Stat224: Applied Regression Analysis	Due date: October 22
Homework 3	
Section 1: MWF 1:30–2:20pm (Eckhart 133)	*Autumn 2025*

Please submit your homework pdf on Gradescope by 6pm on Wednesday, October 22.

Reminder: Your method of reasoning is important, even if your final answer is incorrect. Where a numerical answer is requested, please show how you arrived at your final answer. If you refer to a plot, make sure the plot is made clearly visible alongside your answer in your homework writeup.

Problem 1

In economic theory, the Cobb-Douglas production function models the relationship between the input variables physical capital (K) and labor (L), and the output V that can be produced by those inputs. Its most standard form is

$$V = \alpha K^{\beta_K} L^{\beta_L}, \tag{1}$$

where α is a positive constant corresponding to expected output for unit values of physical capital and labor ($L = K = 1$), and β_K, β_L are explained next. From calculus, we see that

$$\frac{\mathrm{d}V}{\mathrm{d}K} = \alpha\beta_K K^{\beta_K - 1} L^{\beta_L} = \beta_K \frac{V}{K} \Rightarrow \frac{\mathrm{d}V/V}{\mathrm{d}K/K} = \beta_K.$$

In words, a 1% increase in capital usage ($\mathrm{d}K/K = 1\%$) leads to a β_K% increase in output ($\mathrm{d}V/V = \beta_K\%$). The parameter β_K is thus called the output elasticity of capital. Likewise, $\beta_L = \frac{\mathrm{d}V/V}{\mathrm{d}L/L}$ is called the output elasticity of labor.

The measured output V is subject to noise and so will not be given exactly by the Cobb-Douglas production function (1). We may model this noise multiplicatively as follows:

$$V = \alpha K^{\beta_K} L^{\beta_L} \eta.$$

Upon taking logs, the production function becomes a multiple linear regression model:

$$\log(V) = \log(\alpha) + \beta_K \log(K) + \beta_L \log(L) + \varepsilon, \tag{2}$$

where $\varepsilon = \log(\eta)$ are assumed to be independent (additive) errors with mean 0 and constant variance σ^2.

Download the file food.txt, which gives information on capital, labor, and output (value added) of the economic sector for food and kindred products, between 1972 and 1986. We may read this file in R via the command:

```
food = read.table("food.txt", h=T)
```

(a) Fit the model

$$\log(V) = \log(\alpha) + \beta_K \log(K) + \beta_L \log(L) + \varepsilon,$$

report the least squares estimates for β_K, β_L. Test the null hypothesis $H_0 : \beta_K = 0$ at the $\alpha = 5\%$ level and interpret the conclusion in the current context.

(b) Report the residual standard error, multiple R^2 and F-statistic with degrees of freedom from the regression in part (a). Using this information, construct the analysis of variance (ANOVA) table for the overall F-test $H_0 : \beta_K = \beta_L = 0$. Show your work.

Source	df	Sum of squares	Mean Square	F-statistic	p-value
Regression					
Error					

Table 1: ANOVA table for the overall F-test

(c) For the model in part (a), test the hypothesis $H_0 : \beta_K = \beta_L$ using an F-test that compares nested models. Please specify the full model and the reduced model, compute their RSS values, and show how the F-statistic is computed from these RSS values. Compute a p-value and state your conclusion in the current context.

(d) The Cobb-Douglas function (1) suggests doubling capital K and labor L will increase the output V by a factor of $2^{\beta_K+\beta_L}$. So, if $\beta_K + \beta_L = 1$, then doubling the inputs will double the output. In this case, we say the production has *constant returns to scale*. Show that under $H_0 : \beta_K + \beta_L = 1$, the model (2) is equivalent to

$$Y = \log(\alpha) + \beta_K X + \varepsilon, \tag{3}$$

where $Y = \log(V) - \log(L)$ is the new response variable, $X = \log(K) - \log(L)$ is the new predictor variable, and β_K has the same meaning as it does in the original model.

(e) Test whether the food sector has constant returns to scale, i.e. $H_0 : \beta_K + \beta_L = 1$, using an F-statistic computed from the RSS in the full model and the RSS obtained by fitting the derived model (3). You may use the `anova()` function for this part.

