MAT 130, Handout 27: Prediction intervals for y & mean of y *(Ch 24, 25)*

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 `29Nov-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.

```
> 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?

Solution: We plug in 1 for `sexM` when the individual is male, and 0 for `sexM` when the individual is female. On average, a male is 5.22 inches taller than a female.

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?

Solution:

```
> predict(height_lm, data.frame(mother=62, father=65, sex='F'))

       1
61.66603
```

With $R_a^2 = 64\%$, we recognize that this estimate may not be extremely accurate. It is difficult with the tools we have used, to quantify just how good or bad this estimate might be.

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?

Solution: Because the ϵ is a random quantity specific to each child. There's no way to predict its exact value by studying other children.

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 **confidence interval** for the average value of the response variable.
- The second of these intervals is called a **prediction interval** 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?

Solution: The prediction interval is always wider. A particular person could certainly take on any of the values that are plausible for the average, but this person may not be average. They might be taller or shorter than average, so the interval to describe their height must be even wider than the interval that describes the height of the average person.

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?

Solution: The width of the prediction interval is 8.50 inches while the confidence width interval is just 0.88 inches. The difference is 7.63 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.

Solution:

```
> mpg_lm = lm(mpg~wt+cyl+hp, data = mtcars)
> predict(mpg_lm, data.frame(wt=1.940, cyl=4, hp=75))

       1
27.48853
```

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.)

Solution:

```
> summary(mpg_lm)

Call:
lm(formula = mpg ~ wt + cyl + hp, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9290 -1.5598 -0.5311  1.1850  5.8986

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 38.75179    1.78686  21.687  < 2e-16 ***
wt          -3.16697    0.74058  -4.276 0.000199 ***
cyl         -0.94162    0.55092  -1.709 0.098480 .
hp          -0.01804    0.01188  -1.519 0.140015
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 2.512 on 28 degrees of freedom
Multiple R-squared:  0.8431,        Adjusted R-squared:  0.8263
F-statistic: 50.17 on 3 and 28 DF,  p-value: 2.184e-11
```

$R_a^2 = 0.83$ suggests that this model should make reasonably accurate predictions. How reliable? It's difficult to make a quantitative statement about how far off this estimate might be without additional statistical tools.

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.

Solution:

```
> predict(mpg_lm, data.frame(wt=1.940, cyl=4, hp=75), level = 0.99,
+         interval = "confidence")

       fit      lwr      upr
1 27.48853 25.40999 29.56708

> predict(mpg_lm, data.frame(wt=1.940, cyl=4, hp=75), level = 0.99,
+         interval = "prediction")

       fit      lwr      upr
1 27.48853 20.24389 34.73318
```

Even after we have an interval that we are confident contains the average, we need to make that interval even wider to account for the fact that this particular model might be above or below average.

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:

Solution:

```
> summary(mtcars)

      mpg             cyl             disp             hp
 Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0
 1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5
 Median :19.20   Median :6.000   Median :196.3   Median :123.0
 Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7
 3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0
 Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0
      drat             wt             qsec             vs
 Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000
 1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000
 Median :3.695   Median :3.325   Median :17.71   Median :0.0000
 Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375
 3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000
 Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000
       am              gear            carb
 Min.   :0.0000   Min.   :3.000   Min.   :1.000
 1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000
 Median :0.0000   Median :4.000   Median :2.000
 Mean   :0.4062   Mean   :3.688   Mean   :2.812
 3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000
 Max.   :1.0000   Max.   :5.000   Max.   :8.000

> predict(mpg_lm, data.frame(wt=3.217, hp=146.7, cyl=6.188),
+         interval='confidence')

       fit      lwr      upr
1 20.09072 19.18126 21.00018

> predict(mpg_lm, data.frame(wt=3.217, hp=146.7, cyl=6.188),
+         interval='prediction')

       fit      lwr      upr
1 20.09072 14.86628 25.31516
```

- wt: 3.217
- hp: 146.7
- cyl: 6.188
- CI: (19.2, 21.0)
- PI: (14.9, 25.3)

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:

Solution: The `summary()` function gives the min values. Then use `predict()` again to get the intervals.

```
> predict(mpg_lm, data.frame(wt=1.513, hp=52.0, cyl=4.000),
+         interval='confidence')

       fit      lwr      upr
1 29.25571 27.30804 31.20337

> predict(mpg_lm, data.frame(wt=1.513, hp=52.0, cyl=4.000),
+         interval='prediction')

       fit     lwr      upr
1 29.25571 23.7547 34.75671
```

- wt: 1.513
- hp: 52.0
- cyl: 4.000
- CI: (27.3, 31.2)
- PI: (23.8, 34.8)
- Compare widths: The intervals are higher and wider than the ones from the previous problem.

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?

Solution: When the values of the predictor variables move further away from the average the new fitted value becomes skewed.