Economics 241B Exercise 2 Proposed Solution
Best Linear Prediction and Regression

1. Assume $\mathbb{E}(y) < \infty$.

a. Prove

$$\mathbb{E}(\mathbb{E}(y|x)) = \mathbb{E}(y).$$

Proof. We have

$$\mathbb{E}(\mathbb{E}(y|x)) = \int_{\mathbb{R}^k} \left(\int y f_{y|x}(y|x)\, dy \right) f_x(x)\, dx.$$

Because $f_{y|x}(y|x) f_x(x) = f_{y,x}(y,x)$,

$$\int_{\mathbb{R}^k} \left(\int y f_{y|x}(y|x)\, dy \right) f_x(x)\, dx = \int_{\mathbb{R}^k} \int y f_{y,x}(y,x)\, dy dx = \int y f_y(y)\, dy,$$

because $\int_{\mathbb{R}^k} f_{y,x}(y,x)\, dx = f_y(y)$.

b. Prove

$$\mathbb{E}(\mathbb{E}(y|x_1,x_2)|x_1) = \mathbb{E}(y|x_1).$$

Proof. We have

$$\mathbb{E}(\mathbb{E}(y|x_1,x_2)|x_1) = \int_{\mathbb{R}^{k_2}} \left(\int y f_{y|x_1,x_2}(y|x_1,x_2)\, dy \right) f_{x_2|x_1}(x_2|x_1)\, dx_2.$$

Observe that (dropping subscripts to avoid notational clutter)

$$\begin{aligned} f(y|x_1,x_2) f(x_2|x_1) &= \frac{f(y,x_1,x_2)}{f(x_1,x_2)} \frac{f(x_1,x_2)}{f(x_1)} \\ &= f(y,x_2|x_1). \end{aligned}$$

Thus

$$\begin{aligned} \int_{\mathbb{R}^{k_2}} \left(\int y f_{y|x_1,x_2}(y|x_1,x_2)\, dy \right) f_{x_2|x_1}(x_2|x_1)\, dx_2 &= \int_{\mathbb{R}^{k_2}} \int y f_{y,x_2|x_1}(y,x_2|x_1)\, dy dx_2 \\ &= \int y f_{y|x_1}(y|x_1)\, dy, \end{aligned}$$

because $\int_{\mathbb{R}^{k_2}} f_{y,x_2|x_1}(y,x_2|x_1)\, dx_2 = f_{y|x_1}(y|x_1)$.

2. Consider a dependent variable y for which

$$\begin{aligned}\mathbb{E}(y|x) &= \beta_2 x^2 + \beta_1 x + \beta_0, \\ y &= \beta_2 x^2 + \beta_1 x + \beta_0 + e,\end{aligned}$$

where $e \sim \mathcal{N}(0, \sigma^2(x))$.

a. Determine the distribution of y given x.

Answer. Given x, $\beta_2 x^2 + \beta_1 x + \beta_0 := m(x)$ is a real number and the distribution of $m(x) + e$ is simply $\mathcal{N}(m(x), \sigma^2(x))$.

b. For any $h(x)$ such that $\mathbb{E}|h(x)e| < \infty$, prove the following statements:

$$\begin{aligned} i)\ \mathbb{E}(e|x) &= 0, \\ ii)\ \mathbb{E}(h(x)e) &= 0.\end{aligned}$$

Clearly state why the condition $\mathbb{E}|h(x)e| < \infty$ is needed. Do these statements imply that the covariate x is uncorrelated with the (conditional expectation function) error e?