(f) Sometimes the model

$$V_t = \alpha \rho^t K_t^{\beta_K} L_t^{\beta_L} \eta_t \tag{4}$$

is used to account for the different values of capital, labor, and output at times t (year), where ρ is a parameter that accounts for *technological development*. Taking logs on both sides yields the multiple regression model

$$\log(V_t) = \log(\alpha) + t \log \rho + \beta_K \log K_t + \beta_L \log L_t + \varepsilon_t.$$

$\rho > 1$, or equivalently $\log \rho > 0$, is interpreted to mean there was technological progress over the time period over which measurements were taken. Test the null hypothesis that there was no technological progress from 1972 to 1986, i.e. $H_0 : \log \rho = 0$, versus the alternative that there was, i.e. $H_1 : \log \rho > 0$. Report the estimated coefficient, its standard error, and make a conclusion based on the one-sided p-value.

Problem 2

The price of a car varies depending on many factors, but in this problem we focus on two: the engine horsepower and the country in which the car was made. The predictor variable `horsepower` and response variable `price` are quantitative. The `country` variable has 4 categories: USA, Japan, Germany, and Other. Inclusion of this categorical variable as a predictor in a linear regression model requires creating 3 indicator variables, one for each of USA, Japan and Germany. Some regression output for fitting a linear regression of `price` (in thousands of U.S. dollars) onto `country` and `horsepower` is shown below.

Variable	Coefficient	Standard error	t-statistic	p-value
Constant (Intercept)	−4.117	1.582	−2.6	0.0109
Horsepower	0.174	0.011	16.6	0.0001
USA	−3.162	1.351	−2.34	0.0216
Japan	−3.818	1.357	−2.81	0.0061
Germany	0.311	1.871	−0.166	0.8682

Table 2: Regression output for car price data.

(a) Using the regression output, estimate the price of a German car with a 100 horsepower engine.

(b) Estimate the price of a car with a 100 horsepower engine from “Other” countries (not USA, Japan, or German).

(c) Interpret the coefficient -3.818 for Japan.

(d) Describe the effect of `horsepower` on `price` of car.

Problem 3

In this problem, we will justify the claim from lecture that

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{S_{yy}} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{\text{SSreg}} + \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\text{RSS}}, \tag{5}$$

and (optionally) the interpretation of the multiple R^2 value as the squared correlation between the response and the fitted values:

$$R^2 = \text{cor}^2(y, \hat{y}) = \frac{\left(\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{y})\right)^2}{S_{yy} \cdot \text{SSreg}}.$$

(a) Recall the definition of the OLS estimator $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$ for the multiple regression model $y = X\beta + \varepsilon$,

$$\hat{\beta} = \underset{b=(b_0,b_1,\ldots,b_p)}{\text{argmin}} \sum_{i=1}^{n}(y_i - b_0 - b_1 x_{i1} - \cdots - b_p x_{ip})^2, \tag{6}$$

where $X = (x_{ij})$ for $i = 1, \ldots, n$ and $j = 1, \ldots, p$, with $x_{i0} \equiv 1$ for any i. Recall the residuals are defined $e_i = y_i - \hat{y}_i$, or in vector form $e = y - X\hat{\beta}$. Show that for each $j = 1, \ldots, p$,

$$\sum_{i=1}^{n} e_i x_{ij} = 0, \quad \text{and} \quad \sum_{i=1}^{n} e_i = 0.$$

(Hint: Differentiate the RSS in (6) individually with respect to each of $b_0, \ldots, b_p$, and set the partial derivatives equal to zero).

(b) Using your conclusion in the previous part, show that

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0, \tag{7}$$

i.e. the residuals are uncorrelated with the fitted values.

(c) Using the conclusion (7), show that

$$S_{yy} = \text{SSreg} + \text{RSS},$$

where each of these terms is defined at the start of this problem in equation (5). (Hint: Add and subtract $\hat{y}_i$ in each term of the sum for S_{yy}, and open the square while keeping $y_i - \hat{y}_i$ together.)

(d) **Optional (not graded).** Show that $R^2 = \frac{\text{SSreg}}{S_{yy}}$, may equivalently be expressed as

$$R^2 = \text{cor}^2(y, \hat{y}) = \frac{\left(\sum_{i=1}^{n} (y_i - \bar{y})(\hat{y}_i - \bar{y}) \right)^2}{S_{yy} \cdot \text{SSreg}}.$$

(Hint: Add and subtract $\hat{y}_i$ in the first term in the product of the numerator in the right hand side expression above, distribute terms appropriately, and apply (7).)