140.615.HW.11.Jin.Vincent

Vincent Jin

2023-05-03

Homework 11

Vincent Jin

```
library(SPH.140.615)
copper <- copper
```

1.

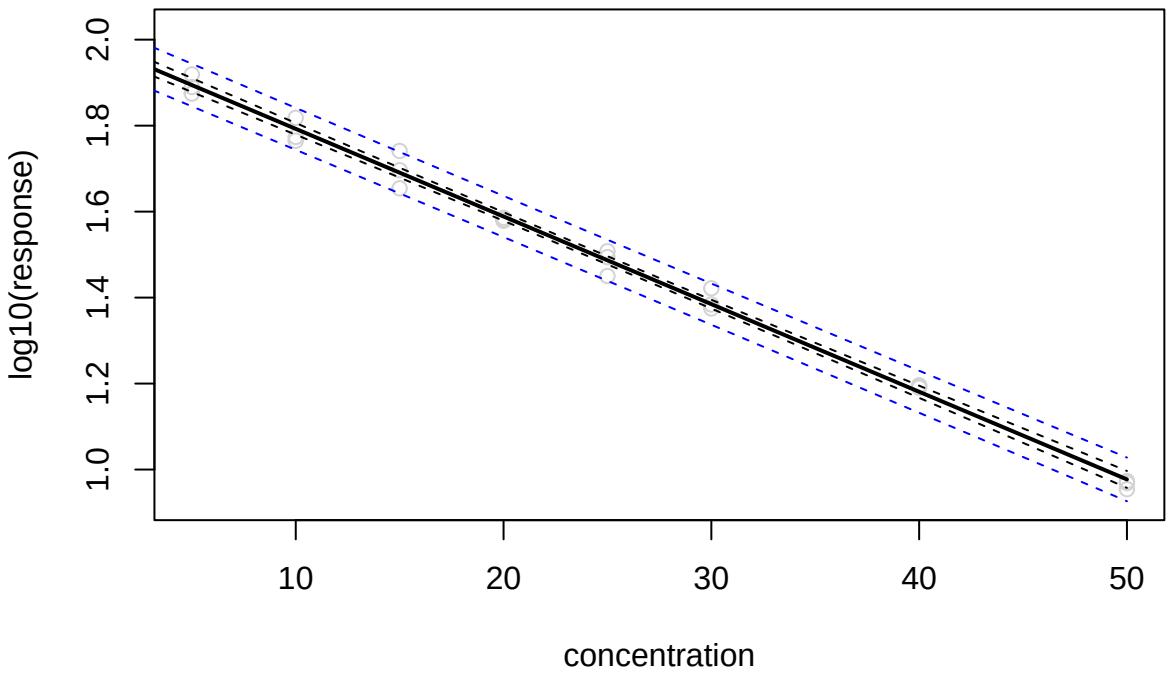
The percent transmittance for a sample with unknown copper concentration was measured to be 35.6%. Use your fitted calibration line from the previous homework to estimate the copper concentration in this sample. Calculate a 95% confidence interval for the copper concentration in this sample. Make a figure (including the calibration line) that visualizes your results.

