STA502: Math & Stat for MBA

Problem Set 1

Question 1. Consider the simple linear regression model: $y_t = \beta_1 + \beta_2 x_t + \epsilon_t, \quad t = 1, ..., T.$

1. Obtain formulae for $\hat{\beta}_1$ and $\hat{\beta}_2$, the least squares estimators, from the general result: $\hat{\beta} = (X'X)^{-1}X'y$.
2. Obtain formuale for $var(\hat{\beta}_1), var(\hat{\beta}_2)$ and $covar(\hat{\beta}_1, \hat{\beta}_2)$ from the general OLS result: $V(\hat{\beta}) = \sigma_\epsilon^2 (X'X)^{-1}$.

Solution.
Denote $X = [i \;\; x]$. Then

$$X'X = \begin{bmatrix} i'i & i'x \\ x'i & x'x \end{bmatrix} = \begin{bmatrix} T & \sum_t x_t \\ \sum_t x_t & \sum_t x_t^2 \end{bmatrix}$$

$$X'y = \begin{bmatrix} \sum_t y_t & \sum_t x_t y_t \end{bmatrix}'$$

It follows that:

$$(X'X)^{-1} = \begin{bmatrix} \sum_t x_t^2 & -\sum_t x_t \\ -\sum_t x_t & T \end{bmatrix} / \triangle$$

where $\triangle = |X'X| = T\sum_t x_t^2 - (\sum_t x_t)^2 = T\sum_t (x_t - \bar{x})^2$.

1. Carrying out the matrix multiplications gives

$$\begin{aligned} \hat{\beta}_1 &= \frac{(\sum_t x_t^2 \sum_t y_t - \sum_t x_t \sum_t x_t y_t)}{\triangle} \\ &= \frac{(\bar{y}\sum_t x_t^2 - \bar{x}\sum_t x_t y_t)}{\sum_t (x_t - \bar{x})^2} \\ &= \frac{\bar{y}\sum_t (x_t - \bar{x})^2 + T\bar{x}^2\bar{y} - \bar{x}\sum_t x_t y_t}{\sum_t (x_t - \bar{x})^2} \\ &= \bar{y} - \frac{\bar{x}\sum_t (x_t - \bar{x}) y_t}{\sum_t (x_t - \bar{x})^2} \\ \hat{\beta}_2 &= \frac{\sum_t (x_t - \bar{x}) y_t}{\sum_t (x_t - \bar{x})^2} \end{aligned}$$

We have used $\bar{w} = \sum_t w_t / T$ and $\sum_t (w_t - \bar{w})^2 = \sum_t w_t^2 - (\sum_t w_t)^2 / T$ for these results.

2.

$$\begin{aligned} V(\hat{\beta}) &= \sigma_\epsilon^2 \begin{pmatrix} \sum_t x_t^2 & -\sum_t x_t \\ -\sum_t x_t & T \end{pmatrix} / \triangle \\ &= \frac{\sigma_\epsilon^2}{\sum_t (x_t - \bar{x})^2} \begin{pmatrix} \sum_t x_t^2 / T & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} \end{aligned}$$

Question 2. In the article "Returns to Scale in Electricity Supply", the following cost function was estimated by Ordinary Least Squares (OLS) using data on different firms:

$$\begin{aligned} \log(\text{Cost}) = &- \underset{(1.77)}{3.53} + \underset{(0.0175)}{0.720}\log(\text{Output}) + \underset{(0.291)}{0.436}\log(\text{Price of Labor}) \\ &+ \underset{(0.339)}{0.220}\log(\text{Cost of Capital}) + \underset{(0.100)}{0.427}\log(\text{Price of Fuel}) + \hat{\epsilon} \end{aligned}$$

where $R^2 = 0.93$. Standard errors in parantheses, with 145 Number of observations.

