Lab 7 Tutorial

Nicholas Nagle

April 13, 2015

Homework

Here is a tutorial looking at a dataset using logistic regression.

The data

Tutorial

```
solea <- read.table('Solea.txt', header=TRUE)
head(solea) # See what we've got.
```

```
##   Sample season month Area depth temperature salinity transparency gravel
## 1      1      1     5    2   3.0          20       30           15   3.74
## 2      2      1     5    2   2.6          18       29           15   1.94
## 3      3      1     5    2   2.6          19       30           15   2.88
## 4      4      1     5    4   2.1          20       29           15  11.06
## 5      5      1     5    4   3.2          20       30           15   9.87
## 6      6      1     5    4   3.5          20       32            7  32.45
##   large_sand med_fine_sand   mud Solea_solea
## 1      13.15         11.93 71.18           0
## 2       4.99          5.43 87.63           0
## 3       8.98         16.85 71.29           1
## 4      11.96         21.95 55.03           0
## 5      28.60         19.49 42.04           0
## 6       7.39          9.43 50.72           0
```

```
# Turn the categorical variables into categorical variables
solea$month <- as.factor(solea$month)
solea$Area <- as.factor(solea$Area)
```

These data represent presence/absence of sole (Solea solea) in the Tagus Estuary of Portugal. The sole is an economically valuable species. Spawning occurs on the contentental shelf, but the young will migrate to coastal areas to develop for a few years. The city of Lisbon lies along the Tagus Estuary, and the area is heavily urbanized.

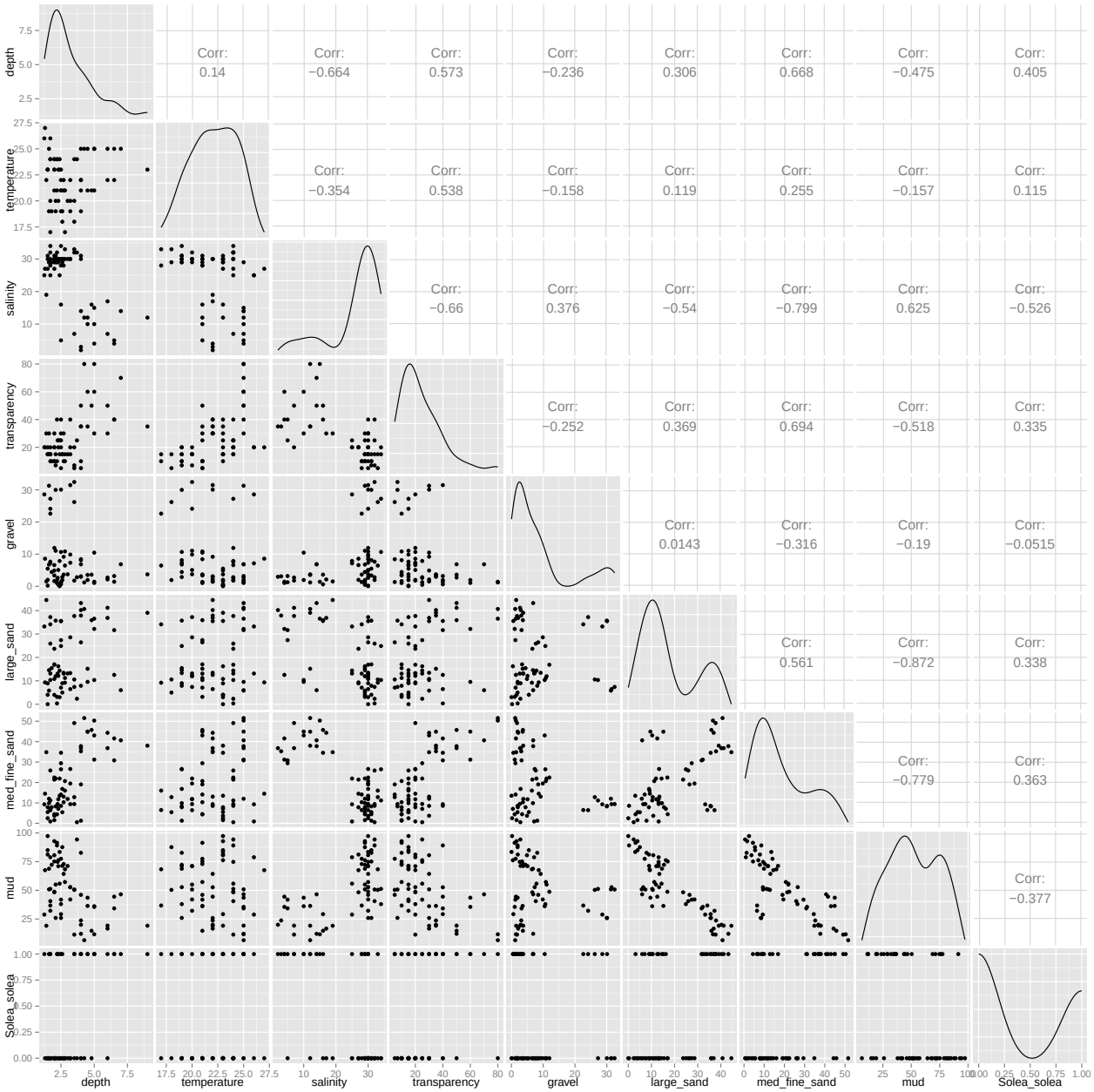
We've got variables on:

1. Season (1=Spring, 2=Summer)
2. Month
3. Station ID
4. Depth (m)
5. Temperature (C)
6. Salinity (ppt)

7. water Transparency (cm)
8. % gravel in sediment
9. % large sand
10. % medium and fine sand
11. % %mud
12. Area - the data were sampled in four different areas.

```
library(dplyr)
library(ggplot2)
library(GGally)

solea %>% select(depth, temperature, salinity, transparency, gravel, large_sand, med_fine_sand, mud, So
```



