**BIOST 518**

**HW #6**

**Answers**

1a. Methods: We investigate the association between time to death and serum LDL. We evaluate the linearity relationship between log hazard function and LDL using a Cox proportional-hazards model with robust standard errors. As predictors we include LDL serum levels and 4 indicator variables for different interval segments of LDL levels (in mg/dl 70 ≤ LDL < 100, 100 ≤ LDL < 130, 130 ≤ LDL < 160 and 160 ≤ LDL < 400; therefore subjects with LDL below 70 are used as reference). The coefficients estimates of the indicator are used to test the null hypothesis that a linear association log hazard ratio and untransformed LDL are adequate to describe the data. A global Wald test for the indicator variables is employed as statistical evidence for at least one of the indicator variables coefficients to be different from zero; thus a nonlinear association. 95% confidence intervals and two-sided p values were computed using Wald statistics based on the Huber-White sandwich estimator for the standard error. In all tests the significance level is set to 5%.

Time to event on the original dataset is in days. We transformed it to years considering that each year has 365.25 days. LDL serum levels have been subtracted 1 mg/dl.

Results: There are 735 subjects in the dataset but 10 miss LDL serum data. Among the 725 subjects we recorded 3582.94 person-years of follow up and found 131 deaths as shown on table 1. The global Wald-test is 15.84 corresponding to a p-value of 0.0073 for 5 degrees of freedom. So we reject the null hypothesis of all coefficients of the model to be zero (in log hazard scale); so there is an association between time to death and serum LDL. The global Wald-test for LDL levels categorized has p-value > 5% we fail to reject the null hypothesis that all indicator variables (from the categorized LDL) have coefficients equal to zero (in log hazard scale); so a linear relationship is adequate based on this model.

Table 1 – Cox proportional-hazard regression estimates with LDL and LDL categorized as predictors

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Hazard-Ratio** | **95% CI** | **p-value** |
| Total records in the model | 725 |  |  |
| Person-years | 3582.94 |  |  |
| Total deaths | 131 |  |  |
|  |  |  |  |
| Untransformed LDL – 1(mg/dl) | 0.996 | 0.979 - 1.013 | 0.621 |
|  |  |  |  |
| LDL levels indicators |  |  |  |
| < 70 | 1.000 | - | - |
| 70 - 100 | 0.456 | 0.185 - 1.124 | 0.088 |
| 100 - 130 | 0.508 | 0.150 - 1.722 | 0.277 |
| 130 - 160 | 0.429 | 0.083 - 2.227 | 0.314 |
| 160 - 400 | 0.465 | 0.047 - 4.601 | 0.513 |
|  |  |  |  |
| Overall model Wald-test p-value |  |  | 0.0073 |
| Overall LDL categorized Wald-test p-value |  |  | 0.3609 |

1b. Methods: We investigate the association between time to death and serum LDL. We evaluate the linearity relationship between log hazard function and LDL using a Cox proportional-hazards model with robust standard errors. As predictors we include LDL serum levels and squared LDL serum levels (quadratic term). The coefficient estimate of the quadratic term is used to test the null hypothesis that a linear association log hazard ratio and untransformed LDL are adequate to describe the data. A Wald test for the quadratic term is employed as statistical evidence for coefficient to be different from zero; thus a nonlinear association. 95% confidence intervals and two-sided p values were computed using Wald statistics based on the Huber-White sandwich estimator for the standard error. In all tests the significance level is set to 5%.

Time to event on the original dataset is in days. We transformed it to years considering that each year has 365.25 days. LDL serum levels have been subtracted 1 mg/dl.

Results: There are 735 subjects in the dataset but 10 miss LDL serum data. Among the 725 subjects we recorded 3582.94 person-years of follow up and found 131 deaths as shown on table 2. The global Wald-test is 15.28 corresponding to a p-value of 0.0005 for 2 degrees of freedom. So we reject the null hypothesis of all coefficients of the model to be zero (in log hazard scale); so there is an association between time to death and serum LDL. The Wald-test for LDL squared term has p-value > 5%. So we fail to reject the null hypothesis that the quadratic term has coefficient equal to zero (in log hazard scale); so a linear relationship is adequate based on this model.

