Contents

1 Introduction **1**

2 Basic Probability and Statistics **3**

3 Linear Regression: The Basics **7**

Chapter 1

Introduction

Chapter 2

Basic Probability and Statistics

Problem 1

Think about what it means to scale this data. We have a set of data X and this data has a mean μ_x and standard deviation σ. We want to linearly transform the data which means:

$$\begin{aligned} Y &= a + bX \\ \mu_y &= a + b\mu_x \\ \sigma_y &= |b|\sigma_x \end{aligned}$$

Now to answer (a) and (b) we can just solve these equations. Let's solve for standard deviation.

$$\begin{aligned} 15 &= |b|10 \\ 1.5 &= |b| \end{aligned}$$

This means that $b = \pm 1.5$ and we can now solve for the mean to get that $a = 47.5$ or $a = 152.5$. We can transform it with either the first set of values or the second. You don't want to use the second because that will reverse the ordering in the data.

Problem 2

(a)

Doing this in R is easy:

```
girls <- c(.4777, .4875, .4874, .4859, .4754, .4864,.4813,.4787,.4895,
          .4797,.4876,.4859,.4857,.4907, .5010,.4903,.4860,.4911,.4871,
```

```
            .4725,.4822,.4870,.4823,.4973)

num_births <- 3903
std <- sd(girls)
avg <- mean(girls)
expected_std <- sqrt(avg * (1 - avg) / num_births)
```

We get that std= .0064 and expected_std= .008

Let's quickly prove this formula for the expected standard error of a proportion.

Proof. We are trying to prove that $SE(X) = \sqrt{\frac{p(1-p)}{n}}$ for binary variables x_i which take on only values 0 and 1.

Begin by assuming that we have a population with m instances where $x = 1$ and $n - m$ instances where $x = 0$ We know that the standard deviation of a population is:

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$$

Let's just concern ourselves with the summation portion of this equation and see if we can massage it into the desirable form.

$$\begin{aligned}
\sum (x_i - \bar{x})^2 &= \sum (x_i - \bar{x})(x_i - \bar{x}) \\
&= \sum x_i^2 - 2x_i\bar{x} + \bar{x}^2 \\
&= \sum_{x_i=0} \bar{x}^2 + \sum_{x_i=1} 1 - 2\bar{x} + \bar{x}^2 \\
&= (n-m)\bar{x}^2 + m - 2m\bar{x} + m\bar{x}^2 \\
&= (n-m)\frac{m^2}{n^2} + m - 2m\frac{m}{n} + m\frac{m^2}{n^2} && \bar{x} = \frac{m}{n} \\
&= m + \frac{m^2}{n} - \frac{m^3}{n^2} - \frac{2m^2}{n} + \frac{m^3}{n^2} \\
&= m - \frac{m^2}{n} \\
&= m\left(1 - \frac{m}{n}\right) \\
&= np(1-p) && p = m/n
\end{aligned}$$

So we can finish by substituting this back into our original equation!

$$\begin{aligned}
\sigma &= \sqrt{\frac{np(1-p)}{n}} \\
&= \sqrt{p(1-p)}
\end{aligned}$$

Since the sample proportion is a mean the standard error is calculated like normal yielding

$$\begin{aligned} SE(X) &= \frac{\sigma_x}{\sqrt{n}} \\ &= \sqrt{\frac{p(1-p)}{n}} \end{aligned}$$

□

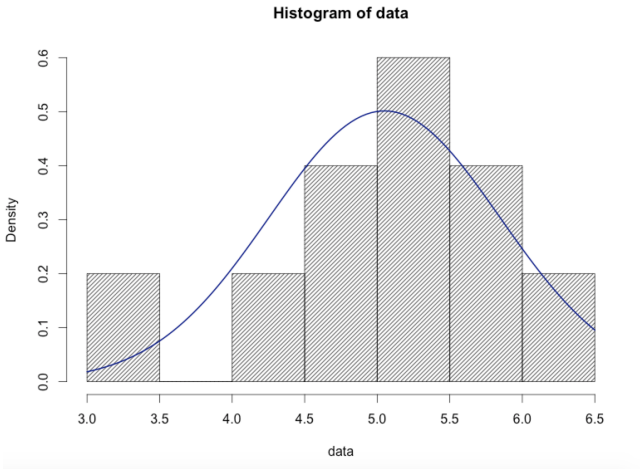
(b)

No they are not significant. If we run a χ^2 test we get $\chi^2 = .0019$ which is no where near what we need for a significant result.

