

**Homework 4 Solutions:**

**Chapter 5: #3, 5, 8ab, 18a, 30, 31, and show the hat matrix  $\mathbf{H}$  is idempotent**

**Handout: Problem 6.1, and use the data from Problem 6.9 to answer questions posted on the website.**

**Assigned Mon, Oct 19:**

5.3 These can be written as follows:

$$(1) \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{bmatrix} - \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \hat{Y}_3 \\ \hat{Y}_4 \end{bmatrix} = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{bmatrix} \quad \text{and} \quad (2) \begin{bmatrix} X_1 & X_2 & X_3 & X_4 \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{bmatrix} = [0]$$

5.5 As noted in the hint, see page 185 for formulas for these in more familiar notation.

$$(1) \mathbf{Y}'\mathbf{Y} = \sum Y_i^2 = 1,259 \quad (2) \mathbf{X}'\mathbf{X} = \begin{bmatrix} 6 & 17 \\ 17 & 55 \end{bmatrix} \quad (3) \mathbf{X}'\mathbf{Y} = \begin{bmatrix} 81 \\ 261 \end{bmatrix}$$

- 5.8 a. The columns are linearly dependent (not linearly independent). Note that  $\mathbf{C}_1 = \mathbf{C}_2 + \mathbf{C}_3$   
 b. The rank is 2, the number of columns that are linearly independent.

5.18 a. 
$$\begin{bmatrix} W_1 \\ W_2 \end{bmatrix} = \begin{bmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/2 & 1/2 & -1/2 & -1/2 \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{bmatrix}$$

5.30 There are a few ways you can do this problem. One is to recognize that  $\hat{Y}_h$  is a scalar (number), so if you transpose it, you get the same thing. Therefore, by formula 5.32 (p. 193),  $\hat{Y}_h = (\mathbf{X}_h'\mathbf{b})' = \mathbf{b}'\mathbf{X}_h$ . You can also write it out directly:  $\mathbf{b}'\mathbf{X}_h = [b_0 \ b_1] \begin{bmatrix} 1 \\ X_h \end{bmatrix} = b_0 + b_1X_h = \hat{Y}_h$ .

5.31

$$\begin{aligned} \sigma^2\{\hat{\mathbf{Y}}\} &= \mathbf{H}\sigma^2\{\mathbf{Y}\}\mathbf{H}' && \text{[by (5.46)]} \\ &= \mathbf{H}\sigma^2\mathbf{I}\mathbf{H} && \text{(since } \mathbf{H} \text{ is symmetric)} \\ &= \sigma^2\mathbf{H} && \text{(since } \mathbf{H}\mathbf{H} = \mathbf{H}) \end{aligned}$$

**Use matrix algebra to show that the hat matrix  $\mathbf{H}$  is idempotent.**

The hat matrix is  $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ , so  $\mathbf{H}\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{X}\mathbf{I}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{H}$ .