Answer

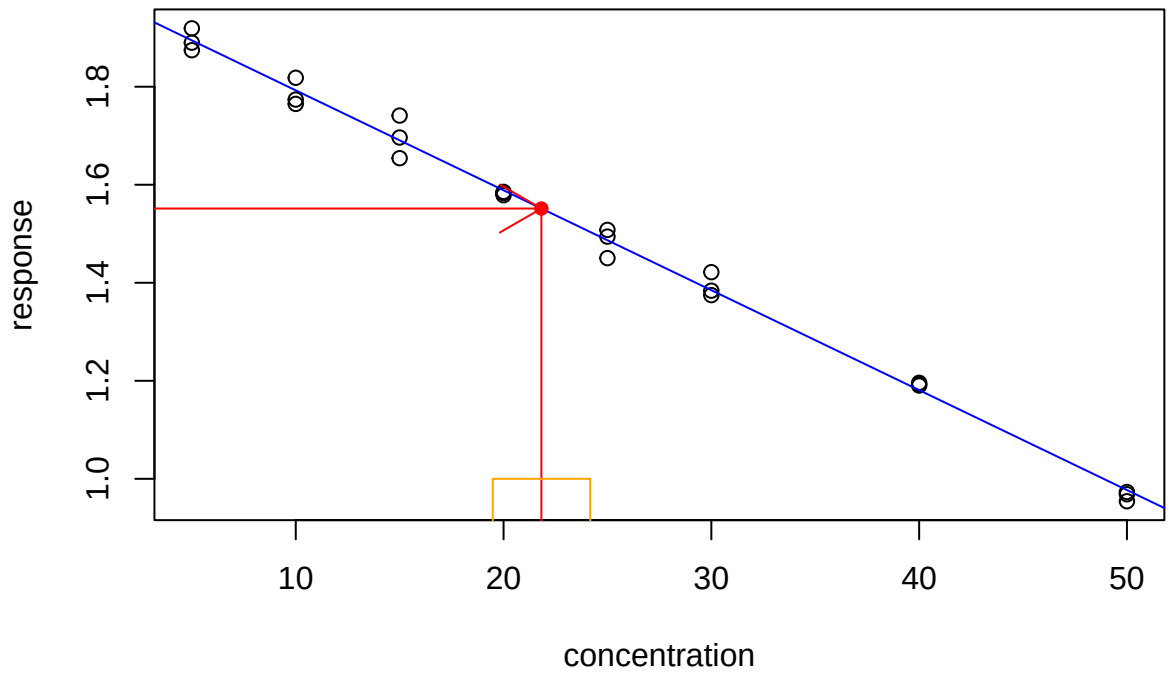
```
# regression line from previos homework
lm.fit <- lm(log10(response)~concentration, data = copper)

# Calibration
calibrate(copper$concentration, log10(copper$response), log10(35.6))
```

```
##      est       lo       hi
## 21.82474 19.48050 24.16898
```

```
# Visualization
xx <- seq(1, 50, by = 0.1)
predict.new <- predict(lm.fit, data.frame(concentration=xx), interval="prediction")
predict.mean <- predict(lm.fit, data.frame(concentration=xx), interval="confidence")
plot(copper$concentration, log10(copper$response), xlab="concentration", ylab="log10(response)", ylim=ra
lines(xx, predict.mean[,1], lwd=2)
lines(xx, predict.mean[,2], lty=2)
lines(xx, predict.mean[,3], lty=2)
lines(xx, predict.new[,2], lty=2, col="blue")
lines(xx, predict.new[,3], lty=2, col="blue")
```




```
# Visualization 2
plot(copper$concentration, log10(copper$response), xlab="concentration", ylab="response")
abline(lm.fit, col = "blue")
points(x = 21.82, y = log10(35.6), col = "red", pch = 16)
arrows(x0 = 21.82, y0 = log10(35.6), x1 = 21.82, y1 = 0, lty = 1, col = "red")
arrows(x0 = 0, y0 = log10(35.6), x1 = 21.82, y1 = log10(35.6), lty = 1, col = "red")
segments(x0 = 19.48, y0 = 1.0, x1 = 24.17, y1 = 1.0, col = "orange")
arrows(x0 = 19.48, y0 = 1.0, x1 = 19.48, y1 = 0, col = "orange")
arrows(x0 = 24.17, y0 = 1.0, x1 = 24.17, y1 = 0, col = "orange")
```



The copper concentration in this sample may be 21.82 with a confidence interval (19.48, 24.17).

2.

The data for this problem are in the dat.xy data frame in the SPH.140.615 package. Here, X is the predictor, and Y is the response.

(a)

Fit a regression line using the function lm, and provide estimates for the intercept β_0, the slope β_1, and the residual standard deviation σ.

Answer

```
xy <- dat.xy
lm.fit <- lm(Y ~ X, data = xy)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = Y ~ X, data = xy)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -129.65  -60.53  -13.03   58.57  247.90
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -227.309    121.862  -1.865   0.0685 .
## X             14.098      2.127   6.629 3.29e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 80.88 on 46 degrees of freedom
## Multiple R-squared:  0.4886, Adjusted R-squared:  0.4774
## F-statistic: 43.94 on 1 and 46 DF,  p-value: 3.29e-08
```

The β_0 is -227.31 and β_1 is 14.10. The residual standard deviation σ is 80.88.

(b)

Provide 95% confidence intervals for β_0 and β_1.