Table 2 – Cox proportional-hazard regression estimates with LDL and squared LDL as predictors

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Hazard-Ratio** | **95% CI** | **p-value** |
| Total records in the model | 725 |  |  |
| Person-years | 3582.94 |  |  |
| Total deaths | 131 |  |  |
|  |  |  |  |
| Untransformed LDL – 1 (mg/dl) | 0.974 | 0.956 - 0.993 | 0.007 |
| Squared LDL (mg/dl) | 1.000 | 1.000 - 1.000 | 0.055 |
|  |  |  |  |
| Overall model Wald-test p-value |  |  | 0.0005 |
| Overall squared LDL Wald-test p-value | |  | 0.0550 |

1c. Methods: We investigate the association between time to death and serum LDL. We evaluate the linearity relationship between log hazard function and LDL using a Cox proportional-hazards model with robust standard errors. As predictors we include LDL serum levels, squared LDL serum levels and cubic LDL serum levels. The coefficients estimates of the squared and cubic term are used to test the null hypothesis that a linear association log hazard ratio and untransformed LDL are adequate to describe the data. A Wald test for the squared and cubic terms is employed as statistical evidence for coefficient to be different from zero; thus a nonlinear association. 95% confidence intervals and two-sided p values were computed using Wald statistics based on the Huber-White sandwich estimator for the standard error. In all tests the significance level is set to 5%.

Time to event on the original dataset is in days. We transformed it to years considering that each year has 365.25 days. LDL serum levels have been subtracted 1 mg/dl.

Results: There are 735 subjects in the dataset but 10 miss LDL serum data. Among the 725 subjects we recorded 3582.94 person-years of follow up and found 131 deaths as shown on table 3. The global Wald-test is 28.89 corresponding to a p-value < 0.0001 for 3 degrees of freedom. So we reject the null hypothesis of all coefficients of the model to be zero (in log hazard scale); so there is an association between time to death and serum LDL. The Wald-test for LDL squared and cubic terms has p-value 0.0164 (< 5%). So we reject the null hypothesis that the squared and cubic terms have coefficients equal to zero (in log hazard scale); so a non linear relationship is present between log hazards and LDL serum levels.

Table 3 – Cox proportional-hazard regression estimates with LDL, squared LDL and cubic LDL as predictors

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Hazard-Ratio** | **95% CI** | **p-value** |
| Total records in the model | 725 |  |  |
| Person-years | 3582.94 |  |  |
| Total deaths | 131 |  |  |
|  |  |  |  |
| Untransformed LDL – 1 (mg/dl) | 0.959 | 0.925 - 0.995 | 0.024 |
| Squared LDL (mg/dl) | 1.000 | 1.000 - 1.001 | 0.248 |
| Cubic LDL (mg/dl) | 1.000 | 1.000 - 1.000 | 0.497 |
|  |  |  |  |
| Overall model Wald-test p-value |  |  | 0.0005 |
| Squared and Cubic LDL Wald-test p-value |  |  | 0.0164 |

1d. Methods: We investigate the association between time to death and serum LDL. We evaluate the linearity relationship between log hazard function and LDL using a Cox proportional-hazards model with robust standard errors. As predictors we include 5 linear splines (representing interval segments of LDL defined in mg/dl at 0 ≤ LDL < 70, 70 ≤ LDL < 100, 100 ≤ LDL < 130, 130 ≤ LDL < 160 and 160 ≤ LDL < 400) of LDL serum levels. The coefficients estimate of the spline terms are used to test the null hypothesis that a linear association log hazard ratio and untransformed LDL are adequate to describe the data. A global Wald test for the term is employed as statistical evidence for at least one of the coefficients to be different from zero; thus there is association between the log hazard and LDL serum levels. As evidence for curvilinear relationship we use another Wald-test for the null hypothesis of similar coefficients on each segments defined by the knots of spline. 95% confidence intervals and two-sided p values were computed using Wald statistics based on the Huber-White sandwich estimator for the standard error. In all tests the significance level is set to 5%.

Time to event on the original dataset is in days. We transformed it to years considering that each year has 365.25 days. LDL serum levels have been subtracted 1 mg/dl.

