Inference for Predictions

Math 430, Winter 2017

Recall: The fitted model

```
climate.lm <- lm(globaltemp ~ co2, data = climate)
summary(climate.lm)
```

```
##
## Call:
## lm(formula = globaltemp ~ co2, data = climate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.24377 -0.08048  0.01431  0.07905  0.22558
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.9083486  0.1943286  -14.97   <2e-16 ***
## co2          0.0087761  0.0005527   15.88   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1016 on 54 degrees of freedom
## Multiple R-squared:  0.8236,    Adjusted R-squared:  0.8204
## F-statistic: 252.2 on 1 and 54 DF,  p-value: < 2.2e-16
```

Prediction in Regression

Key Question:

What do we want to predict?

- the mean response (μ_Y) for a particular value of x?
- or the response ($\widehat{y}$) for an individual (future) case?

Point Estimate:

- We use $\widehat{\beta}_0 + \widehat{\beta}_1 x$ to obtain our "best guess" in both situations.
- But the two situations are *very* different, which is reflected in their SEs

Intervals for Predictions

SEs:

Intervals for Predictions

Jargon:

- **Confidence interval** for the mean response, μ_Y
- **Prediction interval** for a single (future) observation, y

Intervals for Predictions in R

Suppose we wish to predict the global temperature for a CO_2 level of 400

```
new.df <- data.frame(co2 = 400)
predict(climate.lm, newdata = new.df, interval = "confidence")
```

```
##         fit       lwr       upr
## 1 0.6021035 0.5411327 0.6630742
```

```
predict(climate.lm, newdata = new.df, interval = "predict")
```

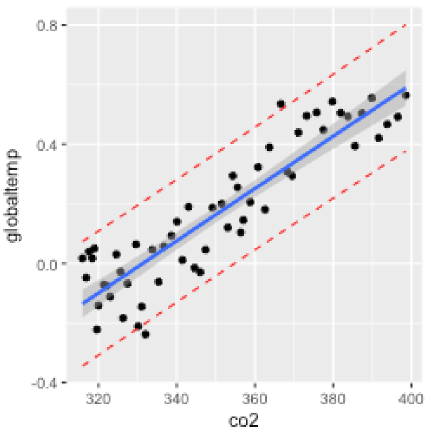
```
##         fit       lwr       upr
## 1 0.6021035 0.3895533 0.8146536
```

which interval should we choose? Why?

Intervals for Predictions in R

We can also obtain predictions for all observations in our data set

```
predict(climate.lm, interval = "confidence")
predict(climate.lm, interval = "prediction")
```



Regression Assumptions

What happens if our assumptions aren't valid?

- **Linearity**: if nonlinear, everything breaks!
- **Independence**: estimates are still unbiased (i.e. we fit the right line) but measures of the accuracy of those estimates (the SEs) are typically too small
- **Normality**: estimates are still unbiased (i.e. we fit the right line), SEs are correct BUT confidence/prediction intervals are wrong (we can't use t-distribution)
- **Constant error variance**: estimates are still unbiased but standard errors are wrong (and we don't know how wrong)