

Stat 411/511

INFERENCE IN SIMPLE LINEAR
REGRESSION

Nov 24 2014

DA #2

Avoid story telling...

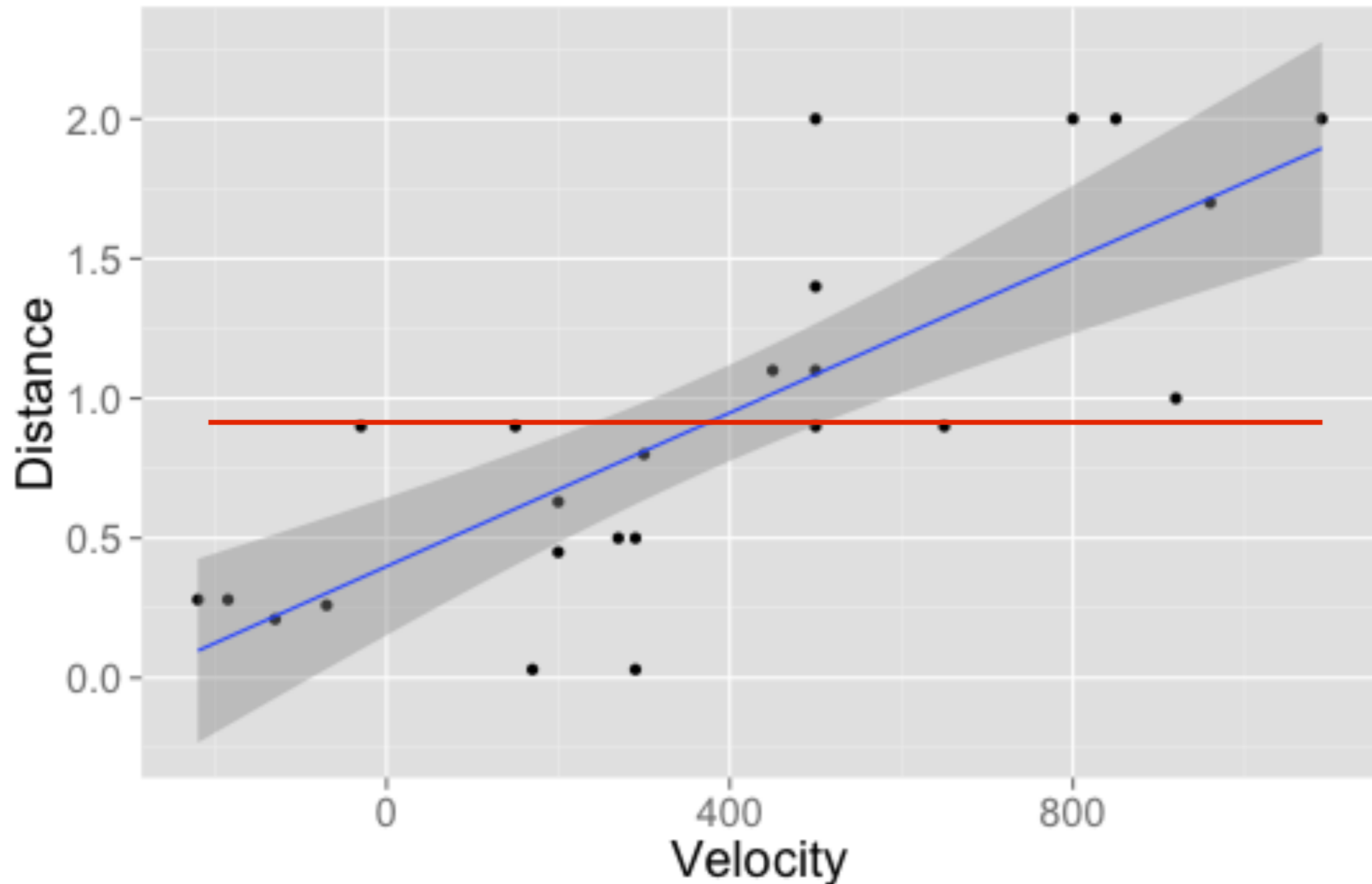
First I ..., then I I noticed so I....

