GBA 462R Lab Session 6 (Week 8)

Yulin Hao

2022-10-20

Logistics

- Office hour: Thu 7-9 pm
- Final review session: Sunday 10 am-noon

How to prepare for the final?

- Put more focus on regression-related stuff
  - **calculation questions matter, tricks matter**
    - Try to relate concepts (e.g., F~R^2, F~t, SER~SSR)
- If you have really tight schedule, prioritize (**ALL**) lab materials
  - Formulas shown up in lab materials are almost all you need
- Come to the office hour and schedule meetings with me
  - I will be available throughout the weekend
  - Group (in-person) meeting is a good idea

Case Study: Employee tenure and store performance (Similar to HW5)

You run a family business that sell fresh farm goods and you have opened serveral stores in the local area. To boost the financial performance, you want to have a better idea of the relationship of employee retention and store profit. A file named **lab_data.csv** contains data you have regarding the store-level characteristics in the following 9 aspects: *Profit*, *MTenure* (on average, how many months have a manager been with this store), *CTenure* (on average, how many months have a crew member been with this store), *Comp* (number of competitiors per 100,000 people in the city), *Pop* (city population), *Visiblity* (visibility rating from 1 to 5, with 5 being the highest), *PedCount* (pedestrian foot traffic rating from 1 to 5, with 5 being the highest), *Hours24* (indicator equal to 1 if the store opens 24 hours), *Res* (indicator equal to 1 if the city is largely a residential, as opposed to industrial area).

1.Run a regression of *Profit* on all the explanatory variables above. Based on your regression results, intepret the coefficient of *MTenure* and *CTenure* and evaluate their statistical significance at the 1% significance level. What about *Res*?

```
####### import data
df = read.csv("C:/Users/adminPC/Desktop/lab_data.csv")
head(df)
```

```
##   Store    Sales   Profit   MTenure   CTenure   Pop     Comp Visibility
## 1     1 5301.470 1766.760   0.80000 25.804930  7535 3.637254          4
## 2     2 8099.370 2826.713  87.02219  7.636550  8630 5.506221          5
## 3     3 5499.605 1484.900  24.68854  6.026694  9695 5.843066          4
## 4     4 5269.300 1400.813   0.80000  6.371663  2797 5.530130          5
## 5     5 6139.205 2003.200   4.67737  7.866530 20335 2.146773          3
## 6     6 8515.700 3127.000 150.73590 12.351130 16926 4.139997          4
##   PedCount Res Hours24
## 1        3   1       1
## 2        3   1       1
## 3        3   1       1
## 4        2   1       1
## 5        5   0       1
## 6        4   1       0
```

```
names(df)
```

```
##  [1] "Store"      "Sales"      "Profit"     "MTenure"    "CTenure"
##  [6] "Pop"        "Comp"       "Visibility" "PedCount"   "Res"
## [11] "Hours24"
```

```
reg=lm(Profit~MTenure+CTenure+Pop+Comp+Visibility+PedCount+Res+Hours24, data=df)
summary(reg)
```

```
##
## Call:
## lm(formula = Profit ~ MTenure + CTenure + Pop + Comp + Visibility +
##     PedCount + Res + Hours24, data = df)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -654.75 -236.98  -49.85  234.63  767.31
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  145.33906  560.15408   0.259 0.796436
## MTenure        5.48563    0.98766   5.554 1.34e-06 ***
## CTenure        5.31205    3.03213   1.752 0.086452 .
## Pop            0.03186    0.01073   2.970 0.004717 **
## Comp        -123.98985   30.20495  -4.105 0.000164 ***
## Visibility    36.37311   74.28626   0.490 0.626720
## PedCount     142.43616   71.77287   1.985 0.053179 .
## Res          709.32388  335.53789   2.114 0.039968 *
## Hours24      485.40570  151.76856   3.198 0.002503 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 368.3 on 46 degrees of freedom
## Multiple R-squared:  0.6662, Adjusted R-squared:  0.6081
## F-statistic: 11.47 on 8 and 46 DF,  p-value: 9.008e-09
```