Estimated Covariance matrix of the estimated coefficients

Constant	3.148				
Log(Output)	-0.437E-02	0.305E-03			
Log(Price of Labor)	-0.141	-0.455E-03	0.0847		
Log(Cost of Capital)	0.591	0.323E-03	0.0237	0.115	
Log(Price of Fuel)	0.715E-02	0.315E-03	-0.0109	-0.663E-02	0.0101

Note that when a number has an "E" postfix, this means it should be multiplied by 10 to the power of the postfix, for example:

$$0.43\text{E}-02 = 0.43 \times 10^{-2} = 0.0043 \text{ and } 0.43\text{E}+02 = 0.43 \times 10^{2} = 43$$

1. Test the hypothesis that the electricity industry displays constant returns to scale against the alternative that it displays increasing returns at the 5% significance level.
2. We would expect the cost function to be homogeneous of degree one in prices. What does this imply for the coefficients of the cost function? Test the hypothesis that the cost function is homogeneous of degree one (i.e. produce just as much input invested) against the alternative that it is not at the 5% significance level.

Solution.

1. To test at 5% significance level $H_0 : \beta_2 = 1$ vs. $H_1 : \beta_2 < 1$, we define the one-sided test procedure:
   $\tau = (\hat{\beta}_2 - 1)/SE(\hat{\beta}_2) < t^*_{5\%}(140) = -1.655$
   Do not reject H_0 if $\tau \geq t^*$.
   In this case, $\tau = (0.720 - 1)/0.0175 = -16 < -1.655$, thus we reject H_0.
2. To test the price homogeneity of degree one $H_0 : \beta_3 + \beta_4 + \beta_5 = 1$ vs. $H_1 : \beta_3 + \beta_4 + \beta_5 \neq 1$ at 5% significance level, we follow the decision rule:
   Do not reject H_0 if $|\tau| = |(\hat{\beta}_3 + \hat{\beta}_4 + \hat{\beta}_5 - 1)/SE(\hat{\beta}_3 + \hat{\beta}_4 + \hat{\beta}_5)| < t^*_{2.5\%}(140) = 1.977$;
   Reject H_0 otherwise.
   Where $SE(\hat{\beta}_3 + \hat{\beta}_4 + \hat{\beta}_5) = \sqrt{\hat{V}_3 + \hat{V}_4 + \hat{V}_5 + 2(\hat{V}_{34} + \hat{V}_{35} + \hat{V}_{45})}$. In this case, $\tau = (0.436 + 0.220 + 0.427 - 1)/\sqrt{0.0847 + 0.115 + 0.0101 + 2(0.0237 - 0.0109 - 0.00663)} = 0.1761$, thus, H_0 is not rejected.

Question 3. In the two-variable model

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, \quad i = 1, ..., 11$$

Suppose that $x_1^{'}x_1 = 2, x_2^{'}x_2 = 2, x_1^{'}x_2 = 1, x_1^{'}y = 2, x_2^{'}y = 1, y^{'}y = 7/3$ where x_1, x_2 and y are the column vectors with typical elements x_{1i}, x_{2i} and y_i respectively.
Assume $\epsilon_i \sim$ i.i.d.$N(0, \sigma^2_\epsilon)$. Suppose you would like to make out-of-sample predictions about the left-hand-side (dependent) variable for two hypothetical observations with the following characteristics:

Obs.	x_1	x_2
12	5	-2
13	3	-7

1. Construct 80% prediction interval for the dependent variable y_{12} and y_{13}.
2. Construct 80% prediction interval for the expected value of y_{12} and y_{13}.
3. Do the answers above differ? Why?

Solution.
In the matrix notation, the least squares estimator of β is $\hat{\beta} = (X'X)^{-1}X'y$.
Using the formula, we get

$$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}^{-1} \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}$$

Insert the obersavions in the formula :

$$\hat{y_{12}} = X'_{12}\hat{\beta} = \begin{bmatrix} 5 \\ -2 \end{bmatrix}' \begin{bmatrix} 1 \\ 0 \end{bmatrix} = 5,$$

$$\hat{y_{13}} = X'_{13}\hat{\beta} = \begin{bmatrix} 3 \\ -7 \end{bmatrix}' \begin{bmatrix} 1 \\ 0 \end{bmatrix} = 3,$$

$$\hat{\sigma^2} = \frac{RSS}{n-p} = \frac{Y'Y - \hat{\beta}'X'y}{9} = \frac{1}{27}.$$

1. A 80% Prediction interval for the future observation is

$$\hat{y_f} - t_{(0.1,9)}\sqrt{V(y_f)} \leq y_f \leq \hat{y_f} + t_{(0.1,9)}\sqrt{V(y_f)}.$$

