DSSH 6301 - HW 07 Solutions

```r
data <- data.frame(age=c(23, 18, 10, 45), iq=c(100, 105, 95, 120))
data
```

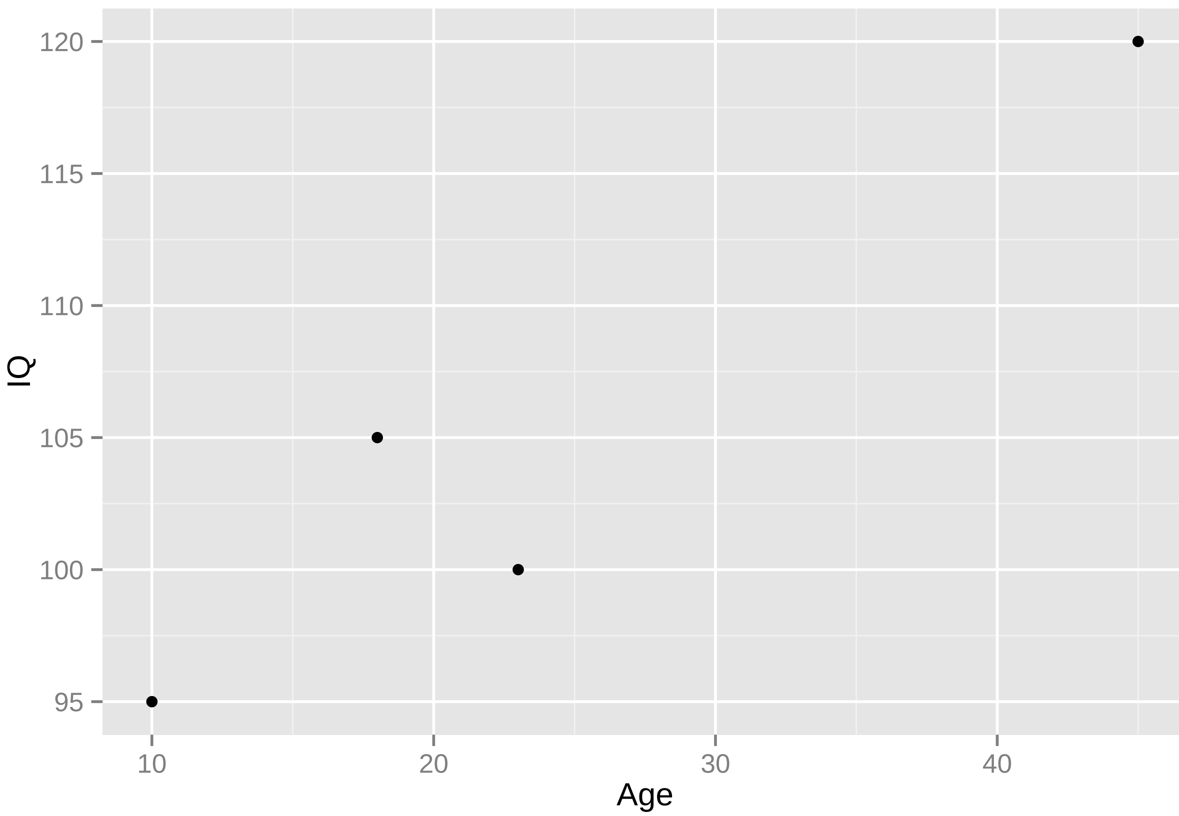
```
##   age  iq
## 1  23 100
## 2  18 105
## 3  10  95
## 4  45 120
```

Problem 1

```r
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```r
ggplot(data, aes(x=age, y=iq)) + geom_point() + xlab("Age") + ylab("IQ")
```



Problem 2

$$\mathrm{Cov}(x,y) = \frac{1}{(n-1)} \sum_i (x_i - \bar{x})(y_i - \bar{y})$$

$$n = 4$$

$$\bar{x} = 24$$

$$\bar{y} = 105$$

$$\text{Cov}(x, y) = \frac{1}{3}\sum_i (x_i - 24)(y_i - 105)$$

$$\text{Cov}(x, y) = \frac{1}{3}(-1 * -5 + -6 * 0 + -14 * -10 + 21 * 15) = 153.3333$$

```
cov(data$age, data$iq)
```

```
## [1] 153.3333
```

Problem 3

There is a high correlation coefficient for this data. This indicates the variables might be significantly related.

