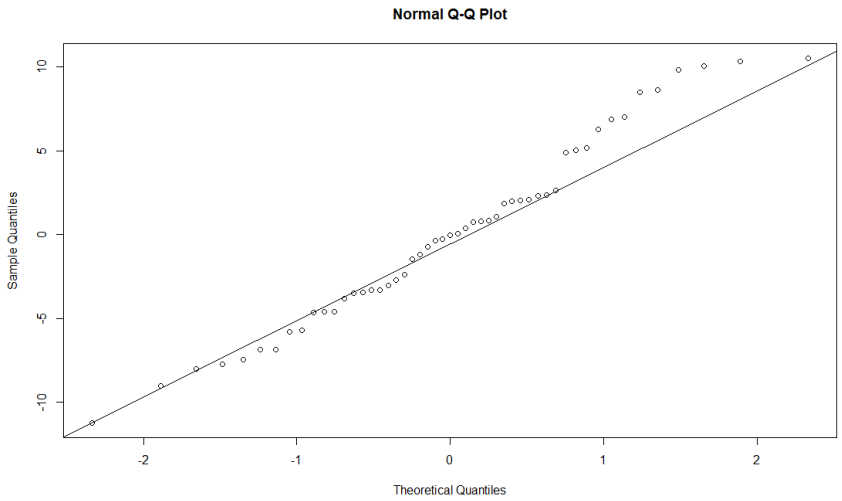
Solutions STA 101 Homework 02

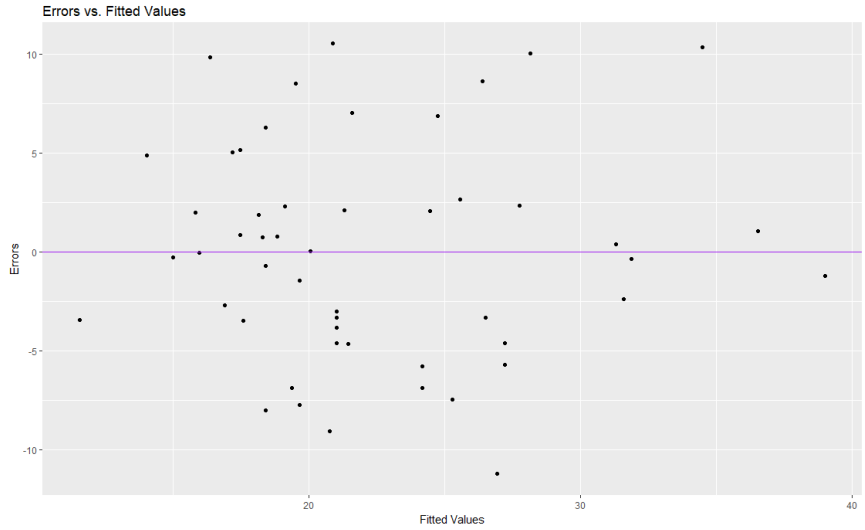
Dr. Erin K. Melcon

1. (a) The plot follows:



The p-value for the Shapiro-Wilks test is: 0.3323856. Since this is relatively large, we fail to reject the null hypothesis and conclude that the data is approximately normal.

(b) The plot follows:



The p-value for the Fligner-Killeen test is: 0.2942374. Since this is relatively large p-value, we fail to reject the null, and conclude the variances are the same in both the upper and lower groups.

(c) There does appear to be at least two outliers - one that is a very small value, and one that is very high. They correspond to the rows:

	Location	PovPct	Brth15to17	ei	yhat	Group
30	New Hampshire	5.30	8.10	-3.45	11.55	Upper
32	New Mexico	25.30	37.80	-1.21	39.01	Lower

(d) The slope ($\hat{\beta}_1$) before was: 1.3733454, and the slope after was: 1.3566767. Thus, the absolute difference is: $|1.3733454 - 1.3566767| = 0.0166687$.