Proof. For statement $i)$,

$$\mathbb{E}(y|x) = \beta_2 x^2 + \beta_1 x + \beta_0 + \mathbb{E}(e|x),$$

hence $\mathbb{E}(e|x) = 0$ is implied by the initial definition of $\mathbb{E}(y|x)$.
For statement $ii)$, if $\mathbb{E}|h(x)e| < \infty$, then $\mathbb{E}(h(x)e)$ exists, so

$$\begin{aligned}\mathbb{E}(h(x)e) &= \int\int h(x)\, e f_{e,x}(e,x)\, dedx \\ &= \int h(x) \left(\int e f_{e|x}(e|x)\, de \right) f_x(x)\, dx \\ &= \int h(x) (\mathbb{E}(e|x)) f_x(x)\, dx = 0,\end{aligned}$$

where the second equality follows from the fact $f_{e,x}(e,x) = f_{e|x}(e|x) f_x(x)$.
Under statement $ii)$, all functions of x are uncorrelated with e, hence $h(x) = x$ implies that the covariate x is uncorrelated with e.

c. We have shown (in class) that $\beta_2 x^2 + \beta_1 x + \beta_0 := m(x)$ is the predictor of y that minimizes the mean-squared prediction error. Consider predicting e^2 and write the mean-squared error of a predictor $g(x)$. Show that $\sigma^2(x)$ minimizes this mean-squared error.

Answer. The mean-squared error is

$$\begin{aligned}\mathbb{E}\left(e^2 - g(x)\right)^2 &= \mathbb{E}\left(e^2 - \sigma^2(x)\right)^2 + 2\mathbb{E}\left[\left(e^2 - \sigma^2(x)\right)\left(\sigma^2(x) - g(x)\right)\right] + \mathbb{E}\left(\sigma^2(x) - g(x)\right)^2 \\ &= \mathbb{E}\left(e^2 - \sigma^2(x)\right)^2 + \mathbb{E}\left(\sigma^2(x) - g(x)\right)^2,\end{aligned}$$

where the second equality follows from the fact that $\mathbb{E}\left[\left(e^2 - \sigma^2(x)\right)|x\right] = 0$. From the second term, the prediction error is minimized if $g(x) = \sigma^2(x)$.

3. Let $g(\cdot) : \mathbb{R}^m \to \mathbb{R}$ be a convex function.

a. For any random vector x, if $\mathbb{E}\|x\| < \infty$ and $\mathbb{E}|g(x)| < \infty$, prove (Jensen's Inequality)

$$g(\mathbb{E}(x)) \leq \mathbb{E}(g(x)).$$

Proof. Because $g(x)$ is a convex function, at each point x there is at least one linear function (called a subderivative) that touches $g(x)$ at x and that lies below the function at all points. Let $a + b^{\mathrm{T}}x$ be such a function at $x = \mathbb{E}(x)$, so

$$a + b^{\mathrm{T}}\mathbb{E}(x) \leq g(x).$$

Applying expectations,

$$\mathbb{E}\left(a + b^{\mathrm{T}}\mathbb{E}(x)\right) \leq \mathbb{E}(g(x)),$$

yet $a + b^{\mathrm{T}}\mathbb{E}(x) = g(\mathbb{E}(x))$, so $g(\mathbb{E}(x)) \leq \mathbb{E}(g(x))$ as stated.

b. With $m = 1$, use Jensen's Inequality to bound $(\mathbb{E}(x))^2$.

Answer. We have $g(u) = u^2$, so Jensen's Inequality states

$$(\mathbb{E}(x))^2 \leq \mathbb{E}\left(x^2\right).$$

c. For any random vectors (y, x), if $\mathbb{E}\|y\| < \infty$ and $\mathbb{E}|g(y)| < \infty$, prove (Conditional Jensen's Inequality)

$$g(\mathbb{E}(y|x)) \leq \mathbb{E}(g(y)|x).$$

Proof. Because $g(y)$ is a convex function, at each point y there is at least one linear function (called a subderivative) that touches $g(y)$ at y and that lies below the function at all points. Let $a + b^{\mathrm{T}} y$ be such a function at $y = \mathbb{E}(y|x)$, so

$$a + b^{\mathrm{T}} \mathbb{E}(y|x) \leq g(y).$$

Applying conditional expectations

$$\mathbb{E}\left(\left(a + b^{\mathrm{T}} \mathbb{E}(y|x)\right)|x\right) \leq \mathbb{E}(g(y)|x),$$

yet $a + b^{\mathrm{T}} \mathbb{E}(y|x) = g(\mathbb{E}(y|x))$, so $g(\mathbb{E}(y|x)) \leq \mathbb{E}(g(y)|x)$ as stated. Note the conditional expectations exist because $\mathbb{E}\|y\| < \infty$ and $\mathbb{E}|g(y)| < \infty$.

d. With $m = 1$, use the Conditional Jensen's Inequality to bound $(\mathbb{E}(y|x))^2$.