Problem 3

```
cols <- 20
rows <- 1000
x <- replicate(cols, runif(rows))
data <- apply(a, 1, sum)
avg <- mean(data)
std <- sd(data)
hist(data, density=20)
curve(dnorm(x, mean=avg, sd=std),
      col="darkblue", lwd=2, add=TRUE, yaxt="n")
```

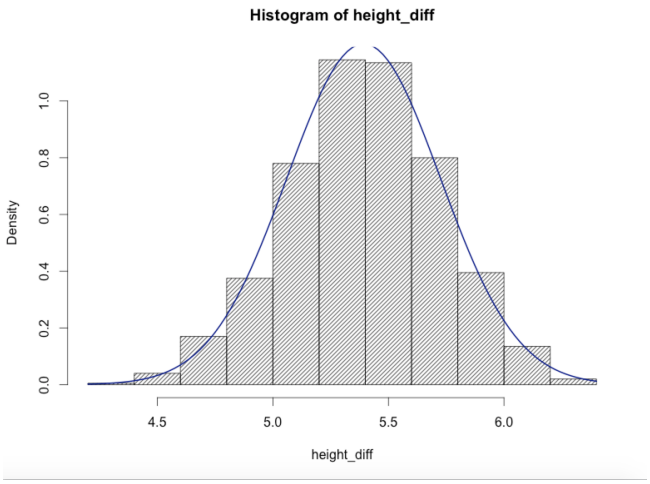
The plot is:



Problem 4

```
i = 1
height_diff = seq(1000) * 0
while (i <= 1000) {
  men_height <- rnorm(100, 69.1, 2)
  women_height <- rnorm(100, 63.7, 2.7)
  height_diff[i] = mean(men_height) - mean(women_height)
  i = i + 1
}

avg <- mean(height_diff)
std <- sd(height_diff)
hist(height_diff, density=20, freq=FALSE)
curve(dnorm(x, mean=avg, sd=std),
      col="darkblue", lwd=2, add=TRUE, yaxt="n")
```

The mean of the difference is 5.05 and the standard deviation is .79 The plot is:



Problem 5

Expectation is linear so:

$$\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

As to standard deviation:

$$\begin{aligned} \mathrm{Var}[X+Y] &= \mathrm{Var}[X] + \mathrm{Var}[Y] + \mathrm{Cov}[X,Y] \\ &= \mathrm{Var}[X] + \mathrm{Var}[Y] + \mathrm{Corr}[X,Y]\sigma_x\sigma_y \end{aligned}$$

Chapter 3

Linear Regression: The Basics

Problem 1

(a)

```
library(arm)
setwd('~/Code/workspace/Gelman/')
data <- read.table('exercise2.1.txt', header=TRUE)
fit.1 <- lm(y ~ x1 + x2, data=data)
summary(fit.1)
```

(b)

```
beta.hat <- coef(fit.1)
beta.sim <- sim(fit.1)

par(mfrow=c(1, 2))
plot(data$x1, data$y)
apply(coef(sim(fit.1)), 1, function(beta) {curve(cbind(1, x, mean(data$x2)) %*% beta, add
curve(cbind(1, x, mean(data$x2)) %*% coef(fit.1), add=TRUE)

plot(data$x2, data$y)
apply(coef(sim(fit.1)), 1, function(beta) {curve(cbind(1, mean(data$x1), x) %*% beta, add
curve(cbind(1, mean(data$x1), x) %*% coef(fit.1), add=TRUE)
```