---

2. (a) The confidence intervals follow:

	5 %	95 %
(Intercept)	-0.2584	9.4175
PovPct (X_1)	1.0001	1.7132

(b) We are 90% confident that when the percentage of people in poverty increases by 1%, we expect the birth rate per 1000 females 15 to 17 years old to increase by between 1.0001 and 1.7132 on average, holding all other variables constant.

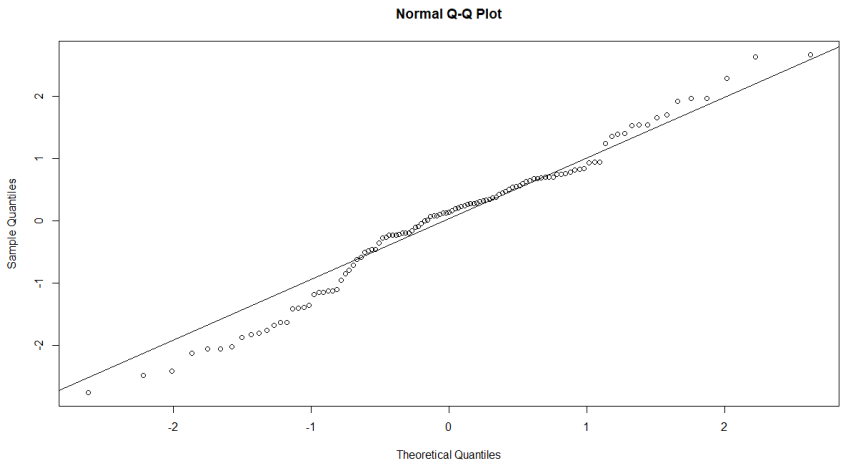
(c) Since this interval does not contain zero, it does suggest that poverty percentage has a linear relationship with the birth rate per 1000 females 15 to 17 years old.

(d) The test-statistic is: 6.384832 with corresponding p-value 0.0000001.

(e) If in reality there was no linear relationship between percentage of people in poverty and the birth rate, we would see our data or more extreme (more linearly related) with probability: 0.0000001

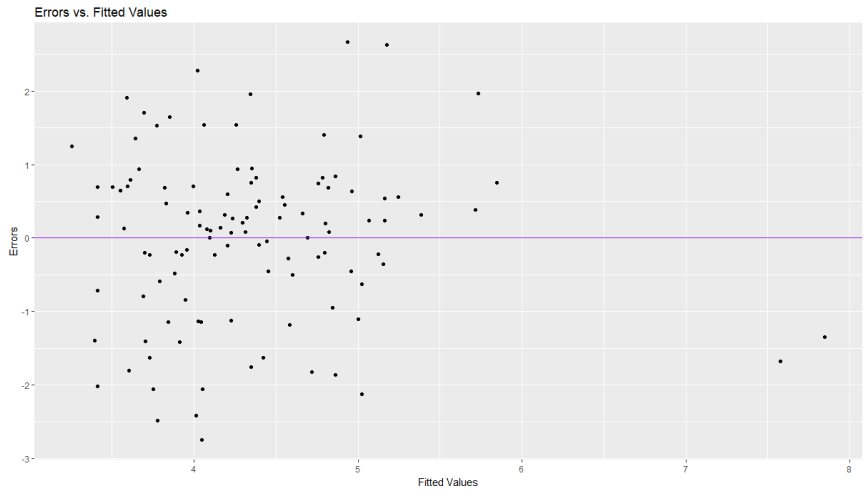
---

3. (a) The plot follows:



The p-value for the Shapiro-Wilks test is: 0.0860007. Since this is relatively small, we reject the null hypothesis and conclude that the data is not approximately normal.

(b) The plot follows:



The p-value for the Fligner-Killeen test is: 0.065866. Since this is a relatively small p-value, we reject the null, and conclude the variances are not the same in both the upper and lower groups.

