Problem Set: Linear Regression 2

Problem 1. *Consider the model that satisfies the classical linear model assumptions (linear in parameters, random sampling, no perfect collinearity, zero conditional mean, and homoskedasticity):*

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i,$$

and the null $H_0 :\ \beta_1 - 3\beta_2 = 1$.

1. *Let* $\hat{\beta}_1, \hat{\beta}_2$ *be the OLS estimates. Derive*

   $$Var(\hat{\beta}_1 - 3\hat{\beta}_2)$$

   in terms of $Var(\hat{\beta}_1)$, $Var(\hat{\beta}_2)$, *and* $Cov(\hat{\beta}_1, \hat{\beta}_2)$. *Hence give* $SE(\hat{\beta}_1 - 3\hat{\beta}_2)$.

2. *Write the t-statistic for testing* H_0 *and state the reference distribution (degrees of freedom).*

3. *Define* $u_1 \equiv \beta_1 - 3\beta_2$ *and* $\hat{u}_1 \equiv \hat{\beta}_1 - 3\hat{\beta}_2$. *Write a regression equation involving* β_0, u_1, β_2, *and* β_3 *that allows you to directly obtain* $\hat{u}_1$ *and its standard error.*

Problem 2. *Regression analysis can be used to test whether the market efficiently uses information in valuing stocks. Let return be a firm's 4-year stock return. The efficient markets hypothesis implies return should not be systematically related to variables known in the beginning of the 4-year period: debt-to-capital ratio (dkr), earnings per share (eps), net income (netinc), and CEO pay (salary). To test this hypothesis, we estimated the following regression equation (standard errors in parentheses):*

$$\widehat{return} = \underset{(6.89)}{-14.37} + \underset{(0.201)}{0.321}\, dkr + \underset{(0.078)}{0.043}\, eps - \underset{(0.0047)}{0.0051}\, netinc + \underset{(0.0022)}{0.0035}\, salary$$

$$n = 142,\ R^2 = 0.0395$$

1. *Test whether dkr, eps, netinc, salary are jointly significant at the 5% level (report the* F*-test). Are any coefficients individually significant at 5%?*

2. *Refit with logs for netinc and salary:*

   $$\widehat{return} = \underset{(39.37)}{-36.30} + \underset{(0.203)}{0.327}\, dkr + \underset{(0.080)}{0.069}\, eps - \underset{(3.39)}{4.74}\, ln(netinc) + \underset{(6.31)}{7.24}\, ln(salary)$$

   $$n = 142,\ R^2 = 0.0330.$$

   Do your conclusions about joint/individual significance change?

3. *Some firms have* $dkr = 0$ *and negative eps. Should* $\log(dkr)$ *or* $\log(eps)$ *be used to improve fit? Briefly justify.*

4. *Overall, is the evidence for predictability of returns by 1990 information strong or weak?*

Problem 3. *We have the estimated regression:*

$$\widehat{rdintens} = \underset{(1.429)}{2.613} + \underset{(0.00014)}{0.0003}\, sales - \underset{(0.0000000037)}{0.000000007}\ sales^2, \qquad n = 32,\ R^2 = 0.1484,$$

where rdintens is R&D spending (as percent of sales) and sales is firm sales (millions $)

1. *At what sales level does the marginal effect of sales on rdintens become negative?*

2. *Would you retain the quadratic term in the model? Justify your answer.*

3. *Define* $salesbil \equiv sales/1000$ *(billions of dollars). Rewrite the estimated equation using salesbil and* $salesbil^2$ *as regressors, reporting the transformed coefficients, standard errors, and the same* R^2*.* Hint: $salesbil^2 = sales^2/1000^2$.

