Homework 3 Solutions

Constructing Confidence Intervals

```
sleep = read.csv("https://www.macalester.edu/~ajohns24/Data/SleepStudy2.csv")
head(sleep, 2)
```

```
##   X Subject   Day0   Day2 Difference
## 1 1       1 249.56 250.80       1.24
## 2 2       2 222.73 202.98     -19.75
```

Exercise 1

```
#a
mean(sleep$Difference)
```

```
## [1] 8.710556
```

```
#b
sd(sleep$Difference)/sqrt(18)
```

```
## [1] 7.494941
```

```
#c
#95% lucky, 5% unlucky
```

Exercise 2

```
#a
#95%

#b
mean(sleep$Difference) - 2*sd(sleep$Difference)/sqrt(18)
```

```
## [1] -6.279327
```

```
mean(sleep$Difference) + 2*sd(sleep$Difference)/sqrt(18)
```

```
## [1] 23.70044
```

```
#c
t.test(sleep$Difference, conf.level=0.95)
```

```
##
##  One Sample t-test
##
## data:  sleep$Difference
## t = 1.1622, df = 17, p-value = 0.2612
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -7.102388 24.523499
## sample estimates:
## mean of x
##  8.710556
```

Exercise 3

0 is within the CI, thus our data don't signifcantly support the claim that that the average reaction time after 2 days of sleep deprivation is greater (slower) than after 0 days of sleep deprivation.

Interpreting Confidence Level

```
#attach the library
suppressPackageStartupMessages(library(gapminder))

#load the data
data(gapminder)

#examine the codebook
?gapminder
```

Exercise 4

```
#a
#each case is a country within a given year
dim(gapminder)
```

```
## [1] 1704    6
```

```
#b
summary(gapminder$year)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1952    1966    1980    1980    1993    2007
```

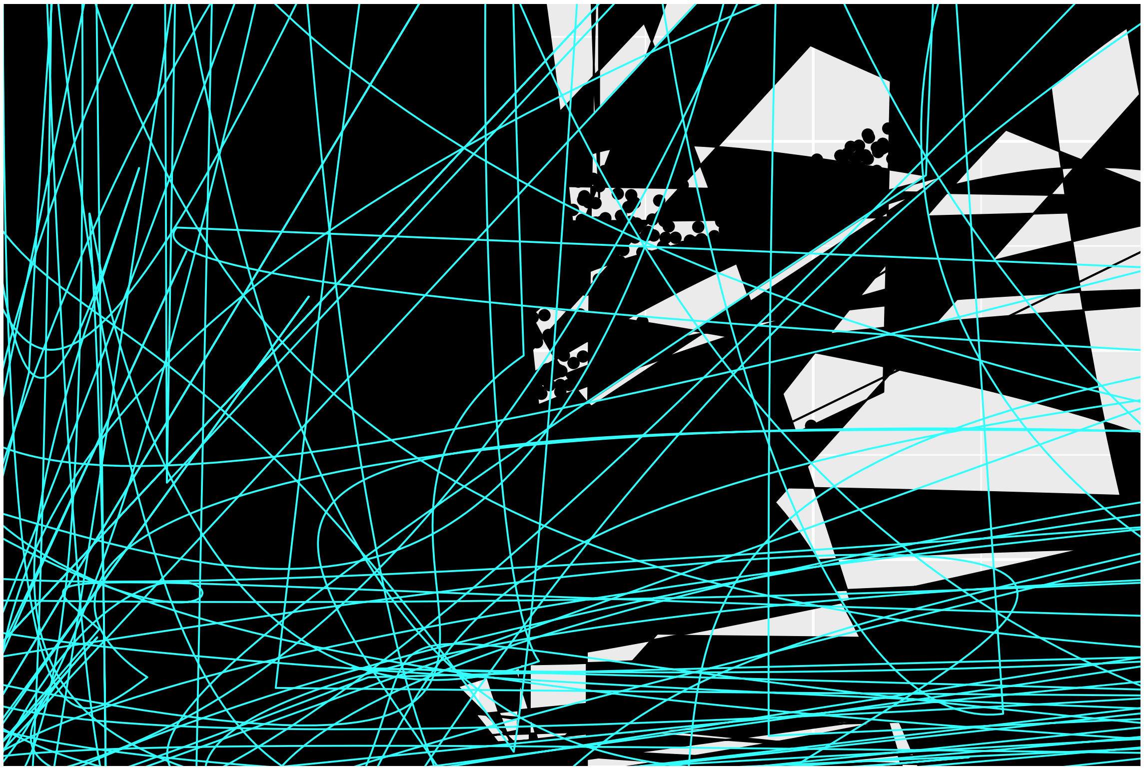
```
#c
levels(gapminder$continent)
```

```
## [1] "Africa"   "Americas" "Asia"     "Europe"   "Oceania"
```

Exercise 5

```
suppressPackageStartupMessages(library(ggplot2))

#a
ggplot(gapminder, aes(y=lifeExp, x=gdpPercap)) +
    geom_point() +
    geom_smooth(method="lm")
```

```r
#b
#residuals don't have mean 0 across the model (lack of fit)

#c
ggplot(gapminder, aes(y=lifeExp, x=log(gdpPercap))) +
    geom_point() +
    geom_smooth(method="lm")
```



```r
#d
suppressPackageStartupMessages(library(dplyr))
gap2007 <- gapminder %>%
    filter(year==2007)

ggplot(gap2007, aes(y=lifeExp, x=log(gdpPercap), color=continent, size=pop)) +
    geom_point() +
    lims(y=c(15,90))
```

