E572 Empirical Problems

Yongseok Kim

February 26, 2020

1. (a) See Figure 1 and 2.
   (b) See Figure 3 and 4.
   (c) From the exercised in (a) and (b), we can conclude that log can resolve a right-skewed data and residuals to have an approximately normal distribution.
   (d) Table 1 reports the regression results. The relevent t-test is reported in parentheses, which is
   $$t(\hat{\beta}) = \frac{\hat{\beta} - \beta}{\sqrt{\hat{\sigma}^2 / \sum (x_i - \bar{x})^2}}$$
   where
   $$\hat{\sigma}^2 = \frac{1}{n}\hat{\varepsilon}'\hat{\varepsilon}$$
   (e) Table 2 reports the regression results. Increased adjusted R-squared suggests that observed skill levels can explain wage differences a lot. However, after controlling skill levels, we can still observe wage gaps between races.
2. (a) Table 3 and 4 report baseline and augmented models, respectively. In the first model, we test
   $$H_0 : \mathbb{E}[\log wage | female = 1, skills] = \mathbb{E}[\log wage | female = 0, skills]$$

Table 1: log(wage) | const + black

	Dependent variable:
	log(wage)
black	-0.202^{***} (0.015)
Constant	2.338^{***} (0.005)
Observations	13,593
R^2	0.013
Adjusted R^2	0.013
Residual Std. Error	0.538 (df = 13591)
F Statistic	181.781^{***} (df = 1; 13591)
Note:	*p<0.1; **p<0.05; ***p<0.01