Although the residuals showed clear non-Normality, the large sample size gives robustness to this violation of the ANOVA assumptions. The residuals were also examined for

Three types of inference

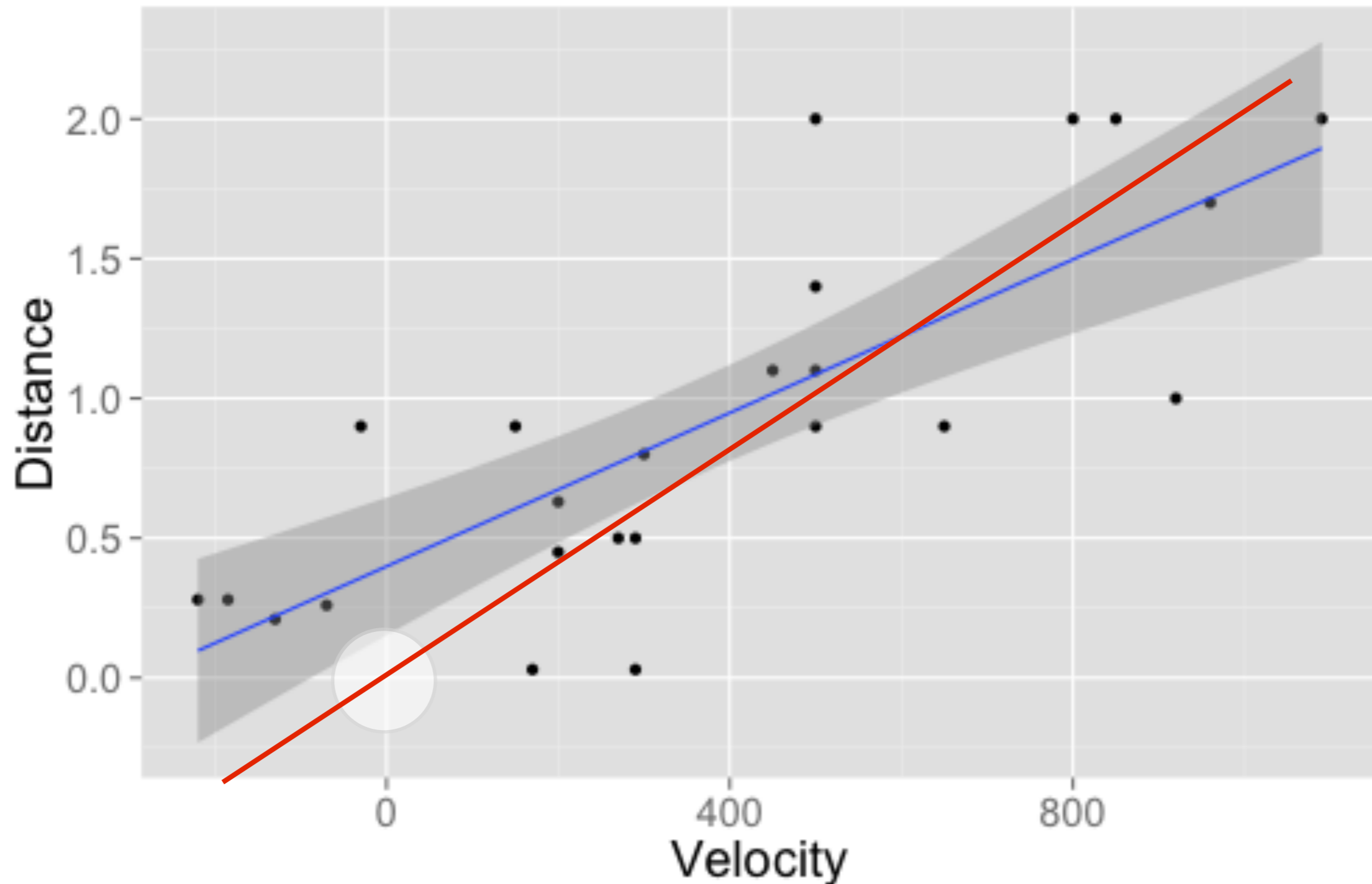
- ① **Inference on the slope or intercept**
uncertainty comes from sampling error in a single parameter
- ② **Inference about the mean response
(at a given explanatory value)**
uncertainty comes from sampling error in both parameters
- ③ **Prediction of a new response
(at a given explanatory value)**
uncertainty comes from sampling error in both parameters
and variability in subpopulations

