

# Homework Assignment

## Week 9

Tengyuan Liang  
Business Statistics  
Booth School of Business

### Problem 1: Baseball

Using our baseball data (**RunsPerGame.xls**), regress  $R/G$  on a binary variable for league membership (League = 0 if National and League = 1 if American) and  $OBP$ .

$$R/G = \beta_0 + \beta_1 League + \beta_2 OBP + \epsilon$$

1. Based on the model assumptions, what is the expected value of  $R/G$  given  $OBP$  for teams in the AL? How about the NL?
2. Interpret  $\beta_0$ ,  $\beta_1$  and  $\beta_2$ .
3. After running the regression and obtaining the results, can you conclude with 95% probability that the marginal effect of  $OBP$  on  $R/G$  (after taking into account the League effect) is positive?
4. Test the hypothesis that  $\beta_1 = 0$  (with 99% probability). What do you conclude?

1. The expected value of  $R/G$  given  $OBP$  is

$$E[R/G|OBP, League = 0] = \beta_0 + \beta_2 OBP$$

for the NL and

$$E[R/G|OBP, League = 1] = (\beta_0 + \beta_1) + \beta_2 OBP$$

for the AL.

2.  $\beta_0$  is the number of runs per game we expect a team from the National League to score if their  $OBP$  is zero.

We expect a team in the American League to score  $\beta_1$  more runs per game on average than a team in the National League with the same  $OBP$ .

$\beta_2$  tells us how  $R/G$  scales with  $OBP$ . For every unit increase in  $OBP$  there will be a  $\beta_2$  increase in  $R/G$ .

3. The 95% confidence interval for  $\beta_2$  is  $37.26 \pm 2 * 2.72 = (31.82; 42.70)$  hence, yes, it is greater than zero.

4. The best guess of  $\beta_1$  is  $b_1 = 0.01615$  with standard error 0.06560. Thus the 99% confidence interval is  $b_1 \pm 3 * s_{b_1} = [-0.18, 0.21]$ , which includes zero. Since zero is in our interval of reasonable values we cannot conclude that  $\beta_1 \neq 0$ .

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	-7.72065	0.93031	-8.299	6.59e-09	***
LeagueAmerican	0.01615	0.06560	0.246	0.807	
OBP	37.26060	2.72081	13.695	1.14e-13	***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1712 on 27 degrees of freedom

Multiple R-squared: 0.8851, Adjusted R-squared: 0.8765

F-statistic: 103.9 on 2 and 27 DF, p-value: 2.073e-13

## Problem 2: Beauty Pays!

Professor Daniel Hamermesh from UT's economics department has been studying the impact of beauty in labor income (yes, this is serious research!!).

First, watch the following video:

<http://www.thedailyshow.com/watch/mon-november-14-2011/ugly-people>

It turns out this is indeed serious research and Dr. Hamermesh has demonstrated the effect of beauty into income in a variety of different situations. Here's an example: in the paper "*Beauty in the Classroom*" they showed that "...instructors who are viewed as better looking receive higher instructional ratings" leading to a direct impact in the salaries in the long run.

By now, you should know that this is a hard effect to measure. Not only one has to work hard to figure out a way to measure "beauty" objectively (well, the video said it all!) but one also needs to "*adjust for many other determinants*" (gender, lower division class, native language, tenure track status).

So, Dr. Hamermesh was kind enough to share the data for this paper with us. It is available in our class website in the file "**BeautyData.csv**". In the file you will find, for a number of UT classes, course ratings, a relative measure of beauty for the instructors, and other potentially relevant variables.

1. Using the data, estimate the effect of "beauty" into course ratings. Make sure to think about the potential many "*other determinants*". Describe your analysis and your conclusions.