Let's try the "kitchen sink" regression first - the regression with everything in it. There is no reason to

include the sample id variable, nor season, (since we also have month)

```
mod <- glm(Solea_solea ~ . - Sample - season, data=solea,
           family=binomial(link='logit'))
summary(mod)
```

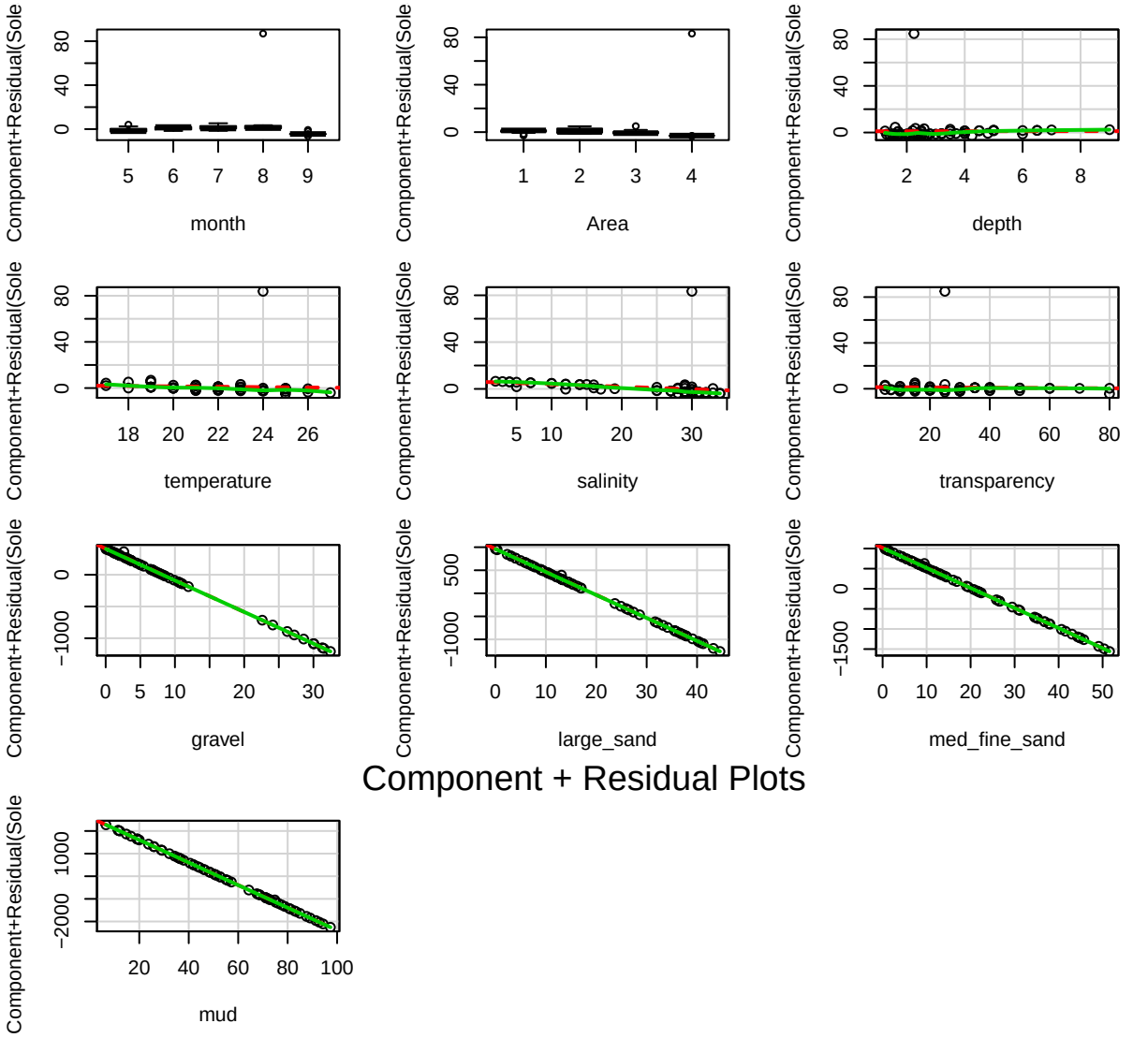
```
##
## Call:
## glm(formula = Solea_solea ~ . - Sample - season, family = binomial(link = "logit"),
##     data = solea)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.520  -0.634  -0.116   0.566   2.981
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   5007.4138  8036.7253    0.62    0.533
## month6           1.7042     1.3992    1.22    0.223
## month7           2.4601     2.7286    0.90    0.367
## month8           2.7990     2.5740    1.09    0.277
## month9          -2.5674     1.5615   -1.64    0.100
## Area2            0.8045     4.1565    0.19    0.847
## Area3           -0.2369     4.1348   -0.06    0.954
## Area4           -2.2860     4.0305   -0.57    0.571
## depth            0.2373     0.5245    0.45    0.651
## temperature     -0.6062     0.4839   -1.25    0.210
## salinity        -0.2522     0.1500   -1.68    0.093 .
## transparency    -0.0216     0.0411   -0.53    0.599
## gravel         -49.7958    80.3675   -0.62    0.536
## large_sand     -49.8947    80.3549   -0.62    0.535
## med_fine_sand  -49.9096    80.3554   -0.62    0.535
## mud            -49.9065    80.3676   -0.62    0.535
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 87.492  on 64  degrees of freedom
## Residual deviance: 47.877  on 49  degrees of freedom
## AIC: 79.88
##
## Number of Fisher Scoring iterations: 6
```

The enormous values and standard errors on gravel, sand, and mud are curious. Let's look at component plus residual plots:

```
library(car)
```

```
## Warning: package 'car' was built under R version 3.1.3
```

```
crPlots(mod)
```