Questions that might be of interest in a regression setting



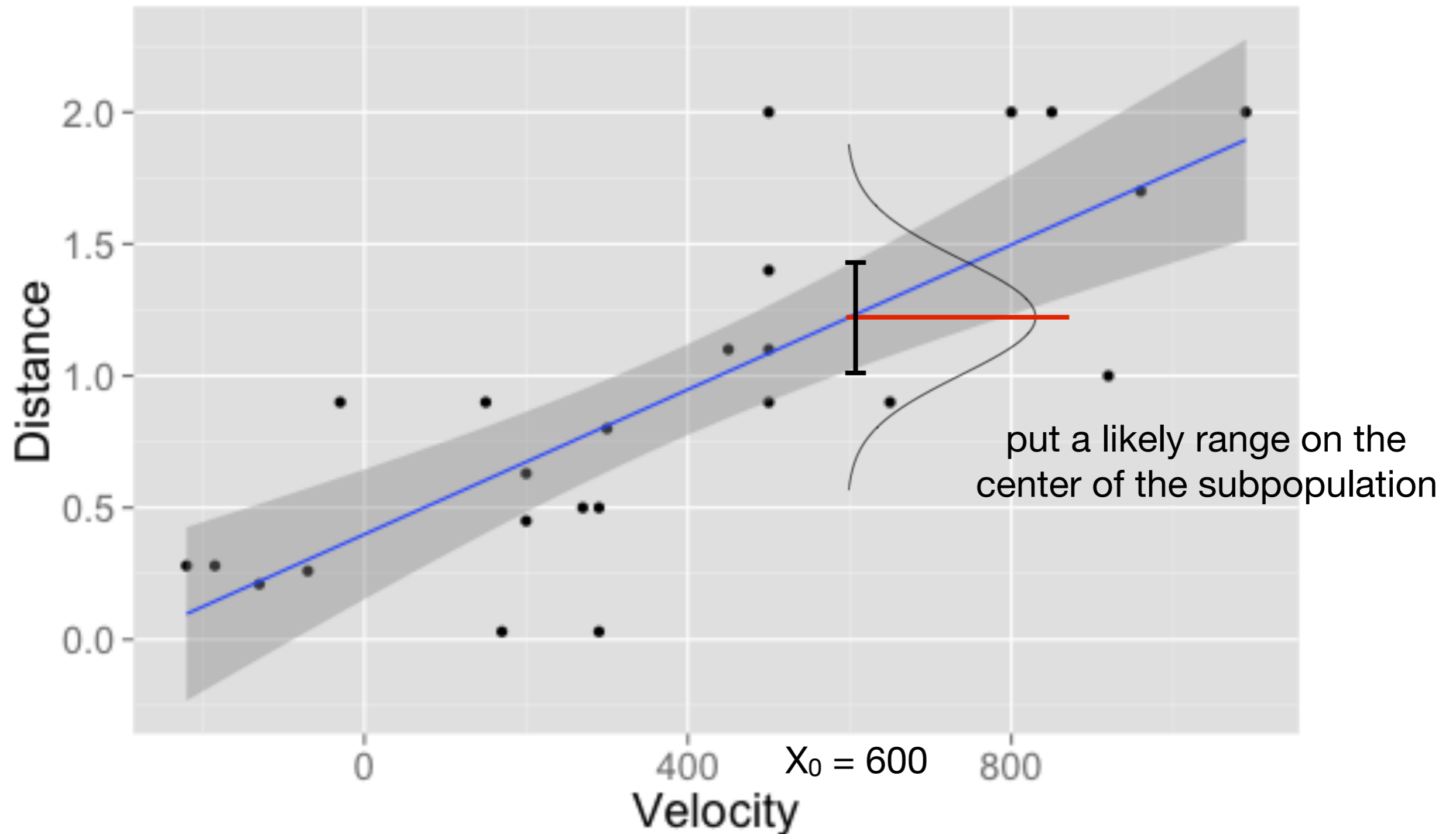
Is the true slope zero? If yes, then the mean response doesn't depend on the explanatory variable.

Questions that might be of interest in a regression setting



Is the true intercept zero? If yes, then the mean response is zero when the explanatory variable is zero.