(c)

```
par(mfrow=c(1,2))
plot(data$x1[1:40], fit.1$residuals, xlab='X1', ylab='Residuals')
```

```
plot(data$x2[1:40], fit.1$residuals, xlab='X2', ylab='Residuals')

plot(predict(fit.1, data[1:40, ]), fit.1$residuals)
```

For X1 it appears that all of the assumptions are met. However, for X2 we can see some heteroscedasticity so we could potentially improve our model with some additional feature or a variable transformation.

d)

```
predictions <- data.frame(predict(fit.1, data[41:dim(data)[1], ], level=.95, interval="pr
predictions
```

I feel pretty good about these predictions. the standard error seems to be about 1, so the 95% confidence interval contains ± 2 of our estimate.

Problem 2

(a)

We want to find the coefficients of the following equation:

$$\log(y) = \alpha + \beta \log(h)$$

We are given $\beta = .8$ now we just need to solve for alpha and get $\alpha = 6.957$

To calculate the standard deviation of the residuals. Notice that our 95% confidence interval is within a factor of 1.1 of our prediction. in other words it is $[x/1.1, 1.1x]$. Examining this on the log scale:

$$\begin{aligned} 2\hat{\sigma} &= \log(1.1x) \\ &= \log(x) + \log(1.1) \end{aligned}$$

Which means that $\hat{\sigma} = .047$

(b)

We can just use the equation for R^2

$$\begin{aligned} R^2 &= 1 - \frac{\hat{\sigma}^2}{\sigma^2} \\ &= 1 - \frac{.047}{.05} \\ &= .94 \end{aligned}$$

Problem 3

(a)

```
var1 <- rnorm(1000, 0 , 1)
var2 <- rnorm(1000, 0 , 1)
fit.3 <- lm(var1 ~ var2)
summary(fit.3)
```

No the p-value of the intercept is .226, indicating it does not meet the widely accepted .05 significance level.

(b)

```
z.scores.all <- rep(NA, 1000)
for (j in 1:1000) {
  z.scores <- rep(NA, 100)
  for (k in 1:100) {
    var1 <- rnorm(1000, 0, 1)
    var2 <- rnorm(1000, 0, 1)
    fit <- lm(var1 ~ var2)
    z.scores[k] <- coef(fit)[2] / se.coef(fit)[2]
  }
  z.scores.all[j] <- sum(z.scores >= 2)
}

z.scores.sorted <- sort(z.scores.all)
median(z.scores.sorted)
z.scores.sorted[50]
z.scores.sorted[950]
```

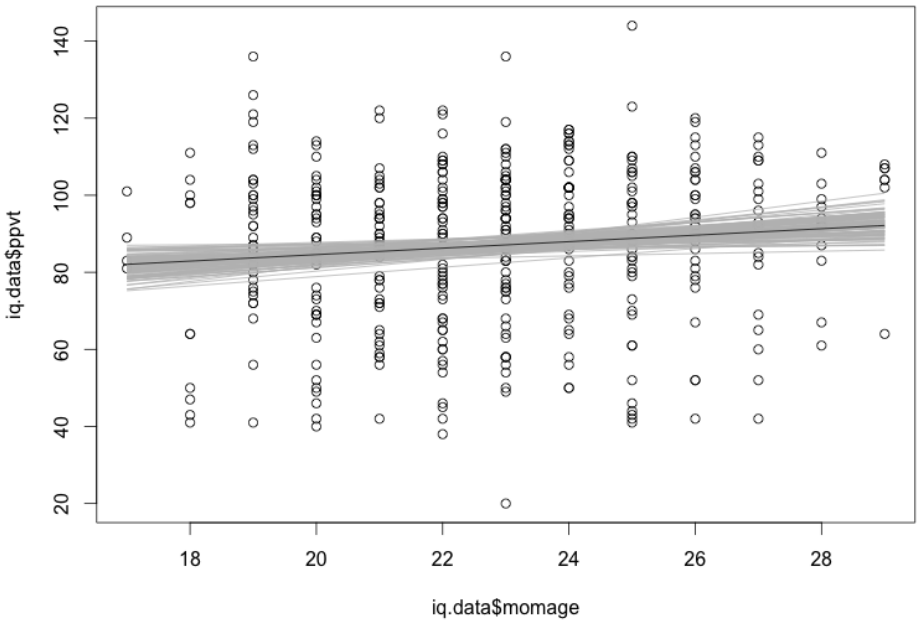
I ran the analysis for 1000 trials. I find that the median value is 2. I find that the 95% confidence interval is [0, 5]