Since $y_f = X_f'\hat{\beta} + \hat{\epsilon_f}, V(y_f) = X_f' V(\hat{\beta}) X_f + V(\hat{\epsilon})$. Insert the $V(\hat{\beta}) = (X'X)^{-1}\hat{\sigma^2}$, we obtain

$$\hat{y_{12}} - t_{(0.1,9)}\sqrt{\hat{\sigma^2}(X_{12}'(X'X)^{-1}X_{12} + 1)} \leq y_{12} \leq \hat{y_{12}} + t_{(0.1,9)}\sqrt{\hat{\sigma^2}(X_{12}'(X'X)^{-1}X_{12} + 1)},$$
$$\hat{y_{13}} - t_{(0.1,9)}\sqrt{\hat{\sigma^2}(X_{13}'(X'X)^{-1}X_{13} + 1)} \leq y_{13} \leq \hat{y_{13}} + t_{(0.1,9)}\sqrt{\hat{\sigma^2}(X_{13}'(X'X)^{-1}X_{13} + 1)}.$$

2. A 80% Confidence interval for the fitted value is

$$\hat{y_f} - t_{(0.1,9)}\sqrt{V(\hat{y_f})} \leq E(y_f) \leq \hat{y_f} + t_{(0.1,9)}\sqrt{V(\hat{y_f})}.$$

Since the fitted value of future observation doesn't include an explicit error term ϵ,
$\hat{y_f} = X_f'\hat{\beta}, V(\hat{y_f}) = X_f' V(\hat{\beta}) X_f$. Insert the $V(\hat{\beta}) = (X'X)^{-1}\hat{\sigma^2}$, we obtain

$$\hat{y_{12}} - t_{(0.1,9)}\sqrt{\hat{\sigma^2}(X_{12}'(X'X)^{-1}X_{12})} \leq E(y_{12}) \leq \hat{y_{12}} + t_{(0.1,9)}\sqrt{\hat{\sigma^2}(X_{12}'(X'X)^{-1}X_{12})},$$
$$\hat{y_{13}} - t_{(0.1,9)}\sqrt{\hat{\sigma^2}(X_{13}'(X'X)^{-1}X_{13})} \leq E(y_{13}) \leq \hat{y_{13}} + t_{(0.1,9)}\sqrt{\hat{\sigma^2}(X_{13}'(X'X)^{-1}X_{13})},$$

3. Yes, the two intervals are different. Confidence interval for the fitted value is narrower than prediction interval for the true value, due to excluding the uncertainty(the error term ϵ). Generally, the farther the point is from the centroid of the x's, the greater the width of the two intervals.

Question 4. Consider the multiple linear regression model $y = X\beta + \epsilon$ with k explanatory variables in X. Show the following:

1. If all the observations on a particular explanatory variable are multiplied by λ, then the residuals of the regression are unchanged while the corresponding regression coefficient is multiplied by $1/\lambda$. Use this result to explain what will happen when a particular explanatory variable is measured in thousands of kgs instead of millions of kgs.

2. If a constant λ is added to all observations of a particular explanatory variable in a regression containing a constant term, then the corresponding regression coefficient is unchanged. Is any other coefficient affected? Use this result to explain that the coefficient of an explanatory variable appearing in a regression in logarithmic form, the corresponding coefficient is independent of the units in which the variable is measured.

Solution.

1. By unfolding the model, we can write the model like :

$$y = \beta_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon$$

Consider the particular explanatory variable are multiplied by λ. (e.g. λx_k) Since the model is "linear", we can reparameterize the model :

$$y = \beta_1 + \beta_2 x_2 + ... + (\frac{\beta_k}{\lambda})(\lambda x_k) + \epsilon$$

Hence, if we run the model, the estimated coefficient of x_k is $\frac{\hat{\beta_k}}{\lambda}$. If the unit of measure is reduced by 1 / 1000, $\lambda = 1000$ and estimated coefficient also reduced by 1 / 1000.

