedom
AIC: 2272.2

Number of Fisher Scoring iterations: 4
```

We have Residual deviance of 2268.22 on 2001 degrees of freedom. The ratio is fairly close to 1, showing no overdispersion.
The coefficient, $\beta = 0.422$. The value is statistically significant. So, we can conclude that there is evicence for spatial correlation. Given the value of β, we can interpret that having 1 extra neighbor with high cancer rate, increases the odds of a place having high cancer rate by 0.525.

1.3

We now look at permutation test to test the null hypohtesis that $H_0 : \beta = 0$, i.e. the 619 pixels with high cancer rates do not cluster. For this, we picked 619 random pixels out of 2293 that corresponding to Eastern US and assigned them high cancer rate and fit the same model as in part 1.1. The total number of ways to permute is out of the scope of the computer. So, we ran just 5000 random samples and approximated the distribution of β using those samples. The histogram of the sampled β is shown in Fig. 2.

Also, the true β is not included in the 99% confidence interval from permutation test. This again supports our result from the previous part that there is spatial correlation in cancer rates.

```
      0.5%      99.5%
-0.1636156  0.1592380
```

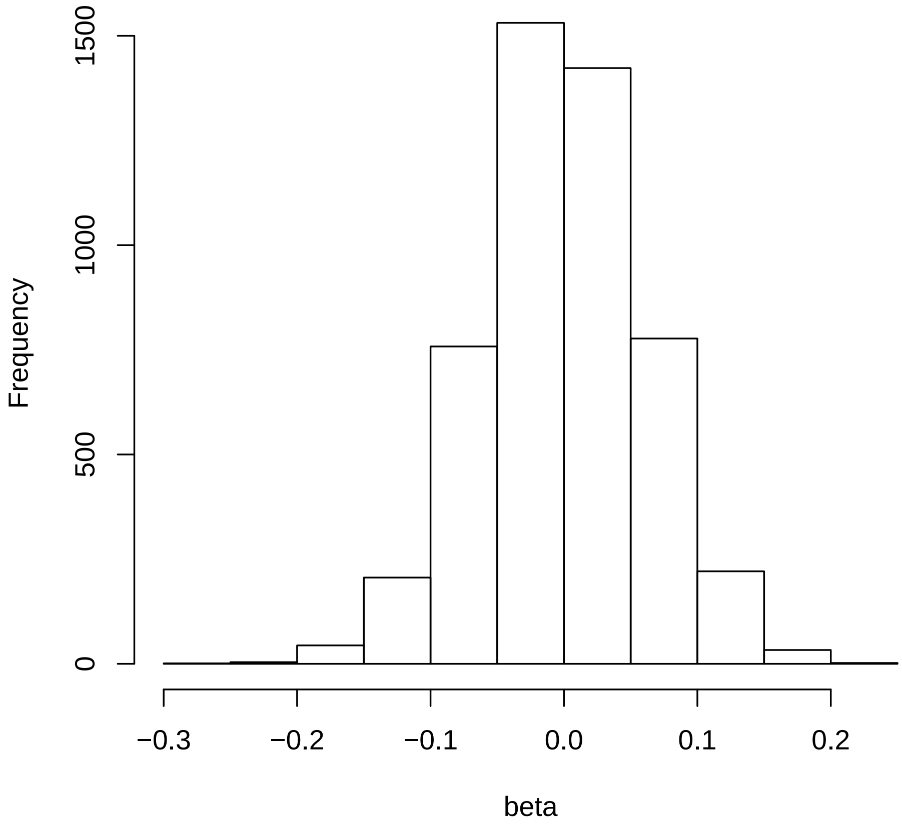

Figure 2: Distribution of β from permuted samples

1.4

We want to test the effect of wind blowing from West to East, which might lead to pixels being more similar to neighbors on the East and West than to the neighbors on the North and South. To test this we fit an autologistic model specified as below:

$$logit(P(Y)) = \alpha + \beta_1(Y_E + Y_S + Y_W + Y_N) + \beta_2(Y_E + Y_W)$$

The summary of this model fit is given below.