Assigned Wed, October 21

6.1

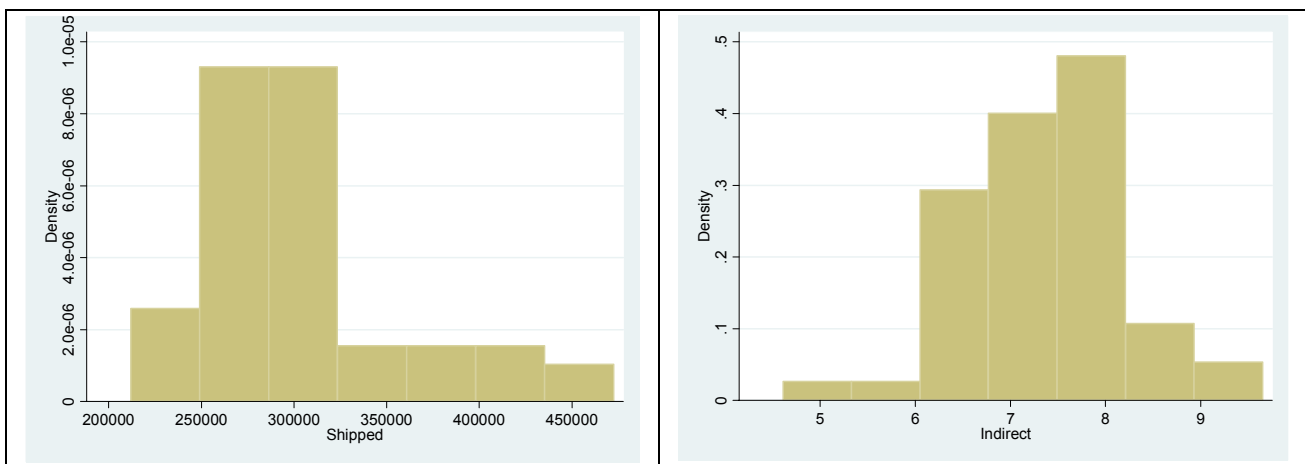
$$a. \quad \mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{11}X_{12} \\ 1 & X_{21} & X_{21}X_{22} \\ 1 & X_{31} & X_{31}X_{32} \\ 1 & X_{41} & X_{41}X_{42} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

$$b. \quad \mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} \\ 1 & X_{21} & X_{22} \\ 1 & X_{31} & X_{32} \\ 1 & X_{41} & X_{42} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

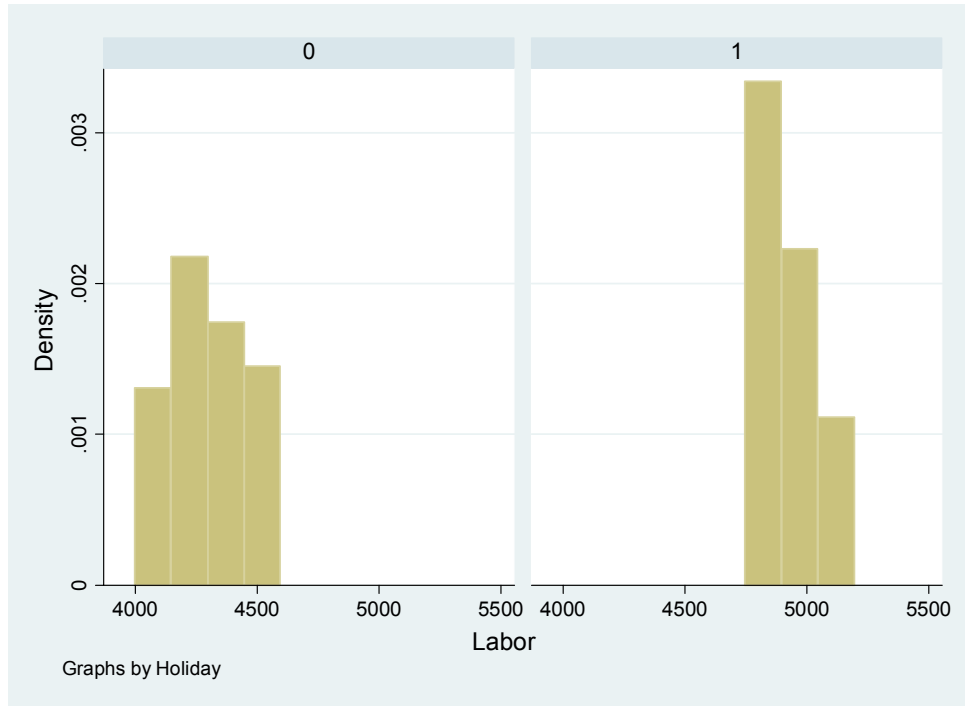
**1. Make appropriate plots before running a regression analysis. Comment on whether you see any problems.**

You should check histograms for the variables  $X_1$  and  $X_2$ , and separately for Labor Hours for holiday weeks and non-holiday weeks. You should also look at a scatter plot matrix. See if there are any unusual observations in any of these plots, and if so, investigate. The plots are below. The only interesting feature is that Labor Hours are higher for holiday weeks than for non-holiday weeks, but that does not indicate a problem. It's hard to tell if the relationships between  $Y$  and  $X_1$  and  $X_2$  are linear, but no other pattern shows up. NOTE: You could do different graphs than the ones shown here, as long as you can use them to investigate the data for initial problems.