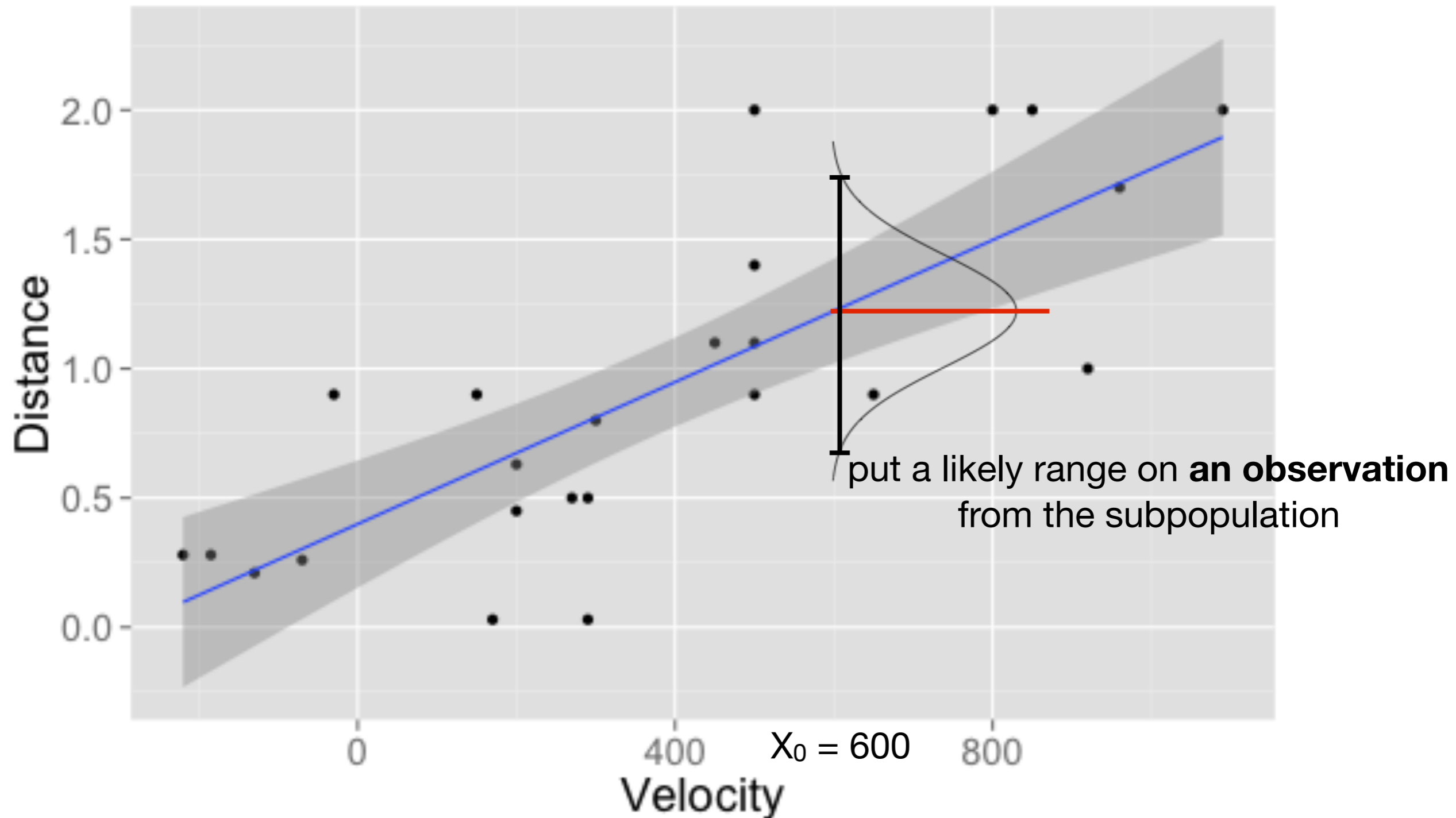
Questions that might be of interest in a regression setting



What's the mean response when the explanatory variable = X_0 ?

2

Questions that might be of interest in a regression setting



What's a likely response for an observation when the explanatory variable = X_0 ? (Prediction) ③

For all three types of inference under our assumptions

$\frac{\text{Estimate} - \text{True value}}{\text{SE}_{\text{Estimate}}}$ has a t-distribution with $n-2$ degrees of freedom

Leads to...

95% CI/Pis: $\text{Estimate} \pm t_{n-2}(0.975) \times \text{SE}_{\text{Estimate}}$

Tests of the null hypothesis: true value = 0

t-ratio = $\frac{\text{Estimate}}{\text{SE}_{\text{Estimate}}}$ and p-values like usual
i.e. $2*(1 - \text{pt}(\text{abs}(\text{t.stat}), n-2))$

but you wouldn't test in ③

1

Inference about slope or intercept

uncertainty comes from sampling variability in a single parameter

From last lecture we can find estimates and their standard errors for the slope and intercept

t-statistics for the null: $\beta_0 = 0$,

$$(\hat{\beta}_0 - 0) / SE_{\hat{\beta}_0} \quad \text{Same for } \beta_1$$

and p-values like usual

i.e. $2*(1 - pt(abs(t.stat), n-2))$

Individual 95% confidence intervals:

$$\hat{\beta}_0 \pm t_{n-2}(0.975) SE_{\hat{\beta}_0}$$

$$\hat{\beta}_1 \pm t_{n-2}(0.975) SE_{\hat{\beta}_1}$$

Some examples of tests

Null: The slope is zero.