Hmmm. The data fitting on the line for gravel, sand and mud are unsual .
I suspect that multicollinearity is a problem here. These are all fractions. Do they add up to 100%? I.e. are they perfectly collinear? If so, then R should have detected and corrected it. But let's check.

```
solea %>% select(gravel, large_sand, med_fine_sand, mud) %>% rowSums()
```

```
##  [1] 100.00  99.99 100.00 100.00 100.00  99.99 100.00 100.01 100.00 100.00
## [11] 100.01 100.00 100.01 100.00  99.99 100.00 100.00 100.00 100.00 100.00
## [21] 100.00 100.00 100.00 100.01 100.00 100.00 100.00 100.00 100.00 100.00
## [31] 100.00 100.00 100.00 100.00 100.00 100.00 100.00 100.00 100.00 100.00
## [41] 100.01  99.99 100.00 100.00 100.00 100.01 100.00  99.99 100.01 100.00
## [51] 100.00 100.00  99.99 100.00 100.00 100.01 100.00 100.00 100.00 100.00
## [61] 100.00  99.99 100.01 100.00 100.00
```

Ahh. That explains a lot. These are fractions, and they add up to 100, apart from what must be roundoff error. Since they don't add up exactly to 100, R couldn't tell that they were perfectly collinear. We must remove one as the "baseline" category. I'll remove mud.

```
mod <- glm(Solea_solea ~ . - Sample - season - mud, data=solea,
           family=binomial(link='logit'))
summary(mod)
```

```
##
## Call:
## glm(formula = Solea_solea ~ . - Sample - season - mud, family = binomial(link = "logit"),
##     data = solea)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.532  -0.641  -0.149   0.480   2.936
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   16.99395   11.30056    1.50    0.133
## month6         1.65241    1.38575    1.19    0.233
## month7         2.27250    2.68477    0.85    0.397
## month8         2.62042    2.53146    1.04    0.301
## month9        -2.53954    1.52903   -1.66    0.097 .
## Area2          0.59342    4.10899    0.14    0.885
## Area3         -0.54211    4.06367   -0.13    0.894
## Area4         -2.44040    4.00532   -0.61    0.542
## depth          0.16539    0.49164    0.34    0.737
## temperature   -0.58399    0.48454   -1.21    0.228
## salinity      -0.25340    0.15063   -1.68    0.093 .
## transparency  -0.01864    0.04019   -0.46    0.643
## gravel         0.11128    0.05444    2.04    0.041 *
## large_sand     0.00386    0.04091    0.09    0.925
## med_fine_sand -0.01072    0.07066   -0.15    0.879
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 87.492  on 64  degrees of freedom
## Residual deviance: 48.265  on 50  degrees of freedom
## AIC: 78.27
##
## Number of Fisher Scoring iterations: 6
```

We see that gravel is strongly significant. What about the sands? Can we drop those? An ecologist might argue that gravel/mud is the most significant difference, and argue to drop these. We should listen to the subject matter expert. But in the absence of that, is there statistical evidence to drop these?

```
mod2 <- glm(Solea_solea ~ . - Sample - season - mud - large_sand - med_fine_sand, data=solea,
           family=binomial(link='logit'))
summary(mod2)
```

```
##
## Call:
## glm(formula = Solea_solea ~ . - Sample - season - mud - large_sand -
```

```
##     med_fine_sand, family = binomial(link = "logit"), data = solea)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.505  -0.620  -0.135   0.488   2.972
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   17.1494    11.1589    1.54    0.124
## month6         1.6895     1.3651    1.24    0.216
## month7         2.3270     2.6606    0.87    0.382
## month8         2.6556     2.5102    1.06    0.290
## month9        -2.5963     1.5007   -1.73    0.084 .
## Area2          0.7914     3.4557    0.23    0.819
## Area3         -0.3553     3.5768   -0.10    0.921
## Area4         -2.2959     3.4650   -0.66    0.508
## depth          0.1366     0.4519    0.30    0.762
## temperature   -0.5958     0.4789   -1.24    0.213
## salinity      -0.2574     0.1458   -1.77    0.077 .
## transparency  -0.0198     0.0390   -0.51    0.612
## gravel         0.1140     0.0521    2.19    0.029 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 87.492  on 64  degrees of freedom
## Residual deviance: 48.293  on 52  degrees of freedom
## AIC: 74.29
##
## Number of Fisher Scoring iterations: 6
```

```
anova(mod2, mod, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Solea_solea ~ (Sample + season + month + Area + depth + temperature +
##     salinity + transparency + gravel + large_sand + med_fine_sand +
##     mud) - Sample - season - mud - large_sand - med_fine_sand
## Model 2: Solea_solea ~ (Sample + season + month + Area + depth + temperature +
##     salinity + transparency + gravel + large_sand + med_fine_sand +
##     mud) - Sample - season - mud
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        52       48.3
## 2        50       48.3  2   0.0277     0.99
```

```
# Compare the coefficients
coef(mod2)
```

```
##  (Intercept)       month6       month7       month8       month9
##     17.14937      1.68950      2.32704      2.65564     -2.59626
##        Area2        Area3        Area4        depth  temperature
##      0.79139     -0.35526     -2.29593      0.13658     -0.59582
```

```
##     salinity transparency       gravel
##     -0.25744     -0.01981      0.11397
```

```
coef(mod)
```

```
##   (Intercept)        month6        month7        month8        month9
##      16.99395       1.65241       2.27250       2.62042      -2.53954
##         Area2         Area3         Area4         depth   temperature
##       0.59342      -0.54211      -2.44040       0.16539      -0.58399
##      salinity  transparency        gravel    large_sand med_fine_sand
##      -0.25340      -0.01864       0.11128       0.00386      -0.01072
```

There are three statistical indications that we can drop the sands:

1. The AIC drops when they are omitted,
2. This is supported by the ANOVA test that doesn't detect a difference between the models, and
3. the coefficients don't change by very much when they are dropped.