Table 4 – Cox proportional-hazard regression estimates with LDL and linear splines of LDL as predictors

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Hazard-Ratio** | **95% CI** | **p-value** |
| Total records in the model | 725 |  |  |
| Person-years | 3582.94 |  |  |
| Total deaths | 131 |  |  |
|  |  |  |  |
| Splines of LDL (mg/dl) |  |  |  |
| < 70 | 0.978 | 0.960 - 0.996 | 0.019 |
| 70 - 100 | 0.979 | 0.953 - 1.006 | 0.131 |
| 100 - 130 | 0.999 | 0.978 - 1.021 | 0.934 |
| 130 - 160 | 0.998 | 0.974 - 1.022 | 0.875 |
| 160 - 400 | 0.994 | 0.966 - 1.023 | 0.678 |
|  |  |  |  |
| Overall model Wald-test p-value |  |  | p < 0.0001 |

Results: There are 735 subjects in the dataset but 10 miss LDL serum data. Among the 725 subjects we recorded 3582.94 person-years of follow up and found 131 deaths as shown on table 4. The global Wald-test is 30.15 corresponding to a p-value < 0.0001 for 5 degrees of freedom. So we reject the null hypothesis of all coefficients of the model to be zero (in log hazard scale); so there is an association between time to death and serum LDL. The Wald-test for similar coefficients in each segment defined by splines knots has p-value of 0.1172 (> 5%). So we fail to reject the null hypothesis that all LDL spline terms have the same coeficient; so a linear relationship is adequate based on this model.

1e. Methods: We investigate the association between time to death and serum LDL. We evaluate the linearity relationship between log hazard function and LDL using a Cox proportional-hazards model with robust standard errors. As predictors we include LDL serum levels and natural logarithm of LDL. The coefficient estimate of the logarithm is used to test the null hypothesis that a linear association log hazard ratio and untransformed LDL are adequate to describe the data. A Wald test for the logarithm term is employed as statistical evidence for coefficient to be different from zero; thus a nonlinear association. 95% confidence intervals and two-sided p values were computed using Wald statistics based on the Huber-White sandwich estimator for the standard error. In all tests the significance level is set to 5%.

Time to event on the original dataset is in days. We transformed it to years considering that each year has 365.25 days.

Results: There are 735 subjects in the dataset but 10 miss LDL serum data. Among the 725 subjects we recorded 3582.94 person-years of follow up and found 131 deaths as shown on table 5. The global Wald-test is 33.15 corresponding to a p-value < 0.0001 for 2 degrees of freedom. So we reject the null hypothesis of all coefficients of the model to be zero (in log hazard scale); so there is an association between time to death and serum LDL. The Wald-test for LDL logarithm term has p-value < 5%; so we reject the null hypothesis that the logarithm term has coefficient equal to zero (in log hazard scale); so there is a non-linear relationship is according to this model.

Table 5 – Cox proportional-hazard regression estimates with LDL and natural logarithm of LDL as predictors

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Hazard-Ratio** | **95% CI** | **p-value** |
| Total records in the model | 725 |  |  |
| Person-years | 3582.94 |  |  |
| Total deaths | 131 |  |  |
|  |  |  |  |
| Untransformed LDL (mg/dl) | 1.002 | 0.992 - 1.012 | 0.650 |
| Natural logarithm of LDL (mg/dl) | 0.360 | 0.181 - 0.716 | 0.004 |
|  |  |  |  |
| Overall model Wald-test p-value |  |  | p < 0.0001 |
| Natural logarithm LDL Wald-test p-value |  |  | 0.0036 |

1f) Methods: Scatter plots are used to display the fitted values from each model on exercise 1a, 1b, 1c, 1d and 1e against the LDL levels. Also the scatter from the fitted values from a Cox proportional hazards model with centred LDL is included. The Y-axis is shown in logarithm scale given the multiplicative nature of the models discussed. The specific methods are described above on respective section for the model. For note, all models have LDL centered at 1mg/dl in order to have the same reference on the graphic comparison we intent.

Results: The fitted values are displayed on figure 1.

All models agree in a descent relationship of the relative hazard with increasing levels of LDL; although the quadratic model show some descent pattern until 150-160 mg/ldl (of LDL) and then show some increase.

q1f.tif

Figure 1 – Fitted values for LDL centered at 1mg/dl for models discusses on 1a, 1b, 1c, 1d and 1e. Vertical blue lines corresponds to 69, 99, 129 and 159mg/dl (matching the respective spline after centering LDL).

The models with quadratic and the splines are the closest to each other compared to any other pair of models.

The logarithm model is consistently too far below other models. It is for note that in the log scale of the relative hazard the logarithm, cubic and quadratic are nearly parallel between 1 and 150 mg/dl centered LDL.

