REGRESSION QUANTILE ANALYSIS OF CLAIM TERMINATION RATES FOR INCOME PROTECTION INSURANCE

Abstract

This paper investigates the use of censored regression quantiles in the analysis of claim termination rates for income protection (IP) insurance. The paper demonstrates the importance of modeling quantiles given the growing interest of regulators and others in stochastic approaches to valuation of insurance liabilities and risk margins.

1. Introduction

Actuarial interest in quantiles other than the median has increased considerably in recent years. Notable is the Australian Prudential Regulatory Authority (APRA) standard for the valuation of general insurance liabilities, GPS210, introduced as part of the Australian General Insurance Reform (2001). This standard requires that a risk margin should be established "on a basis that is intended to secure the insurance liabilities of the insurer at a given level of sufficiency – that level is 75 per cent". Previously the Australian General Insurance Act (1973) was considerably less prescriptive on the level of risk (or prudential) margin that insurers were required to hold.

Given that general insurance actuaries in Australia are now required to estimate a 75^{th} percentile of the distribution of outstanding claims for recording in profit and loss statements, it becomes important that the impact of potential risk factors on various quantiles of the distribution of outstanding claims provisions be considered. This is in addition to estimating the impact of risk factors on the mean of the outstanding claims provision.

In the context of disability income insurance claim termination rates, Pitt (2006) gives an illustration of how various insured characteristics impact claim termination rates differently at different quantiles of the distribution of claim duration. In that paper, the use of mixture models showed that the smoker status, which has not been included in previous Australian industry tables for claim termination, has a statistically significant impact on the probability that an IP insurance claim will continue indefinitely and that the claimant will never return to work. This is evidence that smoker status is a statistically significant predictor for claim termination rates (leading to a reduction in claim termination rates) for the very longest duration claims. In other words, a traditional regression which considers only the impact of rating factors on the mean would not find that smoker status is statistically significant, however closer examination of the impact of smoker status in the tail of the probability distribution of claim durations indicates that smoker status is critical for long duration claims.

The importance of understanding the impact of potential rating factors and the different impacts they have across the claim duration distribution is of particular importance in reserving and pricing IP insurance contracts. Failure to properly assess the impact of a rating factor in the tail of the probability distribution of claim durations will lead to serious underestimation of claim reserves in respect of disabled lives, particularly those lives who have been disabled for longer than, say, six months.

2. Regression Quantiles

One way of extending the linear model to allow for prediction of various quantiles of the distribution of the claim duration is the method of regression quantiles of Koenker and Bassett

(1978). This methodology has recently been extended to allow for standard right censoring and therefore can provide an alternative to the Cox Model or mixture models, see Portnoy (2003).

Traditional statistical and actuarial analysis has focused on sample averages as estimates of the population mean. Variability has generally been considered using sample standard deviations and the assumption of normality or, more recently, other parametric assumptions have been made. It has long been argued, (Galton, 1889), that any complete analysis of the "full variety of an experience requires the entire distribution of a trait, not just a measure of its central tendency." We therefore consider the use of regression quantiles as a method for identifying heterogeneity among subpopulations by considering the behaviour of the percentiles as a function of their associated probability τ.

For a random variable Y of measurements from some population, the population quantile is defined to be the value $Q_Y(\tau)$ satisfying

$$P\{Y \le Q_Y(\tau)\} = \tau \quad \text{for } 0 \le \tau \le 1. \tag{1}$$

Next, we describe the generalisation of this quantile to a regression context through the use of the conditional quantile. The conditional quantile, $Q_{Y|X}(\tau; x)$, is defined such that

$$P\{Y \le Q_{Y|X}(\tau; x) \mid X = x\} = \tau. \tag{2}$$

Whereas traditional regression analysis provides a single regression curve, for example the conditional mean function, in this regression quantile context we can let τ vary, and therefore consider a family of conditional quantile curves to provide a clearer picture of the dependencies present in the data.

To simplify the analysis, Koenker and Bassett (1978) suggest the estimation of conditional quantile curves under the assumption that, after appropriate transformations, they are linear in the covariates. This assumption has the advantage of allowing easier interpretation of coefficient estimates and also permits significantly faster computation. The estimation of the conditional quantile functions involves finding the solution to the problem of choosing ξ to minimise

$$R_T(\xi) = \sum_{i=1}^{n} \rho_T(Y_i - \xi), \tag{3}$$

where ρ_T is the piecewise linear "check" function,

$$\rho_T(u) = u(\tau - I(u < 0)) = \tau u^+ + (1-\tau)u^-, \tag{4}$$

and where u^+ and u^- are the positive and negative parts of u taken as positive values, respectively.