What about area, can we drop that?

```
mod3 <- glm(Solea_solea ~ . - Sample - season - mud - large_sand - med_fine_sand - Area, data=solea,
          family=binomial(link='logit'))
summary(mod3)
```

```
##
## Call:
## glm(formula = Solea_solea ~ . - Sample - season - mud - large_sand -
##     med_fine_sand - Area, family = binomial(link = "logit"),
##     data = solea)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.848  -0.751  -0.231   0.553   2.007
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   25.0041    10.3046    2.43   0.0152 *
## month6         2.1098     1.2368    1.71   0.0880 .
## month7         3.8529     2.2732    1.69   0.0901 .
## month8         4.5289     2.2227    2.04   0.0416 *
## month9        -2.3554     1.4450   -1.63   0.1031
## depth          0.0917     0.3603    0.25   0.7991
## temperature   -0.9161     0.4356   -2.10   0.0355 *
## salinity      -0.3034     0.0976   -3.11   0.0019 **
## transparency  -0.0282     0.0333   -0.85   0.3978
## gravel         0.0703     0.0402    1.75   0.0808 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 87.492  on 64  degrees of freedom
```

```
## Residual deviance: 53.412  on 55  degrees of freedom
## AIC: 73.41
##
## Number of Fisher Scoring iterations: 6
```

```
anova(mod3, mod2, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Solea_solea ~ (Sample + season + month + Area + depth + temperature +
##     salinity + transparency + gravel + large_sand + med_fine_sand +
##     mud) - Sample - season - mud - large_sand - med_fine_sand -
##     Area
## Model 2: Solea_solea ~ (Sample + season + month + Area + depth + temperature +
##     salinity + transparency + gravel + large_sand + med_fine_sand +
##     mud) - Sample - season - mud - large_sand - med_fine_sand
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        55       53.4
## 2        52       48.3  3     5.12     0.16
```

```
# compare the coefficients:
coef(mod2)
```

```
##  (Intercept)       month6       month7       month8       month9
##     17.14937      1.68950      2.32704      2.65564     -2.59626
##        Area2        Area3        Area4        depth  temperature
##      0.79139     -0.35526     -2.29593      0.13658     -0.59582
##     salinity transparency       gravel
##     -0.25744     -0.01981      0.11397
```

```
coef(mod3)
```

```
##  (Intercept)       month6       month7       month8       month9
##     25.00409      2.10979      3.85285      4.52891     -2.35540
##        depth  temperature     salinity transparency       gravel
##      0.09170     -0.91613     -0.30341     -0.02816      0.07027
```

This is a tough call. Based on the p-value and the AIC, we should drop Area. But, the coefficients do change by a bit when Area is dropped.
In particular, temperature becomes significant. NOTE: Other analysis points to the "40" location as being very different than the others, for no apparent reason. If this point is removed, then that leads to the different conclusion that we should keep Area in the model. I'll ignore this point for now and proceed by dropping Area based the AIC and p-statistics. I am not completely happy with this situation, however. But in practice, you should investigate this observation further. What I suspect is happening is that a lot of the difference between Areas is due to systematic differences between temperature, salinity, etc. But neither can temperature and salinity completely explain the remaining differences between area. When we remove Area, we see that temperature and salinity are significant, but also, by removing Area, we can no longer capture that there may be remaining differences between the areas.

I'm going to cautiously proceed with the model without Area. Next, are transparency and depth important? I'll do this in one step, rather than two, for no reason other than that this is getting to be a long tutorial.

```
anova(mod3, update(mod3, . ~ . -transparency - depth), test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Solea_solea ~ (Sample + season + month + Area + depth + temperature +
##     salinity + transparency + gravel + large_sand + med_fine_sand +
##     mud) - Sample - season - mud - large_sand - med_fine_sand -
##     Area
## Model 2: Solea_solea ~ month + temperature + salinity + gravel
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        55       53.4
## 2        57       54.1 -2   -0.733     0.69
```

```
mod4 <- update(mod3, . ~ . -transparency - depth)
```

The AIC and ANOVA says to drop these, but I'm not so sure. Removing transparency and depth slightly affect the salinity and gravel measure, but the change is within the level of uncertainty. It turns out that dropping transparency affects the measure of gravel, and dropping depth affects the measure of salinity. The following code shows that:

```
coef(mod3)
```

```
##  (Intercept)       month6       month7       month8       month9
##     25.00409      2.10979      3.85285      4.52891     -2.35540
##        depth  temperature     salinity transparency       gravel
##      0.09170     -0.91613     -0.30341     -0.02816      0.07027
```

```
coef(update(mod3, . ~ . - transparency))
```

```
## (Intercept)      month6      month7      month8      month9       depth
##  23.3617061   2.0097832   3.3694143   3.8672599  -2.2064041  -0.0006583
## temperature    salinity      gravel
##  -0.8929936  -0.2640518   0.0661313
```

```
coef(update(mod3, . ~ . - depth))
```

```
##  (Intercept)       month6       month7       month8       month9
##     25.25825      2.08325      3.71353      4.45593     -2.36707
##  temperature     salinity transparency       gravel
##     -0.90873     -0.30945     -0.02565      0.07014
```

At this point, you have four slightly different models to choose from.
Whether you drop some of these insignificant variables or leave them in would depend on whether you believed that dropping them might be creating omitted variables bias or not.
If the science behind including these variables is weak, then I would drop them. On the other hand, if the science is pretty strong, then I would leave them in despite their non-significance.
Remember, non significance doesn't mean that they don't belong, it just means that you don't have enough data to tell. Another consideration is that statistical significance is not scientific significance. It could be that there is a large effect, but that the standard error is also large. Or it could be that the effect is small, close to zero. Stastistically, there is no difference between the two, but scientists might disagree.