The linear model has fitted values within maximum and minimum of the step model.

The splines model suggest a great decrease of mortality associated with increase of LDL below 100 mg/dl range; almost null decrease between 70 and 160mg/dl. Then another decrease (smaller compared to the LDL below 70 mg/dl) above 160 mg/dl per increase of LDL.

For all this I would choose the logarithm transformed based model (1e) for few parameters to estimate, a smooth curve and biologically plausible.

2.a) First, let’s establish are minimum concept of a spline LDL. We have the LDL splines divided as 0 ≤ LDL < 70, 70 ≤ LDL < 100, 100 ≤ LDL < 130, 130 ≤ LDL < 160 and 160 ≤ LDL < 400. In each interval the value of spline is how much higher the LDL is above the minimum of the interval and that maximum should not exceed the maximum difference (maximum – minimum) for the interval. If it exceeds should count on the next interval. For example someone with 110 mg/dl counts on the first spline (0 ≤ LDL < 70) has 70, on the second (70 ≤ LDL < 100) has 30, on the third (100 ≤ LDL < 130) has 10 and on the rest following splines has 0. The sums of the splines should be equal to the original variable.

To compute the effect of association on each interval covered by the spline we multiply the linear centered coefficient with the respective coefficient of spline. With the right software or utility such as lincom (on Stata) we can obtain also the 95% confidence interval and p-value of this multiplication. For easier interpretation here we subtract 1 to hazard ratio and call it relative hazard and write as percentage.

With that said:

* Among subjects with LDL below 70mg/dl, 1 mg/dl increase of LDL is associated with 2.18% relative decrease on the hazard of death; it wouldn’t be a surprise if this relative hazard were found between 0.35% and 3.97%. We reject the null hypothesis of this relative hazard not to be different from zero given the p-value is 0.019.
* Subjects with LDL between 70 and 100mg/dl have 2.07% relative decrease on the hazard of death associated with 1mg/dl increased LDL; it wouldn’t be a surprise if this relative hazard were found between -0.63% and 4.69%. We can’t reject the null hypothesis of this relative hazard to be different from zero (p-value = 0.131).
* Subjects with LDL between 100 and 130mg/dl have 0.09% relative decrease on the hazard of death associated with 1mg/dl increased LDL; it wouldn’t be a surprise if this relative hazard were found between -2.08% and 2.22%. We can’t reject the null hypothesis of this relative hazard to be different from zero (p-value = 0.934).
* Subjects with LDL between 130 and 160mg/dl have 0.19% relative decrease on the hazard of death associated with 1mg/dl increased LDL; it wouldn’t be a surprise if this relative hazard were found between -2.25% and 2.58%. We can’t reject the null hypothesis of this relative hazard to be different from zero (p-value = 0.875).
* Subjects with LDL above 160mg/dl have 0.61% relative decrease on the hazard of death associated with 1mg/dl increased LDL; it wouldn’t be a surprise if this relative hazard were found between -2.31% and 3.45%. We can’t reject the null hypothesis of this relative hazard to be different from zero (p-value = 0.410).

2b) Based on the regression with spline predictors we couldn’t find evidence for a U-shaped relationship. This comes from the fact that slopes from each segment of LDL defined by the splines are descending (on the log hazard ratio scale they are all negative).

3. Methods (general for all questions): We investigate the association between time to death and serum LDL. We evaluate the relationship between log hazard function and LDL adjusted for race using a Cox proportional-hazards model with robust standard errors. As predictors we include natural log-transformed LDL serum levels, 3 indicator variables for race (for African Americans, Asians and other; White category is used as reference), interaction terms for each indicator variable for race multiplied by natural log-transformed LDL . We report the coefficients, the 95% confidence intervals and two-sided p values computed using Wald statistics based on the Huber-White sandwich estimator for the standard error. In all tests the significance level is set to 5%.

Time to event on the original dataset is in days. We transformed it to years considering that each year has 365.25 days.