$$r = \frac{\text{Cov}(x, y)}{s_x s_y}$$

$$s^2 = \frac{1}{n-1}\sum_i^n (x_i - \bar{x})^2$$

$$s_x^2 = \frac{1}{3}(-1^2 + -6^2 + -14^2 + 21^2) = \frac{674}{3}$$

$$s_x = \sqrt{\frac{674}{3}} = 14.98888$$

$$s_y^2 = \frac{1}{3}(-5^2 + 0^2 + -10^2 + 15^2) = \frac{350}{3}$$

$$s_y = \sqrt{\frac{350}{3}} = 10.80123$$

$$r = \frac{153.3333}{14.98888 * 10.80123} = 0.9470957$$

```
cor(data$age, data$iq)
```

```
## [1] 0.9470957
```

Problem 4

$$\beta_1 = \frac{\text{Cov}(x,y)}{\text{Var}(x)}$$

$$\beta_1 = \frac{-1*-5+-6*0+-14*-10+21*15}{674} = 0.6824926$$

$$\beta_0 = \bar{y} - \beta_1\bar{x} = 105 - 0.6824926*24 = 88.62018$$

```
b1 <- cov(data$age, data$iq) / var(data$age)
b1
```

```
## [1] 0.6824926
```

```
b0 <- mean(data$iq) - mean(data$age)*b1
b0
```

```
## [1] 88.62018
```

$$y_i = \beta_0 + \beta_1 x_i = 88.62018 + 0.6824926x_i$$

Problem 5

$$y_1 = 88.62018 + 0.6824926x_1 = 88.62018 + 0.6824926*23 = 104.3175$$
$$y_2 = 88.62018 + 0.6824926x_2 = 88.62018 + 0.6824926*18 = 100.9050$$
$$y_3 = 88.62018 + 0.6824926x_3 = 88.62018 + 0.6824926*10 = 95.4451$$
$$y_4 = 88.62018 + 0.6824926x_4 = 88.62018 + 0.6824926*45 = 119.3323$$

```
y <- b0 + b1*data$age
y
```

```
## [1] 104.3175 100.9050  95.4451 119.3323
```

Problem 6

$$R^2 = r^2 = 0.9470957^2 = 0.8969903$$

```
tss <- sum((data$iq - mean(data$iq))^2)
tss
```

```
## [1] 350
```

```
sse <- sum((data$iq - y)^2)
sse
```

```
## [1] 36.05341
```

```r
r_sq <- (tss - sse) / tss
r_sq
```

```
## [1] 0.8969903
```

```r
cor(data$age, data$iq)^2
```

```
## [1] 0.8969903
```

In this bivariate regression, the R^2 term is the square of the correlation between x and y. Much of the variation is explained by our model.

Problem 7

$$se_{\hat{y}} = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}}$$

$$se_{\hat{y}} = \sqrt{\frac{-4.3175074^2 + 4.0949555^2 + -0.4451039^2 + 0.6676558^2}{2}}$$

$$se_{\hat{y}} = \sqrt{18.02671} = 4.245787$$

$$se_{b1} = se_{\hat{y}} \frac{1}{\sqrt{\sum(x_i - \bar{x})^2}}$$

$$se_{b1} = 4.245787 \frac{1}{\sqrt{674}} = 0.1635416$$

$$t = \frac{\beta_1}{se_{b1}} = \frac{0.6824926}{0.1635416} = 4.173205$$

```r
n <- length(data$iq)
df <- n - 2

qt(c(0.025, 0.975), df)
```

```
## [1] -4.302653  4.302653
```

The test statistic not in either of the rejection regions, therefore β_1 is not significant at the $\alpha = 0.05$ level.

```r
se_y <- sqrt(sum((data$iq - y)^2) / df)

se_b1 <- se_y / sqrt(sum((data$age - mean(data$age))^2))
se_b1
```

```
## [1] 0.1635416
```

```
t <- b1 / se_b1
t
```

```
## [1] 4.173205
```

Problem 8

```
p_val <- pt(t, df, lower.tail=F)*2
p_val
```

```
## [1] 0.05290431
```

β_1 is not significant at the $\alpha = 0.05$ level.

Problem 9

$$\text{CI} = \beta_1 \pm 4.302653 * se_{\beta 1} = 0.6824926 \pm 4.302653 * 0.1635416$$
$$\text{CI} = [-0.02117013, 1.38615529]$$

```
ci <- b1 + qt(c(0.025, 0.975), df)*se_b1
ci
```

```
## [1] -0.02117013  1.38615529
```

The CI fails to exclude 0 for β_1, and thus we cannot conclude that it is statistically significantly different from 0.

