Survival Rate of Breast Cancer: A Causal Analysis

Anna Sofie Christensen (202107978)
Carrie Susanne Lovejoy (202106470)
Sofie Schubert Elving (202105513)

December 2022

1. Introduction

According to the American Cancer Society, breast cancer is the most common cancer in women in the United States, except skin cancers [1]. They estimate that the average risk of a woman in the United States developing breast cancer sometime in her life is 13%. Most breast cancers are found in women who are over 50 years old.

The goal of this paper is to determine the factors that have a causal effect on how long a breast cancer patient survives by using multivariate regression on a data set examining breast cancer patients in the United States. Our analysis uses methods for censoring and truncation to account for limits in our data set, and examines the functional form of certain variables, which will help us answer the following research question:

What factors have a causal effect on the survival time of breast cancer patients?

We found that several factors were significant, where particularly factors as race, grade, hormone status and number of regional positive lymph nodes seem to have the biggest effect on the survival time of breast cancer patients. However, we cannot be certain of causal effect due to uncertainties.

A quick search on Google Scholar with the search term 'breast cancer survival "seer database"' yields a number of reports and research articles that take their data from the same source as we do - the SEER Program of the NCI. The patient's age at diagnosis often comes up as an important factor in surviving breast cancer, although a 2005 report using an older version of the SEER database finds that tumor size, tumor grade, race, and year of diagnosis all have significant effects. [5].

2. The data

To answer our causal question we use a data set of breast cancer patients obtained from the 2017 November update of the SEER Program of the NCI [9]. The data set contains data from female patients diagnosed with infiltrating duct and lobular carcinoma breast cancer in 2006-2010. Rows with missing values have already been removed, along with patients whose survival months were less than 1 month. No further data cleaning is necessary. The data set therefore includes 4024 patients.

A few alterations have been made to the data set. This includes altering the values of our *Age* variable to be age after 30 (since all patients are over 30), changing binary variables to contain the values 0 or 1, and renaming variables.

Our variables after these changes can be seen in *Table 2.1*, where in case of dummy variables, the variable will be assigned the value 1 if true according to the variable name and 0 otherwise.

2.1. Variable description

Variable:	**Description:**
Age	Patient's age at diagnosis. Age is in years above 30 years of age.
Race	Patient's ethnicity. Either black, white or other.
MaritalStatus	Patient's Marital Status. Either Married, single, divorced, widowed or separated.
TStage	Describes 4 different stages, T1, T2, T3 and T4, of the tumor regarding size and location of the tumor.
NStage	Describes 3 different stages, N1, N2 and N3, based on number of found cancerous lymph nodes and their placement.
6thStage	Overall evaluation of cancer stage based on several variables; T stage, N stage, tumor grade, metasis and testing. Used classifications are IIA, IIB, IIIA, IIIB, IIIC.
Differentiate	How well differentiated (how normal looking) the cancer cells are. Normal looking cancer cells (well differentiated) grow slower than poorly differentiated cells.
Grade	Describes 4 stages of differentiation. Grade 1 = well differentiated, grade 2 = moderately diff., grade 3 = poorly diff., grade 4 = undiff.
AStage	Whether the cancer is regional or has extended distantly, for example to distant organs and/or distant lymph nodes.
TumorSize	Tumor size in millimeters.
EstroPos	Estrogen Positive is whether the cancer cells have estrogen receptors. If estrogen positive the cancer cells may receive signals from the hormone to grow. Patients who are hormone positive can receive hormone treatment.
ProgPos	Progesterone Positive is whether the cancer cells have progesterone receptors. If progesterone positive the cancer cells may receive signals from the hormone to grow. Patients who are hormone positive can receive hormone treatment.
RegNEx	Total number of examined and removed lymph nodes.
RegNPos	Total number of examined lymph nodes that are found to contain metastases.
SurvivalMonths	Number of months from cancer diagnosis till passing, or number of months from cancer diagnosis till most recent follow-up before study cut-off date.
Dead	Whether patient is dead as of study cut-off date.

TABLE 2.1: Description of variables contained in the used dataset.

3. Naive OLS Model

We initially regressed survival months on all available variables in R. However, many of the the staging variables, *TStage, NStage* and *6thStage*, are categorical variables defined by a combination of other variables in the data set such as *TumorSize, Grade, AStage* and *RegNPos*. As these staging variables cause high multicollinearity if included in the model, they are excluded. Moreover, we are more interested in the marginal effects of the individual biological marker variables than those of the staging variables.