(The mean response doesn't depend on the explanatory variable).

Null: The intercept is zero.

Null: The slope is 1.

(problem specific, maybe you expect a specific slope from theory)

Alternatives will generally be "... is not equal to"

```

> library(Sleuth3)
> fit <- lm(Distance ~ Velocity, data = case0701)
> summary(fit)

```

```

Call:
lm(formula = Distance ~ Velocity, data = case0701)

```

```

Residuals:
    Min       1Q   Median       3Q      Max
-0.7632 -0.2352 -0.0088  0.2072  0.9144

```

p-value, two-sided
for null hypothesis
parameter = 0

```

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.3990982  0.1184697   3.369  0.00277 **
Velocity     0.0013729  0.0002274   6.036  4.48e-06 ***
---

```

```

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

```

```

Residual standard error: 0.405 on 22 degrees of freedom
Multiple R-squared: 0.6235,    Adjusted R-squared: 0.6064
F-statistic: 36.44 on 1 and 22 DF,  p-value: 4.477e-06

```

t-statistic =
Estimate/Std. Error

Your turn

Using this output from R, construct a 95% CI for the slope and intercept.

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	0.3990982	0.1184697	3.369	0.00277	**
Velocity	0.0013729	0.0002274	6.036	4.48e-06	***

`qt(0.975, 22) = 2.073873`

```
> confint(fit)
                2.5 %      97.5 %
(Intercept) 0.1534070224 0.644789404
Velocity    0.0009012456 0.001844627
```

With 95% confidence, the **mean** distance of a nebula with **zero velocity** is between 0.153 and 0.645 parsecs from Earth.

With 95% confidence, **an increase** in velocity of **1km/sec** is associated with an **increase in mean distance** between 0.0009 and 0.0018 parsecs.

2

Inference about mean response

uncertainty comes from sampling variability
in both parameters

Estimate of the mean

We've already seen to estimate the mean response at an explanatory value, say X_0 , we just substitute our estimates into the line equation,

$$\hat{\mu}\{Y|X_0\} = \hat{\beta}_0 + \hat{\beta}_1 X_0$$

To make inferences we need to know the standard error on this estimate.

Standard error on the estimated mean

$$SE_{\hat{\mu}\{Y|X_0\}} = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{(n-1)s_X^2}} \quad \text{d.f.} = n - 2$$

Depends on how far our new point is from the average of the explanatory values.

Once again, the estimate minus the parameter, divided by the standard error, has a Student's t-distribution with $n - 2$ degrees of freedom.

Leads to 95% CI

$$\hat{\beta}_0 + \hat{\beta}_1 X_0 \pm t_{n-2}(0.975) SE_{\hat{\mu}\{Y|X_0\}}$$

What is a 95% CI for the mean distance of a nebula with velocity of 600 km/sec?

Make a new data.frame with the explanatory variable we want to estimate

```
> newdata <- data.frame(Velocity = 600)
```

```
> newdata
  Velocity
1      600
```

