Problem Set 7 – MATH392

Theodore Dounias

3/25/2018

9.4

If I consider -5Y to be an rv of its own, I can apply the following identity:

$$Var[2X - 5Y] = Var[2X] + Var[-5Y] + 2Cov[2X, -5Y]$$

To find the covariance, I will use the following:

$$Cov[2X, -5Y] = E[-10XY] - E[2X]E[-5Y] = -10(E[XY] - E[X]E[Y]) = -10Cov[X, Y] = -20$$

Plugging in to the first equation I have:

$$Var[2X - 5Y] = 12 + 150 - 20 = 142$$

9.7

A.

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

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

B.

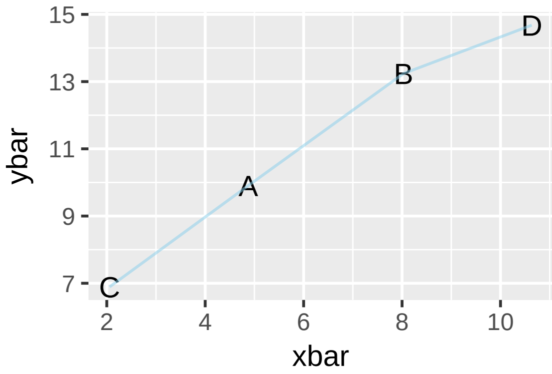
```r
data <- data %>%
  group_by(Z) %>%
  summarize(xbar = mean(X), ybar = mean(Y))

data %>%
  kable() %>%
  kable_styling()
```

Z	xbar	ybar
A	4.875843	9.906436
B	8.029427	13.240133
C	2.056802	6.892129
D	10.635826	14.678636

C.

```r
ggplot(data, aes(x = xbar, y = ybar)) +
  geom_text(aes(label = Z)) +
  geom_line(alpha = .5, col = "skyblue")
```



9.9

I will use the fact that $\beta_1 = \bar{y} - \beta_2\bar{x}$ and $\hat{y} = \beta_1 + \beta_2 x$:

$$\sum_{i=1}^{n}(\hat{y} - y_i) = \sum_{i=1}^{n}(y_i - (\beta_1 + \beta_2 x_i)) = -\beta_1 n + \sum_{i=1}^{n}(y_i - \beta_2 x_i) = -\beta_1 n + n\sum_{i=1}^{n}(\frac{y_i}{n} - \beta_2\frac{x_i}{n})$$

By using the aforementioned equations I have:

$$\sum_{i=1}^{n}(\hat{y} - y_i) = -n(\bar{y} - \beta_2\bar{x}) + n(\bar{y} - \beta_2\bar{x}) = 0$$

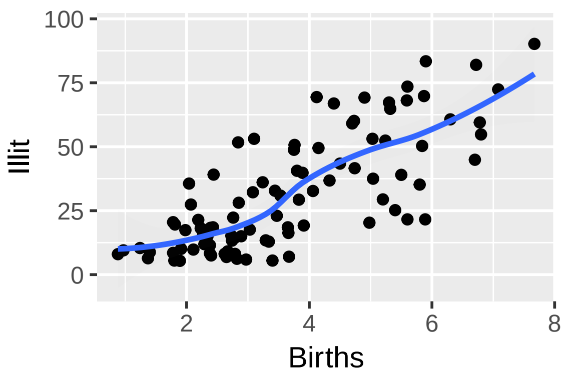
9.14

A.

```
data <- Illiteracy


ggplot(data, aes(y = Illit, x = Births)) +
  geom_point() +
  geom_smooth(alpha = .02)
```

```
## `geom_smooth()` using method = 'loess'
```



The relationship appears to be linear; higher birth rates and higher illiteracy increase simultaneously.

B.

```r
lm1 <- lm(data = data, Illit ~ Births)

tidy(lm1)
```

```
##          term  estimate std.error statistic      p.value
## 1 (Intercept) -8.239759 3.7508638 -2.196763 3.054992e-02
## 2      Births 10.836417 0.9401581 11.526165 1.502910e-19
```

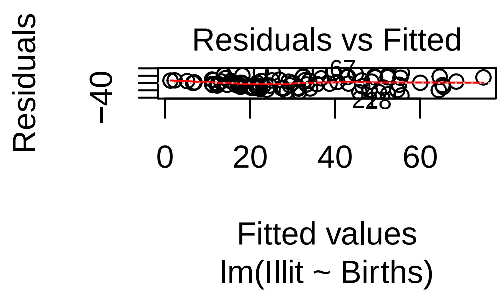
```r
glance(lm1)[1]
```

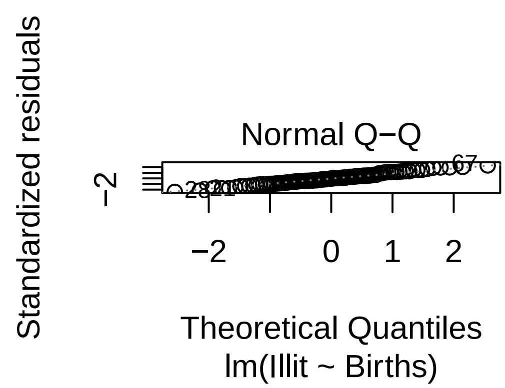
```
##   r.squared
## 1 0.5908428
```

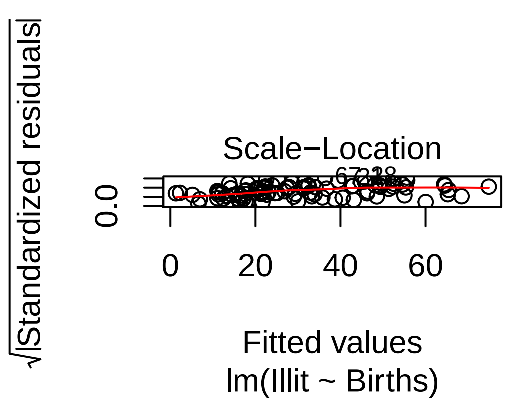
The slope coefficient is statisticaly significant, positive, and equal to 10.83; this means that for an increase of one birth, there is approximatelly a 10.83 increase in illiteracy rates. The residual sum of squares is about .6, which signifies the error rate that the output line has in predicting the data.

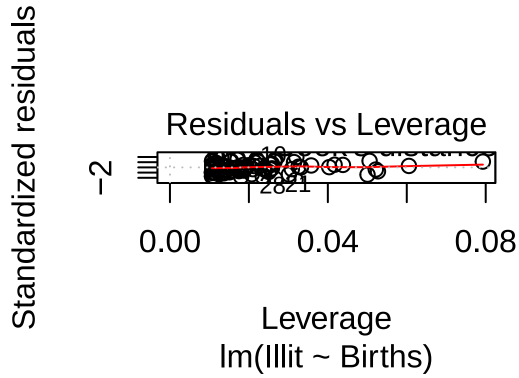
C.

```r
plot(lm1)
```









If we consider these diagnostic plots, it is obvious that the residuals present an increase in variance as the values of illiteracy themselves increase. This might suggest that using a log value for the response variable might be appropriate, but the linear model is a fine approximation.

D. Not necessarily. To consider policy solutions we would have to run a wider array of models that include other variables. What we are seeing in this simple model might be an example of collinearity, a relationship that exists due to an underlying cause. In that case, we should try to find the true root of the issue.