Problem 4

(a)

```
Residuals:
    Min      1Q  Median      3Q     Max
-67.109 -11.798   2.971  14.860  55.210

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  67.7827     8.6880   7.802 5.42e-14 ***
momage        0.8403     0.3786   2.219    0.027 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 20.34 on 398 degrees of freedom
Multiple R-squared:  0.01223,Adjusted R-squared:  0.009743
F-statistic: 4.926 on 1 and 398 DF,  p-value: 0.02702
```

The slope coefficient implies that for each unit increase in momage the child's test score will increase by .8403 units. It looks like mothers should give birth later in life. In making these recommendations I'm assuming that momage is the only feature relevant to the test performance of the children.



(b)

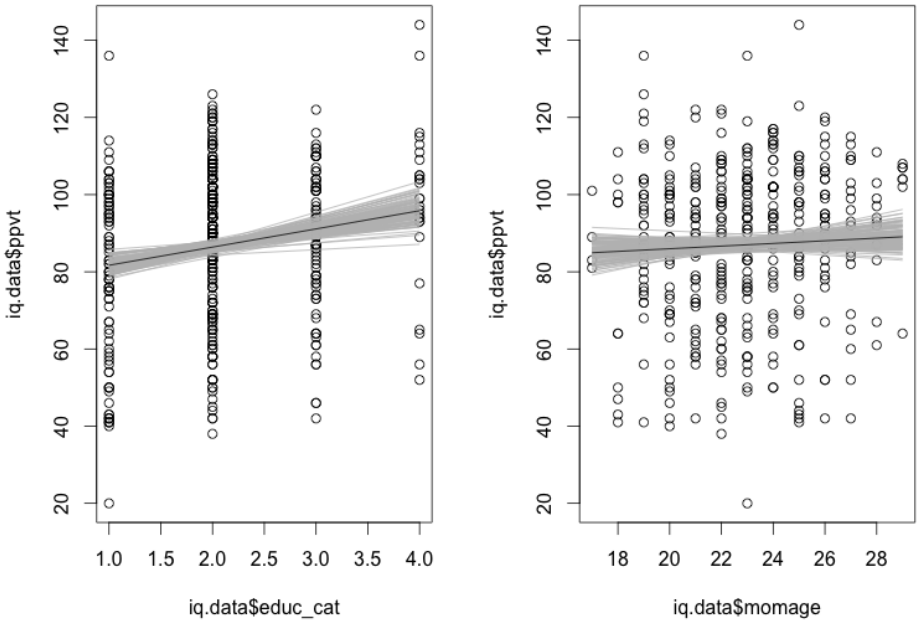
```
library("foreign")
iq.data <- read.dta("child.iq.dta")
fit.4 <- lm(ppvt ~ ., data=iq.data)
summary(fit.4)

Residuals:
    Min      1Q  Median      3Q     Max
-61.763 -13.130   2.495  14.620  55.610

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  69.1554     8.5706   8.069 8.51e-15 ***
educ_cat      4.7114     1.3165   3.579 0.000388 ***
momage        0.3433     0.3981   0.862 0.389003
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 20.05 on 397 degrees of freedom
Multiple R-squared:  0.04309,Adjusted R-squared:  0.03827
F-statistic: 8.939 on 2 and 397 DF,  p-value: 0.0001594
```

From the Summary statistics we can see that momage does not appear to contribute significantly to the model. To test this let's run ANOVA.

```
fit.4.noage <- lm(ppvt ~ educ_cat, data=iq.data)
anova(fit.4.noage, fit.4)

Analysis of Variance Table

Model 1: ppvt ~ educ_cat
Model 2: ppvt ~ educ_cat + momage
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    398 159816
2    397 159517  1    298.82 0.7437  0.389
```

These ANOVA results indicate that the addition of the momage variable does not add any predictive power. Thus we should change our conclusions in (a). This information suggests that it does not matter when a mother gives birth.

