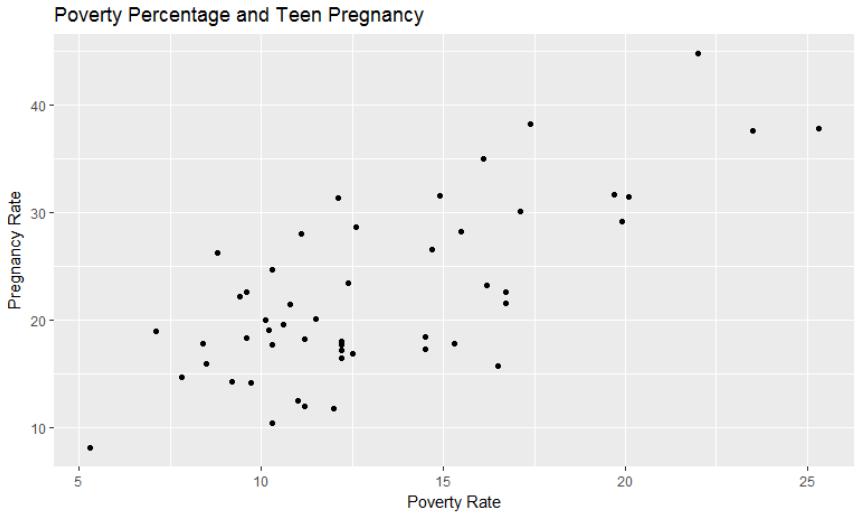
Solutions STA 101 Homework 01

Dr. Erin K. Melcon

1. (a) The plot follows:



There does appear to be a linear relationship, however it appears as if there are outliers present as well.

(b) The estimated correlation is: 0.7303 . This value is pretty close to 1, so a strong relationship is suggested.

(c) The five number summary is:
Min: 8.1, Q_1 : 17.25, median : 20, Q_3: 28.1, Max: 44.8.

(d) Yes - there is a state that has unusally small values in both the response and the explanatory variable. There may also be a state with unusually high poverty rate.

(e) Since the range of birth rate is between 8.1 and 44.8, this is the most appropriate range to predict the response variable with.

---

2. (a) The estimated regression line is: $\hat{y} = 4.2673 + (1.3733)X_1$

(b) When the poverty percentage increases by 1%, the birth rate (per 1000) tends to increase by 1.3733 on average for 15 to 17 year olds.

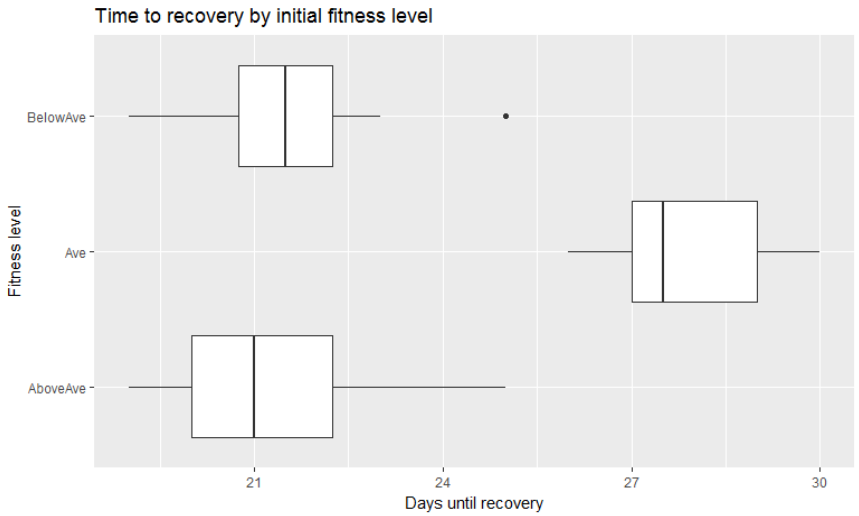
(c) Since it is very unlikely to have a poverty of 0 per 1000, and it is not within the range of our explanatory variable, it is not appropriate to predict at $X_1 = 0$.

(d) The predicted value is: $y^* = 4.2673 + (1.3733)10 = 18.0007467$

(e) From R, $SSR = 1725.2594896$, and $SSTO = 3234.8941176$. Thus, $R^2 = \frac{1725.2594896}{3234.8941176} = 0.5333$.
53.33% of the variance in birth rates (per 1000) of 15 to 17 year olds is explained by the linear relationship with poverty rate.

---

3. (a) The plot follows:



It seems like there is a difference in the average time to recovery - specifically the below average group and the above average group seem to recover faster than the average group.

(b) Yes - the below average group has an unusally large observation.

(c) There will be two betas, since there are three categories. Say β_1, with corresponding $X_1 = 1$ if the level is `Ave`, and β_2, with the corresponding $X_2 = 1$ if the level is `BelowAve`.

(d) The five number summary is:
Min: 19, Q_1 : 21, median : 22.5, Q_3: 27, Max: 30.