Problem 10

```
mod <- lm(data$iq ~ data$age)
summary(mod)
```

```
##
## Call:
## lm(formula = data$iq ~ data$age)
##
## Residuals:
##       1       2       3       4
## -4.3175  4.0950 -0.4451  0.6677
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  88.6202     4.4623  19.860  0.00253 **
## data$age      0.6825     0.1635   4.173  0.05290 .
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.246 on 2 degrees of freedom
## Multiple R-squared:  0.897,  Adjusted R-squared:  0.8455
## F-statistic: 17.42 on 1 and 2 DF,  p-value: 0.0529
```

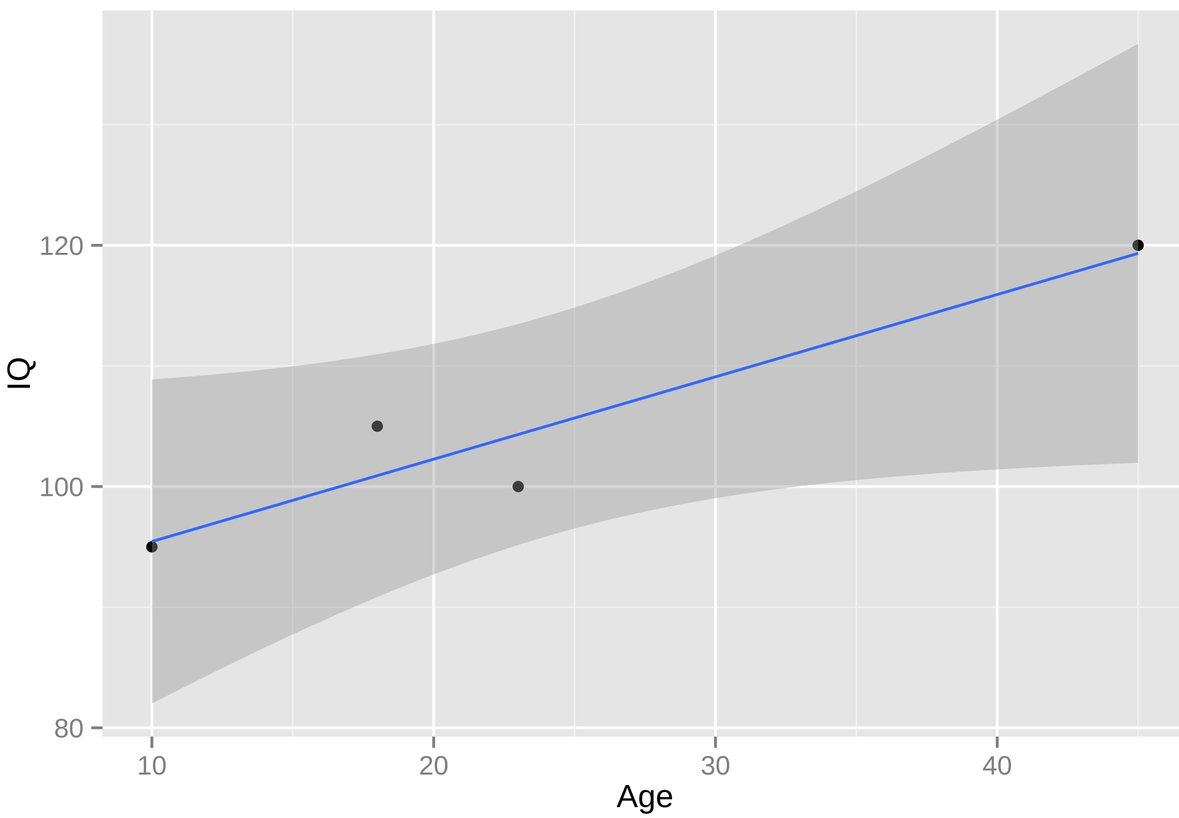
```r
confint(mod)
```

```
##                   2.5 %     97.5 %
## (Intercept) 69.42036991 107.819986
## data$age    -0.02117013   1.386155
```

All of the lm results closely match what we have calculated by hand.

Problem 11

```r
ggplot(data, aes(x=age, y=iq)) + geom_point() + xlab("Age") + ylab("IQ") +
  geom_smooth(method=lm)
```



Problem 12

Based on these data, we cannot conclude that there is a statistically significant relationship between age and IQ.