Exercise 6

```r
#set the seed
set.seed(39)

#take a sample
samp1 <- sample_n(gap2007, size=25)

#calculate the sample mean & sd
mean(samp1$lifeExp)
```

```
## [1] 72.49796
```

```r
sd(samp1$lifeExp)
```

```
## [1] 9.723725
```

```r
#a
mean(samp1$lifeExp) - 2*sd(samp1$lifeExp)/sqrt(25)
```

```
## [1] 68.60847
```

```r
mean(samp1$lifeExp) + 2*sd(samp1$lifeExp)/sqrt(25)
```

```
## [1] 76.38745
```

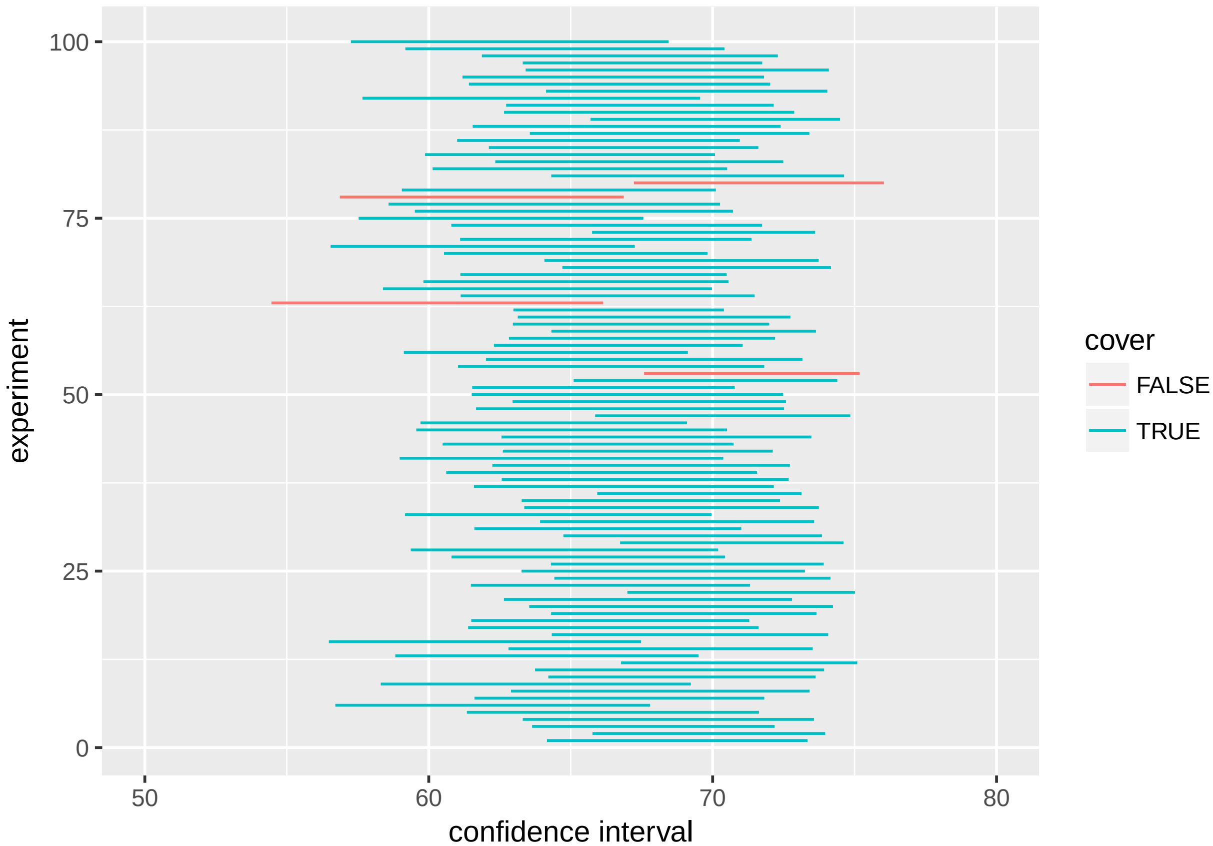
```r
#b
t.test(samp1$lifeExp, conf.level=0.95)
```

```
##
##  One Sample t-test
##
## data:  samp1$lifeExp
## t = 37.279, df = 24, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  68.48420 76.51172
## sample estimates:
## mean of x
##  72.49796
```

```
#c
#No, the true mean mu falls below our CI
```