Please notice that the other explanatory variables and y, ϵ doesn't change. $R^2 = 1 - \frac{RSS}{TSS}$ should be same, neither RSS nor TSS changed. Besides, t-statistic for the significance of coefficient also be same because

$$t = \frac{\frac{1}{\lambda}\hat{\beta_k} - 0}{\frac{1}{\lambda}s.e.(\hat{\beta_k})} = \frac{\hat{\beta_k} - 0}{s.e.(\hat{\beta_k})}$$

Notice that the unit of the explanatory variable we're trying to interpret power of coefficient changed. $E(y)$ increased by $\frac{1}{\lambda}\hat{\beta_k}$ if λx increased to $\lambda(x+1)$.

2. Consider the particular explanatory variable are added by λ. (e.g. $(x_k + \lambda)$) Since the model is "linear", we can reparameterize the model :

$$y = \alpha_1 + \alpha_2 x_2 + ... + \alpha_k(x_k + \lambda) + \epsilon = (\alpha_1 + \lambda\alpha_k) + \alpha_2 x_2 + ... + \alpha_k x_k + \epsilon$$

Hence $\hat{\alpha_2} = \hat{\beta}_2, ..., \hat{\alpha_k} = \hat{\beta}_k$, but $(\hat{\alpha_1} + \lambda\hat{\alpha_k}) = \hat{\beta}_1$, hence $\hat{\alpha_1}$ has smaller values than $\hat{\beta}_1$ by $\lambda\hat{\alpha_k}$. As mentioned above, the other explanatory variables and y, ϵ doesn't change. $R^2 = 1 - \frac{RSS}{TSS}$ and t-statistic for the significance of coefficient also be same.
In logarithmic form, the corresponding coefficient is independent of the units since

$$y = \beta_1 + \beta_2 \ln x_2 + ... + \beta_k \ln \lambda x_k + \epsilon = (\beta_1 + \lambda\beta_k) + \beta_2 \ln x_2 + ... + \beta_k \ln x_k + \epsilon$$

Question 5. A researcher has collected a set of data (= 100 observations) containing a single predictor and a quantitative response. She then fits a linear regression model to the data, as well as a separate cubic regression, i.e., $Y = \beta_0 + \beta_1 X^1 + \beta_2 X^2 + \beta_3 X^3 + \epsilon$.

1. Suppose that the true relationship between X and Y is linear, i.e. $Y = \beta_0 + \beta_1 X + \epsilon$. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

2. Suppose that the true relationship between X and Y is not linear, but we don't know how far it is from linear. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

Solution.

1. If true relationship is linear, there's no significance of β_2 and β_3, hence we cannot reject the null hyphothesis that $\beta_2 = \beta_3 = 0$.
However, the training RSS always decrease or equal, since RSS = $\text{argmin}_\beta(y - \beta_0 - \beta_1 x - \beta_2 x^2 - \beta_3 x^3)^2$.
If $\hat{\beta}_2 = \hat{\beta}_3 = 0$, estimated RSS is equal to linear model, else RSS always decrease otherwise. Therefore $R^2 = 1 - \frac{RSS}{TSS}$ always equal or increase.

   Some analysts prefer to use an "adjusted R^2 statistic", defined as

$$R^2_{adj} = 1 - \frac{RSS/(n-p)}{TSS/(n-1)}$$

   Since R^2_{adj} will only increase on adding a "significant" variable to the model if the variable reduces the RSS, or R^2_{adj} will decrease because of RSS is divided by $(n-p)$, which is a penalizing term.
Consequences for the including non-significant term will bring about "over-fitting", which harms generalization of the model. The considerable other penalty terms that was discussed in lecture, are AIC and BIC.

2. If true relationship is non-linear, β_2 and β_3 will be significant, hence we can reject the null hypothesis that $\beta_2 = \beta_3 = 0$. F-statistic is

$$F = \frac{(RSS_{restricted} - RSS_{full})/2}{RSS_{full}/(100-4)} \sim F_{(2,96)}$$

   and the test statistic is larger than critical value.

We can also plot the (Standardized) residual versus the corresponding fitted values $\hat{y}$. If the model is adequate, resudials can be contained in a horizontal band around of 0. That means there's no pattern in residuals.
However, if the model cannot explain y sufficiently, residual plot shows us the specific pattern, a curved plot indicating non-linearity. This could mean that other non-linear regressor variables are needed.

How could we determine the k, the order of polynomial regression? It's important to keep the order as low as possible to prevent over-fitting. Polynomial regression models are extremely hazardous for extrapolation, which predicts out-of-sample observations that far away from the centroid. We always should remember that the sense of "Occam's razor", the simplest model is usually the best one. plotting the residual versus the corresponding fitted values $\hat{y}$ should help.

