Model selection

We'll start just fitting a model with just an intercept, and then add in effects of gears,

```
null.model = glm(SEX ~ 1, family = "binomial", data=d)
gear.model = glm(SEX ~ gear, family = "binomial", data=d)
```

We can add time in in one of two ways, either as a linear trend or a factor. [Note, we could also use non-linear models, like GAMs, but those give a pretty linear estimate because of small sample sizes]

```
gearTime.model01 = glm(SEX ~ gear + as.factor(Year), family = "binomial", data=d)
gearTime.model02 = glm(SEX ~ gear + Year, family = "binomial", data=d)
```

We can also try a few flavors of GAMs,

```
gam.model = gam(SEX ~ gear + s(Year), family = "binomial", data=d)
gam.model.02 = gam(SEX ~ s(Year, by=as.factor(gear)), family = "binomial", data=d)
```

Model	AIC	delta_aic
Gear, age(f), time(f)	163299.0	0.00
Gear, age(s), time(f)	163302.9	3.84
Gear, age, time(f)	163312.5	13.50
Gear, age(f), time(s)	163511.5	212.46
Gear, age(s), time(s)	163524.2	225.20
Gear, age, time(s)	163531.0	231.95
Gear, age(f), time	163614.3	315.32
Gear, age(s), time	163627.0	327.97
Gear, age, time	163635.4	336.34
Null	164406.7	1107.66

Based on this table (lower AICs = better models) we'd say that the best model includes time as factor, but both that model and the model with the temporal trend are better than the model without time as a predictor. This model also does better than GAM models, with smooth terms (either shared, or by gear). To visualize what the sex ratio looks like, we can look at the output of a GAM.