Answer. We have $g(u) = u^2$, so the Conditional Jensen's Inequality states

$$(\mathbb{E}(y|x))^2 \leq \mathbb{E}\left(y^2|x\right).$$

4. You are asked to determine how the conditional mean of a discrete variable y depends on a (continuous) conditioning variable x. With a discrete dependent variable, the assumption about the form of the conditional mean is replaced with an assumption about the entire conditional distribution for y. You need to consider two cases.

Case 1: y takes only 2 values, $y \in \{0, 1\}$. Assume

$$\mathbb{P}(y = 1|x) = x^{\mathrm{T}} \beta_0.$$

(The conditional distribution of y given x is Bernoulli.)

Case 2: y takes positive integer values, $y \in \{0, 1, 2, \ldots\}$. Assume

$$\mathbb{P}(y = k|x) = \frac{\exp\left(-x^{\mathrm{T}} \beta_0\right)\left(x^{\mathrm{T}} \beta_0\right)^k}{k!} \quad k = 0, 1, 2, \ldots$$

(The conditional distribution of y given x is Poisson.)

a) For Case 1, compute $\mathbb{E}(y|x)$. Does this justify a linear regression model of the form $y = x^{\mathrm{T}} \beta_0 + u$?

Answer. Because $\mathbb{P}(y = 1|x) = x^{\mathrm{T}} \beta_0$,

$$\mathbb{E}(y|x) = 0 \cdot \mathbb{P}(y = 0|x) + 1 \cdot \mathbb{P}(y = 0|x) = x^{\mathrm{T}} \beta_0.$$

This justifies a linear regression model of the form $y = x^{\mathrm{T}}\beta_0 + u$ with two features. First, $x^{\mathrm{T}}\beta_0$ is a probability and thus should be restricted to $[0, 1]$. Yet there is no such restriction in the functional form, so for certain values of x the probability would lie outside of $[0, 1]$. Second, $\mathbb{E}(u|x) = 0$ by construction, because the error is also bivariate with

$$\mathbb{E}(u|x) = -x^{\mathrm{T}}\beta_0 \left(1 - x^{\mathrm{T}}\beta_0\right) + \left(1 - x^{\mathrm{T}}\beta_0\right) x^{\mathrm{T}}\beta_0 = 0.$$

b) For Case 1, compute $Var(y|x)$. Does this justify an alternative estimator to OLS?

Answer. Because the error is bivariate with conditional mean 0,

$$\begin{aligned} Var(u|x) &= \mathbb{E}\left(u^2|x\right) = \left(x^{\mathrm{T}}\beta_0\right)^2 \left(1 - x^{\mathrm{T}}\beta_0\right) + \left(1 - x^{\mathrm{T}}\beta_0\right)^2 x^{\mathrm{T}}\beta_0 \\ &= x^{\mathrm{T}}\beta_0 \left(1 - x^{\mathrm{T}}\beta_0\right). \end{aligned}$$

Thus the errors are heteroskedastic. With heteroskedastic errors, and with the form of the heteroskedasticity known, weighted least squares is justified. The only catch, if $x^{\mathrm{T}}\beta_0$ lies outside $[0, 1]$, then the conditional variance is negative. Once again, this is only a useful way of capturing the effect of x on the distribution of y for values of x for which $x^{\mathrm{T}}\beta_0 \in [0, 1]$.