- Coefficient of MTenure is roughly 5.5. Increase manager tenure by 1 month **is associated with** an increase in profit of 5.5 **holding other things constant**; MTenure is statistically significant at the 1% level (p-value < 0.01)
- Interpreation of CTenure is similar to MTenure, CTenure is significant at the 10% level (p-value < 0.1)
- Res: Compared with industrial area, locating at residential area would bring a higher profit by 709 **holding other things constant**, it is significant at the 5% level (p-value < 0.05);

Note: - Remember how to interpret dummy variables!

2. Based on your results, what is your estimate of the impact of a 1.412-month crew tenure increase on store profits? How would you proceed with estimate to better evaluate your earlier contracting plan?

```r
#extract regression coefficients
reg$coefficients["CTenure"]*1.412
```

```
##  CTenure
## 7.500618
```

```r
#Alternatively,
reg$coefficients[3]*1.412
```

```
##  CTenure
## 7.500618
```

Note: - What is the effect on profit if crew tenure increase from 1 to 4? - Change 1.412 to 3

3.Construct a 95% CI for the effect of *Comp* on profits and interpret the CI.

```
#95% CI for all variables
confint.default(reg, level=0.95)
```

```
##                    2.5 %        97.5 %
## (Intercept) -952.5427580 1243.22086963
## MTenure        3.5498550    7.42140838
## CTenure       -0.6308188   11.25492313
## Pop            0.0108357    0.05288129
## Comp        -183.1904630  -64.78923001
## Visibility  -109.2252807  181.97150309
## PedCount       1.7639255  283.10839110
## Res           51.6816991 1366.96606139
## Hours24      187.9447848  782.86660997
```

```
#95% CI  for com
confint.default(reg, level=0.95)["Comp",]
```

```
##      2.5 %     97.5 %
## -183.19046  -64.78923
```

- With 95% probability, we believe the **population coefficient** of Com is between -184.8 and -63.2. In other words, the **true effect** of Com on profit is between -184.8 and -63.2 with 95% probability.

Note: - Remember how to interpret CIs!

4.Among the explanatory variables, *Comp*, *Pop*, *Visiblity*, *PedCount*, *Hours24* and *Res* are location-based. What are the roles of these location-based variables? Do you expect them to have a certain sign? Rerun the earlier regression with these location-based variables dropped. What does the results tell you?

```
reg2=lm(Profit~MTenure+CTenure, data=df)
summary(reg2)
```

```
##
## Call:
## lm(formula = Profit ~ MTenure + CTenure, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -803.45 -319.22  -44.98  268.40 1257.98
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1596.646    101.035  15.803  < 2e-16 ***
## MTenure        4.438      1.199   3.701  0.00052 ***
## CTenure        4.803      3.662   1.311  0.19545
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 517.6 on 52 degrees of freedom
## Multiple R-squared:  0.2546, Adjusted R-squared:  0.2259
## F-statistic: 8.881 on 2 and 52 DF,  p-value: 0.0004807
```

- Adj-R2 dropped from 0.6 to 0.23, model fit is worse (R^2 decrease for sure! Not very helpful here).
- Omitting location-based variables causes omitted variable bias (estimates of MTenure and CTenure change), also lowers estimate precision (e.g., CTenure's standard error)