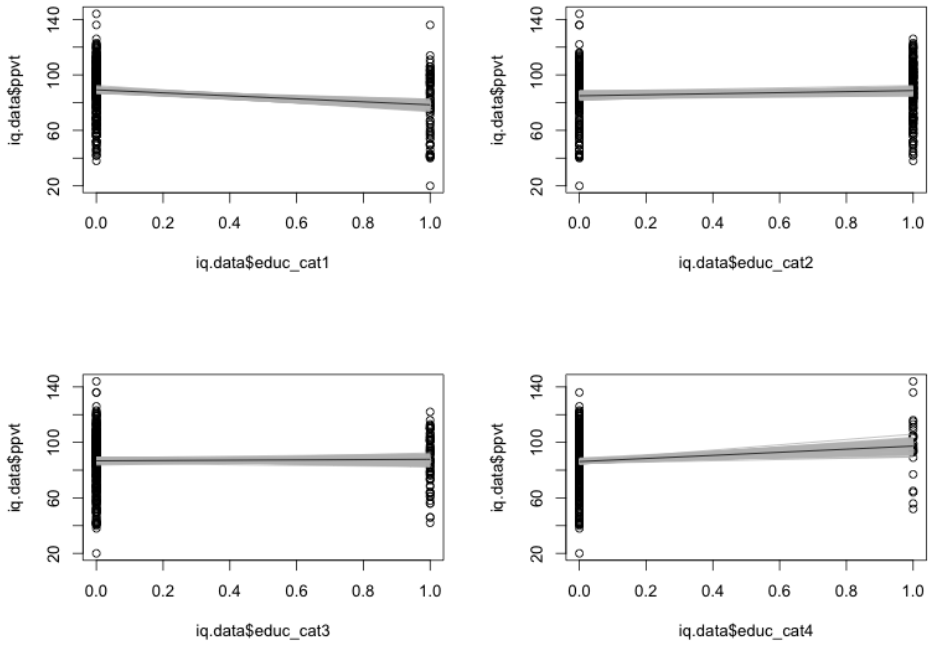
Next I plot the regression line for both variables.



(c)

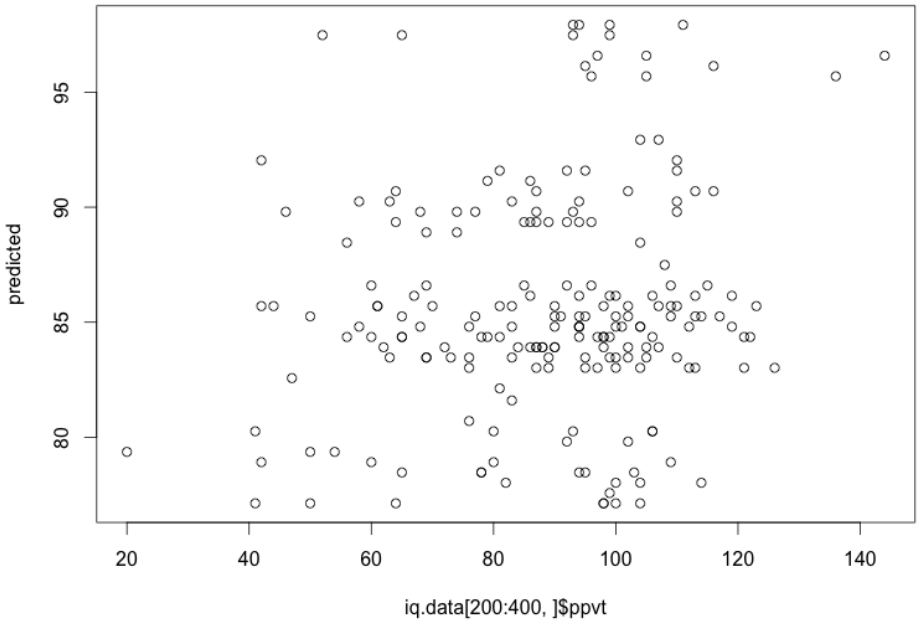
```
# Create Highschool completion factor
iq.data$educ_cat <- as.factor(iq.data$educ_cat)
temp <- model.matrix( ~ educ_cat - 1, data=iq.data )
iq.data <- cbind(iq.data, temp)
head(iq.data)

fit.4.c <- lm(ppvt ~ momage + momage*educ_cat2, data=iq.data)
summary(fit.4.c)

lm(formula = ppvt ~ momage + momage * educ_cat2, data = iq.data)

Residuals:
    Min      1Q  Median      3Q     Max
-65.041 -11.594   2.896  14.886  56.995

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       62.4546    11.8408   5.275  2.2e-07 ***
momage             0.9820     0.5132   1.914   0.0564 .
educ_cat2          9.6726    17.4080   0.556   0.5788
momage:educ_cat2  -0.2517     0.7587  -0.332   0.7402
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
Residual standard error: 20.29 on 396 degrees of freedom
Multiple R-squared:  0.02175,Adjusted R-squared:  0.01433
F-statistic: 2.934 on 3 and 396 DF,  p-value: 0.0333
```



