



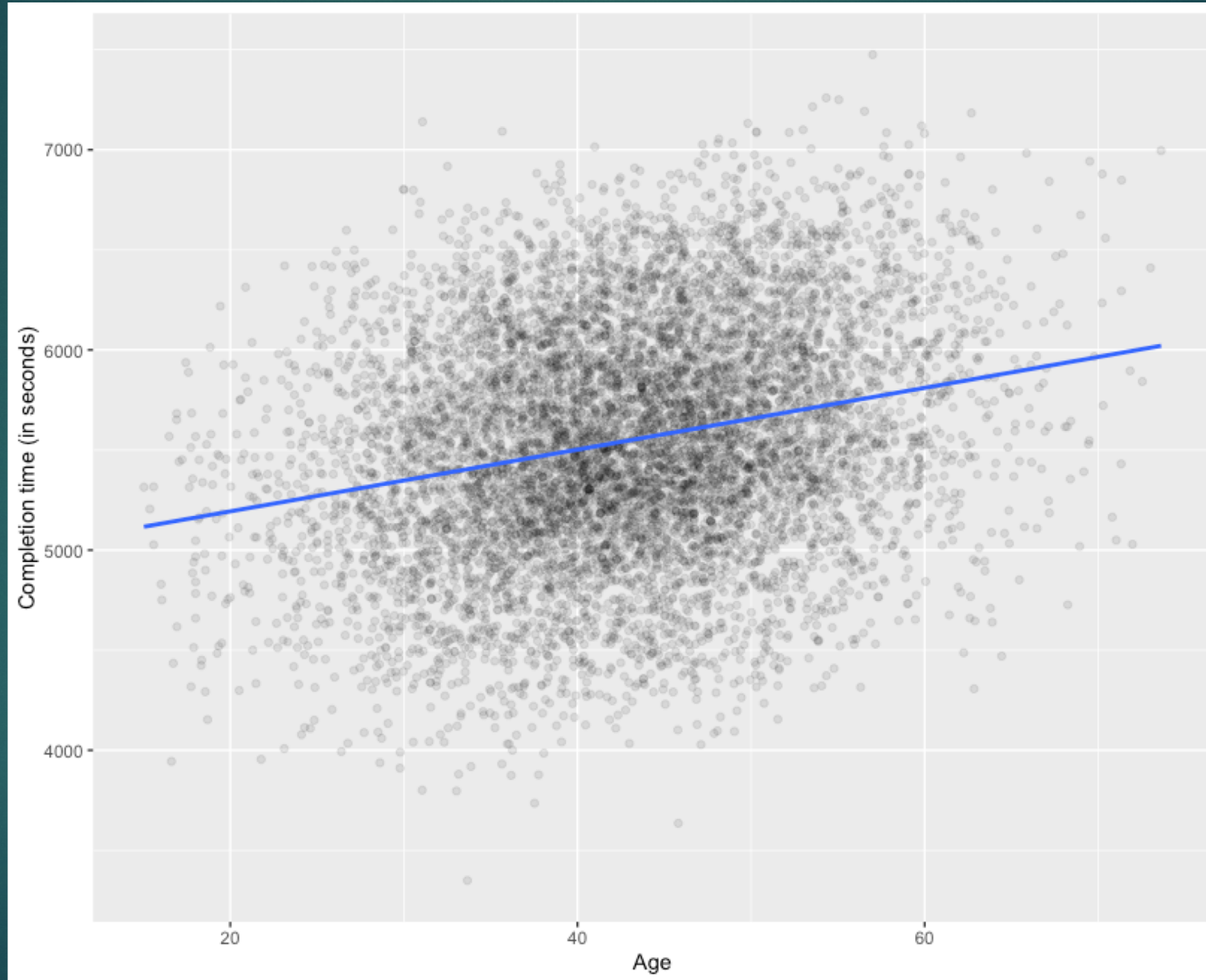
Multiple Linear Regression

QAC 201

- 
- Suppose you are interested in how age relates to length of time it takes runners to complete a 10 mile race.

What is your preliminary approach to answering this question?