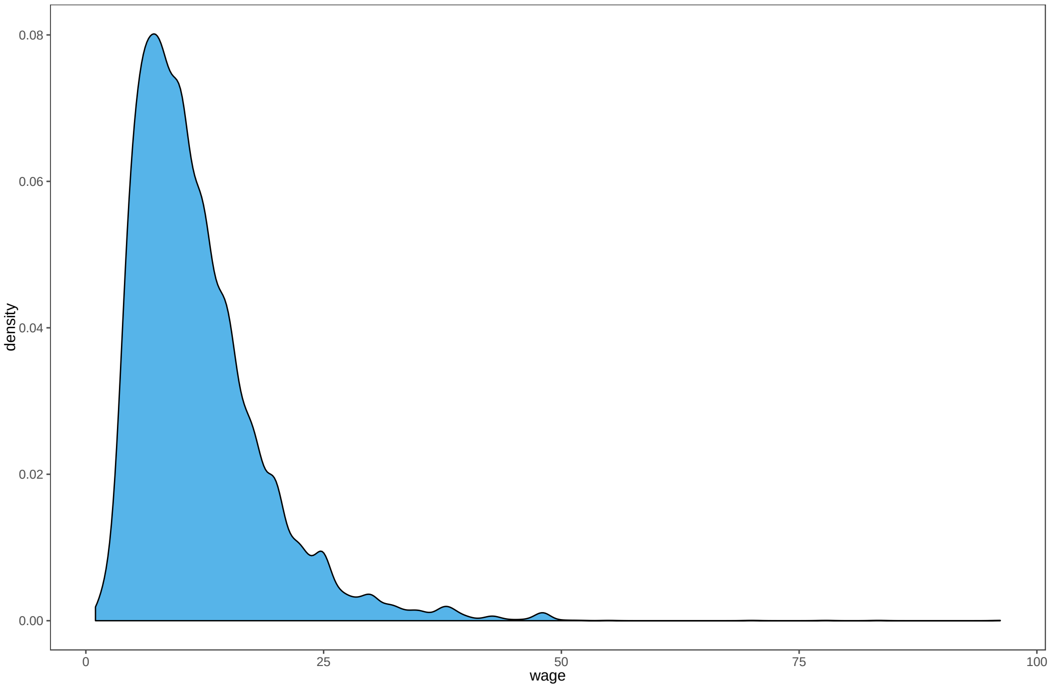

Figure 1: the level of hourly earnings

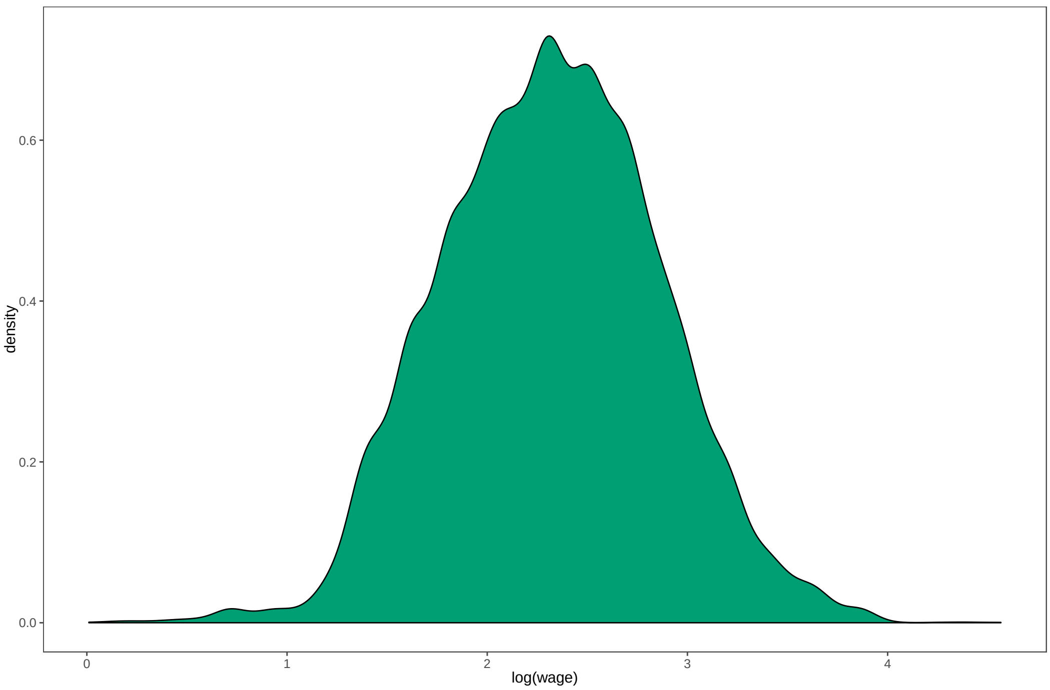

Figure 2: the log of hourly earnings

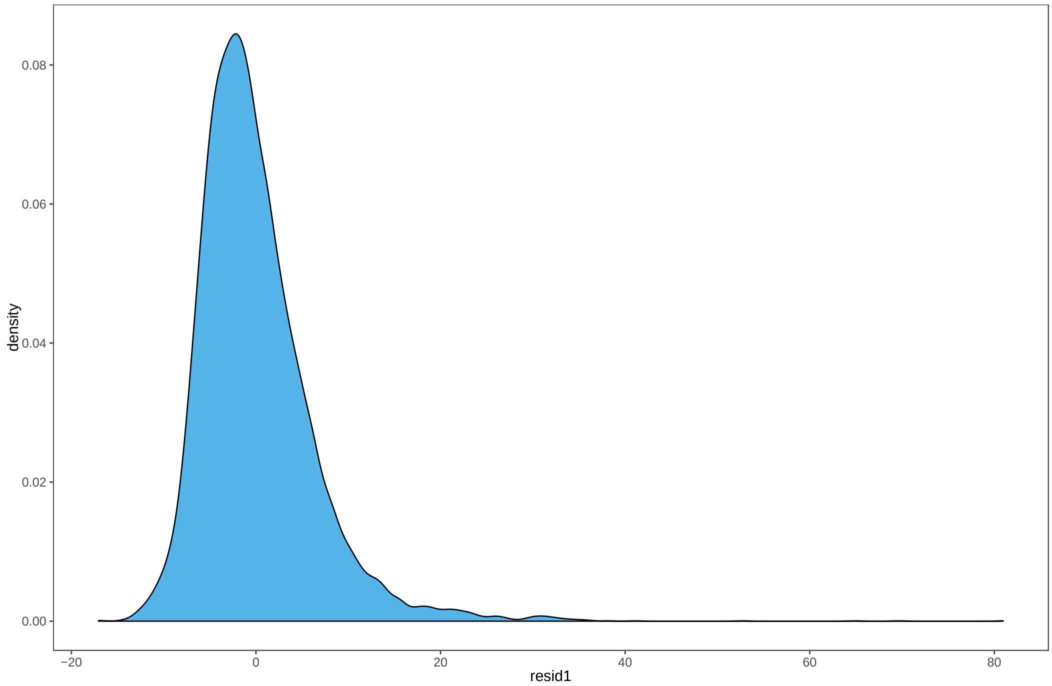

Figure 3: resid1

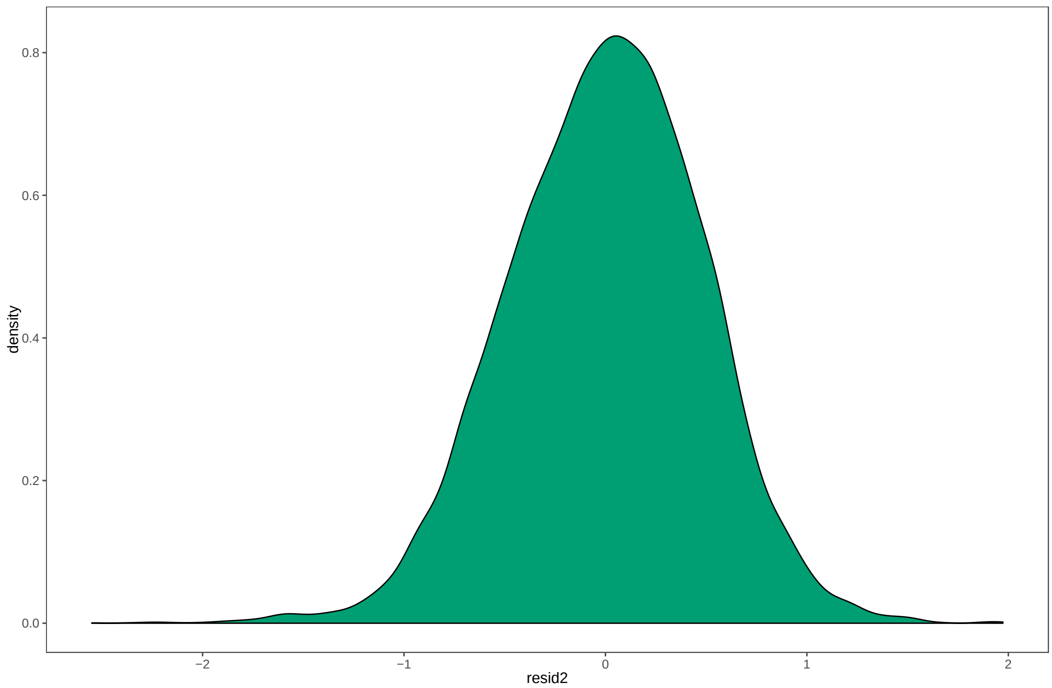

Figure 4: resid2

Table 2: log(wage) | const + black + educ + exp + exp2

	Dependent variable:
	log(wage)
black	-0.149^{***} (0.013)
educ	0.094^{***} (0.002)
poly(exp, 2)1	9.221^{***} (0.513)
poly(exp, 2)2	-2.280^{***} (0.481)
Constant	1.088^{***} (0.022)
Observations	13,593
R^2	0.213
Adjusted R^2	0.212
Residual Std. Error	0.480 (df = 13588)
F Statistic	917.631^{***} (df = 4; 13588)
Note:	*p<0.1; **p<0.05; ***p<0.01

and this null hypothesis is rejected by t-test on the coefficient of $black$. In the second model, we test

$$H_0 : \mathbb{E}[\log wage|female = 1, black = 1, skills] = \mathbb{E}[\log wage|female = 1, black = 0, skills]$$

and this null hypothesis is rejected by t-test on the coefficient of $I(female * black)$.

