פתרון

שאלה 1

א. ערכו של β_6 הוא ההפרש בערכו של Y עבור גבר משכיל, כאשר מוסיפים יחידה ל-X.

ב. המודל הסופי כולל רק את U_1 (המגדר), X, האינטראקציה ביניהם והחותך.

ג. כיון שמשתנה הדמי של ההשכלה וכל האינטראקציות שלו עם המשתנים האחרים לא נכנסו למודל, ניתן לומר ע"ס הניתוח הסטטיסטי שרמת ההשכלה לא עוזרת לחזות את המשתנה Y. לעומת זאת, המגדר תורם לחיזוי הן ישירות והן מבחינת אופן ההתחשבות במשתנה X.

ד. תחת ההנחה ערך החותכים בכל אחת מתתי האוכלוסיות הוא אפס. לכן אין צורך לאמוד את חותך כללי, או מקדמים למשתני הדמי. האינטראקציות בין משתני הדמי למשתנה המסביר משמשות להבחנה בין השיפועים במודלים השונים ולכן מקדמי האינטראקציות כן ייכנסו למודל, כלומר במודל יהיו ארבעה פרמטרים.

שאלה 2

א.

$$Cov\left(Y_i,\hat{Y}_i\right)=Cov\left(Y_i,\hat{\beta}_0+\hat{\beta}_1X_i\right)=Cov\left(Y_i,\overline{Y}+\hat{\beta}_1\left(X_i-\overline{X}\right)\right)=Cov\left(Y_i,\overline{Y}\right)+Cov\left(Y_i,\hat{\beta}_1\left(X_i-\overline{X}\right)\right)$$

$$=\frac{\sigma^2}{n}+\left(X_i-\overline{X}\right)Cov\left(Y_i,\hat{\beta}_1\right)=\frac{\sigma^2}{n}+\left(X_i-\overline{X}\right)Cov\left(Y_i,\frac{\sum_{j=1}^{n}(X_j-\overline{X})Y_j}{SXX}\right)=$$

$$\frac{\sigma^2}{n}+\frac{X_i-\overline{X}}{SXX}Cov(Y_i,(X_i-\overline{X})Y_i)=\frac{\sigma^2}{n}+\frac{\left(X_i-\overline{X}\right)^2}{SXX}\sigma^2=\sigma^2\left(\frac{1}{n}+\frac{\left(X_i-\overline{X}\right)^2}{SXX}\right)$$

ב. נכפיל את ערכי X בקבוע a. נסמן $\tilde{X}_i=\mathrm{a}X_i$ וכן נסמן את אומדי הרגרסיה המקוריים ב-$\hat{\beta}_0,\hat{\beta}_1$ כרגיל. כעת נרצה למזער את

$$\sum_{i=1}^{n}\left(Y_i-\tilde{\beta}_0-\tilde{\beta}_i\tilde{X}_i\right)^2=\sum_{i=1}^{n}\left(Y_i-\tilde{\beta}_0-\tilde{\beta}_i\mathrm{a}X_i\right)^2=\sum_{i=1}^{n}\left(Y_i-\tilde{\beta}_0-\left[\tilde{\beta}_i\mathrm{a}\right]X_i\right)^2$$

בעיה זו נפתרה כבר כאשר המקדמים הממזערים הם $\tilde{\beta}_0=\hat{\beta}_0$ ו-$\tilde{\beta}_1\mathrm{a}=\hat{\beta}_1$ כלומר $\tilde{\beta}_1=\frac{\hat{\beta}_1}{\mathrm{a}}$.

ה-MSE כמובן נותר ללא שינוי, כיון שהבעיה הכתובה כאן זהה למקורית.

