1. (a) **(3 points)** An estimator is a function of a sample of data to be drawn randomly from a population. An estimate is the numerical value of the estimator when it is actually computed using data from a specific sample.

   (b) **(3 points)** X and Y being uncorrelated means they have no linear relationship, but they can have a nonlinear relationship and be not independent.

   (c) **(3 points)** $SE\left(\widehat{\beta}_1\right)$ gets smaller when the variance of X_i gets larger.

   (d) **(3 points)** The OLS estimator is a linear functions of $Y_1, \ldots, Y_n$, i.e. $\widehat{\beta}_1 = \sum_{i=1}^{n} a_i Y_i$.

   (e) **(3 points)** Under the conditional mean independence assumption, control variables can be correlated with residuals.

2. (a) **(5 points)** Let X be the test score for Korean students. We know

$$n_1 = 500$$
$$\overline{X}_1 = 250$$
$$s_1 = 20.$$

Then, a 90% confidence interval for the average test score for Korean students is

$$\begin{aligned}\overline{X}_1 \pm Z_{0.05} \cdot \frac{s_1}{\sqrt{n_1}} &= 250 \pm 1.645 \cdot \frac{20}{\sqrt{500}} \\ &\approx [248.5287, 251.4713].\end{aligned}$$

We have 90% confidence that the true average test score for Korean students falls between 248.5287 and 251.4713.

   (b) **(5 points)** We know

$$n_2 = 500$$
$$\overline{X}_2 = 262$$
$$s_2 = 13.$$

Prepare:

$H_0 : \mu_1 - \mu_2 = 0$ v.s. $H_1 : \mu_1 - \mu_2 \neq 0$, where μ_1 is mean score with original group, and μ_2 is mean score with training group.

Let the significance level be 10%.

Calculate:

$$\begin{aligned} t^* &= \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \\ &= \frac{250 - 262}{\sqrt{\frac{20^2}{500} + \frac{13^2}{500}}} \\ &\approx -11.2489 \end{aligned}$$

Conclude:

Because

$$|t^*| > Z_{0.05} = 1.645,$$

we reject H_0. There is significant evidence that the course helped.

(c) **(5 points)** We know

$$\begin{aligned} n &= 500 \\ \overline{X}_{diff} &= 5 \\ s_{diff} &= 20. \end{aligned}$$

Prepare:

$H_0 : \mu_{diff} = 0$ v.s. $H_1 : \mu_{diff} \neq 0$

Let the significance level be 10%.

Calculate:

$$\begin{aligned} t^* &= \frac{\overline{X}_{diff} - 0}{\sqrt{\frac{s_{diff}^2}{n}}} \\ &= \frac{5 - 0}{\sqrt{\frac{20^2}{500}}} \\ &\approx 5.5902 \end{aligned}$$

Conclude:

Because

$$|t^*| > Z_{0.05} = 1.645,$$

we reject H_0. There is significant evidence that students perform better on their second attempt after taking course.

3. (a) **(5 points)** We know

$$
\begin{aligned}
Y &\sim \left(\mu, \sigma^2\right) \\
\overline{Y} &\sim \left(\mu, \frac{\sigma^2}{n}\right).
\end{aligned}
$$

Then,

$$
\begin{aligned}
E\left(W\right) &= E\left[\left(\frac{n-1}{n}\right)\overline{Y}\right] \\
&= \frac{n-1}{n}E\left(\overline{Y}\right) \\
&= \frac{n-1}{n}\mu \neq \mu.
\end{aligned}
$$

That is, W is biased. Moreover, the bias is

$$
E\left(W\right) - \mu = \frac{n-1}{n}\mu - \mu = -\frac{1}{n}\mu.
$$

(b) **(5 points)** By WLLN, we have

$$
\overline{Y} \xrightarrow{p} \mu.
$$

Also,

$$
\frac{n-1}{n} \xrightarrow{p} 1.
$$

Thus,

$$
W = \frac{n-1}{n}\overline{Y} \xrightarrow{p} \mu.
$$

That is, W is consistent.

(c) **(5 points)**

$$
\begin{aligned}
\text{Var}\left(W\right) &= \text{Var}\left[\left(\frac{n-1}{n}\right)\overline{Y}\right] \\
&= \left(\frac{n-1}{n}\right)^2 \text{Var}\left(\overline{Y}\right) \\
&= \left(\frac{n-1}{n}\right)^2 \frac{\sigma^2}{n} \\
&= \frac{(n-1)^2}{n^3}\sigma^2
\end{aligned}
$$

4. (a) **(3 points)** The objective function is

$$
\min_{\widehat{\alpha}} \quad \sum_{i=1}^{n} \left(Y_i - \widehat{\alpha}\right)^2.
$$