(c) There are two clear outliers which have a fitted value above 7, so we will remove those. These correspond to rows:

	InfctRsk	MedSchool	Stay	ei	yhat	Group
47	6.50	No	19.56	-1.35	7.85	Lower
112	5.90	Yes	17.94	-1.68	7.58	Lower

(d) For the two slopes (not the intercept) we have:

	Old $\hat{\beta}_i$	New $\hat{\beta}_i$	Absolute Difference
Stay	0.3572	0.4776	0.1204
MedSchoolYes	0.3056	0.2453	0.0603

---

4. (a) The confidence intervals follow:

	5 %	95 %
(Intercept)	-1.4273	0.9414
Stay	0.3519	0.6032
MedSchoolYes	-0.2799	0.7705

(b) We are 90% confident that when the hospital is affiliated with medical school, there is no affect on the infection rate (since the interval covers zero), holding all other variables constant.

(c) Since this interval does contain zero, it does not suggest that the type of school has an effect on infection rate.

(d) The test-statistic is: 0.7748948 with corresponding p-value 0.4400951.

(e) If in reality there was no relationship between infection rate and affiliation with medical school, we would see our data or more extreme (more linearly related) with probability: 0

---

5. (a) The model with an interaction term is: $\hat{y} = -0.1031 + (0.4625)X_1 + (-1.7873)X_2 + (0.1943)X_1X_2$

(b) The confidence intervals are:

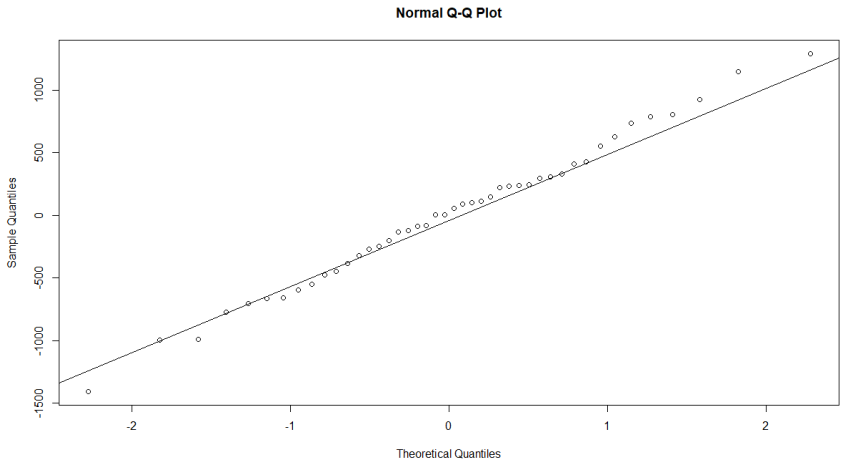
	2.5 %	97.5 %
(Intercept)	-1.5786	1.3724
Stay	0.3058	0.6192
MedSchoolYes	-7.7128	4.1382
Stay:MedSchoolYes	-0.3690	0.7576

(c) Based on the above confidence intervals, only those where the confidence intervals do not contain zero should remain, so that would be the intercept, and β_1.

(d) Looking at only the confidence intervals that do not contain zero (the significant ones), the largest change we could expect is from X_1, and it is the upper bound of the confidence interval: 0.6192.

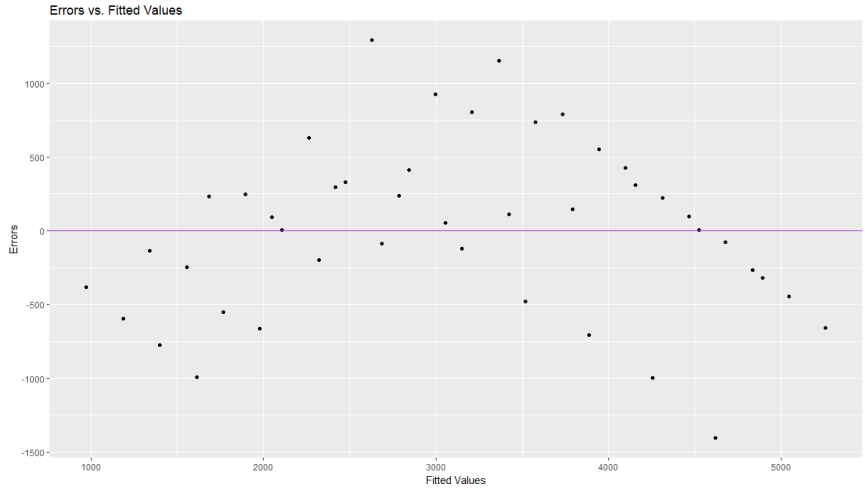
---

6. (a) The plot follows:



The p-value for the Shapiro-Wilks test is: 0.9979569. Since this is relatively large, we fail to reject the null hypothesis and conclude that the data is approximately normal.