Portnoy (2003) next describes a general linear response model where $\{Y_i, \underline{x}_i\}$ denotes a sample of responses Y and explanatory variables $\underline{x}$ (in m dimensions), and suppose

$$Y_i = \underline{x}_i' \underline{\beta} + z_i, \quad i = 1, 2, ..., n, \tag{5}$$

where $\underline{\beta}$ is an m-dimensional parameter and z_i is the random error term. If we then minimise

$$R_\tau(\beta) \equiv \sum_{i=1}^{n} \rho_\tau\left(Y_i - \underline{x}_i'\underline{\beta}\right) \tag{6}$$

by varying $\underline{\beta}$ we obtain the regression quantiles. Note that the estimated regression quantile parameters implicitly depend on the probability, τ. In particular, the *j*th coordinate of $\hat{\underline{\beta}}(\tau)$ gives the predicted marginal effect of a unit change in the *j*th explanatory variable, $x^{(j)}$, on the conditional τth-quantile of the response.

If the model predicts that the $\underline{\beta}$ coefficients change with τ, then we have evidence of heterogeneity in the population. This heterogeneity is often the result of unequal variances (heteroscedasticity).

Throughout the analysis which follows we will make use of the R library crq. This library contains a function which allows the user to fit censored regression quantile models and assess the extent of heterogeneity in the covariate effect over the range of claim durations.

3. Regression Quantiles and Claim Termination Rates

The aim of this section is to illustrate the application of censored regression quantiles (Portnoy 2003), to claim termination rates for IP insurance. The heterogeneity of the effect of the covariates age, occupation and deferred period across the distribution of claim durations will be analysed. The benefits of using censored regression quantile analysis as compared to more traditional Cox regression in this context will also be explored.

The potential output from censored regression analysis can be extremely voluminous. This issue arsies due to the wide range of possible conditional quantile curves that can be estimated. In order to make the interpretation of results simpler, we restrict ourselves in this paper to the consideration of the effect of age, occupation class (C or D vs A or B), see Pitt(2006) for a description of occupation classes, and deferred period (greater than or equal to 28 days or less than 28 days).

A censored regression quantile model was fit using the entire dataset of claim durations described in Pitt(2006). The R command used to fit the model is

```
crq(Surv(log(durn3),terminate) ~ age + occupnew + defpdnew, data = termrates2),
```
(7)

where occupnew is an indicator variable for occupation classes C and D, and defpdnew is an indicator variable for deferred period in excess of 27 days.

Mathematically, the form of the fitted censored regression model is

$$\log(\text{Time to return to work}) = \beta_0 + \beta_1(\text{Age}) + \beta_2(\text{Occupation Class}) + \beta_3(\text{Deferred Period}), \tag{8}$$

where separate models of the above form are fit to quantiles corresponding to breakpoints in the claim duration data.

To consider the impact of age on the log of claim duration we create a graph of the predicted censored regression quantile relationship between log of duration and age. See Figure 1.