Answer

```
confint(lm.fit)
```

```
##                   2.5 %   97.5 %
## (Intercept) -472.603932 17.98570
## X              9.817299 18.37954
```

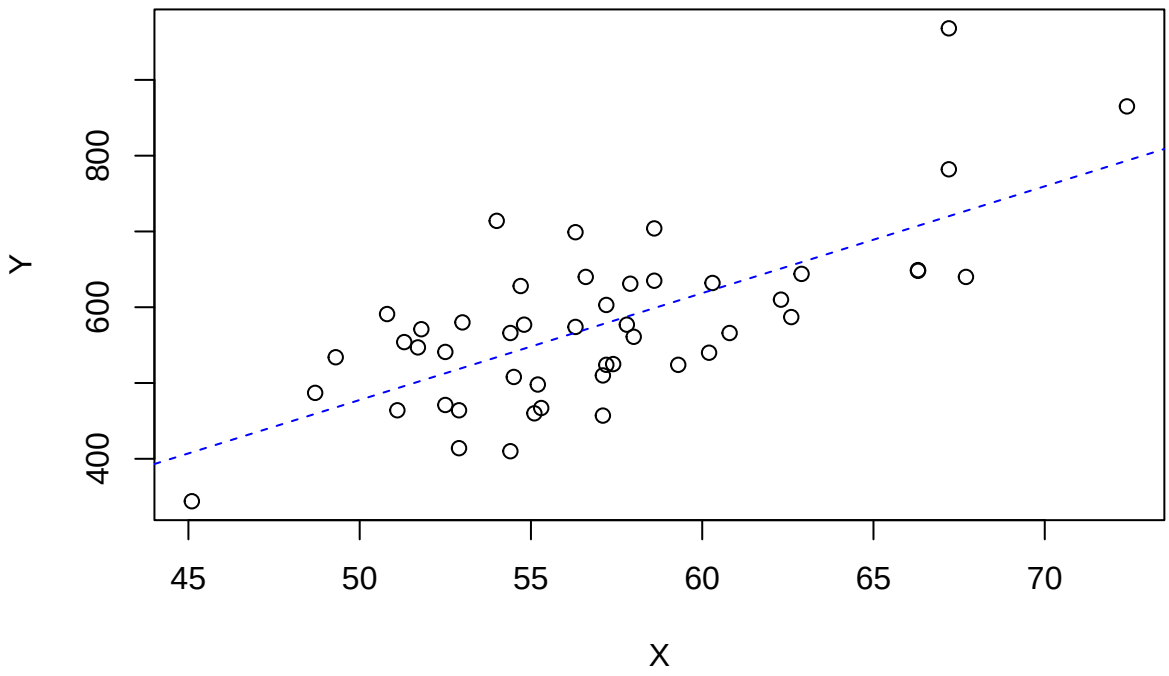
The confidence intervals for β_0 is (-472.60, 17.99), and for β_1 is (9.82, 18.38)

(c)

Plot the data, and show the regression line.

Answer

```
plot(xy$X, xy$Y, xlab="X", ylab="Y")
abline(lm(xy$Y ~ xy$X), col="blue", lty=2)
```



(d)

What is the expected response for X = 60? What is the expected response for X = 70? Provide 95% confidence intervals using the function predict.

Answer

```
predict(lm.fit, data.frame(X = c(60, 70)), interval = "confidence")
```

```
##        fit      lwr      upr
## 1 618.5961 591.8847 645.3076
## 2 759.5804 699.2997 819.8611
```

The expected response for X = 60 was 618.60 with a confidence interval of (591.88, 645.31) and for X = 70 was 759.58 with a confidence (699.30, 819.86).

(e)

For both X = 60 and X = 70, also provide 95% prediction intervals.

Answer

```
predict(lm.fit, data.frame(X = c(60, 70)), interval = "prediction")
```

```
##        fit      lwr      upr
## 1 618.5961 453.6148 783.5775
## 2 759.5804 585.9742 933.1866
```

For X = 60 and 70 the prediction intervals are: (453.615,783.578), (585.974, 933.187) respectively.

(f)

Comment on the lengths of the four intervals.

Answer

The prediction interval seems to be wider than the 95% confidence interval.