3a) Results: There are 735 subjects in the dataset but 10 miss LDL serum data. Among the 725 subjects we recorded 3582.94 person-years of follow up and found 131 deaths as shown on table 6. Interpreting each coefficient:

* Among whites a 10% increase of LDL serum levels is associated with 7.11% decrease in relative hazard of death
* Among subjects with 1mg/dl LDL serum levels, African Americans compared to Whites have 0.1545 times higher hazard of death
* Among subjects with 1mg/dl LDL serum levels, Asians compared to Whites have 304.98 times higher hazard of death
* Among subjects with 1mg/dl LDL serum levels, people from other races compared to Whites have 333 million times higher hazard of death
* Compared to White subjects among African Americans subjects the ratio of instantaneous hazard-ratios is 1.0428 in groups differing in 10% higher of LDL, for LDL serum levels different from 1mg/dl.
* Compared to White subjects among Asians subjects the ratio of instantaneous hazard-ratios is 0.8944 in groups differing in 10% higher of LDL, for LDL serum levels different from 1mg/dl.
* Compared to White subjects among subjects of other races the ratio of instantaneous hazard-ratios is 0.6816 in groups differing in 10% higher of LDL, for LDL serum levels different from 1mg/dl.

Table 6 - Cox proportional-hazard regression estimates with natural logarithm of LDL, indicator variables for race, and interaction terms between natural logarithm of LDL and race as predictors

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Hazard-Ratio** | **95% CI** | **p-value** |
| Total records in the model | 725 |  |  |
| Person-years | 3582.94 |  |  |
| Total deaths | 131 |  |  |
|  |  |  |  |
| Natural log-transformed LDL | 0.461 | 0.300 - 0.709 | < 0.001 |
|  |  |  |  |
| Race |  |  |  |
| White | 1.000 | - | - |
| African Americans | 0.154 | 0.000 - 425.629 | 0.644 |
| Asian | 304.980 | 0.281 - 331470.200 | 0.109 |
| Other | 3.33 × 108 | 142.855 – 7.77 × 1014 | 0.009 |
|  |  |  |  |
| Interaction Race and natural log-transformed LDL |  |  |  |
| White | 1.000 | - | - |
| African Americans | 1.553 | 0.297 - 8.109 | 0.602 |
| Asian | 0.310 | 0.066 - 1.450 | 0.137 |
| Other | 0.018 | 0.001 - 0.500 | 0.018 |
|  |  |  |  |
| Overall model Wald-test p-value |  |  | p < 0.0001 |

3b) In addition to above described methods we use Wald-test to test the null hypothesis that all interaction terms have a null regression coefficient. We test at a significance of 5% and two-sided p-values are reported.

The Wald statistic is 8.04 for 3 degrees of freedom that corresponds to a p-value of 0.0452. This p-value is below 5% so we reject the null hypothesis. Therefore race modifies the association between time to death and serum LDL.

3c) In addition to above described methods we use Wald-test to test the null hypothesis that all interaction and log transformed terms have a null regression coefficient. We test at a significance of 5% and two-sided p-values are reported.

The Wald statistic is 26.84 for 4 degrees of freedom that corresponds to a p-value < 0.0001. We reject the null hypothesis. Therefore there is association between time to death and serum LDL.

3d) In addition to above described methods we use Wald-test to test the null hypothesis that all interaction and indicator variables for race have a null regression coefficient. We test at a significance of 5% and two-sided p-values are reported.

The Wald statistic is 42.40 for 6 degrees of freedom that corresponds to a p-value < 0.0001. We reject the null hypothesis. Therefore there is association between time to death and race.

3e) In addition to above described methods we use Wald-test to test the null hypothesis that all interaction Afro-American/log-transformed LDL and the indicator variable Afro-America race terms have a null regression coefficient. We test at a significance of 5% and two-sided p-values are reported.

The Wald statistic is 1.23 for 2 degrees of freedom that corresponds to a p-value 0.5416. We reject do not reject the null hypothesis. There is no difference in the distribution of time to death between whites and blacks.

4.a) Methods: We investigate the association between gender and mean salary increase. We use a linear regression model with salary as dependent variable and as predictors we include female dummy variable, calendar year as continuous and an interaction term between indicator female and year. We report the coefficients, the 95% confidence intervals and two-sided p values computed using Wald statistics based on robust standard-errors that allow for intragroup correlation. In all tests the significance level is set to 5%. To account for

We use a dataset based on a census of 1995 faculty members. We restrict the dataset to members who were hired within 1 year of their highest degree since 1990. For easy interpretation year of salary information have centered to start count since 1990.

Results: The original dataset had 1597 members, after applying the restriction above described only 118 members were hired since 1990 onwards within 1 year of their highest degree, of whom 62 (52.5%) were female. Table 7 shows the linear regression coefficients and figure 2 shows the scatter plot of salary and the fitted values of the regression.