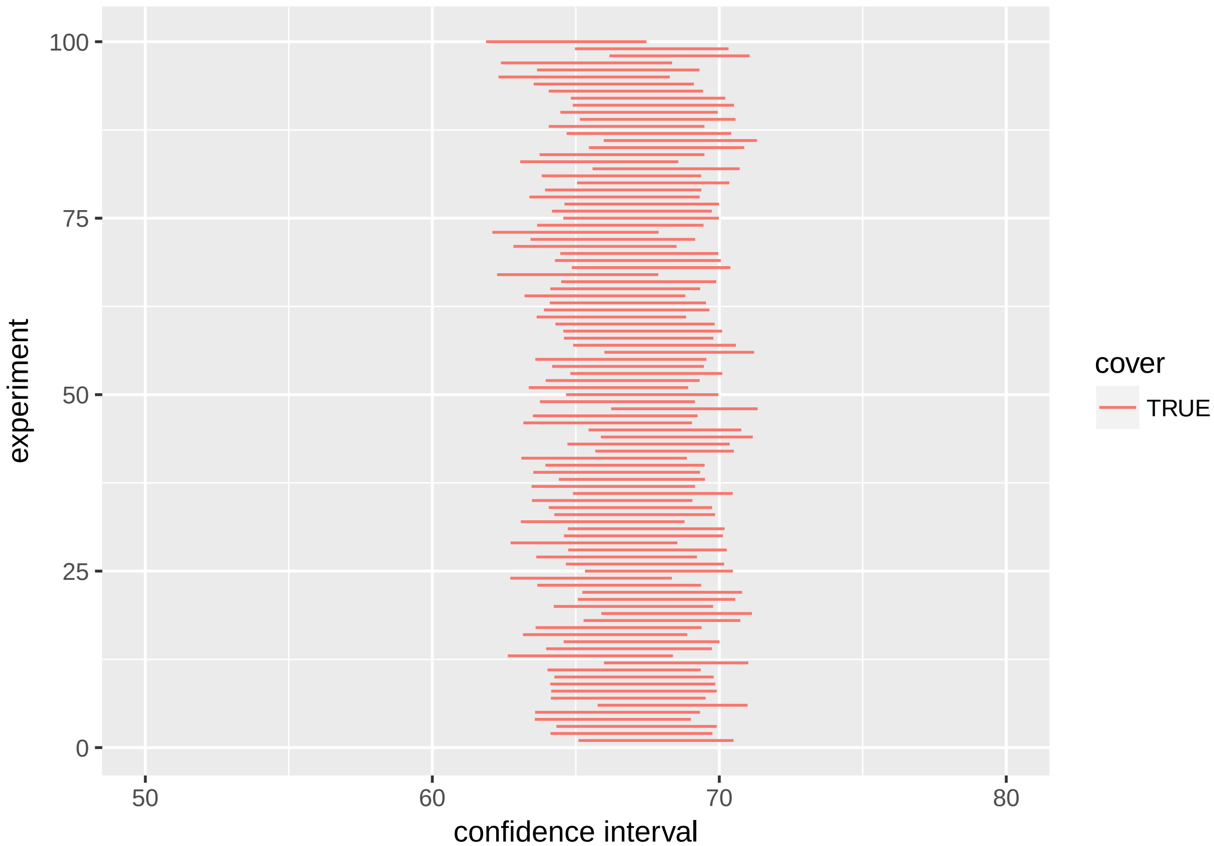
Exercise 7

```
#set the seed
set.seed(90)

#initialize the for loop
CIsimulation <- data.frame(lower=rep(0,100), upper=rep(0,100))

#run the for loop
for(i in 1:100){
    samp <- sample_n(gap2007, size=25)
    CIsimulation[i,] <- t.test(samp$lifeExp, conf.level=0.95)$conf.int
}

#b
CIsimulation <- CIsimulation %>%
    mutate(experiment=c(1:100), cover=(lower < 67.00742) & (upper > 67.00742))

#c
#94% cover mu
ggplot(CIsimulation, aes(y=experiment, x=lower, color=cover)) +
    geom_segment(aes(x=lower, xend=upper, y=experiment, yend=experiment)) +
    labs(x="confidence interval") +
    lims(x=c(50,80))
```



Exercise 8

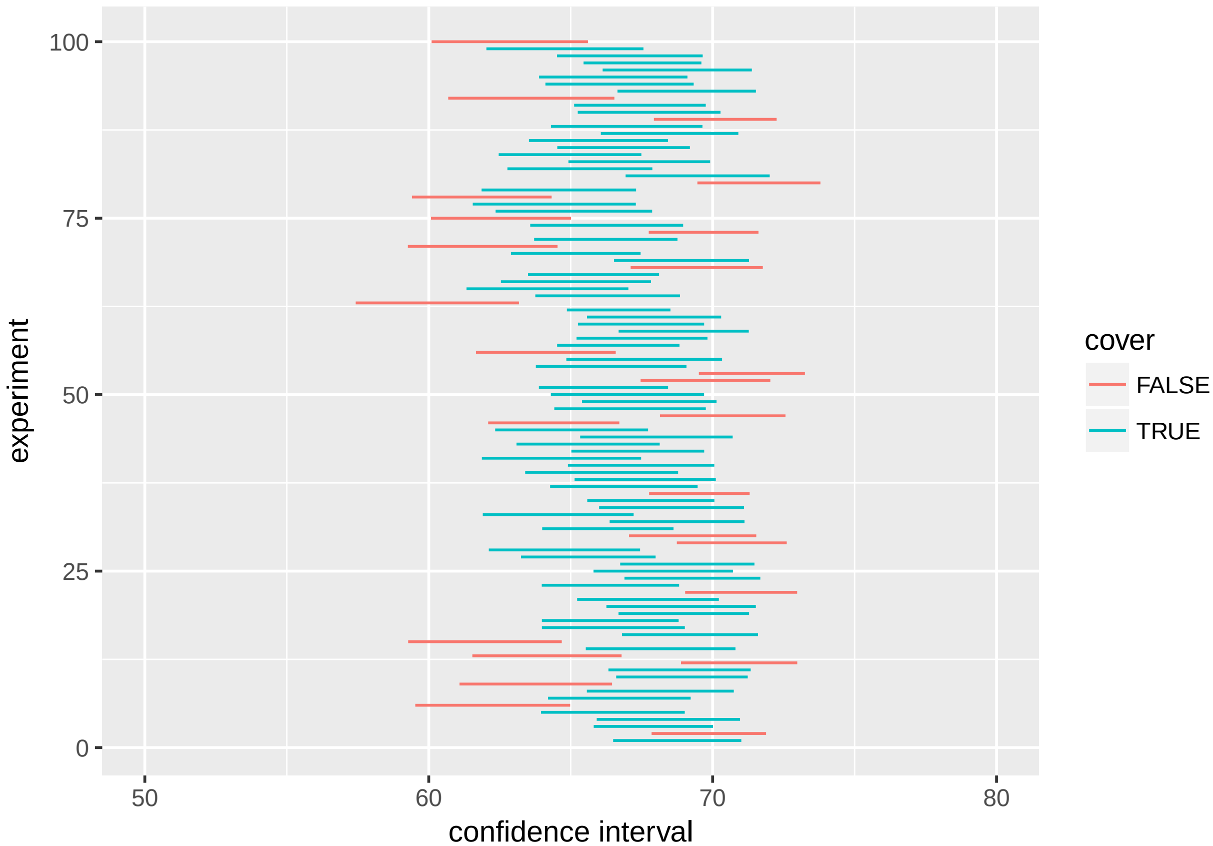
```
#a: intuition

#b
set.seed(90)
CIsimulation <- data.frame(lower=rep(0,100), upper=rep(0,100))
for(i in 1:100){
    samp <- sample_n(gap2007, size=75)
    CIsimulation[i,] <- t.test(samp$lifeExp, conf.level=0.95)$conf.int
}
CIsimulation <- CIsimulation %>%
    mutate(experiment=c(1:100), cover=(lower < 67.00742) & (upper > 67.00742))
ggplot(CIsimulation, aes(y=experiment, x=lower, color=cover)) +
    geom_segment(aes(x=lower, xend=upper, y=experiment, yend=experiment)) +
    labs(x="confidence interval") +
    lims(x=c(50,80))
```