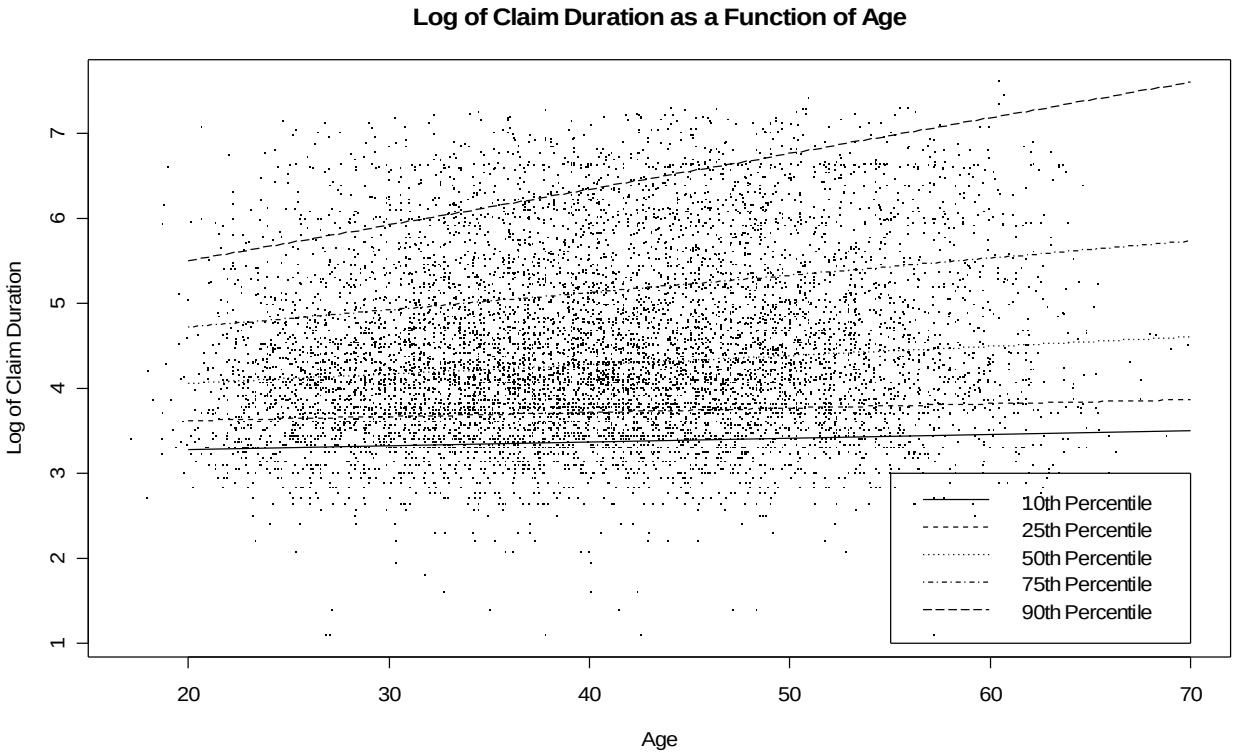

Figure 1 Log of Claim Duration as a Function of Age for Five Different Percentiles

Immediately obvious from Figure 1 is the increasing slope of the regression lines at higher percentiles of the distribution of log claim duration. This suggests that the age sensitivity of (the log of) claim duration is greater for longer duration claims. In particular, the effect of increasing age increases the 90th percentile of the distribution of claim durations for a given age much more than the same increase in age increases the 10th percentile of the distribution of claim durations. The regression coefficients for age, occupation class and deferred period for various percentiles are given below, in Table 1. Note that the effect of occupation class on claim duration also varies significantly with the percentile of the distribution being considered. In particular, the effect of being in occupation class C or D in reducing the predicted duration of disability is more pronounced at the higher percentiles of the distribution of the claims distribution.

Percentile	Censored Regression Quantile Coefficient (Age)	Censored Regression Quantile Coefficient (Occupation)	Censored Regression Quantile Coefficient (Deferred Period)
10th Percentile	0.00419	-0.01846	0.60475
25th Percentile	0.00494	-0.06772	0.59322
50th Percentile	0.01068	-0.12023	0.59366
75th Percentile	0.02048	-0.23410	0.71315
90th Percentile	0.04206	-0.29936	0.62723

Table 1 Censored Regression Quantile Coefficients

The above results have clear implications for the determination of the disabled life reserve (DLR). This quantity is the reserve held by an insurer in respect of an insured who is currently claiming benefits at the date of the valuation. Insurers will always have a material proportion of their portfolio relating to insured lives who are currently disabled and who have been disabled for a reasonable period of time at the date of valuation. The insurer is required to determine the amount of money that needs to be held in respect of these disabled and insured lives at a particular instant in time. Clearly the amount of money required depends on the future disability status of the insured life. Table 1 shows that the effect of the insured being older or of being in occupation class A or B on the claim duration is more significant for longer duration claims; that is, for claims that have extended into higher percentiles of the claim duration probability distribution.

A model considered in Pitt(2006) that is often used in survival analysis is the Cox Regression model, (Cox, 1972). This model estimates the impact of rating factors such as age and occupation class on the dependent variable, claim duration, by considering the impact of these rating factors collectively across all quantiles of the claim duration distribution. It is therefore of interest to assess the difference in the predicted sensitivities of claim duration to each of the insured characteristics from the Cox model and the censored regression quantile method.

4. Comparison of Cox Regression and Censored Regression Quantiles for Claim Termination Rates

To begin this section, we fit a Cox regression model to our data using age, occupation (class C or D indicator) and deferred period (greater than 27 days indicator) to the claim duration data. The dependent variable is the log of the claim duration and the usual right censoring in the data is used within the analysis. The output for this regression model is given below in Table 2.

coxph(formula = Surv(log(durn3), terminate) ~ age + occupnew + defpdnew, data = termrates2)					
	coef	**exp(coef)**	**se(coef)**	**z score**	**p-value**
age	-0.0124	0.988	0.00119	-10.40	0.0e+00
occupnew	0.1285	1.137	0.02524	5.09	3.5e-07
defpdnew	-0.4668	0.627	0.02482	-18.80	0.0e+00
Likelihood ratio test=631 on 3 df, p=0 n= 8863					

Table 2 Censored Regression Quantile Output

The three rating factors are clearly highly statistically significant and the overall model indicates that age, occupation class and deferred period are jointly statistically significant.

In order to compare the Cox regression model to the censored regression quantile model, it is necessary to compare the predicted sensitivities of the quantiles of the claim duration distribution under the two models. For the Cox model, we have that the predicted hazard function for the *i*th individual in the sample, $h_i(t)$, is

$$h_i(t) = h_0(t)e^{x_i\beta}, \quad i = 1,...,n, \tag{9}$$

where $h_0(t)$ is the baseline hazard function. Given this form for the hazard function, the survival function can be written as

$$S_i(t) = \exp\left(-H_0(t)e^{x_i\beta}\right), \tag{10}$$

where $H_0(t) = \int_0^t h_0(s)\,ds$.

