MAT 130, Handout 29: Prediction intervals for y & mean of y

LEARNING OBJECTIVES. *After this class you should be able to...*

- *use R to create confidence intervals for the average value of the response variable given a set of values for the predictor variables.*
- *use R to create prediction intervals for the value of the response variable for a particular individual given a set of values for the predictor variables.*
- *explain the difference between these two kinds of intervals.*
- *determine how the width of these intervals changes as the sample size changes, the confidence level changes, and the values of the predictor variables change.*

Save the script file from today as `H29-Prediction_intervals.R`

Example 1. We previously used the `Galton` data frame, to build a linear model for the heights (in inches) of adult children as a function of the heights of their parents and their sex at birth in the 1880s.

```
> library(mosaicData)
> height_lm = lm(height~father+mother+sex, data = Galton)
> coef(height_lm)
```

```
(Intercept)      father      mother        sexM
 15.3447600   0.4059780   0.3214951   5.2259513
```

$$height = 15.34 + 0.41 \cdot father + 0.32 \cdot mother + 5.22 \cdot sex_M$$

Question 1. For the variable `sex`, how did R pair Female & Male with 0 & 1? How much taller, on average, is a Male than a Female according to this model?

Question 2. Use the `predict` command to find the expected height of the daughter of a 62 inch tall mother and a 65 inch tall father from this time period. Based on the value of R_a^2 (found in the `summary`), how reliable do you think this prediction is?

Regression Model

Our regression model for $y = height$ is

$$y = \beta_0 + \beta_{mother} \cdot mother + \beta_{father} \cdot father + \beta_{sex_M} \cdot sex_M + \epsilon$$

- ϵ represents how far a particular child's height differs from the regression equation value given by the other terms involving the β's.
- The β's themselves are unknown parameters.
- We know how to find confidence intervals for each β that capture our uncertainty in their values.

Question 3. Even if we could somehow know the values of the β's exactly, why would it still not be possible to predict the height of a particular child exactly?

Interval Estimates: There are two different questions that we can answer with interval estimates, and their difference is *subtle*.

- What is the average height of all children of a particular sex whose parents heights are particular values?
- What is the height of one particular child of a particular sex whose parents heights are particular values?
- The first of these intervals is called a ______________ ______________ for the average value of the response variable.
- The second of these intervals is called a ______________ ______________ for a particular value of the response variable.

Question 4. One of these intervals is **always** wider than the other. Which one do you think is wider and why?

Interval Estimates in R The same `predict` command that we have used to find point estimates can also be used to find confidence intervals for the average value of y as well as prediction intervals for the value of a particular y. The syntax is identical to what we have seen except for the extra flag at the end of the command to specify: `interval = ‘confidence’`, or `interval = ‘prediction’`. In addition, we have the ability to specify the confidence level of the interval with the same `level` flag as we have previously.

```
predict(height_lm, data.frame(mother=62, father=65, sex=‘F’), interval = ‘confidence’)
predict(height_lm, data.frame(mother=62, father=65, sex=‘F’), interval = ‘prediction’)
```

Question 5. How much wider is the 95% prediction interval for a particular son’s height than the 95% confidence interval for the average of all sons’ heights born to a father who is 72 inches and a mother who is 70 inches?

Example 2. Predicting *mpg* using *wt*, *cyl*, and *hp* in the `mtcars` data frame.

Question 6. Consider the 1974 Saab Sonett III, a car that was not included in the `mtcars` data frame, and one for which it is difficult to find any record of its fuel efficiency. This vehicle weighed 1940 lb and had a 4 cylinder, 75 hp engine. Use the `predict` function to find a point estimate for the fuel efficiency of this car.

Question 7. Based on the value of R_a^2, do you think that this prediction is reliable? How reliable? (Saab did not publish an official value for this vehicle, so we have no official value that we can compare our prediction to.)

Question 8. Find a 99% confidence interval for the average value of the fuel efficiency of vehicles with the values specified for the 1974 Saab Sonett III. Also find a 99% prediction interval for an individual 1974 Saab Sonnett III. Explain why the prediction interval is wider.

Remark. An internet blog from a Saab enthusiast reports that the actual fuel efficiency of this model vehicle is 27.2 mpg.

Question 9. Find the mean value of each predictor variable from the **mtcars** data frame. Record their values below. Then write down the 95% confidence interval for the average *mpg* as well as the 95% prediction interval for the *mpg* of a vehicle with these values of the predictors.

- wt:
- hp:
- cyl:
- CI:
- PI:

Question 10. Find the minimum value of each predictor variable from the **mtcars** data frame. Record their values below. Then write down the 95% confidence interval for the average *mpg* as well as the 95% prediction interval *mpg* for a vehicle with these values of the predictors. How does the width of these intervals compare to the ones in the previous question?

- wt:
- hp:
- cyl:
- CI:
- PI:
- Compare widths:

Remark. Interval estimates always grow wider as the values of the predictor variables move further away from the sample averages. Does this behavior make intuitive sense? Why or why not?