5.How would you design the regression if the impact of *MTenure* and *CTenure* on *Profit* varies with the level of tenure? Run the regression and find the point where an additional month of *MTenure* leads to lower profit.

```r
reg3=lm(Profit~MTenure+I(MTenure^2)+CTenure+I(CTenure^2)+Pop+Comp+Visibility+PedCount+Res+Hours24, data=
summary(reg3)
```

```
##
## Call:
## lm(formula = Profit ~ MTenure + I(MTenure^2) + CTenure + I(CTenure^2) +
##     Pop + Comp + Visibility + PedCount + Res + Hours24, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -577.99 -251.07   -6.58  190.62  608.83
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.807e+01  5.100e+02   0.035 0.971897
## MTenure       1.368e+01  2.260e+00   6.052 2.83e-07 ***
## I(MTenure^2) -3.705e-02  9.644e-03  -3.842 0.000388 ***
## CTenure       1.632e+01  7.726e+00   2.112 0.040384 *
## I(CTenure^2) -1.115e-01  7.904e-02  -1.410 0.165484
## Pop           2.416e-02  9.783e-03   2.470 0.017468 *
## Comp         -1.348e+02  2.677e+01  -5.036 8.55e-06 ***
## Visibility    6.022e+01  6.838e+01   0.881 0.383320
## PedCount      1.853e+02  6.478e+01   2.861 0.006436 **
## Res           4.892e+02  3.230e+02   1.515 0.136969
## Hours24       4.488e+02  1.335e+02   3.363 0.001606 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 322.4 on 44 degrees of freedom
## Multiple R-squared:  0.7553, Adjusted R-squared:  0.6996
## F-statistic: 13.58 on 10 and 44 DF,  p-value: 1.833e-10
```

- To find the point where the profit starts to decrease, we just need to find the maximal point, because the profit first increases, then decreases.
  - profit= 13.68MTenure–0.037$MTenure^2$
  - Take derivtive wrt MTenure, 13.68- 0.074MTenure=0 → $MTenure = 184$

Pratical questions from final samples

1. *Omitted variable bias primarily arises from having too few observations (i.e. a small n).*

- False
- OVB: Omitting important variables Z which is correlated with **both** X and Y
- No OVB If Z is a good predictor of Y, but uncorrelated with X.
- Small observations affect standard error (larger SE), but do not affect unbiasedness of the (point) estimates ($\hat{\beta}$).
  - Recall that $SE(\hat{\beta}_1) = \frac{SER}{\sqrt{nVar(X)}}$

2. *Including additional covariates in an OLS regression cannot increase the adjusted* R^2

- False
- Including additional (even bad) covariates increase R^2 for sure, but only including relevant predictors can increase adjusted R^2. Including bad covariates can decrease the adjusted R^2.
- adj$R^2 = 1 - \frac{n-1}{n-p-1}\frac{SSR}{SST}$, adjusted R^2 punishes the number of covariates you include.
  - Both the numerator and denominator decreases as your p increases, which makes the changes in adjusted R^2 uncertain.
- Note :Adjusted R2 is always less than or equal to R2! Adjusted R2 can be negative!

3. *You cannot use* R^2 *to compare the fit of the log-log and quadratic regression models*

- True
- log-log model: $log(Y_i) = \beta_0 + \beta_1 log(X_i) + u_i$
- quadratic model: $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_i$
- Cannot use R^2(or adj R^2) to compare models with different dependent variables (you are predicting different things). But you can use R^2(or adj R^2) to compare models with identical dependent variables but different independent variables.

4. *Suppose that, for the population mean and using a standard normal distribution, you calculate a confidence interval of (0.968,3.032). If the value of the standard error associated with the sample mean was .4, what was the size of this confidence interval*

- CI is symmetric around sample parameters (here is sample mean)
- $\bar{X} = \frac{0.968+3.032}{2} = 2$
- $\bar{X} - zSE(\bar{X}) = LB$, $z = \frac{\bar{X}-LB}{SE(\bar{X})} = \frac{2-0.968}{0.4} = 2.58$
- 2.58 is the critical value for 99% CI (1.96 for 95%, 1.64 for 90%).
- **Note:** You should be able to obtain sample parameters, SE, t-stat and even F-stat (for simple regression) given CIs

5. *Including both X and $3 + 3X^2$ as covariates in a regression will violate OLS Assumption 4 (no perfect collinearity)*

- False
- X and $3 + 3X^2$ are not perfectly col-linear.
- When include X, only including linear function of X will fail
  - You CANOT include 2X, X+2, 2X+2 if you already have X in the model
  - You CAN include X^2, $log(X)$, $exp(X)$