↑
Name needs to match
column in original data.frame

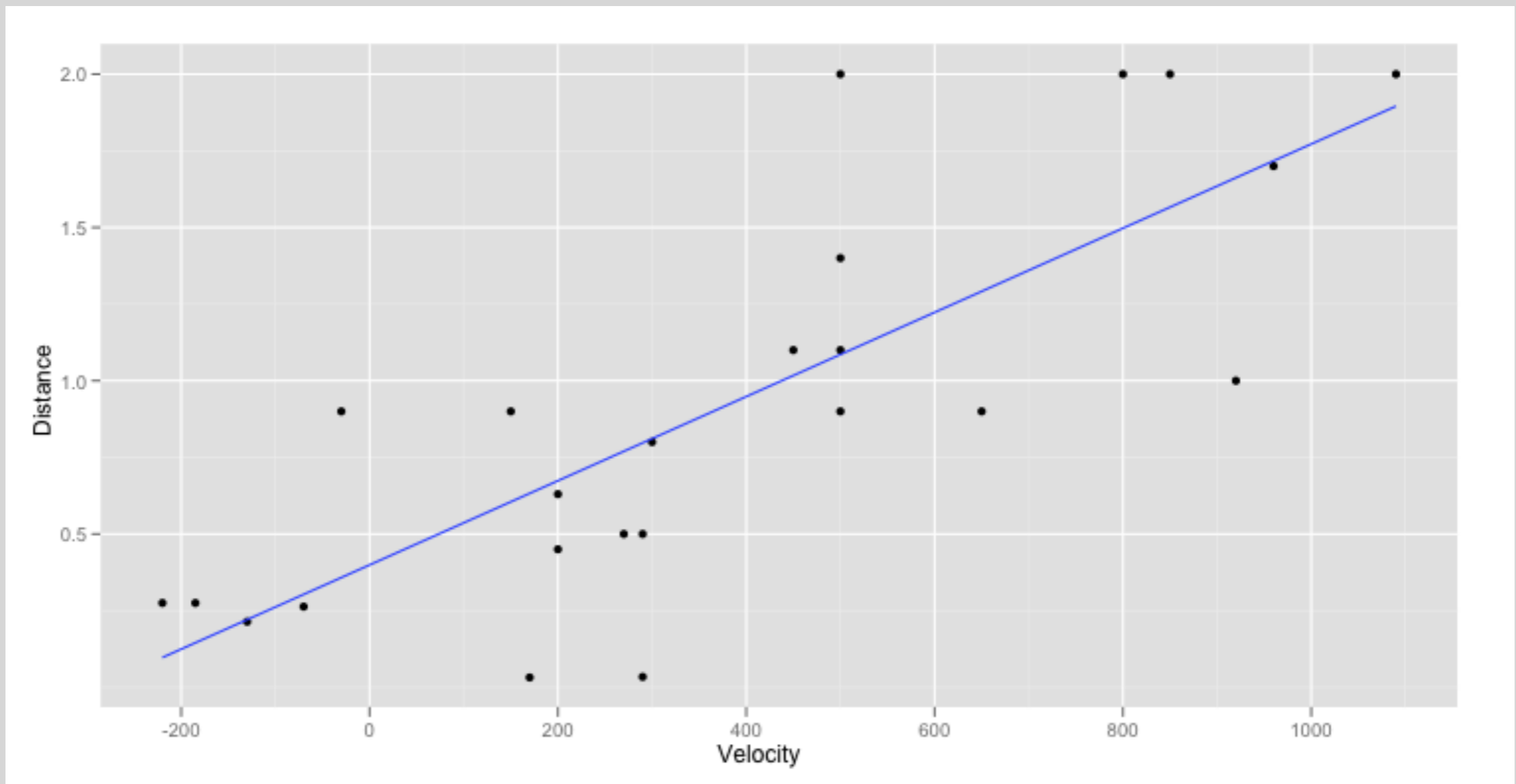
```
> predict(fit, newdata, interval = "confidence")
```

```
      fit      lwr      upr
1 1.22286 1.02077 1.42495
```

With 95% confidence, the **mean distance** of a nebula with a velocity of 600km/sec is between 1.02 and 1.42 parsecs from Earth.

Your turn

Which will have the larger standard error:
estimating the mean distance at a velocity of 200km/sec,
or estimating the mean distance at a velocity of 1000km/sec?



Your turn

Two values

```
> newdata2 <- data.frame(Velocity = c(200, 1000))
```

```
> predict(fit, newdata2, se = TRUE)
```