Begin by visualizing the association:



Or you could test the association with regression:

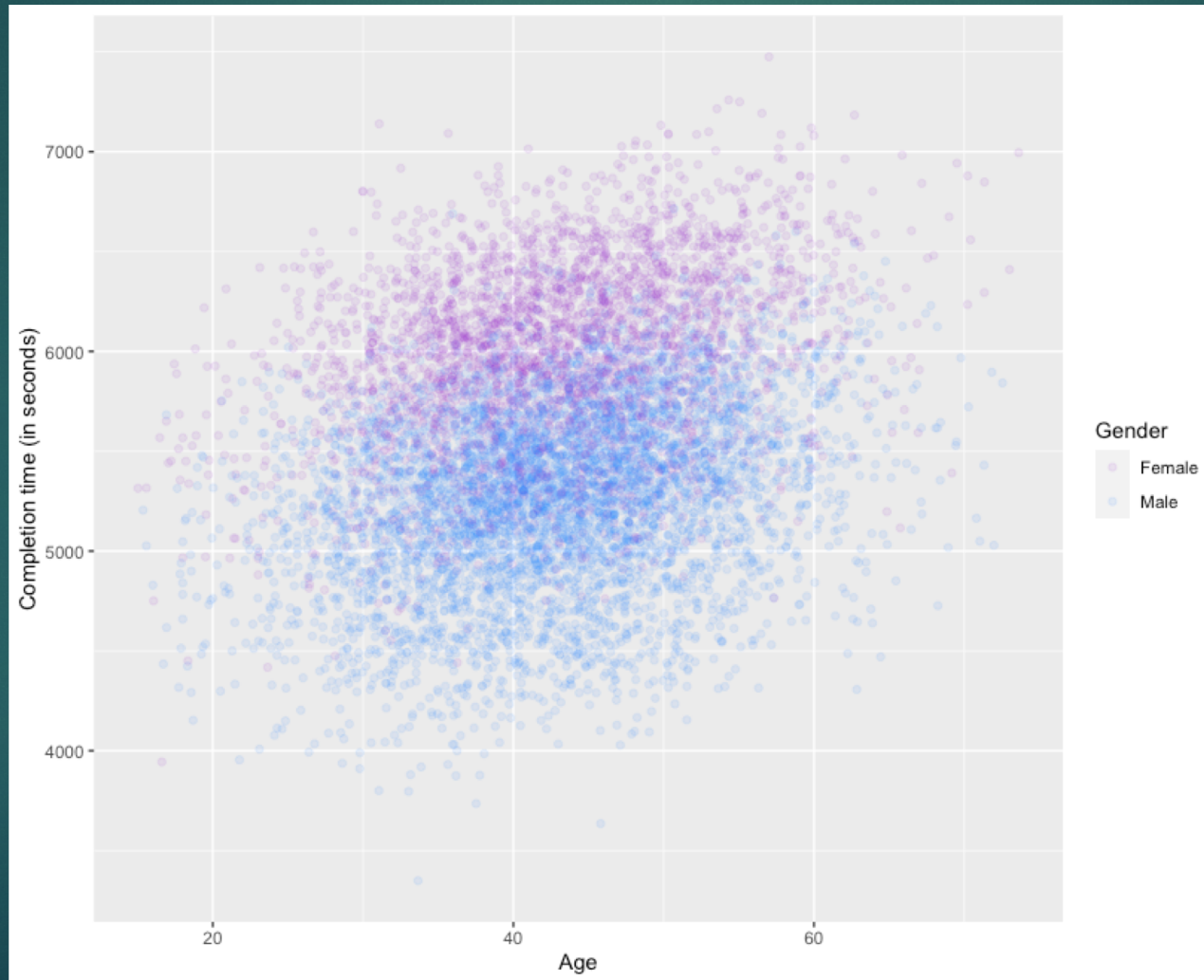
```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 4884.8024    28.8822  169.13  <2e-16 ***
age          15.4285     0.6583   23.44  