It might be helpful to report confidence intervals on the coefficients and on the odds ratios:

```
cbind(coef(mod4), LOR = confint(mod4)) # Log Odds ratio
```

```
## Waiting for profiling to be done...
```

```
##                          2.5 %  97.5 %
## (Intercept) 23.35880  6.536827 45.2884
## month6       2.00988 -0.262837  4.5907
## month7       3.37017 -0.631279  7.9752
## month8       3.86737  0.219814  8.3345
## month9      -2.20621 -5.132753  0.1958
## temperature -0.89304 -1.848969 -0.1452
## salinity    -0.26398 -0.423805 -0.1449
## gravel       0.06613 -0.009653  0.1486
```

```
exp(cbind(coef(mod4), OR = confint(mod4))) # Odds ratio
```

```
## Waiting for profiling to be done...
```

```
##                           2.5 %    97.5 %
## (Intercept) 1.395e+10 690.0932 4.661e+19
## month6      7.462e+00   0.7689 9.856e+01
## month7      2.908e+01   0.5319 2.908e+03
## month8      4.782e+01   1.2458 4.165e+03
## month9      1.101e-01   0.0059 1.216e+00
## temperature 4.094e-01   0.1574 8.649e-01
## salinity    7.680e-01   0.6546 8.652e-01
## gravel      1.068e+00   0.9904 1.160e+00
```

```
round(exp(cbind(coef(mod4), OR = confint(mod4)))[-c(1:5),], digits=2) # a little prettier
```

```
## Waiting for profiling to be done...
```

```
##                  2.5 % 97.5 %
## temperature 0.41  0.16   0.86
## salinity    0.77  0.65   0.87
## gravel      1.07  0.99   1.16
```

The interpretation of these numbers is that - for exampple - each 1 degree Celsius increase in temperature multiples the odds ratio by .15 to .86, i.e. decreases the odds ratio of sole presence 14% ((1-.86)100) to 85% ((1-.15)100). A large margin for sure, but we can be certain that the effect is to reduce the probability of presence for sole. The other coefficients can be interpreted similarly.

In my final analysis, I would probably report the Odds Ratios for the model with depth and transparency ommitted, just for comparison (because this model has the lowest AIC value).

```
round(exp(cbind(coef(mod3), OR = confint(mod3)))[-c(1:5),], digits=2) # a little prettier
```

```
## Waiting for profiling to be done...
```

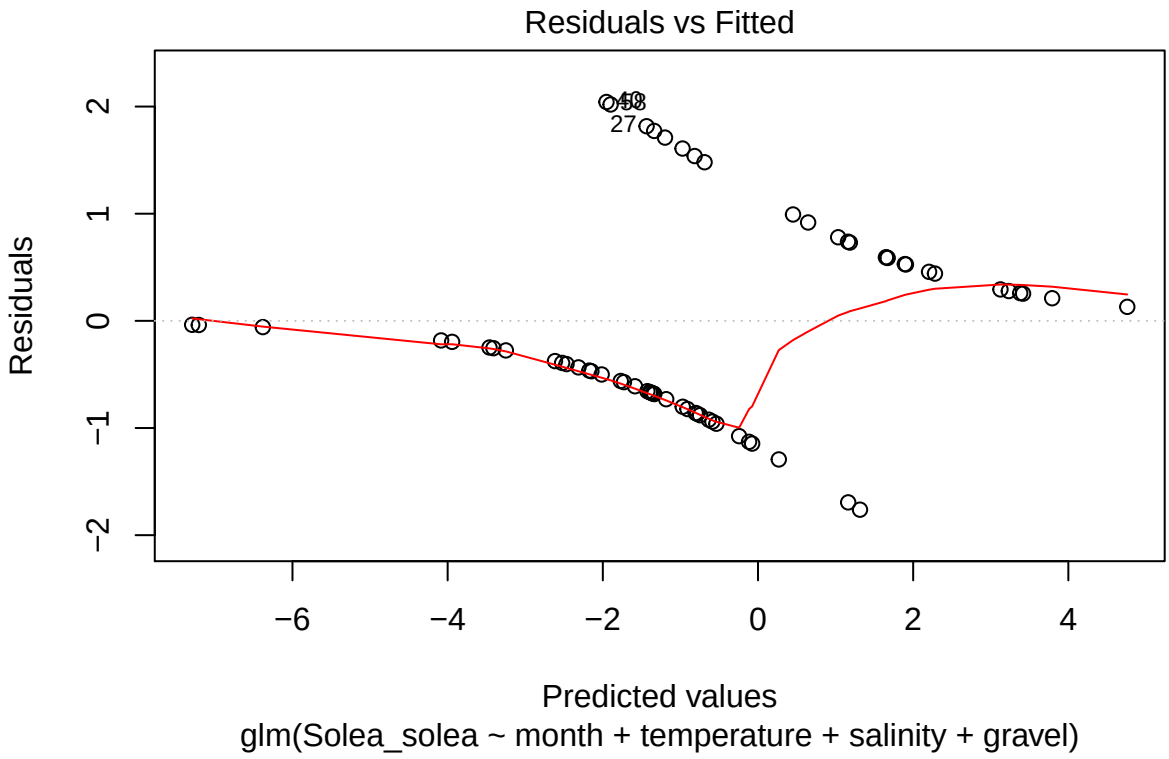
```
##                   2.5 % 97.5 %
## depth        1.10  0.56   2.35
## temperature  0.40  0.15   0.86
## salinity     0.74  0.59   0.87
## transparency 0.97  0.91   1.04
## gravel       1.07  0.99   1.17
```