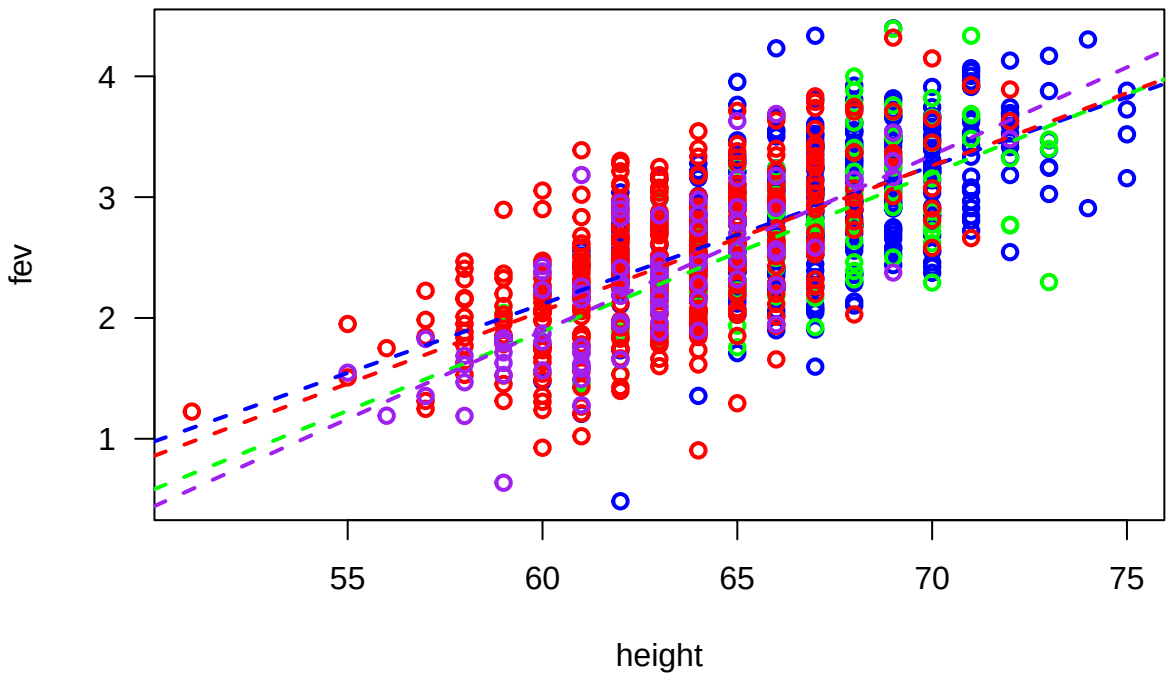
#3. The data for this problem are in the fev data frame in the SPH.140.615 package. The response is a measure of lung function (FEV, in liters). The other variables are the gender of a subject (male/female), the height (in inches), and whether or not that person is a smoker (yes/no). Analyze the data using linear regression (note: there are no interactions in these data, so you don't need to worry about them), and report your findings. In particular:

(a)

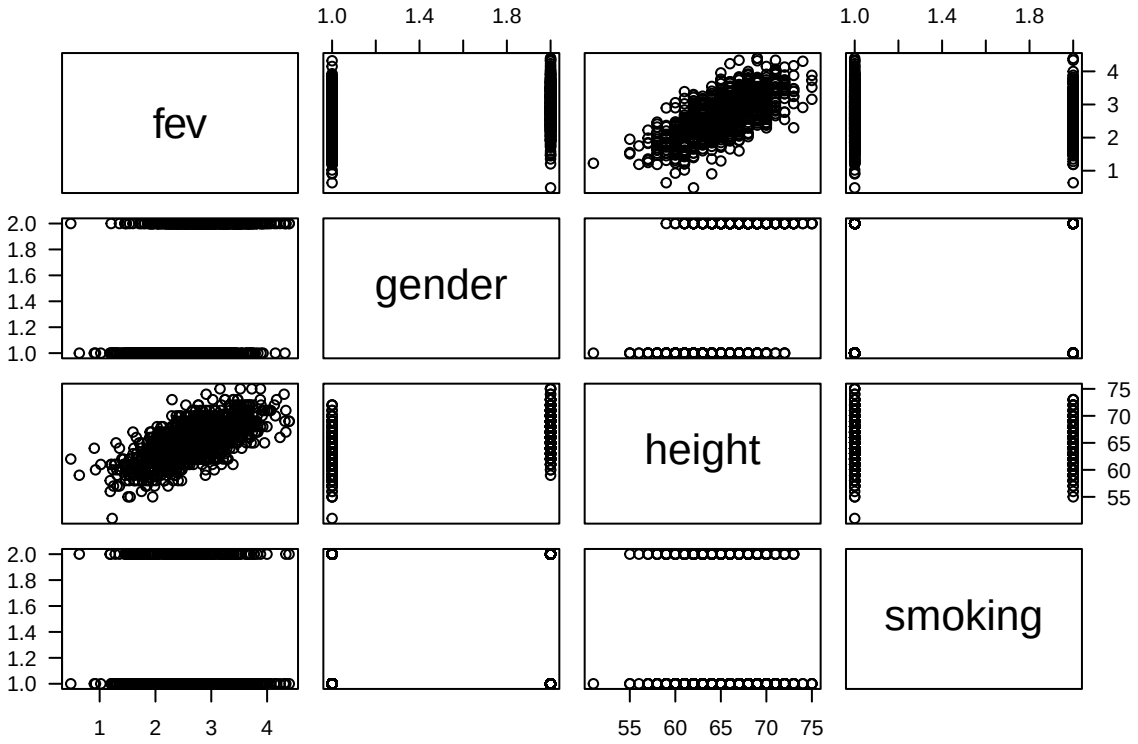
Explore the data by plotting them in a meaningful way, and comment.