ג. נכפיל את ערכי Y בקבוע a. נסמן $\tilde{Y}_i=\mathrm{a}Y_i$. כעת נרצה למזער את

$$\sum_{i=1}^{n}\left(\tilde{Y}_i-\tilde{\beta}_0-\tilde{\beta}_i\tilde{X}_i\right)^2=\sum_{i=1}^{n}\left(\mathrm{a}Y_i-\tilde{\beta}_0-\tilde{\beta}_iX_i\right)^2=\sum_{i=1}^{n}\left(\mathrm{a}\left[Y_i-\frac{\tilde{\beta}_0}{\mathrm{a}}-\frac{\tilde{\beta}_iX_i}{\mathrm{a}}\right]\right)^2=\mathrm{a}^2\sum_{i=1}^{n}\left(Y_i-\frac{\tilde{\beta}_0}{\mathrm{a}}-\frac{\tilde{\beta}_iX_i}{\mathrm{a}}\right)^2$$

בעיה זו נפתרה בבעיה המקורית כאשר $\frac{\tilde{\beta}_0}{a} = \hat{\beta}_0$ כלומר $\tilde{\beta}_0 = \hat{\beta}_0 a$ ובאותו האופן $\tilde{\beta}_1 = \hat{\beta}_1 a$. ערך ה MSE יוכפל ב a^2.

<u>שאלה 3</u>

א. המשתנה שיוצא מהמודל הוא זה עם ערך F חלקי הקטן ביותר.

לפי הנוסחה, $F_{partial} = \frac{SSR_{full} - SSR_{reduced}}{MSE_{full}}$ כאשר $SSR_{reduced}$ הוא ה-SSR של המודל בלי המשתנה שהוצאנו. כיון שעבור כל אחד מהמשתנים הערכים MSE_{full}, SSR_{full} הם אותם ערכים, ה-F חלקי יהיה קטן אם ה-SSR של המודל עם המשתנה שנשאר יותר גדול. אם ה-SSR יותר גדול אז גם ה-$R^2 = \frac{SSR}{SST}$ יותר גדול כיון שה-SST הוא לא תלוי מודל. מכיון ש-R^2 הוא ריבוע מקדם המתאם, הטענה הוכחה.

ב. מדד ה-VIF הוא $\frac{1}{1-R_j^2}$ מצד שני ראינו שהאיבר ה-j על האלכסון של $\left(X^T X\right)^{-1}$ הוא

$\frac{1}{\left(SX_j X_j\right)\left(1-R_j^2\right)}$ (כך למעשה בנינו את ה-VIF)

נחשב את סכום הריבועים של X_1. באלכסון של המטריצה $X^T X$ נמצא $\sum_{i=1}^{200} X_{i1}^2 = 15702$ ממנו צריך להפחית את ריבוע הסכום חלקי n: $\left(\sum_{i=1}^{200} X_{i1}\right)^2 / 200 = 1724^2 / 200$ ולכן $SX_j X_j = 841.12$

לכן, $\frac{1}{\left(1-R_1^2\right)} = 841.12 \frac{144}{100000} = 1.211$.

ג. $C_p = \frac{SSE_{reduced}}{MSE_{full}} - (n-2p) = \frac{SSE_{reduced}}{14.5625} - (200-4)$ כדי לחשב את ה-SSE של המודל החלקי, נשתמש ב- $\hat{\beta}_1^2 SXX = SSR$ כיון שנתון לנו צד שמאל: $SSR = 6.6^2 * 841.12 = 36639.2$ נרצה למצוא את SST.

כיון שהוא לא תלוי מודל הוא מתקבל ע"י הנתונים מהמודל המלא:

$$0.9688 = R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \Rightarrow 0.0322 = \frac{2861.72}{SST} \Rightarrow SST = 88873.3$$

ולכן ה-SSE של המודל החלקי הוא 88873.3-36639.2=52234.1

$$C_p == \frac{52234.1}{14.5265} - (200-4) = 3399 .$$

<u>בונוס</u>

ננתח את התצפיות הממורכזות (כלומר אחרי שהחסרנו את הממוצע) .

במודל עם חותך:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2} \text{ ו- } \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} = 0$$

במודל ללא חותך $\hat{\beta}_0 = 0$ בהגדרה.

סכום הפרשי הריבועים במודל ללא חותך הוא $\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_1 X_i\right)^2$ הנגזרת היא

$2\sum_{i=1}^{n} X_i\left(Y_i - \hat{\beta}_1 X_i\right)$ ולכן היא מתאפסת כאשר $\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}$.

מסקנה: המקדמים בשני המקרים זהים.