6. *Suppose that, in a regression with 203 observations and two regressors, you find that the sum of squared residuals is one-fourth of the explained sum of squares. What is the F-statistic that is automatically reported by Excel equal to*

- $SSR = \frac{1}{4}ESS$
- Let's say ESS=4, SSR=1, SST=5, then $R^2 = \frac{ESS}{SST} = 0.8$
- $F = \frac{R^2/k}{(1-R^2)/(n-k-1)} = \frac{0.8/2}{0.2/200} = 400$

7. *Suppose we have data on Sales and Advertising. In a univariate regression of Sales on ln(Advertising), you find the intercept to be -4000, the slope coefficient to be 2500. Also, the F-statistic reported automatically by Excel is 100. Consider increasing advertising expenditures from 200 to 204. What is the 90% confidence interval for the change in Sales ΔY*

- $\hat{Y} = -4000 + 2500ln(X)$
- In simple regression, $F = t^2 = 100$, so $t = 10$
- $t = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} \rightarrow se(\hat{\beta}_1) = 2500/10 = 250$
- **To construct CI for ΔY, we need to know $\Delta\hat{Y}$ and $se(\Delta\hat{Y})$**
  - $\Delta\hat{Y} = \hat{\beta}_1 \Delta ln(X) = 2500\Delta ln(X) = 2500 \times \frac{204-200}{200} = 50$ since $\Delta ln(ad) = \frac{\Delta ad}{ad}$ by linear approximation
  - $se(\Delta\hat{Y}) = se(\hat{\beta}_1) \times \Delta ln(X) = 250 \times \frac{204-200}{200} = 5$
  - 90% CI: $[\Delta\hat{Y} - 1.64se(\Delta\hat{Y}), \Delta\hat{Y} + 1.64se(\Delta\hat{Y})] = [50 - 1.64 \times 5, 50 + 1.64 \times 5] = [41.8, 58.2]$
- **Note:** Remember the approximation trick

8. If $Cov(X1, X2) > 0, \alpha_1 > 0, \alpha_2 < 0$,then $\beta_1 < \alpha_1$ in the regressions below because of the omitted variable bias.

$$Y = \beta_0 + \beta_1 X_1 + u$$

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + u$$

- True
- Equation 1 is the estimated model (wrong model), Equation 2 is the true model
- The direction of the OVB depends on the actual signs of both α_2 (the true impact of X2(omitted variable) on Y) and $cov(X1, X2)$
- $Cov(X1, X2) > 0, \alpha_2 < 0 \rightarrow$ negative bias(underestimate)$\rightarrow \beta_1 < \alpha_1$
- **Note 1:** $++$ or $--$ $\rightarrow$ positive bias (overestimate); $+-$ or $-+$ $\rightarrow$ negative bias (underestimate)
- **Note 2:** Sometimes you need to infer the sign of cov(X1,X2) and α_2 based on intuition/economic reasoning!

9. *Suppose that, in a quadratic regression of Y on X and X^2, you find the coefficient on X to be 5.1 and the coefficient on X^2 to be -.05. What is the expected impact on Y of increasing X from 10 to 12*

- $\hat{Y} = \beta_0 + 5.2X - 0.05X^2$
- $E(Y|X = 12) - E(Y|X = 10) = (\beta_0 + 5.1 \times 12 - 0.05 \times 144) - (\beta_0 + 5.1 \times 10 - 0.05 \times 100) = 54 - 46 = 8$

10. *When the dataset is small, the t-stat calculated based on the normal distribution will be different from t-stat reported by Excel.*

- False
- When the dataset is small, the p-value for the t-stat calculated based on the normal distribution will be different from that automatically reported by Excel (which uses t-distribution).
  - Remember how we obtain p-value in R: 2*pnorm(-abs(t))
- T-stat itself does not depend on the distribution (recall t-statistic formula: $t = \frac{\hat{\beta} - \beta^0}{se(\hat{\beta})}$)