Answer

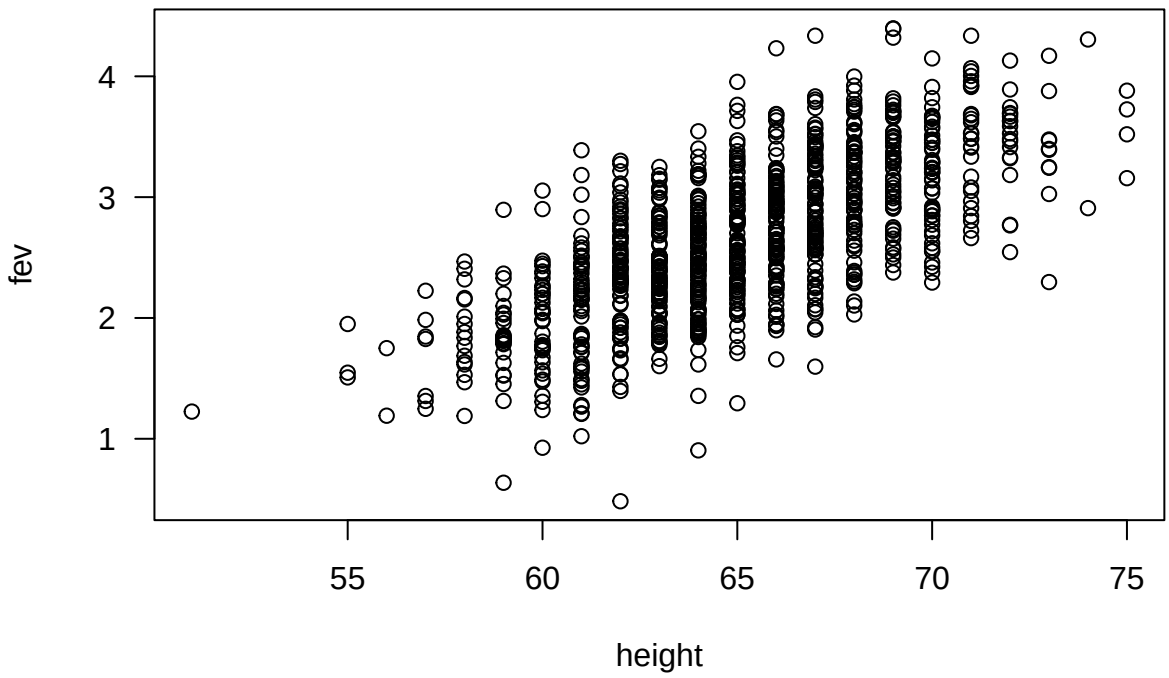
```
lm.outA <- lm(fev ~ height, data=fev, subset=(gender == 'M' & smoking == 'N'))
lm.outB <- lm(fev ~ height, data=fev, subset=(gender == 'M' & smoking == 'Y'))
lm.outC <- lm(fev ~ height, data=fev, subset=(gender == 'F' & smoking == 'N'))
lm.outD <- lm(fev ~ height, data=fev, subset=(gender == 'F' & smoking == 'Y'))
par(las=1)
plot(fev ~ height, data=fev, type="n", xlab="height", ylab="fev")
points(fev ~ height, data=fev, subset=(gender == 'M' & smoking == 'N'), col="blue", lwd=2)
points(fev ~ height, data=fev, subset=(gender == 'M' & smoking == 'Y'), col="green", lwd=2)
points(fev ~ height, data=fev, subset=(gender == 'F' & smoking == 'N'), col="red", lwd=2)
points(fev ~ height, data=fev, subset=(gender == 'F' & smoking == 'Y'), col="purple", lwd=2)
abline(lm.outA$coef, col="blue", lty=2, lwd=2)
abline(lm.outB$coef, col="green", lty=2, lwd=2)
abline(lm.outC$coef, col="red", lty=2, lwd=2)
abline(lm.outD$coef, col="purple", lty=2, lwd=2)
```