(d)

```
par(mfrow=c(1, 1))
iq.data <- read.dta("child.iq.dta")
fit.4.d <- lm(ppvt ~ ., data=iq.data[1:200, ])
predicted = predict(fit.4.d, iq.data[200:400, ])
plot(iq.data[200:400, ]$ppvt, predicted)
```



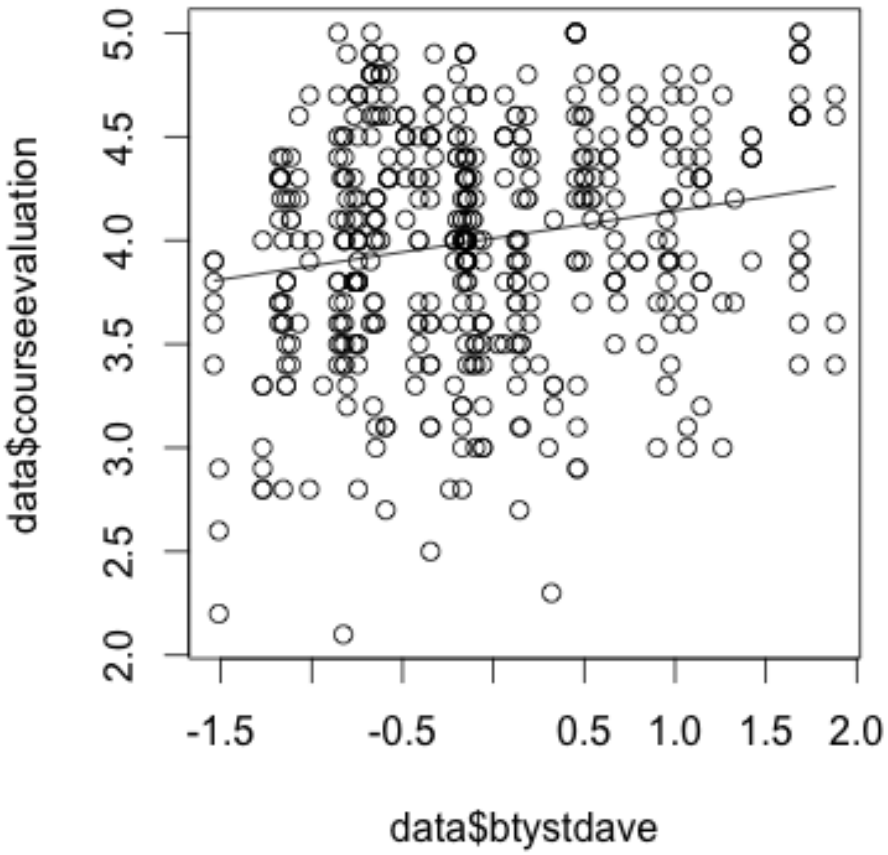
Problem 5

(a)

```
# Do more beautiful professors get better marks
par(mfrow=c(1, 1))
plot(data$btystdave, data$courseevaluation)

fit.5.beauty <- lm(courseevaluation ~ btystdave, data=data)
curve(fit.5.beauty$coef[1] + fit.5.beauty$coef[2] * x, add=TRUE)

# Add dotted lines to show +/- 1 standard deviation
curve (fit.5.beauty$coef[1] + fit.5.beauty$coef[2]*x + summary(fit.5.beauty)$sigma, lty=2
curve (fit.5.beauty$coef[1] + fit.5.beauty$coef[2]*x - summary(fit.5.beauty)$sigma, lty=2

# Look at male and female professors
fit.5.gender <- lm(courseevaluation ~ btystdave + female, data=data)

par(mfrow=c(1,2))
plot(data$btystdave[data$female == 0], data$courseevaluation[data$female == 0],
     xlab='beauty', ylab='average teaching evaluation', main='Men')
curve(fit.5.gender$coef[1] + fit.5.gender$coef[2] * x, add=TRUE)