Short questions

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.856482284
R Square	0.733561902
Adjusted R Square	0.730411452
Standard Error	8.292030107
Observations	**A**

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression		112068.5857	16009.79795	**B**	2.016E-165
Residual	592	40704.59587	68.7577633		
Total		152773.1815			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-0.630215426	0.341387709	-1.846040178	0.065385416	-1.300693808	0.040262955
X1	4.433171956	**C**	**D**	5.7014E-38	3.804351664	**E**
X2	3.680847373	0.330036266	11.15285729	2.37693E-26	3.03266299	4.329031756
X3	1.201344326	0.333161524	3.60589155	0.00033732	0.547022	1.855666652
X4	2.342426732	0.320332235	7.312491465	8.55316E-13	1.713300865	2.971552599
X5	**F**	**G**	**H**	4.46856E-42	4.43417383	5.798336855
X6	4.865040805	0.346828134	14.02723808	8.4316E-39	4.183877541	5.546204068
X7	9.049965484	0.346901367	26.08800757	1.8659E-100	8.368658393	9.731272576

Figure 1: Question A3

- A: obs=total df +1(for constant) = df(for residual)+df(for regression)+1=592+7+1=600
- B: $F = \frac{R^2/k}{(1-R^2)/(n-k-1)}$, $R^2 = 0.73$, $k = 7$, $n = 600 \rightarrow$ F=232.84
- E: $\hat{\beta}_1 = 4.43 = \frac{LB+UB}{2} = \frac{3.80+UB}{2}$, $\rightarrow UB = 5.06$
- C: $\hat{\beta}_1 - 1.96se(\hat{\beta}_1) = LB$, $\hat{\beta}_1 = 4.43$, $LB = 3.80$, $\rightarrow se(\hat{\beta}_1) = 0.32$
- D: $t = \frac{\hat{\beta}_1 - 0}{se(\hat{\beta}_1)} = 4.43/0.32 = 13.8$
- F: $\hat{\beta}_5 = \frac{LB+UB}{2} = 5.12$
- G: $\hat{\beta}_5 - 1.96se(\hat{\beta}_5) = LB$, $se(\hat{\beta}_5) = 0.348$
- H: $t = \frac{\hat{\beta}_5}{se(\hat{\beta}_5)} = 5.116/0.348 = 14.70$

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.842620357
R Square	0.710009066
Adjusted R Square	0.704322969
Standard Error	8.201325033
Observations	**A**

ANOVA

	df	SS	MS	F	Significance F
Regression	4	**B**		124.8675668	1.06913E-53
Residual	204	**C**			
Total	208	**D**			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-17.72102617	**E**	**F**		-23.5288946	**G**
X1	0.005587134	0.001357196	4.116674079	5.57974E-05	0.002911204	0.008263065
X2	-0.215358409	0.0662176	-3.252283506	0.001339621	-0.345917062	-0.084799757
X3	0.162075267	0.015475629	10.4729356	8.15111E-21	0.131562574	0.192587959
X4	**H**	0.79430469	-0.290575576	0.771671126	-1.796905034	1.335293949

Figure 2: Question A3

- A: obs=total df +1(for constant) = 208+1=209
- C: SER=RMSE(root-mean-square error)=$\sqrt{\frac{SSR}{n-p-1}} = \sqrt{\frac{SSR}{\text{df for residual}}} = 8.20$, SSR=$SER^2 \times$ (df for the residual) $= 8.20^2 \times 204 = 13716.96$
- D: $R^2 = \frac{ESS}{SST} = 1 - \frac{SSR}{SST} = 1 - \frac{13716.96}{SST} = 0.71 \rightarrow$ SST=47299.86
- B: $ESS = SST - SSR = 47299.86 - 13716.96 = 33582.9$
- E, F, H, G standard questions

Hint

- Relate ANOVA with Regression statistics
  - e.g, $F\&R^2$, $SSR\&SER$, $df\&obs$
- For the regression coefficient table:
  - Only one trick: the point estimate is the median point of CI!!!