So the conditional quantile for claim duration, *T*, at $\underline{x}$ becomes

$$Q_{\text{Cox}}(\tau \mid x) = H_0^{-1}\left(-\log(1-\tau)e^{-\underline{x}_i\beta}\right). \tag{11}$$

The quantity $Q_{\text{Cox}}(\tau \mid x)$ defined in Portnoy (2003) is therefore the predicted time since claim inception, under the Cox Regression model, when a proportion τ of those insureds who claim from their IP insurance contract will have returned to work. The censored quantile regression coefficients give the predicted change in various quantiles of the distribution of the log of claim duration when various rating factors are increased by one unit. It is therefore possible to directly compare the coefficients estimated using censored regression quantiles with the derivative of the expression at (11). Consequently we compare the $\hat{\beta}(\tau)$ with the quantity

$$\frac{\partial}{\partial x} Q_{\text{Cox}}(\tau \mid x) = \frac{\partial}{\partial x} H_0^{-1}\left(-\log(1-\tau)e^{-\underline{x}_i\beta}\right). \tag{12}$$

To calculate the above derivatives we need to use numerical differentiation owing to the irregularities present in the inverse cumulative baseline hazard function, $H_0^{-1}(t)$. For the calculation of this derivative, we use the model in Table 2. From this model we calculate the fitted claim continuance probabilities at each of the times that a person in the sample returns to work for a life aged 40.53, the mean of the ages in the sample. Denote these values $S_1(t_i)$. We also use the model in Table 2 to calculate the fitted claim continuance probabilities for a life aged 41.53 (one plus the mean of the ages in the sample). Denote these values $S_2(t_i)$. Next we determine the log of claim duration that corresponds to each of the values of $S_1(t_i)$ for a life aged 41.53. These values are calculated using linear interpolation and the S-Plus function, out2, which performs this calculation (amongst other calculations) is given in full in Appendix A. It is then straightforward to numerically estimate the predicted quantile sensitivity based on the Cox Regression model. The difference between the survival times for a given quantile of the log claim duration distribution estimates the sensitivity of various quantiles of the log of claim duration distribution under the Cox Model.

It is useful to compare the sensitivities of various quantiles of the log of claim duration distribution from the use of censored regression quantiles and the more conventional Cox regression model. Figures 2 and 3 compare these quantile sensitivities for changes in age and occupation class. The sensitivity labelled on the y-axis of the graphs in Figures 2 and 3 refers to the expression in (12).