c) For Case 2, compute $\mathbb{E}(y|x)$. Does this justify a linear regression model of the form $y = x^{\mathrm{T}}\beta_0 + u$? Hint:

$$\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = \exp(\lambda).$$

Answer. Because of the Poisson assumption

$$\begin{aligned}
\mathbb{E}(y|x) &= \sum_{k=0}^{\infty} k \frac{\exp\left(-x^{\mathrm{T}}\beta_0\right)\left(x^{\mathrm{T}}\beta_0\right)^k}{k!} \\
&= \exp\left(-x^{\mathrm{T}}\beta_0\right)\sum_{k=1}^{\infty} k \frac{\left(x^{\mathrm{T}}\beta_0\right)^k}{k!} \quad \text{because the } k=0 \text{ term is } 0 \\
&= x^{\mathrm{T}}\beta_0 \exp\left(-x^{\mathrm{T}}\beta_0\right)\sum_{k=1}^{\infty} \frac{\left(x^{\mathrm{T}}\beta_0\right)^{k-1}}{(k-1)!} \quad \text{divide bottom by } k \\
&= x^{\mathrm{T}}\beta_0 \exp\left(-x^{\mathrm{T}}\beta_0\right)\sum_{k=0}^{\infty} \frac{\left(x^{\mathrm{T}}\beta_0\right)^{k}}{k!} \\
&= x^{\mathrm{T}}\beta_0 \exp\left(-x^{\mathrm{T}}\beta_0\right)\exp\left(x^{\mathrm{T}}\beta_0\right) = x^{\mathrm{T}}\beta_0.
\end{aligned}$$

This does indeed justify use of a linear regression model of the form $y = x^{\mathrm{T}}\beta_0 + u$. The restriction here is that, for a Poisson density, $x^{\mathrm{T}}\beta_0 > 0$, which is the reason that Poisson regression models are not specified in this way.

d) For Case 2, compute $Var(y|x)$. Does this justify an alternative estimator to OLS? Hint:

$$\mathbb{E}\left(y^2|x\right) = \mathbb{E}\left[\left(y(y-1)+y\right)|x\right].$$

Answer. To compute the variance observe that $\mathbb{E}(y^2) = \mathbb{E}\left[y(y-1)+y\right]$, so that we need to calculate

$$\begin{aligned}
\mathbb{E}\left(y(y-1)|x\right) &= \sum_{k=0}^{\infty} k(k-1)\frac{\exp\left(-x^{\mathrm{T}}\beta_0\right)\left(x^{\mathrm{T}}\beta_0\right)^k}{k!} \\
&= \exp\left(-x^{\mathrm{T}}\beta_0\right)\sum_{k=2}^{\infty} k(k-1)\frac{\left(x^{\mathrm{T}}\beta_0\right)^k}{k!} \quad \text{because the } x=0 \text{ and } x=1 \text{ terms are } 0 \\
&= \left(x^{\mathrm{T}}\beta_0\right)^2 \exp\left(-x^{\mathrm{T}}\beta_0\right)\sum_{k=2}^{\infty}\frac{\left(x^{\mathrm{T}}\beta_0\right)^{k-2}}{(k-2)!} \quad \text{divide bottom by } k(k-1) \\
&= \left(x^{\mathrm{T}}\beta_0\right)^2 \exp\left(-x^{\mathrm{T}}\beta_0\right)\sum_{k=0}^{\infty}\frac{\left(x^{\mathrm{T}}\beta_0\right)^{k}}{k!} \\
&= \left(x^{\mathrm{T}}\beta_0\right)^2.
\end{aligned}$$

Thus

$$
\begin{aligned}
Var\left(y|x\right) &= \mathbb{E}\left(y^2|x\right) - \left(\mathbb{E}\left(y|x\right)\right)^2 \\
&= \mathbb{E}\left[\left(y\left(y-1\right)+y\right)|x\right] - \left(\mathbb{E}\left(y|x\right)\right)^2 \\
&= \left(x^{\mathrm{T}}\beta_0\right)^2 + x^{\mathrm{T}}\beta_0 - \left(x^{\mathrm{T}}\beta_0\right)^2 = x^{\mathrm{T}}\beta_0.
\end{aligned}
$$

Thus the errors are heteroskedastic. With heteroskedastic errors, and with the form of the heteroskedasticity known, weighted least squares is justified. The only catch, $x^{\mathrm{T}}\beta_0$ must be positive for all values of the covariates.