plot(data$btystdave[data$female == 1], data$courseevaluation[data$female == 1],
     xlab='beauty', ylab='average teaching evaluation', main='Females')
```

```
curve(fit.5.gender$coef[1] + fit.5.gender$coef[2] * x + fit.5.gender$coef[3] * 1, add=TRU
```

Regressing on just beauty gives us an $R^2 = .035$ so the model fit isn't great. Below I plot the fit for beauty vs course evaluation.

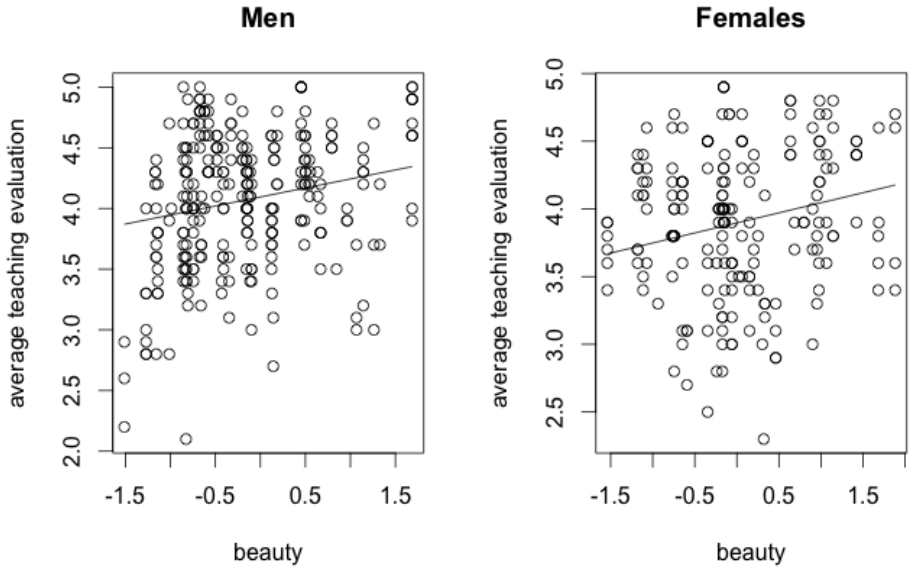


Now let's see if gender plays a role in the courseevaluations. We find that our R^2 has doubled and both beauty and gender seem to be statistically significant. We can interpret the coefficients as follows:

I- (Intercept) For a female teacher with 0 beauty we can expect a course rating between [4.02, 4.16] with 95% confidence

- (beauty) We can expect that for a female instructor an increase in beauty of 1 will result in an increase in the course evaluation by .148

- (female) A male teacher with 0 beauty will have coursevaluation reduced by .198 on average.


Men
Females
average teaching evaluation
beauty
2.0
2.5
3.0
3.5
4.0
4.5
5.0
-1.5
-0.5
0.5
1.5