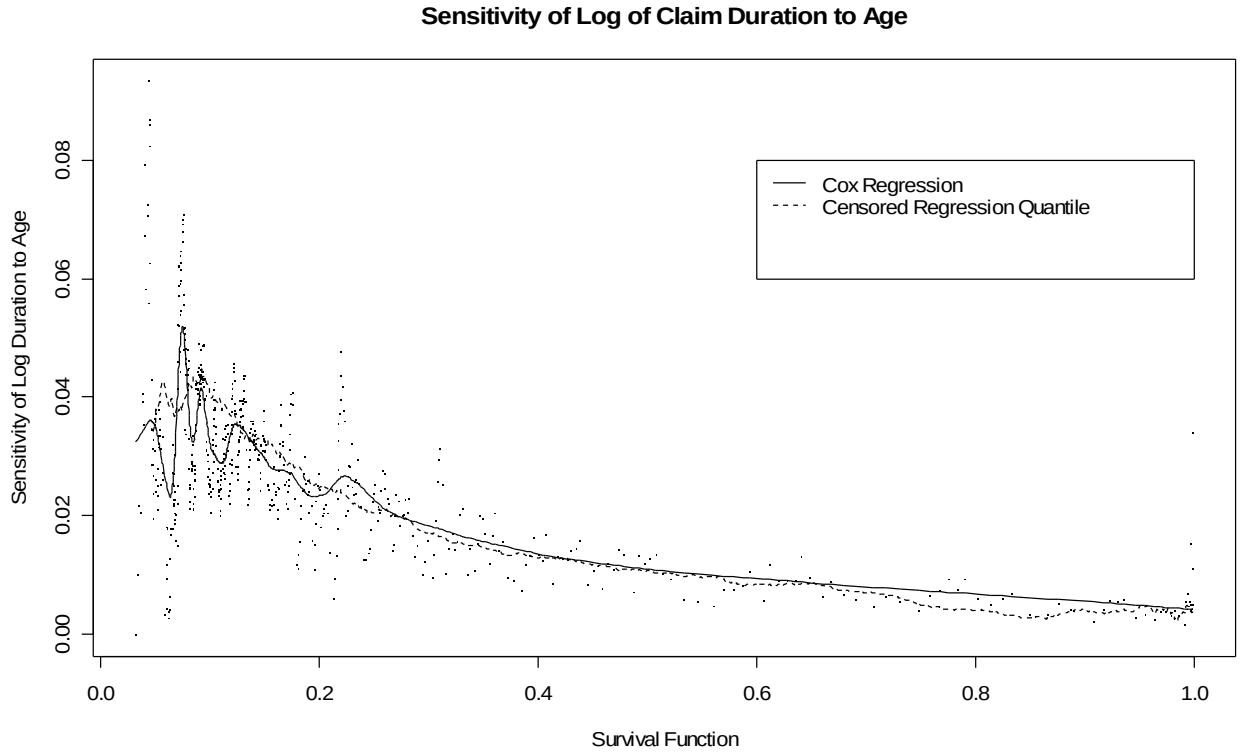

Figure 5.2 Comparison of Sensitivities of Log Duration to Age

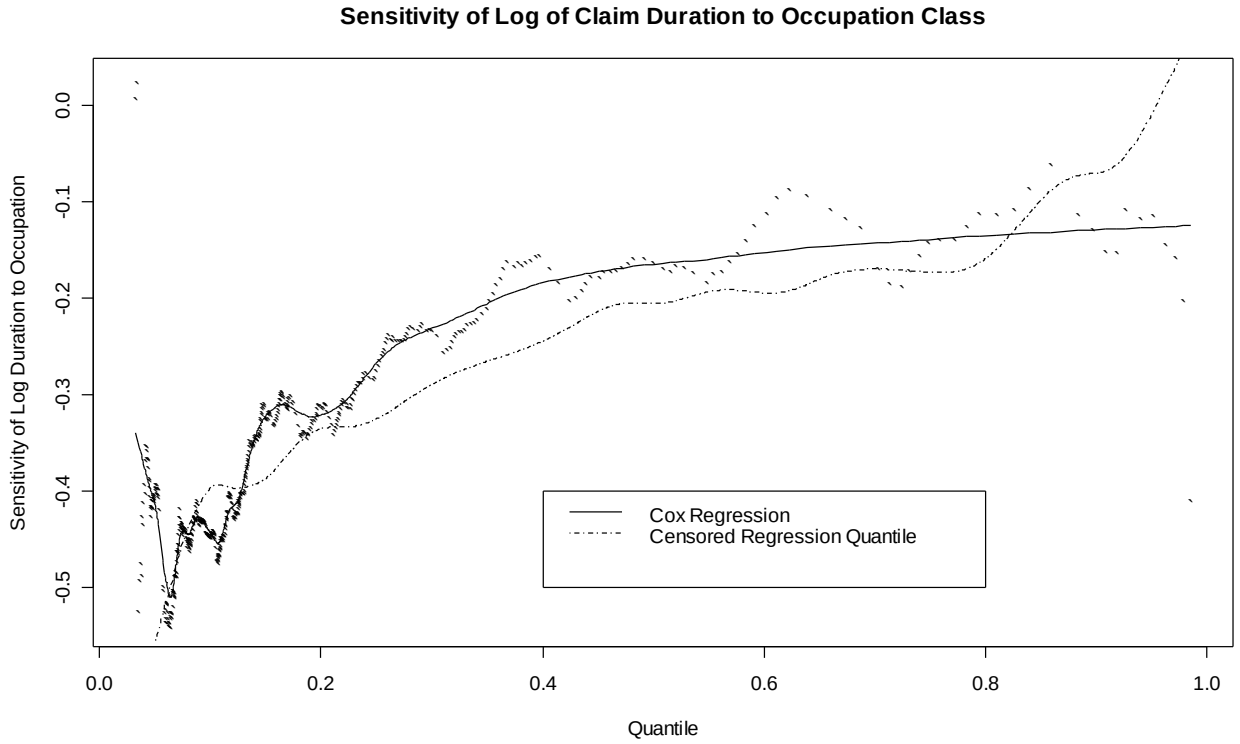

Figure 3 Comparison of Sensitivities of Log Duration to Occupation Class

From Figure 2 it is clear that the predicted age sensitivity of lower quantiles (at higher levels of the survival function) is higher under the Cox model than for the censored regression quantile analysis. There is considerable variability in the censored regression quantile coefficients for the higher quantiles (survival function between 0 and 0.2). This volatility is primarily due to the small number of claims that are still continuing at these claim durations.

From Figure 3, there is a clear bias in the estimation of the quantile sensitivities of log of claim duration to occupation class for the Cox regression model. The Cox regression model predicts a greater reduction in claim duration for occupation classes C and D than the censored regression quantile analysis over most of the range of the log of claim duration distribution. This result is driven in part by the inappropriateness of the proportional hazards assumption in the case of the disability data being considered here that underlies the Cox Regression model.

5. Assessing the Comparison between Cox Regression and Censored Regression Quantiles using Subsampling

Figure 2 also clearly demonstrates that using Cox regression alone can lead to incorrect conclusions about the age sensitivity of the log of claim duration particularly for shorter duration claims. It is of interest to see whether the disparity between predicted age sensitivities of claim duration between the two approaches is likely to occur with most sets of income protection insurance data or whether the difference is more a feature of the particular set of Australian industry claim duration data that is being analysed.