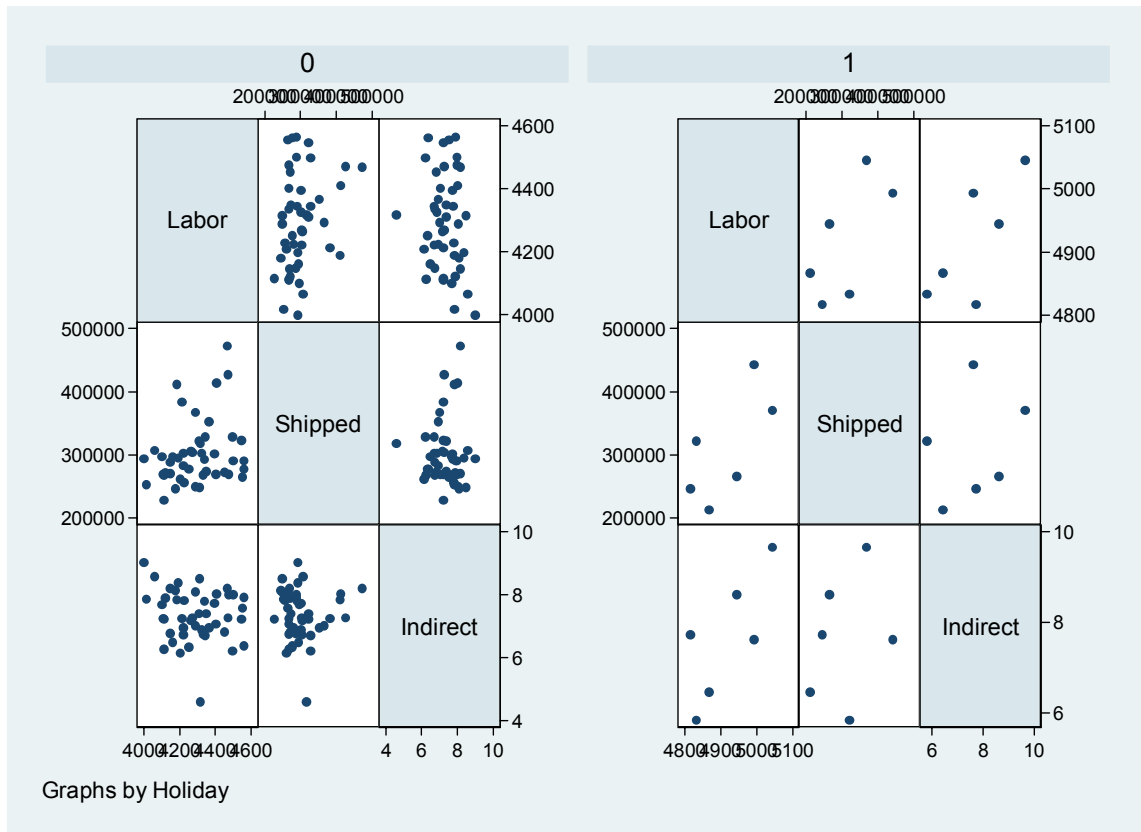
Histograms of  $X_1$  and  $X_2$ :



Histogram of Y for the non-holiday and holiday weeks:



Scatter plot matrix, separately for non-holiday and holiday weeks:



**2. Even if you thought there were problems from step 1 (this isn't a hint as to whether or not you should have found any), use all of the data to perform a regression analysis. Write out the regression equation you found.**

```
. regress Labor Shipped Indirect Holiday
```

Source	SS	df	MS			
Model	2176606.18	3	725535.393	Number of obs =	52	
Residual	985529.745	48	20531.8697	F( 3, 48) =	35.34	
Total	3162135.92	51	62002.6652	Prob > F =	0.0000	
				R-squared =	0.6883	
				Adj R-squared =	0.6689	
				Root MSE =	143.29	