By FOC, we have

$$\begin{aligned}\sum_{i=1}^{n} 2\left(Y_i-\widehat{\alpha}\right)(-1) &= 0\\ \sum_{i=1}^{n}\left(Y_i-\widehat{\alpha}\right) &= 0\\ \sum_{i=1}^{n} Y_i-n\widehat{\alpha} &= 0.\end{aligned}$$

Thus, the least squares estimator of α is

$$\widehat{\alpha}=\frac{\sum_{i=1}^{n} Y_i}{n}=\overline{Y}.$$

(b) **(3 points)** By FOC, we know

$$\begin{aligned}\sum_{i=1}^{n}\widehat{u}_i &= \sum_{i=1}^{n}\left(Y_i-\widehat{Y}_i\right)\\ &= \sum_{i=1}^{n}\left(Y_i-\widehat{\alpha}\right)=0.\end{aligned}$$

5. **(4 points)** Consider the simple regression

$$SALARY_i=\alpha_0+\alpha_1 GPA_i+v_i.$$

We know

$$\begin{aligned}&\operatorname{Cov}(GPA, BEAUTY)<0\\ &\operatorname{Cov}(BEAUTY, SALARY)>0 \quad\Rightarrow\quad \beta_2>0.\end{aligned}$$

Then,

$$\widehat{\alpha}_1 \xrightarrow{p} \alpha_1+\beta_2\cdot\frac{\operatorname{Cov}(GPA, BEAUTY)}{\sigma_{GPA}^2}<\alpha_1.$$

6. (a) **(5 points)** The proportion of fast-food restaurants in the data located in Pennsylvania is

$$1-0.81=0.19=19\%.$$

(b) **(5 points) Prepare**:

$H_0:\beta_{PrpBlack}=0$ v.s. $H_1:\beta_{PrpBlack}\neq 0$

Let the significance level be 5%.

Calculate:

$$t^*=\frac{0.0649-0}{0.0240}\approx 2.7042$$

Conclude:

Because

$$|t^*| > Z_{0.025} = 1.96,$$

we reject H_0. There is significant evidence that the coefficient on $PrpBlack$ in Model 1 is different from 0. If the share of blacks increases by 10 percentage points, $PriceSoda$ is expected to increase by \$0.0649 × 0.1 ≈ \$0.0065 on average.

(c) **(5 points)**

$$\widehat{PriceSoda}_i = 0.9563 + 0.1150 PrpBlack_i + 0.0000016 Income_i$$

(d) **(5 points)** When we control for income, the coefficient on $PrpBlack$ in Model 2 is larger than in Model 1. Because

$$\text{Cov}\,(PrpBlack, Income) < 0$$
$$\text{Cov}\,(Income, PriceSoda) > 0,$$

$\widehat{\beta}_{PrpBlack}$ in Model 2 is more positive than in Model 1.

(e) **(5 points) Prepare**:

$H_0 : \beta_{NJ} = 0$ v.s. $H_1 : \beta_{NJ} \neq 0$

Let the significance level be 5%.

Calculate:

$$t^* = \frac{0.0775 - 0}{0.0098} \approx 7.9082$$

Conclude:

Because

$$|t^*| > Z_{0.025} = 1.96,$$

we reject H_0. There is significant evidence that the coefficient on NJ in Model 3 is different from 0. $\widehat{\beta}_{NJ}$ means that other things being equal, $PriceSoda$ of New Jersey is \$0.0775 higher than that of Pennsylvania on average.

(f) **(5 points)**

$$\begin{aligned}\overline{R}^2 &= 1 - \frac{n-1}{n-k-1}\left(1 - R^2\right) \\ &= 1 - \frac{401-1}{401-3-1}(1 - 0.1689) \approx 0.1626\end{aligned}$$

It means that the model explains 16.26% of the variability in $PriceSoda$.

(g) **(5 points) Prepare**:

$H_0 : \beta_{KFC} = 0$ v.s. $H_1 : \beta_{KFc} \neq 0$

Let the significance level be 5%.

Calculate:

$$t^* = \frac{0.0252 - 0}{0.0118} \approx 2.1356$$

Conclude:

Because

$$|t^*| > Z_{0.025} = 1.96,$$

we reject H_0. There is significant evidence that the coefficient on KFC in Model 4 is different from 0. $\widehat{\beta}_{KFC}$ means that other things being equal, $PriceSoda$ in Kentucky Fried Chicken is \$0.0252 higher than in Wendy's.

(h) **(5 points) Prepare**:

$H_0 : \beta_{BK} = \beta_{KFC} = \beta_{RR} = 0$ v.s. H_1 : At least one of the coefficients in H_0 is not 0.

(i) **(5 points)** Yes, we can. The estimated coefficients are

$$\widehat{\gamma}_4 = 0.8954 + 0.0796 = 0.975$$
$$\widehat{\gamma}_5 = 0.8954 + 0.0252 = 0.9206$$
$$\widehat{\gamma}_6 = 0.8954 + 0.1274 = 1.0228$$
$$\widehat{\gamma}_7 = 0.8954.$$