To explore this, we consider a subsampling approach whereby 84 different datasets, each of size 400, chosen from the original set of 8863 data points. These 84 different datasets contain records 1 to 400, 101 to 500, …, 8401 to 8800. Since the data is in no particular order, with respect to the variable of interest namely claim duration, this is similar to analyzing 84 different sets of randomly chosen disability income insurance claim duration data each of size 400 records.

For each of these datasets of size 400, we fit both a Cox regression model, equivalent to the model in Table 2, and also a censored regression quantile model. These models both use age, occupation class and deferment period as the only covariates.

We then compare the censored regression quantile age coefficient for each of the 400 models to the Cox regression age sensitivities for a range of quantiles. We are interested in assessing the absolute difference between the censored regression quantile coefficients and the Cox regression quantile derivative function. In order to make the comparison more straightforward, we averaged the Cox regression quantile sensitivities over survival function bands, namely [0,0.2), [0.2,0.4), [0.4,0.6),[0.6,0.8) and [0.8,1.0]. We also averaged the censored regression quantile function over the same bands for the survival function. The difference in the mean sensitivities for each of the 84 models were then calculated. A density of these differences was then created. The program which performs this subsampling is the S-Plus function, out3, shown in full in Appendix A.

The density for the difference in mean sensitivities for age and Survival Function in excess of 0.8 is shown below. The mean of the average differences between the sensitivities is 0.0008570 which is 20.1% of the censored regression quantile sensitivity. Hence the Cox Regression predicts an age sensitivity of log of claim duration that is 20.1% higher than the censored regression quantile method for the shortest 20% of claim durations.

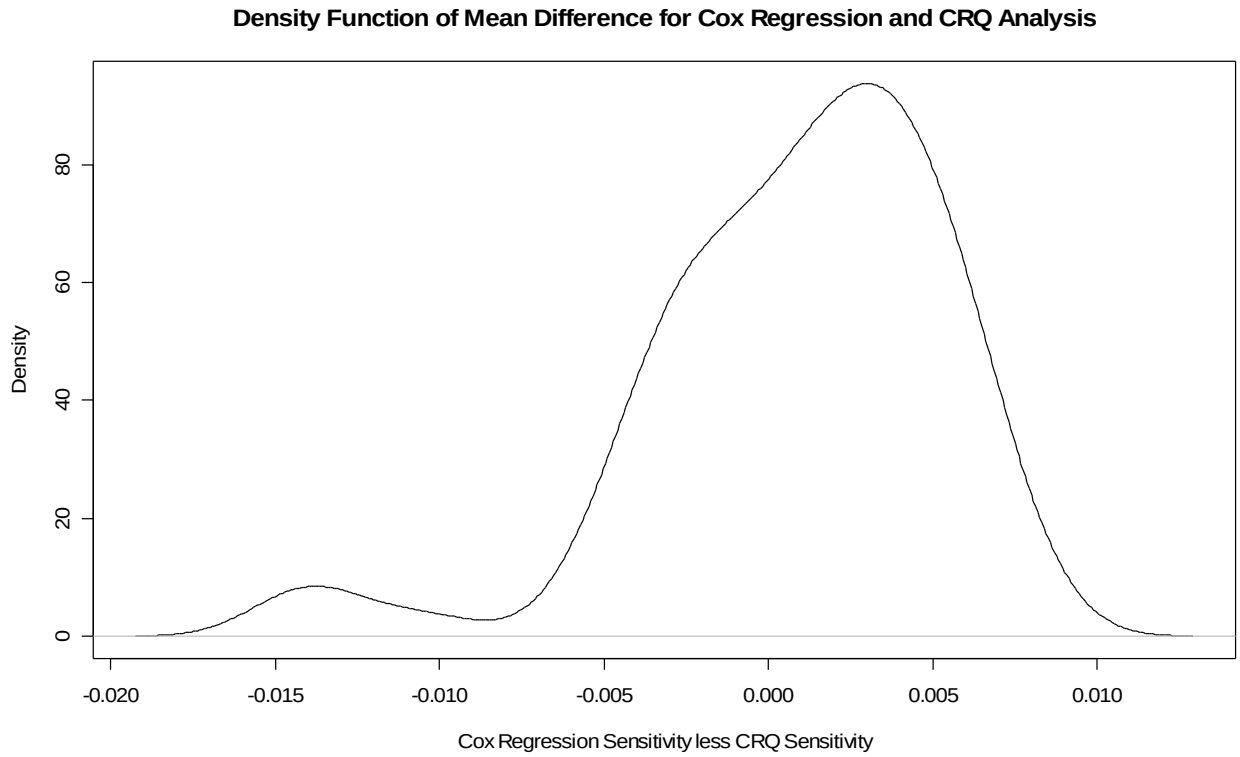

Figure 4 Density Function of Mean Difference for Cox Regression and CRQ Analysis for the Survival Function on the range [0.8,1.0]

Similarly, Figure 5 on the following page shows an empirical density function of mean differences in predicted quantile sensitivities from Cox Regression and censored regression quantiles over the [0.6,0.8) band of the survival function. The mean difference is 0.0024, or 35.6% of the censored regression quantile analysis. This finding indicates again that Cox Regression predicted sensitivities of the return to work hazard rate to covariates are consistently higher than their censored regression quantile counterparts over the 60% to 80% region for the claim duration survival function.