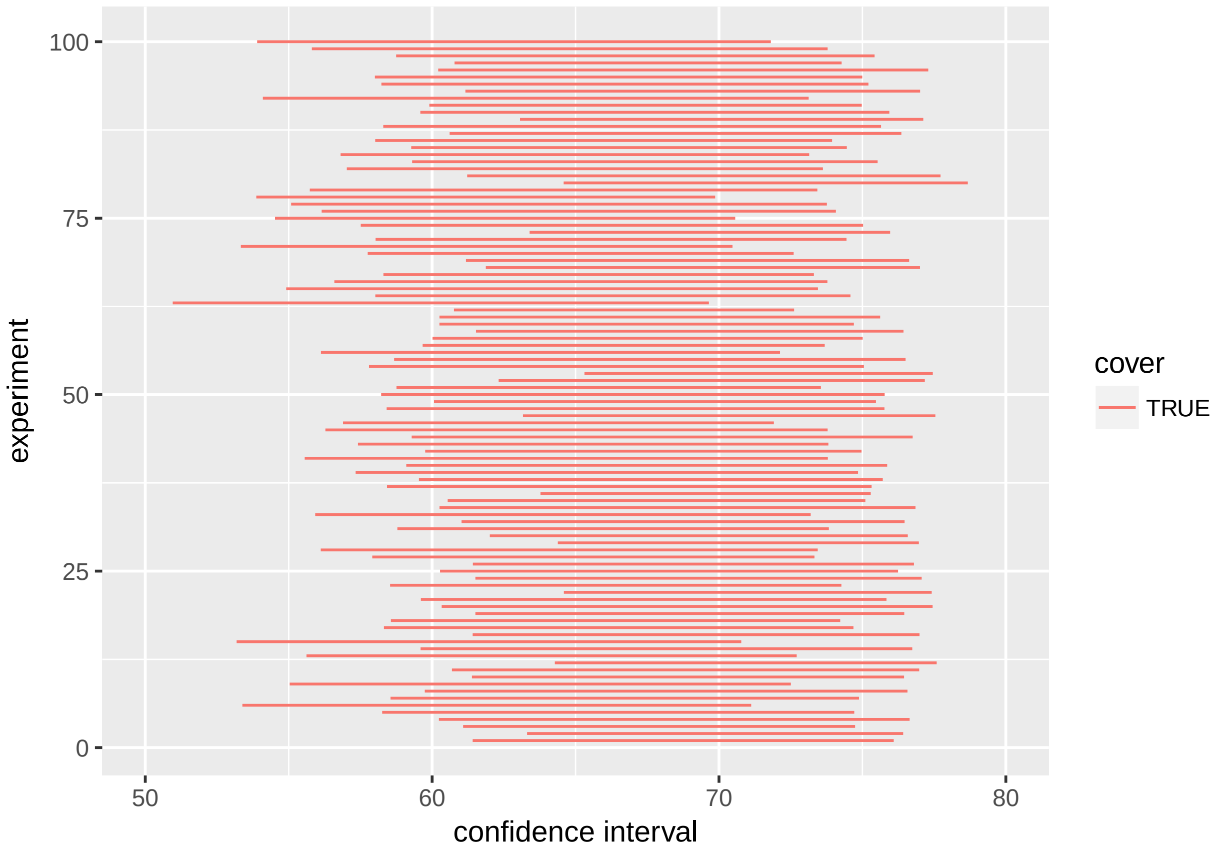

```
#c
#larger sample size = narrower CI
```

