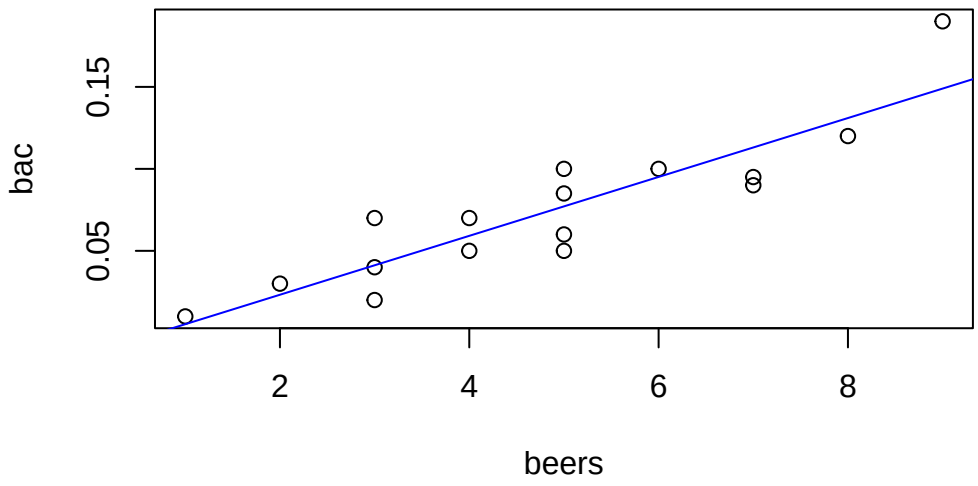
Lab 2

Brad Staples

1b

```
plot(beers, bac)
reg<-lm(bac~beers)
abline(reg, col="blue")
```




```
confint(reg, level=0.95)
```

```
                  2.5 %     97.5 %
(Intercept) -0.03980535 0.01440414
beers        0.01281262 0.02311490
```

The range for our confidence interval for 95% is between -0.03980535 and 0.01440414, so when a person has consumed zero beers their bac should be within this range. Realistically an individual who has consumed zero beers should have a BAC of zero, which lies within this interval, which supports the baseline bac being at zero.

1g

A 98% confidence interval for the mean blood alcohol level after 1 beer would be wider than the 95% confidence interval after 4 beers There are two reasons for this, firstly because a 98% confidence interval means we want to be more certain that our interval contains true mean values, which implies a larger margin of error which automatically makes the interval a little wider. The second reason is because that the 95% confidence interval is based on a mean of 4 beers, which is closer to the middle of the dataset, where values are the less prone to extrapolation and this extra precision translates to a narrower confidence despite a 95% confidence level.

1h

```
bac_data<-predict(reg, list(beers=c(1,3,7)), interval="confidence", level=0.98)
```

The interval for three beers contains a mean of 0.0411907, a lower value of 0.0235725 and a upper value of 0.0588088. This means after three beers consumed we are 98% confident the mean BAC for the population of people who consume 3 beers is 0.0411907, with the range of the mean being between 0.0235725 and 0.0588088.

1j

```
bac_data2<-predict(reg, list(beers=c(1,3,5,7)), interval="prediction", level=0.9)
```

This time the interval for three beers contains a mean of 0.0411907, a lower value of 0.003296 and a upper value of 0.0790853. This means after 3 beers are consumed we can predict with 90% confidence the the mean bac of a single individual consuming the 3 beers will average 0.0411907, and this bac could be as low as 0.003296 or as high as 0.0790853.

1k

The confidence interval is telling us what the mean bac of our entire population, giving us an overall idea of what the mean bac will be at a given number of beers consumed, based on the confidence level we set. Its a way to look at our dataset and find the mean y of a given x value. The prediction interval is more focused on what the bac of a new individual in the dataset would be after 3 beers, so its based more on adding in a new data point to the set, and relies a bit more on assumption.

2b

```
reg2<-lm(time~dist)
coef(reg2)
```

```
(Intercept)        dist
  -4.840720    8.330456
```

The intercept of -4.8407202 means that the predicted record race time for a race with a distance of zero miles is -4.8407202 minutes. This value may initally seem meaningless or incorrect context of hill running because a distance of zero miles is impossible and time cannot be negative. However, this intercept does serve as a starting point for our regression line in this context of the graph, so its negative values make more sense in that context.

2c

To test if there is significant evidence at the alpha level of 0.05 that an additional mile of the race increases the record winning time by more than 7 minutes, we will use the following calculated values.