This gives us the following OLS model:

$$\begin{aligned} SurvivalMonths_i = \beta_0 &+ \beta_1 Age_i + \beta_2 Race_i + \beta_3 MaritalStatus_i + \beta_4 Grade_i \\ &+ \beta_5 AStage_i + \beta_6 TumorSize_i + \beta_7 EstroPos_i + \beta_8 ProgPos_i \\ &+ \beta_9 RegNPos_i + \beta_{10} RegNEx_i + u_i \end{aligned} \tag{3.1}$$

In our OLS regression, it appears that most of our variables are statistically significant or close to it. though *Age*, *MaritalStatus*, *Grade* and *ProgPos* are not significant. *AStage* and *TumorSize* are also slightly insignificant. We will leave out *MaritalStatus* in later models, as it seems to be irrelevant and will continue to be quite insignificant, as well as other reasons (such as omitted variable bias) that will be brought up in the discussion.

We do however have some problems with our data set, where taking these problems into consideration could possibly leave *Grade* to be interesting.

4. Limited dependent variables

The first problem we have is that our data set is truncated from below as all patients who survived less than one month have been excluded. This means some more extreme observations are unaccounted for.

4.1. Truncation

Truncated regression follows the model [8, slide 20]

$$y = X\beta + u$$

However since all data points with survival months less than 1 have been excluded, we have a limitation on our data set from below, where all our points $y \geq c$, where c is 1 month.

From *Lecture 5* we know that to estimate β we need the distribution of y given X and that $y \geq c$.

Given this, we can then write the distribution of y given that $y \geq c$ and X as:

$$\frac{f_{X\beta,\sigma^2}(y)}{1 - F_{X\beta,\sigma^2}(c)}$$

To maximize the log likelihood function of this distribution to obtain our β coefficients we use the *truncreg*-function in R on the model in (3.1).

After accounting for the truncation, we get approximately the same results as in our naive OLS model, meaning the truncation of our data set does not appear to have a big effect on our estimation of parameters for *SurvivalMonths*.

Our data set suffers from another limited dependent variable problem; our data set is heavily censored (approximately 85%). The variable *SurvivalMonths* denotes survival months from diagnosis until most recent check-up before the study cut-off date, meaning we do not know the actual survival time of patients who survived past the study cut-off date.

4.2. Censoring

Censored regression follows the model [8, slide 6]:

$$y^* = X\beta + \epsilon$$

where we assume $\epsilon \sim N(0, \sigma^2)$ (visual inspection shown in A.3) and where we do not observe y^* but instead observe $y_i = min(y_i^*, c_i)$ as we have censoring from above in our data set instead of censoring from below, because we have no information on our data points' death after our study cut-off. y^* is the actual number of survival months after diagnosis of breast cancer and our observation y_i is the minimum of when a patient potentially dies or when the observation of the individual is stopped. This observation period varies for each individual as it is observed as the time between diagnosis and most recent check-up. Given this, we will observe different c_i's for each data point.

Given that we have different study times (c_i's) for each data point we must assume that each c_i is independent of the time until death (y_i^*) and therefore also independent of each error term ϵ_i.[8, slide 6]

As described in [8, slide 10], the log-likelihood for a censored regression model with below censoring is [8]:

$$\ln L(\beta, \sigma; y_i) = \sum_{i=1}^{n} I(y_i = c) \ln\left(\Phi\left(\frac{c - X_i\beta}{\sigma}\right)\right) + I(y_i > c) \ln\left(\frac{1}{\sigma}\phi\left(\frac{y_i - X_i\beta}{\sigma}\right)\right)$$

where $I(\cdot)$ is the indicator function, $\phi(\cdot)$ is the standard normal pdf and $\Phi(\cdot)$ is the standard normal CDF.

However, this is the model for below censoring and a single c-value for all observations, which means we need to rewrite this for above censoring and individual c_i's. To do this we start with the given densities of y_i given X_i and c_i given by *Wooldridge [10, eq. 17.38 & 17.39]* that fits our exact censoring situation:

$$f(y|x_i, c_i) = \begin{cases} 1 - \Phi\left(\frac{c_i - X_i\beta}{\sigma}\right), & y = c_i \\ \frac{1}{\sigma}\phi\left(\frac{y - X_i\beta}{\sigma}\right), & y < c_i \end{cases} \tag{4.1}$$