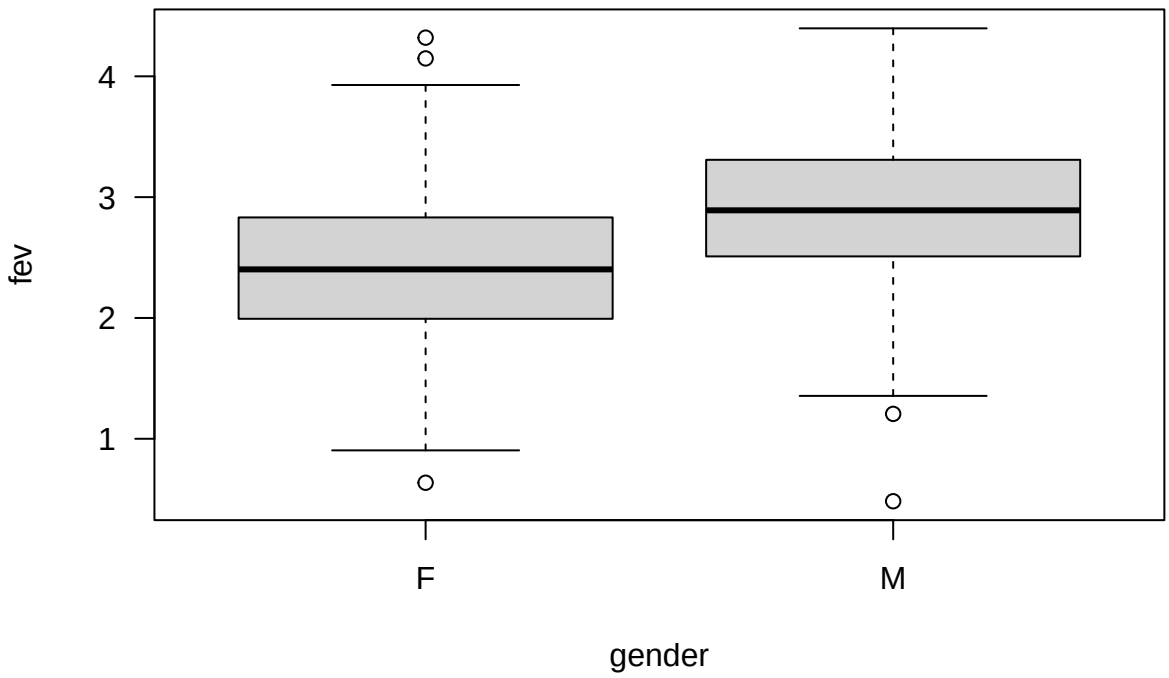

```
plot(fev)
```



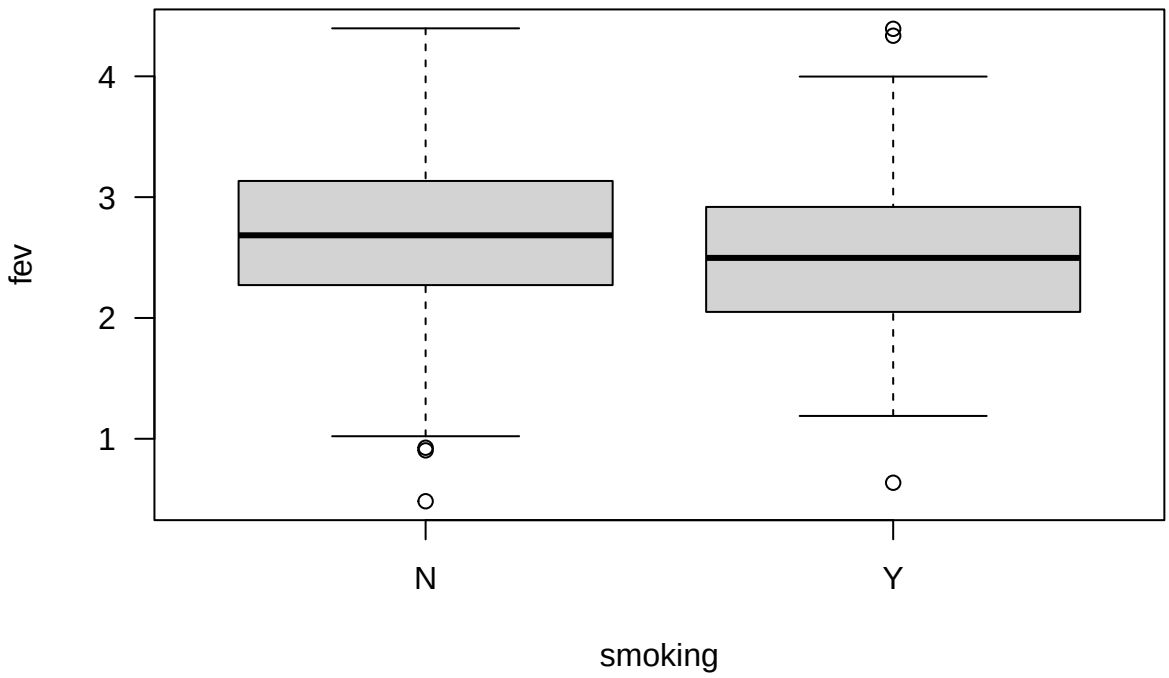

```
plot(fev ~ height, data = fev)
```




```
plot(fev ~ gender, data = fev)
```




```
plot(fev ~ smoking, data = fev)
```



Based on plots, we can see that there seems to be positive relationship between height and fev and gender and fev, and negative relationship between smoking and fev. Overall, the fev scores for male and female, smoker and smokers were overlapping with each other. Regardless the gender and smoking status combination, fev increases as height increases.

(b)

Fit your statistical models, and select the parameters for your final model.

```
lm.full <- lm(fev ~ gender + height + smoking, data = fev)
summary(lm.full)
```

```
##
## Call:
## lm(formula = fev ~ gender + height + smoking, data = fev)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.82118 -0.30188  0.00575  0.30685  1.44124
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.244425   0.312132 -16.802  < 2e-16 ***
## genderM      0.006640   0.034809   0.191  0.84875
## height       0.121645   0.004921  24.720  < 2e-16 ***
```

```
