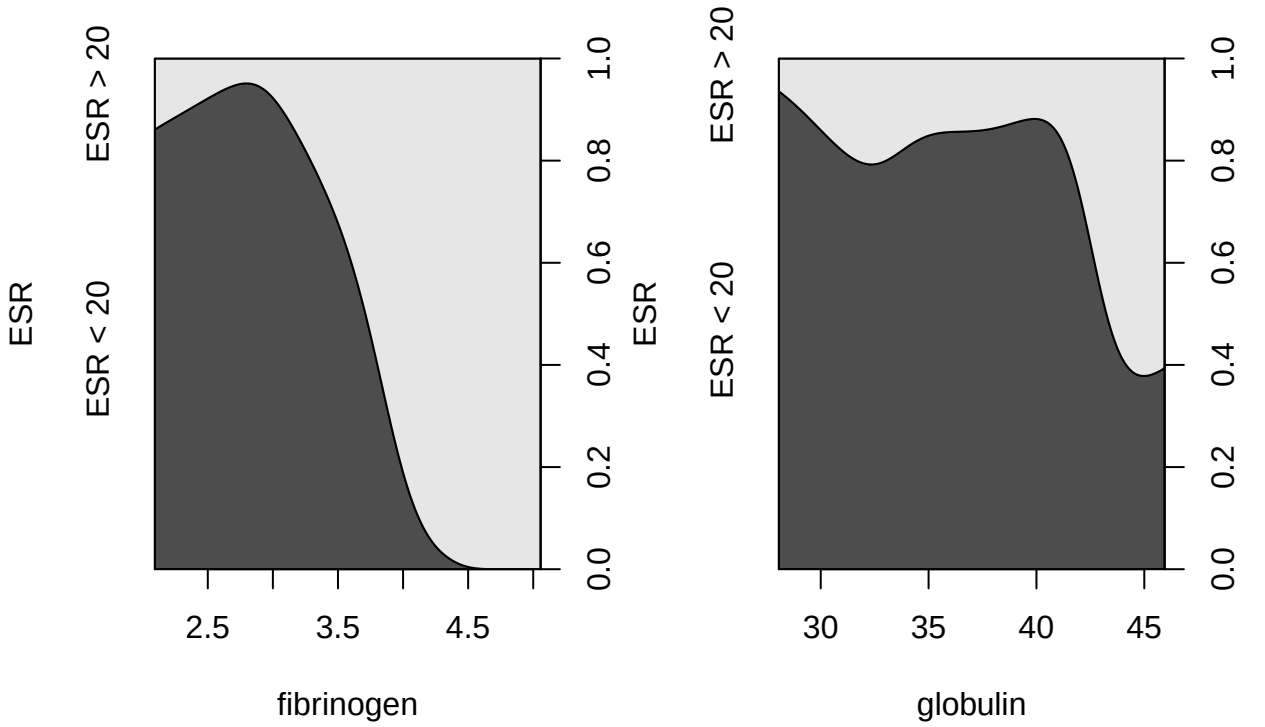
# Lab 5

## Analysis of Blood plasma data

Conditional density plots of the data:

```
data("plasma", package = "HSAUR3")
layout(matrix(1:2, ncol = 2))
cdplot(ESR ~ fibrinogen, data = plasma)
cdplot(ESR ~ globulin, data = plasma)
```




We can see that higher levels of each protein are associated with ESR values above 20. Now, fit a logistic regression model to the data with the glm function, including only the single variable fibrinogen:

```
plasma.glm.1 <- glm(ESR ~ fibrinogen, data = plasma, family = binomial())
summary(plasma.glm.1)
```

```
##
## Call:
## glm(formula = ESR ~ fibrinogen, family = binomial(), data = plasma)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9298  -0.5399  -0.4382  -0.3356   2.4794
##
## Coefficients:
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -6.8451     2.7703  -2.471   0.0135 *
## fibrinogen    1.8271     0.9009   2.028   0.0425 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 30.885  on 31  degrees of freedom
## Residual deviance: 24.840  on 30  degrees of freedom
## AIC: 28.84
##
## Number of Fisher Scoring iterations: 5
```

```
head(plasma$ESR)
```

```
## [1] ESR < 20 ESR < 20 ESR < 20 ESR < 20 ESR < 20 ESR < 20
## Levels: ESR < 20 ESR > 20
```

