Transformed Regression, Prediction Interval, and T test

(Jing Li, Miami University)

1. This note discusses how to obtain prediction interval (or interval forecast) and run T test with the technique of transformed regression.

2. Consider House data, and we want to predict the average price for a house with two bathrooms. The R codes and results of regressing price onto baths are below

```
> ad = "https://www.fsb.miamioh.edu/lij14/400_house.txt"
> da = read.table(url(ad), header=T)
> m = lm(rprice~baths,data=da)
> summary(m)$coef
            Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 14510.80   4299.715  3.374829 8.297909e-04
baths       29582.67   1745.859 16.944481 2.225878e-46
```

3. The general formula for prediction is

$$\hat{y} \equiv \hat{E}(y|x = c) = \hat{\beta}_0 + \hat{\beta}_1 c \tag{1}$$

For this example, the predicted average price when baths equal two is 73676.15:

```
> # predict E(y|x=c)
> c = 2
> yhat = summary(m)$coef[1,1]+summary(m)$coef[2,1]*c
> yhat
[1] 73676.15
```

The R built-in function **predict** can be applied here:

```
> newda = data.frame(baths=c)
> predict(m,newdata=newda)
73676.15
```

4. It is harder to obtain a prediction interval that looks like

$$\hat{y} \pm criticalvalue * se \tag{2}$$

The challenge is finding the standard error (se), the square root of variance of $\hat{y}$. Let us try using math to derive the variance. First, rewrite $\hat{y}$ as

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 c = (\bar{y} - \hat{\beta}_1 \bar{x}) + \hat{\beta}_1 c = \bar{y} + \hat{\beta}_1 (c - \bar{x}) \tag{3}$$

where we use the OLS formula for the intercept $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$. Let $\sigma^2 = var(u|x)$ be the conditional variance of error term, which is also the <u>conditional</u> variance of y. Then it follows that

$$var(\hat{y}) = var(\bar{y}) + var(\hat{\beta}_1(c - \bar{x})) = \sigma^2 \left(\frac{1}{n} + \frac{(c - \bar{x})^2}{\sum_i (x_i - \bar{x})^2} \right) \tag{4}$$

where we use the facts that $var(\bar{y}) = \frac{\sigma^2}{n}$ and $var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}$. The R codes to obtain the standard error (square root of $var(\bar{y})$) and prediction intervals are

```
> varyhat = summary(m)$sigma^2*(1/length(da$baths)+(c-mean(da$baths))^2/sum((da$bat
> seyhat = sqrt(varyhat)
> seyhat
[1] 1468.147

> yhat-qt(0.975,319)*seyhat
[1] 70787.67
> yhat+qt(0.975,319)*seyhat
[1] 76564.62
```

where the critical value is from the T distribution with $n - k - 1 = 319$ degree of freedom.

5. In this case, the 95 percent prediction interval of rprice for a house with two bathrooms are $(70787.67, 76564.62)$

6. It turns out that there is an equivalent but simpler way to compute the standard error and obtain prediction interval. The key is running a transformed regression.

   (a) First, define a new regressor w as the difference between the original regressor and the value used for prediction

   $$w = x - c$$

   Then algebra rearrangement of original regression leads to a transformed regression using w as regressor:

   $$y = \beta_0 + \beta_1 x + u = \beta_0 + \beta_1(w + c) + u = (\beta_0 + \beta_1 c) + \beta_1 w + u \tag{5}$$

   (b) Notice that the constant term (intercept) $\beta_0 + \beta_1 c$ in the transformed regression is the same as yhat. Unsurprisingly, its standard error gives us the se of yhat

```
> da$w = da$baths-c
> m.t = lm(rprice~w,data=da)
> summary(m.t)$coef
            Estimate Std. Error  t value      Pr(>|t|)
(Intercept) 73676.15   1468.147 50.18308 1.952704e-153
w           29582.67   1745.859 16.94448  2.225878e-46
```

   As expected, the standard error of intercept 1468.147 is what we seek.

   (c) Even better, we can use built-in function **confint** to obtain the confidence interval for the intercept, which is also the prediction interval for yhat

```
> confint(m.t)
               2.5 %   97.5 %
(Intercept) 70787.67 76564.62
w           26147.82 33017.53
```

   (d) Function**predict** provides the same interval

```
> predict(m,newdata=newda,interval="confidence",level=0.95)
       fit      lwr      upr
1 73676.15 70787.67 76564.62
```

(e) We may predict <u>individual</u> price other than average price for a house with two bathrooms

```
> # predict individual y
> yhat-qt(0.975,319)*sqrt(summary(m.t)$coef[1,2]^2+summary(m)$sigma^2)
[1] 26243.5
> yhat+qt(0.975,319)*sqrt(summary(m.t)$coef[1,2]^2+summary(m)$sigma^2)
[1] 121108.8
> predict(m,newdata=data.frame(baths=c(2)),interval="prediction",level=0.95)
       fit      lwr      upr
1 73676.15 26243.5 121108.8
```

In this case, the standard error of individual yhat is

$$se(individual\ \hat{y}) = \sigma\sqrt{\left(1+\frac{1}{n}+\frac{(c-\bar{x})^2}{\sum_i(x_i-\bar{x})^2}\right)} \tag{6}$$

7. Now consider how to obtain a transformed regression in order to test the single hypothesis of <u>linear</u> combination of coefficients such as

$$H_0 : \beta_1 + \beta_2 = c \tag{7}$$

after fitting a multiple regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \tag{8}$$

(a) We can define a new parameter γ as

$$\gamma = \beta_1 + \beta_2 - c \tag{9}$$

Note that under H_0 we have $\gamma = 0$.

(b) Next replace β_1 with $\gamma - \beta_2 + c$ in the original regression

$$y = \beta_0 + (\gamma - \beta_2 + c)x_1 + \beta_2 x_2 + u \tag{10}$$

Simplify. Then we have

$$y - cx_1 = \beta_0 + \gamma x_1 + \beta_2(x_2 - x_1) + u \tag{11}$$

(c) For the transformed regression, the dependent variable is $y - cx_1$, and regressors are x_1 and $x_2 - x_1$.

(d) Now we need to test the hypothesis

$$H_0 : \gamma = 0$$

and the t value is automatically reported by R (for coefficient of x_1) after we run the transformed regression.

(e) For example, considering testing $H_0 : \beta_1 + \beta_2 = 1$ for the multiple regression that regresses rprice onto age and baths

```
> da$newy = da$rprice-da$age
> da$newx1 = da$age
> da$newx2 = da$baths-da$age
> summary(lm(newy~newx1+newx2,data=da))$coef
            Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 19865.82   4871.209  4.07821 5.741383e-05
newx1       27965.67   1872.823 14.93236 1.417865e-38
newx2       28067.13   1856.694 15.11673 2.786054e-39
```

The t test is the t value of newx1, 14.93236.

(f) The squared t value 222.9754 is F test, which be obtained by **linearHypothesis** function in **car** package

```
> 14.93236^2
[1] 222.9754
> library(car)
> m = lm(rprice~age+baths,data=da)
> linearHypothesis(m, matrix(c(0, 1, 1), nrow = 1), 1)
Hypothesis: age  + baths = 1
Model 1: restricted model
Model 2: rprice ~ age + baths

  Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
1    319 3.0917e+11
2    318 1.8174e+11  1 1.2743e+11 222.98 < 2.2e-16 ***
```