Table 7 – Linear regression with untransformed salary and female dummy variable, year hired and interaction between female dummy and year hired.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Coefficient** | **95% CI** | **p-value** |
| Total records in the model | 456 |  |  |
|  |  |  |  |
| Female | -602.7 | -1,209.1 - 3.8 | 0.051 |
| Year | 169.2 | 77.3 - 261.1 | < 0.001 |
| Female-Year interaction | 11.8 | -98.4 - 122.0 | 0.833 |
| Intercept | 4422.9 | 3,936.3 - 4,909.6 | < 0.001 |
|  |  |  |  |
| Overall model F-test p-value |  |  | p < 0.0001 |

linreg_qa.tif

Figure 2 – Fitted values of the linear regression without untransformed salary

In 1990 males had a salary of 4422.9 USD in average whereas woman had 602.7 USD less (p-value = 0.051, no statistically significant difference in initial salaries). Yearly males had a mean increase of salary of 169.2 USD whereas women had 181.0 USD increase. The difference on salary yearly increase is not statistically different between males and females (p-value = 0.833); so no sex bias involved for increasing the salaries. We see graphically almost parallel fitted lines of males and females.

4.b) Methods: We investigate the association between gender and mean salary increase. We use a linear regression model with natural log-transformed salary as dependent variable and as predictors we include female dummy variable, calendar year as continuous and an interaction term between indicator female and year. We report the coefficients, the 95% confidence intervals and two-sided p values computed using Wald statistics based on robust standard-errors that allow for intragroup correlation. In all tests the significance level is set to 5%. To account for

We use a dataset based on a census of 1995 faculty members. We restrict the dataset to members who were hired within 1 year of their highest degree since 1990. For easy interpretation year of salary information have centered to start count since 1990.

Results: The original dataset had 1597 members, after applying the restriction above described only 118 members were hired since 1990 onwards within 1 year of their highest degree, of whom 62 (52.5%) were female. Table 8 shows the linear regression coefficients and figure 3 shows the scatter plot of salary and the fitted values of the regression.

Table 8 – Linear regression with natural log-transformed salary and female dummy variable, year hired and interaction between female dummy and year hired (coefficients are exponentiated).

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Geometric mean Ratio** | **95% CI** | **p-value** |
| Total records in the model | 456 |  |  |
|  |  |  |  |
| Female | 0.865 | 0.758 - 0.988 | 0.032 |
| Year | 1.037 | 1.017 - 1.057 | < 0.001 |
| Female-Year interaction | 1.007 | 0.983 - 1.031 | 0.576 |
| Intercept | 4301.374 | 3,870.904 - 4,779.715 | < 0.001 |
|  |  |  |  |
| Overall model F-test p-value |  |  | p < 0.0001 |

linreg_qb.tif

Figure 2 – Fitted values of the linear regression with natural log-transformed salary

In 1990 males had a geometric mean salary of 4301.4 USD whereas woman had 86.50% of that salary (p-value = 0.032, a statistically significant difference in initial salaries). Yearly males had a mean a relative increase of salary of 3.66% whereas women had 4.37 relative increase. The difference of year salary relative increase is not statistically different between males and females (p-value = 0.576); so no sex bias involved for increasing the salaries. We see graphically almost parallel fitted lines of males and females.

4.c) On both methods we have accounted for correlated observations. Method on a gives an idea of absolute increase on salaries so easily we see the magnitude of gap between salaries in dollar amount that could be easily used to make changes in polices and compensation packages. Whereas method on b is about ratio of geometric means (a relative increase).

We know that salaries increase based on the amount of the previous year so I would prefer method b.

4d) Methods as same on 4a and 4b except for standard errors. Here we use Huber-White sandwich estimator. Table 9 shows the same regression as 4a (untransformed salary as outcome) and table 10 shows the same as 4b (natural log-transformed salary outcome). On tables 11 and 12 we compare the standard errors of both approaches for standard errors for transformed and untransformed salary outcome.