Exercise 9

```
#a
#estimate +/- 1 standard error
#68% are narrower

#b
set.seed(90)
CIsimulation <- data.frame(lower=rep(0,100), upper=rep(0,100))
for(i in 1:100){
    samp <- sample_n(gap2007, size=25)
    CIsimulation[i,] <- t.test(samp$lifeExp, conf.level=0.68)$conf.int
}
CIsimulation <- CIsimulation %>%
    mutate(experiment=c(1:100), cover=(lower < 67.00742) & (upper > 67.00742))
ggplot(CIsimulation, aes(y=experiment, x=lower, color=cover)) +
    geom_segment(aes(x=lower, xend=upper, y=experiment, yend=experiment)) +
    labs(x="confidence interval") +
    lims(x=c(50,80))
```




```
#c
set.seed(90)
CIsimulation <- data.frame(lower=rep(0,100), upper=rep(0,100))
for(i in 1:100){
    samp <- sample_n(gap2007, size=25)
    CIsimulation[i,] <- t.test(samp$lifeExp, conf.level=0.997)$conf.int
}
CIsimulation <- CIsimulation %>%
    mutate(experiment=c(1:100), cover=(lower < 67.00742) & (upper > 67.00742))
ggplot(CIsimulation, aes(y=experiment, x=lower, color=cover)) +
    geom_segment(aes(x=lower, xend=upper, y=experiment, yend=experiment)) +
    labs(x="confidence interval") +
    lims(x=c(50,80))
```




```
#d
set.seed(90)
CIsimulation <- data.frame(lower=rep(0,100), upper=rep(0,100))
for(i in 1:100){
    samp <- sample_n(gap2007, size=25)
    CIsimulation[i,] <- t.test(samp$lifeExp, conf.level=1)$conf.int
}
head(CIsimulation)
```

```
##   lower upper
## 1  -Inf   Inf
## 2  -Inf   Inf
## 3  -Inf   Inf
## 4  -Inf   Inf
## 5  -Inf   Inf
## 6  -Inf   Inf
```

```
#e
#The greater the confidence level, the greater our confidence that the CI covers the population paramet
#but the wider (and less useful) the interval.
#Why 95%?  Tradition.
```

Confidence & Prediction Intervals for Regression Models

Exercise 10

```
set.seed(60)
samp50 <- sample_n(gap2007, size=50)
sampMod <- lm(lifeExp ~ log(gdpPercap), data=samp50)

#a
#R^2 = 0.6687
#GDP explains 67% of the variability in life expectancies
summary(sampMod)
```

```
##
## Call:
## lm(formula = lifeExp ~ log(gdpPercap), data = samp50)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.7163  -2.7433   0.9611   4.2167  12.8711
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3.9392     6.5152   0.605    0.548
## log(gdpPercap)     7.3636     0.7482   9.842 4.26e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.775 on 48 degrees of freedom
## Multiple R-squared:  0.6687, Adjusted R-squared:  0.6618
## F-statistic: 96.86 on 1 and 48 DF,  p-value: 4.256e-13
```

```
#b

#c
3.9392 - 2*6.5152
```

```
## [1] -9.0912
```

```
3.9392 + 2*6.5152
```

```
## [1] 16.9696
```

```
7.3636 - 2*0.7482
```

```
## [1] 5.8672
```

```
7.3636 + 2*0.7482
```

```
## [1] 8.86
```

```
#d
confint(sampMod, level=0.95)
```

```
##                     2.5 %    97.5 %
## (Intercept)     -9.160472 17.038776
## log(gdpPercap)   5.859247  8.867894
```

```r
#e
#Yes, the interval for the GDP coef is above 0
```

Exercise 11

```r
#a.
3.9392 + 7.3636*log(42951.65)
```

```
## [1] 82.49284
```

```r
#b
#intuition

#c
predict(sampMod, newdata=data.frame(gdpPercap=42951.65),
     interval="confidence", level=0.95)
```