```
Call:
glm(formula = cbind(y, 1 - y) ~ y1 + y2, family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2632  -0.7720  -0.6379   1.2694   1.8398

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.48900    0.08264 -18.018  < 2e-16 ***
y1           0.41305    0.07779   5.310  1.1e-07 ***
y2           0.01807    0.11390   0.159    0.874
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2334.9  on 2002  degrees of freedom
Residual deviance: 2268.2  on 2000  degrees of freedom
AIC: 2274.2

Number of Fisher Scoring iterations: 4
```

We see that there is insignificant gain from just the east and west neighbors compared to all the neighbors so, we can say that the effect of wind blowing from west to east in the cancer rates is negligible.

2 Problem 2

In this problem, we're looking at the relation between parental socioeconomic status and the mental health of children. Given the data where response is ordered factor, we fit the proportional-odds logistic regression model given by:

$$logitP(Y <= j|x) = \alpha_j + SES'\beta$$

In this model, we have different intercepts for each mental health status, but β is shared across all. Even though Socioeconomic status of parents is an ordered factor, for simplicity in interpretation, it is taken to be unordered factor but the effect should be apparent in inference.
In R, we fit the model as:
polr(status $\sim$ ses, weights = count, data = data2)

This actually fits the model:

$$logitP(Y <= j|x) = \alpha_j - SES'\beta$$

So, the coefficients will have opposite sign than expected. The summary is shown below.

```
Call:
polr(formula = status ~ ses, data = data2, weights = count)

Coefficients:
        Value Std. Error t value
sesB -0.02084     0.1596 -0.1306
sesC -0.20394     0.1540 -1.3244
sesD -0.35888     0.1452 -2.4720
sesE -0.61709     0.1582 -3.9013
sesF -0.72095     0.1669 -4.3197

Intercepts:
               Value    Std. Error t value
impaired|mild  -1.5127   0.1198   -12.6314
mild|moderate   0.0853   0.1136     0.7512
moderate|well   1.1874   0.1186    10.0144

Residual Deviance: 4455.018
AIC: 4471.018
```

In this model, our baseline is SES A, i.e., high socioeconomic status of parents. We can see by looking at different α_j that for the baseline group, there's much higher odds of falling into "well" mental health status. If we look at the different β values, we can see that lower the socio economic group, higher the odds of having worse mental health status.

3 Problem 3

Moved to next homework as per instruction.

4 Problem 4

This problem looks at the effect of abode type, education level and years since first marriage on the number of children a woman of Indian race in Fiji has.

4.1

An appropriate model to fit in this case seems to be log-linear model with poisson family (log link). We're looking at count data if we look at total children born to each group of women. So, poission family of log-linear model seems appropriate. For simplicity in interpretability, we assumed unordered factors for both marriage and education even though they should be ordered factors. Also, we rounded the total number of children to the nearest whole number as the rounding to two decimal places in the averages caused some totals to be decimals. The model we fit is:
log E(total number of children) = log(total women) + marriage + abode + education,
where log(total women) is the offset.
In R, the code is
glm(count∼marriage+edu+abode, family=poisson(link=log), offset = log(tot), subset=(tot>0), data = data4)
The summary is shown below. There seems to be slight underdispersion in this model. We also estimated

σ^2 and plugged that into estimating equation to check for model adequacy.

```
Call:
glm(formula = count ˜ marriage + edu + abode, family = poisson(link = log),
    data = data4, subset = (tot > 0), offset = log(tot))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.01562  -0.48720  -0.07375   0.55851   1.28367

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.08410    0.05876  -1.431   0.1524
marriage2    0.99715    0.05869  16.990   <2e-16 ***
marriage3    1.41220    0.05648  25.003   <2e-16 ***
marriage4    1.66473    0.05628  29.580   <2e-16 ***
marriage5    1.84435    0.05667  32.546   <2e-16 ***
marriage6    2.03017    0.05546  36.606   <2e-16 ***
edu2         0.03750    0.02450   1.531   0.1259
edu3        -0.04710    0.03377  -1.395   0.1631
edu4        -0.21093    0.06147  -3.431   0.0006 ***
abode2       0.06107    0.02484   2.458   0.0140 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 3035.429  on 45  degrees of freedom
Residual deviance:   29.831  on 36  degrees of freedom
AIC: 346.3

Number of Fisher Scoring iterations: 4

Call:
glm(formula = count ˜ marriage + edu + abode, family = poisson(link = log),
    data = data4, subset = (tot > 0), offset = log(tot))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.01562  -0.48720  -0.07375   0.55851   1.28367