Labor	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
Shipped	.0007871	.0003646	2.16	0.036	.0000541	.0015201
Indirect	-13.16602	23.09173	-0.57	0.571	-59.59506	33.26302
Holiday	623.5545	62.64095	9.95	0.000	497.6064	749.5025
_cons	4149.887	195.5654	21.22	0.000	3756.677	4543.098

The regression equation is  $\hat{Y}_i = 4149.89 + 0.00079 X_{i1} - 13.17 X_{i2} + 623.55 X_{i3}$

Note that  $X_{i3}$  is 0 for non-holiday weeks and 1 for holiday weeks, so you could write the equation separately for those two cases, by removing the last term and writing the intercept as 4149.89 for non-holiday weeks and as  $4149.89 + 623.55$  for holiday weeks.

**3. Explain the results, including:**

- Interpretation of each of the coefficients in the context of this example
- A test for each coefficient, with an explanation in words of the test and the results
- $R^2$  and what it means
- Root MSE and what it means
- An overall test for whether regression is useful in this example.

**a. Interpretation of each of the coefficients in the context of this example:**

The coefficient .00079 for “Shipped” indicates that for a *fixed* combination of “Indirect” and type of week (holiday or not), the predicted labor hours go up .00079 on average for each increase of 1 for the variable “shipped.”

Similarly, the coefficient of -13.17 for “Indirect” indicates that predicted labor hours go *down* 13.17 hours when the indirect cost percentage goes *up* by 1, for a *fixed* combination of “shipped” and type of week.

The coefficient of 623.55 for “holiday” indicates that on average, when the other two variables are fixed, labor hours are predicted to be about 623.55 hours higher for holiday weeks than for non-holiday weeks.

**b. A test for each coefficient, with an explanation in words of the test and the results:**

Hypotheses:	t*	p-value	conclusion
$H_0: \beta_0 = 0, H_a: \beta_0 \neq 0$	21.22	0.000	Reject $H_0$ , conclude $\beta_0 \neq 0$
$H_0: \beta_1 = 0, H_a: \beta_1 \neq 0$	2.16	0.036	Reject $H_0$ , conclude $\beta_1 \neq 0$
$H_0: \beta_2 = 0, H_a: \beta_2 \neq 0$	-0.57	0.571	Do not reject $H_0$ , conclude $\beta_2$ could be 0
$H_0: \beta_3 = 0, H_a: \beta_3 \neq 0$	9.95	0.000	Reject $H_0$ , conclude $\beta_3 \neq 0$

In words, we conclude that even with indirect percentage and type of week in the model, the “shipped” variable is useful in predicting labor hours. We also conclude there is a significant difference between holiday and non-holiday weeks, after accounting for the “shipped” and “indirect” variables. However, the non-significant test for the “indirect” variable indicates that including the indirect percentage does not significantly help in predicting or estimating labor hours once “shipped” and holiday (or not) have been included in the model. The test for the constant term doesn’t have a useful interpretation in this example.

**c.  $R^2$  and what it means:**

The value of  $R^2$  is .6883, or 68.83%. This tells us that almost 69% of the variability in weekly labor hours can be explained by the variables used in this regression model, compared with using only the mean weekly labor hours as an estimate. Technically (not required) almost 69% of SSTO is in SSR and the remaining 31% is in SSE.

**d. Root MSE and what it means**

Root MSE = 143.29 hours =  $\sqrt{MSE} = s$ , and is an unbiased estimate of the population standard deviation  $\sigma$ . This tells us that for the (hypothetical) population of weeks in any given combination of “shipped”, “indirect” and “holiday” the standard deviation for the number of labor hours is about 143.29 hours. This represents the natural variability in the populations, not measurement error or uncertainty in our estimates.