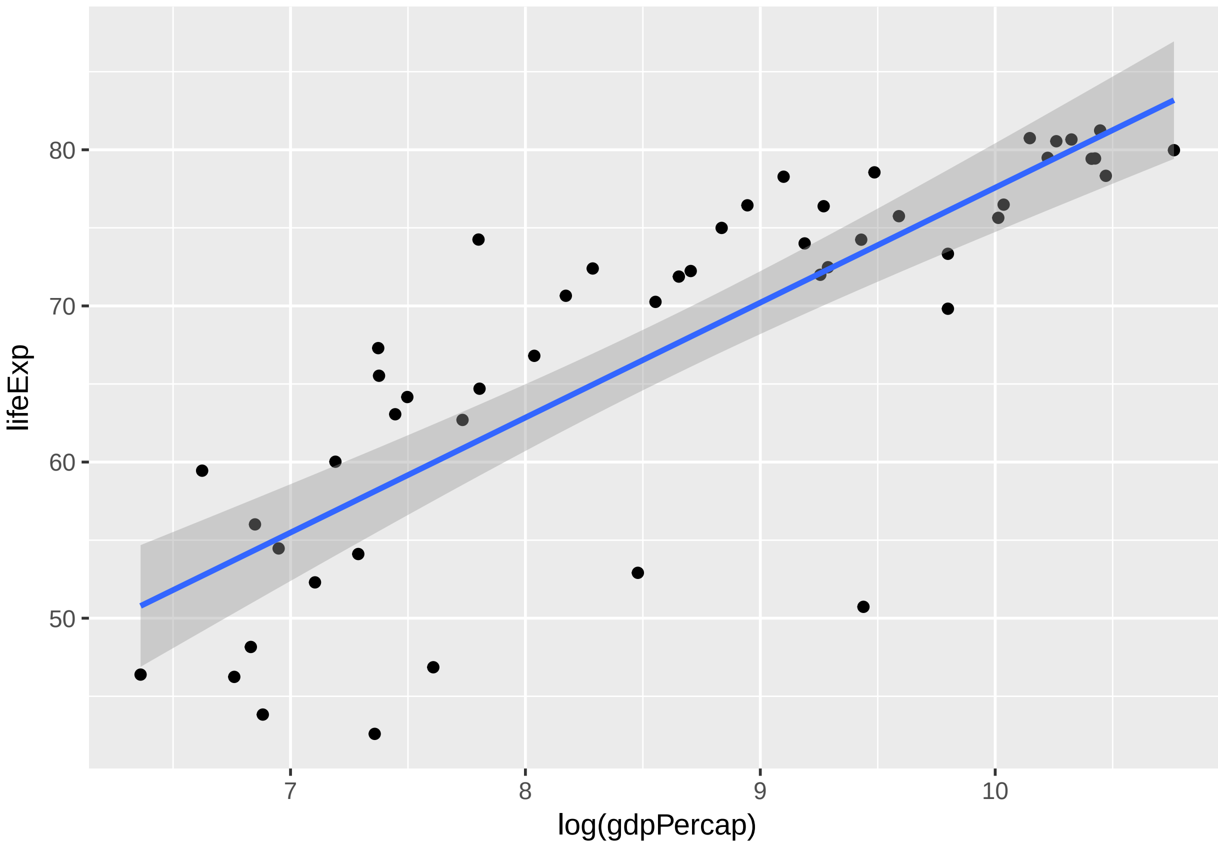
```
##        fit     lwr      upr
## 1 82.49247 78.8506 86.13434
```

```r
#d
predict(sampMod, newdata=data.frame(gdpPercap=42951.65),
    interval="prediction", level=0.95)
```

```
##        fit      lwr      upr
## 1 82.49247 68.39238 96.59256
```

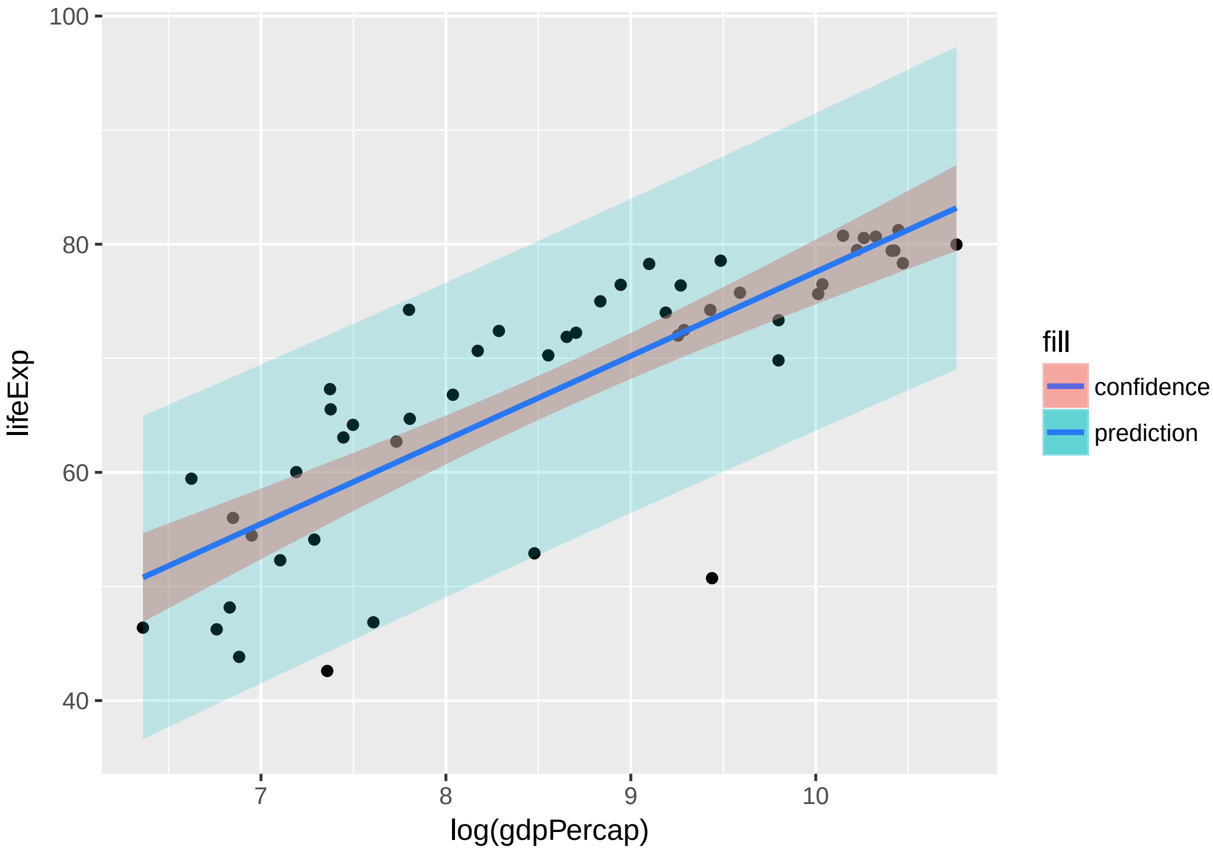
Exercise 12

```r
#a
#these cover the trend but not a high % of individual cases
ggplot(samp50, aes(x=log(gdpPercap), y=lifeExp)) +
    geom_point() +
    geom_smooth(method="lm")
```




```r
#b
#Calculate and store prediction intervals for every GDP value
PredInt = data.frame(samp50, predict(sampMod, newdata=data.frame(gdpPercap=samp50$gdpPercap),
    interval = "prediction"))

#Plot model with confidence and prediction bands
ggplot(PredInt, aes(x=log(gdpPercap), y=lifeExp)) +
    geom_point() +
    geom_smooth(method="lm", aes(fill="confidence"), alpha = 0.5) +
    geom_ribbon(aes(y=fit, ymin=lwr, ymax=upr, fill="prediction"), alpha=0.2)
```

```
## Warning: Ignoring unknown aesthetics: y
```




```r
#c
#These are the means (where we tend to have more data)
mean(samp50$lifeExp)
```

```
## [1] 67.36386
```