Table 9 - Linear regression with untransformed salary and female dummy variable, year hired and interaction between female dummy and year hired.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Hazard-Ratio** | **95% CI** | **p-value** |
| Total records in the model | 456 |  |  |
|  |  |  |  |
| Female | -602.7 | -1,094.5 - -110.8 | 0.016 |
| Year | 169.2 | 61.2 - 277.2 | 0.002 |
| Female-Year interaction | 11.8 | -127.9 - 151.5 | 0.868 |
| Intercept | 4422.9 | 4,037.9 - 4,807.9 | < 0.001 |
|  |  |  |  |
| Overall model F-test p-value |  |  | p < 0.0001 |

Table 10 – Linear regression with natural log-transformed salary and female dummy variable, year hired and interaction between female dummy and year hired (coefficients are exponentiated).

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Geometric mean Ratio** | **95% CI** | **p-value** |
| Total records in the model | 456 |  |  |
|  |  |  |  |
| Female | 0.865 | 0.782 - 0.957 | 0.005 |
| Year | 1.037 | 1.015 - 1.059 | 0.001 |
| Female-Year interaction | 1.007 | 0.979 - 1.035 | 0.633 |
| Intercept | 4301.374 | 3,974.224 - 4,655.455 | < 0.001 |
|  |  |  |  |
| Overall model F-test p-value |  |  | p < 0.0001 |

Table 11 – Untransformed salary model coefficients and standard errors (IGC accounting for intra group correlation; HW using Huber-White estimator)

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **SE** | |
| **Variable** | **Coefficient** | **IGC** | **HW** |
|  |  |  |  |
| Female | -602.7 | 306.228 | 250.272 |
| Year | 169.2 | 46.393 | 54.956 |
| Female-Year interaction | 11.8 | 55.650 | 71.091 |
| Intercept | 4422.9 | 245.734 | 195.910 |
|  |  |  |  |

Table 12 – Natural log-transformed salary model coefficients and standard errors (IGC accounting for intra group correlation; HW using Huber-White estimator)

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **SE** | |
| **Variable** | **Coefficient** | **IGC** | **HW** |
|  |  |  |  |
| Female | 0.865 | 0.058 | 0.044 |
| Year | 1.037 | 0.010 | 0.011 |
| Female-Year interaction | 1.007 | 0.012 | 0.014 |
| Intercept | 4301.374 | 229.021 | 173.141 |
|  |  |  |  |

We would reach different conclusions on the salary untransformed outcome for initial salaries given that the standard errors for these two coefficient decreases substantially so Huber-White would not be conservative for initial salaries. For year changes Huber-White estimator inflated the standard errors (we would be more conservative).

For the natural log-transformed salary model we have the same pattern but the changes are very small for terms including year (more conservative). The intercept and main female term standard-errors have important changes but would not change the conclusions (less conservative).

**Appendix**

**Stata code**

clear

set more off

cd /mnt/disco/data/cursos/washington/biostat518/hw6

infile ptid mridate age male race weight height packyrs yrsquit ///

alcoh physact chf chd stroke diabetes genhlth ldl alb crt plt ///

sbp aai fev dsst atrophy whgrd numinf volinf obstime death ///

using mri.txt

drop in 1

// Question 1

g logldl= log(ldl)

g cldl= ldl - 1

g cldlsqr= cldl^2

g cldlcub= cldl^3

egen ldlctg= cut(ldl), at(0,70,100,130,160,400) icode label

mkspline sldlA 70 sldlB 100 sldlC 130 sldlD 160 sldlE = ldl

// Use the centered

stset obstime, failure(death) scale(365.5)

stdes

log using q1.log, replace

// q1\_a

stcox cldl i.ldlctg, robust nolog

testparm i.ldlctg

predict f\_q1a

// q1\_b

stcox cldl cldlsqr, robust

testparm cldlsqr

predict f\_q1b

// q1\_c

stcox cldl cldlsqr cldlcub, robust

testparm cldlsqr cldlcub // Revise your HW

predict f\_q1c

// q1\_d

stcox sldlA sldlB sldlC sldlD sldlE, robust

// stcox sldlA sldlB sldlC sldlD sldlE, robust

// testparm sldl\*

test sldlA=sldlB=sldlC=sldlD

predict f\_q1d

// q1\_e

stcox cldl logldl, robust

testparm logldl

predict f\_q1e

// q1\_f

stcox cldl, robust

predict f\_q1f

log close

scatter f\_q1a f\_q1b f\_q1c f\_q1d f\_q1e f\_q1f cldl, ///

legend(order(1 "Step (1a)" ///

2 "Quadratic (1b)" ///

3 "Cubic (1c)" ///

4 "Splines (1d)" ///

5 "Logarithmic (1e)" ///

6 "Continuous LDL" ///

) ///

rows(3)) ///

xtitle(Centered LDL (mg/dl)) ///

ytitle(Relative Hazard) ///

ylabel(0.05 0.1 0.2(0.2) 1.0, angle(horizontal) format(%3.2f)) ///

yscale(log) ///

xlabel(0 10 25(25) 250) ///

xline(69 99 129 159, lcolor(ltblue))