(e) Yes, since days is a numeric variable it is possible that there is an unusually high or low value.

---

4. (a) The estimated regression function is: $\hat{y} = 21.4167 + (6.3333)X_1 + (0.0833)X_2$

(b) The average number of days until recovery when the group was of `Ave` fitness was : 6.3333

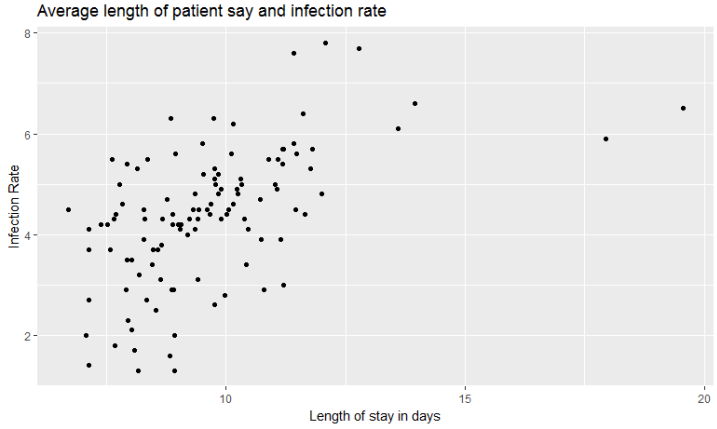
(c) The average number of days until recovery when the group was of `BelowAve` fitness was : 0.0833

(d) The average number of days until recovery when the group was of `AboveAve` fitness was : 21.4167

(e) The three models are:
For `AboveAve` Fitness: $\hat{y} = 21.4167$
For `Ave` Fitness: $\hat{y} = 27.75$
For `BelowAve` Fitness: $\hat{y} = 21.5$

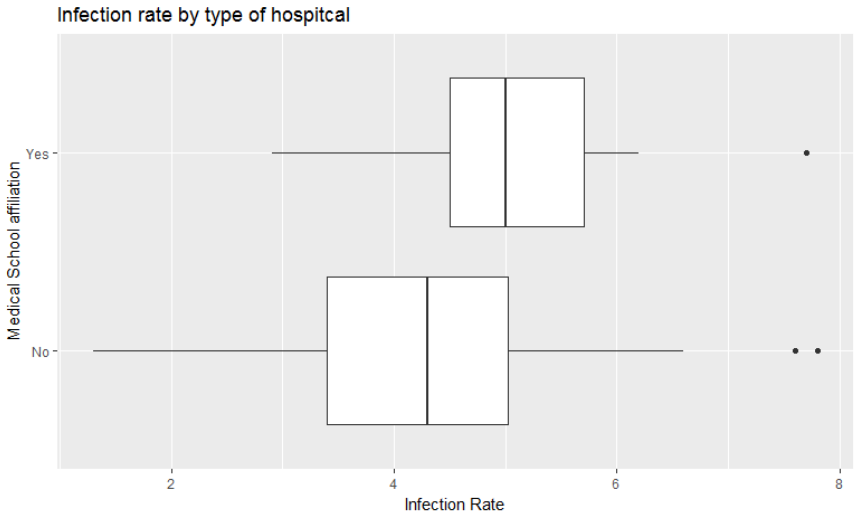
---

5. (a) The plot follows:



There does appear to be a linear relationship; specifically, a positive linear relationship. However, there are clearly outliers present.

(b) The plot follows:



There may be a difference in the groups; specifically that teaching schools seem to have a higher rate of infection.

(c) The estimated correlation is: 0.5334

(d) The outliers are arguably: Length of stay over 15, or infection risk over 7. If we say the outliers are length of stay over 15, the corresponding rows of the dataset are:

row	InfctRsk	MedSchool	Stay
47	6.50	No	19.56
112	5.90	Yes	17.94

If we say the outliers are those with infection rate above 7, we have:

row	InfctRsk	MedSchool	Stay
13	7.70	Yes	12.78
53	7.60	No	11.41
54	7.80	No	12.07

(e) Exluding the outliers, a reasonable range appears to be approximately 5 to 15.

---

6. (a) The estimated regression function with no interaction term is:
$\hat{y} = 0.8628 + (0.3572)X_1 + (0.3056)X_2$

(b) The simplified models are:
For Medical School Affiliation: $\hat{y} = 1.1684 + (0.3572)X_1$
For No Medical School Affiliation: $\hat{y} = 0.8628 + (0.3572)X_1$

(c) For the medical school affiliation model : When the average length of stay of patients increases by one day, the infection risk tends to increase by 0.3572 on average.
For the no medical school affiliation model : When the average length of stay of patients increases by one day, the infection risk tends to increase by 0.3572 on average.