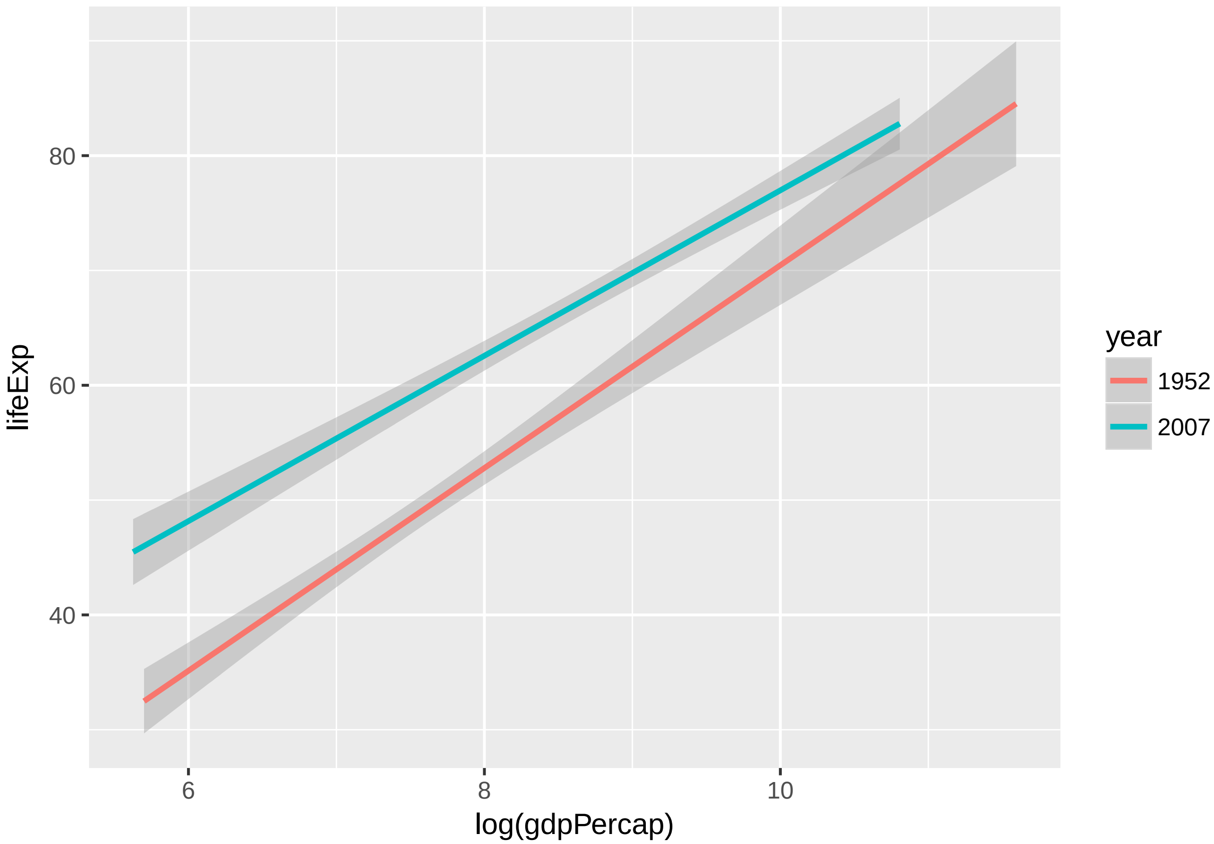
```r
mean(log(samp50$gdpPercap))
```

```
## [1] 8.613309
```

More Practice with Models and Confidence Intervals

Exercise 13

```r
#a
yeargap <- gapminder %>%
    filter(year==1952 | year==2007)

#b
yeargap$year <- as.factor(yeargap$year)

#c
ggplot(yeargap, aes(y=lifeExp, x=log(gdpPercap), color=year)) +
         geom_smooth(method="lm")
```




```r
#d
#For fixed GDP, life expectancies were 9.85 years higher in 2007 than in 1952
yearmod <- lm(lifeExp ~ log(gdpPercap) + year, yeargap)
summary(yearmod)
```

```
##
## Call:
## lm(formula = lifeExp ~ log(gdpPercap) + year, data = yeargap)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.8265  -4.2775   0.5244   5.8996  13.6033
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -10.0583     2.9501  -3.409 0.000746 ***
## log(gdpPercap)   7.8020     0.3799  20.536  < 2e-16 ***
## year2007         9.8450     0.9950   9.894  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.696 on 281 degrees of freedom
## Multiple R-squared:  0.7419, Adjusted R-squared:  0.7401
## F-statistic:   404 on 2 and 281 DF,  p-value: < 2.2e-16
```

```r
9.8450 - 2*0.9950
```

```
## [1] 7.855
```