(b) Table 5 reports a model that allows each of the four groups to have a different intercept. Specifically,

$$Const = \text{single and male}$$

$$Const + married = \text{married and male}$$

$$Const + female = \text{single and female}$$

$$Const + married : female = \text{married and female}$$

To test a joint null hypothesis that there is no difference between the four groups, I estimate a restricted model

$$log(wage)|const + educ + exp + exp2$$

and construct $F =_d F(3, n - k)$. From the test, I obtain

$$F = 584.6024$$

$$p = 0.0000$$

suggesting the null hypothesis is rejected. Next, I test a null hypothesis that there is no difference between single males and single females. From t-test on the coefficient of $female$, we can conclude that this null hypothesis is rejected.

3. (a) First, we have

$$\hat{\alpha} = [S'(I - P_X)S]^{-1}S'(I - P_X)y$$

Table 3: log(wage) | const + female + educ + exp + exp2

	Dependent variable:
	log(wage)
female	-0.310^{***} (0.008)
educ	0.095^{***} (0.002)
poly(exp, 2)1	9.504^{***} (0.488)
poly(exp, 2)2	-2.432^{***} (0.458)
Constant	1.201^{***} (0.021)
Observations	13,593
R^2	0.287
Adjusted R^2	0.287
Residual Std. Error	0.457 (df = 13588)
F Statistic	$1{,}369.810^{***}$ (df = 4; 13588)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 4: log(wage) | const + female + female * black + educ + exp + exp2

	Dependent variable:
	log(wage)
female	-0.317^{***} (0.008)
black	-0.181^{***} (0.019)
educ	0.094^{***} (0.002)
poly(exp, 2)1	9.391^{***} (0.486)
poly(exp, 2)2	-2.472^{***} (0.456)
female:black	0.111^{***} (0.026)
Constant	1.233^{***} (0.021)
Observations	13,593
R^2	0.293
Adjusted R^2	0.293
Residual Std. Error	0.455 (df = 13586)
F Statistic	938.296^{***} (df = 6; 13586)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 5: log(wage) | const + female + married + female * married + educ + exp + exp2

	Dependent variable:
	log(wage)
married	0.159^{***} (0.012)
female	-0.196^{***} (0.014)
educ	0.094^{***} (0.002)
poly(exp, 2)1	8.587^{***} (0.490)
poly(exp, 2)2	-2.129^{***} (0.456)
married:female	-0.160^{***} (0.017)
Constant	1.108^{***} (0.022)
Observations	13,593
R^2	0.296
Adjusted R^2	0.296
Residual Std. Error	0.454 (df = 13586)
F Statistic	953.995^{***} (df = 6; 13586)
Note:	*p<0.1; **p<0.05; ***p<0.01

Second,

$$\begin{aligned}\hat{\gamma} &= [S'(I-P_\iota)S]^{-1}S'(I-P_\iota)\hat{v} \\ &= [S'(I-P_\iota)S]^{-1}S'(I-P_\iota)(I-P_X)y \\ &= [S'(I-P_\iota)S]^{-1}S'(I-P_X)y\end{aligned}$$

since

$$\mathcal{R}(X)^\perp \subset \mathcal{R}(\iota)^\perp$$

(b) Note that

$$\frac{\hat{\alpha}}{\hat{\gamma}} = \frac{S'(I-P_\iota)S}{S'(I-P_X)S} = \frac{1}{1-\bar{R}^2} \geq 1$$

suggesting

$$\hat{\alpha} \geq \hat{\gamma}$$

(c) No, simple is the best. The Urn Model complicates the simple regression without adding any interpretation and may underestimate the sex differential.

4. (a) Note that

$$u_t = \varepsilon_t + \alpha\varepsilon_{t-1} + \alpha^2\varepsilon_{t-2} + \cdots$$

$$\text{var}(u_t) = \frac{\sigma^2}{1-\alpha^2}$$

$$\begin{aligned}\gamma(k) &= \text{cov}(u_t, u_{t-k}) \\ &= \alpha^k \text{cov}(u_{t-k}, u_{t-k}) \\ &= \alpha^k \frac{\sigma^2}{1-\alpha^2}\end{aligned}$$

Then

$$\begin{aligned}\rho_u(k) &= \frac{\gamma(k)}{\text{var}(u_t)} \\ &= \alpha^k\end{aligned}$$