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.084099   0.003127  -26.89   <2e-16 ***
marriage2    0.997150   0.003124  319.20   <2e-16 ***
marriage3    1.412196   0.003006  469.77   <2e-16 ***
marriage4    1.664732   0.002995  555.75   <2e-16 ***
marriage5    1.844349   0.003016  611.49   <2e-16 ***
marriage6    2.030168   0.002952  687.77   <2e-16 ***
edu2         0.037497   0.001304   28.76   <2e-16 ***
edu3        -0.047102   0.001798  -26.20   <2e-16 ***
edu4        -0.210932   0.003272  -64.47   <2e-16 ***
abode2       0.061067   0.001322   46.18   <2e-16 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 0.002832845)

    Null deviance: 3035.429  on 45  degrees of freedom
Residual deviance:   29.831  on 36  degrees of freedom
AIC: 346.3

Number of Fisher Scoring iterations: 4
```

4.2

We're assuming all three covariates(abode, edu, marriage) to be unordered factors in the model. So, for any coefficient β, it implies that being in the certain category leads to increase in mean by e^{β} times or in our problem, mean number of children is e^{β} times more than that from baseline for each parameter. The baseline is urban woman(abode=1) with no education(edu=1) and less than 5 years of marriage(marriage=1). So, we have coefficients for abode2, marriage2-6 and educ2-4 and the intercept. β_{abode2} tells extra mean number of children rural women have compared to urban women. With $\beta_{abode2} = 0.061$, we have that rural women have 1.06 times more children than urban woman with everything else remaining the same. Similarly $exp(\beta_{marriage5})$ gives the times more children the women with 20-24 years since first marriage are likely to have compared to women with less than 5 years of marriage and so on.
Time since first marriage seems to be the major factor affecting the number of children. We do see some variation due to education level and abode type with more educated women tending to have fewer children and rural women having more children than urban women. But most of the difference seem to be because of the length of the marriage, which is to be expected.

4.3

In the model, urban woman(abode=1) with no education(edu=1) and less than 5 years of marriage(marriage=1) is the baseline. We're looking at mean number of children born to urban woman(baseline) with upper elementary education(edu=3) after ten years of marriage(marriage=3). Here we assumed that after ten years of marriage meant that the women were in group 3(10-14). To get the mean, we calculated the mean of the linear combination of random variables $\alpha + \beta_{educ3} + \beta_{marriage3}$ and took the exponential of the mean.

$$\mu = exp(\alpha + \beta_{educ3} + \beta_{marriage3})$$

For the 95% confidence interval for the mean, we calculated the variance of the linear combination of the random variable and sampled 5000 from the univariate normal distribution and calculated the inner 95% quantile of the samples.
Mean and 95% confidence interval for the mean are shown below in the results.

```
> varcovar=vcov(fit4)
> cov= varcovar[c("(Intercept)","edu3", "marriage3"),
+           c("(Intercept)","edu3", "marriage3")]
> var = cov[1,1]+cov[2,2]+cov[3,3]+2*cov[2,1]+ 2*cov[3,1]+2*cov[3,2]
> mu = fit4$coefficients["(Intercept)"] + fit4$coefficients["edu3"]+
+       fit4$coefficients["marriage3"]
> samples = exp(rnorm(5000, mu, sqrt(var)))
> mean = exp(mu)
> mean

(Intercept)
   3.600221

> quantile(samples, c(.025,.975))
```

```
    2.5%    97.5%
3.331905 3.878460
```

4.4

This problem is similar to the one in previous part with only changes being that now we're looking at lifetime(marriage=6) mean number of children born to rural women(abode=2) with secondary education(edu=4). The analysis and procedure is the same as above with mean given by:

$$\mu = exp(\alpha + \beta_{abode2} + \beta_{educ4} + \beta_{marriage6})$$

```
> cov2 = varcovar[c("(Intercept)","abode2","edu4", "marriage6"),
+               c("(Intercept)","abode2","edu4", "marriage6")]
> var2 = cov2[1,1]+cov2[2,2]+cov2[3,3]+cov2[4,4]+2*cov2[2,1]+ 2*cov2[3,1]+2*cov2[3,2]+2*cov2[4,1]+
+       2*cov2[4,2]+2*cov2[4,3]
> mu2 = fit4$coefficients["(Intercept)"] + fit4$coefficients["abode2"]+
+   fit4$coefficients["edu4"] +fit4$coefficients["marriage6"]
> samples2 = exp(rnorm(5000, mu2, sqrt(var2)))
> mean =exp(mu2)
> mean

(Intercept)
   6.026728

> quantile(samples2, c(.05,.95))

      5%      95%
5.434657 6.708552
```