Firstly we need to compute the t-statistic with the formula $t = \frac{\text{Observed Value} - \text{Hypothesis}}{\text{Std. Error}}$, which for our values given would be $t = \frac{8.3305-7}{0.6196} = 2.147$

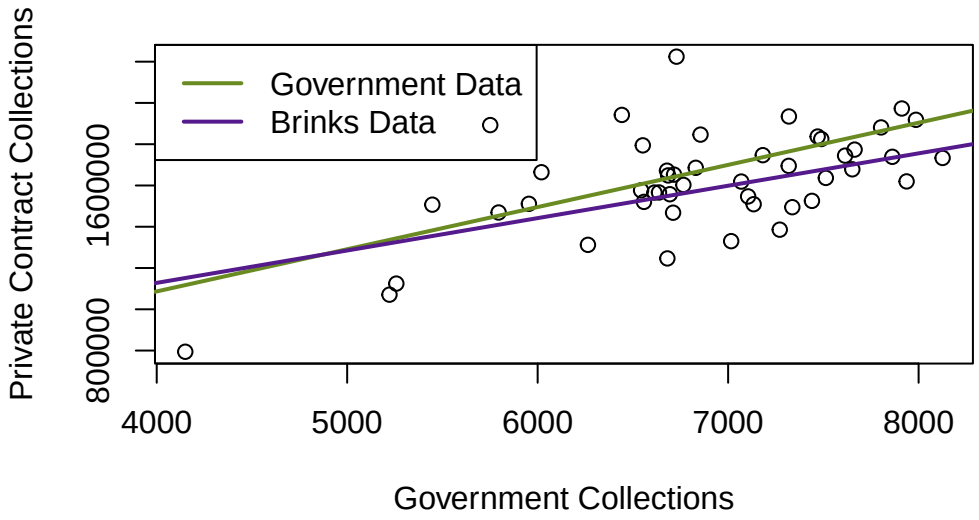
Now to determine if there is evidence to our hypothesis we need to calculate our critical t-value for a one tailed test with 33 degrees of freedom, uisng the r function abs(qt(0.05,33)) gives us a critical t-value of 1.6923603. Now we can compared our t-value to our critical t-value, 2.147>1.692 and we can see that the t value is larger than the critical value and the results falls into the rejection region, so we can reject our null hypo

This means there is significant evidence at the 0.05 level that increasing the race distance by one mile increases the record winning time by more than 7 minutes. This result also makes sense in context: since an average mile typically takes at least seven minutes in long-distance

racing, adding another mile would be expected to increase the total winning time by a similar or slightly greater amount.

Question 5

```
meter_data<-read.csv("meter_data.csv")
attach(meter_data)
govt_data<-subset(meter_data, brinks==0)
brinks_data<-subset(meter_data, brinks==1)
plot(priv~govt, data=meter_data, xlab="Government Collections", ylab="Private Contract Colle
govt_reg<-lm(priv~govt, data=govt_data)
brinks_reg<-lm(priv~govt, data=brinks_data)
abline(govt_reg, col="olivedrab4", lwd=2)
abline(brinks_reg, col="purple4", lwd=2)
legend("topleft",
       legend=c("Government Data", "Brinks Data"),
       col=c("olivedrab4", "purple4"),
       lty=1,
       lwd=2)
```



The two regression lines start off relatively close when Government Collections (x) are low, around 4,000, and they intersect near x = 4,800. After the crossover, the lines begin to

separate, and by x = 5,800 the difference becomes more visible. The Government's regression line (green) grows at a consistently faster rate than the Brinks line, indicating that when Brinks was the private contractor collecting coins, they reported lower collections compared to months when other contractors were used.

Question 6

```
confint(govt_reg, level=0.9)
```

```
                    5 %        95 %
(Intercept) -349167.416 886994.9981
govt            113.856    294.9707
```

```
confint(brinks_reg, level=0.9)
```

```
                   5 %         95 %
(Intercept) -93226.610 1094145.4441
govt            71.591     242.1931
```

There is substantial overlap between the 90% confidence intervals for the intercepts of the two regressions. The overlapping region (approximately from –93,226.610 to 886,994.9981) suggests that the baseline amount of money collected by both groups is relatively similar, meaning these values are not statistically different. Therefore, we cannot conclude that Brinks had a systematically higher or lower baseline collection level than other private companies during their collection months

Question 7

The 90% confidence intervals for the slopes also overlap (from roughly 113.856 to 242.1931). This means that we cannot say, with 90% confidence, that there is a statistically significant difference between the slopes. However, the Brinks interval lies mostly below that of the non-Brinks interval, suggesting that Brinks' collections tended to increase at a slower rate than those of other contractors. While this does not provide conclusive proof of wrongdoing, it supports the suspicion that Brinks may have collected or reported less money than expected.