We can then compute the likelihood (defined as $L(\beta) = \prod_{i=1}^{n} f(y_i, \beta)$) and log-likelihood function for individual c_i's and above censoring using (4.1):

$$L(\beta, \sigma; y_i) = \prod_{i=1}^{n} \left(1 - \Phi\left(\frac{c_i - X_i\beta}{\sigma}\right)\right)^{I(y_i = c_i)} \cdot \left(\frac{1}{\sigma}\phi\left(\frac{y_i - X_i\beta}{\sigma}\right)\right)^{I(y_i < c_i)}$$

$$\ln L(\beta, \sigma; y_i) = \sum_{i=1}^{n} I(y_i = c_i) \ln\left(1 - \Phi\left(\frac{c_i - X_i\beta}{\sigma}\right)\right) + I(y_i < c_i) \ln\left(\frac{1}{\sigma}\phi\left(\frac{y_i - X_i\beta}{\sigma}\right)\right)$$

We see that this function is quite similar to the function from given for censoring in [8, slide 10]

To maximise the log-likelihood function we can use the *survreg*-function for a Gaussian distribution in R.

We use the *survreg*-function on the model in *equation (3.1)*.

Running our censored regression model in R the regressors *Age*, $Grade_1$, $Grade_2$, *TumorSize* and *ProgPos* are all significant unlike in the naive OLS model. *MaritalStatus* is still quite insignificant and *AStage* is still slightly insignificant.

Our parameter values are generally larger than the ones in our OLS model, meaning most of our parameters now seem to have a bigger effect on the survival length.

5. Functional Form and Interaction Terms

Given the information we have on our variables, we have reason to suspect that some of our variables effects' might be dependent on each other, and we wish to inspect potential interaction between certain regressors.

5.1. Estrogen and Progesterone Interaction

ProgPos and *EstroPos* both affect cancer behavior and whether you can receive hormonal treatment or not, and so we suspect their interaction may have a combined effect on survival months. The added interaction term in our model (as seen in Equation 5.1) becomes $\beta_{11}(ProgPos \cdot EstroPos)$. *ProgPos* and *EstroPos* are dummy variables with the value 1 if positive and 0 if not. The supposed causal effect of *ProgPos* becomes:

$$\frac{\partial E[SurvivalMonths|X]}{\partial Progpos} = \beta_{10} + \beta_{11} EstroPos = \begin{cases} \beta_{10} + \beta_{11}, & \text{if } EstroPos = 1 \\ \beta_{10}, & \text{if } EstroPos = 0 \end{cases}$$

And likewise for *EstroPos*.

If we look at our model intercepts (for readability all other regressors than the *EstroPos, ProgPos* and their interaction have been omitted) we get the following values added to the intercept for the following combination of groups:

$$E[SurvivalMonths|ProgPos = 0, EstroPos = 1] = ... + \beta_7$$
$$E[SurvivalMonths|ProgPos = 1, EstroPos = 0] = ... + \beta_8$$
$$E[SurvivalMonths|ProgPos = 1, EstroPos = 1] = ... + \beta_8 + \beta_7 + \beta_9$$

Having either positive *ProgPos* or *EstroPos* allows the cancer patient to receive hormonal treatment, which, according to the National Cancer Institute, can stop or slow cancer's growth and reduce the chance it will return. With this knowledge, we expect our regression to show that these regressors have strongly positive coefficients, hereby extending a patient's survival time, or even eliminating that patient's cancer entirely (in other words, extending their survival time indefinitely as far as the study is concerned).

5.2. Regional Node Examined and Regional Node Positive Interaction

We also have *RegNEx* and *RegNPos*. Here, *RegNPos* is limited by *RegNEx*, as we cannot record more positive lymph nodes than we have removed and examined. This issue is illustrated by Figure A.1. To account for this, we have tried adding an interaction term $\beta_{11}(RegNPos \cdot RegNEx)$, whilst excluding $RegNEx$ from the model. Intuitively, we do not expect *RegNEx* to have a direct effect on survival

months, and we expect for this variable to mostly have an effect on *SurvivalMonths* through *RegNPos*. Adding the interaction term to the model, the potentially causal effect of *RegNPos* becomes:

$$\frac{\partial E[SurvivalMonths|X]}{\partial RegNPos} = \beta_{10} + \beta_{11} RegNEx$$

The interaction term is significant in our final model, with β_{10} having a negative effect on survival months, whilst β_{11} contributes positively. This means that whilst having many positive lymph nodes has a negative impact on survival months, the effect of this is lessened by having a large number of examined lymph nodes.