We talked about this one in class. The main point here is that in order to isolate the effect of beauty into class ratings we need to CONTROL for other potential determinants of ratings. From the data available it looks like all the other variables are relevant so we should be running the following regression:

$$Ratings = \beta_0 + \beta_1 BeautyScore + \beta_2 Female + \beta_3 Lower + \beta_4 NonEnglish + \beta_5 TenureTrack + \epsilon$$

Here are the results:

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	4.06542	0.05145	79.020	< 2e-16 ***
BeautyScore	0.30415	0.02543	11.959	< 2e-16 ***
female	-0.33199	0.04075	-8.146	3.62e-15 ***
lower	-0.34255	0.04282	-7.999	1.04e-14 ***
nonenglish	-0.25808	0.08478	-3.044	0.00247 **
tenuretrack	-0.09945	0.04888	-2.035	0.04245 *

So, as discussed in class it makes sense for some of these coefficients to be negative, right? For example, if an instructor is not a native english speaker he/she might have a harder time communicating the material and hence lower teaching evaluations. Same goes for lower division classes; most people have to take those classes whether they want or not which leads to lower ratings as students are potentially less interested in the materials to begin with. Now, the results for females is a bit surprising. Why are (holding all else equal) females instructors receiving lower ratings on average? Are there any reasons for us to believe females are not as capable as males to teach? Probably not, right? So, this data demonstrates a potential negative bias that people have in evaluating women.

Finally, with all of that taken into account we find that the higher the beauty score of the instructor the higher their ratings!

2. In his paper, Dr. Hamermesh has the following sentence: *“Disentangling whether this outcome represents productivity or discrimination is, as with the issue generally, probably impossible”*. Using the concepts we have talked about so far, what does he mean by that?

The question here is: are beautiful people indeed better teachers or are they just perceived to be better teachers because of their looks? This analysis can't answer this question! In my opinion the results are very suggestive that this is just discrimination as I dont really believe that beauty relates to one's ability to teach. But, until we run an controlled experiment or find a “natural experiment” (like the one in question 3) we can't conclusively prove this point. What would be a potential natural experiment here? Wouldn't it be nice if we had data on blind students taking these classes? Why would that help?

### Problem 3: Housing Price Structure

The file `MidCity.xls`, available on the class website, contains data on 128 recent sales of houses in a town. For each sale, the file shows the neighborhood in which the house is located, the number of offers made on the house, the square footage, whether the house is made out of brick, the number of bathrooms, the number of bedrooms, and the selling price. Neighborhoods 1 and 2 are more traditional whereas 3 is a more modern, newer and more prestigious part of town. Use regression models to estimate the pricing structure of houses in this town. Consider, in particular, the following questions and be specific in your answers:

1. Is there a premium for brick houses everything else being equal?
2. Is there a premium for houses in neighborhood 3?
3. Is there an extra premium for brick houses in neighborhood 3?
4. For the purposes of prediction could you combine the neighborhoods 1 and 2 into a single “older” neighborhood?

There may be more than one way to answer these questions.

- (1) To begin we create dummy variable *Brick* to indicate if a house is made of brick and  $N_2$  and  $N_3$  to indicate if a house came from neighborhood two and neighborhood three respectively. Using these dummy variables and the other covariates, we ran a regression for the model

$$Y = \beta_0 + \beta_1 \text{Brick} + \beta_2 N_2 + \beta_3 N_3 + \beta_4 \text{Bids} + \beta_5 \text{SqFt} + \beta_6 \text{Bed} + \beta_7 \text{Bath} + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2).$$

and got the following regression output.

```

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2159.498   8877.810    0.243 0.808230
BrickYes     17297.350  1981.616   8.729 1.78e-14 ***
N2          -1560.579  2396.765  -0.651 0.516215
N3           20681.037  3148.954   6.568 1.38e-09 ***
Offers      -8267.488  1084.777  -7.621 6.47e-12 ***
SqFt         52.994     5.734    9.242 1.10e-15 ***
Bedrooms    4246.794  1597.911   2.658 0.008939 **
Bathrooms   7883.278  2117.035   3.724 0.000300 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

              (Intercept) -15417.94711 19736.94349
BrickYes       13373.88702 21220.81203
N2             -6306.00785  3184.84961
N3             14446.32799 26915.74671
Offers        -10415.27089 -6119.70575
SqFt           41.64034    64.34714
Bedrooms      1083.04162  7410.54616
Bathrooms     3691.69572 12074.86126

Residual standard error: 10020 on 120 degrees of freedom
Multiple R-squared:  0.8686,    Adjusted R-squared:  0.861
F-statistic: 113.3 on 7 and 120 DF,  p-value: < 2.2e-16

```

To check if there is a premium for brick houses given everything else being equal we test the hypothesis that  $\beta_1 = 0$  at the 95% confidence level. Using the regression output we see that the 95% confidence interval for  $\beta_1$  is [13373.89, 21220.91]. Since this does not include zero we conclude that brick is a significant factor when pricing a house. Further, since the entire confidence interval is greater than zero we conclude that people pay a premium for a brick house.