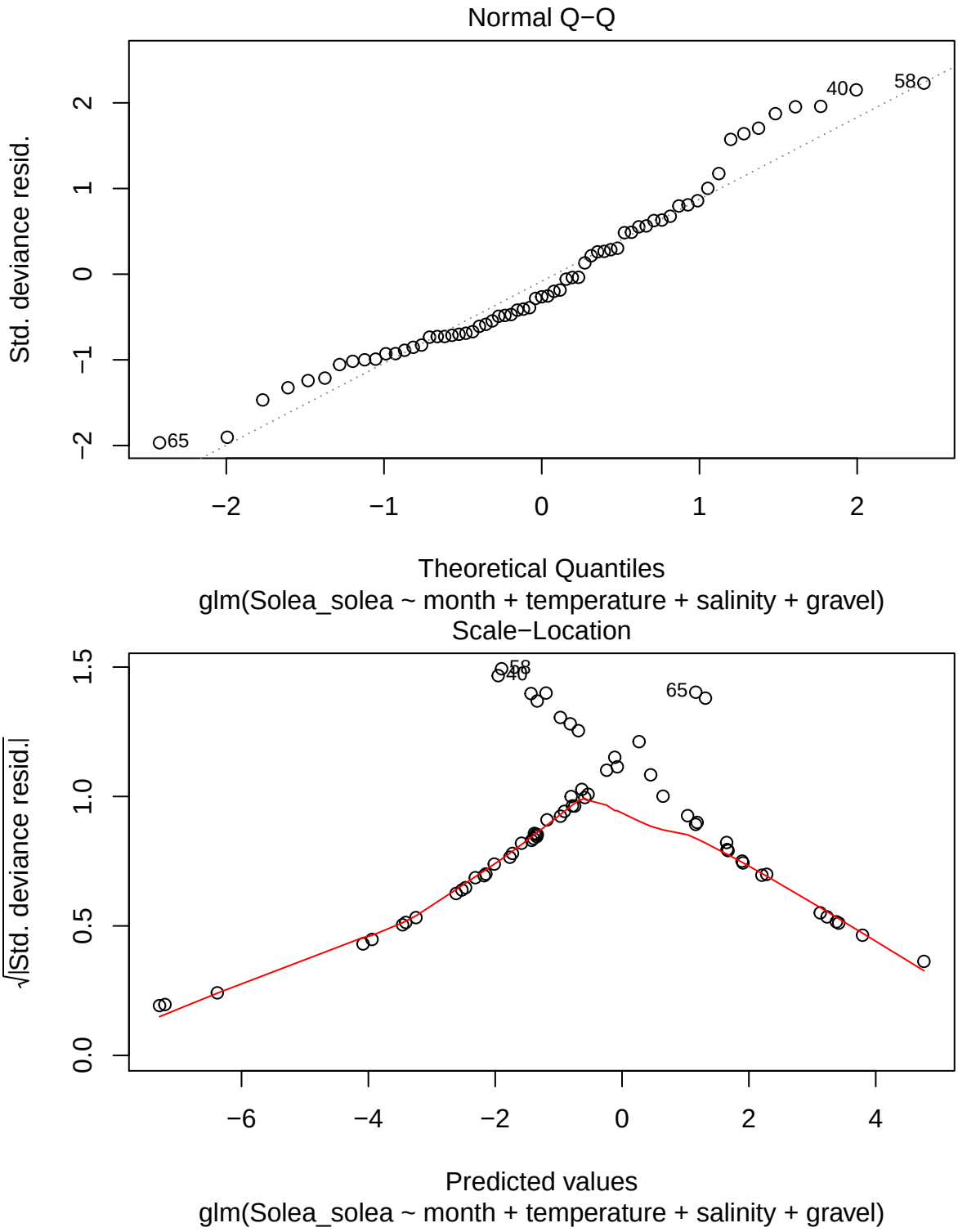
In terms of Odds Ratios, it appears that we can drop those variables and hardly change the result.

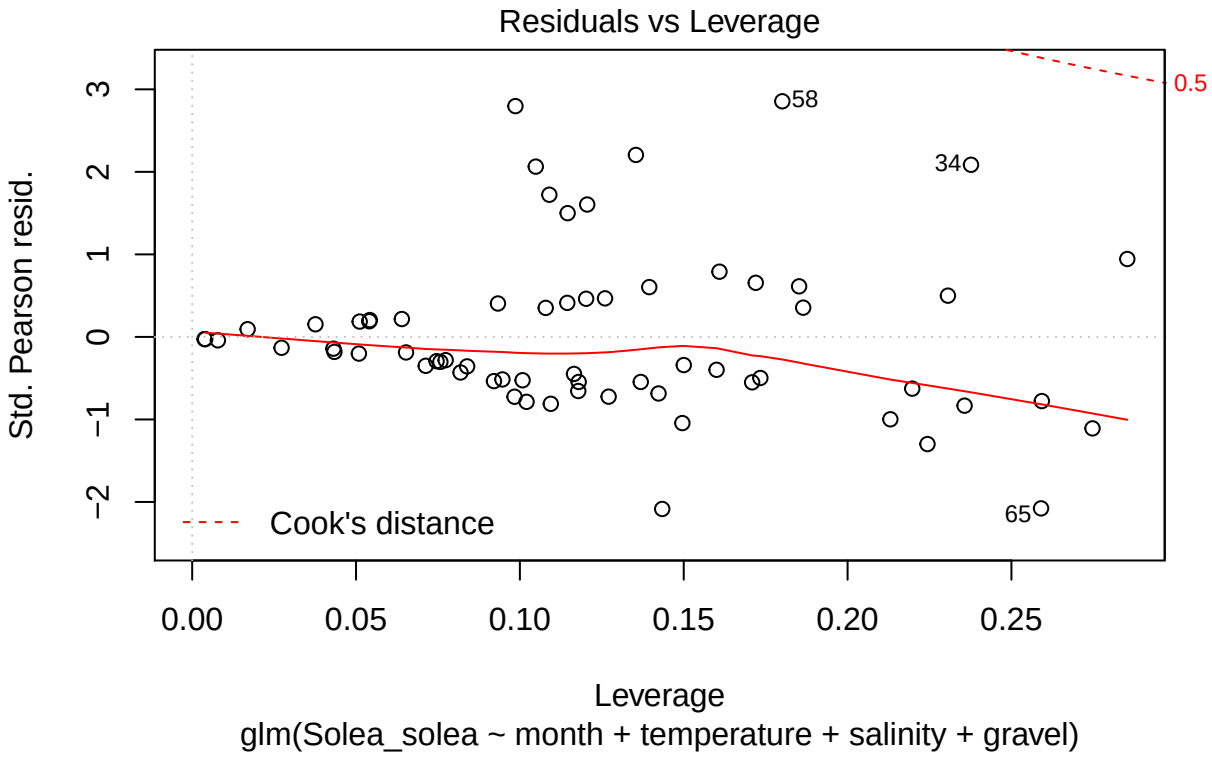
But note that the effect of depth, while statistically insignificant, could be quite large, scientifically. A one meter increase in depth could change the odds anywhere from -44% to +235%. Only more data or better data would tell for certain. In contrast, transparency is not only statistically insignificant, but the effect sizes on the Odds ratios are not very different from 1, either.