However, we are uncertain of how to appropriately account for the relationship between $RegNEx$ and $RegNPos$. If we simply include $RegNEx$ in our model, the regressor is significant and if we include both $RegNEx$, $RegNPos$ and their interaction all regressors are significant, but $RegNEx$ and their interaction is less significant.

5.3. Quadratic Age Effect

Age was not not significant in the initial OLS model, however, it was significant in our censored regression model. We still wish to examine if age could potentially have a quadratic effect on survival. A research article published in 2016 concludes that breast cancer is more aggressive in younger women compared to older women [4]. Moreover, available treatment and the effectiveness of treatment depends on menopause status of the individual. However, it would be unrealistic to expect age to solely have a positive linear effect on survival time as aging leads to a weakened immune system over time.

To include the quadratic effect on the model we add the variable Age^2 so that we get $SurvivalMonths_i = \beta_0 + \beta_1 Age_i + \beta_2 Age_i^2 + ... + u_i$. The causal effect of *Age* on *SurvivalMonths* would then be:

$$\frac{\partial E[SurvivalMonths|X]}{\partial Age} = \beta_1 + 2\beta_2 Age$$

Where X denotes the regressors used in our model (Equation 5.1).

When included in the censored model quadratically, age becomes significant with values of $\beta_1 = 2.66$ and $\beta_2 = -0.07$. The parameters indicate that women around 40 years of age have the longest survival period after cancer diagnosis.

Including these interaction terms leads us to our new regression model given by:

$$\begin{aligned} SurvivalMonths_i = \beta_0 &+ \beta_1 Age_i + \beta_2 Age_i^2 + \beta_3 Race_i + \beta_4 Grade_i + \beta_5 AStage_i \\ &+ \beta_6 TumorSize_i + \beta_7 EstroPos_i + \beta_8 ProgPos_i \\ &+ \beta_9 (EstroPos_i \cdot ProgPos_i) \\ &+ \beta_{10} RegNPos_i + \beta_{11} (RegNPos_i \cdot RegNEx_i) + u_i \end{aligned} \tag{5.1}$$

We run this model as both a naive OLS regression (for comparison purposes), truncated regression and censored regression separately.

Running the model 5.1 as OLS, we get that the only very insignificant parameter is *Grade* and all of our interaction terms are quite significant. The truncated regression continues to be very similar to the naive OLS.

For our censored regression, significance level changes for the interaction terms and the related variables. Particularly *ProgPos* becomes slightly insignificant and the interaction term between *ProgPos* and *EstroPos* becomes very insignificant. Because of this we decide to remove this interaction term and run the censored regression on the final model:

$$\begin{aligned} SurvivalMonths_i = \beta_0 &+ \beta_1 Age_i + \beta_2 Age_i^2 + \beta_3 Race_i + \beta_4 Grade_i + \beta_5 AStage_i \\ &+ \beta_6 TumorSize_i + \beta_7 EstroPos_i + \beta_8 ProgPos_i + \beta_9 RegNPos_i \\ &+ \beta_{10}(RegNPos_i \cdot RegNEx_i) + u_i \end{aligned} \tag{5.2}$$

6. Results

When interpreting the coefficients of our variables we consider the full model, where our β_j measures the partial effect of x_j on $E[y^*|X]$ (number of survival months til the patient is deceased), and not $E[y|X]$. If we wish to examine the partial effect on $E[y|X]$ can be computed by[10]:

$$\frac{\partial E[y|X]}{\partial x_j} = \beta_j \Phi\left(\frac{X\beta}{\sigma_j}\right)$$

Where $\Phi(X\beta/\sigma_j)$ is the adjustment factor.

However, for now we examine partial effect of x_j on $E[y^*|X]$. This choice will be discussed later.

Our results are shown in the table below.

Variable	Coefficient	p-value
Intercept	46.04	0.01802
Age	2.66	0.00027
Agesq	-0.07	$1.5 \cdot 10^{-6}$
$\text{Race}_{\text{Other}}$	31.40	$1.4 \cdot 10^{-5}$
$\text{Race}_{\text{White}}$	18.87	$9.6 \cdot 10^{-5}$
Grade_1	49.43	0.00186
Grade_2	34.03	0.02516
Grade_3	22.71	0.13513
$\text{AStage}_{\text{Regional}}$	13.66	0.07144
TumorSize	-0.22	0.00071
EstroPos	25.13	$3.6 \cdot 10^{-6}$
ProgPos	19.41	$8.0 \cdot 10^{-7}$
RegNPos	-5.14	$< 2 \cdot 10^{-16}$
RegNPos × RegNEx	0.09	$2.4 \cdot 10^{-6}$

Table 2: Full Model Results

We generally see that the only variables that are insignificant according to our model are $Grade_3$ and $AStage_{Regional}$. Both of these are, however, not very insignificant as they have p-values of 0.14 and 0.07 respectively which is not that far from our usual 0.05-significance level - particularly for $AStage_{Regional}$.