- (2) To check that there is a premium for houses in Neighborhood three, given everything else we repeat the procedure from part (1), this time looking at  $\beta_3$ . The regression output tells us that the confidence interval for  $\beta_3$  is [14446.33, 26915.75]. Since the entire confidence interval is greater than zero we conclude that people pay a premium to live in neighborhood three.

- (4) We want to determine if Neighborhood 2 plays a significant role in the pricing of a house. If it does not, then it will be reasonable to combine neighborhoods one and two into one “old” neighborhood. To check if Neighborhood 2 is important, we perform a hypothesis test on  $\beta_2 = 0$ . The null hypothesis  $\beta_2 = 0$  corresponds to the dummy variable  $N_2$  being unimportant. Looking at the confidence interval from the regression output we see that the 95% confidence interval for  $\beta_2$  is  $[-6306, 3184]$ , which includes zero. Thus we can conclude that it is reasonable to let  $\beta_2$  be zero and that neighborhood 2 may be combined with neighborhood 1.
- (3) To check that there is a premium for brick houses in neighborhood three we need to alter our model slightly. In particular, we need to add an interaction term  $Brick \times N_3$ . This more complicated model is

$$Y = \beta_0 + \beta_1 Brick + \beta_2 N_2 + \beta_3 N_3 + \beta_4 Bids + \beta_5 SqFt + \beta_6 Bed + \beta_7 Bath + \beta_8 Brick \cdot N_3 + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2).$$

To see what this interaction term does, observe that

$$\frac{\partial E[Y|Brick, N_3]}{\partial N_3} = \beta_3 + \beta_8 Brick.$$

Thus if  $\beta_8$  is non-zero we can conclude that consumers pay a premium to buy a brick house when shopping in neighborhood three. The output of the regression which includes the interaction term is below.

Coefficients:					0.5 %			99.5 %		
	Estimate	Std. Error	t value	Pr(> t )						
(Intercept)	3009.993	8706.264	0.346	0.73016	(Intercept)	-19781.05615	25801.04303			
BrickYes	13826.465	2405.556	5.748	7.11e-08 ***	BrickYes	7529.25747	20123.67244			
N2	-673.028	2376.477	-0.283	0.77751	N2	-6894.11333	5548.05681			
N3	17241.413	3391.347	5.084	1.39e-06 ***	N3	8363.62557	26119.20030			
Offers	-8401.088	1064.370	-7.893	1.62e-12 ***	Offers	-11187.37034	-5614.80551			
SqFt	54.065	5.636	9.593	< 2e-16 ***	SqFt	39.31099	68.81858			
Bedrooms	4718.163	1577.613	2.991	0.00338 **	Bedrooms	588.32720	8847.99967			
Bathrooms	6463.365	2154.264	3.000	0.00329 **	Bathrooms	823.98555	12102.74436			
BrickYes:N3	10181.577	4165.274	2.444	0.01598 *	BrickYes:N3	-722.17781	21085.33248			
---										
					(Intercept)	-19781.05615	25801.04303			
					BrickYes	7529.25747	20123.67244			
					N2	-6894.11333	5548.05681			
					N3	8363.62557	26119.20030			
					Offers	-11187.37034	-5614.80551			
					SqFt	39.31099	68.81858			
					Bedrooms	588.32720	8847.99967			
					Bathrooms	823.98555	12102.74436			
					BrickYes:N3	-722.17781	21085.33248			
					(Intercept)	-19781.05615	25801.04303			
					BrickYes	7529.25747	20123.67244			
					N2	-6894.11333	5548.05681			
					N3	8363.62557	26119.20030			
					Offers	-11187.37034	-5614.80551			
					SqFt	39.31099	68.81858			
					Bedrooms	588.32720	8847.99967			
					Bathrooms	823.98555	12102.74436			
					BrickYes:N3	-722.17781	21085.33248			

Residual standard error: 9817 on 119 degrees of freedom  
Multiple R-squared: 0.8749, Adjusted R-squared: 0.8665  
F-statistic: 104 on 8 and 119 DF, p-value: < 2.2e-16

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

To see if there is a premium for brick houses in neighborhood three we check that the 95% confidence interval is greater than zero. Indeed, we calculate that the 95% confidence interval is  $[1933, 18429]$ . Hence we conclude that there is a premium at the 95% confidence level. Notice however, that the confidence interval at the 99% includes zero. Thus if one was very stringent about drawing conclusions from statistical data, they may accept the claim that there is no premium for brick houses in neighborhood three.

