Cover Page

Given Name: Animesh

Family Name: Tirkey

CCID: tirkey

Student ID: 1789443

Please answer all questions on the test page in the space provided. For short answers questions, a sentence with suffice. There is an extra scratch page at the end. Please submit all pages when you are done.

Income by Occupation (15 Points)

Consider a linear regression with sample size $n = 100$ different occupations. We aim to model average income on the log-scale (y) based on average education (x_1), percentage of women in the occupation (x_2), a prestige score (x_3), and the type of job (categorical, μ_i).

$\log(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \mu_i$

1. (**2 points**)

   There are 3 types of jobs in the dataset: Technical, Blue collar, White Collar. How many parameters are required to fit a one-way ANOVA model

   $$y_{i,j} = \mu_i + \varepsilon_{i,j} \qquad \text{(model 1)}$$

   where μ_i is the mean for the ith type of job.

   \# parameters = $\boxed{3}$

2. (**4 points**)

   If we include the continuous predictor variables, we get

   $$y_{i,j} = \mu_i + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon_{i,j} \qquad \text{(model 2).}$$

   What are the degrees of freedom for the partial F-test comparing model 1 to model 2?

   DoFs = $\boxed{3}$

3. Another model is fit without the categorical variable:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon_i \qquad \text{(model 3).}$$

The estimated coefficients are in the following table:

	Estimate	StdErr	t-value	p-value
intercept, β_0	3.3980	0.0559	60.73	<2e-16
education, β_1	-0.0038	0.0096	-0.40	0.69
women, β_2	-0.0035	0.0004	-7.98	2.82e-12
prestige, β_3	0.0108	0.0015	7.02	2.90e-10

(a) (**2 points**)
What is the relationship between log-income and education level?

1 unit increase = -0.0038 in log income

(b) (**2 points**)
What is the relationship between log-income and percentage of women in the occupation?

"

(c) (**2 points**)
What is the relationship between log-income and the prestige score?

"

4. True or False, the following models can be compared using a partial F-test:

   (a) (**1 points**)

   Model 1 and Model 2? T

   (b) (**1 points**)

   Model 1 and Model 3? T

   (c) (**1 points**)

   Model 2 and Model 3? T

Computing Variances (15 Points)

First, consider the standard linear model $Y = X\beta + \varepsilon$.

1. (**1 point**) Write down the least squares estimator $\hat{\beta}$ in terms of the design matrix X and the vector Y.

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

2. (**2 point**) Assuming that $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, derive the distribution of $\hat{\beta}$.

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$
$$= (X^T X)^{-1} X^T (X\beta + \varepsilon)$$
$$= (X^T X)^{-1} X^T X \beta + (X^T X^{-1}) X^T \varepsilon$$
$$= \beta + (X^T X)^{-1} X^T \varepsilon$$
$$E\{\hat{\beta}\} = \beta + (X^T X)^{-1} X^T E[\varepsilon]$$
$$= \beta + (X^T X)^{-1} X^T \cdot 0$$
$$= \beta$$
$$\mathrm{Var}\{\hat{\beta}\} = \mathrm{Var}\{(X^T X)^{-1} X^T \varepsilon\}$$
$$=$$

Next, consider the linear regression $y_i = a + bx_i + cz_i + \varepsilon_i$ for $i = 1, \ldots, n$, and assume that $\sum_{i=1}^n x_i = \sum_{i=1}^n z_i = \sum_{i=1}^n x_i z_i = 0$.

3. (**4 points**)

   Derive the variance for a predicted value, $\hat{a} + \hat{b}x + \hat{c}z$, at some points $x, z \in \mathbb{R}$.

   $\mathrm{Var}[\hat{a} + \hat{b}x + \hat{c}z]$

4. (**4 points**)

   Derive the covariance between two predicted values, $\hat{a} + \hat{b}x + \hat{c}z$ and $\hat{a} + \hat{b}u + \hat{c}v$, at some points $x, z, u, v \in \mathbb{R}$.

5. (**4 points**) Recalling that the general formula for a confidence ellipsoid is

$$\frac{(\hat{\beta}-\beta)^{\mathrm{T}}X^{\mathrm{T}}X(\hat{\beta}-\beta)/(p+1)}{SS_{\text{res}}/(n-p-1)} \le F,$$

Show that the confidence ellipsoid for $(\hat{a}, \hat{b}, \hat{c})$ can be written as

$$\frac{n-3}{3SS_{\text{res}}}\left[(\hat{a}-a)^2 n + (\hat{b}-b)^2\sum_{i=1}^{n} x_i^2 + (\hat{c}-c)^2\sum_{i=1}^{n} z_i^2\right] \le F.$$

Scratch Paper