(d) A patient could reasonably stay 0 days (out patient procedures, for example), although in this dataset the minimum was 6.7 days. Therefore, you could argue either way. If you argued that you can interpret them:
For the medical school affiliation model : When the average length of stay of patients is 0, the infection risk is 1.1684 on average.
For the medical school affiliation model : When the average length of stay of patients is 0, the infection risk is 0.8628 on average.

(e) The predicted value is: $y^* = 3.3629964$.

7. (a) The estimated regression function with an interaction term is:
$\hat{y} = 0.6368 + (0.3812)X_1 + (1.7132)X_2 + (-0.1315)X_1X_2$

   (b) The simplified models are:
For Medical School Affiliation: $\hat{y} = 2.35 + (0.2497)X_1$
For No Medical School Affiliation: $\hat{y} = 0.6368 + (0.3812)X_1$

   (c) For the medical school affiliation model : When the average length of stay of patients increases by one day, the infection risk tends to increase by 0.2497 on average.
For the no medical school affiliation model : When the average length of stay of patients increases by one day, the infection risk tends to increase by 0.3812 on average.

   (d) A patient could reasonably stay 0 days (out patient procedures, for example), although in this dataset the minimum was 6.7 days. Therefore, you could argue either way. If you argued that you can interpret them:
For the medical school affiliation model : When the average length of stay of patients is 0, the infection risk is 2.35 on average.
For the medical school affiliation model : When the average length of stay of patients is 0, the infection risk is 0.6368 on average.

   (e) The predicted value is: $y^* = 3.3050921$.

---

8. (a) FALSE. It only measures linear relationships.

   (b) FALSE. In general, it will allow for a different slope and intercept for each line, and will change your overall prediction.

   (c) TRUE. The sign in the estimated correlation should agree with the slope, or else they would be suggesting different types of linear relationships for the same varaible.

   (d) TRUE. Since an error is defined as $e_i = y_i - \hat{y}_i$, when you have a positive error that means $y_i > \hat{y}_i$, in other words we underestimated the true data point.

```r
#Problem 1
library(ggplot2)
poverty = read.csv("C:/GitHub/Teaching-Materials/STA-101/STA-101-2017-Spring/Datasets/HW-2/poverty.csv")
#Part (a)
#qplot(PovPct,Brth15to17 , data = poverty) +  ggtitle("Poverty Percentage and Teen Pregnancy") +
  #ylab("Pregnancy Rate") + xlab("Poverty Rate")
#Part (b)
r = round(cor(poverty$PovPct,poverty$Brth15to17),4)
#Part (c)
fn = fivenum(poverty$Brth15to17)
#Problem 2
#Part (a)
lin.model = lm( Brth15to17 ~  PovPct, data = poverty)
the.betas = round(lin.model$coefficients,4)
#part(d)
xs = data.frame(PovPct = 10)
ys = predict(lin.model,xs, SE = FALSE)
#Part(e)
anova.table = anova(lin.model)
SSR = anova.table[1,2]
SSTO = sum(anova.table[,2])
R2 = round(SSR/SSTO,4)
#Problem 3
rehab = read.csv("C:/GitHub/Teaching-Materials/STA-101/STA-101-2017-Spring/Datasets/HW-2/rehab.csv")
#Part (a)
#ggplot(rehab,aes(y = days,x =fitness)) + geom_boxplot() + ylab("Days until recovery") +
  #xlab("Fitness level") + ggtitle("Time to recovery by initial fitness level") + coord_flip()
#Part (d)
fn = fivenum(rehab$days)
#Problem 4
#Part (a)
lin.model = lm( days ~  fitness, data = rehab)
the.betas = round(lin.model$coefficients,4)
#Problem 5
hospital = read.csv("C:/GitHub/Teaching-Materials/STA-101/STA-101-2017-Spring/Datasets/HW-2/hospital.csv")
#qplot(Stay, InfctRsk, data = hospital) +  ggtitle("Average length of patient say and infection rate") +
  #xlab("Length of stay in days") + ylab("Infection Rate")
#ggplot(hospital,aes(y = InfctRsk,x =MedSchool)) + geom_boxplot() + ylab("Infection Rate") +
  #xlab("Medical School affiliation") + ggtitle("Infection rate by type of hospitcal") + coord_flip()
#(a)
r = round(cor(hospital$InfctRsk,hospital$Stay),4)
#(b)
outliers1 = hospital[which(hospital$Stay > 15),]
outliers2 = hospital[which(hospital$InfctRsk > 7),]
#Problem 6
#(a)
model1 = lm(InfctRsk ~ Stay + MedSchool, data = hospital)
the.betas = round(model1$coefficients,4)
#(e)
xs = data.frame(Stay = 7, MedSchool= "No")
ys = predict(model1, xs)
#Problem 7
#(a)
model2 = lm(InfctRsk ~ Stay + MedSchool + Stay*MedSchool, data = hospital)
the.betas = round(model2$coefficients,4)
#(e)
```

```
xs = data.frame(Stay = 7, MedSchool= "No")
ys = predict(model2, xs)
```