1. [10] Donald Trump estimates the simple regression $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. He has 300 observations, $\sum x_i = 300$, $\sum y_i = 0$, $\|x\|^2 = 30000$, $\sum x_i y_i = 100$, $\|y\| = 100$.
   (a) [4] Estimate coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$.
   (b) [4] Calculate 95% confidence interval for β_1.
   (c) [2] Test hypothesis that $\beta_1 = 1$ against alternative $\beta_1 \neq 1$ using 5% significance level.

2. [10] Consider the model $y = X\beta + u$ where β is non-random, $\mathbb{E}(u \mid X) = 0$. The matrix X of size $n \times k$ has rank $X = k$ and $\mathbb{V}\mathrm{ar}(u \mid X) = \sigma^2 I$. Let $\hat{\beta}$ be the standard OLS estimator of β.
   (a) [2] Find $\mathbb{E}(\hat{y} \mid X)$.
   (b) [4] Find $\mathbb{V}\mathrm{ar}(\hat{y} \mid X)$ and $\mathbb{V}\mathrm{ar}(\hat{u} \mid X)$.
   (c) [4] Prove that $H_{ii} \in [0; 1]$ if $H = X(X^T X)^{-1} X^T$.

3. [10] The whole dataset of $n = 603$ observations is split into two parts: 600 observations and 3 separated observations. Donald Trump estimated two regressions.

   Regression $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 w_i$ on the whole dataset with SST $= 800$ and $\mathrm{SS}^{\mathrm{res}} = 650$. The same regression on the first part of 600 observations with $\mathrm{SS}^{\mathrm{res}}_1 = 600$.
   (a) [4] Test H_0: $\beta_1 = 0$ and $\beta_2 = 0$ on the whole dataset against H_1: $\beta_1 \neq 0$ or $\beta_2 \neq 0$.
   (b) [6] Test H_0 that the separated observations are not outliers against an alternative that the linear model is valid only for the first 600 observations.

You are free to use these 5% critical values: $F_{1,597} = 3.9$, $F_{2,597} = 3.0$, $F_{3,597} = 2.6$, $F_{4,597} = 2.4$, $F_{5,597} = 2.2$, $F_{6,597} = 2.1$, $F_{1,591} = 3.9$, $F_{2,591} = 3.0$, $F_{3,591} = 2.6$, $F_{4,591} = 2.4$, $F_{5,591} = 2.2$, $F_{6,591} = 2.1$, $F_{2,600} = 3.0$, $F_{3,600} = 2.6$.

4. [10] The true model is $y_i = \beta_0 + \beta_1 x_i + u_i$ with $\mathbb{V}\mathrm{ar}(u \mid x) = \sigma^2 I$. Observations are independent. Winnie-the-Pooh observes y but does not observe x. Instead of x he observes a and b, $a_i = x_i + v_i^a$, $b_i = x_i + v_i^b$. Random variables u_i, v_i^a and v_i^b are independent, $\mathbb{E}(u \mid x) = \mathbb{E}(v^b \mid x) = \mathbb{E}(v^a \mid x) = 0$.

   Consider regression A: $\hat{a}_i = \hat{\gamma}_0 + \hat{\gamma}_1 b_i$ and regression B: $\hat{y}_i = \hat{\delta}_0 + \hat{\delta}_1 b_i$.

   (a) [3] Find $\operatorname{plim} \hat{\gamma}_1$ in terms of $\mathbb{V}\mathrm{ar}(v_i^b)$ and $\mathbb{V}\mathrm{ar}(x_i)$.

   (b) [5] Find $\operatorname{plim} \hat{\delta}_1$ in terms of β_1, $\mathbb{V}\mathrm{ar}(v_i^b)$ and $\mathbb{V}\mathrm{ar}(x_i)$.

   (c) [2] Construct a consistent estimator of β_1 using $\hat{\gamma}_1$ and $\hat{\delta}_1$.

5. [10] Observations are independent. The vector of regressor and random error (x_i, u_i) is uniformly distributed inside the parallelogram $ABCD$, $A = (-3, 0)$, $B = (3, 2)$, $C = (3, 0)$, $D = (-3, -2)$.

   (a) [4] Find $\mathbb{E}(u_i \mid x_i)$, $\mathbb{V}\mathrm{ar}(u_i \mid x_i)$.

   (b) [3] Find $\mathbb{E}(u_i)$, $\mathbb{C}\mathrm{ov}(x_i, u_i)$.

   (c) [1] Which Gauss — Markov assumptions are violated?

   (d) [2] Is the OLS estimator $\hat{\beta}_1$ in the model $y_i = \beta_0 + \beta_1 x_i + u_i$ consistent?

6. [10] (from LSE past exams)

   Let's consider how workplace smoking ban affect the incidence of smoking. We use the data on $n = 10000$ US indoor workers from 1991 to 1993. The data is taken from the article 'Do workplace Smoking Bans Reduce Smoking' by Evans et al.

   The smoker is a dummy variable (1 if a worker smokes and 0 if no). The smkban is a dummy variable (1 if there is a ban on smoking, 0 otherwise).

   The first regression is

$$\widehat{\text{smoker}}_i = \underset{0.007}{0.3} - \underset{0.009}{0.078}\text{smkban}_i, R^2 = 0.0078, \text{SS}^{\text{res}} = 1820.$$

   Standard errors are in parentheses.

   (a) [2] Interpret the parameter estimates of the coefficient on smkban.

   (b) [2] Provide the approximate 95% confidence interval for the coefficient on smkban.

   (c) [1] Test the hypothesis that $\beta_{\text{smkban}} = 0$.

   The second regression is

$$\widehat{\text{smoker}}_i = \underset{(0.02)}{0.2} - \underset{(0.045)}{0.009}\text{smkban}_i + \underset{(0.009)}{-0.033}\text{fem}_i - \underset{(0.0003)}{0.001}\,\text{age}_i - \underset{(0.016)}{-0.027}\text{black}_i - \underset{(0.014)}{0.1}\,\text{hisp}_i +$$
$$+ \underset{(0.02)}{0.3}\,\text{e}_{1i} + \underset{(0.012)}{0.2}\,\text{e}_{3i} + \underset{(0.012)}{0.16}\,\text{e}_{3i} + \underset{(0.012)}{0.042}\text{e}_{4i}, R^2 = 0.0526, \text{SS}^{\text{res}} = 1736.$$

   Here e_1 is a dummy for highschool dropout, e_2 — for highschool graduate, e_3 — for some college, e_4 — for college graduate, e_5 — for master degree or above.

   (d) [2] Compare the coefficient on smkban in the long and short models. Explain why the estimates differ.

   (e) [3] Interpret the estimate of the coefficient on e_2. Explain how would you obtain its p-value and what information p-value provides.