The variables with the biggest positive impact on the survival length would be *Race, Grade, EstroPos* and *ProgPos*.

The results show that being white or of other race increases your survival length with 18.87 and 31.40 months respectively compared to being black. A possible interpretation of this is included in the discussion.

Similarly, we see that an increase in *Grade* will reduce your survival length with between 11.32 and 22.71 months depending on the *Grade* step-down. Generally having $Grade_1$ increases your survival length with 49.43 months compared to having the most aggressive $Grade_4$.

Being *EstroPos* (estrogen receptor positive) and *ProgPos* (progesterone receptor positive) increases your survival length with 25.13 and 19.41 months respectively compared to those who are estrogen receptor negative and progesterone receptor negative.

We see that *Age* (with *Agesq*) follows the polynomial shown in Appendix A.2, where *Age* has a positive effect on *SurvivalMonths* from 0-10 years (30-40 years of age) and then has a negative effect on *SurvivalMonths* from 10-30 (40-60 years of age).

We also have a negative effect of both *TumorSize* and *RegNPos* on *SurvivalMonths*. This meaning that an increase of 1mm in *TumorSize* reduces your survival length with 0.22 months (about 1 week), whereas for each additional positive regional lymph node (*RegNPos*) it will reduce the survival length with 5.14 months.

The interaction term ($RegNPos \times RegNEx$) is also significant and we see that the effect of one additional RegNPos is given by: $effect = -5.15 + RegNEx \cdot 0.09$.

We generally see that our estimated coefficient values are as expected, when it comes to positive or negative effect.

7. Discussion

As much information as the data set gives us on each individual patient's condition, we found its lack of information to be problematic during our analysis. The dependent variable of our analysis is the time (in months) a patient survived from time of diagnosis during this study. Our data set only informs us of the patient's condition at the time of diagnosis. We do not know anything about the treatment(s) they were given, only the biological markers and observable factors such as age and race, so we assumed that all possible treatments were attempted independent of external factors such as affordability. It was not stated whether individuals were responsible for covering the costs of their own treatment, in which case we would have a major issue with omitted variable bias (OVB) due to income being omitted from the model - and race, while widely believed to be correlated, is *not* an adequate substitute!

Furthermore, since we do not know if the study involved giving patients equal access to treatment, the race variable may have more weight than it should if we are strictly considering it as a biological marker and not as an indicator of socioeconomic status or lifestyle patterns. we instead consider it as the latter, it points to a fundamental problem in how (North American) society treats its poorest individuals - problems which our analysis does not account for. [3]

Similarly, we have no way of knowing how to interpret marital status of a patient. Marital status on its own should intuitively mean very little, but it can be an *indicator* of potentially significant factors such as social support and an additional income (or the absence of it), and these can affect how the cancer develops - and this brings us to the issue that our data set merely contains "snapshots" of how each individual was faring at one particular point. We may have been able to form a better idea of the importance of marital status if we had been working with panel data. However, the fact is that according to our analysis, marital status has little statistical significance on the time a patient survives, although there are studies that suggest otherwise.

Additionally, the data set does not tell us whether any patients were ever declared "cancer-free". We only know whether a patient was dead or alive at the time of their latest checkup before the study

cut-off date. The data set fails to capture the full picture.

Accounting for truncation does not seem to have a significant effect on our estimated parameters from the naive OLS. We generally see that not many individuals die within a couple of months from they get diagnosed with breast cancer, meaning that we most likely haven't excluded (truncated) very many data points from the original data set who had died within a month of diagnosis and thus accounting for truncation will not make any significant changes to our naive OLS. We do not include the correction for truncation in our censored model, meaning our censored model may vary slightly, as this model is still truncated. Potentially, it might have a larger effect on our censored regression than uncensored, since the truncated values would be more extreme 'outliers' when we compare to the actual survival time of patients, and not the shortened observation time.