(b) The plot follows:



The p-value for the Fligner-Killeen test is: 0.1128873. Since this is relatively large p-value, we fail to reject the null, and conclude the variances are the same in both the upper and lower groups. However, there is a clear pattern in this plot!!

(c) There does not appear to be any significant outliers.

(d) Since there were no ouliers, there is no difference to compare.

---

7. (a) The confidence interavals are:

	5 %	95 %
(Intercept)	1970.5646	5837.9674
Age	22.7822	29.6992
Temp	-174.4891	-38.3382

(b) We are 90% confident that when the age of the fish increases by 1 day, the length of the fish increases by between 22.7822 and 29.6992 mm on average, holding all other variables constant.

(c) Since this interval does not contain zero, it does suggest the age of the fish has an effect on the length of the fish.

(d) The test-statistic is: 6.3057294 with corresponding p-value 0.

(e) If in reality there was no relationship between the age of the fish and the length of the fish, we would see our data or more extreme (more linearly related) with probability: 0

8. (a) The model with an interaction term is: $\hat{y} = -174.2399 + (75.2719)X_1 + (39.2473)X_2 + (-1.7511)X_1X_2$

   (b) The confidence intervals are:

	5 %	95 %
(Intercept)	-4128.3490	3779.8693
Age	33.2555	117.2883
Temp	-101.5227	180.0173
Age:Temp	-3.2469	-0.2553

   (c) Based on the above confidence intervals, only those where the confidence intervals do not contain zero should remain. However, when an interaction term does not contain zero, both single terms that are involved should remain in the model. Because of this, all β's should remain.

   (d) Looking at only the confidence intervals that do not contain zero (the significant ones), the largest change we could expect is from X_1, and it is the upper bound of the confidence interval: 117.2883.

---

9. (a) TRUE. Each X will explain some small amount of variance in Y, so that the SSE will be lowest for the largest model, and thus R^2 will be highest.

   (b) FALSE. This would mean that a slope of zero, indicating no change in Y when X changes, is a plausible value.

   (c) TRUE. They tend to significantly change the slopes of the regression line, since they are dragged toward the outlier.