Question 6. Describe the null hypotheses to which the p-values given in Table below correspond, with response sales being regressed on predictor TV, radio, and newspaper. Explain what conclusions you can draw based on these p-values. Your explanation should be phrased in terms of sales, TV, radio, and newspaper, rather than in terms of the coefficients of the linear model.

.	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	<0.0001
TV	0.046	0.0014	32.81	<0.0001
radio	0.189	0.0086	21.89	<0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

Note: For the Advertising data, least squares coefficient estimates of the multiple linear regression of number of units sold on radio, TV, and newspaper advertising budgets.

Solution.
(1) Check the coefficient and their sign :
By assumption, coefficient sign of TV, radio, newspaper should be positive, but the coefficient sign of newspaper not. We should test the significance of newspaper.

(2) Check the standard error :
$s.e.(\hat{\beta}) = \sqrt{(X'X)^{-1}\sigma^2}$. If $(X'X)$ is large, $s.e.(\hat{\beta})$ smaller. In the table, all the variables' standard error isn't high significantly, we should make conclusion that the range of the variables is wide enough and there's no symptoms of multicollinearity.

(3) Check the p-value :
By definition, p-value = P(statistic T is more extreme than the critical value | when H_0 is true). the p-value of newspaper is higher than significance level $\alpha = 0.05$, hence we couldn't reject the null hypothesis that there's no significance of newspaper.
Test statistic T for the significance of coefficient : $T = \frac{\hat{\beta}-0}{s.e.(\hat{\beta})} \sim t_{(n-4)}$.

(4) Check the significance of whole model :
H_0 : All regressors expect intercept term is not significant / H_a : Not H_0
coefficient of TV, radio is significant, hence we could reject the null hypothesis. (F test is more accurate, but we couldn't know the RSS on the table)

Question 7. Suppose we have a data set on flights, with three predictors, X_1 = Distance (to destination in kilometers), X_2 = Holiday (1 if Yes, 0 for No), X_3 = Interaction between Distance and Holiday. The response is the flight ticket price (in US$). Suppose we use least squares to fit the model, and get $\hat{\beta}_0 = 100$, $\hat{\beta}_1 = 0.2$, $\hat{\beta}_2 = 20$, $\hat{\beta}_3 = 0.05$.

1. Which answer is correct, and why?

   - For a fixed value of Distance, on average tickets are more expensive on holidays than on usual days.
   - For a fixed value of Distance, on average tickets are more expensive on usual days than on holidays.
   - For a fixed value of Distance, on average tickets are more expensive on usual days than on holidays, provided that Distance is long enough.
   - For a fixed value of Distance, on average tickets are more expensive on holidays than on usual days, provided that Distance is long enough.

2. Predict the average holiday price of a ticket for a flight that travels 1,000km to destination.

3. True, false, or uncertain: Since the coefficient for the Distance/Holiday interaction term is pretty small, there is no evidence of an interaction effect. Justify your answer.

Solution.

1. the regression model is : $\hat{y_i} = 100 + 0.2D_i + 20H_i + 0.05(D_i * H_i)$. All the sign of estimated coefficients including interaction term is positive, we could say that "for a fixed value of distance, on "average" tickets are more expensive on holidays than on usual days".

2. Put the future value of Holiday = 1 and Distance = 1000, the fitted value $\hat{y_f} = 100 + 0.2 * 1000 + 20 * 1 + 0.05(1000 * 1) = 370$. If we want to interval estimation, the 95% prediction interval for the future observation is :

$$370 - t_{(0.025,n-4)}\sqrt{\hat{\sigma^2}(X_f'(X'X)^{-1}X_f + 1)} \leq y_f \leq 370 + t_{(0.025,n-4)}\sqrt{\hat{\sigma^2}(X_f'(X'X)^{-1}X_f + 1)}$$

3. We couldn't test the significant of interaction effect, since we don't know the "Standard Error" of interaction effect. It's difficult to judge from the size of estimated coefficient alone.
Test statistic T for the significance of coefficient : $T = \frac{0.05-0}{s.e.(\hat{\beta})} \sim t_{(n-4)}$.