In our final Tobit (censored) model we choose to measure the partial effect of x_j on $E[y^*|X]$, meaning the number of survival months til the patient is deceased. It might not make much sense to examine the long term effect on survival time for decades after a breast cancer diagnosis, if patients have been declared cancer free. And so it can be of interest to examine the partial effect of x_j on the observed survival months $E[y|X]$. To find the partial effect on $E[y|X]$ we have computed the adjustment factors (the adjustemt factors can be found in appendix A.1). However, many of the adjustment factors are close to 1, and so the partial effect of the regressors on $E[y|X]$ compared to $E[y^*|X]$ are nearly identical. This might be due to the observation period (approximately from 2006 to follow up in 2017) is relatively long when it comes to breast cancer survival, as the probability of a patient dying from breast cancer after 10 years is relatively low[2]. It should be noted that the adjustment factors for $Race$, $TumorSize$ and $Grade_1$ is moderately adjusted.

Compared to our OLS model, the coefficient values in our censored regression are significantly larger, as expected. When accounting for censoring, we would expect individuals, who are alive as of study cut-off and have a less aggressive breast cancer diagnosis, to have their survival time prolonged. Therefore we would see bigger difference in survival time between individuals with severe cancer and milder cancer cases, and so the effect of our regressors would be magnified.

The Tobit model relies on normality and homoskedasticity, as shown in section 4.2, *Censoring*, to be unbiased and consistent. The assumptions of normality and homoskedasticity have been visually inspected in figure A.3 and A.4. The plots show that the residuals are seemingly normally distributed, however, figure A.4 shows that the absolute value of the residuals systematically increase, as the fitted values increase. This is probably due to the residuals being the difference between observed y and fitted y, where our fitted y's are unsurprisingly larger than the observed due to censoring. We have not been able to evaluate whether the homoskedasticity assumption holds, however, for only moderate departures from the assumption, the Tobit model will still provide good estimates.[10]

Another approach to account for censoring could have been to drop censored individuals, i.e. removing individuals who are marked as alive after study cut-off. However, this means we would reduce our sample size by approximately 85% and move into the truncated model area, where using naive OLS would cause other issues.

During our analysis, we decided to exclude the interaction between $EstroPos$ and $ProgPos$, due to their insignificance. The variables *EstroPos* and *ProgPos* are moderately/highly correlated, and our data shows that most individuals are both estrogen and progesterone positive. Particularly it seems that individuals, who are progesterone positive, are almost always estrogen positive as well (see A.3). This leads to our interaction term $EstroPos \cdot ProgPos$ (a dummy variable) to be severely correlated with *ProgPos* (correlation of almost 0.98). Due to the strong correlation between the interaction term and $ProgPos$ the variance of our coefficient estimates will increase and our p-tests can be unreliable.[10] In the final model the interaction term has been dropped due to its uncertainty and

insignificance when we correct for censoring.

8. Conclusion

By accounting for limitations such as truncation and censoring in our data set, as well as examining the functional form of our variables, we find that our results of multivariate regressions fit those found in the 2005 report [5]. We see that age, race, tumor size, grade, hormone receptor status and number of positive lymph nodes are statistically significant in our regression. The most influential parameters are race, grade, positive hormone status and number of positive lymph nodes, adding several months to a cancer patients survival time, or in the case of positive lymph nodes, reduces survival time significantly.

However, our model suffers from omitted variable bias, as our data set lacks information about the patients socioeconomic status, i.e. factors such as income, access to treatment, general health, lifestyle etc, and so we see variables such as race become highly significant. We also have not been able to fully validate whether or not our model satisfies the necessary assumptions of homoskedasticity and normality of the error term, which could harm our parameter estimates if the assumptions are violated.

Overall, we see indications of causal effect, however, the size and significance is uncertain.

9. Bibliography

[1] American Cancer Society. Key statistics for breast cancer, 2022. URL https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html#references.

[2] American Society of Clinical Oncology. Breast cancer: Statistics. *Cancer.Net*, Jan. 2022. URL https://www.cancer.net/cancer-types/breast-cancer/statistics.

[3] BreastCancer.org. Race/Ethnicity, Dec. 2022. URL https://www.breastcancer.org/risk/risk-factors/race-ethnicity?fbclid=IwAR0lCx46Il6kQb0EM65ffYm6h4dL_hCh2hXqmCTatq3xzTQ3J_5XSSJgBf8. [Online; accessed 31. Dec. 2022].

[4] H.-l. Chen, M.-q. Zhou, W. Tian, K.-x. Meng, and H.-f. He. Effect of age on breast cancer patient prognoses: A population-based study using the seer 18 database. *PLOS ONE*, 11(10):1–11, 10 2016. doi: 10.1371/journal.pone.0165409. URL https://doi.org/10.1371/journal.pone.0165409.