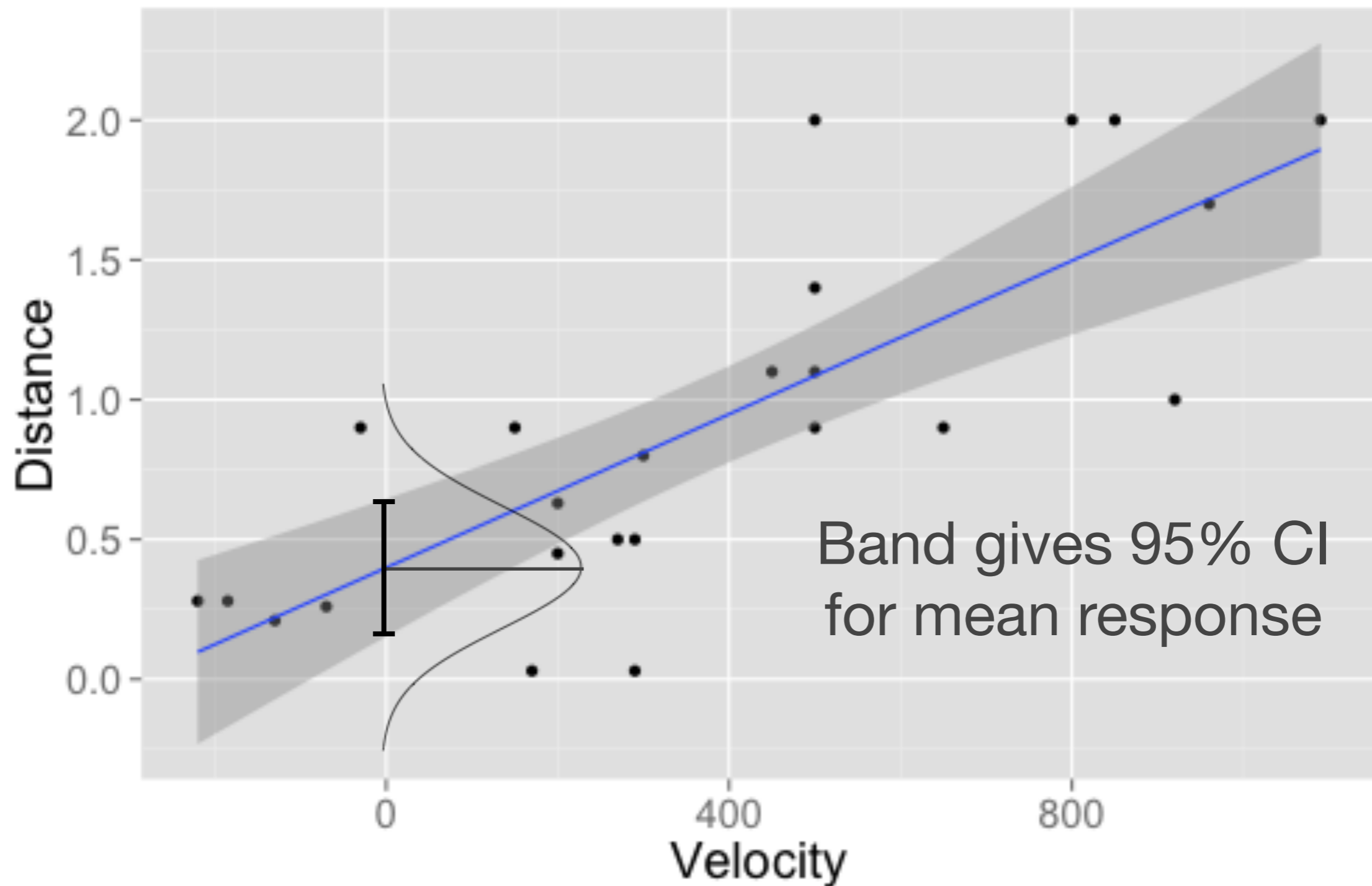
```
$fit
```

```
      1      2  
0.6736854 1.7720343
```

```
$se.fit
```

```
      1      2  
0.09156133 0.16480835
```

```
qplot(Velocity, Distance, data = case0701) +  
  geom_smooth(method = "lm")
```



CI on mean response: likely range for the center of our subpopulations

3

Prediction of new response

uncertainty comes from sampling variability
in both parameters
and variability in subpopulations

Predicting a new response

For a new response, its estimate will be the estimated mean at the explanatory value.

$$\text{Pred}(Y|X_0) = \hat{\mu}\{Y|X_0\} = \hat{\beta}_0 + \hat{\beta}_1 X_0$$

Its standard error will be larger because we need to add the uncertainty due to the **variation of the response around the mean (σ)**.

Standard error on prediction

$$SE_{Pred(Y|X_0)} = \sqrt{\hat{\sigma}^2 + SE_{\hat{\mu}\{Y|X_0\}}^2}$$

always bigger than the SE on the mean response

We are more uncertain about the response of a single unit with explanatory value X_0 , than we are about the mean of all units with the explanatory value, X_0

Once again, the estimate minus the parameter, divided by the standard error, has a Student's t-distribution with $n - 2$ degrees of freedom.