Finally, you should look at the battery of diagnostic plots to check for obvious problems. I see none. Maybe a nonlinear effect in salinity?

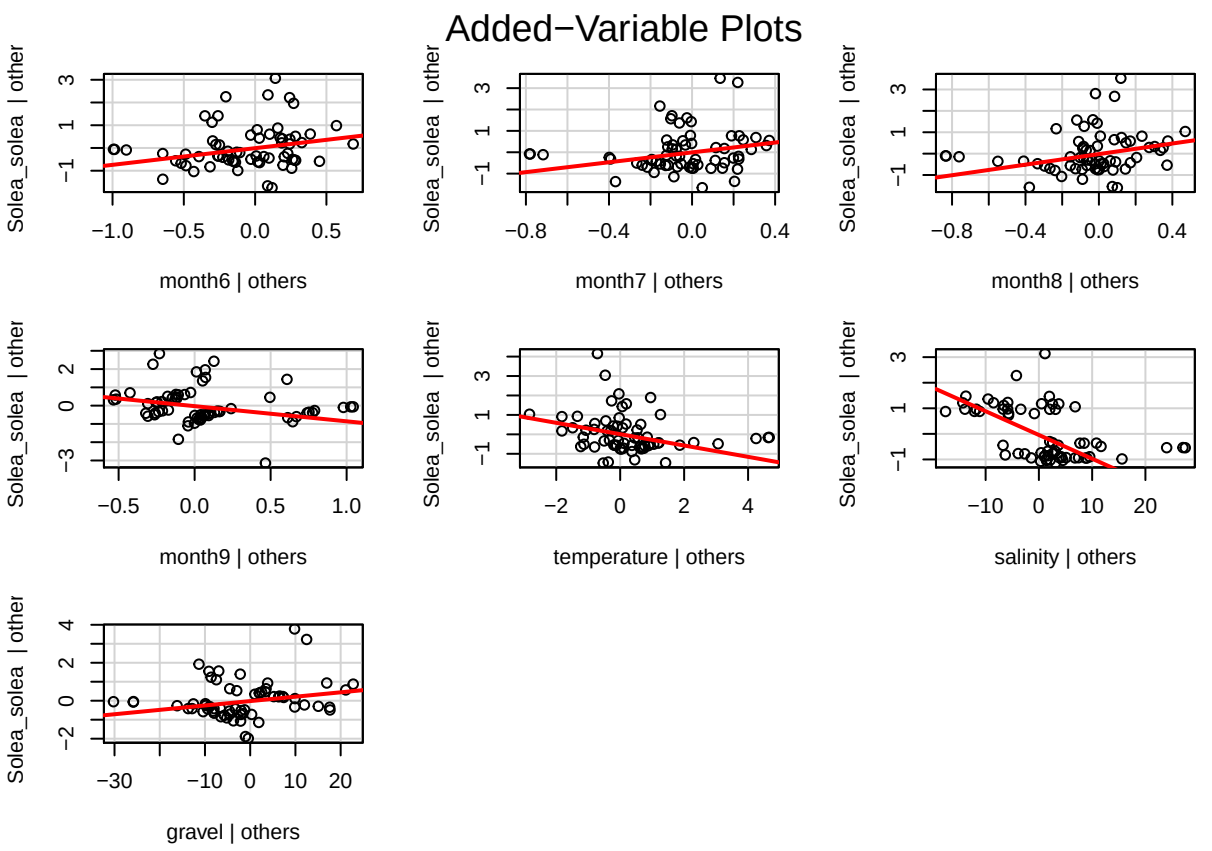
```
plot(mod4)
```



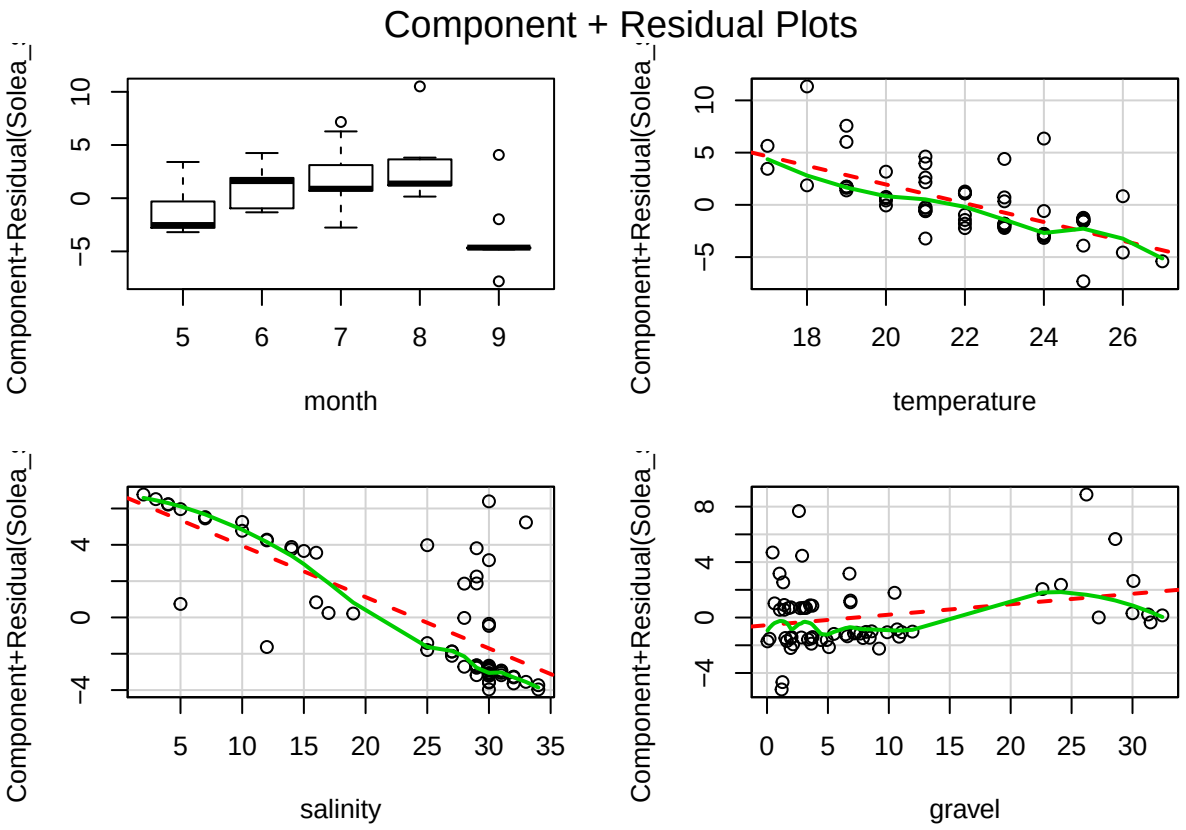

Normal Q−Q
Std. deviance resid.
65
40
58
−2
−1
0
1
2
Theoretical Quantiles
glm(Solea_solea ~ month + temperature + salinity + gravel)
Scale−Location
√|Std. deviance resid.|
58
40
65
0.0
0.5
1.0
1.5
−6
−4
−2
0
2
4
Predicted values
glm(Solea_solea ~ month + temperature + salinity + gravel)




