Problem Set: Linear Regression

Problem 1. *The following equation relates housing price (price) to the distance from a recently built garbage incinerator (dist):*

$$\widehat{ln(price)} = 9.40 + 0.312\, ln(dist) \quad n = 30, R^2 = 0.162$$

1. *Interpret the coefficient on ln(dist). Is the sign of this estimate what you expect it to be?*
2. *Do you think simple regression provides an unbiased estimator of the ceteris paribus elasticity of price with respect to dist?*
3. *What other factors about a house affect its price? Might these be correlated with distance from the incinerator?*

Problem 2. *Consider*

$$sav = \beta_0 + \beta_1\, inc + u, \qquad u = \sqrt{inc} \cdot e,$$

with $E[e] = 0$, $Var(e) = \sigma_e^2$, *and* $e \perp inc$.

1. *Verify the zero conditional mean:* $E[u \mid inc] = 0$.
2. *Show* $Var(u \mid inc) = inc \cdot \sigma_e^2$. *Conclude the model is heteroskedastic and* $Var(sav \mid inc)$ *rises with income.*
3. *Give an economic rationale for why savings variability increases with income.*

Problem 3. *Let $\widehat{\beta}_0, \widehat{\beta}_1$ be the OLS intercept and slope from the regression of y_i on x_i using n observations.*

1. *Let c_1 and c_2 be constants with $c_2 \neq 0$. Let $\tilde{\beta}_0, \tilde{\beta}_1$ be the intercept and slope from the regression of $c_1 y_i$ on $c_2 x_i$. Show that*

$$\tilde{\beta}_1 = \frac{c_1}{c_2}\,\widehat{\beta}_1 \quad \text{and} \quad \tilde{\beta}_0 = c_1\,\widehat{\beta}_0.$$

2. *Now let $\tilde{\beta}_0, \tilde{\beta}_1$ be from the regression of $(c_1 + y_i)$ on $(c_2 + x_i)$ (no restriction on c_1, c_2). Show that*

$$\tilde{\beta}_1 = \widehat{\beta}_1 \quad \text{and} \quad \tilde{\beta}_0 = \widehat{\beta}_0 + c_1 - c_2\,\widehat{\beta}_1.$$

3. *Let $\widehat{\beta}_0, \widehat{\beta}_1$ be the OLS estimates from the regression $\log(y_i)$ on x_i (assume $y_i > 0$). For $c_1 > 0$, let $\tilde{\beta}_0, \tilde{\beta}_1$ be from the regression of $\log(c_1 y_i)$ on x_i. Show that*

$$\tilde{\beta}_1 = \widehat{\beta}_1 \quad \text{and} \quad \tilde{\beta}_0 = \log(c_1) + \widehat{\beta}_0.$$

4. *Assume $x_i > 0$ for all i. Let $\tilde{\beta}_0, \tilde{\beta}_1$ be from the regression of y_i on $\log(c_2 x_i)$. How do these compare with the intercept and slope from the regression of y_i on $\log x_i$?*

Problem 4. *The dataset CEOSAL2 (click to download) contains the information about CEOs. Let salary be total annual compensation (in thousands of dollars) and ceoten be prior tenure as CEO (years). Data description is here.*

1. *Compute the sample means of salary and ceoten.*

2. *How many CEOs have ceoten $= 0$? What is the maximum tenure observed?*

3. *Estimate the log–linear model*

$$\log(salary_i) = \beta_0 + \beta_1\, ceoten_i + u_i.$$

Report the results in the usual format ($\hat{\beta}_0, \hat{\beta}_1$, standard errors, R^2, and n). Interpret $\hat{\beta}_1$ as the approximate percent change in salary from one additional year as CEO.

Problem 5. *Consider*

$$sleep_i = \beta_0 + \beta_1\, totwrk_i + \beta_2\, educ_i + \beta_3\, age_i + u_i,$$

where sleep and totwrk (total work) are minutes/week and educ, age are years.

1. *If adults trade off sleep for work, what sign do you expect for* β_1*?*
2. *What signs do you expect for* β_2 *and* β_3*? Briefly justify.*
3. *Using data, the estimated equation is*

   $$\widehat{sleep} = 3638.25\ -\ 0.148\, totwrk\ -\ 11.13\, educ\ +\ 2.20\, age, \quad n = 706,\ R^2 = 0.113.$$

   If totwrk rises by 5 hours/week ($= 300$ *minutes), by how many minutes is sleep predicted to fall? Is this a large tradeoff?*
4. *Interpret the sign and magnitude of the coefficient on educ (units: minutes/week per additional year of schooling).*
5. *Do totwrk, educ, and age explain much of the variation in sleep? What other factors might matter, and could they be correlated with totwrk?*

Problem 6. *Using the* ***CEOSAL2*** *data on* $n = 177$ *CEOs:*

1. *Estimate the constant–elasticity model*

   $$\log(salary_i) = \beta_0 + \beta_1 \log(sales_i) + \beta_2 \log(mktval_i) + u_i.$$

   Report the estimates in equation form.
2. *Augment (i) with profits:*

   $$\log(salary_i) = \beta_0 + \beta_1 \log(sales_i) + \beta_2 \log(mktval_i) + \beta_3 profits_i + u_i.$$

   Why is $\log(profits)$ *unsuitable? Do these performance variables explain most variation in salary?*
3. *Add tenure:*

   $$\log(salary_i) = \beta_0 + \beta_1 \log(sales_i) + \beta_2 \log(mktval_i) + \beta_3 profits_i + \gamma\, ceoten_i + u_i.$$

   What is the estimated percentage return for another year of CEO tenure, holding other factors fixed?
4. *Compute the sample correlation* $\text{corr}\big(\log(mktval),\, profits\big)$*. Are these variables highly correlated? What does this say about the OLS estimators?*

Problem 7. *Download WAGE1 (description). Verify the partialling out interpretation of the OLS coefficient on educ:*

1. *Regress* $educ_i$ *on* $exper_i$ *and* $tenure_i$*:*

$$educ_i = \alpha_0 + \alpha_1\, exper_i + \alpha_2\, tenure_i + v_i,$$

and save the residuals $r_i \equiv \widehat{v}_i$.

2. *Regress* $\log(wage_i)$ *on* r_i*:*

$$\log(wage_i) = \delta_0 + \delta_1 r_i + e_i.$$

3. *Estimate the full model:*

$$\log(wage_i) = \beta_0 + \beta_1\, educ_i + \beta_2\, exper_i + \beta_3\, tenure_i + u_i.$$

4. *Compare* $\hat{\delta}_1$ *with* $\hat{\beta}_1$.