// xaxis(1 2) xline(69 99 129 159, lcolor(ltblue) axis(2))

// Talk about smoothness. Polynomial ones are smooth

// Linear splines are more similar to LOWESS (lowess uses weighting)

// The knots have been chosen scientfically

graph export q1f.tif, replace

// Question 2

// 2a - splines

stcox sldlA sldlB sldlC sldlD sldlE, robust nolog

// 2b No statistical evidence

/////////////////////////////////////////////////////////////////////////////////

// Question 3a

lab define race 1 white 2 black 3 asian 4 other

lab val race race

stcox i.race##c.logldl, robust nolog

// .di log(1.1)

// .09531018

//

// . di log(2)

// .69314718

lincom .09531018\*logldl, eform // For 10% increase in LDL

lincom .69314718\*logldl, eform // For 2 fold increase in LDL

lincom .09531018\*2.race#c.logldl, eform

lincom .09531018\*3.race#c.logldl, eform

lincom .09531018\*4.race#c.logldl, eform

// 3b

testparm i.race#c.logldl

// 3c

testparm i.race#c.logldl logldl

// 3d

testparm i.race#c.logldl i.race

// 3e

testparm 2.race#c.logldl 2.race

////////////////////////////////////////////////////////////////////////////////

//

clear

use salary\_original.dta

// Tag records to be kept

gen keep = startyr >= 90 & abs(yrdeg - startyr) <= 1

// egen keep = cond(startyr >= 90 & abs(yrdeg - startyr) <= 1, 1, 0), by(id)

// Keep the records that were tagged

keep if keep == 1

// What we do with id == 4?

sort id year

// Miss some years

gen gap = year - year[\_n-1] != 1 if id == id[\_n-1] // No jumping years

sort id year

by id: gen nrecs = \_N

unique id

// 4a

// Independent clusters

gen female = sex == "F"

gen logsal = log(salary)

gen yr\_cent = year - 90

regress salary female yr\_cent i.female#c.yr\_cent, cluster(id) robust

lincom yr\_cent + 1.female#c.yr\_cent // Mean salary per year increase among females

lincom yr\_cent + 0\*1.female#c.yr\_cent // Mean salary per year increase among males

predict fit\_meansal\_a

twoway (scatter salary year if female==0, jitter(1)) ///

(scatter salary year if female==1, jitter(1)) ///

(line fit\_meansal\_a year if female == 0) ///

(line fit\_meansal\_a year if female == 1) ///

, ylab(0 2000(2000) 10000, angle(horizontal)) ///

ytitle(Salary (USD)) ///

xtitle(Calendar year) ///

legend(order(1 "Male" 2 "Female" 3 "Fitted (Male)" 4 "Fitted (Female)"))

graph export linreg\_qa.tif, replace

// 4b

regress logsal female yr\_cent i.female#c.yr\_cent, cluster(id) robust eform(GR)

lincom \_cons + female, eform

lincom yr\_cent + 1.female#c.yr\_cent, eform // Mean salary per year increase among females

lincom yr\_cent + 0\*1.female#c.yr\_cent, eform // Mean salary per year increase among males

predict fit\_geomesal\_a

replace fit\_geomesal\_a = exp(fit\_geomesal\_a)

twoway (scatter salary year if female==0, jitter(1)) ///

(scatter salary year if female==1, jitter(1)) ///

(line fit\_geomesal\_a year if female == 0) ///

(line fit\_geomesal\_a year if female == 1) ///

, ylab(0 2000(2000) 10000, angle(horizontal)) ///

ytitle(Salary (USD)) ///

xtitle(Calendar year) ///

legend(order(1 "Male" 2 "Female" 3 "Fitted (Male)" 4 "Fitted (Female)"))

graph export linreg\_qb.tif, replace

// 4c

// This is pretending that each record is independent

regress salary female yr\_cent i.female#c.yr\_cent, robust

regress logsal female yr\_cent i.female#c.yr\_cent, robust eform(GR)