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

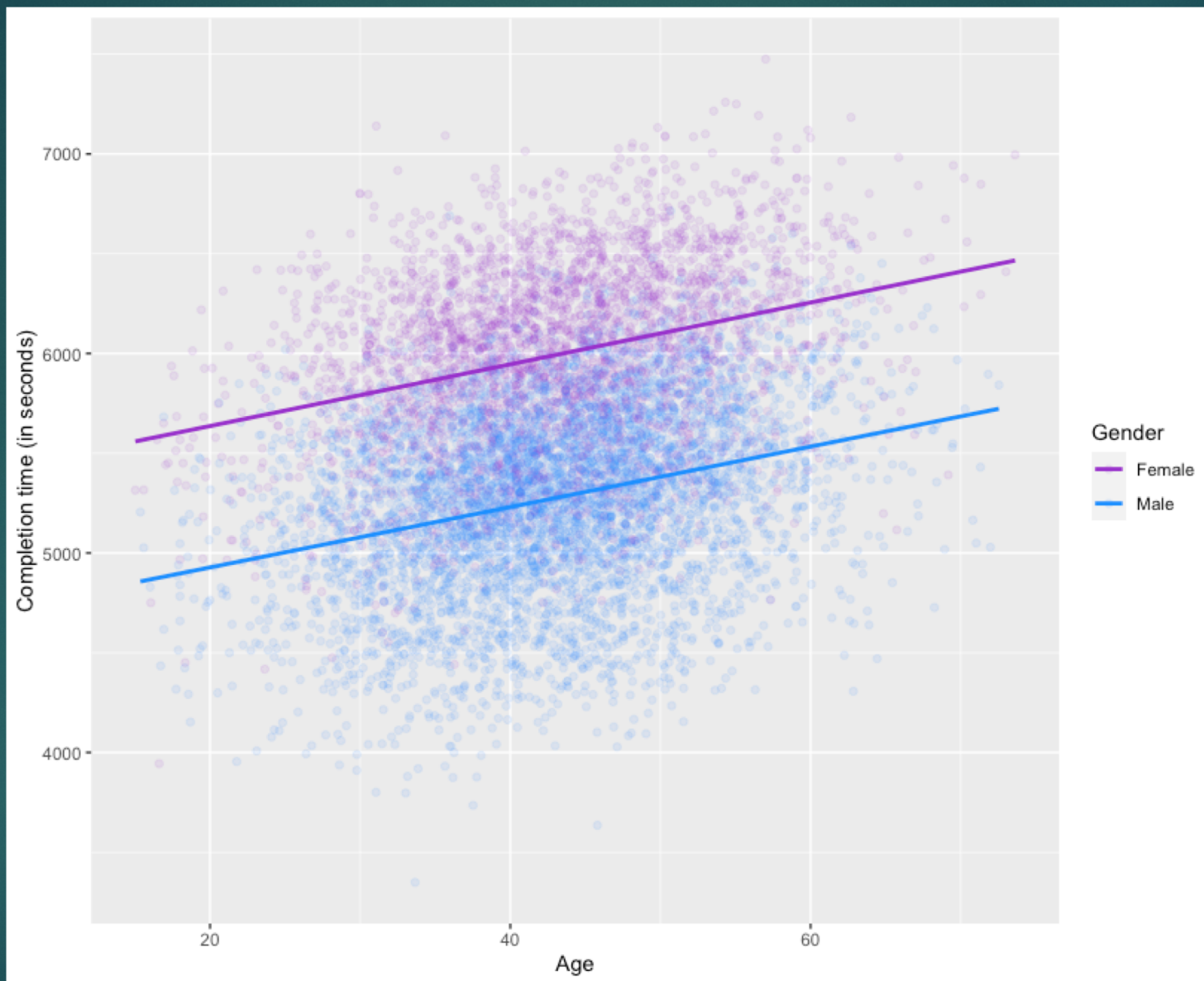
Residual standard error: 555.8 on 8634 degrees of freedom
Multiple R-squared:  0.05982,    Adjusted R-squared:  0.05971
F-statistic: 549.3 on 1 and 8634 DF,  p-value: < 2.2e-16
```