**e. An overall test for whether regression is useful in this example**

The appropriate test is given in the output as  $F(3, 48) = 35.34$ . This is the test statistic  $F^* = MSR/MSE$  for testing  $H_0: \beta_1 = \beta_2 = \beta_3 = 0$  versus  $H_a$ : they are not all 0. The p-value for the test is 0.0000, so we reject the null hypothesis and conclude that at least one of the X variables is useful in predicting labor hours. In other words, at least one of those 3  $\beta$ 's is not 0.

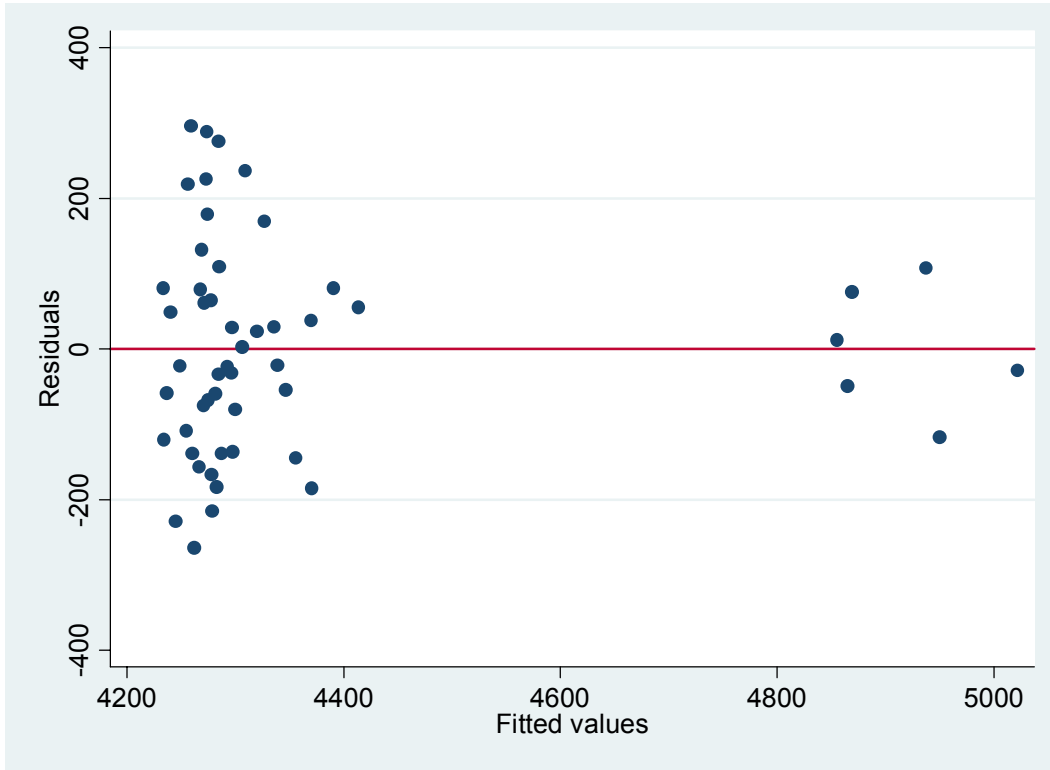
**4. Provide appropriate plots of residuals and comment on them.**

We could plot the residuals or semi-studentized residuals versus each X variable, but we will just plot them against the fitted values. We should also look at a normal probability plot. These two plots are on the next page.

The plot of residuals versus fitted values illustrates that there are two separated clumps of predicted values. This is because labor hours are predicted to much higher during holiday weeks, as indicated by the coefficient of 623.55 for “holiday.” This amount is added to the predicted labor hours when it’s a holiday week. It’s okay that there are two clumps, but the issue of concern is the variance. Clearly, the points are less spread out for the upper values (the holiday weeks). It’s hard to tell with so few points, but it looks like the variance is decreasing in general as predicted labor hours increase.

The normal probability plot looks good. It isn’t quite linear, but it’s very close, indicating that there are no major outliers in the residuals, and no extreme skewness.

**Plot of residuals versus fitted values:**



**Normal probability plot of the residuals:**