Are these changes biologically meaningful? We can make predictions with a new data frame,

```
pred.C = predict(best_model, newdata = data.frame("Year"=2005, "gear"="C", "AGE" = 3:11),
  type="response", se.fit=TRUE)
pred.P = predict(best_model, newdata = data.frame("Year"=2005, "gear"="P", "AGE" = 3:11),
  type="response", se.fit=TRUE)

df_C = data.frame("Age"=3:11, pred = pred.C$fit, low = pred.C$fit - 2*pred.C$se.fit,
  hi = pred.C$fit + 2*pred.C$se.fit, gear = "C", stringsAsFactors = FALSE)
df_P = data.frame("Age"=3:11, pred = pred.P$fit, low = pred.P$fit - 2*pred.P$se.fit,
  hi = pred.P$fit + 2*pred.P$se.fit, gear = "P", stringsAsFactors = FALSE)
df = rbind(df_C, df_P)

names(df)[which(names(df)=="gear")] = "Gear"
df$Gear[which(df$Gear=="C")] = "Cast net"
df$Gear[which(df$Gear=="P")] = "Purse seine"

g1 = ggplot(df, aes(Age, pred, group = Gear)) +
  geom_line() +
```
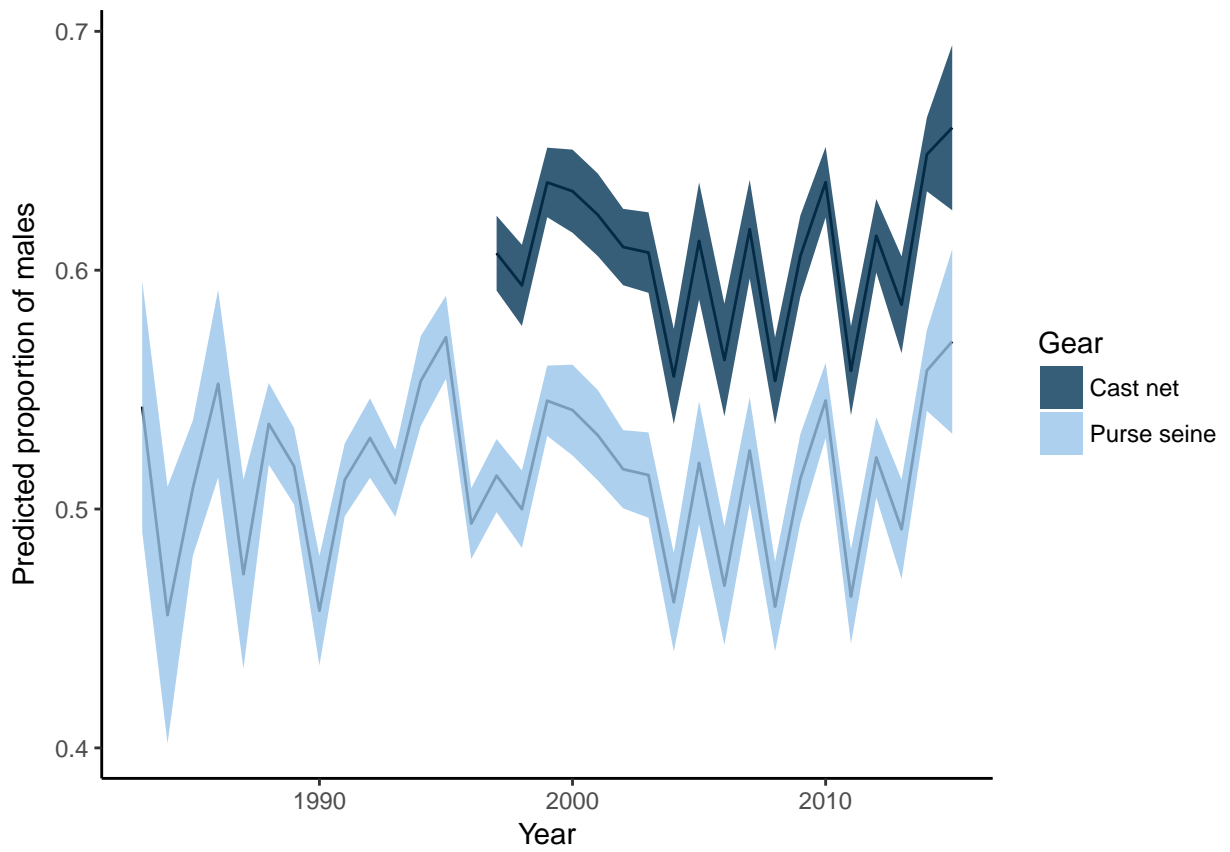
Figure 1: Estimated percent male for PWS herring, for 2 gears sampling in spring (March - April). Estimates shown from best-fit GLM (Table 1), where gear is treated as a factor/offset.

```
  xlab("Age") + ylab("Predicted proportion of males") + scale_fill_manual(values=c("#043657", "#98c5ea")
  geom_ribbon(aes(ymin=low, ymax=hi, fill=Gear), alpha=0.8) +
  theme_bw() + theme(panel.border = element_blank(), panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(), axis.line = element_line(colour = "black"))
g1
```
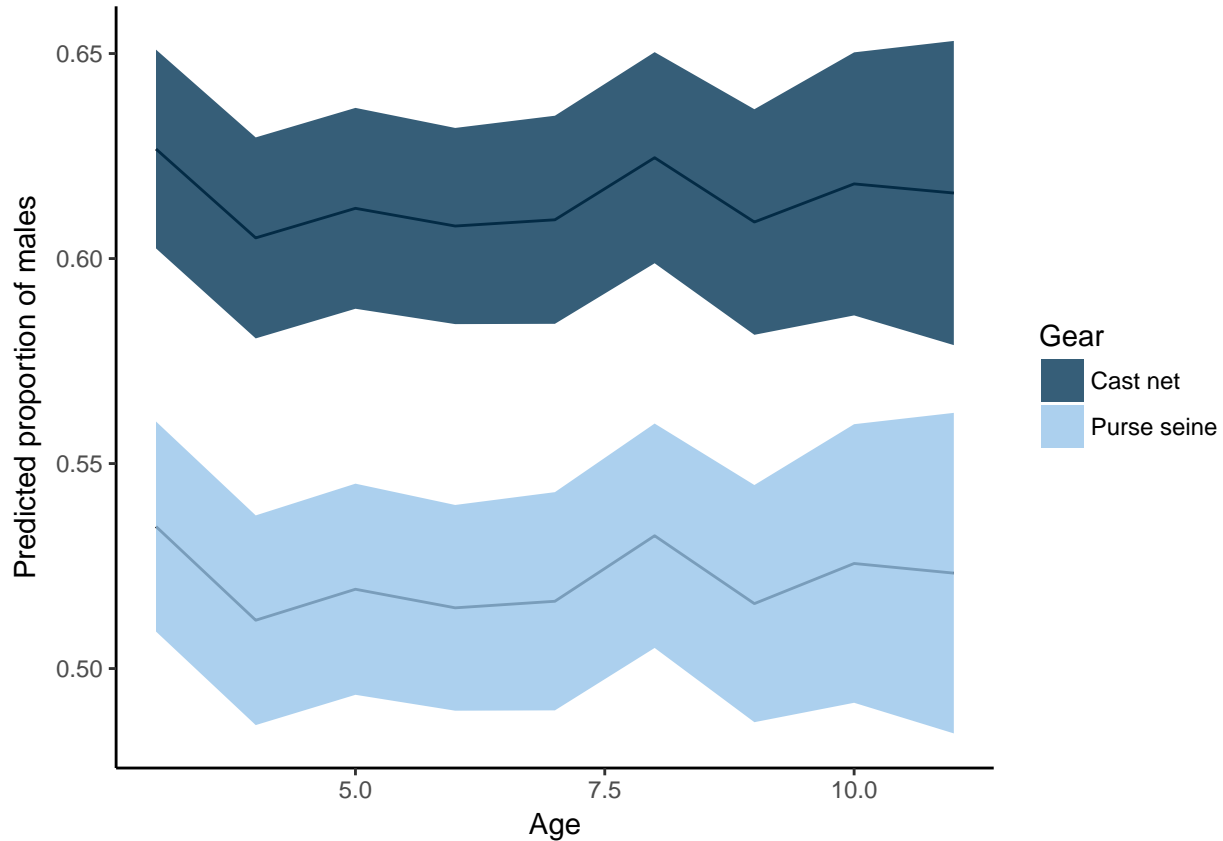


```
pdf("../figures/Figure 05.pdf")
g1
dev.off()
```

```
## pdf
##   2
```

