MATH 392 Problem Set 7

EJ Arce

26 March 2018

4

$$Var(X) = 5$$

$$Var(Y) = 7$$

$$Cov(X, Y) = 2$$

$$\begin{aligned} Var(2X - 5Y) &= Var(2X) + Var(5Y) - 2(2)(5)Cov(X, Y) \\ &= 4Var(X) + 25Var(Y) - 20Cov(X, Y) \\ &= 20 + 175 - 40 \\ &= 155 \end{aligned}$$

7

```
corrExerciseB <- corrExerciseB
```

a

```
cov <- cov(corrExerciseB$X, corrExerciseB$Y)
sigx <- sd(corrExerciseB$X)
sigy <- sd(corrExerciseB$Y)
rho <- cov/(sigx*sigy)
rho
```

```
## [1] 0.4996089
```

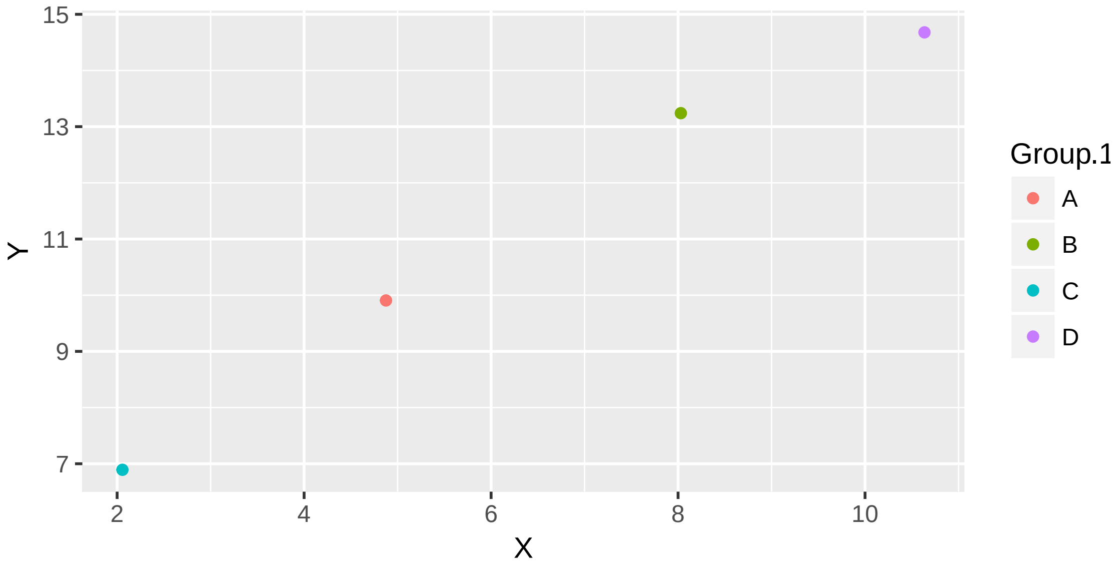
b

```
d <- aggregate(corrExerciseB[, 1:2], list(corrExerciseB$Z), mean)
d
```

```
##   Group.1         X         Y
## 1       A  4.875843  9.906436
## 2       B  8.029427 13.240133
## 3       C  2.056802  6.892128
## 4       D 10.635826 14.678636
```

c

```r
ggplot(d, aes(x = X, y = Y)) +
  geom_point(aes(color = Group.1))
```




```r
cor(d$X,d$Y)
```

```
## [1] 0.9921153
```

The correlation coefficient is much higher between means of X and Y than the correlation coefficient between each observation.

9

Show that $\sum_{i=1}^{n} y_i - \hat{y}_i = 0$.

$\sum_{i=1}^{n} y_i - \hat{y}_i = \sum_{i=1}^{n} y_i - (a + bx_i)$, where a and b are found using the function $g(a, b) = \sum_{i=1}^{n} (y_i - (a + bx_i))^2$, setting $\frac{\partial g}{\partial a} = 0$ and solving for a and b. Thus,

$$g(a, b) = \sum_{i=1}^{n} y_i^2 - 2y_i(a + bx_i) + a^2 + 2abx_i + (bx_i)^2$$

$$\frac{\partial g}{\partial a} = 0 = \sum_{i=1}^{n} -2y_i + 2a + 2bx_i$$

$$0 = 2\sum_{i=1}^{n} a + bx_i - y_i$$

$$0 = 2na + 2\sum_{i=1}^{n} bx_i - y_i$$

$$a = \frac{1}{n}\sum_{i=1}^{n} y_i - bx_i$$

Now plug a into the original equation:

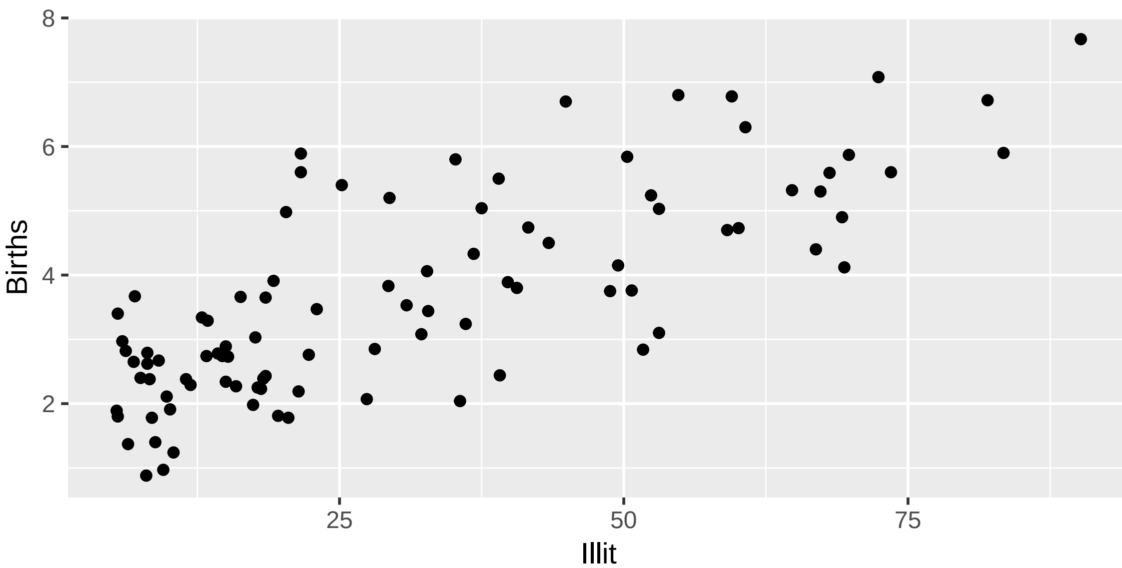
$$
\begin{aligned}
\sum_{i=1}^{n} y_i - (a + bx_i) &= -na + \sum_{i=1}^{n} y_i - bx_i \\
&= -n\frac{1}{n}\sum_{i=1}^{n} y_i - bx_i + \sum_{i=1}^{n} y_i - bx_i \\
\sum_{i=1}^{n} y_i - \hat{y}_i &= 0
\end{aligned}
$$

14

```
df <- Illiteracy
```

a

```
ggplot(df,aes(x=Illit, y = Births)) + geom_point()
```



At first glance, the scatterplot appears to show a positive relationship between female birth rate and illiteracy.

b

```
m1 <- lm(Births ~ Illit, data = df)
m1$coefficients
```

```
## (Intercept)       Illit
##  1.94873703  0.05452382
```

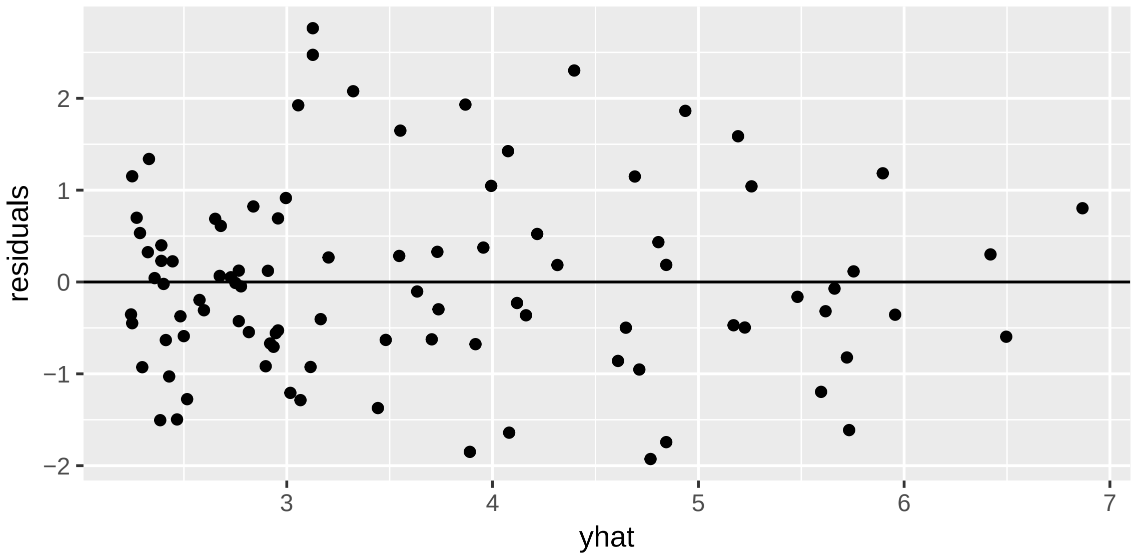
```
summary(m1)$r.squared
```

```
## [1] 0.5908428
```

The least-squares equation is $\hat{y} = 1.9488 + .0545x$. A 1% increase in the illteracy of women is associated with an increase of .0545 in the number of births per woman. With an r^2 of .5908, about 59% of the variance in the number of births per woman can be explaned by the variance in illteracy rate.

c

```
yhat <- predict(m1)
resid.d <- data.frame(residuals = m1$residuals,
                      y = yhat)
ggplot(resid.d, aes(x = yhat, y = residuals)) + geom_point() + geom_hline(yintercept = 0)
```



The residuals are randomly scattered when plotted against the predicted values of births. There is no clear linearity nor outliers, so this straight-line model is appropriate.

d

Although the scatterplot, regression model, and residual plot all imply a positive relationship between births and illiteracy rate, we cannot conclude that the variance in one varaible causes the variance in another variable. For example, it is possible that a third variable causes the variance in both births and illiteracy.