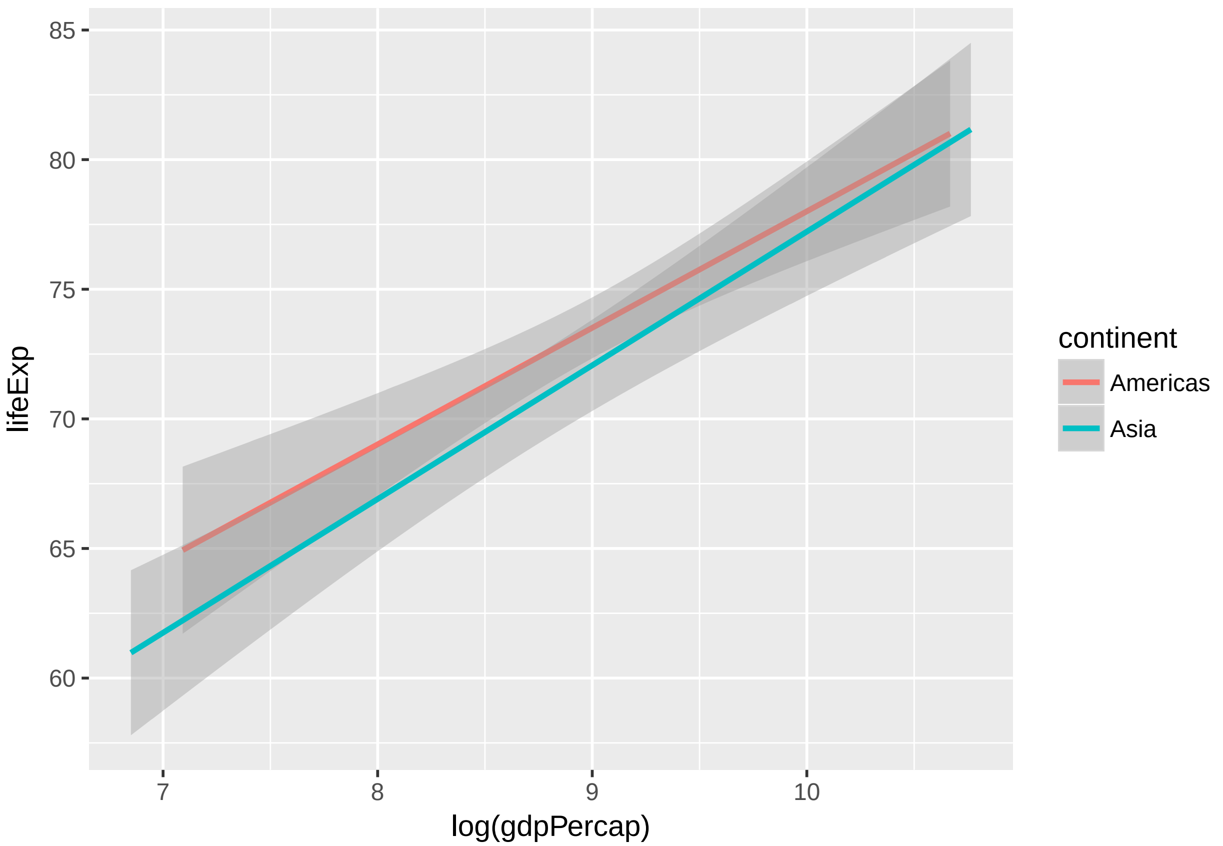
```
9.8450 + 2*0.9950
```

```
## [1] 11.835
```

```
#e
#Yes.  The interval is far above 0.
```

Exercise 14

```
#a
AAgap <- gap2007 %>%
    filter(continent=="Americas" | continent=="Asia")

#b
ggplot(AAgap, aes(y=lifeExp, x=log(gdpPercap), color=continent)) +
          geom_smooth(method="lm")
```




```
#c
#For fixed GDP, life expectancies are 1.5 years shorter among Asian countries than North/South American
AAmod <- lm(lifeExp ~ log(gdpPercap) + continent, AAgap)
summary(AAmod)
```

```
##
## Call:
## lm(formula = lifeExp ~ log(gdpPercap) + continent, data = AAgap)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -17.5928  -1.4674   0.1152   2.5441   8.2293
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      28.4335     4.7140   6.032 1.43e-07 ***
## log(gdpPercap)    5.0075     0.5146   9.731 1.47e-13 ***
## continentAsia    -1.4740     1.0938  -1.348    0.183
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.089 on 55 degrees of freedom
## Multiple R-squared:  0.6491, Adjusted R-squared:  0.6363
## F-statistic: 50.87 on 2 and 55 DF,  p-value: 3.109e-13
```

```r
-1.4740 - 2*1.0938
```

```
## [1] -3.6616
```

```r
-1.4740 + 2*1.0938
```

```
## [1] 0.7136
```

```r
#d
#0 is in the interval
#thus there's not significant evidence that life expectancies
#are significantly higher in the Americas than in Asia
```