[5] J. Rosenberg, Y. L. Chia, and S. Plevritis. The effect of age, race, tumor size, tumor grade, and disease stage on invasive ductal breast cancer survival in the U.S. SEER database. *Breast Cancer Res. Treat.*, 89(1):47–54, Jan. 2005. ISSN 1573-7217. doi: 10.1007/s10549-004-1470-1. URL https://doi.org/10.1007/s10549-004-1470-1.

[6] L. N. Taylor. *Lecture 2: Multivariate Regression.* Aarhus University, 2022.

[7] L. N. Taylor. *Lecture 3: Functional Form and Specification.* Aarhus University, 2022.

[8] L. N. Taylor. *Lecture 5: MLE - Limited Dependent Variables.* Aarhus University, 2022.

[9] J. Teng. Seer breast cancer data, 2019. URL https://dx.doi.org/10.21227/a9qy-ph35.

[10] J. M. Wooldridge. *Introductory Econometrics.* South-Western Cengage Learning, Boston, MA, USA, Sept. 2012.

Appendix

A. Figures

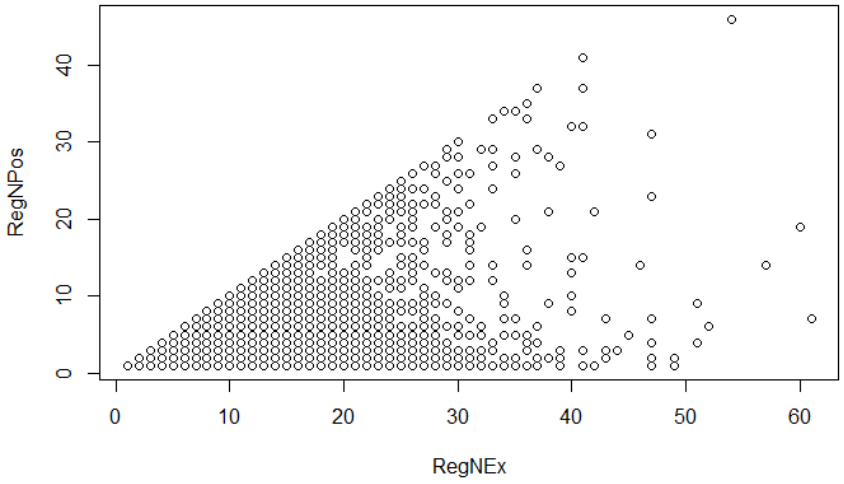

FIGURE A.1: Figure illustrating the relationship between the variables *RegNPos* and *RegNEx*

