Discussion 3

Sam Reeves

3.3

Using the teengamb data, fit a model with gamble as response and the other variables as predictors.

```
lmod <- lm(gamble ~ ., teengamb)
```

(a) Which variables are statistically significant at the 5% level?

```
summary(lmod)
```

```
##
## Call:
## lm(formula = gamble ~ ., data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.082 -11.320  -1.451   9.452  94.252
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.55565   17.19680   1.312   0.1968
## sex         -22.11833    8.21111  -2.694   0.0101 *
## status        0.05223    0.28111   0.186   0.8535
## income        4.96198    1.02539   4.839 1.79e-05 ***
## verbal       -2.95949    2.17215  -1.362   0.1803
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.69 on 42 degrees of freedom
## Multiple R-squared:  0.5267, Adjusted R-squared:  0.4816
## F-statistic: 11.69 on 4 and 42 DF,  p-value: 1.815e-06
```

Only income and sex!

(b) What interpretation should be given to the coefficient for sex?

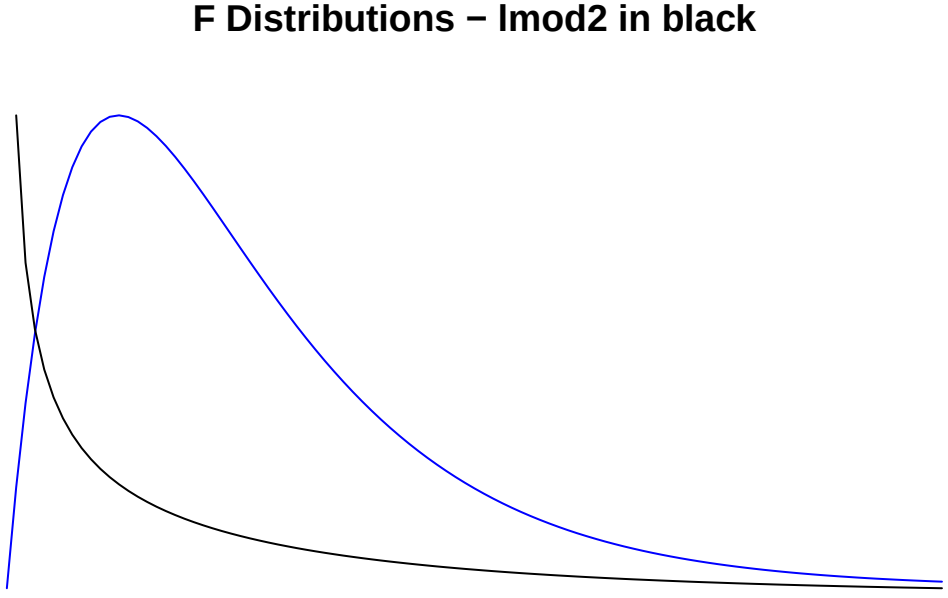
It has significant predictive value, however income will be a bigger deciding factor alone, and it will affect the predictive power of sex.

(c) Fit a model with just income as a predictor and use an F-test to compare it to the full model.

```
lmod2 <- lm(gamble ~ income, teengamb)
summary(lmod2)
```

```
##
## Call:
## lm(formula = gamble ~ income, data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.020 -11.874  -3.757  11.934 107.120
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -6.325      6.030  -1.049      0.3
## income         5.520      1.036   5.330 3.05e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.95 on 45 degrees of freedom
## Multiple R-squared:  0.387,  Adjusted R-squared:  0.3734
## F-statistic: 28.41 on 1 and 45 DF,  p-value: 3.045e-06
```

```
curve(df(x, df1=4, df2=42), from=0, to=4, xlab = '', ylab = '', col = 'blue', main = 'F Distributions -
par(new=TRUE)
curve(df(x, df1=1, df2=45), from=0, to=4, xlab = '', ylab = '', axes = FALSE)
```



```
nullmod <- lm(gamble ~ 1, teengamb)
anova(nullmod, lmod)
```

```
## Analysis of Variance Table
##
## Model 1: gamble ~ 1
## Model 2: gamble ~ sex + status + income + verbal
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1     46 45689
## 2     42 21624  4     24066 11.686 1.815e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(nullmod, lmod2)
```

```
## Analysis of Variance Table
##
## Model 1: gamble ~ 1
## Model 2: gamble ~ income
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1     46 45689
## 2     45 28009  1     17681 28.407 3.045e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The model with only income seems about a third as effective.