```
avPlots(mod4)
```

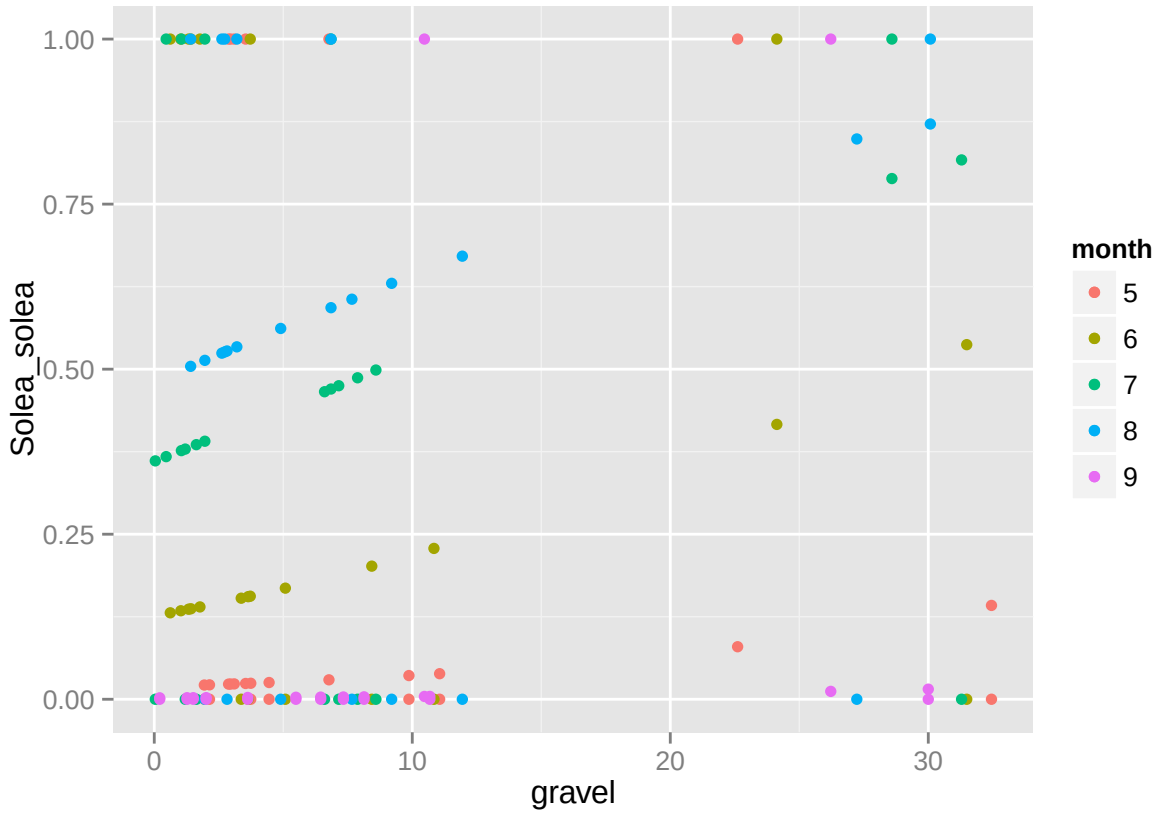


```
crPlots(mod4)
```



Finally, it might be nice to produce a plot showing the effect of something like gravel on probability.

```
pred.data <- solea
pred.data$salinity=29
pred.data$temperature=22
pred.data$pred <- predict(mod4, pred.data, type='response')
ggplot(data=pred.data) + geom_point(aes(y=Solea_solea, x=gravel, color=month)) + geom_point(aes(y=pred,
```


1.00
0.75
0.50
0.25
0.00
Solea_solea
0
10
20
30
gravel
month
5
6
7
8
9