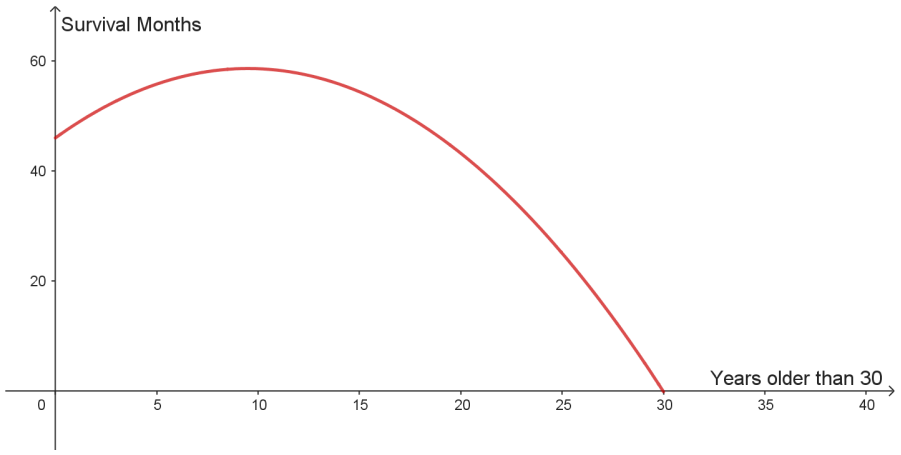

FIGURE A.2: Quadratic effect of age on survival months

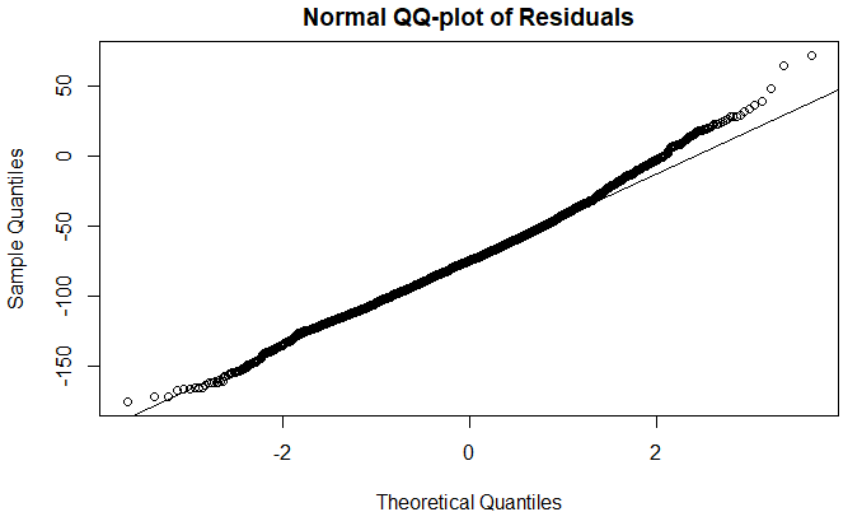

FIGURE A.3: QQ Plot

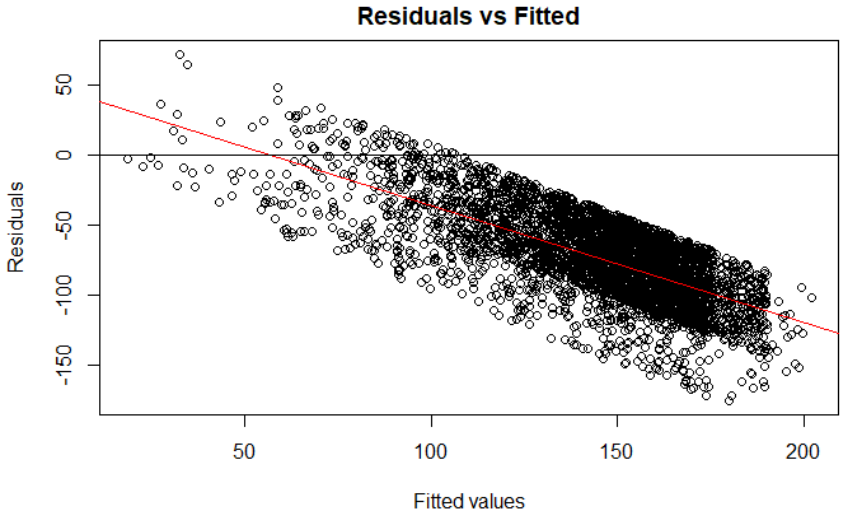

FIGURE A.4: Residual plot

A.1. Tables

Variable	Adjustment Factor
Intercept	1.00
Age	1.00
Agesq	1.00
$\text{Race}_{\text{Other}}$	0.70
$\text{Race}_{\text{White}}$	0.84
Grade_1	0.67
Grade_2	0.90
Grade_3	0.80
$\text{AStage}_{\text{Regional}}$	1.00
TumorSize	0.89
EstroPos	1.00
ProgPos	1.00
RegNPos	1.00
RegNPos × RegNEx	1.00

TABLE A.1: Adjustment Factors.

Variable	Coefficient	Adjusted Coefficient	p-value
Intercept	46.04	46.04	0.01802
Age	2.66	2.66	0.00027
Agesq	−0.07	−0.07	$1.5 \cdot 10^{-6}$
$\text{Race}_{\text{Other}}$	31.40	21.89	$1.4 \cdot 10^{-5}$
$\text{Race}_{\text{White}}$	18.87	15.91	$9.6 \cdot 10^{-5}$
Grade_1	49.43	33.10	0.00186
Grade_2	34.03	30.78	0.02516
Grade_3	22.71	18.08	0.13513
$\text{AStage}_{\text{Regional}}$	13.66	13.66	0.07144
TumorSize	−0.22	−0.20	0.00071
EstroPos	25.13	25.13	$3.6 \cdot 10^{-6}$
ProgPos	19.41	19.41	$8.0 \cdot 10^{-7}$
RegNPos	−5.14	−5.14	$< 2 \cdot 10^{-16}$
RegNPos × RegNEx	0.09	0.09	$2.4 \cdot 10^{-6}$

TABLE A.2: Full Model Result

	Estro positive	Estro negative
Prog positive	3299	27
Prog negative	456	242

TABLE A.3: Number of individuals who are progesterone and hormone positve

A.2. R code

See the attached RMarkdown file.