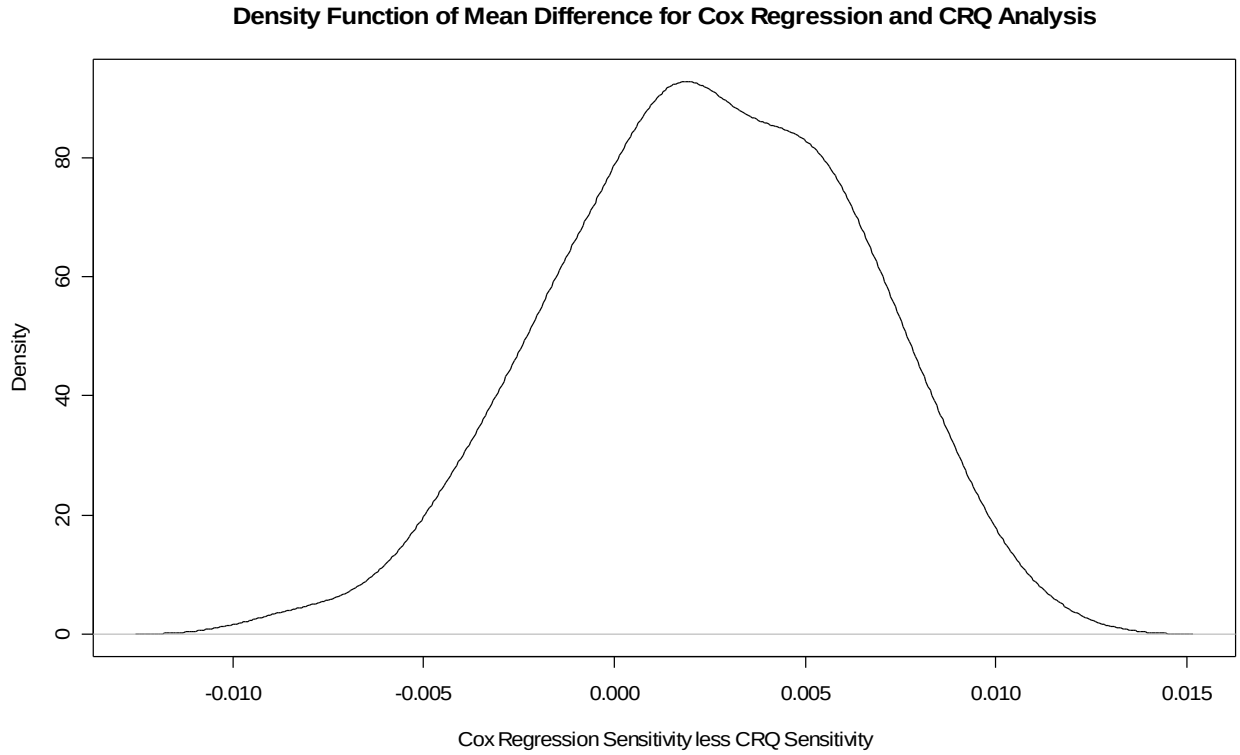

Figure 5 Density Function of Mean Difference for Cox Regression and CRQ Analysis for the Survival Function on the range [0.6,0.8)

This chapter has demonstrated an additional technique that can be used to detect heterogeneity in claim duration data. Censored regression quantiles therefore provide a more reliable method for assessing the impact of covariates in the tail of the probability distribution of claim durations than do other more commonly adopted methods from survival analysis, such as Cox Regression.

6. Conclusion

This paper has demonstrated the value of censored regression quantile analysis in the valuation of outstanding claims for IP insurance claims. In particular the paper has highlighted the magnitude of the distortion that Cox Regression analysis, when used inappropriately with IP insurance claims data, can have on predicted regression coefficients.

References

Australian Prudential Regulation Authority GPS 210 (2005) Liability Valuation for General Insurers. General Insurance Prudential Standard.

Australian Prudential Regulation Authority, (2001) General Insurance Reform Act.

Booth, P.; Chadburn, R.; Cooper, D.; Haberman, S.; James, D. (2000) Modern Actuarial Theory and Practice, *Chapman and Hall, CRC*, London.

Galton, F. (1889) Natural Inhertiance, *Macmillan*, London.

Insurance and Superannation Commission, (1973) Insurance Act.