Leads to 95% prediction intervals

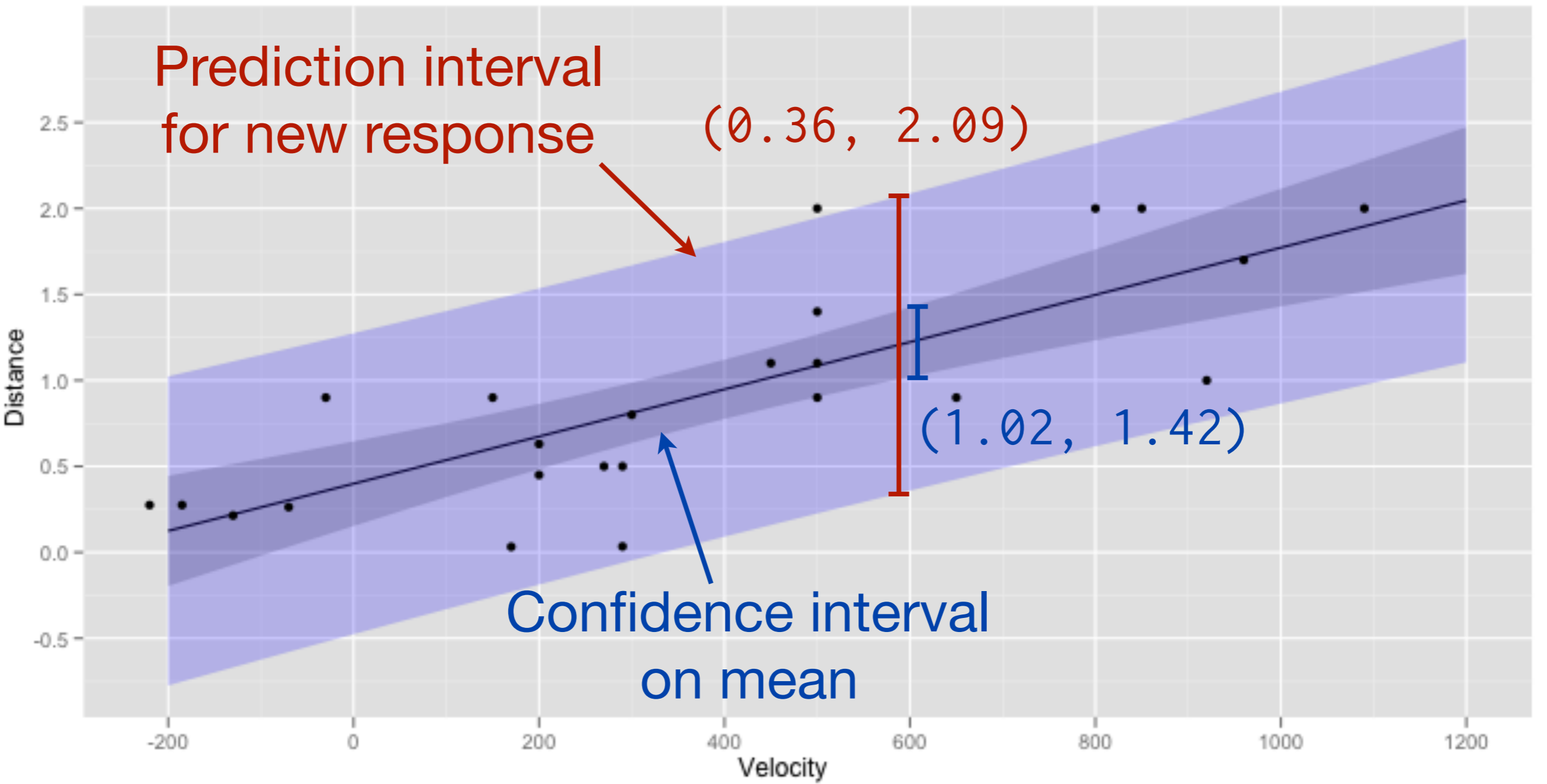
$$\hat{\beta}_0 + \hat{\beta}_1 X_0 \pm qt(0.975, n - 2) SE_{Pred\{Y|X_0\}}$$

What is a 95% prediction interval for the distance of a nebula with velocity of 600 km/sec?

```
> predict(fit, newdata, interval = "prediction")
```

```
      fit      lwr      upr
1 1.22286 0.3590542 2.086666
```

A 95% prediction interval for the **distance** of a nebula with a velocity of 600km/sec is between 0.36 and 2.09 parsecs from Earth.



PI on response: likely range for observations from our subpopulations