(b) Note that

$$\begin{aligned}v_t &= u_t - u_{t-1} \\ &= \alpha v_{t-1} + \varepsilon_t - \varepsilon_{t-1} \\ &= \alpha^2 v_{t-2} + \varepsilon_t - \varepsilon_{t-1} + \alpha(\varepsilon_{t-1} - \varepsilon_{t-2}) \\ &= \varepsilon_t + (\alpha - 1)\varepsilon_{t-1} + \alpha(\alpha-1)\varepsilon_{t-2} + \cdots\end{aligned}$$

$$\begin{aligned}\text{var}(v_t) &= \sigma^2 + (\alpha-1)^2 \frac{\sigma^2}{1-\alpha^2} \\ &= \frac{2(1-\alpha)}{1-\alpha^2}\sigma^2\end{aligned}$$

$$\begin{aligned}\text{cov}(v_t, v_{t-k}) &= \text{cov}(\alpha^k v_{t-k} - \varepsilon_{t-k}, v_{t-k}) \\ &= \alpha^k \text{var}(v_{t-k}) - \sigma^2\end{aligned}$$

Thena

$$\begin{aligned}\rho_v(k) &= \alpha^k - \frac{1-\alpha^2}{2(1-\alpha)} \\ &= \alpha^k - \frac{1}{2}(1+\alpha)\end{aligned}$$

(c) Note that

$$\rho_u(k) - \rho_v(k) = \frac{1}{2}(1+\alpha)$$

which is increasing in α.

5. (a) First of all, log scale reduces the effect of outliers on estimators. Second, log scale centers a data when it is right-skewed.

   (b) From the coefficient of sex variable, we can conclude that men receive on average approximately 6.38% more than women in the sample. This model is reasonable, but cannot capture cross-departmental differences in gender pay gaps. From $exp(-0.7011) = 0.4960$, we can interpret -0.7110 that people pediatrics department on average earn 50% of the sample average salary.

   (c) Since it is not specified, I assume that unspecified regressors are 0. First, we have

$$\hat{\mathbb{E}}[\log salary84 | sex = 0, cert = 1, assistn = 1, exper = 5, expersq = 25, clin = 1]$$

is equal to

$$12.0696 + 1 \times 0.1854 + 1 \times (-0.1908) + 5 \times 0.0292 + 25 \times (-0.000335) + 1 \times 0.1591 = 12.36092$$

In the case of the male professor, we have

$$12.36092 + 0.0638 = 12.42472$$

The difference in the average salaries in actual dollars would be

$$\exp(12.42472) - \exp(12.36092) = \$15,382.6$$

(d) Defenote R_1^2 is from the unrestricted model, and R_2^2 is from the restricted model. Then

$$\frac{R_1^2 - R_2^2}{1 - R_1^2} = \frac{\tilde{\varepsilon}'\tilde{\varepsilon} - \hat{\varepsilon}'\hat{\varepsilon}}{\hat{\varepsilon}'\hat{\varepsilon}}$$

Note that

$$\frac{R_1^2 - R_2^2}{1 - R_1^2}\frac{261 - 14}{2} =_d F(2, 261 - 14)$$

under the null hypothesis that rank is insignificant. Then, we have

$$F = 21.9174$$

$$p = 0.0000$$

suggesting the null hypothesis is rejected.

(e) First, the coefficient means that after controlling other variables, males earn on average $exp(0.1083) = 1.114832$ times as much as females earn. Second, a much higher t-statistics are driven from the increased coefficient value. Since coefficients of $assistn$ and $associa$ are negative, we can infer that sex and $assistn, associa$ are negatively correlated, which means that males are more likely to be a full professor than females in the sample. Third, under the null hypothesis

$$H_0 : \beta = 0.2083$$

we can construct t-statistics that

$$\begin{aligned} t(\hat{\beta}) &= \frac{0.1083 - 0.2083}{0.0352} \\ &= -2.8409 \end{aligned}$$

By the rule of thumb, we may conclude that the data is inconstent with the null hypothesis.

A Appendix

For R codes to get the above results, see `https://github.com/ysugk/Course-ECON572/blob/master/code/EmpiricalHW.R`