**Assignment #3— Due Friday, 12 Nov.**

*This homework covers Chapter 5 (*Problem* 1-3) and Chapter 6 (*Problem* 3-4). Submit your homework on Canvas or send it to our TA, Mr. LYU Zhongyuan (zlyuab@connect.ust.hk).

*No late homework will be accepted for credit.

*Append the R codes you used to your submission. *If the problem does not need R or is not explicitly stated to complete in R, then you should just do it by hand with a calculator.*

*In case of rounding error, keep 3 figures after the decimal point.

**Problem 1** The following table shows a regression output obtained from fitting the model $Y = \beta_0 + \beta_1 X + \varepsilon$ to a set of data consisting of $n$ workers in a given company, where $Y$ is the weekly wages in \$100 and $X$ is the gender. The Gender variable is coded as 1 for Males and 0 for Females.

**Regression Output from the Regression of the Weekly Wages, $Y$, on $X$ (Gender: 1 = Male, 0 = Female)**

ANOVA Table

| Source | Sum of Squares | df | Mean Square | $F$-Test |
|---|---|---|---|---|
| Regression | 98.8313 | 1 | 98.8313 | 14 |
| Residual | 338.449 | 48 | 7.05101 | |

Coefficients Table

| Variable | Coefficient | s.e. | $t$-Test | $p$-value |
|---|---|---|---|---|
| Constant | 15.58 | 0.54 | 28.8 | < 0.0001 |
| $X$ | –2.81 | 0.75 | –3.74 | 0.0005 |

(a) How many workers are there in this data set ?

(b) Compute the sample variance of $Y$?

(c) Given that $\bar{X} = 0.52$, what is $\bar{Y}$?

(d) Given that $\bar{X} = 0.52$, how many women are there in this data set?

(e) What percentage of the variability in $Y$ can be accounted for by $X$?

(f) Compute the correlation coefficient between $Y$ and $X$?

(g) What is your interpretation of the estimated coefficient $\hat{\beta}_1$?

(h) What is the estimated weekly wages of a man chosen at random from the workers in the company?

(i) What is the estimated weekly wages of a woman chosen at random from the workers in the company?

(j) Construct a 95% confidence interval for $\beta_1$.

(k) Test the hypothesis that the average weekly wages of men is equal to that of the women. [Specify (a) the null and alternative hypotheses, (b) the test statistic, (c) the critical value, and (d) your conclusion.]

**Problem 2** (Use R) Presidential Election Data (1916-1996): The data in the following table, stored in the file *Presidential_Eelection_Data.txt* , were kindly provided by Professor Ray Fair of Yale University, who has found that the proportion of votes obtained by a presidential candidate in a U.S.A. presidential election can be predicted accurately by three macroeconomic variables, incumbency, and a variable which indicated whether the election was held during or just after a war. The variables considered are given in Table 5.20. All growth rates are annual rates in percentage points. Consider fitting the following initial model to the data:

$$V = \beta_0 + \beta_1 I + \beta_2 D + \beta_3 W + \beta_4 (G \cdot I) + \beta_5 P + \beta_6 N + \varepsilon$$

Presidential Election Data (1916–1996)

| Year | $V$ | $I$ | $D$ | $W$ | $G$ | $P$ | $N$ |
|---|---|---|---|---|---|---|---|
| 1916 | 0.5168 | 1 | 1 | 0 | 2.229 | 4.252 | 3 |
| 1920 | 0.3612 | 1 | 0 | 1 | −11.463 | 16.535 | 5 |
| 1924 | 0.4176 | −1 | −1 | 0 | −3.872 | 5.161 | 10 |
| 1928 | 0.4118 | −1 | 0 | 0 | 4.623 | 0.183 | 7 |
| 1932 | 0.5916 | −1 | −1 | 0 | −14.901 | 7.069 | 4 |
| 1936 | 0.6246 | 1 | 1 | 0 | 11.921 | 2.362 | 9 |
| 1940 | 0.5500 | 1 | 1 | 0 | 3.708 | 0.028 | 8 |
| 1944 | 0.5377 | 1 | 1 | 1 | 4.119 | 5.678 | 14 |
| 1948 | 0.5237 | 1 | 1 | 1 | 1.849 | 8.722 | 5 |
| 1952 | 0.4460 | 1 | 0 | 0 | 0.627 | 2.288 | 6 |
| 1956 | 0.4224 | −1 | −1 | 0 | −1.527 | 1.936 | 5 |
| 1960 | 0.5009 | −1 | 0 | 0 | 0.114 | 1.932 | 5 |
| 1964 | 0.6134 | 1 | 1 | 0 | 5.054 | 1.247 | 10 |
| 1968 | 0.4960 | 1 | 0 | 0 | 4.836 | 3.215 | 7 |
| 1972 | 0.3821 | −1 | −1 | 0 | 6.278 | 4.766 | 4 |
| 1976 | 0.5105 | −1 | 0 | 0 | 3.663 | 7.657 | 4 |
| 1980 | 0.4470 | 1 | 1 | 0 | −3.789 | 8.093 | 5 |
| 1984 | 0.4083 | −1 | −1 | 0 | 5.387 | 5.403 | 7 |
| 1988 | 0.4610 | −1 | 0 | 0 | 2.068 | 3.272 | 6 |
| 1992 | 0.5345 | −1 | −1 | 0 | 2.293 | 3.692 | 1 |
| 1996 | 0.5474 | 1 | 1 | 0 | 2.918 | 2.268 | 3 |