## Problem 4: What causes what??

Listen to this podcast:

<http://www.npr.org/blogs/money/2013/04/23/178635250/episode-453-what-causes-what>

1. Why can't I just get data from a few different cities and run the regression of "Crime" on "Police" to understand how more cops in the streets affect crime? ("Crime" refers to some measure of crime rate and "Police" measures the number of cops in a city)
2. How were the researchers from UPENN able to isolate this effect? Briefly describe their approach and discuss their result in the "Table 2" below.
3. Why did they have to control for METRO ridership? What was that trying to capture?

The problem here is that data on police and crime cannot tell the difference between more police leading to crime or more crime leading to more police... in fact I would expect to see a potential positive correlation between police and crime if looking across different cities as mayors probably react to increases in crime by hiring more cops. Again, it would be nice to run an experiment and randomly place cops in the streets of a city in different days and see what happens to crime. Obviously we can't do that!

What the researchers at UPENN did was to find a natural experiment. They were able to collect data on crime in DC and also relate that to days in which there was a higher alert for potential terrorist attacks. Why is this a natural experiment? Well, by law the DC mayor has to put more cops in the streets during the days in which there is a high alert. That decision has nothing to do with crime so it works essentially as a experiment. From table 1 we see that controlling for ridership in the METRO, days with a high alert (this was a dummy variable) have lower crime as the coefficient is negative for sure. Why do we need to control for ridership in the subway? Well, if people were not out and about during the high alert days there would be fewer opportunities for crime and hence less crime (not due to more police). The results from the table tells us that holding ridership fix more police has a negative impact on crime.

Still we can't for sure prove that more cops leads to less crime. Why? Well, imagine the criminals are afraid of terrorists and decide not to go out to "work" during a high alert day... this would lead to a reduction in crime that is not related to more cops in the streets. But again, I dont believe that is a good line of reasoning so these results are building a very strong circumstantial case that more cops reduce crime.

4. In the next page, I am showing you "Table 4" from the research paper. Just focus on the first column of the table. Can you describe the model being estimated here? What is the conclusion?

In table 4 they just refined the analysis a little further to check whether or not the effect of high alert days on crime was the same in all areas of town. Using interactions between location and high alert days they found that the effect is only clear in district 1. Again, this makes a lot of sense as most of the potential terrorists targets in DC are in District 1 and that's where more cops are most likely deployed to. The effect in the

other districts is still negative but small and given the standard error in parenthesis we conclude it can still be zero (why? check the confidence interval!).

### EFFECT OF POLICE ON CRIME

TABLE 2

TOTAL DAILY CRIME DECREASES ON HIGH-ALERT DAYS

	(1)	(2)
High Alert	-7.316* (2.877)	-6.046* (2.537)
Log(midday ridership)		17.341** (5.309)
$R^2$	.14	.17

Figure 1: The dependent variable is the daily total number of crimes in D.C. This table present the estimated coefficients and their standard errors in parenthesis. The first column refers to a model where the only variable used in the High Alert dummy whereas the model in column (2) controls form the METRO ridership. \* refers to a significant coefficient at the 5% level, \*\* at the 1% level.

TABLE 4

REDUCTION IN CRIME ON HIGH-ALERT DAYS: CONCENTRATION ON THE NATIONAL MALL

	Coefficient (Robust)	Coefficient (HAC)	Coefficient (Clustered by Alert Status and Week)
High Alert × District 1	-2.621** (.044)	-2.621* (1.19)	-2.621* (1.225)
High Alert × Other Districts	-.571 (.455)	-.571 (.366)	-.571 (.364)
Log(midday ridership)	2.477* (.364)	2.477** (.522)	2.477** (.527)
Constant	-11.058** (4.211)	-11.058 (5.87)	-11.058 <sup>+</sup> (5.923)

Figure 2: The dependent variable is the daily total number of crimes in D.C. District 1 refers to a dummy variable associated with crime incidents in the first police district area. This table present the estimated coefficients and their standard errors in parenthesis.\* refers to a significant coefficient at the 5% level, \*\* at the 1% level.