## smokingY    -0.106989   0.037457  -2.856  0.00437 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4585 on 996 degrees of freedom
## Multiple R-squared:  0.4762, Adjusted R-squared:  0.4747
## F-statistic: 301.9 on 3 and 996 DF,  p-value: < 2.2e-16
```

```
lm.reduced <- lm(fev ~ height + smoking, data = fev)
summary(lm.reduced)
```

```
##
## Call:
## lm(formula = fev ~ height + smoking, data = fev)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8162 -0.3042  0.0052  0.3090  1.4442
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.274963   0.267828 -19.695  < 2e-16 ***
## height       0.122163   0.004099  29.803  < 2e-16 ***
## smokingY    -0.106547   0.037367  -2.851  0.00444 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4583 on 997 degrees of freedom
## Multiple R-squared:  0.4762, Adjusted R-squared:  0.4752
## F-statistic: 453.2 on 2 and 997 DF,  p-value: < 2.2e-16
```

```
anova(lm.full, lm.reduced)
```

```
## Analysis of Variance Table
##
## Model 1: fev ~ gender + height + smoking
## Model 2: fev ~ height + smoking
##   Res.Df    RSS Df  Sum of Sq      F Pr(>F)
## 1    996 209.41
## 2    997 209.42 -1 -0.0076513 0.0364 0.8487
```