Year coefficients

The year coefficients from the best model show an interesting pattern of oscillations over time, mostly between odd / even years. Extracting these coefficients,

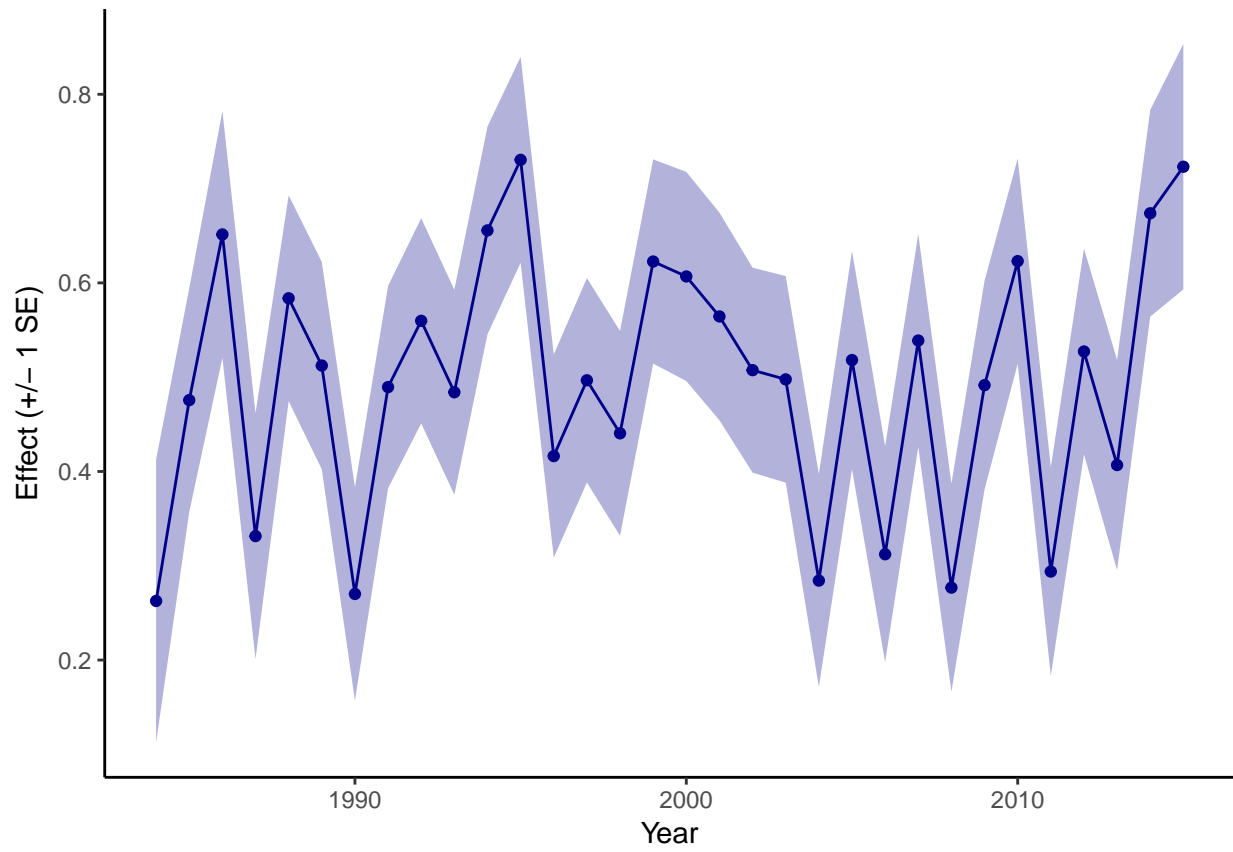We can also look at these in the raw data, calculating proportions by age and gear.

Figure 2: Estimated year coefficients from best-fit GLM (Table 1), where gear is treated as a factor/offset.