Variables for the Presidential Election Data (1916–1996)

| Variable | Definition |
|---|---|
| YEAR | Election year |
| $V$ | Democratic share of the two-party presidential vote |
| $I$ | Indicator variable (1 if there is a Democratic incumbent at the time of the election and −1 if there is a Republican incumbent) |
| $D$ | Categorical variable (1 if a Democratic incumbent is running for election, −1 if a Republican incumbent is running for election, and 0 otherwise) |
| $W$ | Indicator variable (1 for the elections of 1920, 1944, and 1948, and 0 otherwise) |
| $G$ | Growth rate of real per capita GDP in the first three quarters of the election year |
| $P$ | Absolute value of the growth rate of the GDP deflator in the first 15 quarters of the administration |
| $N$ | Number of quarters in the first 15 quarters of the administration in which the growth rate of real per capita GDP is greater than 3.2% |

(a) Write the regression model corresponding to each of three possible values of $D$ and interpret the regression coefficient of $D$ ($\beta_2$).

(b) Do we need to keep the variable $I$ in the above model?

(c) Do we need to keep the interaction variable $(G \cdot I)$ in the above model?

(d) Examine different models to produce the model or models that might be expected to perform best in predicting future presidential elections. Include interaction terms if needed.

**Problem 3** (Use R) Refer to the *Presidential Election Data* in Problem 2, where the $D$ is a categorical variable with three categories. Now, if we replace $D$ by two indicator variables such as

$$D_1 = 1 \text{ if } D = 1 \text{ (Democratic incumbent is running) and 0 otherwise, and}$$
$$D_2 = 1 \text{ if } D = -1 \text{(Republican incumbent is running) and 0 otherwise}$$

Then an alternative to the model in Problem 2 is

$$V = \beta_0 + \beta_1 I + \alpha_1 D_1 + \alpha_2 D_2 + \beta_3 W + \beta_4 (G \cdot I) + \beta_5 P + \beta_6 N + \varepsilon$$

(a) Write the regression model corresponding to each of the three possible values of $D$ in the above model and interpret the regression coefficient of $D_1$ and $D_2$.

(b) Show the model in Problem 2 can be obtained as a special case of the model in Problem 3 by assuming that $\alpha_1 = -\alpha_2$.

(c) Do the *Presidential_Election_Data.txt* support the assumption that $\alpha_1 = -\alpha_2$?

**Problem 4** Two variables, $Y$ and $X$, are believed to be strongly nonlinearly related. A power transformation $Y^\lambda$ was thought to make the relationship between $Y^\lambda$ and $X$ linear for some value of $\lambda$. The following table gives the value of the correlation coefficient between $Y^\lambda$ and $X$ for some values of $\lambda$.

Correlation Coefficient Between $Y^\lambda$ and $X$ for Some Values of $\lambda$

| $\lambda$ | 1 | 0.5 | 0.001 | –0.001 | –0.5 | –1 | –2 |
|---|---|---|---|---|---|---|---|
| Correlation | –0.777 | –0.852 | –0.930 | 0.930 | 0.985 | 0.999 | 0.943 |

(a) What is the correlation coefficient between $Y$ and $X$? Explain.

(b) Observing the trend in the above table, what is the best (and easy to explain for interpret) value of $\lambda$? Explain.

(c) Using your choice of $\lambda$ in (b), write the equation that related $Y$ to $X$.

**Problem 5** (Use R) Refer to the *Presidential Election Data* in Problem 2, where the response variable $V$ is the proportion of votes obtained by a presidential candidate in the United States. Since the response is in a proportion, it has a value between 0 and 1. The transformation $Y = \log[V/(1-V)]$ takes the variable $V$ with values between 0 and 1 to a variable $Y$ with values between $-\infty$ to $+\infty$. It is therefore more reasonable to expect that $Y$ satisfies the normality assumption than does $V$. Consider then fitting the model in Problem 2 but replacing $V$ by $Y$.

(a) For each of the two models ($V \sim .$ and $Y \sim .$ using the model in Problem 2), examine the appropriate residual plots discussed in Chapter 4 to determine which model satisfies the standard assumptions more than the other, the original variable $V$ or the transformed variable $Y$.

(b) What does the fitted model above imply about the form of the model relating the original variables $V$ in terms of the predictor variables? That is, find the form of the function

$$V = f(\beta_0 + \beta_1 I + \beta_2 D + \beta_3 W + \beta_4 (G \cdot I) + \beta_5 P + \beta_6 N + \varepsilon)$$

[Hint: This is a nonlinear function referred to as the logistic function, which is discussed in Chapter 9.]

**Problem 6** (Use R) Repeat Problem 5 but fitting the model in *Problem 3* (rather than Problem 2) but replacing $V$ by $Y$, and compare the result with that of Problem 5.