   (d) TRUE. Since AIC uses the negative log-likelihood, we want to minimize AIC.

```r
#Problem 1
library(xtable)
library(ggplot2)
poverty = read.csv("C:/GitHub/Teaching-Materials/STA-101/STA-101-2017-Spring/Datasets/HW-3/poverty.csv")
#Part (a)
the.model = lm( Brth15to17 ~  PovPct, data = poverty)
poverty$ei = the.model$residuals
poverty$yhat = the.model$fitted.values
#qqnorm(lin.model$residuals)
#qqline(lin.model$residuals)
ei = the.model$residuals
the.SWtest = shapiro.test(ei)
#Part (b)
#qplot(yhat, ei, data = poverty) +  ggtitle("Errors vs. Fitted Values") + xlab("Fitted Values") +
  #ylab("Errors") + geom_hline(yintercept = 0,col = "purple")
Group = rep("Lower",nrow(poverty)) #Creates a vector that repeats "Lower" n times
Group[poverty$Brth15to17 < median(poverty$Brth15to17)] = "Upper" #Changing the appropriate values to "Uppe
Group = as.factor(Group) #Changes it to a factor, which R recognizes as a grouping variable.
poverty$Group = Group
the.FKtest= fligner.test(poverty$ei, poverty$Group)
#Part (c)
Outliers = which(poverty$PovPct > 24 | poverty$PovPct < 6)
new.poverty = poverty[-Outliers,]
new.model = lm( Brth15to17 ~  PovPct, data = new.poverty)
options(scipen = 10)
#Problem 2
#Part (a)
alpha = 0.10
all.CIs = confint(new.model,level = 1-alpha)
#Part (d)
CIb1 = round(all.CIs[2,],4)
all.tests = summary(new.model)$coefficients
#Problem 3
hospital = read.csv("C:/GitHub/Teaching-Materials/STA-101/STA-101-2017-Spring/Datasets/HW-3/hospital.csv")
#Part (a)
the.model = lm(InfctRsk ~ Stay + MedSchool, data = hospital)
hospital$ei = the.model$residuals
hospital$yhat = the.model$fitted.values
#qqnorm(the.model$residuals)
#qqline(the.model$residuals)
ei = the.model$residuals
the.SWtest = shapiro.test(ei)
#Part (b)
#qplot(yhat, ei, data = hospital) +  ggtitle("Errors vs. Fitted Values") + xlab("Fitted Values") +
  #ylab("Errors") + geom_hline(yintercept = 0,col = "purple")
Group = rep("Lower",nrow(hospital)) #Creates a vector that repeats "Lower" n times
Group[hospital$InfctRsk < median(hospital$InfctRsk)] = "Upper" #Changing the appropriate values to "Upper"
Group = as.factor(Group) #Changes it to a factor, which R recognizes as a grouping variable.
hospital$Group = Group
the.FKtest= fligner.test(hospital$ei, hospital$Group)
#Part (c)
Outliers = which(hospital$yhat > 7)
#xtable(hospital[Outliers,])
new.hospital = hospital[-Outliers,]
new.model = lm(InfctRsk ~ Stay + MedSchool, data = new.hospital)
betas = round(the.model$coefficients[-1],4)
```

```
new.betas = round(new.model$coefficients[-1],4)
abs.diff = abs(betas-new.betas)
#xtable(cbind(betas,new.betas,abs.diff))
#Problem 2
#Part (a)
alpha = 0.10
all.CIs = confint(new.model,level = 1-alpha)
#Part (d)
CIb1 = round(all.CIs[2,],4)
all.tests = summary(new.model)$coefficients
# Problem 5
# Part (a)
I.model = lm(InfctRsk ~ Stay + MedSchool + Stay*MedSchool, data = new.hospital)
betas = round(I.model$coefficients,4)
all.CIs = round(confint(I.model),4)
#Problem 6
fish = read.csv("C:/GitHub/Teaching-Materials/STA-101/STA-101-2017-Spring/Datasets/HW-3/fish.csv")
#Part (a)
the.model = lm(Length ~ Age + Temp, data = fish)
fish$ei = the.model$residuals
fish$yhat = the.model$fitted.values
#qqnorm(the.model$residuals)
#qqline(the.model$residuals)
ei = the.model$residuals
the.SWtest = shapiro.test(ei)
#Part (b)
#qplot(yhat, ei, data = fish) +  ggtitle("Errors vs. Fitted Values") + xlab("Fitted Values") +
  #ylab("Errors") + geom_hline(yintercept = 0,col = "purple")
Group = rep("Lower",nrow(fish)) #Creates a vector that repeats "Lower" n times
Group[fish$Length < median(fish$Length)] = "Upper" #Changing the appropriate values to "Upper"
Group = as.factor(Group) #Changes it to a factor, which R recognizes as a grouping variable.
fish$Group = Group
the.FKtest= fligner.test(fish$ei, fish$Group)
#Part (c)
library(MASS)
sr = stdres(the.model)

#Problem 7
#Part (a)
alpha = 0.10
all.CIs = confint(the.model,level = 1-alpha)
#Part (d)
CIb1 = round(all.CIs[2,],4)
all.tests = summary(new.model)$coefficients
print(xtable(all.CIs,digits = 4))
# Problem 5
# Part (a)
I.model = lm(Length ~ Age + Temp + Age*Temp, data = fish)
betas = round(I.model$coefficients,4)
all.CIs = round(confint(I.model,level = 0.90),4)
print(xtable(all.CIs,digits = 4))
```