$$\widehat{Time} = 4884.8 + 15.4(Age)$$

There is a significant association between age and completion of race (Beta=15.4, p-value<0.001). When examining the association between age and time completion, for each additional year increase in age, the completion time is expected to increase by 15.4 seconds on average.

Ultimately, we may want to try to improve the accuracy of our predictions by including another variable that is expected to relate to our response:





Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	5335.6073	23.3235	228.76	<2e-16	***
age	15.2681	0.5134	29.74	<2e-16	***
male	-716.4446	9.6078	-74.57	<2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

$$\widehat{Time} = 5335 + 15.3(\text{Age}) - 716(\text{Male})$$

Both age and gender are significantly related to average time completion of a 10 mile race.

How does the model estimate that male times compare to female times?

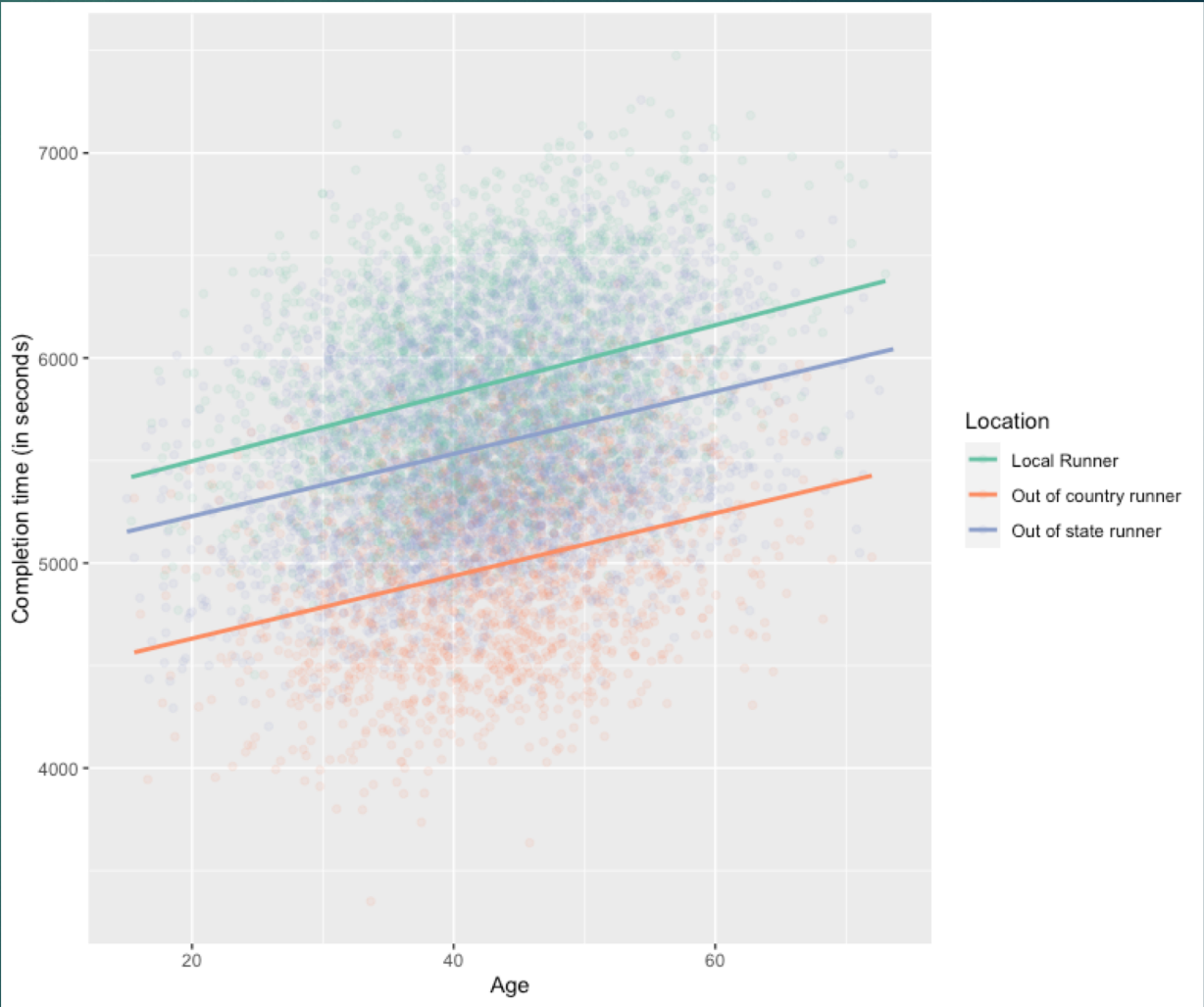
What does the model predict to be the average time completion for a 25 year old female?

Now suppose instead we want to look at where the runner traveled from to complete the race:

- Local runner
- Out of state runner
- Out of country runner

Sample of data:

ID	Age	Location	Time
1	47	Local	6532
2	49	Out of state	4997
3	40	Out of state	6026
4	51	Out of state	5045
5	20	Out of country	4501



Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	4905.5015	24.4373	200.74	<2e-16	***
age	15.6400	0.5445	28.72	<2e-16	***
local	299.8215	11.4251	26.24	<2e-16	***
out_of_country	-594.9430	13.0585	-45.56	<2e-16	***

What is the estimated equation?

How does age relate to completion time?

How does location relate to completion time?

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	4905.5015	24.4373	200.74	<2e-16	***
age	15.6400	0.5445	28.72	<2e-16	***
local	299.8215	11.4251	26.24	<2e-16	***
out_of_country	-594.9430	13.0585	-45.56	<2e-16	***

What is the estimated completion time for a local 25 year old runner?

What is the estimated completion time for an out-of-country 45 year old runner?