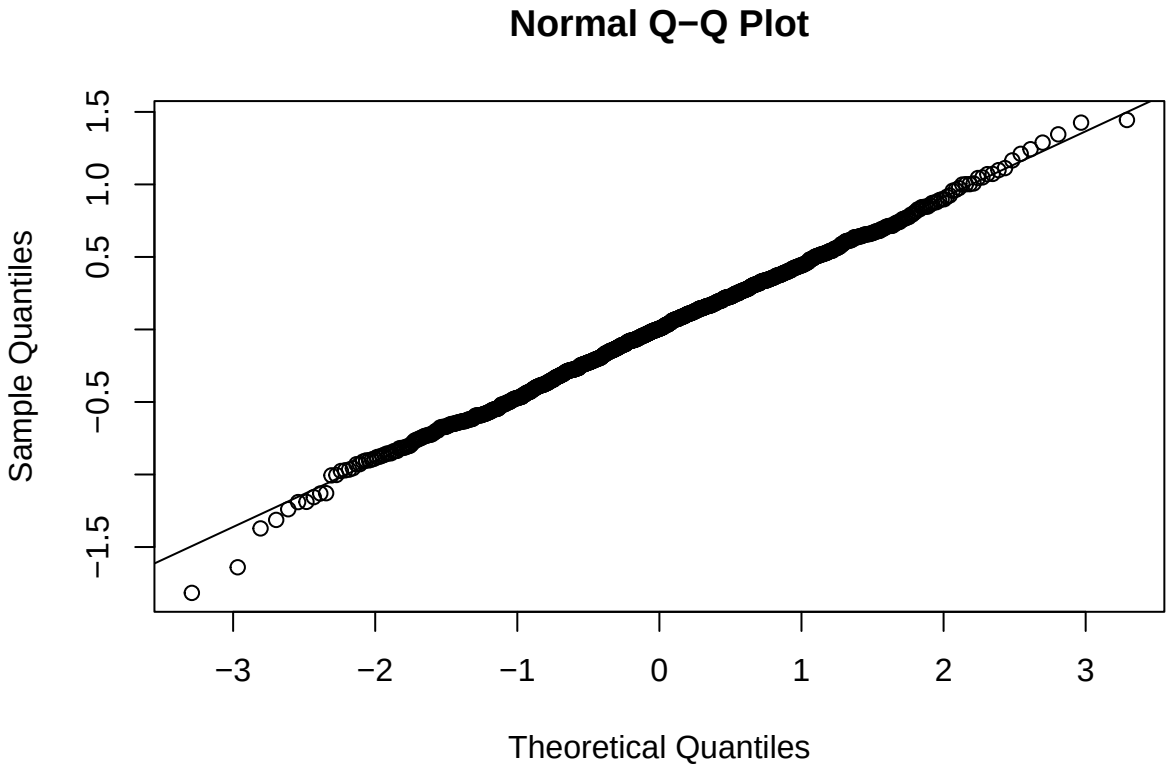
Answer

My final model is: $fev = \beta_0 + \beta_1 * height + \beta_2 * smoking$, where $\beta_1 = 0.12$ and $\beta_2 = -0.11$. Parameters selected are height and smoking. Gender was not selected as the coefficient in the full model was not significant, and also when testing the full model and reduced model, the p-value of 0.85 suggested that the coefficient for gender was not significantly different from 0.

(c)

Make a qq-plot of the residuals. Do they look normal?

```
qqnorm(lm.reduced$residuals)
qqline(lm.reduced$residuals)
```



Answer

The qq-plot suggested that the residuals are normal.

(d)

Report your findings, and give an interpretation of the parameter estimates for the variables that you used in your final model.

Answer

Based on the analysis, the model I fitted was: $fev = -5.27 + 0.12 * height - 0.11 * smoking$ Parameters selected in this final model are height and smoking. Gender was not selected as the coefficient in the full model was not significant, and also when testing the full model and reduced model, the p-value of 0.85 suggested that the coefficient for gender was not significantly different from 0.

Therefore we can interpret the parameter estimates is that: β_0: The expected measure of lung function in the form of FEV in liters among non-smokers who has a height of 0 inch is -5.27. However, the interpretation for β_0 is not much meaningful without centering for height as people always have some heights. β_1: The expected measure of lung function in the form of FEV in liters will increase 0.12 for every 1 inch increase in height, adjusting for smoking status (or holding the smoking status the same). β_2: The expected measure of lung function in the form of FEV in liters for smokers is 0.11 less than the lung function in the form of FEV in liters for non-smoker, adjusting for height of the person.

(e)

How much of the variability in FEV can you explain with your predictors?

```
summary(lm.reduced)$r.squared
```

```
## [1] 0.4762189
```

Answer

47.62% (0.4762 * 100%) of the variability in FEV can be explained with our predictors.