# Homework Set 1, ECO 311

### Due Date: check syllabus

**Instruction**: There are eight questions. Each question is worth 2 points. You need to submit the answers of **only five questions** which you choose. The maximum point you can get is 10 points. I will grade only the first five questions if you answer more than five questions. For the purpose of preparing exam, you need to understand all questions.

I will discuss the homework on the due date. Please do not ask me to go through the homework before the due date. However, you can discuss the homework with your classmates. You need to submit the homework individually though.

## Q1: Summation (Sigma Notation)

Statistics and Econometrics involve a lot of summations, and the shorthand for summation is $\Sigma$. By definition

$$\sum_{i=1}^{i=n} y_i \equiv y_1 + y_2 + \dots y_n$$

where $i$ is the index for the observation, and $n$ is the number of observations (sample size). For example, we can use the sigma notation to define the sample mean as

$$\bar{y} \equiv \frac{y_1 + y_2 + \dots y_n}{n} = \frac{\sum_{i=1}^{i=n} y_i}{n}$$

More discussion about summation is in appendix A.1 of the textbook. Now please prove

1. $\sum_{i=1}^{i=n} c y_i = c \sum_{i=1}^{i=n} y_i$ where $c$ is a constant (1 point)
2. $\sum_{i=1}^{i=n} (y_i - \bar{y}) = 0$ (0.5 point)
3. $\sum_{i=1}^{i=n} (y_i - \bar{y})^2 = \sum_{i=1}^{i=n} (y_i - \bar{y}) y_i$ (0.5 point)

The second problem indicates that the sample mean measures the central location, in the sense that the total deviation of $y$ from the sample mean is zero. The third problem shows how to simplify the calculation of sample variance.

## Q2: Sample Mean is OLS Estimator

Consider running a strange regression and only estimating an intercept (i.e., there is no independent variable $X$ on the right hand side of the regression). The regression is

$$y = \beta_0 + u.$$

1. Please show that the estimated intercept is the sample mean

$$\hat{\beta}_0 = \bar{y}$$

2. Please show the sum of residual equals zero

## Q3: Interpretation, Prediction, Residual

Use House data to answer this question. We are interested in the relation between rprice and area. The R reports following simple regression result

```
> summary(lm(rprice~area))$coef
               Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 16225.83153 4381.25679  3.703465 2.505499e-04
area           32.03807    1.97527 16.219593 1.448699e-43
```

1. How to interpret the estimated slope $\hat{\beta}_1 = 32.03807$? Does area have significant effect on rprice? Why?

2. Please find the predicted rprice for a house with area=1000. Suppose for that house the residual is positive, what does that mean?

3. Compute the residual for the most expensive house in the sample. How to interpret that residual? Do you want to buy that house? (Hint: list the most expensive house using `data[rprice==max(rprice),]`)

## Q4: T test, P value and Confidence Interval

Please use House data and run the simple regression using R function `lm`

$$\texttt{rprice} = \beta_0 + \beta_1 \texttt{area} + u$$

Consider a null hypothesis

$$H_0 : \beta_1 = 31$$

versus the one-tailed alternative hypothesis

$$H_1 : \beta_1 > 31$$

Please (a) interpret the null and alternative hypotheses; (b) find the t statistic, and use function `pnorm` to find the p-value and draw a conclusion; (c) please use function `qnorm` to find the appropriate critical value and construct the 90% confidence interval for $\beta_1$.

### Q5: R-Squared

The overall performance of a regression can be evaluated by the R-squared, which is also called the coefficient of determination. Eco 311 emphasizes causality, so we downplay the importance of R-squared. The formula for R-squared is

$$R^2 = 1 - \frac{SSR}{SST}$$

where $SSR = \sum_i \hat{u}_i^2$ is the sum of squared residual, and $SST = \sum_i (y_i - \bar{y})^2$ is the total sum of squares, see equation 2.38 in the textbook. Basically, R-squared can be interpreted as the fraction of the sample variation in $y$ that can be explained by $x$ (or explained by the regression). Everything else equal, we want to use a regression that produces greater R-squared. Please use House data for this problem. Consider two simple regressions:

$$\texttt{rprice} = \beta_0 + \beta_1 \texttt{baths} + u \quad \texttt{(regression A)}$$

$$\texttt{rprice} = \beta_0 + \beta_1 \texttt{area} + u \quad \texttt{(regression B)}$$

Notice that the two regressions have the same dependent variable; otherwise the R-squared is not comparable. Please (a) find the residual for regression A, compute SSR and SST, and then find the R-squared (Hint: SST can be computed as
`sst = (length(rprice)-1)*var(rprice)`); (b) interpret the R-squared in each regression; (c) explain which model has better fit.

## Q6: Standard error and T value

Most often, we run a regression in the hope that some key regressor is statistically significant. For example, we may hope to find that cutting class size has significant effect on learning outcome. In terms of statistics, that means we hope to obtain a big t value for the slope coefficient (so that we can reject the default null hypothesis that the independent variable has no effect $H_0 : \beta_1 = 0$). Notice that a big t value requires small standard error of $\hat{\beta}_1$. Please use House data for this problem. Consider running two simple regressions

$$\texttt{rprice}_i = \beta_0 + \beta_1 \texttt{baths}_i + u_i, \quad (i = 1, 2, \ldots 100)$$

$$\texttt{rprice}_i = \beta_0 + \beta_1 \texttt{baths}_i + u_i, \quad (i = 1, 2, \ldots 321)$$

So the first regression uses only the first 100 houses (observations), while the second regression uses all 321 houses in the sample. Please explain in detail why the second regression has smaller standard error and bigger t value for $\hat{\beta}_1$ than the first regression. Be specific! (Hint: see Theorem 2.2 on textbook; the first one hundred observations of rpice is `rprice[1:100]`). Finally, a student suggests this R function

```
lm(rprice[baths==2]~ baths[baths==2])
```

In other words, this student wants to use only the house with two bathrooms to run the regression. Please explain why you will see strange result after trying that command.

## Q7: Omitted Variable Bias

Use the House data. Please run a simple regression $\texttt{rprice} = \beta_0 + \beta_1 \texttt{baths} + u$. Suppose the error term $u$ includes only one omitted variable `area`. Please explain in detail whether $\hat{\beta}_1$ in the simple regression underestimates or overestimates the marginal effect of `baths` on `rprice`. You need to explicitly specify the sign of $\beta_2$ and $cov(x, w)$ in the OVB formula in the handout. How to empirically check the sign of correlation between `baths` and `area`?

## Q8: Frisch-Waugh Theorem

This exercise intends to show how to apply the Frisch-Waugh Theorem.

1. Use House data and report the regression $\texttt{rprice} = \beta_0 + \beta_1 \texttt{area} + \beta_2 \texttt{age} + u$. Interpret $\hat{\beta}_1$.

2. Suppose the key regressor is `area`. Please show the R code and result that implement the two-step procedure of the Frisch-Waugh Theorem to obtain $\hat{\beta}_1$