4. *For reporting, which specification do you prefer—the original in dollars or the rescaled in billions—and why?*

Problem 4. *Use the data set WAGE1 (description).*

1. *Estimate by OLS:*

$$\log(wage) \;=\; \beta_0 + \beta_1\, educ + \beta_2\, exper + \beta_3\, exper^2 + u,$$

and report the results in the usual format (point estimates, standard errors, n, R^2*).*

2. *Test whether* $exper^2$ *is statistically significant at the* 1% *level (i.e., test* $H_0 : \beta_3 = 0$*).*

3. *Using the semi-elasticity approximation*

$$\%\Delta wage \;\approx\; 100\big(\hat{\beta}_2 + 2\hat{\beta}_3\, exper\big)\, \Delta exper,$$

compute the approximate return to the fifth year of experience and to the twentieth year (exper is measured in years).

4. *At what value* $exper^*$ *does an additional year of experience begin to* reduce *predicted* $\log(wage)$*? How many workers in the sample have* $exper > exper^*$*?*

Problem 5. *An OLS regression explaining CEO salary yields*

$$\widehat{\log(salary)} = \underset{(0.302)}{4.59} + \underset{(0.032)}{0.257}\log(sales) + \underset{(0.042)}{0.011}\,roe + \underset{(0.089)}{0.158}\,finance + \underset{(0.085)}{0.181}\,consprod + \underset{(0.099)}{0.283}\,utility,$$

with $n = 209$ *and* $R^2 = 0.357$, *where finance, consprod, and utility are binary variables indicating the financial, consumer products, and utilities industries. The omitted industry is transportation. The data are from* ***CEOSAL1***.

1. *Holding* **`sales`** *and* **`roe`** *fixed, compute the* approximate *percentage difference in expected salary between the utility and transportation industries. Is this difference statistically significant at the 1% level?*

2. *Using the exact log–level transformation, obtain the* exact *percentage difference for utility versus transportation, and compare it with part (i).*

3. *What is the* approximate *percentage difference between the consumer products and finance industries? Write an equation that would allow you to test whether the difference is statistically significant.*

Problem 6. *Let* $d \in \{0,1\}$ *indicate treatment and let the potential outcomes be* $y(0)$ *and* $y(1)$. *Assume complete random assignment:* $d \perp\!\!\!\perp \{y(0), y(1)\}$. *Define*

$$m_0 \equiv \mathbb{E}[y(0)], \quad m_1 \equiv \mathbb{E}[y(1)], \quad \sigma_0^2 \equiv \mathrm{Var}[y(0)], \quad \sigma_1^2 \equiv \mathrm{Var}[y(1)].$$

1. *Define the observed outcome* $y \equiv (1-d)\,y(0) + d\,y(1)$. *Let* $\tau \equiv m_1 - m_0$ *be the average treatment effect. Show that*

   $$y \;=\; m_0 + \tau\, d \;+\; (1-d)\,v(0) \;+\; d\,v(1),$$

   where $v(0) \equiv y(0) - m_0$ *and* $v(1) \equiv y(1) - m_1$.

2. *Let* $u \equiv (1-d)\,v(0) + d\,v(1)$ *be the error term in*

   $$y \;=\; m_0 + \tau\, d + u.$$

   Show that $\mathbb{E}[u \mid d] = 0$. *What does this imply about the unbiasedness and consistency of the OLS estimator of* τ *from the simple regression* y_i *on* d_i *in a random sample of size* n*? What happens as* $n \to \infty$*?*

3. *Show that*

   $$\mathrm{Var}(u \mid d) \;=\; \mathbb{E}\big[u^2 \mid d\big] \;=\; (1-d)\,\sigma_0^2 \;+\; d\,\sigma_1^2.$$

   Is the error variance generally heteroskedastic?

4. *If you suspect* $\sigma_1^2 \neq \sigma_0^2$ *and* $\hat{\tau}$ *is the OLS estimator, how would you obtain a valid standard error for* $\hat{\tau}$*?*

5. *After obtaining the OLS residuals* $\hat{u}_i$ *for* $i = 1, \ldots, n$, *propose a regression that allows consistent estimation of* σ_0^2 *and* σ_1^2. [Hint: First square the residuals.]