Koenker, R. and Bassett, G. (1978) Regression quantiles, *Econometrica,* Vol.46, pp. 33-50.

Pitt, DGW (2006) Modeling the Claim Duration of Income Protection Insurance Policyholders using Parametric Mixture Models. University of Melbourne Centre for Actuarial Studies Working Paper Series No. 132.

Portnoy, S. (2003) Censored regression quantiles, *Journal of the American Statistical Assocation,* pp.1001-1012.

Appendix A -Selected S-Plus Functions

> out2

```
function(a) {
        for(i in 3:(length(coxcrv$time)-1)) {
                A<<-max(coxcrv1$surv[coxcrv1$surv<coxcrv$surv[i]])
                B<<-min(coxcrv1$surv[coxcrv1$surv>coxcrv$surv[i]])
                C1<<-coxcrv1$time[coxcrv1$surv==A]
                D1<<-coxcrv1$time[coxcrv1$surv==B]
                timeout[i]<<-log(D1)*(coxcrv$surv[i]-A)/(B-A)+log(C1)*(B-coxcrv$surv[i])/(B-
        A)
                derivcvout[i]<<-timeout[i]-log(coxcrv$time[i])
                E1=length(coxcrv$surv[coxcrv$surv>0.8])
                F1=length(coxcrv$surv[coxcrv$surv>0.6])-E1
                G1=length(coxcrv$surv[coxcrv$surv>0.4])-(E1+F1)
                H1=length(coxcrv$surv[coxcrv$surv>0.2])-(E1+F1+G1)
                I1=length(coxcrv$surv[coxcrv$surv>0])-(E1+F1+G1+H1)
                E1mean<<-mean(derivcvout[1:E1])
                F1mean<<-mean(derivcvout[(E1+1):(E1+F1)])
                G1mean<<-mean(derivcvout[(E1+F1+1):(E1+F1+G1)])
                H1mean<<-mean(derivcvout[(E1+F1+G1+1):(E1+F1+G1+H1)])
                I1mean<<-mean(derivcvout[(E1+F1+G1+H1+1):(E1+F1+G1+H1+I1)])
        }}
```

> out3

```
function(a) {
        for(j in 1:84) {
                durn3a<<-durn3[(100*(j-1)+1):(100*j+300)]
                terminatea <<- terminate[(100*(j-1)+1):(100*j+300)]
                agea <<- age[(100*(j-1)+1):(100*j+300)]
                occupnewa<<-occupnew[(100*(j-1)+1):(100*j+300)]
                defpdnewa<<-defpdnew[(100*(j-1)+1):(100*j+300)]
                tempcox<<-
                coxph(Surv(durn3a,terminatea)~agea+occupnewa+defpdnewa,data=termrates2)
                tempcrq<<-
                crq(Surv(log(durn3a),terminatea)~agea+occupnewa+defpdnewa,data=termrates2)
                coxcrv<<-summary(survfit(tempcox,newdata=temp))
                coxcrv1<<-summary(survfit(tempcox,newdata=temp1))
                out2(5)
                E1means[j]<<-E1mean
                F1means[j]<<-F1mean
                G1means[j]<<-G1mean
                H1means[j]<<-H1mean
                I1means[j]<<-I1mean
                M1<<-length(tempcrq[tempcrq$sol[1,]<0.2])
                N1<<-length(tempcrq[tempcrq$sol[1,]<0.4])-M1
                O1<<-length(tempcrq[tempcrq$sol[1,]<0.6])-(M1+N1)
                P1<<-length(tempcrq[tempcrq$sol[1,]<0.8])-(M1+N1+O1)
                Q1<<-length(tempcrq[tempcrq$sol[1,]<1])-(M1+N1+O1+P1)
                M1mean<<-mean(tempcrq$sol[3,(1:M1)])
```

```
                N1mean<<-mean(tempcrq$sol[3,((M1+1):(M1+N1))])
                O1mean<<-mean(tempcrq$sol[3,((M1+N1+1):(M1+N1+O1))])
                P1mean<<-mean(tempcrq$sol[3,((M1+N1+O1+1):(M1+N1+O1+P1))])
                ifelse(M1+N1+O1+P1+1<=length(tempcrq$sol[3,]),Q1mean<<-
                mean(tempcrq$sol[3,((M1+N1+O1+P1+1):(M1+N1+O1+P1+Q1))]),Q1mean<<-
                0)
                M1means[j]<<-M1mean
                N1means[j]<<-N1mean
                O1means[j]<<-O1mean
                P1means[j]<<-P1mean
                Q1means[j]<<-Q1mean
                R1means[j]<<-E1means[j]-M1means[j]
                S1means[j]<<-F1means[j]-N1means[j]
                T1means[j]<<-G1means[j]-O1means[j]
                U1means[j]<<-H1means[j]-P1means[j]
                V1means[j]<<-I1means[j]-Q1means[j]
}}
```