We see significance at the 5% level. The regression equation here is

$$\log\left\{\frac{\pi(x)}{1-\pi(x)}\right\} = \beta_0 + \beta_1 \cdot \text{fibrinogen} + \text{error}.$$

Then exponentiate both sides to see that the $\beta_1$ would correspond to the increase in the odds ratio, if we were to let fibrinogen increase by one unit and hold everything else fixed. The 90% CI on this coefficient is:

```
confint(plasma.glm.1, parm = "fibrinogen")
```

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

```
##     2.5 %    97.5 %
## 0.3387619 3.9984921
```

```
coef(plasma.glm.1)["fibrinogen"]
```

```
## fibrinogen
##   1.827081
```

```
exp(coef(plasma.glm.1)["fibrinogen"])
```

```
## fibrinogen
##   6.215715
```

```
exp(confint(plasma.glm.1, parm = "fibrinogen"))
```

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

```
##     2.5 %    97.5 %
##  1.403209 54.515884
```

One reason why the confindence interval could be large is the small sample size, with large variability in response.

Now, using both variables:

```
plasma.glm.2 <- glm(ESR ~ fibrinogen + globulin, data = plasma, family = binomial())
anova(plasma.glm.1, plasma.glm.2, test = "Chisq")
```

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

```
## Model 1: ESR ~ fibrinogen
## Model 2: ESR ~ fibrinogen + globulin
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        30     24.840
## 2        29     22.971  1   1.8692   0.1716
```

Coefficient of globulin not significantly different than 0 as the associated p-value is above the 0.05 threshold.

## Analysis of Colon polyps data

```
data("polyps", package = "HSAUR3")
mu_polyp <- mean(polyps$number)
var_polyp <-var(polyps$number)
```

Recall that the Poison assumption requires that $\mu = \sigma^2$, while we have $\mu$ =24.05 and $\sigma^2$ = 434.6815789. Even by looking at the mean we see that the variance is much larger than the mean, hence indicating overdispersion.

```
polyps.glm.1 <- glm(number ~ treat + age, data = polyps, family = poisson())
summary(polyps.glm.1)
```

```
##
## Call:
## glm(formula = number ~ treat + age, family = poisson(), data = polyps)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.2212  -3.0536  -0.1802   1.4459   5.8301
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.529024   0.146872   30.84  < 2e-16 ***
## treatdrug   -1.359083   0.117643  -11.55  < 2e-16 ***
## age         -0.038830   0.005955   -6.52 7.02e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 378.66  on 19  degrees of freedom
## Residual deviance: 179.54  on 17  degrees of freedom
## AIC: 273.88
##
## Number of Fisher Scoring iterations: 5
```

Overdispersion can also be seen by looking that the deviance and the degrees of freedom. Thus, modeling this data using the Poisson family as assumption yield erroneous p-values regarding model variables.

```
polyps.glm.2 <- glm(number ~ treat + age, data = polyps, family = quasipoisson())
summary(polyps.glm.2)
```

```
##
## Call:
## glm(formula = number ~ treat + age, family = quasipoisson(),
##     data = polyps)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.2212  -3.0536  -0.1802   1.4459   5.8301
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.52902    0.48106   9.415 3.72e-08 ***
## treatdrug   -1.35908    0.38533  -3.527  0.00259 **
## age         -0.03883    0.01951  -1.991  0.06284 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 10.72805)
##
##     Null deviance: 378.66  on 19  degrees of freedom
## Residual deviance: 179.54  on 17  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

The treatment is significant in both cases, indicating that the drug may be effective at reducing the number of polyps (notice negative relationship from $\beta_1 < 0$). The difference is that when we take overdispersion into account, age is no longer significant. One reason why this might be the case is that the additional overdispersion parameter $\phi = 10.73$ might capture a large amount of thes variation of the response variable, therefore making the relation with age now insignificant. Note that since the underlying data is the same, the residuals deviance is the same.